Università
della
Svizzera
italiana

Institute
of
Digital
Technologies
for
Communication

Final Report

Governance of Digital Hate Speech from the Users' Perspective
Governance digitaler Hassrede aus Perspektive der Nutzer*innen

April 2024

Prof. Katharina Lobinger
Università della Svizzera italiana (USI)

Federico Lucchesi
Università della Svizzera italiana (USI)

Dr. Rebecca Venema
University of Amsterdam (UvA)

Acknowledgements
We would like to thank the Federal Office of Communications (OFCOM) for the financial support that made this project possible in the first place. We would also like to thank Thomas Häussler for his ongoing guidance and support during the implementation of the project. We are grateful to Dr. Seraina Tarnutzer, who contributed to the project by conducting the French- and German-speaking focus groups interviews and by participating in the affordance analysis. Furthermore, we would like to express our gratitude to Dahlia Pilar Navarro Hoffleuchter, Eléonore Fortier, Auréliane Dubuis, and Nina Gambone, the student assistants who have collaborated with us and assisted us in all steps of the project.

Data Availability Statement
The verbatim transcriptions of the focus groups' interviews, conducted in Italian, French, and Swiss German, will be made available on the open data repository *SWISSUBase.*

Table of Contents

1. BACKGROUND, STATE OF RESEARCH, AND RESEARCH GAPS 5

2. PROJECT DESCRIPTION 7

3. METHODOLOGICAL IMPLEMENTATION 7

4. RESULTS 10

4.1. Affordances Analysis (Work Package I) 10
4.2. Focus Groups: Imagined Affordances (Work Package II) 11

5. CONCLUSIONS 14

6. CONTACT INFORMATION 16

7. BIBLIOGRAPHY 17

Abstract

Dealing with digital hate speech is a fundamental societal and political challenge and task (see, e.g., Kuehn & Salter, 2020; European Commission, 2016) not least to the high prevalence of hate speech in current societies (for Swiss data see e.g., Stahel et al., 2022). The question arises as to which actors *are* responsible for the governance of digital hate speech or *should be* responsible, and what possibilities for intervention they have (see, e.g., Helberger et al., 2018; Heldt, 2019). This project examines the hitherto less focused action options and perspectives of users in research. The project investigates affordances, i.e., functions and features, as well as information for users on possibilities for intervening against hate speech on social media platforms and in comment sections of news sites (Package I, affordance analysis). In addition, it examines how users perceive such functions as well as the perceived potential but also the problems in the governance of hate speech (Package II, focus groups). Thus, the project identifies important starting points for political and platform-side measures and can demonstrate whether and where adjustments or, for example, communication with users, information campaigns, or media educational interventions are appropriate to strengthen the fight against digital hate speech and the agency of users. The results of the project can thus complement approaches to regulation, co-regulation, and self-regulation for the governance of digital hate speech.

Our research reveals a notable gap between social media and Swiss news sites with respect to the perceived affordances and the expected responsibility for tackling hate speech. Users exhibit a lack of knowledge concerning reporting hate speech on Swiss news sites, which can be attributed to, on the one hand, the limited information provided by Swiss news sites and, on the other hand, a firm belief in their editorial responsibility. The findings can be read as a need for augmenting transparency on Swiss news sites to develop moderation practices that align with user expectations for combatting online hate speech more effectively. Conversely, users are familiar with the mechanisms for reporting inappropriate content on social media platforms. However, they express concerns about the effectiveness of these mechanisms, highlighting the limitations of automated content moderation and its susceptibility to errors. Despite recognizing these flaws in addressing hate speech online, users often do not hold social media platforms accountable. This underscores the importance of enhancing digital literacy to foster a more responsible view of social media platforms in their critical role in combating hate speech and protecting online spaces.

1. Background, State of Research, and Research Gaps

The governance of digital hate speech concerns structures and rules of social communication (Katzenbach, 2021). In a broad understanding of governance (Burris et al., 2008; Woolgar & Neyland, 2013), governance is the “totality of various, coexisting forms of (collective) regulation of societal issues; a form of reflexive coordination in which actors negotiate rules of coexistence, mutual expectations, and legitimate institutional structures” (Katzenbach, 2021, p. 3, also see Katzenbach, 2018, translated by authors). The governance of platforms, and specifically the governance of digital hate speech, is closely linked to the preservation and negotiation of values and norms in digital societies (van Dijck, 2020). It involves an interplay between platforms as providers and “architects” of digital infrastructures, users as individuals who use platforms in certain ways and decide on their behavior and practices, (supra)national political actors who establish the respective legal-regulatory frameworks, and civil society actors who demand for or contribute interventions (Gorwa, 2019; Siapera & Viejo-Otero, 2021). Governance is thus a multifaceted process (Helberger et al., 2018), in which different actors bear responsibility. This can make it difficult to hold individual actors accountable for actions, which Helberger et al. (2018) discuss as the “problem of many hands,” and Lobinger and Brantner (2022) as an issue of “distributed responsibility.”

Models of regulation, co-regulation, or self-regulation (see, e.g., Stockmann, 2022; Brousseau et al., 2012) focus on (supra)national political legislation as well as measures and obligations of platforms or journalistic institutions. In recent years, attention has increasingly turned to the possibilities and limitations of algorithmic governance in both public and academic debates (e.g., Elkin-Koren, 2020; Gillespie, 2020; Gorwa et al., 2020; Katzenbach & Ulbricht, 2019, see also the project “Stop Hate Speech” funded by Innosuisse with the so-called “Bot Dog”, https://stophatespeech.ch/). Similarly, users and their “individual responsibility to make the choices that help create social order online” (Johnson et al., 2004, p. 33) play a role in the "Internet governance mosaic" (Dutton & Peltu, 2007, p. 63) and in the governance of hate speech.

The present project, therefore, explores options for users to intervene against hate speech. The frequently sought counter-speech can be an important but potentially very exposing strategy for users. Therefore, we particularly focus on low-threshold options for action on social media platforms and in the comment sections of news sites. These are the places where users most commonly observe or are confronted with hate speech (e.g., Chen, 2017; Stahel, 2020).

How users understand and play their role in the governance of hate speech is strongly influenced by the rules and (technical) governance mechanisms established and provided by social media platforms and news sites. These include, for example, community guidelines, terms of service, but also technical architectures and the so-called affordances of platforms and comment sections, i.e., functions and features that enable or prevent certain uses or interventions and thus shape actions (e.g., Bucher & Helmond, 2018; deNardis & Hackl, 2015; Evans et al., 2017; Flyverbom, 2016; Gillespie, 2018b). Indeed, affordances can amplify so-called “disinhibition effects.” This means they can facilitate people doing or saying things in digital environments that they otherwise would not do or say. This applies to both toxic communication practices like hate speech and prosocial practices (Springer et al., 2022). Another crucial element of governance is content moderation (Gillespie, 2018a; Roberts, 2019). In this regard, users play a key role as initiators; for example, through so-called “flagging” (Crawford & Gillespie, 2016). Through this feature, users can signal and report content such as hate speech or content that violates the established rules or personal and human rights (Crawford & Gillespie, 2016). Users thus actively participate in the continuous affirmation or renegotiation of shared (communicative) norms. In fact, an important form of content moderation is reactive (Klonick, 2018, p. 1638) or collaborative, interactive moderation (see Klonick, 2018; Springer & Naab, 2022): This means that platforms examine reported

contents or accounts/persons and may consequently act against “inacceptable behaviors,” e.g., by deleting or blocking them. Therefore, it is essential to examine the intervention options regarding digital hate speech that users have on social media platforms and news sites.

This project addresses two central research gaps. In recent years, numerous studies have focused on community guidelines, affordances, and rules of individual platforms or news sites for dealing with controversial and problematic content such as hate speech (Crawford & Gillespie, 2016; Fiesler et al., 2018; Jiang et al., 2020; Katzenbach, 2021; Konikoff, 2021; Maddox & Malson, 2020; Siapera & Viejo-Otero, 2021). These studies provide an essential basis for the present project. However, such affordances and rules are always “moving targets,” meaning they are subject to continuous change (Bucher & Helmond, 2018), for example, due to changing user behavior and preferences, but also due to changes in ownership (as currently discussed in the case of X, formerly Twitter). These changes also affect the ways in which hate posts and misinformation are handled or can be handled. We argue that, therefore, a current examination of platform- and site-specific affordances and information for intervention against hate speech is necessary.

Furthermore, affordances are always relational, meaning they must be considered in interaction with users. More fundamental than actual affordances are “imagined affordances” (Nagy & Neff, 2015), which refer to how users perceive and imagine certain affordances, such as those for intervening against hate speech, because these perceptions and imaginaries significantly shape users' actual actions. After all, their actions and the use of platform- or site-specific affordances depend on whether they know that certain options exist and on their beliefs about what these functions could accomplish. There is still a lack of studies on the perspective of users that focus on how they imagine affordances, how they perceive their personal options and room for maneuver for intervening against hate speech on social media platforms and news sites, or what other functions users would expect or wish for. This applies especially to the Swiss context. Instead, previous international research has shown, for example, how journalists and representatives of political parties perceive and use (technical) options for moderation and dealing with hate speech (e.g., Boberg et al., 2018; Frischlich et al., 2019; Kalsnes & Ihlebæk, 2021; Ksiazek & Springer, 2020). Additionally, researchers have focused on factors influencing the reporting of hate comments (Wilhelm et al., 2020). Studies have also examined how individuals perceive and imagine governance when they themselves are affected by governance measures; for example, when their social media accounts are suspended or their posts are deactivated (Myers West, 2018; Savolainen, 2022), when they are affected by moderation measures in comment sections of news sites (Løvlie et al., 2018), or when they are subjected to harassment, discrimination, and oppression (Duguay et al., 2020). Furthermore, studies have examined the attributed legitimacy of governance practices of various actors (Pan et al., 2022; Suzor et al., 2018) and have identified the lack of transparency as an important problem and obstacle regarding governance practices (Crawford & Gillespie, 2016; Gillespie, 2018b; Gorwa & Garton Ash, 2020). However, we argue that also the knowledge and beliefs of “regular” users regarding the functionalities and options for action in governance practices are crucial for them to engage in the fight against digital hate speech.

2. Project Description

Against the backdrop of the identified research gaps, in *Work Package 1*, we have analyzed the affordances, i.e., functions and features, as well as information for users regarding intervention opportunities against hate speech. In this first step, we have thus analyzed what information is provided regarding the intervention against hate speech and what specific respective options are available to users on social media platforms and in the comment sections of Swiss news sites.

This analysis was complemented with a focus on the perspective of users. In *Work Package 2*, we have explored the "imagined affordances"; that is, how users perceive and understand affordances and options for intervention against hate speech, what experiences they have had, what potentials and problems they see and which functions or procedures they wish for. This allowed us to explore why users use specific functions or choose not to use them.

In the following we will describe the methodological design of both studies.

3. Methodological Implementation

For Work Package 1 (Analysis of affordances), we have performed an analysis of platform- and site-specific affordances, self-descriptions of platforms/news sites, community guidelines, as well as terms of service and usage agreements. Considering both supranational platforms and nationally operating Swiss news sites equally has been one of the most innovative strengths of the project as it fulfilled a crucial research gap.

We focus the analysis on the most used social media platforms in Switzerland, i.e., Facebook, YouTube, Instagram, Twitter, TikTok (Statista, 2022; Kemp, 2023), as well as popular Swiss online news sites (Reuters Institute, 2021; WEMF, 2021). This includes high-traffic news sites from media institutions with different funding models and political orientations. We have included news sites in the three Swiss languages German, French, and Italian. The analysis of affordances was conducted between December 2022 and January 2023.

Overall, the sample thus included:

Social media platforms

- Facebook
- YouTube
- Instagram
- Twitter
- TikTok

German language Swiss news sites

- 20 Minuten online
- SRF online
- Blick
- Watson
- Neue Zürcher Zeitung
- Tages-Anzeiger

French language Swiss news sites

- 20 Minutes online
- RTS online
- Le Matin (including Sunday)
- Bluewin.ch
- 24heures
- Le Temps

Italian language Swiss news sites

- 20 minuti/Ticinonline
- RSI online
- Corriere del Ticino
- La Regione

For analyzing the affordances of the social media platforms and Swiss news sites (see also Bucher & Helmond, 2018), we have developed an overview of the functions and governance mechanisms based on the approach of Crawford and Gillespie (2016). For each platform and news site, we examined the technological features for reporting and all the respective information, such as guidelines, FAQs, and user agreements. We conducted the examination of the technological features both via mobile devices, such as smartphones, and via desktop. For testing the social media platforms, we created new profile accounts based in Switzerland. For testing the Swiss news sites, we created profiles, and we paid for subscriptions, where necessary, to test eventual changes linked to the users' enrolment.

For the analysis of affordances of the platforms and news sites (see also Bucher & Helmond, 2018), we have developed an overview of the functions and governance mechanisms based on the approach of Crawford and Gillespie (2016). For each social media platform and news site, by using a content-analytical qualitative approach (Kuckartz, 2014), we examined the technological features for reporting and all the respective information (such as guidelines, self-descriptions of platforms and news sites, FAQs, user agreements, terms of services). One of the main areas of focus was on the accessibility and cost structure of the platform or news site. We scrutinized whether it operated on a paid, free, or freemium model, and looked at the user registration and authentication processes. Specifically, we probed the requirements for creating profiles and logging in, including the types of information solicited, such as real names, pseudonyms, Swiss phone numbers, or emails, along with considerations of double authentication mechanisms. Another critical area of investigation regarded user-generated content and interactions. We investigated who could publish content, comment, or interact and the parameters governing such activities. This investigation included inquiries into the permissions granted to registered and non-registered users and the incentivization strategies employed to encourage engagement, such as commenting features such as statistics or symbolic rewards.

Furthermore, we explored community guidelines, netiquette standards, and measures for handling undesirable content, including hate speech. We probed into disseminating information regarding community guidelines, definitions of hate speech, and procedures for reporting objectionable content, alongside elucidating the responsibilities of both users and the platform in fostering a safe online environment. A crucial aspect scrutinized was the moderation framework deployed, encompassing human and/or machine-based moderation, the timing of moderation — whether pre- or post-publication — and the locus of moderation, whether it was done internally or externally of the platform/site. Moreover, we looked at the efficacy of moderation mechanisms in addressing reported content, including the options available to users for flagging content, the categories for specifying the reasons behind the flagging, and the possibility of providing explanatory remarks. Finally, we considered the repercussions of reported content, encompassing the feedback provided to users, actions

taken on reported content, such as removal or concealment, and the potential penalties imposed on content authors in the wake of moderation decisions. In sum, the analysis grasped the intricate interplay between platform affordances, user behaviors, and regulatory mechanisms, providing a nuanced understanding of the dynamics shaping user engagement and content moderation within social media platforms and Swiss news sites. Overall, we provided a current overview of platform- and site-specific affordances and information for intervention against hate speech.

For work package II (Analysis of imagined affordances), we conducted seven focus group interviews in three language regions of Switzerland (French-, German-, and Italian-speaking regions), to explore users' imagined affordances for addressing hate speech, as well as perceived problems and suggested improvements regarding platforms' governance. Each focus group consisted of 5-7 participants (34 participants, 18-62 y.o., M= 30,45 y.o.) and took place in Lugano, Lausanne, Zug, Fribourg, and Zürich. For composing the sample, we aimed for maximum structural variation regarding (regional) origin, age, education level, professional contexts, and the use of social media platforms and news sites. The focus groups were conducted to explore possibilities, experiences, desires, and problems regarding combating hate speech from the perspective of users. Visual elicitation techniques were used to foster discussion and narrations among the participants.

Following a semi-structured interview guide, the participants engaged with the interviewer in a structured exploration of their interactions with newspapers and social media platforms, discussing various dimensions of imagined affordances and perceived moderation mechanisms. Following a brief warm-up session, wherein the interviewer introduced the topic and procedures, participants were questioned about their usage patterns, eliciting insights into the platforms they use, the frequency, and the reasons behind their usage. Delving deeper into the perceived affordances, participants discussed their perspectives on combating hate speech and problematic content, reflecting on their experiences reporting such instances and the technical imagined functionalities available for flagging hate speech and problematic content. The participants were asked to describe the imagined reporting practice in detail by discussing the step-by-step process they expected to encounter when they wanted to report something on the platforms. The discussion then shifted towards perceived moderation practices, with participants sharing their beliefs regarding the timing and processes involved in content moderation and their expectations regarding feedback and actions taken by moderators and platforms. Finally, participants voiced their opinions on existing challenges and potential improvements in content moderation. The interview concluded with participants allowed to express any additional thoughts or concerns regarding hate speech on Swiss news sites and social media platforms.

All the interviews were video and audio recorded after proper consent was obtained from the participants. At the end of the focus groups, the interviews were transcribed *verbatim* and anonymized to guarantee the privacy of the participants. The interviews were then analyzed by using NVivo software with a content-analytical qualitative approach (Kuckartz, 2014) and adopting a combination of deductive-inductive category system (Schreier, 2014).

4. Results

4.1. Affordances Analysis (Work Package I)

In the affordances analysis we have examined 5 social media platforms and 16 Swiss news sites to understand users' possibilities for intervening against hate speech. The overall aim was to identify the current policies, reporting procedures, and available features and to understand whether platforms and news sites provide users with the necessary tools and resources to intervene against hate speech and promote a more positive and inclusive online environment.

Overall, the analysis showed that social media platforms have developed comprehensive and user-friendly mechanisms for reporting different kinds of content. On each platform, several dedicated pages provide step-by-step instructions to guide experienced and inexperienced users through the process of reporting and flagging inappropriate content. Furthermore, the platforms typically have dedicated guidelines on how to avoid the spread of hate speech, and how to effectively counter it. The reporting mechanism on social media platforms provides users with designated "report" buttons. The ways in which these flagging features are presented follow similar principles across various platforms, except for TikTok. Users initiate the process by clicking these buttons, which then guide them to a predetermined list of reasons for reporting, supplied by the platform itself. Sometimes, users can specify additional motivations by filling in text fields to specify reasons that are not included in the predefined categories. Social media content is then moderated a-posteriori by a combination of automated filters and algorithms, and human-based moderation.

Swiss news sites usually have community guidelines or netiquettes for commenting that users should read and must agree on before commenting or when creating a user profile. On various news sites, the guidelines also contain a specific definition of hate speech that can potentially increase the awareness of the users about the issue. In general, everyone can see the comments published online below the articles, except for NZZ that allows this option only for registered users. On Swiss news sites, flagging content is not based on fixed categories as is typical for social media platforms. Users are instead required to explain their reasons for reporting, mostly using an open text field or in some cases by writing an email. However, apart from these usual common points, the analysis of the guidelines and affordances of Swiss news sites in different linguistic regions showed decisive differences in their approaches towards user interaction and content moderation.

In the Italian-speaking region, commenting on articles is not possible on most of the examined news sites, except for 20 minuti/Ticinonline. The latter allows for user comments and for reporting hate speech via e-mail, however without any on-site features. On the other hand, Swiss news sites in the German-speaking area allow for commenting but have stricter protocols for participation. Users must typically register with their full names and surnames to engage in commenting and reporting activities such as flagging inappropriate content. The commenting feature is selective and is typically enabled only for a few limited specific articles. On these articles, moderation is primarily done by human moderators and is done a priori, with automated systems supplementing during off-hours. Users are encouraged to report inappropriate content using various mechanisms, such as flagging and e-mails. Not all news sites provide sufficiently explicit guidelines or a clear description of content moderation practices, potentially resulting in ambiguity for users. Finally, in the French-speaking parts of Switzerland news sites allow for registering under nicknames or pseudonyms, thus permitting anonymous comments and interactions. Typically, only registered users can interact with published user comments and thus report them. One exception is 20 minutes, which allows also non-registered users to report comments published by others. In this linguistic region, the

moderation strategy typically occurs a posteriori, is done by humans and is not systematic. In other words, problematic content is checked only after its publication, and the news sites do not guarantee that any content will be checked, nor that moderation will occur within a specific amount of time. Usually, if the checked content is found to be offensive, the news sites claim that internal moderators will remove the inappropriate content.

Generally, neither social media platforms nor Swiss news sites provide users with feedback about the outcomes of their reporting actions. Typically, after users report content, they do not receive further updates – just a short message of confirmation mentioning that their request has been received and will be examined – leaving users uninformed about any subsequent consequences or changes. However, YouTube was an exception in this regard: beyond receiving a “thankful” message after reporting a content, users can follow the outcome of the moderation process on a dedicated page called “Report history,” where the updates concerning all the flagged content are grouped.

Overall, the differences in moderation strategies and options for commenting and reporting across linguistic regions invite to reflect about the diverse approaches to managing public discourse on news platforms in Switzerland, which might be linked to discrepancies in cultural or regional preferences of users and media organizations.

4.2. Focus Groups: Imagined Affordances (Work Package II)

As the analysis of affordances has shown, the features that are provided (or not provided) for flagging hate speech, potentially yield a decisive difference in user experience between social media platforms and Swiss news sites. As found in our focus groups interviews, many users of Swiss news sites are in fact uncertain about how to report inappropriate content and have never done so, and many users are not even aware of the available reporting options. The problem is further exacerbated by the fact that respondents perceive Swiss news sites as lacking clear and detailed instructions for reporting, resulting in a gap in user awareness and action. Therefore, enhancing and making reporting processes more visible, for example with dedicated buttons and features, could be a way for Swiss news sites to increase user engagement and trust.

On the other hand, users are highly aware of reporting mechanisms on social media and have typically already reported content on various platforms. They consider social media platforms as providing more user-friendly and straightforward reporting mechanisms, except for TikTok, where users described the flagging features to be less intuitive. This lack of knowledge regarding TikTok, however, might be linked to the “newness” of the platform, which is mostly used by younger people.

There are some contrasting positions on the role of user comments, even though generally our participants agree on their important role and value. For example, some respondents suggest that disabling comments on Swiss news sites altogether might reduce harmful discussions. However, at the same time, the (same) participants perceived this strategy as potentially detrimental to free speech, underlining that eliminating interactive features, such as commenting, might not be the most effective approach. In particular, users underline the importance and relevance of having discussion spaces that enable the encounter of different points of views that might create stimulating conversations between users. And they generally state that they enjoy reading user comments. This finding suggests that the most prominent strategy adopted by Swiss news sites in the Italian-speaking part of Switzerland (i.e., totally disabling comments sections on the websites) is considered rather inappropriate by the users.

It must be emphasized that we have not examined the comment sections related to articles stemming from Swiss news sites that are shared or published on social media platforms; typically on dedicated social media accounts of the media organization. Research has, in fact, shown that there is a tendency of news sites and news organizations to move user interaction and thus also user comments to social media platforms for various reasons (see e.g., Springer & Naab, 2022).

Furthermore, our results suggest that users have different expectations regarding the perceived responsibilities of social media platforms and Swiss news sites in combating hate speech. For Swiss news sites, there is a strong expectation that the responsibility for addressing hate speech should rest entirely with the editorial and moderation teams rather than the users. These sites are expected to effectively moderate comments to ensure the comment sections remain respectful spaces of discussion. At the same time, they are also expected to respond promptly and transparently, providing proper feedback. Indeed, Swiss news sites are considered highly accountable for maintaining a healthy community and a safe discussion space that mirrors the quality of their editorial content. However, users also relate the aim of conserving serious and professional discussion spaces to specific editorial policies that might differ from news organization to news organization. For example, users expect that outlets which they consider to be more “serious” (mentioning particularly SRF or NZZ) care more about having healthy discussion spaces than newspapers which they perceive as less serious or as more “aggressive” (such as 20 Minuten online). Interestingly, our respondents believe that the volume of users’ comments on Swiss News sites should be manageable, and that for this reason Swiss news sites should have the necessary staff, expertise, and commitment to conduct careful, human-based moderation. Moreover, these sites should facilitate direct communication between moderators and users for reporting and feedback. However, users are also aware that moderation requires an economic investment. Moreover, it must be emphasized that we did not investigate the economic struggles of news journalism in general. In other words, we did not discuss with our participants whether their high expectations are realistic with respect to current economic models of journalistic organizations.

Quite differently to Swiss news sites, our participants consider the role of users on social media platforms as crucial, due to the overwhelming volume of content that is typically shared on such platforms. Most of our respondents believe that the vast amount of material necessitates and is addressed with the deployment of automated moderation systems. However, these systems are not perceived as trustworthy, as they could struggle to accurately identify inappropriate content because algorithms cannot always understand context and semantic nuances. Consequently, the role of users for reporting and flagging content is viewed as an essential mechanism for signaling hate speech in contents that might not be immediately “obvious” to algorithms.

These findings highlight a clear gap between the responsibilities assigned to Swiss news sites and to social media platforms in the realm of hate speech moderation. First of all, users have higher expectations regarding Swiss news sites than regarding social media. In this regard, due to presence of “tangible” individuals, such as journalists and editors on Swiss news sites, there is a clear line of accountability and personification that influences expectations. Users perceive a direct relationship with these individuals, fostering expectations of more considerate or personalized moderation practices.

Most importantly, when combating hate speech online, users demand clarification regarding the effectiveness of reporting practices. Generally, we found a lack of knowledge and a high level of insecurity about what happens after reporting or flagging. However, it must be emphasized that users have different positions on this aspect. While for some users, reporting unacceptable content is considered enough. Particularly after having reported content on social media platforms, they consider “their job done.” They are not interested in what happens afterwards, arguing that from now on it is the responsibility of the platforms. Generally,

however, our respondents would like to receive more precise feedback after reporting problematic content. Firstly, the lack of feedback makes users doubtful and reduces their trust in the platforms' commitment to addressing hate speech effectively. People feel that moderation decisions do not often meet their expectations of fairness and justice, further eroding their confidence in these systems. Moreover, users believe that individual efforts to report hate speech might be ineffective. Indeed, there is a common uncertainty regarding the usefulness of reporting content unless there is a large mass of users that reports the same problematic content. This belief contributes to a general disenchantment with the reporting process. As a result, many users tend to disengage from the reporting mechanism entirely. This disengagement takes several forms, including withdrawing from participating in comment sections on social media and news sites or hiding, ignoring, or unfollowing content and users that propagate hate speech instead of reporting. This trend poses a significant challenge for online platforms: Maintaining user engagement in moderation processes while ensuring that these processes meet user expectations for fairness and effectiveness. Comparing this perceived lack of feedback with the analysis of affordances shows that, effectively, the information provided on flagging and reporting usually stops with the description of features for reporting and how to use them. The steps after flagging, on the other hand, remain vague and totally in the hands of platforms and the news sites.

Overall, these results provide valuable insights into how, according to the users, Swiss news sites and social media platforms can improve their moderation practices to combat online hate speech more effectively and to better align with users' expectations. The suggestions for improvements are multi-faceted and address various aspects of the moderation process. Firstly, there is a strong call for providing more comprehensive feedback to users who report problematic content. This transparency could help build trust and encourage more active participation in the moderation process. Additionally, on Swiss news sites, expanding the size of moderation teams with economical investments ad-hoc, when doable, and ensuring they receive specialized training, is seen as crucial by some respondents. This would improve the efficiency of handling flags and accuracy in dealing with complex hate speech issues. Another recommended improvement is to increase the visibility of moderation efforts (e.g., make moderators' interventions visible). Making such efforts visible could act as a deterrent against users posting hateful comments. Especially for Swiss news sites, facilitating direct communication between users and moderators is also suggested, which would provide a clearer and more immediate way to discuss concerns in a more detailed manner and with more possibilities for explaining a personal point of view.

Concerning the adjustments aimed at improving 'flagging' processes, insights from the focus groups suggest that simplifying the reporting process on Swiss news sites and allowing non-logged-in users to flag inappropriate content would be highly appreciated. For example, this could involve implementing reporting mechanisms similar to those already known from social media platforms. For example, the introduction of 'dislike' or 'reporting' buttons (that are just available on few news sites) could make it easier for users to express disapproval and flag content. Generally, it has been emphasized that simple and straightforward flagging without additional barriers increases the willingness to engage in the moderation process. For instance, it was noted that on Swiss news sites, unregistered users often face limitations in reporting inappropriate content. The requirement to create a profile was identified as a significant barrier due to its time-consuming nature. Rather, the consensus was that anyone who encounters hateful content should be able to report it promptly. In contrast, social media platforms were mentioned as a good example. However, an intriguing aspect emerged: Users emphasized that no registration was required on social media platforms, as "You are already in". This aspect illustrates the perception that social media platforms are now viewed as everyday environments for which no registration is necessary. Participants also mentioned suggestions for improving their agency and role in reporting hate speech on social media. For example, they wish for spaces to explain their motivations for reporting, even when selecting from pre-defined categories, arguing that this would enrich the provided context needed by

human moderators, which eventually would lead to more nuanced decision-making. Finally, participants expect platforms and political actors to further develop strategies aimed at increasing educational literacy about digital media and hate speech prevention across different age groups. This could include media education measures and information campaigns, which would help build a more informed user base capable of recognizing and countering hate speech online when interacting on Swiss news sites and social media.

Overall, these comprehensive expectations and desires could be addressed to not only enhance the technical aspects of moderation but also foster more informed and proactive communities in combating online hate speech.

5. Conclusions

In conclusion, our project highlights certain aspects of Swiss news sites and social media platforms that, if improved, could be valuable tools in the battle against hate speech.

One of the key findings of our study is that Swiss news sites and social media platforms are perceived differently in terms of their responsibility in addressing hate speech. Social media platforms, due to the sheer volume of posts and user-generated content, are seen as unable to adequately moderate all contents. Therefore, the role of users in flagging and reporting is considered particularly important, in the sense of a reactive, collaborative, or interactive moderation. Moreover, automated moderation is perceived as inevitable but at the same time considered to be prone to errors and lacking oversight. These errors are considered, and are basically accepted, as an inevitable consequence of the quantity of data shared. In contrast, Swiss news sites are generally expected to ensure adequate moderation conducted by journalists themselves or by human moderators with appropriate training. This attribution of responsibility is associated with a high expectation of quality and trust in news sites. Users believe that proper and effective online content moderation is, in fact, just one of the many ways a news site conveys an image of professionalism and authority. Therefore, users expect news sites to be active, precise and timely in moderation.

Our research indicates a need for enhanced user awareness regarding the mechanisms available to address hate speech on Swiss news sites. Based on our focus group interviews, it appears that participants lack understanding of the processes involved in reporting and flagging on these platforms. This deficiency in awareness is, on the one hand, associated with the considerable responsibility attributed to Swiss news sites. Consequently, users anticipate and desire proactive and timely moderation by the news sites themselves, diminishing the perceived importance of the users' role and contribution in the moderation process. Furthermore, our affordance analysis suggests room for improvement in the clarity of how hate speech reporting procedures are presented on Swiss news sites. Reporting procedures need to be clearly described, and users advocate for quick and straightforward affordances for both registered and non-registered users. Remarkably, our examination across Switzerland's linguistic regions reveals divergent strategies implemented by Swiss news sites. In conclusion, we recommend enhancing the availability of information for users, such as moderation policies, guidelines, and reporting options, to empower them in addressing hate speech on Swiss news sites.

Another important finding concerns the lack of information regarding what happens after reporting, applicable to both Swiss news sites and social media platforms. There is uncertainty about whether reporting hate speech is effective at all. This result highlights that guidance and explanations regarding the reporting of hate speech should not end with the click of the reporting button. An essential component is feedback on the outcome of the evaluation. This signals to the user that their concern has indeed been addressed and conveys a result.

Furthermore, transparent information is necessary regarding whether reporting by an individual user can indeed be effective, or if mass flagging is necessary. This is a speculation that has been repeatedly voiced, indicating that the perceived efficacy of users is very limited and sometimes even implies a perception of disempowerment.

These findings emphasize the vital need for political, social, and educational institutions to prioritize enhancing digital literacy on social media. We argue that literacy is particularly required regarding the role and responsibility of platforms. As said in the introduction, the governance of platforms, and specifically the governance of digital hate speech, is closely linked to the preservation and negotiation of values and norms in digital societies (van Dijck, 2020). In this respect it is highly important that social media platforms are also perceived as responsible actors. In other words, we consider it important to make users become more conscious of the significant responsibility that social media platforms bear and should bear in addressing these issues, which directly impact our lives in digitized and datafied societies.

6. Contact Information

Prof. Dr. Katharina Lobinger
Institute of Digital Technologies for Communication
Faculty of Communication, Culture and Society

Università della Svizzera italiana
Via G. Buffi 13
CH-6900 Lugano
Tel: +41(0) 58 666 4544
katharina.lobinger@usi.ch
http://usi.to/wz3

7. Bibliography

Barbour, R. S. (2018). *Doing focus groups*. SAGE.

Boberg, S., Schatto-Eckrodt, T., Frischlich, L., & Quandt, T. (2018). The moral gatekeeper? Moderation and deletion of user-generated content in a leading news forum. *Media and Communication*, *6*(4), 58–69. https://doi.org/10.17645/mac.v6i4.1493

Brousseau, E., Marzouki, M., & Méadel, C. (Eds.). (2012). *Governance, regulations and powers on the Internet*. Cambridge University Press.

Bucher, T., & Helmond, A. (2018). The affordances of social media platforms. In J. Burgess, T. Poell, & A. Marwick (Eds.), *The SAGE handbook of social media* (pp. 233–253). SAGE.

Burris, S., Kempa, M., & Shearing, C. (2008). Changes in governance: A cross-disciplinary review of current scholarship. *Akron Law Review*, *41*(1), 1–61.

Chen, G. M. (2017). *Online incivility and public debate: Nasty talk*. Palgrave Macmillan.

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. https://doi.org/10.1177/1461444814543163

deNardis, L., & Hackl, A. M. (2015). Internet governance by social media platforms. *Telecommunications Policy*, *39*, 761–770.

Duguay, S., Burgess, J., & Suzor, N. (2020). Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. *Convergence: The International Journal of Research into New Media Technologies*, *26*(2), 237–252. https://doi.org/10.1177/1354856518781530

Dutton, W. H., & Peltu, M. (2007). The emerging internet governance mosaic: Connecting the pieces. *Information Polity*, *12*(1–2), 63–81.

Elkin-Koren, N. (2020). Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence. *Big Data & Society*, *7*(2), 1–13. https://doi.org/10.1177/2053951720932296

European Commission. (2016). *The EU code of conduct on countering illegal hate speech online*. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

Evans, S. K., Pearce, K. E., Vitak, J., & Treem, J. W. (2017). Explicating affordances: A conceptual framework for understanding affordances in communication research. *Journal of Computer-Mediated Communication*, *22*(1), 35–52. https://doi.org/10.1111/jcc4.12180

Fiesler, C., Jiang, J. A., McCann, J., Frye, K., & Brubaker, J. R. (2018). Reddit rules! Characterizing an ecosystem of governance. *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*, 72–81.

Flyverbom, M. (2016). Disclosing and concealing: Internet governance, information control and the management of visibility. *Internet Policy Review*, *5*(3), 1–15. https://doi.org/10.14763/2016.3.428

Frischlich, L., Boberg, S., & Quandt, T. (2019). Comment sections as targets of dark participation? Journalists' evaluation and moderation of deviant user comments. *Journalism Studies*, *20*(14), 2014–2033. https://doi.org/10.1080/1461670X.2018.1556320

Gillespie, T. (2018a). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T. (2018b). Governance of and by platforms. In J. Burgess, T. Poell, & A. Marwick (Eds.), *The SAGE handbook of social media* (pp. 254–278). SAGE.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2), 1–5. https://doi.org/10.1177/2053951720943234

Gorwa, R. (2019). What is platform governance? *Information, Communication & Society*, *22*(6), 854–871. https://doi.org/10.1080/1369118X.2019.1573914

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 1–15. https://doi.org/10.1177/2053951719897945

Gorwa, R., & Garton Ash, T. (2020). Democratic transparency in the platform society. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy: The state of the field, prospects for reform* (pp. 286–312). Cambridge University Press. https://doi.org/10.1017/9781108890960

Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *The Information Society*, *34*(1), 1–14. https://doi.org/10.1080/01972243.2017.1391913

Heldt, A. (2019). Let's meet halfway: Sharing new responsibilities in a digital age. *Journal of Information Policy*, *9*, 336–369.

Jiang, J. "Aaron", Middler, S., Brubaker, J. R., & Fiesler, C. (2020). Characterizing community guidelines on social media platforms. *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, 287–291. https://doi.org/10.1145/3406865.3418312

Johnson, D. R., Crawford, S. P., & Palfrey, J. G. (2004). The accountable net: Peer production of internet governance. *Virginia Journal of Law & Technology*, *9*(9), 1–33.

Kalsnes, B., & Ihlebæk, K. A. (2021). Hiding hate speech: Political moderation on Facebook. *Media, Culture & Society*, *43*(2), 326–342. https://doi.org/10.1177/0163443720957562

Katzenbach, C. (2018). *Die Regeln digitaler Kommunikation. Governance zwischen Norm, Diskurs und Technik*. Springer VS.

Katzenbach, C. (2021). Die Governance sozialer Medien. In J.-H. Schmidt & M. Taddicken (Eds.), *Handbuch Soziale Medien* (pp. 1–24). Springer Fachmedien. https://doi.org/10.1007/978-3-658-03895-3_26-1

Katzenbach, C., & Ulbricht, L. (2019). Algorithmic governance. *Internet Policy Review*, *8*(4), 1–18. https://doi.org/10.14763/2019.4.1424

Kemp, S. (2023). Digital 2023: Global overview report. DataReportal. https://datareportal.com/reports/digital-2023-global-overview-report

Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, *131*, 1598–1670.

Konikoff, D. (2021). Gatekeepers of toxicity: Reconceptualizing Twitter's abuse and hate speech policies. *Policy & Internet*, 502–521. https://doi.org/10.1002/poi3.265

Ksiazek, T. B., & Springer, N. (2020). *User comments and moderation in digital journalism: Disruptive engagement*. Routledge.

Kuckartz, U. (2014). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung*. Beltz Juventa.

Kuehn, K. M., & Salter, L. A. (2020). Assessing digital threats to democracy, and workable solutions: A review of the recent literature. *International Journal of Communication*, *14*, 2589–2610.

Lobinger, K., & Brantner, C. (2022). „Niemand muss diese Videos zeigen". Der medienethische Diskurs über die visuelle Berichterstattung zum Terroranschlag 2020 in Wien (pp. 253-277). In U. Autenrieth & C. Brantner (Eds.), *It's all about Video. Visuelle Kommunikation im Bann bewegter Bilder*. Herbert von Halem Verlag.

Løvlie, A. S., Ihlebæk, K. A., & Larsson, A. O. (2018). User experiences with editorial control in online newspaper comment fields. *Journalism Practice*, *12*(3), 362–381. https://doi.org/10.1080/17512786.2017.1293490

Lüthje, C. (2016). Die Gruppendiskussion in der Kommunikationswissenschaft. In S. Averbeck-Lietz & M. Meyen (Eds.), *Handbuch nicht standardisierte Methoden in der Kommunikationswissenschaft* (pp. 157–173). Springer VS.

Maddox, J., & Malson, J. (2020). Guidelines without lines, communities without borders: The marketplace of ideas and digital manifest destiny in social media platform policies. *Social Media + Society*, *6*(2), 1–10. https://doi.org/10.1177/2056305120926622

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383. https://doi.org/10.1177/1461444818773059

Nagy, P., & Neff, G. (2015). Imagined affordance: Reconstructing a keyword for communication theory. *Social Media + Society*, *1*(2), 1–9. https://doi.org/10.1177/2056305115603385

Pan, C. A., Yakhmi, S., Iyer, T. P., Strasnick, E., Zhang, A. X., & Bernstein, M. S. (2022). Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–31. https://doi.org/10.1145/3512929

Reuters Institute. (2021). *Reuters institute digital news report 2021*. University of Oxford. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Savolainen, L. (2022). The shadow banning controversy: Perceived governance and algorithmic folklore. *Media, Culture & Society*, 1–19. https://doi.org/10.1177/01634437221077174

Schreier, M. (2014). Varianten qualitativer Inhaltsanalyse: Ein Wegweiser im Dickicht der Begrifflichkeiten. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *15*(1). https://doi.org/10.17169/fqs-15.1.2043

Siapera, E., & Viejo-Otero, P. (2021). Governing hate: Facebook and digital racism. *Television & New Media*, *22*(2), 112–130. https://doi.org/10.1177/1527476420982232

Springer, N., Brantner, C., Wilhelm, C., Engelmann, I., Stehle, H., Detel, H., & Lobinger, K. (2022, 26-30 Mai). *The online communication disinhibition model: Toward a holistic understanding of (benign and toxic) online communication*. Hybrid 72nd Annual ICA Conference “One World, One Network?”, Paris.

Springer, N., & Naab, T. K. (2022). Hass in Kommentaren: Blockieren oder Einmischen? In G. Weitzel & S. Mündges (Eds.), Hate Speech (pp. 199–216). Springer. https://doi.org/10.1007/978-3-658-35658-3_10

Stahel, L. (2020). *Status quo und Massnahmen zu rassistischer Hassrede im Internet: Übersicht und Empfehlungen*. Eidgenössisches Departement des Innern.

Stahel, L., Weingartner, S., Lobinger, K., & Baier, D. (2022). Digitale Hassrede in der Schweiz: Ausmass und sozialstrukturelle Einflussfaktoren (p. 50). Universität Zürich. https://doi.org/10.21256/zhaw-26867

Statista. (2022). *Meistgenutzte Soziale Medien in der Schweiz nach monatlicher Nutzungsrate im Jahr 2021*. https://de.statista.com/statistik/daten/studie/467634/umfrage/bekanntheit-und-nutzung-von-ausgewaehlten-social-media-plattformen/

Stockmann, D. (2022). Tech companies and the public interest: The role of the state in governing social media platforms. *Information, Communication & Society*, 1–15. https://doi.org/10.1080/1369118X.2022.2032796

Suzor, N., Van Geelen, T., & Myers West, S. (2018). Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette*, *80*(4), 385–400. https://doi.org/10.1177/1748048518757142

van Dijck, J. (2020). Governing digital societies: Private platforms, public values. *Computer Law & Security Review*, *36*, 1–4. https://doi.org/10.1016/j.clsr.2019.105377

WEMF. (2021). *WEMF Auflagenbulletin 2021*. https://wemf.ch/media/wemf_auflagebulletin.pdf

Wilhelm, C., Joeckel, S., & Ziegler, I. (2020). Reporting hate comments: Investigating the effects of deviance characteristics, neutralization strategies, and users’ moral orientation. *Communication Research*, *47*(6), 921–944. https://doi.org/10.1177/0093650219855330

Woolgar, S., & Neyland, D. (2013). *Mundane governance: Ontology and accountability*. Oxford University Press.