This item is the archived peer-reviewed author-version of:

Features for hate? Using the Delphi method to explore digital determinants for online hate perpetration and possibilities for intervention

Reference:
Weber Ina, Vandebosch Heidi, Poels Karolien, Pabian Sara.- Features for hate? Using the Delphi method to explore digital determinants for online hate perpetration and possibilities for intervention
Cyberpsychology, behavior, and social networking - ISSN 2152-2723 - (2023), p. 1-10
Full text (Publisher's DOI): https://doi.org/10.1089/CYBER.2022.0195
To cite this reference: https://hdl.handle.net/10067/1959990151162165141

uantwerpen.be
Institutional repository IRUA

Features for hate? Using the Delphi method to explore digital determinants for online hate perpetration and possibilities for intervention.

Ina Weber[a], Heidi Vandebosch[a], Karolien Poels[a], & Sara Pabian[a, b]

Affiliations

[a] Department of Communication Studies, University of Antwerp, Antwerp, Belgium

[b] Department of Communication and Cognition, Tilburg University, Tilburg, Netherlands

Corresponding Author: Ina Weber, ina.weber@uantwerpen.be

Author Contributions (CRediT): I.W.: Conceptualization, Methodology, Formal Analysis, Writing—Original Draft; **H.V.:** Methodology, Writing— Review and Editing, Supervision, Funding Acquisition; **K.P.:** Conceptualization, Methodology, Supervision; **S.P.:** Methodology, Supervision, Writing—Review and Editing.

Cite as:

Weber, I., Vandebosch, H., Poels, K., & Pabian, S. (2023). Features for Hate? Using the Delphi Method to Explore Digital Determinants for Online Hate Perpetration and Possibilities for Intervention. *Cyberpsychology, Behavior, and Social Networking*.

https://doi.org/10.1089/cyber.2022.0195

Abstract

Online hate speech on social media platforms causes harm to those who are victimized as well as society at large. The prevalence of hateful content has thus prompted numerous calls for improved countermeasures and prevention. For such interventions to be effective, it is necessary to gain a nuanced understanding of influences that facilitate the spread of hate speech. This study does so by investigating what are relevant digital determinants for online hate perpetration. Moreover, the study explores possibilities of different technology-driven interventions for prevention. Thereby, the study specifically considers the digital environments in which online hate speech is most often produced and disseminated, namely social media platforms. We apply frameworks related to the concept of digital affordances to focus on the role that technological features of these platforms play in the context of online hate speech. Data was collected using the Delphi method in which a selected sample of experts from both research and practice answered multiple rounds of surveys with the goal of reaching a group consensus. The study encompassed an open-ended collection of initial ideas, followed by a multiple-choice questionnaire to identify, and rate the most relevant determinants. Usefulness of the suggested intervention ideas was assessed through the three lenses of human-centered design. The results of both thematic analysis and non-parametric statistics yield insights on how features of social media platforms can be both determinants that facilitate online hate perpetration as well as crucial mechanisms of preventive interventions. Implications of these findings for future intervention development are discussed.

Keywords: Online hate speech, social media, Delphi study, determinants, interventions, digital affordances

Introduction

Hate speech is a form of prejudice-motivated communication towards (individual members of) a group defined by shared characteristics such as race, ethnicity, gender, or sexual identity.[1,2] It aims to discriminate, disparage or intimidate the targeted group.[3,4] In the online context, speech also includes (audio)visual forms of expression instantaneously communicated through digital channels.[5,6] As a context-dependent manifestation of prejudice, hate speech varies with regards to its explicitness and targets.[7,8] However, to the concern of scholars and field experts, the number of detected online hate speech incidents of any form on social media from Twitter to LinkedIn has been rising, giving importance to viewing the issue in its entirety.[9–11]

Digital spaces in which online hate speech is produced and disseminated are not static but subject to change. Social media companies regularly adjust their policies and functions, while new platforms with novel features emerge and users adapt their behavior in response.[12] The platforms' functionalities are products of organizational choices, economic interests, and institutional regulations impacting users individually, but also society at large.[13,14] In this dynamic context, technological functionalities do not determine user behavior, but are influential factors in a complex interplay between users and their social environments. With this study, we aim to give an overview of technological functionalities related to online hate perpetration and analyze them from an affordance perspective.

Previous studies looked in detail at specific technological functionalities in relation to harmful behavior or content (e.g., interface features and (in)civility in comment sections[15] or algorithmic recommendations promoting extremist content[16]). Yet there is no systematic overview of platform-independent functionalities and their relevance for the production and dissemination of online hate speech. Given the multidisciplinary nature of the subject, compiling this overview requires the consideration of perspectives from social sciences, as well as technology- and policy-oriented research.[17]

An effective way of achieving such an interdisciplinary overview is a Delphi study.[18] We used this consensus method combining survey-based and qualitative research techniques to address RQ1: *What do experts with knowledge on the topic of online hate speech agree on constitute the most relevant digital determinants for online hate perpetration?*

With increased attention dedicated to online hate speech, numerous interventions are being envisioned by researchers and practitioners in civil society organizations and think tanks, but also by social media companies themselves.[19] These fast-paced changes make it difficult for researchers to map recent developments solely through published research. Therefore, practitioners were also invited as participants, ensuring high actuality and applicability of the results.

To identify directions for future solutions in hate speech prevention, RQ2 asks: *What do experts with knowledge on the topic of online hate speech agree on make technology-driven interventions useful for preventing online hate perpetration?* Technology-driven interventions are understood as tools and strategies using technological features or being implemented into social media environments to mitigate influences that facilitate online hate perpetration. While mostly focused on prevention, these interventions may include detection and countering of hate speech as part of their preventive effort.

Besides being facilitated by technological features, the production and dissemination of online hate speech results from individual user behaviors and business models of social media platforms focused on maximizing engagement, e.g., by popularizing emotionally charged content.[20,21] To allow for a holistic approach to RQ2, experts provide their evaluations of intervention possibilities by using the three lenses of human-centered design (HCD). This framework proposes that effective solutions to complex problems are found when user needs are met under consideration of the possibilities and limitations of technology and the economic interests of stakeholders.[22–24]

Theory

Despite contributing to the spread of hate speech, technological features of social media platforms such as comment functions or pop-up notifications have also been used to counter online aggressive behaviors.[25,26] Thus, technological features can serve as determinants for online hate perpetration, but can also be crucial for its prevention. The concept of digital affordances helps to approach this multiplicity.

Digital affordances encompass both the features of a digital environment and the users. They refer to actions made possible through the relations and interactions between users and objects.[27] In other words, digital affordances are user behaviors enabled or constrained through technological features.[28] These features are not deterministic: They provide possibilities for action, but realizing these actions partly depends on the perceptions and capabilities of the user.[27,29] In the following, we describe two affordance frameworks for the analysis of technological features in the context of online hate speech.

The mechanisms and conditions framework by Davis moves beyond binary conceptions of whether a feature serves a particular function or not by examining how it enables user actions, for whom and under which circumstances.[29] While mechanisms of affordances explain the directionality and intensity with which features make actions possible (e.g., by initiating or responding to an action request they push or pull user behavior in a certain direction), conditions of affordances specify the user to whom action possibilities are afforded (i.e., accounting for the user's knowledge, skills, and social and political contexts).[29]

The mechanisms and conditions framework considers affordances from the perspective of the individual user. However, posting online hate speech on social media is not always an individual act, but can also result from group dynamics and social norms.[30–32] Thus, it is relevant to investigate how technological features on social media afford the formation of groups, how they structure interactions, distribute agency within these groups and affect social norms.[33]

An approach for doing so is provided by the taxonomy of social network platform affordances for group interactions, consisting of interaction and intervention affordances.[33] Group interactions, enabled by technological features, are distinguished by their degree of openness and by how much the involved actors are connected. Intervention affordances consider how features give agency to users. Both types of affordances influence norms for socially acceptable behavior, which bears consequences for the handling of hate speech.[33]

Method

Delphi studies allow for a structured compiling of experts' viewpoints through surveys, making it an effective method for collecting insights from various disciplines.[34]The main aim of a Delphi study is to synthesize these insights by striving for a group consensus.[35] Through multiple rounds of anonymous surveys building up on each other's findings, data are systematically narrowed down from a broad overview to a concise voting.[36] Between rounds, participants receive overviews of their own answers and preliminary results.[37] The method has been applied to understand harmful behaviors such as bullying, its antecedents, and possible intervention measures.[18,38,39] It has, to our best knowledge, not been applied to online hate speech yet.

Table 1. Field of study (researchers) and occupations (practitioners)

Field of study	Occupation	Count	Percentage (of *n* = 28)
Sociology		6	21.4 %
Communication Studies		5	17.9 %
Criminology		4	14.3 %
Computer Science		3	10.7 %
Psychology		2	7.1 %
Law		2	7.1 %
Cultural Studies		1	3.6 %
Subtotal		16	57.0%
	Monitoring expert	5	17.9 %
	(Vice-)CEO of an NGO	2	7.1 %
	Advisor on online hate related topics	1	3.6 %
	Content creator	1	3.6 %
	Online activism project founder	1	3.6 %
	Staff member non-governmental sector	1	3.6 %
	Trainer and course-developer specialized in countering extremism	1	3.6 %
	Subtotal	11	39%
Unknown	Unknown	1	3.6 %

Note: multiple mentioning of fields of study or occupation allowed.
Unknown refers to survey participant who did not provide demographic data but completed survey otherwise.

In contrast to conventional surveys, Delphi studies include elements of qualitative research, such as open-ended questions, and work with comparatively small sample sizes (less than 30 participants are common).[40] The sampling process emphasizes purposeful selection based on participants' expertise rather than probability sampling.[38,41] It is possible to also include practitioners in the sample, which is important when studying a timely topic such as online hate speech whose recent developments may not be reflected in academic publications yet.[42]

This study specifically discusses technology-related determinants and intervention ideas. Beyond that, the surveys included questions on personal and social determinants. The overall aim was to establish a holistic view on determinants for online hate perpetration across different levels of influence. The analysis presented here concentrates on the particular role of digital determinants as digitization permeates every area of daily life, enabling many possible interactions of digital determinants with personal and social influences.

In the first open-ended survey participants were asked to list and explain digital determinants for online hate perpetration and ways of using technology to prevent users from posting online hate speech. The results were processed through thematic analysis and provided the structure and content of the second survey,[42] in which participants rated each determinant for relevance on a 5-point Likert scale from *not at all relevant* to *extremely relevant* (answer option *cannot specify* was added). Participants also assessed technology-driven intervention ideas through the three lenses of

by indicating whether an idea was desirable for users, technologically feasible, and economically viable. To compensate for the lack of nuance of the dichotomous questions, we added comment fields for further elaboration. Both surveys provided definitions for hate speech to avoid misunderstandings. Basic demographics and occupational information were recorded. The surveys were administered through Qualtrics from June 2021 to January 2022 and approved by the institutional ethics review committee.

In a multi-stage process of purposive sampling, we first invited the authors of systematic reviews related to the keywords "hate speech", "online hate" and "cyberhate", as well as the first authors of the studies discussed in the reviews. The same search terms were used on Web of Science to identify additional researchers. They were invited to our sample if their academic profiles (publication record, CV, description of research interests) emphasized online hate speech as primary field of expertise. Websites such as the

International Network Against Cyber Hate (INACH) were searched for organizations active in monitoring, awareness raising, or prevention of hate speech to determine practitioners. We used snowball sampling techniques by asking participants and members of our professional networks to nominate other knowledgeable experts.

We contacted 75 researchers, practitioners, and organizations of whom 28 individuals (57% researchers, 38% female) participated and 13 completed both surveys. Response rates were 26.7% in the first and 53.2% in the second round[a]. The majority of the sample has over 14 years of experience in their current position (21%) and holds a Master's degree (41%). For further information on fields of expertise, see Table 1. Participants are affiliated with institutions and organizations in 16 countries across Europe, North America, and Australia.

Results

Determinants for online hate perpetration

The first survey resulted in 272 comments on determinants and intervention ideas. Thematic analysis was used to evaluate the results, allowing for an inductive discovery of structural patterns in the data.[43] In this process, an initial set of codes was established from the data to iteratively create themes that describe subsets of the data. Each theme was illustrated by examples of concrete determinants. These served as items in the second survey, which was analyzed by using non-parametric statistics. Considering the small sample size, the median serves as measure for relevance (*median* > 4 indicating high relevance).[41,44] The interquartile range shows how much opinions deviated (*IQR* ≤ 1 indicating group

[a] Due to the high survey duration in the second round, a monetary compensation was offered which might have contributed to the increased response rate.

consensus, see Table 2).[45] For an overview of all determinants, themes, and relevance ratings, see Table 2.

Table 2. Digital determinants for online hate perpetration

Theme	Survey item	Median	IQR
Communicative affordances (1)	***Share function***	***5***	***1***
Functionalities	*Like function*	*4*	*0.5*
	Extended emoji reactions such as laughing (may appear offensive or demeaning, potentially triggering hate reactions)	3	1
	Comment function	***5***	***1***
	Follow or befriend function	3	2.5
Community-building practices	*“Small-world” enabled by ICTs: global connectedness allowing to find like-minded people around the world*	*4*	*1.5*
	Homophily: tendency to form ties with similar others	***5***	***1***
	Isolated online communities, formation of echo chambers	*4*	*1*
Design and discursive affordances (2)	No or low requirements to create user account (e.g., no identity verification)	4	2
Design choices and user perceptions	*No or low accountability for online actions caused by anonymity or pseudonymity*	*4*	*1*
	Protection from persecution due to anonymity or pseudonymity	***4***	***1***
	Distance from victims due to anonymity or pseudonymity	***4.5***	***1***
	Immediacy of online communication	4	1.5
	Absence of design friction (e.g., two-step verification before posting)	*4*	*1*
Norms	*Perceived positive reputational impact (gaining notoriety on a platform through conspicuous posts)*	*4*	*0.75*
	Dominance of visible users (affirmation through volume)	*4*	*0.5*
	Normalisation or acceptability of hate through availability	***5***	***1***
	Perceived legitimisation due to lack of consequences	4	1.5
	(Perceived) lack of jurisdiction: no or limited recognition of "what's illegal offline is also illegal online"	4	2
Structure of social media (3)	***Personalized recommendation through algorithms***	***5***	***1***
Technological infrastructure	*Escalating exposure through algorithms that recommend similar but more extreme content*	*4*	*1*
	Black box algorithms (their inner workings are hidden)	4	2
	Rewarding emotionally charged language by algorithms aiming to increase engagement	*4*	*1*
Organizational structure	Unpredictable and untransparent moderation practices	4	1.75
	Vague or hardly understandable community guidelines	3	2.75
	Impersonal or condescending tone in moderation messages	3	2
	Privacy (end-to-end encryption) in instant messaging	3	3
	Focus on technological over organizational optimization	4	2

	Public / semi-open character of platforms providing easy access to (hateful) content	*4*	*1*
	Slow or non-elimination of hateful content	***5***	***1***
	Opacity on how social media platforms are organized	3	2
Economic structures	User engagement as a business model	4	2
	Lack of sanctioning from platforms to sustain engagement	4	1.5
	Monetization of hate (allowing producers of hateful content to generate revenue)	***4.5***	***1***
	No or low costs for users to use social media	4	2.5
User agency (4)	*Availability of hateful content produces more hate*	*4*	*1*
	Use bots to increase visibility of content at large scale	4	2
	False news, misinformation, or disinformation	*4*	*1*
	Conspiracy theories	***5***	***1***
	Polarizing content (biased, manipulated or emotionally charged)	***5***	***1***

Italicized determinants represent those upon which consensus was reached and which are considered relevant (*Median* = 4) while italicized and bold determinants reach consensus and were considered highly relevant (*Median* > 4).
Theme descriptions:
(1) Functionalities on social media for interacting and building communities.
(2) Design characteristics of social media platforms and the user reactions and norms they give rise to.
(3) Technological, organizational and economic aspects of how social media platforms are structured and operate
(4) Describes the impact of users through the creation and dissemination of content.

Intervention ideas for online hate perpetration

The analysis of the intervention ideas similarly employed thematic analysis for the data of the first survey. Results from the second survey we compared for how useful the intervention ideas were rated based on the three lenses of HCD. Given the participants' diverse backgrounds, we expected their expertise to vary across the dimensions of the framework (user desirability, technological feasibility, and economic viability) and thus asked them to assess their own expertise on a scale from 0 to 100 (*no - very high expertise*). For each dimension we excluded responses of participants who scored less than 50 to ensure the data was grounded in a sufficient level of expertise (Table 3).

Table 3. Results of expert self-assessment: Votes for possessing ≥ 50% confidence in expertise

Dimension		Included answers (≥ 50)		
		Female *n* = 10	Male *n* = 15	Total *n* = 25
User desirability	Count	10	14	24
	%	100%	93.0%	96.0%
Technological feasibility	Count	8	12	20
	%	80.0%	80.0%	80.0%
Economic viability	Count	5	9	14
	%	50.0%	60.0%	56.0%

Inclusion criterium: rating of expertise at least
Note: It was verified that exclusion of answers did not lead to aggravated gender imbalance in the sample (addressing concerns that female participants may perceive their own expertise on technology and business dimensions as lower due to internalised stereotypes).

We summarized composite scores on desirability, feasibility, and viability of the intervention ideas into a total score that indicates their potential to prevent users from posting hate speech. This allowed us to rate the intervention ideas through the three lenses of HCD, compare them and identify the ones that are seen as most useful by the participants. A score of at least 75% on each dimension was set to indicate high usefulness. Three intervention ideas had an overall score of over 75% but scored lower on economic viability (Table 4). The experts further elaborated on the desirability, feasibility, and viability of the intervention ideas in textual comments. The analysis of these comments complements the numeric data, while also uncovering limitations in the applicability of the framework.

Seven participants (five researchers) were concerned that desirability of an intervention cannot be generalized due to differing motivations for using social media. While some users might welcome interventions exposing them to different viewpoints, others who seek belonging and connection with like-minded individuals might be pushed towards more polarized and less regulated spaces. A solution to this dilemma may be to aim for a balance between what is desirable for most users and what is acceptable for the user groups targeted by an intervention.

Despite positive evaluations of feasibility, skepticism was voiced by six researchers and three practitioners about how much interventions should rely on automated solutions. These concerns pertain to a lack of nuance in AI-driven hate speech detection and a need for more advanced personality profiling algorithms for targeted interventions. One researcher experienced in Natural Language Generation was especially skeptical about AI's abilities to successfully create counter narratives. However, it was proposed by two practitioners to experiment with such solutions in controlled environments or to use technological solutions under supervision of trained moderators.

Three researchers and one practitioner found economic viability difficult to assess and conditioned by user desirability. In their view, interventions such as structural changes would be implemented by social media companies only if desired by their users or made mandatory through legal regulation. However, three practitioners also noted that companies could exert some influence over how functionalities are perceived by users through advertising.

Five participants assigned responsibility for the prevalence of online hate speech to the social media companies by highlighting their shortcomings to realize effective countermeasures. Fittingly, many intervention ideas with positive evaluations – e.g., changing recommendation systems, making algorithms transparent or excluding risk content from being promoted - related to changing technological structures. This emphasizes that experts located the agency and responsibility for effective prevention mostly within the companies.

A recurring concern among five participants was that users may feel personally attacked or restricted in their freedom of speech by invasive interventions. Three practitioners however advocated for a shift in this debate, with one stating: “There need[s] to be a greater emphasis on the right to freedom from fear as a corollary to the right to free speech and the freedoms of expression”. Strong opposition was voiced to the suggestion of limiting anonymity online. Even though anonymity relates to relevant determinants, all seven comments on this topic mentioned the importance of protecting members of minority groups or activists from targeting and persecution online.

Table 4. Intervention Ideas: Themes, descriptions and ratings (total votes and percentages)

Theme	Intervention	Desirability (n = 24)		Feasibility (n = 20)		Viability (n = 14)		Total (n = 58)	
		Votes	%	Votes	%	Votes	%	Votes	%
Design Friction and Nudges (1)	Two-step verification as browser extension checks if text contains hate speech, asks if user really wants to post.	12	50.0%	19	95.0%	8	57.1%	39	67.2%
	Real-time (anti-racism, anti-sexism etc.) notification when a text contains hate speech that points out trigger words and makes suggestions to revise	14	58.3%	17	85.0%	8	57.1%	39	67.2%
Targeted Risk Prevention (2)	Users showing signs of radicalisation do not receive friend recommendations based on like-mindedness to prevent further radicalisation through social circles.	15	62.5%	18	90.0%	10	71.4%	43	74.1%
	Adjust recommendation algorithms to promote counter narratives to users who interact with extremist accounts.	**19**	**79.2%**	**18**	**90.0%**	**12**	**85.7%**	**49**	**84.5%**
	Stricter age controls and limits to content access for minors.	14	58.3%	13	65.0%	6	42.9%	33	56.9%
	Identify influential public accounts that contribute to hate speech escalation and apply stricter regulations for them (e.g., no monetisation, stricter monitoring and moderation).	16	66.7%	18	90.0%	9	64.3%	43	74.1%
	Use technology to provide counselling (through a chatbot) or suggest in-person counselling to users with likeliness of repeated online hate perpetration.	13	56.5%	13	65.0%	9	64.3%	35	60.3%
Changing Technological Structure (3)	**Use technology to make people with different (political) viewpoints more visible to each other and encourage meaningful civil interactions between them.**	**21**	**87.5%**	**19**	**95.0%**	**12**	**85.7%**	**52**	**89.7%**
	Give users recommendations about content, communities or people to follow which they don't know yet, introducing them to new ideas and challenging their viewpoints.	**22**	**91.7%**	**19**	**95.0%**	**13**	**92.9%**	**54**	**93.1%**

	Make algorithms and how they work more transparent by explaining to users why they see the content they are seeing.	*24*	*100%*	*19*	*95.0%*	*5*	*35.7%*	*48*	*82.8%*
	Use algorithms to balance hate with counter narratives. For every bit of hateful material a user is encounters they should see a counter narrative, too.	**21**	**87.5%**	**16**	**80.0%**	**12**	**85.7%**	**49**	**84.5%**
	Make social media platforms open source.	15	68.2%	13	65.0%	2	14.3%	30	51.7%
	Identify risk content and exclude it from being promoted or recommended by algorithms to reduce amplification (cf. Instagram's attempts to reduce amplification of self-harm content).	*19*	*79.2%*	*19*	*95.0%*	*10*	*71.4%*	*48*	*82.8%*
Moderation (4)	*Make moderation processes transparent and justify decisions for content removals to users adequately.*	*22*	*91.7%*	*18*	*90.0%*	*9*	*64.3%*	*49*	*84.5%*
	Remove harmful content but avoid blocking or de-platforming of users (except for bots and trolls).	11	47.8%	17	85.0%	10	71.4%	38	65.5%
	Moderation efforts should dismantle harmful content like fake news and conspiracy theories instead of removing it to avoid notions of censorship.	**18**	**75.0%**	**17**	**85.0%**	**12**	**85.7%**	**47**	**81.0%**
	Include more easily understandable and usable reporting functions on social media platforms to facilitate moderation (limiting visibility of harmful content and decreasing perceived lack of consequences for posting it).	**22**	**91.7%**	**19**	**95.0%**	**13**	**92.9%**	**54**	**93.1%**
Pro-Social Digital Environments (5)	Use Natural Language Generation to automate creation and promotion of positive narratives to increase visibility of pro-social content on social media and overshadow hate.	14	58.3%	14	70.0%	9	64.3%	37	63.8%
	Use pro-social bots to promote values like tolerance, acceptance or diversity to positively influence perceived social norms and reduce prejudice.	13	54.2%	17	85.0%	10	71.4%	40	69.0%

	Establish close cooperation between social media companies and administrators of large, influential communities. Provide tools to these communities to prevent harmful conduct and foster mutual respect.	**22**	**91.7%**	**20**	**100%**	**12**	**85.7%**	**54**	**93.1%**
Other (6)	Decrease opportunities for anonymity by requiring some form of identification from users and revelation of IP addresses in cases of reported hate crimes.	8	33.3%	16	80.0%	8	57.1%	32	55.2%
	Disrupt communicative processes that could evoke hateful behaviour by disabling comment sections.	12	50.0%	19	95.0%	7	50.0%	38	65.5%
	Help users track the time they spend online or on social media, combined with warnings or support to decrease online time.	13	54.2%	19	95.0%	4	28.6%	36	62.1%

Note: Intervention suggestions in bold received ≥ 75% of votes on all dimensions, italicized and underlined intervention suggestions received ≥ 75% of votes in total, but scored < 75% on at least one individual dimension.

Themes descriptions:

(1) Rapid and impulsive online communication could be disrupted by creating moments of reflection in which users could be nudged towards positive online behaviour.

(2) Opposite to one-fits-all solutions, identifying perpetrators or users at risk of perpetration and applying changes to their digital environments to prevent hate speech posting and radicalisation.

(3) Structural changes for example to the algorithms of social media platforms but also the mitigation of the negative influences caused by social media technology (e.g., regarding filter bubbles).

(4) Limiting visibility of hateful content and preventing of engagement with it through automated detection and removal.

(5) Creation of digital communities with prosocial norms built on shared interests that provide users with a sense of belonging and options for meaningful interaction to mitigate normalisation of hate.

(6) Changes in functionalities and design of social media platforms.

Two further themes, *Digital Citizenship Education* and *Organisational Changes*, were excluded from further analysis as the suggested interventions ideas were not technology-driven

Discussion

To answer RQ1, we collected expert assessments of the relevance of digital determinants for online hate perpetration and interpreted them under consideration of different affordance frameworks. Notably, none of the social media features in Table 2 should be understood as direct cause for online hate perpetration because data collected through the Delphi method is not suitable to make inferences about causal relations or the directionality of influence.

Besides, digital determinants interact with personal, social, and contextual factors, meaning that technological features may facilitate online hate perpetration under certain circumstances and for certain users, as suggested by the mechanisms and conditions framework.[29] For example, sharing and commenting content is not inherently harmful, but when combined with an algorithmic logic promoting visibility of emotionally agitative content or used by individuals for the purpose of attacking others, these functionalities can contribute to the spread of hate speech.[20] Personalized content recommendations correlate with online hate when repeatedly exposing users to hateful, radicalizing, or strongly polarizing content but may not do so otherwise.[1]

Companies have agency to intervene by changing recommendation systems, eliminating harmful content, or stopping its monetization. Thus, certain platform features and organizational decisions shape the visibility of content. Another influential feature within control of the platforms are the requirements for personal identification to open an account. While profiles affording anonymity serve to protect users, they can also create personal distance and decrease empathy with victims of hate.[1,27] If a lack of empathy with others is understood as social norm, account requirements can become determinants for online hate perpetration.

To answer RQ2, we asked participants to describe intervention ideas and evaluate their usefulness through the three lenses of HCD. We use the term intervention ideas as we understand the output of this study to reflect current needs for intervention and give directions for future development processes rather than providing concrete tools. Looking at the intervention ideas from an affordance perspective highlights how they can employ features that are determinants of online hate perpetration for the purpose of prevention. For example, comment sections can become places of counter speech and recommendation algorithms can be re-shaped to break filter bubbles and introduce new viewpoints to users. The availability of hateful content can be mitigated by interventions that dismantle or remove it while providing transparent explanations for this procedure.

The responsibility and action possibilities related to many features are located within the social media companies. They have considerable agency over how their platforms afford group interactions.[33] These design choices are intertwined with the group interactions that are fostered and whether these are prosocial or antisocial. However, design choices are not made in a vacuum and depend on how the platforms react to user needs. As discussed by our expert sample, anonymity, for instance, shall not be decreased in order to protect vulnerable users, but intervention strategies can leverage technological features to mitigate the negative impacts of anonymity and foster empathy and pro-social interactions.

Not all suggested intervention ideas represent responses to a specific determinant. In turn, several determinants are not addressed by any intervention idea. This disparity between determinants and mitigating factors could represent missed opportunities for creating more effective interventions. Based on the expert viewpoints this study yields, we suggest for future intervention development to bridge this gap by focusing on the specific aspects that give social media features a facilitating or mitigating influence on online hate perpetration.

The discussed affordance frameworks can help to make sense of the multiplicity of technological features in social media environments and to ensure that the development of interventions accounts for differences between users, their motivations and needs, and the differing levels of agency among actors.[29,33] Furthermore, these perspectives account for how design choices on social media platforms impact user behavior and how interventions could induce behavior changes. Evaluating feasibility and viability could further improve assessments of an intervention's long-term and large-scale implementation possibilities.[46]

To establish concrete guidelines on intervention development, more data on how the suggested intervention ideas are intended to function and be implemented would be helpful. Given that Delphi studies are narrowed down to numerical outcomes in later rounds, this study would have benefitted from obtaining detailed information through in-depth interviews. While this was beyond feasible commitment for our participants, we recommend future research to consider such methodological combinations. With regards to the three lenses of HCD, future research should extend the framework to also address political and ethical considerations as well as differences in user needs and motivations to reflect on the usefulness of interventions. This is especially important in the context of a sensitive topic such as online hate speech.[47]

This study aimed to deepen the understanding of digital determinants for online hate perpetration and possible technology-driven interventions for its prevention by giving a comprehensive overview of what experts in the field of online hate speech regard as relevant determinants and useful intervention ideas. By applying an affordance perspective to data generated by the Delphi Method, we have furthermore shown how different technological features can be determinants or components of interventions. A continuation of this exploration may be a crucial part in the development of future strategies for the prevention of online hate speech.

Conflict of Interest

All authors declare that they have no conflicts of interest to disclose.

Funding statement

This research was conducted as part of a program that receives funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 861047.

References

1. Bilewicz M, Soral W. Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization. Polit Psychol 2020;41(S1):3–33; doi: 10.1111/pops.12670.

2. ECRI European Commission against Racism and Intolerance. ECRI General Policy Recommendation No. 15 on Combating Hate Speech. 2016.

3. Cohen-Almagor R. Fighting Hate and Bigotry on the Internet. Policy Internet 2011;3(3):89–114; doi: 10.2202/1944-2866.1059.

4. Wachs S, Wright MF. Associations between bystanders and perpetrators of online hate: The moderating role of toxic online disinhibition. Int J Environ Res Public Health 2018;15(9); doi: 10.3390/ijerph15092030.

5. Article 19. ‘Hate Speech’ Explained. A Toolkit. 2015.

6. Brown A. What is so special about online (as compared to offline) hate speech? Ethnicities 2018;18(3):297–326; doi: 10.1177/1468796817709846.

7. Gonçalves J, Weber I, Masullo GM, et al. Common sense or censorship: How algorithmic moderators and message type influence perceptions of online content deletion. New Media Soc 2021; doi: 10.1177/14614448211032310.

8. Roussos G, Dovidio JF. Hate Speech Is in the Eye of the Beholder: The Influence of Racial Attitudes and Freedom of Speech Beliefs on Perceptions of Racially Motivated Threats of Violence. Soc Psychol Personal Sci 2018;9(2):176–185; doi: 10.1177/1948550617748728.

9. Vidgen B, Margetts H, Harris A. How Much Online Abuse Is There? A Systematic Review of Evidence for the UK (Policy Briefing). The Alan Turing Institute: London; 2019.

10. De Smedt T, Lemmens S, Cooke A, et al. Creative Counternarratives Against Hate Speech. Technical Report. Detect Then Act; 2021.

Hamann G. Pöbeln, Hassen, Karriere Machen. Zeit Online 2022. Available from: https://www.zeit.de/2022/10/linkedin- hatespeech-corona-leugner [Last accessed: June 28, 2022].

12. boyd danah. Social Network Sites as Networked Publics. Affordances, Dynamics, and Implications. In: A Networked Self. Identity, Community, and Culture on Social Network Sites. (Papacharissi Z. ed) Routledge: New York; 2011.

13. Merrill S, Åkerlund M. Standing Up for Sweden? The Racist Discourses, Architectures and Affordances of an Anti-Immigration Facebook Group. J Comput-Mediat Commun 2018;23(6):332–353; doi: 10.1093/jcmc/zmy018.

14. Munn L. Angry by design: toxic communication and technical architectures. Humanit Soc Sci Commun 2020;7(1):53; doi: 10.1057/s41599-020-00550-7.

15. Bossens E, Geerts D, Storms E, et al. RHETORiC: An Audience Conversation Tool That Restores Civility in News Comment Sections. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts ACM: New Orleans LA USA; 2022; pp. 1–7; doi: 10.1145/3491101.3503560.

16. Schmitt JB, Rieger D, Rutkowski O, et al. Counter-messages as Prevention or Promotion of Extremism?! The Potential Role of YouTube. J Commun 2018;68(4):780–808; doi: 10.1093/joc/jqy029.

17. Tontodimamma A, Nissi E, Sarra A, et al. Thirty years of research into hate speech: topics of interest and their evolution. Scientometrics 2021;126(1):157–179; doi: 10.1007/s11192-020-03737-6.

18. Jacobs NCL, Dehue F, Völlink T, et al. Determinants of adolescents' ineffective and improved coping with cyberbullying: A Delphi study. J Adolesc 2014;37(4):373–385; doi: 10.1016/j.adolescence.2014.02.011.

19. Windisch S, Wiedlitzka S, Olaghere A. PROTOCOL: Online interventions for reducing hate speech and cyberhate: A systematic review. Campbell Syst Rev 2021;17(1); doi: 10.1002/cl2.1133.

20. Brady WJ, Wills JA, Jost JT, et al. Emotion shapes the diffusion of moralized content in social networks. Proc Natl Acad Sci 2017;114(28):7313–7318; doi: 10.1073/pnas.1618923114.

21. Haidt J, Rose-Stockwell T. The Dark Psychology of Social Networks. Why It Feels like Everything Is Going Haywire. Atlantic 2019. Available from: https://www.theatlantic .com/magazine/archive/2019/12/social-media-democracy/ 600763/ [Last accessed: January 30, 2023].

22. Fenn T, Hobbs J. Conceiving and Applying Relationship Models for Design Strategy. In: Research into Design for Communities, Volume 2. (Chakrabarti A, Chakrabarti D. eds). Smart Innovation, Systems and Technologies Springer Singapore: Singapore; 2017; pp. 517–528; doi: 10.1007/978-981-10-3521-0_45.

23. Friis Dam R, Yu Siang T. From Prototype to Product: Ensuring Your Solution Is Feasible and Viable. 2021. Available from: https://www.interaction-design.org/literature/article/from-prototype-to-product-ensuring-your-solution-is-feasible-and-viable. [Last accessed: June 22, 2022].

24. IDEO.org. The Field Guide to Human-Centered Design: Design Kit. 1st edition. IDEO: San Francisco; 2015. Available from: https://www.designkit.org/resources/1.html [Last accessed: February 2, 2023].

25. Taylor SH, DiFranzo D, Choi YH, et al. Accountability and Empathy by Design: Encouraging Bystander Intervention to Cyberbullying on Social Media. Proc ACM Hum-Comput Interact 2019;3(CSCW):1–26; doi: 10.1145/3359220.

26. Van Royen K, Poels K, Vandebosch H, et al. “Thinking before posting?” Reducing cyber harassment on social networking sites through a reflective message. Comput Hum Behav 2017;66:345–352; doi: 10.1016/j.chb.2016.09.040.

27. Evans SK, Pearce KE, Vitak J, et al. Explicating Affordances: A Conceptual Framework for Understanding Affordances in Communication Research: EXPLICATING AFFORDANCES. J Comput-Mediat Commun 2017;22(1):35–52; doi: 10.1111/jcc4.12180.

28. Bucher T, Helmond A. The affordances of social media platforms. In: SAGE Handbook of Social Media (Burgess J, Marwick A, Poell T. eds.) SAGE Publications: London; 2017; pp. 233–253.

29. Davis JL. How Artifacts Afford: The Power and Politics of Everyday Things. Design Thinking, Design Theory. The MIT Press: Cambridge, Massachusetts; 2020.

30. Walther JB. Social media and online hate. Curr Opin Psychol 2022;45:101298; doi: 10.1016/j.copsyc.2021.12.010.

31. Siegel AA, Badaan V. #No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online. Am Polit Sci Rev 2020;114(3):837–855; doi: 10.1017/S0003055420000283.

32. Soral W, Liu JH, Bilewicz M. Media of Contempt: Social Media Consumption Predicts Normative Acceptance of Anti-Muslim Hate Speech and Islamoprejudice. Int J Confl Violence 2020;14:1–13.

33. Van Raemdonck N, Pierson J. Taxonomy of Social Network Platform Affordances for Group Interactions. In: 2021 14th CMI International Conference - Critical ICT

Infrastructures and Platforms (CMI) 2021; pp. 1–8; doi: 10.1109/CMI53512.2021.9663773.

34. Okoli C, Pawlowski SD. The Delphi method as a research tool: an example, design considerations and applications. Inf Manage 2004;42(1):15–29; doi: 10.1016/j.im.2003.11.002.

35. Buettner D, Nelson T, Veenhoven R. Ways to Greater Happiness: A Delphi Study. J Happiness Stud 2020;21(8):2789–2806; doi: 10.1007/s10902-019-00199-3.

36. Hirschhorn F. Reflections on the application of the Delphi method: lessons from a case in public transport research. Int J Soc Res Methodol 2019;22(3):309–322; doi: 10.1080/13645579.2018.1543841.

37. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique: Delphi survey technique. J Adv Nurs 2000;32(4):1008–1015; doi: 10.1046/j.1365-2648.2000.t01-1-01567.x.

38. Lassiter BJ, Bostain NS, Lentz C. Best Practices for Early Bystander Intervention Training on Workplace Intimate Partner Violence and Workplace Bullying. J Interpers Violence 2021;36(11–12):5813–5837; doi: 10.1177/0886260518807907.

39. Rodríguez-Carballeira Á, Solanelles JE, Vinacua BV, et al. Categorization and Hierarchy of Workplace Bullying Strategies: A Delphi Survey. Span J Psychol 2010;13(1):297–308; doi: 10.1017/S1138741600003875.

40. Wahid SS, Ottman K, Hudhud R, et al. Identifying risk factors and detection strategies for adolescent depression in diverse global settings: A Delphi consensus study. J Affect Disord 2021;279:66–74; doi: 10.1016/j.jad.2020.09.098.

41. Wakefield R, Watson T. A reappraisal of Delphi 2.0 for public relations research. Public Relat Rev 2014;40(3):577–584; doi: 10.1016/j.pubrev.2013.12.004.

42. Foster CJ, Plant KL, Stanton NA. A Delphi study of human factors methods for the evaluation of adaptation in safety-related organisations. Saf Sci 2020;131:104933; doi: 10.1016/j.ssci.2020.104933.

43. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77–101; doi: 10.1191/1478088706qp063oa.

44. Setkowski K, van Balkom AJLM, Dongelmans DA, et al. Prioritizing suicide prevention guideline recommendations in specialist mental healthcare: a Delphi study. BMC Psychiatry 2020;20(1):55; doi: 10.1186/s12888-020-2465-0.

45. Ramos D, Arezes P, Afonso P. Application of the Delphi Method for the inclusion of externalities in occupational safety and health analysis. Dyna 2016;83(196):14–20.

46. Jecan D. Innovation by Design: How to Unlock New Business Opportunities. 2019. Available from: https://uxstudioteam.com/ux-blog/innovation-by-design/ [Last accessed: June 22, 2022].

47. Gartside N. Expanding the ‘3 Lenses’: Beyond Viability, Feasibility & Desirability. 2021. Available from: https://uxdesign.cc/expanding-the-three-lenses-the-case-for-

innovation-frameworks-that-look-beyond-viability-12701e2c234a [Last accessed: June 28, 2022].