journalism and media

MDPI

Article

Combating Hate Speech on Social Media: Applying Targeted Regulation, Developing Civil-Communicative Skills and Utilising Local Evidence-Based Anti-Hate Speech Interventions

Stefanie Pukallus [1,*] and Catherine Arthur [2]

[1] School of Journalism, Media and Communication, University of Sheffield, Sheffield S10, UK
[2] Humanitarian and Conflict Response Institute, University of Manchester, Manchester M13, UK; carthur02@qub.ac.uk
* Correspondence: s.pukallus@shef.ac.uk

Abstract: Social media platforms such as Facebook and X (formerly Twitter) set their core aim as bringing people and communities closer together. Yet, they resemble a digital communicative battleground in which hate speech is increasingly present. Hate speech is not benign. It is the communicative driver of group oppression. It is therefore imperative to disarm this digital communicative battlefield by (a) regulating and redesigning social media platforms to prevent them from playing an active and enabling role in the dissemination of hate speech and (b) empowering citizen-users and local civil associations to recognise and actively counter hate speech. This top-down and bottom-up approach necessarily enforces responsibility and builds capacity. This requires that we adapt and combine three aspects of communicative peacebuilding: first, the (re)building of civil-communicative institutions; second, the use of digital citizenship educational programmes to support the development of civil-communicative skills for using social media; and third, the identification and use of local civil capacity and knowledge, which manifests in the present context in the use of local evidence-based anti-hate speech interventions. We argue that this interdisciplinary combinatorial approach has the potential to be effective because it combines two things: it places responsibility on relevant actors to both make social media safer and to navigate it harmlessly and responsibly; and it simultaneously helps build the capacity for actively identifying and countering hate speech in civil societies.

Keywords: hate speech; social media; digital communication; peacebuilding

check for updates

Citation: Pukallus, Stefanie, and Catherine Arthur. 2024. Combating Hate Speech on Social Media: Applying Targeted Regulation, Developing Civil-Communicative Skills and Utilising Local Evidence-Based Anti-Hate Speech Interventions. *Journalism and Media* 5: 467–484. https://doi.org/10.3390/journalmedia5020031

Academic Editor: Andreu Casero-Ripollés

Received: 17 November 2023
Revised: 8 March 2024
Accepted: 25 March 2024
Published: 7 April 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction: Social Media as a Communicative Battleground

Recent academic attention has focused on how social media[1] contributes to division, conflict and violence within society, and how it increases polarisation by reaffirming and amplifying certain views—often with the help of algorithms and bots (Aral 2021; Bar-Ilan 2007; Howard 2020; Noble 2018). Polarisation can be seen as a "natural response to being attracted to similar (positive) things and repulsed by different (strange) [things]" (Coleman 2021, p. 221). However, it becomes a social problem when it leads to groups on opposed poles no longer engaging with each other in a non-violent manner. Polarisation—whether affective, ideological, political, social, perceptual or identity-based (Klein 2020)—thrives on inter-group bias (see Mason 2018), animated by an "us vs. them" mentality (Schmitt [1932] 2007; Mouffe 2005). When this mentality becomes extreme, it can take the form of political tribalism in which group identity and interest trump individual identity and self-interest, even to the point of violence (e.g., Littman and Paluck 2015). Importantly, "the tribal instinct is not just an instinct to belong, it is also an instinct to exclude" (Chua 2018, p. 1). This instinct to exclude or oppress "the enemy" plays out loudly in polarised spaces on social media in the form of coarsened public language, the circulation of extremist views, conspiracy theories and disinformation as well as the rejection of truth and facts. This is facilitated by social media's affordances which are designed to help tech companies achieve

their monetary goals by attracting ever-increasing audiences and gathering ever-more data. The equation of social media companies is simple: the more noise, the more people pay attention, sign up to social media platforms, click and look at posts, like and share. This in turn expands viewership and makes content go viral, boosting social media companies' monetary gains. In short, social media platforms favour and facilitate the circulation of forms of "deviant communication as it violates shared cultural standards, rules, or norms of social interaction in social group contexts" (Castaño-Pulgarín et al. 2021, p. 1) for monetary gains and accept that social media platforms become digital communicative battlegrounds complicit in group oppression through hate speech. Indeed, it is precisely on these digital communicative battlegrounds that hate speech is for the first time artificially and mechanically amplified and disseminated with unprecedented speed, reach and force by an unseen variety of digital military actors, supported by algorithmic megaphones (Reich et al. 2023). It is these that provide users—human, automated and hybrid—with an immense and immediate power to attack the dignity of other users and to "intimidate targeted groups from participating in the deliberative process", which in turn "stymies the depth of pluralistic speech" (Tsesis 2009, p. 499) and leads to harm in the form of group oppression.

In order to prevent such group oppression, we argue that it is imperative to disarm this digital communicative battlefield by (a) regulating and redesigning social media platforms to prevent them from playing an active and enabling role in the dissemination of hate speech and (b) empowering citizen-users and local civil associations to recognise and actively counter hate speech. To do so, we recommend considering the prevention of hate speech as a communicative peacebuilding task and propose a new combinatorial and interdisciplinary approach that combines three activities: first, the (re)building of civil-communicative institutions; second, the use of digital citizenship educational programmes to support the development of civil-communicative skills for using social media; and third, the identification and use of local civil capacity and knowledge, which manifests in the present context in the use of local evidence-based anti-hate speech interventions. Before we turn to this disarming approach, we need to define what we mean by hate speech and explain how it contributes to and enables group oppression.

2. Hate Speech on Social Media as a Means of Group Oppression

2.1. Defining Hate Speech on Social Media

Amidst the various definitions of hate speech, one can delineate three broad definitional approaches. The first defines hate speech in the most expansive way. For example, Carlson (2021, p. 2) adopts a wide definition of hate speech and stipulates that it "includes broad categories of speech, including racism, anti-semitism, homophobia, bigotry against the disabled, political hatred, rumourmongering, misogyny and violent pornography, promotion of terrorism, cyberbullying, harassment, stalking, and the sale and promotion of online products" (also, Gagliardone et al. 2015). The second definitional approach to hate speech focuses on the target of hate speech: specific groups. For example, Davidson et al. (2017, p. 1) argue that "Language that is used to expresses [sic] hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group". Similarly, Cohen-Almagor (2011, p. 1f.) emphasises the target of hate speech and accordingly defines hate speech "as bias-motivated, hostile, malicious speech aimed at a person or a group of people because of some of their actual or perceived innate characteristics. It expresses discriminatory, intimidating, disapproving, antagonistic, and/or prejudicial attitudes toward those characteristics, which include gender, race, religion, ethnicity, color, national origin, disability, or sexual orientation". The third definitional approach, in turn, engages primarily with the aspect of "hate" and the intention of harming the target of hate speech. Hate can be defined as an "extreme, and continuous emotion that is directed at a particular individual or group and denounces them fundamentally and all-inclusively" (Halperin 2011, p. 26). Correspondingly, hate speech is "associated with the aspiration to harm the outgroup as much as possible, and it can lead people to desire total elimination of

the hated outgroup" (ibid.; also, Royzman et al. 2005; Allport 1958; Williams 2021). These three definitional approaches are reflected in the broad definition of hate speech adopted by the United Nations, which states that hate speech is "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" (United Nations n.d.). Crucially, the UN notes that a defining characteristic of this type of speech is the fact that "[it] may threaten social peace" (ibid.), delineating the connections between hate speech, forms of violence and conflict, and peace. Scholars taking this definitional approach focus often on attack, physical harm, discrimination and violence and are therefore more concerned with the potential consequences of hate speech (Waldron 2012; Zhang and Luo 2019; Papcunová et al. 2023).[2]

What underpins these three definitional approaches to hate speech within the realm of social media is a two-fold architecture of mechanical design and purposive intentions of content. With regard to the former, social media is constructed and engineered in such a way that allows (even if originally unintentionally) for the free, rapid circulation and amplification of hateful content. This includes misinformation, rumours, insults, satire, mockery, misogyny, religious intolerance, dehumanisation, stereotyping, generalisations and prejudice. Communicatively speaking, hate speech "is best understood as comprising three communicative dimensions: styles of public communication, rhetorical devices of hate speech and their communicative vehicles for dissemination" (Pukallus 2024b). What this means is that they use a specific style of public communication—pseudo-scientific, affective-emotional, commanding-assertive, pseudo-deliberative rational or teleological-prophetic—to get their message of hate across. Any of these styles can be adapted for use on social media and then be combined with rhetorical devices such as narratives, cause–effect scenarios, coded language, metaphors or visual elements such as memes, humour and satire amongst others. Memes in particular have been adopted by many hate speakers on social media as they are more difficult to detect through content moderation and disguise hate in humour.[3] Finally, hate speech on social media can use a variety of communicative vehicles including the performative and visual arts. These three communicative dimensions of hate speech can be universally defined but will be variously interpreted and combined depending on the targets of hate speech, its context and its intended audience.[4] It is this variety that makes hate speech so effective and attention-grabbing and also increasingly difficult (yet not impossible) to detect and take action against.

The mechanical design of social media facilitates the achievement of the second architectural feature of online hate speech: its purposive intentions. These aim to render members of target groups vulnerable to harm in two ways: first, by attacking their dignity, which Waldron (2012, p. 5) defines as a person's "social standing, the fundamentals of basic reputation that entitle them to be treated as equals in the ordinary operations of society".[5] Attacks on the individual's dignity for belonging to a certain group ultimately threaten the dignity of all group members and make them vulnerable to attacks on their group's reputation (group libel) through denigration based on opinions disguised as facts, disputing the normative basis of equal standing, dehumanisation (see below) and the use of slogans and instructions to degrade members of a group (Waldron 2012). Second, these attacks on individual and group dignity in turn undermine groups' assumed "ontological security". They are forced to live in a state of insecurity and are discursively placed outside the moral universe of a society (Hagan and Rymond-Richmond 2008; Opotow 1990), which means that they can become victims of violence (see Allport 1958; Williams 2021) and be excluded from civil and political life with impunity (see also, Harrison and Pukallus 2018).

In short, hate speech can be defined, for all practical and operational purposes, as mechanically enabled and amplified language with the intention to harm. The most significant group harm that can be brought about by hate speech is that of group oppression as conceived of by Iris Marion Young.

2.2. Group Oppression

For Young (1990), oppression needs to be understood as structural and systematic, with its causes "embedded in unquestioned norms, habits, and symbols, in the assumptions underlying institutional rules and the collective consequences following those rules" (Young 1990, p. 41). For her, the group element is fundamentally important because group "meanings partially constitute people's identities in terms of the cultural forms, social situation, and history that group members know as theirs because these meanings have been either forced upon them or forged by them or both" (Young 1990, p. 44)—they "are real (. . .) as forms of social relations" (ibid.). Hate speech fosters oppression because it supports "a conceptualization of group difference in terms of unalterable essential natures that determine what group members deserve or are capable of, and that exclude groups so entirely from one another that they have no similarities or overlapping attributes" (Young 1990, p. 47). Young identifies five faces of oppression: exploitation, marginalisation, powerlessness, cultural imperialism (where dominant groups and meanings render certain groups' perspectives invisible and negatively stereotype them) and violence thereby seeking to ensure that all aspects of social life are subordinated to a dominant ideological view.[6]

It is clear then that hate speech on social media can act as linguistic violence and therefore a communicative weapon and driver of group oppression, which undermines diversity and plurality of voices in society, opportunities for inclusive community building, as well as the achievement of international policy agendas such as the Sustainable Development Goals (SDGs) that depend on inclusivity and equality. Equally and in the light of the above, it is fair to argue that hate speech on social media can be seen as an attack on the public peace—society's ability to cooperate peacefully despite differences and disagreements and to grant dignity to its members. It is precisely for its potential to undermine non-violent association and cooperation that we argue that interdisciplinary peacebuilding bears genuine potential to be successful in combating hate speech online.

3. A Peacebuilding Combinatorial Approach to Combating Hate Speech on Social Media

If we understand social media as a digital battlefield in which hate speech is a communicative weapon, then effectively combating hate speech and by extension group oppression requires some form of digital decommissioning via a peacebuilding approach. We argue that we should adapt and apply three aspects of communicative peacebuilding—the process of building peaceful relations in a society through the transformative capacity of communication (Pukallus 2022, 2024a)—in an interdisciplinary combinatorial approach. The first aspect is the (re)building of civil-communicative institutions. This requires a change from the top down in the design of social media platforms to disable the free circulation and amplification of hateful content. This necessitates targeted regulation of the mechanics of social media platforms. Second, the use of digital citizenship educational programmes to support the development of civil-communicative skills from the bottom up for using social media in non-hateful and therefore non-harmful ways. Third, the identification and use of local civil capacity and knowledge, which manifests in the present context in the use of local evidence-based anti-hate speech interventions on social media. Each of these three aspects is insufficient to effectively combat hate speech on its own but, combined, they allocate diverse responsibilities to three different groups of agents in both universal and particular contexts—the enablers and carriers of hate speech, the hate speaker(s)[7] and the resisters of hate speech—and build capacity for resilience in the latter two groups.

3.1. Aspect 1: Targeted Regulation of the Mechanics of Social Media Platforms

The aim of this first aspect of our suggested combinatorial and interdisciplinary peacebuilding approach to combating hate speech endorses a specific kind of targeted regulation in order to firmly place responsibility on tech companies for the design and functioning of their platforms, to increase their accountability and to find mechanisms that enable them to uphold their community standards in a routinised manner. It is fair to

assume that social media companies are aware of the dangers inherent in hate speech that is allowed to freely circulate. Both Facebook (Meta) and X (formerly Twitter) set community standards that explicitly address the issue of hate speech. X's (2023) community standards state that (and it is worth quoting at length): "You may not directly attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease" (X 2023) and show awareness of the dangers of hate speech outlined above, including attacks on dignity and the stifling of a plurality of voices. Indeed, X (2023) adds:

> We recognize that if people experience abuse on X, it can jeopardize their ability to express themselves. Research has shown that some groups of people are disproportionately targeted with abuse online. For those who identify with multiple underrepresented groups, abuse may be more common, more severe in nature, and more harmful.

The statement further adds that X is firmly committed "to combating abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalized" (X 2023). In a similar vein, Facebook (n.d.) defines

> hate speech as a direct attack against people (. . .) on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanising speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation

and makes clear that Facebook removes hate speech because it "creates an environment of intimidation and exclusion, and in some cases may promote offline violence" (Facebook n.d.).

However, thus far tech companies' community standards have only led to a reactive and sporadic rather than systematically pro-active approach to tackling hate speech—even in the context of the recent drastic increase in hate speech on X following Elon Musk's takeover of what was formerly Twitter (See for example Frenkel and Conger (2022)). For example, a report by the Center for Countering Digital Hate (2023, cf. p. 5) shows that X failed to adhere to its own community standards by continuing to host posts that had been reported and identified as hate speech. Of the 300 posts that included racist caricatures of Black and Jewish people, engaged in Holocaust denial, that labelled Hitler a hero or memes that cast Black people as inherently violent, 86% remained online and 90% of the accounts they came from remained active. These posts were judged to clearly violate the community standards referred to above and yet, X did not take these posts down nor did it sanction users for the violation of their user agreement. Though this is just one example, it is indicative of a sector-wide failure to moderate hate speech effectively or prevent it from circulating in the first place through changes to the design of the platforms.

The passivity of these tech companies—Facebook/Meta and X are only two examples—and their unwillingness to compromise on capitalist ambition for user safety and dignity mean that these companies need to be legally forced to systematically uphold their own community standards, rather than being allowed to self-regulate. As Ressa (2023, p. 74) points out, tech companies have "largely abdicated [any] gatekeeping role of protecting facts, truths and trust. (. . .) Now, under the technology companies, the information you get is directly determined by the corporation's drive for profit". This is "because their incentive system is built around power and money" (ibid.). And correspondingly, self-regulation has thus far been inadequate and sporadic at best. Legal obligations to uphold standards should include accountability for the ways in which social media platforms are designed and the kind of content their design enables and amplifies in order to fulfil their monetary targets. It is worth noting that this problem has been recognised as serious to the extent that the European Union has drawn up legislation to address the need for more stringent

regulation of tech companies to address their prioritising profit over fact. The EU Artificial Intelligence Act 2024 was drafted with the express intention of regulating the use of artificial intelligence (AI) in digital technologies, precisely because it facilitates the rapid spreading of disinformation as well as posing a number of other policy concerns (European Parliament 2023b). The European Parliament has stated that its priority "is to make sure that AI systems used in the EU are *safe*, transparent, traceable, *non-discriminatory* and environmentally friendly. AI systems should be overseen by people, rather than by automation, to *prevent harmful outcomes*" (European Parliament 2023a, emphasis added). This piece of legislation, drafted to specifically tackle the harmful effects of digital technology and hold tech companies accountable for them, is an important step in addressing issues of mis/disinformation online, including hate speech. It constitutes a clear example of the possible legal obligations that can be created to ensure tech companies take appropriate responsibility and limit their use of algorithmic technology for profit at the expense of the truth and social media users.

The AI Act 2024 demonstrates the seriousness with which EU member states approach the problematic role of digital technology in promoting hate speech and should be considered a strong foundation for tech companies to uphold their own standards, supported by law. Indeed, such external regulation should be welcomed as support in upholding their own mission statements which have variously established social media platforms not as neutral content providers or dissemination platforms, but as civil-communicative institutions; that is, institutions that have an associative, solidarising and civil purpose. For example, Facebook states that its mission is to "give people the power to build community and bring the world closer together" (Facebook n.d.) and aims to "give people a voice", "build connection and community" and "keep people safe" (Meta n.d.). X sees its mission as giving "everyone the power to create and share ideas and information, and to express their opinions and beliefs without barriers" and to "serve the public conversation, which requires representation of a diverse range of perspectives" (X 2023). Both thereby attribute themselves a role that goes beyond providing a dissemination platform without a specific purpose or a specific aim in mind. Consequently, as public institutions with civil aims they have an inherent responsibility to be accountable and to take action to uphold their own standards and aims, and provide clear explanations for when they do not.

While social media platforms do facilitate positive social relations and have increased their efforts to limit hateful interactions online (e.g., via content removal and blocking/removal of accounts) (Gillespie 2018; Siapera and Viejo-Otero 2021; Wan and Kim 2023), these efforts can largely be described as a band-aid approach: they are not systematic in kind or driven by a sincere concern for societal consequences, but only occur when enforced by law or under significant public pressure—rare occurrences and so they often remain unsuccessful to stop the proliferation of hate speech. They are also what Ullmann and Tomalin (2020) have described as reactive rather than pro-active and as "Too Little Too Late" interventions. In short, there has been no genuine and voluntary attempt to fix the root problem of social media platforms from an engineering perspective and to thereby protect users from being choicelessly exposed to hate speech and vulnerable groups from falling victims to oppression.

Thus far, social media companies have not been legally forced to "abandon the pretense that they are 'merely' neutral platform providers" (Singer and Brooking 2018, p. 268). If they were neutral platform providers they would not "use purpose-built algorithms to determine how information gets to people, exactly who gets it, and when" (Woolley 2023, p. 50)—which is exactly what they do. What current legislation shows is not that social media is entirely unregulated but that "the growth of technology's power has not been matched by corresponding growth in legal responsibility" (Susskind 2022, p. 7), meaning that tech companies get away with "serious misconduct" (ibid.) and are largely immune to legal liability. In other words, the mechanical design and engineering of social media platforms is way ahead of any regulating laws and deliberately so—tech companies have made no attempt to adjust their algorithms or mechanical make-up in order to address

concerns over hate speech, nor have they had to. This could, however, be redressed by tweaking Section 230 of the U.S. Communications Decency Act (CDA) (which applies to Facebook and X which are based in the US), which would "allow people to sue platforms for the algorithmic amplification of harmful content" (Hasen 2022, p. 312). It would make tech companies open to lawsuits when, on the balance of probability, harmful content online can be seen to have triggered offline violence—such as the 2019 Christchurch Mosque shootings, election violence, or the 2017 Rohingya genocide in Myanmar—or have led to such psychological harm that suicide ensued, an issue especially pertinent amongst adolescents as the primary users of social media. This would shift some responsibility towards social media companies rather than expecting hate speech victims to engage in lengthy and costly processes to claim their rights and hold tech companies responsible for the harm done by hate content on their platforms.

Indeed, in the US, where hate speech does not exist as a legal concept due to the First Amendment, Section 230 of the CDA[8] provides near complete legal immunity to tech companies such as X and Facebook for any content on their platforms at the time of writing.[9] This means that tech companies have accumulated an unprecedented degree of unaccountable power and influence on societies worldwide. They have the power to "write code [and thereby] to write the rules by which the rest of us live" and "to affect how we perceive the world" (Susskind 2022, p. 4). Not only that, codes introduce the bias of the coder as well as the bias (gender, race, ethnicity, religion) that is contained in the historical data sets used to develop codes and algorithms (Noble 2018; Benjamin 2019) and they influence how much visibility and prominence is given to hate speech. To avoid this, regulation should target social media platforms and tech companies from an engineering problem perspective, as redressing the mechanics can have a significant impact on what kind of content can circulate on social media platforms and be amplified. Put differently, "major platforms should be regulated at the system of design level" (Wylie 2019, p. 295) to redress design, reduce coded biases, make secret algorithms transparent[10] and disable the amplification of false and manipulative information and by extension the sensationalising of negative content particularly, hate speech.[11]

To achieve this, any regulation would apply to social media companies depending on their level of social risk (Susskind 2022) and should include the following four aspects:

First, limitation of the size of platforms in the same way that many countries regulate media ownership and concentration. Such existing regulation already "implicitly recognises that too much concentration of power over what people are told is dangerous to governments and citizens" (Arthur 2021, p. 311) and could be extended to social media platforms.

Second, stipulations on making visible and labelling misinformation, including the requirement for mechanisms to disrupt the spread of hate speech in two ways: (a) by introducing a circuit breaker that ensures "that newly viral content is temporarily stopped from spreading while it is fact-checked" (Fox 2020, no page number) and thereby undermining the contribution of misinformation to spreading hate speech, and (b) to adopt a "quarantining process" for hate speech as developed by Ullmann and Tomalin (2020, p. 75), which "enables the recipients of those messages (or other appropriate moderators) to decide (i) whether they wish to read the messages or not, and (ii) if they decide to read them, whether they wish to allow them to be posted or not". The aim of such a "quarantining process" is to prevent harm by protecting users from exposure to hate speech and by giving them more ownership in how to engage with hate speech directed at them.

Third, imposing a duty to openness which would mandate social media platforms to "disclose the content of the moderation policies, algorithms and human processes used to implement those policies, the principles and practices guiding the ordering of content on the platforms, statistics concerning regulated areas (. . .) and details about how data is gathered and used" (Ullmann and Tomalin 2020, p. 299). Fourth, imposing on social media platforms the obligation to devise community standards as "rules that can be followed, that make sense to users, that give [the platform's] policy team a reasonably clear guide

for deciding what to remove, that leave enough breathing room for questionable content they might want to retain, that can change over time, and that will provide a satisfactory justification for removals if they're disputed, whether by users themselves or in the glare of public scrutiny" (Gillespie 2018, p. 45). This needs to be accompanied by a systematic and legal obligation to enforce their compliance by users in order to protect targeted groups from oppression, libel and offline violence.[12]

To prevent the harm that hate speech causes, it is imperative that tech companies be held legally responsible for what they design and build (the mechanics) and the harm that ensues through the purposive nature of hate speech—which, as we have shown, they are certainly aware of but choose to ignore in the pursuit of monetary gains. Such regulation entails a restriction on their powers, a review of their mechanical design, and tackling free speech fundamentalism.[13] Of course and undoubtedly, this is a difficult undertaking given that tech companies such as X and Facebook though based in the US operate globally—that is in different linguistic, political, cultural and legal environments. However, they have chosen to be global capitalist entities and accordingly, they need to take responsibility for the products they have created and are monetising often to the detriment of their stated civil aims. However, regulation is only part of the story. For sustainable change in communication and behaviour to be realised, steps must be taken from the bottom-up of social media platforms, as well as from the top-down. In other words, social media users must enact change as well as the social media companies who run them in order to build resilience. This can be done through digital citizenship education programmes that focus on the development of civil-communicative skills that enable users to harmlessly and responsibly navigate social media.

3.2. Aspect 2: Developing Civil-Communicative Skills for Responsible and Informed Social Media Use

Digital citizenship can be understood in various ways (see Hintz et al. 2018). We argue that digitally literate citizens are those who understand how the Internet and social media work and what impact everyone's communicative behaviour (including hate speech) has on individuals, groups and society. In Emejulu and McGregor's (2019, p. 140) words, it is "a process by which individuals and groups committed to social justice critically analyse the social, political and economic consequences of digital technologies in everyday life and collectively deliberate and take action to build alternative and emancipatory technologies and technological practices". Doing this effectively to combat hate speech requires a set of four civil-communicative skills: first, civil digital media literacy as a combination of technical skills and civil consciousness; second, the ability to recognise and identify different types of hate speech and their potential consequences for "the other"; third, counter-speech skills; and fourth, discursive civility as a depolarising and non-violent communication skill. The combination of these four skills can be understood as both an ex-ante (not engaging in hate speech) and ex-post (countering hate speech) universal approach to combating hate speech and building resilience. To take each of the four skills in turn:

1. Civil digital media literacy

Despite the emphasis on democracy, access and citizenship in the literature on media literacy (e.g., Wilson 2019), relatively little has been done to connect media literacy to citizenship education, to civil society as a communicative environment and the idea of peaceful association. Media information literacy has often been about the individual in isolation, rather than the individual having a responsibility as a member of civil society and being connected to others by means of communication. However, if we actually consider the wider connections and responsibilities of the individual as a civil actor, we can more fully grasp the issues any understanding of media literacy needs to be able to address, especially with a view to combating hate speech. In line with and developing Mihailidis' (2018, p. 155f.) gap analysis of current media literacy practice, we argue for an understanding of digital media literacy as having a civil role and being digital media-literate to require two basic core skills: first, technical skills and second, an understanding

of the transformative role of communication in civil society for both civil and anti-civil purposes—in short, civil consciousness (see Pukallus 2022).

Regarding the acquisition of technical skills, it is essential that social media users of all ages[14] are educated or trained in how the Internet, social media and search engines are engineered and work. We argue that this requires building a basic curriculum that must address as a minimum our proposed nine themes and questions:

(i) How tech companies are run
(ii) How search engines work, emphasising that the top result is not by any means the most reliable
(iii) How the online and the offline world might or might not be the same, i.e., how online noise distorts reality or reflects society
(iv) What the role of bots and algorithms is in the amplification of posts
(v) What happens when users comment, post, share or like
(vi) How the affordances on social media websites amplify messages/posts and comments and can make them go viral
(vii) The sophisticated ways in which pictures and videos can be audio-visually manipulated, including deep fakes which are nearly impossible to detect with the naked eye (Schick 2020; Woolley 2020)
(viii) The meaning and significance of coded bias in terms of programmed injustices, inequality and discrimination against certain groups
(ix) How any unintentional participation in the spread of hate speech can be avoided.

The acquisition of such technical skills regarding social media is important but on their own, they are not sufficient to combat hate speech online. Instead, such skills need to be complemented by the development of a civil consciousness that understands that communication has transformative capacity. As Dewey (Dewey [1916] 2011, p. 6) points out, "Society not only continues to exist by transmission, by communication, but it may fairly be said to exist in transmission, in communication" and in this way communication and the behaviour of everyone matters to ensure peaceful cooperation; that is, cooperating non-violently despite deep disagreements and differences. Transformative communication can be used to build more inclusive and socially just communities, to find compromise and consensus, to drive solidarity and thereby reduce hate. It gives users agency (Mihailidis 2018) to contribute to activism and advocacy of civil values and causes, where social media can be used to build communities, to show solidarity with the disadvantaged, and trigger social change in the pursuit of equality and justice. As Mihailidis (2018, p. 162) puts it, communication when informed by civil media literacy can "challenge existing systems and structures that restrict or constrain individuals and communities in their pursuit of bringing people together in pursuit of a common good". Yet, such transformative communicative outcomes are subverted by hate speech, raising the issue of how to identify it and its subversive aims.

2. The skill to identify different types of "hate speech" and to understand their potential consequences for "the other"

There are two basic types of and contributors to hate speech: first, stereotyping, prejudice and scapegoating often supported or reinforced by misinformation; and second, dehumanisation. To take each in turn:

(a) Stereotyping, Prejudice and Scapegoating

Stereotypes are best defined as "the pictures we hold in our heads of people from other groups. These pictures tell us something about these groups—their culture, temperament, level of threat—before we even interact with them (...) more often than not, these stereotypes are an exaggerated [negative or positive] view of reality" (Williams 2021, p. 165). Stereotypes in and of themselves do not constitute dangerous speech but they can become dangerous when they are predominantly negative and develop into enemy images (Oppenheimer 2006) and prejudice. Prejudices are entrenched attitudes. Unlike stereotypes, they "are not reversible when exposed to new knowledge. A prejudice, unlike

a simple misconception, is actively resistant to all evidence that would unseat it" (Allport 1958, p. 9) and often leads to scapegoating. Examples of this include hate speech against Muslims where Orientalist stereotypes and prejudices frame them as a threat to society and as having values incompatible with those of the West (see Said 2001), making peaceful cooperation impossible. This is despite the fact that, for example, the Islamic values of generosity, civil hospitality, peace and justice are shared with the Judeo-Christian values that underpin and are associated with the West. Often these stereotypes are formed, enhanced and upheld through mis- or disinformation in the form of lies, rumours or conspiracy theories.[15] Recognising stereotypes, prejudices and scapegoating as core elements of hate speech is important as it is through these that oppression is communicatively prepared, legitimised and facilitated, and that a civil society can be put on a path towards violence against targeted groups. It is essential that citizens are trained and educated in identifying these aspects as well as in how to verify sources, check facts and how to spot disinformation strategies.[16] The reason being that even if misinformation is debunked it is not easily forgotten and influences how citizens behave with each other.[17] As such, training to understand what dangers they can pose to "the other" can help individuals to avoid becoming unintentionally complicit in the preparation of such violence by either passively accepting or actively contributing to the spread of hate speech.[18]

(b) Dehumanising Speech

Dehumanising speech is a particularly dangerous kind of speech for "the other", as it constitutes "a denial that a certain group is 'equally' human, no matter how that 'humanity' is defined. It (. . .) denies that (. . .) members of that group are worthy of the same treatment or consideration that would be afforded to members of the in-group" (Savage 2013, p. 144; also, Neilsen 2015). Once members of a group have been identified by others as less than human, they are no longer seen as worthy of inclusion or protection by civil norms and principles and following that they are then often discriminated against by law and therefore legitimately. Dehumanising speech indoctrinates perpetrators of violence to the "point where they genuinely believe they are doing what is best for society, through purification and elimination of those seen as less than human and who therefore pose a threat to the common goal" (Green et al. 2015, p. 22). Indeed, dehumanisation occurs when there is still hesitation to harm members of these groups (Livingston Smith 2011) and as such, it functions as a tool to diminish inhibition based on the idea that "If the other group is not human, then killing them is not murder" (Stanton 2004, p. 214). However, as Livingston Smith (2020, p. 14) emphasises, dehumanisation "is not always in the service of slaughter. It's also a handmaid of oppression". Dehumanising speech uses a variety of metaphors and terms to show that members of these groups are subhuman (Waller 2007; Pukallus 2022) whether through objectification (Weitz 2005; Savage 2012), animalisation or the use of language of disease, infection, parasitism as well as criminality (also Livingston Smith 2020). For example, the dehumanisation of Black people—hate speech based on race—typically includes animal metaphors such as apes and monkeys, labelling Black people as savages, and references to a subhuman or "biologically inferior" status (Jardina and Piston 2021, p. 6), echoing colonial-era language and racism. When Black people are portrayed as less than human, they are seen as the less valuable other that can "legitimately" become a target of hate speech and oppression. Identifying dehumanising speech and understanding the often ultimately life-threatening danger of this type of speech for targeted groups is vital if the step from dehumanisation to violence is to be prevented.

3. Counter-speech skills

Free-speech advocates often recommend countering hate speech with more speech rather than removing hateful content; as Strossen (2018, p. 164) points out, "it is essential for the well-being of both individuals and society that we encourage and facilitate (. . .) counterspeech rather than adopting the disempowering, anti-democratic censorial approach". Foxman and Wolf (2012, p. 129) define counter-speech as "the dissemination of messages that challenge, rebut, and disavow messages of bigotry and hatred—can serve

a variety of purposes" such as "exposing hate speech for its deceitful and false content, setting the record straight, and promoting the values of respect and diversity". It is through counter-speech that social media users can be exposed to different points of view, see falsehoods corrected and stereotypes or dehumanising speech called out and countered, thereby "piercing the insularity of hateful messages that may lead to more extreme views" (Citron and Norton 2011, p. 1474). More specifically and in a pioneering study, Benesch et al. (2016, p. 17) identify eight counter-speech strategies to combat hate speech that have proven effective in reducing hate speech online and on social media in particular. These are: "(1) presentation of facts to correct misstatements or misperceptions, (2) pointing out hypocrisy or contradictions, (3) warning of possible offline and online consequences of speech, (4) identification with original speaker or target group, (5) denouncing speech as hateful or dangerous, (6) use of visual media, (7) use of humor, and (8) use of a particular tone, e.g., an empathetic one". Some have argued that counter-speech promotes hate speech and possibly even helps it recruit followers by increasing attention given to it. However, evidence points to the contrary and counter-speech therefore constitutes an effective alternative to silence, censorship or counterproductive "cancelling", which runs the risk of creating communicative martyrs or narratives of the unjust persecution of the truth-tellers. For example, and in line with Benesch et al.'s (2016) recommendations, a recent study by Hangartner et al. (2021) tested the effectiveness of three counter-speech strategies in tackling online xenophobic and racist hate speech and persuading the perpetrator to stop disseminating hate speech: humour, empathy and warning of consequences. The study found that each strategy was effective in reducing hate speech creation and increasing hate speech deletion from the social media platforms by the original hate speech creator, with fostering empathy being the most effective tool.[19] Another study has shown that counter-speech can be effective and opined that "citizens wishing to engage in counter-speech would likely increase the effectiveness of their efforts if they organized and participated in discussions in a coordinated way" (Garland et al. 2022, p. 21).

This shows that there is not only a place but a need for counter-speech skills as well as a systematic and coordinated approach to it. The effective practice of counter-speech, however, also requires the learning of communicative skills to avoid unintentional escalation or amplification.[20]

4. Discursive civility

Given the necessity for digital media literacy skills and the ability to identify hate speech, it is imperative to address the practical matter of "how to use" social media harmlessly and responsibly. Here, the concept of discursive civility is crucial. Discursive civility, developed by Pukallus (2022), is achieved when three principles are upheld in communicative engagement: first, managing and constructively using individual negative emotions (emotional forbearance) rather than suppressing them and requiring people to be rational; second, genuinely listening to the other and importantly hearing the other (perspective-taking), rather than having a dismissive or aggressive attitude; and third, making only such contributions that are supportive of the pursuit of the common goal (reasonableness). Discursive civility is the minimal communicative condition for non-violent communication in conflictual contexts, whether these are simply polarised and hateful, physically violent, or concern post-conflict and post-civil war scenarios. Its three principles are universal though locally sensitive, and they can be taught, learnt and adapted to cultural circumstances. Importantly, they do not prescribe content, regulate which topics can be covered and which ones cannot but rather the style of expression. Discursive civility aims to teach individuals how to deal with difficult situations, where emotions are intense, where disagreement is deep, and where conflict could at any point turn into violence. It empowers individuals to de-escalate verbal conflict and to move from enmity and aggression to co-citizenship and peaceful cooperation (see Pukallus 2022, 2024a) and to build resilience to hate speech.

3.3. Aspect 3: Utilising Local Evidence-Based Anti-Hate Speech Interventions on Social Media

Communicative peacebuilding is an inherently local approach, and essential to tackling the problems of hate speech in the digital battlefield. The core principle of local peacebuilding is that those closest to conflict—here closest to hate speech (see below for an example)—are also the ones who are best placed "to create and enact their own solutions to prevent, reduce, and/or transform the conflict, with the support they desire from outsiders" (Locally Driven Peacebuilding 2015, p. 2). It is based on the firm belief that "people at the local level (a) have the capacity to articulate, develop, and enact solutions to their own problems; (b) possess important knowledge and a better understanding about the complexities of the people, situation, context, and culture of the conflict; (c) maintain pre-existing, established relationships that enable them to network and engage with others in the peace process; (d) may be more likely to have the motivation to act, and act quickly, to resolve issues that directly impact them; and (e) have the staying power to sustain peace long after external actors have left" (Locally Driven Peacebuilding 2015, p. 2). It is this kind of local expertise, investment and capacity that needs to also be mobilised in anti-hate speech campaigns. More specifically, to effectively combat hate speech it is essential for local organisations to be involved in (i) identifying hate speech which may include culturally specific pejorative terms or locally symbolic information, seemingly superficially innocent but in fact representing hate, persecution or violence: and (ii) designing anti-hate speech campaigns that resonate with their audiences and engage them effectively. Both (i) and (ii) require local organisations to gather evidence of hate speech on social media and design anti-hate campaigns to counter it. It is here that local organisations might benefit from external expertise and capacity—that of digital peacebuilders. The Digital Media Arts for an Inclusive Public Sphere (DMAPS) project provides an example of local evidence-based anti-hate speech interventions worth examining in more detail.

3.4. Digital Media Arts for an Inclusive Public Sphere (DMAPS)

In 2021/22, a consortium of academic and non-academic organisations led by Build Up and funded by the British Council (British Council n.d.) piloted the DMAPS programme. DMAPS supported "young leaders from [18] creative and media arts organisations across eight countries [Libya, Palestine, Jordan, Yemen, Iraq, the Occupied Palestinian Territories (OPT), Tunisia and Syria] to conduct social media mapping in order to understand how issues of identity, social cohesion and inclusion manifest on and are shaped by social media" (Build Up 2022b, p. 2).[21] The pilot programme had three stages: first, the definition of problem statements; second, the social media mapping process, i.e., data collection and analysis; and third, the design of evidence-based interventions to counter hate speech and polarisation. In the first stage, each of the 18 participants individually or in regional clusters developed a problem statement which essentially focused on two things: (a) a specific conflict or controversial and polarising topic on social media which would then frame the social media mapping and (b) a set of questions that would guide the data collection and analysis and, subsequently, the intervention. The topics varied and included the expressions of religious and ethnic identities; hate speech between different groups; generational conflict; and harassment, hate speech and violence against women.

Based on the problem statement and in the second stage, participants developed a detailed list of accounts (public Facebook pages and groups, Twitter (now X) handles, YouTube channels) to analyse and keywords to search for in public posts on Facebook, tweets or public YouTube videos in order to scrape data. They also developed classification models for topics, actors and sentiment so that AI could then be trained to automatically apply the labelling to posts (for details, see Build Up 2022a). All coded data were subsequently pulled into a dashboard which included tables and graphs and that allowed participants to see where polarisation and hate speech occurred, to then subsequently design anti-hate interventions online to counter the kind of hate speech that occurred. For example, the stereotyping of women was prominent on social media. Participants in Libya "found that patriarchal assumptions about women and gender roles are reflected

in social media, with explicit reference to discrimination against women in the workplace, honour crimes, gender-based violence and harassment, and masculinity. (. . .) Facebook posts that referred to gender or women-related topics were often ridiculed and generated aggressive responses from men, including dehumanising language and religious references to support this position" (Arthur and Pukallus 2022, p. 3). Hate speech in Kurdistan, in turn, was directed at ethnic and religious minority communities (specifically the Yezidi and Christian minorities) and participants "found that hate speech (. . .) has included name calling, mocking religious figures and icons, and statements to challenge/undermine (in particular the Christian community's) faith and religious practice" (Arthur and Pukallus 2022, p. 4). Hate speech targeting ethnic groups and using particularly sectarian and "ethnically-charged" language was also identified in social media posts in Yemen (Arthur and Pukallus 2022, p. 5).

In order to develop effective, evidence-based anti-hate interventions to redress such hate speech, the DMAPS programme provided skills training for partners in non-violent communication, discursive civility, depolarisation and re-humanisation strategies to ensure effective communication and, particularly, to prevent any unintentional participation in or amplification of hate speech by the partners. The partners then designed interventions which had three common elements: first, they all raised awareness of the existence of hate speech and its detrimental impact on society; second, they all shared experiential accounts of those exposed to and targeted by hate speech and third, they all engaged in debunking common beliefs and providing alternative narratives. These carefully and locally designed communicative anti-hate campaigns had a measurable impact in two ways. First, this approach enabled local organisations to directly target hate speech without amplifying any existing divisions, thereby significantly and measurably decreasing the number of hateful comments on relevant social media platforms and groups. Second, users who had previously engaged in hate speech or used violent language to talk about certain issues changed their vocabulary in identifiable ways to voice disagreement or criticism in response to the anti-hate campaigns, resulting in a measurable reduction in hateful messages (Build Up 2022b).

4. Conclusion: Responsibility and Capacity to Prevent Group Oppression

Much research on hate speech and how to combat it is available, but often these works remain in their disciplinary fields. In this paper, we have tried to overcome disciplinary isolation and drawn from insights from computer science, sociology, law, political and public communication as well as peacebuilding. By understanding social media as a communicative battlefield and hate speech as a digital weapon, we have been able to conceive of combating hate speech as an exercise in communicative peacebuilding and to develop the combinatorial approach we propose. This approach has in parts been applied in real-life contexts such as the DMAPS project and peacebuilding activities undertaken by one of the authors in Kenya and has shown empirical success. It is clear and desirable, however, that the approach be tested and used in a variety of real-life contexts in order to confirm its feasibility and identify areas for adaptation or improvement. Importantly and when successful, the combination of the three aspects of communicative peacebuilding enables a focus on two main things: responsibility and capacity. Regarding the former, our approach puts clear responsibility on tech companies for their mechanics and the artificial amplification of harmful content. It requires them to adjust the underlying architecture of social media platforms and regulate content by enforcing community standards systematically through transparent, verifiable and accountable means. It is in this way that tech companies can take responsibility for the harm their social media platforms can cause both online and offline and their contribution to attacks on dignity and group oppression. Concerning the latter, it is essential to build the capacity for confidently recognising, countering or even preventing hate speech in civil societies through education in digital media literacy and discursive civility, alongside tech company regulation. There is no easy way to combat hate speech—othering and attempting to harm the other is part of human nature. However, it is

also part of human nature to connect with others, to empathise and protect. By supporting and strengthening this through education and skills training, it is possible to build the capacity to prevent and to redress hate speech and make a significant start in tackling its fundamental root(s). From the top down and the bottom up, conscious effort is required to promote dignity, respect and equality rather than hate. It is only when hate speech is continuously disabled that it becomes devalued, and this is necessary if one wants to prevent hate speech online from further contributing to group oppression in its various manifestations.

Author Contributions: Conceptualization, S.P. (lead) and C.A.; writing—original draft preparation S.P. (lead); writing—review and editing, S.P. and C.A.; funding acquisition, S.P. and C.A. as part of DMAPS. All authors have read and agreed to the published version of the manuscript.

Funding: British Council, Digital Media Arts for an inclusive Public Sphere (for Section 3.4).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

1 Our focus here is on major platforms such as Facebook, Twitter/X, TikTok and Reddit rather than community boards or chatrooms that have little influence on and power over societies.

2 For an overview covering eight aspects of hate speech that fall into more than one of these approaches see Sellars (2016).

3 See example Williams et al. (2016); Ozalp et al. (2020).

4 For more detail on these three communicative dimensions of hate speech see Pukallus (2024b).

5 Jones (2015, p. 684) argues that "Waldron characterises free expression as itself a demand of dignity, so that the trade-off takes place *within* the confines of dignity, with the implication that dignity will not lose out in the trade". On dignity vs. freedom of expression see also Heyman (2008), Simpson (2013) and Bousquet (2022).

6 See Sherry (2010) on how a particular group—the disabled—is oppressed through hate crimes. Though Sherry (2010) does not use the term oppression, the argument mad resonates with Young's understanding of group oppression.

7 These can be any civil or political actor.

8 We refer to the US as two of the largest social media platforms are based there and operate under US law. It is a matter of fact that the US is where many such large tech companies are based, where the great digital revolution of Silicon Valley took place at the turn of the century, and that it is within this legal context that social media platforms with such global reach operate. This does not render our discussion US-centric. We will cover the MENA region later on in this paper.

9 This is not to suggest that circulation is not also facilitated by the sheer heterogeneity of social media platforms and the emergence of a new alt-tech landscape that is used circumvent moderation and restriction of conventional social media platforms (see Ebner and Guhler 2024).

10 Change the Terms (n.d.) runs a communicative campaign entitled 'Fix the Feed' which includes three main steps: to fix the algorithms to stop 'promoting the most incendiary, hateful content', to 'protect people equally' and 'disclose business models and moderation practices).

11 Other more proactive suggestions include encouraging social media platforms to formulate a purpose such as bridging partisan divides and to reward users that help contribute to the achievement of the purpose (see e.g., Bail 2021).

12 See also Carlson and Rousselle (2020). Our argument is not solely about content moderation and such, we do not address this issue in detail. We simply advocate holding social media companies to account in terms of the enforcement of their community guidelines, it is a responsibility point. However, for more information on the difficulty of content moderation see Gerrard (2018), Gillespie (2018), Wilson and Land (2020), Díaz and Hecht-Felella (2021).

13 There are those who believe in absolute freedom of expression and then there is the generally accepted idea that freedom of expression needs to be limited to prevent harm to others (see e.g., Gorenc 2022). In this paper, we are in favour of freedom of expression within the confines of dignity (in agreement with Waldron) and are advocating for tech companies to deprive users of the algorithmic megaphone.

14 The imperative to teach children such media literacy skills form an early stage has been acknowledged and to some extent pioneered by Common Sense, a nonprofit organisation, that has since 2003 engaged in designing curricula for school children

focusing on a variety of aspects including the importance of language and the harm that can derive form language as well as norm education about community membership.

15 A clear example of Islamophobic stereotyping, rumour and disinformation was the conspiracies that were widely circulated on social media by the Leave.EU campaign in the United Kingdom prior to the 2016 Brexit referendum. Leave.EU propaganda directly linked the topical issue of immigration to Islamic extremism and terrorism in the crude trope 'immigration without assimilation equals invasion' (Bakir 2020, p. 11). The spreading of this Islamophobic conspiracy and its success in persuading popular opinion relied directly on social media users' algorithms. See Bakir (2020) and Caeser (2019).

16 Such training has been developed in various ways. On example is the simulation engine Bad News (Van der Linden 2023), another is Common Cause (n.d.); for lists of training see: RAND (n.d.) and IREX (2023).

17 For examples of what governments across the world are doing to combat mis-/disinformation see Poynter 2024 (Funke and Flamini 2024).

18 Our approach focuses on ordinary civil and political actors rather than professional hate speech producers and disseminators such as those working for troll farms being paid for using hate speech for political gain.

19 A 2023 study by Zheng, Ross and Magdy interestingly tests whether counter-speech generated by ChatGPT can be effective in countering hate speech but comes with challenges and ethical questions. They (Zheng et al. 2023, p. 70) argue: that 'Looking to the future, our analysis shows that the automatic generation of counterspeech remains a challenging task, even for current large language models' and that the 'prospect of using automatically generated counterspeech to counter hate speech on social media raises important ethical questions' (ibid.).

20 This need is also recognised by PEN America (2024) in their *Guidelines for Safely Practicing Counterspeech* when they provide specific recommendations on how to avoid escalation and achieve de-escalation.

21 For reasons of participant security, confidentiality and research ethics and integrity we cannot provide any further information on these partners.

References

Allport, Gordon. 1958. *The Nature of Prejudice*. London: Basic Books.

Aral, Sinan. 2021. *The Hype Machine: How Social Media Disrupts Our Elections, Our Economy and Our Health—And How We Must Adapt*. London: HarperCollins Publishing.

Arthur, Charles. 2021. *Social Warming: The Dangerous and Polarising Effects of Social Media*. Edinburgh: One World.

Arthur, Catherine, and Stefanie Pukallus. 2022. *Theoretical Foundations of the DMAPS Approach*. Position Paper 1. Edgecliff: British Council.

Bail, Chris. 2021. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton: Princeton University Press.

Bakir, Vian. 2020. Psychological operations in digital political campaigns: Assessing Cambridge Analytica's psychographic profiling and targeting. *Frontiers in Communication* 5: 67. [CrossRef]

Bar-Ilan, Judit. 2007. Manipulating search engine algorithms: The case of Google. *Journal of Information, Communication and Ethics in Society* 5: 155–66. [CrossRef]

Benesch, Susan, Derek Ruths, Kelly P. Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Counterspeech on Twitter: A Field Study. Report, Public Safety Canada. Available online: https://dangerousspeech.org/counterspeech-on-twitter-a-field-study/ (accessed on 11 July 2023).

Benjamin, Ruha. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity Press.

Bousquet, Chris. 2022. Words That Harm: Defending the Dignity Approach to Hate Speech Regulation. *The Canadian Journal of Law & Jurisprudence* XXXV: 31–57.

British Council. n.d. Call for Applications—Digital Media Arts for an Inclusive Public Sphere. Available online: https://iraq.britishcouncil.org/en/about/jobs/call-applications-—digital-media-arts-inclusive-public-sphere (accessed on 18 April 2023).

Build Up. 2022a. *Participatory Action Research, Polarisation, and Social Media: Ongoing Lessons from the Digital Maps Program*. Edgecliff: British Council.

Build Up. 2022b. Digital Media Arts for an inclusive Public Sphere. Final Report April 2022 unpublished.

Caeser, Ed. 2019. The Chaotic Triumph of Aaron Banks, the "Bad Boy of Brexit". *The New Yorker*. March 18. Available online: https://www.newyorker.com/magazine/2019/03/25/the-chaotic-triumph-of-arron-banks-the-bad-boy-of-brexit (accessed on 3 April 2024).

Carlson, Caitlin. 2021. *Hate Speech*. Massachusetts: MIT Press.

Carlson, Caitlin, and Hayley Rousselle. 2020. Report and repeat: Investigating Facebook's hate speech removal process. *First Monday* 25. [CrossRef]

Castaño-Pulgarín, Sergio Andrés, Natalia Suárez-Betancur, Luz Magnolia Telano Vega, and Harvey Mauricio Herrera López. 2021. Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behaviour* 58: 101608. [CrossRef]

Center for Countering Digital Hate. 2023. X Content Moderation Failure. Report. September. Available online: https://counterhate.com/wp-content/uploads/2023/09/230907-X-Content-Moderation-Report_final_CCDH.pdf (accessed on 23 January 2024).

Change the Terms. n.d. Fix the Feed. Available online: https://www.changetheterms.org (accessed on 23 January 2024).

Chua, Amy. 2018. *Political Tribes. Group Instinct and the Fate of Nations*. London: Bloomsbury.

Citron, Danielle, and Helen Norton. 2011. Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review* 91: 1435–84. Available online: https://scholarship.law.bu.edu/faculty_scholarship/614 (accessed on 26 March 2024).
Cohen-Almagor, Rafael. 2011. Fighting Hate and Bigotry on the Internet. *Policy & Internet* 3: 6.
Coleman, Peter. 2021. *The Way Out. How to Overcome Toxic Polarization*. New York: Columbia University Press.
Common Cause. n.d. Stop Disinformation Training. Available online: https://www.commoncause.org/stopdisinformationtraining (accessed on 22 January 2024).
Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. Paper Presented at the International AAAI Conference on Web and Social Media, Montreal, QC, Canada, May 15–18; vol. 11.
Dewey, John. 2011. *Democracy and Education. An Introduction to the Philosophy of Education*. New York: Simon and Brown. First published 1916.
Díaz, Ángel, and Laura Hecht-Felella. 2021. Double Standards in Social Media Content Moderation. Brennan Center for Justice. August. Available online: https://www.skeyesmedia.org/documents/bo_filemanager/Double_Standards_Content_Moderation.pdf (accessed on 23 January 2024).
Ebner, Jakob, and Julia Guhler. 2024. Extremism, the extreme right and conspiracy myths on social media. In *Handbook of Conflict and Peace Communication*. Edited by Stacey Connaughton and Stefanie Pukallus. New York: Routledge, Forthcoming.
Emejulu, Akwugo, and Callum McGregor. 2019. Towards a radical digital citizenship in digital education. *Critical Studies in Education* 60: 131–47. [CrossRef]
European Parliament. 2023a. EU AI Act: First Regulation on Artificial Intelligence. *European Parliament News*. December 19. Available online: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (accessed on 23 January 2024).
European Parliament. 2023b. Parliament's Negotiating Position on the Artificial Intelligence Act. *Plenary: At a Glance*. June. Available online: https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/747926/EPRS_ATA(2023)747926_EN.pdf (accessed on 23 January 2024).
Facebook. n.d. Hate Speech. Available online: https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/ (accessed on 18 April 2023).
Fox, Chris. 2020. Social Media: How Might it Be Regulated? *BBC News*. November 12. Available online: https://www.bbc.co.uk/news/technology-54901083 (accessed on 18 April 2023).
Foxman, Abraham, and Christopher Wolf. 2012. *Viral Hate. Containing Its Spread on the Internet*. Basingstoke: Palgrave Macmillan.
Frenkel, Sheera, and Kate Conger. 2022. Hate Speech's Rise on Twitter is Unprecedented, Researchers Find. *New York Times*. December 2. Available online: https://www.nytimes.com/2022/12/02/technology/twitter-hate-speech.html (accessed on 14 November 2023).
Funke, Daniel, and Daniela Flamini. 2024. A Guide to Anti-misinformation Actions around the World. Poynter. Available online: https://www.poynter.org/ifcn/anti-misinformation-actions (accessed on 22 January 2024).
Gagliardone, Iginio, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering Online Hate Speech*. Paris: UNESCO.
Garland, Joshua, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2022. Impact and dynamics of hate and counter speech online. *EPJ Data Science* 11: 3. [CrossRef]
Gerrard, Ysabel. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society* 20: 4492–511.
Gillespie, Tarleton. 2018. *Custodians of the Internet*. Yale: Yale University Press.
Gorenc, Nina. 2022. Hate speech or free speech: An ethical dilemma? *International Review of Sociology* 32: 413–25. [CrossRef]
Green, Penny, Thomas MacManus, and Alicia de la Cour Venning. 2015. Countdown to annihilation: Genocide in Myanmar. International State Crime Initiative. Available online: http://statecrime.org/state-crime-research/isci-report-countdown-to-annihilation-genocide-in-myanmar/ (accessed on 15 November 2023).
Hagan, John, and Wenona Rymond-Richmond. 2008. The Collective Dynamics of Racial Dehumanization and Genocidal Victimization in Darfur. *American Sociological Review* 73: 875–902. [CrossRef]
Halperin, Eran. 2011. Emotional Barriers to Peace: Emotions and Public Opinion of Jewish Israelis about the Peace Process in the Middle East. *Peace and Conflict* 17: 22–45. [CrossRef]
Hangartner, Dominik, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, and et al. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment, Brief Report. *Proceedings of the National Academy of Sciences of the United States of America* 118: e2116310118. [CrossRef] [PubMed]
Harrison, Jackie, and Stefanie Pukallus. 2018. The Politics of Impunity: A Study of Journalists' Experiential Accounts of Impunity in Bulgaria, Democratic Republic of Congo, India, Mexico and Pakistan. *Journalism* 22: 303–19. [CrossRef]
Hasen, Richard. 2022. *Cheap Speech: How Disinformation Poisons Our Politics—And How to Cure It*. Yale: Yale University Press.
Heyman, Steven. 2008. *Free Speech and Human Dignity*. Yale: Yale University Press.
Hintz, Arne, Lena Dencik, and Karin Wahl-Jorgensen. 2018. *Digital Citizenship in a Datafied Society*. Cambridge: Polity Press.
Howard, Philipp. 2020. *Lie Machines*. Yale: Yale University Press.

IREX. 2023. Supporting Information Integrity and Resilience: Tools and Resources. Available online: https://www.irex.org/supporting-information-integrity-and-resilience-tools-and-resources (accessed on 22 January 2024).
Jardina, Ashley, and Spencer Piston. 2021. Hiding in plain sight: Dehumanization as a foundation of white racial prejudice. *Sociology Compass* 15. [CrossRef]
Jones, Peter. 2015. Dignity, Hate and Harm. *Political Theory* 43: 678–86. [CrossRef]
Klein, Ezra. 2020. *Why We Are Polarized*. London: Profile Nooks.
Littman, Rebecca, and Elizabeth Paluck. 2015. The cycle of violence: Understanding individual participation in collective violence. *Political Psychology* 36: 79–99. [CrossRef]
Livingston Smith, David. 2011. *Less than Human. Why We Demean, Enslave, and Exterminate Others*. New York: St. Martin's Griffin.
Livingston Smith, David. 2020. *Making Monsters. The Uncanny Power of Dehumanization*. Cambridge: Harvard University Press.
Locally Driven Peacebuilding. 2015. Signed Letter. March 27. Available online: https://www.cla.purdue.edu/ppp/documents/publications/Locally.pdf (accessed on 15 November 2023).
Mason, Liliana. 2018. *Uncivil Agreement: How Politics Became Our Identity*. Chicago: Chicago University Press.
Meta. n.d. Our Principles. Available online: https://about.meta.com/uk/company-info/ (accessed on 18 April 2023).
Mihailidis, Paul. 2018. Civic media literacies: Re-Imagining engagement for civic intentionality. *Learning, Media and Technology* 43: 152–64. [CrossRef]
Mouffe, Chantal. 2005. *On the Political*. London: Routledge.
Neilsen, Rhiannon. 2015. 'Toxification' as a more precise early warning sign for genocide than dehumanization? An emerging research agenda. *Genocide Studies and Prevention: An International Journal* 9: 83–95. [CrossRef]
Noble, Safia. 2018. *Algorithms of Oppression. How Search Engines Reinforce Racism*. New York: New York University Press.
Opotow, Susan. 1990. Moral Exclusion and Injustice: An Introduction. *Journal of Social Issues* 46: 1–20. [CrossRef]
Oppenheimer, Louis. 2006. The Development of Enemy Images: A Theoretical Contribution. *Peace and Conflict: Journal of Peace Psychology* 12: 269–92. [CrossRef]
Ozalp, Sefa, Matthew Williams, Pete Burnap, Han Liu, and Mohamed Mostafa. 2020. Antisemitism on Twitter: Collective Efficacy and the Role of Community Organisations in Challenging Online Hate Speech. *Social Media + Society* 6: 1–20. [CrossRef]
Papcunová, Jana, Marcel Martončik, Denisa Fedáková, Michal Kentoš, Miroslava Bozogáňová, Ivan Srba, Robert Moro, Matúš Pikuliak, Marián Šimko, and Matúš Adamkovič. 2023. Hate speech operationalization: A preliminary examination of hate speech indicators and their structure. *Complex & Intelligent Systems* 9: 2827–42. [CrossRef]
PEN America. 2024. Guidelines for Safely Practicing Counterspeech. Available online: https://onlineharassmentfieldmanual.pen.org/guidelines-for-safely-practicing-counterspeech (accessed on 22 January 2024).
Pukallus, Stefanie. 2022. *Communication in Peacebuilding. Civil Wars, Civility and Safe Spaces*. Basingstoke: Palgrave Macmillan.
Pukallus, Stefanie. 2024a. Discursive civility. Theory and practice. In *Handbook of Conflict and Peace Communication*. Edited by Stacey Connaughton and Stefanie Pukallus. New York: Routledge, Forthcoming.
Pukallus, Stefanie. 2024b. The three communicative dimension of hate speech. In *Handbook of Conflict and Peace Communication*. Edited by Stacey Connaughton and Stefanie Pukallus. New York: Routledge.
RAND. n.d. Tools that Fight Disinformation Online. Available online: https://www.rand.org/research/projects/truth-decay/fighting-disinformation/search.html (accessed on 23 January 2024).
Reich, Rob, Mehran Sahami, and Jeremy Weinstein. 2023. *System Error: Where Big Tech Went Wrong and How We Can Reboot*. London: Hodder Paperbacks.
Ressa, Maria. 2023. *How to Stand Up to a Dictator*. London: Penguin.
Royzman, Edward, Clark McCauley, and Paul Rozin. 2005. From Plato to Putnam: Four ways to think about hate. In *The Psychology of Hate*. Edited by R. J. Sternberg. Washington, DC: American Psychological Association, pp. 3–35.
Said, Edward. 2001. *Orientalism*. London: Penguin Classics.
Savage, Rowan. 2012. 'With scorn and bias': Genocidal dehumanisation in bureaucratic discourse. In *Genocide Perspectives IV. Essays on Holocaust and Genocide*. Edited by Colin Tatz. Sydney: The Australian Institute for Holocaust & Genocide Studies, pp. 21–64.
Savage, Rowan. 2013. Modern genocidal dehumanization: A new model. *Patterns of Prejudice* 47: 139–61. [CrossRef]
Schick, Nina. 2020. *Deep Fakes and the Infocalypse. What You Urgently Need to Know*. London: Monoray.
Schmitt, Carl. 2007. *The Concept of the Political*. Chicago: Chicago University Press. First published 1932.
Sellars, Andy. 2016. Defining Hate Speech. *The Berkman Klein Center for Internet & Society*. December. Available online: https://cyber.harvard.edu/publications/2016/DefiningHateSpeech (accessed on 23 January 2024).
Sherry, Mark. 2010. *Disability Hate Crimes: Does Anyone Really Hate Disabled People?* London: Routledge.
Siapera, Eugenia, and Paloma Viejo-Otero. 2021. Governing Hate: Facebook and Digital Racism. *Television & New Media* 22: 112–30. [CrossRef]
Simpson, Robert. 2013. Dignity, Harm, and Hate Speech. *Law and Philosophy* 32: 701–28. [CrossRef]
Singer, Peter, and Emerson Brooking. 2018. *LikeWar: The Weaponization of Social Media*. New York: Mariner Books.
Stanton, Gregory. 2004. Could the Rwandan genocide have been prevented? *Journal of Genocide Research* 6: 211–28. [CrossRef]
Strossen, Nadine. 2018. *Hate: Why We Should Resist It with Free Speech, Not Censorship*. Oxford: Oxford University Press.
Susskind, Jamie. 2022. *The Digital Republic: On Freedom and Democracy in the 21st Century*. London: Bloomsbury.

Tsesis, Alexander. 2009. Dignity and Speech: The Regulation of Hate Speech in a Democracy. *Law Review* 44: 497–532. Available online: http://lawecommons.luc.edu/facpubs (accessed on 28 March 2024).
Ullmann, Stefanie, and Marcus Tomalin. 2020. Quarantining online hate speech: Technical and ethical perspectives. *Ethics and Information Technology* 22: 69–80. [CrossRef]
United Nations. n.d. Understanding Hate Speech: What Is Hate Speech. Available online: https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech (accessed on 7 March 2024).
Van der Linden, Sander. 2023. *Foolproof: Why We Fall for Misinformation and How to Build Immunity*. London: Fourth Estate.
Waldron, Jeremy. 2012. *The Harm in Hate Speech*. Cambridge: Harvard University Press.
Waller, James. 2007. *Becoming Evil: How Ordinary People Commit Genocide and Mass Killing*. Oxford: Oxford University Press.
Wan, Sai, and Ki Joon Kim. 2023. Content Moderation on Social Media: Does It Matter Who and Why Moderates Hate Speech? *Cyberpsychology, Behavior, And Social Network* 26: 527–34. [CrossRef]
Weitz, Eric. 2005. *A Century of Genocide. Utopias of Race and Nation*. Princeton: Princeton University Press.
Williams, Amanda, Clio Oliver, Katherine Aumer, and Chanel Meyers. 2016. Racial microaggressions and perceptions of Internet memes. *Computers in Human Behavior* 63: 424–32. [CrossRef]
Williams, Matthew. 2021. *The Science of Hate. How Prejudice Becomes Hate and What We Can Do to Stop It*. London: Faber & Faber Limited.
Wilson, Carolyn. 2019. *Media and Information Literacy: Challenges and Opportunities for the World of Education*. Ontario: The Canadian Commission for UNESCO's IdeaLab, November.
Wilson, Richard Ashby, and Molly Land. 2020. Hate Speech on Social Media: Content Moderation in Context. *Connecticut Law Review* 52: 1029–76. Available online: https://ssrn.com/abstract=3690616 (accessed on 15 November 2023).
Woolley, Samuel. 2020. *The Reality Game. How the Next Wave of Technology Will Break the Truth and What We Can Do about It*. London: Endeavour.
Woolley, Samuel. 2023. *Manufacturing Consensus. Understanding Propaganda in the Era of Automation and Anonymity*. Yale: Yale University Press.
Wylie, Christopher. 2019. *Mindf*ck: Inside Cambridge Analytica's Plot to Break the World*. London: Profile books.
X. 2023. Hateful Conduct Policy. Available online: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy (accessed on 10 November 2023).
Young, Iris Marion. 1990. *Justice and the Politics of Difference*. Princeton: Princeton University press.
Zhang, Ziqi, and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web* 10: 925–45. [CrossRef]
Zheng, Yi, Björn Ross, and Walid Magdy. 2023. What Makes Good Counterspeech? A Comparison of Generation Approaches and Evaluation Metrics. Paper Presented at the 1st Workshop on Counter Speech for Online Abuse (CS4OA), Prague, Czech Republic, September 11–12; Prague: Association for Computational Linguistics, pp. 62–71.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.