Online Hate Speech: User Perception and Experience Between Law and Ethics

Gregor Fischer-Lessiak[1], Susanne Sackl-Sharif[2] and Clara Millner[3]
[1]University of Graz, Graz, Austria
[2]University of Music and Performing Arts, Graz, Austria
[3]Antidiscrimination Office Styria, Graz, Austria
gregor.fischer@uni-graz.at
susanne.sackl-sharif@kug.ac.at
millner@antidiskriminierungsstelle.steiermark.at

Abstract: 'Governance' of online hate speech (OHS) has become a buzzword in social media research and practice. Inputs from a plethora of stakeholders, international organisations, platforms, governments, and NGOs are discussed by academics, (social) media executives and lawmakers around the globe. In these discussions, the opinions of users remain underexplored, and data on their experiences and perceptions is scarce. The present paper focuses on five case studies of model OHS postings in the context of the Austrian OHS governance system. For these case studies, 157 respondents assessed in an online survey whether a posting should be deleted according to their own ethical standards, whether they believed that this posting was currently punishable under Austrian criminal law, and whether it should be punishable. Furthermore, respondents indicated how they deal with OHS in their daily lives when confronted with it on digital platforms. Using social sciences, human rights, and criminal law approaches, we found that OHS-awareness among our respondent group was high and that there is a preference for state regulation, i.e., punishability under national criminal law, and for the deletion of OHS postings. Simultaneously, readiness for counter-speech and reporting of postings for deletion remains relatively low. Thus, OHS postings are hardly ever answered directly or forwarded to specialised organisations and/or the police. If OHS postings are reported, it is mostly done via the channels of the respective platform.

Keywords: freedom of expression, incitement to hatred, criminal law, user perception, online survey

1. Introduction

Online hate speech (OHS) is a virulent social problem that has been challenging democratic discourses in the past years. The storming of the US Capitol in January 2021 is the most prominent high-level case of online disinformation and OHS leading to real-life consequences. As well, by the end of 2021, anti-vaccination propaganda and hate speech against medical staff have gained momentum (Gleicher et al, 2021).

Due to the great relevance of the topic, it is unsurprising that OHS research has been intensifying in recent years. Many studies deal with the question of how OHS can be evaluated from a communication studies or media studies perspective and/or discuss technical conditions of social media and societal changes related to OHS (e.g. Pörksen, 2018; Sponholz, 2018; Zannettou et al, 2020). Other studies address the content of OHS, its impact on those affected as well as coping strategies (e.g. Anderson et al, 2014; Brodnig, 2016; Lumsden and Morgan, 2017). When it comes to the regulation of OHS, which is the focus of this paper, a lot has been written lately about platforms' and states' duties to regulate and counter OHS effectively (e.g. Brown, 2020; Davidson et al, 2017; Waseem and Hovy, 2016). But a central aspect, namely users' opinions on OHS and its regulation, remains under-researched. Especially in the German-speaking area, there are only a few studies on this topic, e.g. Geschke et al (2019) that explore how users perceive state- and platform-made norms on OHS and their implementation.

To contribute to this debate using empirical data, we conducted an online survey (November/December 2020, 157 respondents, Austria) to gather the opinions and experiences of users with OHS based on five OHS model postings. In this paper, we interpret and contextualize their responses based on legal (criminal and human rights law) and sociological approaches. First, we provide the theoretical backdrop. Second, we present our empirical research design. Third, we analyse how respondents assessed our model postings legally and ethically. This includes respondents' willingness to take action against these postings. Fourth, we present cross-case analyses and conclusions.

2. Considerations on 'hate', the law, and the user

2.1 Human rights and online hate

In human rights law, the discussion on the management/moderation of OHS is generally framed around freedom of expression and its boundaries. While there is no authoritative definition of 'hate speech', the term is widely used as an umbrella term. Lately, the United Nations Strategy and Plan of Action on Hate Speech (United Nations, 2020: 8) offers the following definition of 'hate speech': "Any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor. This is often rooted in, and generates, intolerance and hatred, and in certain contexts can be demeaning and divisive."

This Action Plan builds on a plethora of international and regional human rights documents that stipulate state duties to take action against hate speech. For example, Art. 4 (a) ICERD explicitly imposes state duty to prohibit expression that promotes racial hatred, hereby setting boundaries to the freedom of expression to protect the rights of targets/target groups of hate speech. These state obligations apply substantively online and offline. The fulfilment thereof, however, can be procedurally challenging in internationalized online environments (Brown, 2020). Given the lack of immediate state governance, users are confronted mainly with media providers' business ethics, their terms of service and the latter's implementation. To illustrate this fact, we provide the example of Austrian hate speech and platform governance.

2.2 Austria: criminal law fulfils positive obligations

In human rights law, states enjoy a certain leeway, a so-called margin of appreciation, in deciding how to combat hate speech, allowing them to take into account their historical backgrounds as well as legal traditions. For example, denial of the Shoah is outlawed in many EU member states by a variety of norms. In spite of the EU's harmonization efforts (Council Framework Decision 2008/913/JHA, 2008), these norms still differ from state to state – and beyond EU borders, there is even less uniformity. Austria has decided within its margin of appreciation that there are public interests – especially the prevention of a reinvigorated National Socialist movement – justifying the exclusion of these expressions from legal protection, and their prosecution. Postings and comments on social media that fulfil the criteria of these provisions hence lead to law enforcement action just like expressions in the 'analogue' world.

While some argue that social media platforms could be treated as accessories to hate speech crimes in cases of non-deletion of OHS (e.g. Austrian Ministry of Justice, 2016), social media platforms have not (yet) been held criminally liable this way. Contrarily, platforms have profited from OHS, legal and illegal forms alike, as it produces high interaction numbers raising platforms' ad revenue. Social media companies' algorithms have even learned to accelerate the spread of OHS (Zannettou et al, 2020). At the same time, ethical considerations built into terms of service have been meandering between recognition for the problematics of OHS and overstressing US-American freedom of speech doctrine (Kang and Isaac, 2019).

Human rights considerations are more and more included in community standards (e.g. Facebook, 2021a). A strong, strictly legal duty of platforms to moderate according to international human rights law, however, has not emerged yet. While community standards ban hate speech, they lack efficacy until now.

2.3 Laws, platform rules, ethics and the user

Given these shortcomings, states have tried to enhance platform accountability by other means. Since 2021, for example, the Austrian Communication Platforms Act (CPA, BGBl. I Nr. 151/2020), links the moderation practices of large private social media networks to national hate speech provisions. If platforms fail to provide effective reporting tools and transparency reports on their moderation, they might face harsh financial penalties. Austrian national law may hence be able to exert immediate effects on the practices of social media platforms (e.g. Facebook, 2021b). As well, the law may provide for new avenues of cooperation between law enforcement and social media. In the past, investigations against original posters of OHS have been hampered by their non-cooperation and hesitance to provide national authorities with user data needed to trace OHS suspects under national criminal law (Haider and Millner, 2021).

Only lately, platforms have moved towards more comprehensive implementation of due process in their OHS rules by the establishment of court-like entities, e.g., Facebook's/Instagram's/Meta's Oversight Board (2021).

The corporation herein sought to supplement business ethics with legal considerations based on global freedom of expression standards. This work is ongoing. At the same time, and seemingly contrarily, Facebook is challenging the Austrian Communication Platforms Act. The company's motion is currently pending review by the Austrian Administrative Court of Appeal. In this lawsuit, the law's compliance with the EU E-Commerce-Directive (2000/31/EC, 2000) will have to be assessed. Whatever the outcome of the case may be, platforms are not going to be able to hold off extended moderation duties forever. US lawmakers are considering reforms of platform responsibility (Cole et al, 2021), and the EU is preparing new legislation momentarily that is expected to harmonize platform duties in Europe (Digital Services Act, 2020). Amending policies and rethinking business ethics now would help platforms anticipate these new legal developments. More immediate reasons to do so exist: Users affected by OHS report, inter alia, disengagement from societal debates, psychological trauma and even physical effects (e.g. GREVIO, 2021: §12).

In the following, we present our empirical research of user experiences with OHS on social media platforms. We include their views on the applicable laws, their ethical considerations on freedom of expression vis-a-vis OHS and their opinions on reporting mechanisms offered by platforms.

3. Research design

In November and December 2020, we conducted an online survey to gather the public's perceptions of and experiences with OHS. The survey focused on five main topics: definition of OHS, platforms and contents, affected persons and coping strategies, OHS perpetrators, and counter-speech strategies. In this paper, we focus on the evaluation of five model cases that resemble real OHS postings sourced from the data of an Austrian OHS reporting app offered by the Antidiscrimination Office Styria. In the selection of cases, we tried to integrate different OHS target groups, contents, and degrees of intensity. In this context, we were guided by previous empirical results on OHS (e.g. Geschke et al, 2019) and by legal considerations. Respondents were asked to assess the postings with regards to their (perceived) illegality and their (ethical) worthiness of remaining online. We provided 11 items on a 5-point rating scale as well as the opportunity to enter further information in an open response field. The original questionnaire was published in German, the official language in our survey area. All questionnaire responses described below were translated by the authors.

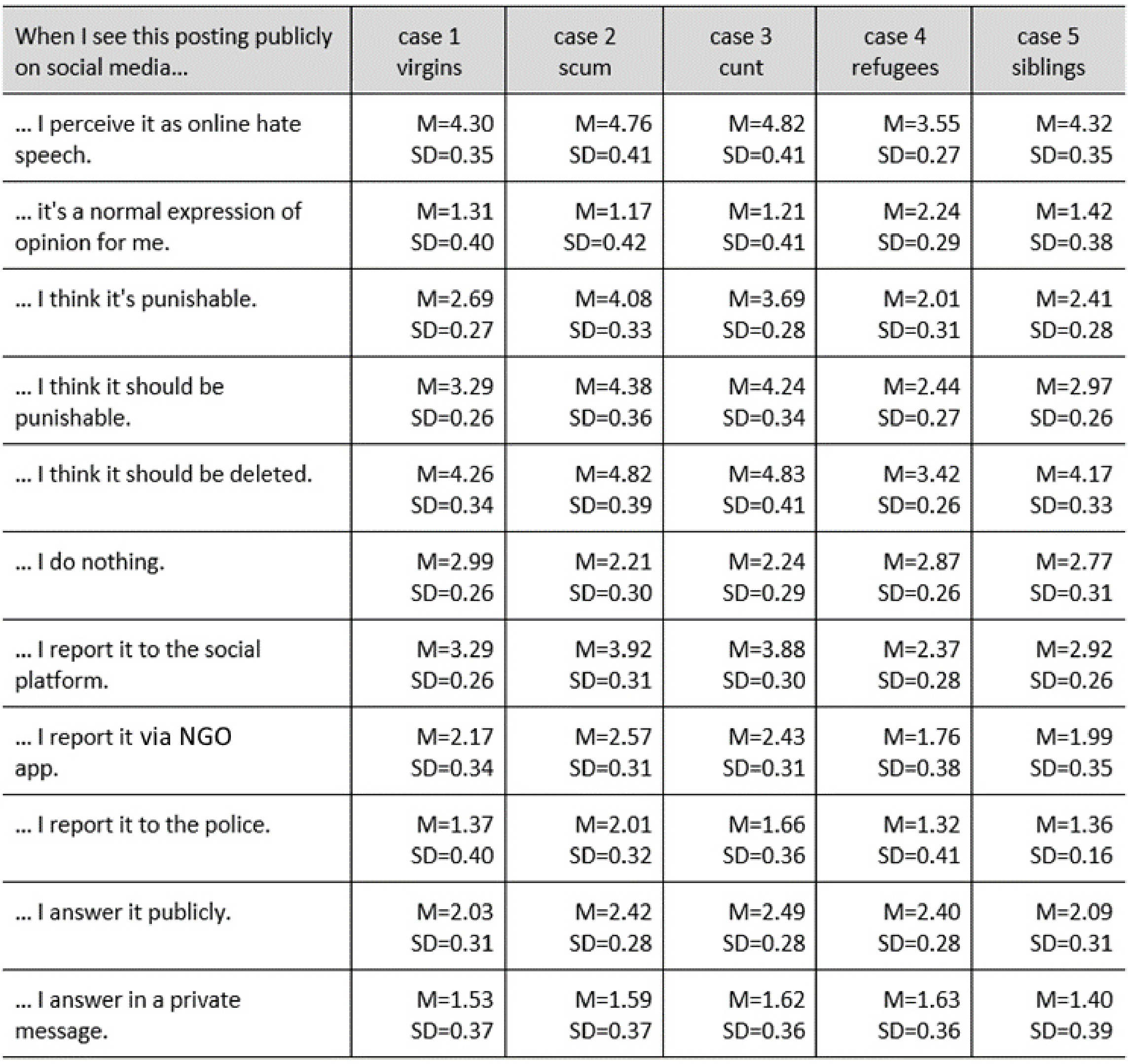
Our purposive sampling strategy focused on two groups of people: 1) young people, as they are particularly frequent users of social media and more likely to be exposed to OHS (Gadringer et al, 2021) and 2) members/employees of organisations dealing with OHS in their work practice. The sample hence mainly includes persons frequently confronted with social media and/or OHS in possession of expert knowledge. Our sample comprises 157 persons. About half of them are between 20 and 29 years old (46%), 5% are younger than 20, 37% are between 30 and 49, and 12% older than 50. Almost 30% are students, almost 20% work in the educational and social sector, 13% as (office) employees and 6% in STEM professions outside the education sector. More than half of the respondents identified as female (59%), 37% as male, and 4% as other genders.

As the complexity of OHS requires an open-explorative and interdisciplinary research approach, the analysis of the five model cases includes social science and legal evaluations. Our social sciences analysis included usual statistical descriptive rations such as arithmetic mean or standard deviation, but also the clustering of open-ended questions according to frequency. Our legal analysis is based largely on Austrian criminal law doctrine, accompanied by human rights law considerations. Each case is analysed individually before essential similarities and differences are worked out in a cross-case analysis.

4. Case studies

In this chapter, we describe and discuss the main characteristics of our five model cases. We assess these cases based on Austrian criminal law and compare this analysis with the responses to our questionnaire. Table 1 gives an overview of the ratings of the 11 items per case (5=complete agreement; 1=complete disagreement). The items can be clustered into three content groups: a) online hate speech vs. expression of opinion (items 1-2), b) evaluation of posting according to punishability (items 3-4), c) responses/strategies (items 5-11).

Table 1: Overview of case studies evaluations, complete model postings below



When I see this posting publicly on social media...	case 1 virgins	case 2 scum	case 3 cunt	case 4 refugees	case 5 siblings
... I perceive it as online hate speech.	M=4.30 SD=0.35	M=4.76 SD=0.41	M=4.82 SD=0.41	M=3.55 SD=0.27	M=4.32 SD=0.35
... it's a normal expression of opinion for me.	M=1.31 SD=0.40	M=1.17 SD=0.42	M=1.21 SD=0.41	M=2.24 SD=0.29	M=1.42 SD=0.38
... I think it's punishable.	M=2.69 SD=0.27	M=4.08 SD=0.33	M=3.69 SD=0.28	M=2.01 SD=0.31	M=2.41 SD=0.28
... I think it should be punishable.	M=3.29 SD=0.26	M=4.38 SD=0.36	M=4.24 SD=0.34	M=2.44 SD=0.27	M=2.97 SD=0.26
... I think it should be deleted.	M=4.26 SD=0.34	M=4.82 SD=0.39	M=4.83 SD=0.41	M=3.42 SD=0.26	M=4.17 SD=0.33
... I do nothing.	M=2.99 SD=0.26	M=2.21 SD=0.30	M=2.24 SD=0.29	M=2.87 SD=0.26	M=2.77 SD=0.31
... I report it to the social platform.	M=3.29 SD=0.26	M=3.92 SD=0.31	M=3.88 SD=0.30	M=2.37 SD=0.28	M=2.92 SD=0.26
... I report it via NGO app.	M=2.17 SD=0.34	M=2.57 SD=0.31	M=2.43 SD=0.31	M=1.76 SD=0.38	M=1.99 SD=0.35
... I report it to the police.	M=1.37 SD=0.40	M=2.01 SD=0.32	M=1.66 SD=0.36	M=1.32 SD=0.41	M=1.36 SD=0.16
... I answer it publicly.	M=2.03 SD=0.31	M=2.42 SD=0.28	M=2.49 SD=0.28	M=2.40 SD=0.28	M=2.09 SD=0.31
... I answer in a private message.	M=1.53 SD=0.37	M=1.59 SD=0.37	M=1.62 SD=0.36	M=1.63 SD=0.36	M=1.40 SD=0.39

4.1 '72 virgins'

You see a picture of a (Muslim) man and a flock of sheep with the caption: "72 virgins just for you; ElitePartner Academics & Singles with standards; Syria edition."

4.1.1 The law

This publication fulfils the requirements for punishability as incitement to hatred (§ 283 (1) no. 1, second case Austrian Criminal Code – ACC). The posting at hand must be assessed in its entirety: Content-wise, the published picture must be treated as an integral part of the publication. The picture and its caption are directed against men of Muslim faith, a protected group (criterion: religion) under § 283 ACC (Plöchl 2020: § 283 no. 8). The posting appeals to the recipients' feelings, instigates hatred against persons of Muslim faith and can evoke a strong feeling of antipathy by attributing a tendency towards sodomy to Muslim men. The pseudo-humoristic context of the posting expresses additional contempt against the target group of the posting (Plöchl 2020: § 283 no. 19).

4.1.2 The data

This model case is clearly perceived as hate speech: 83% of the respondents completely or somewhat agree that this posting constitutes OHS and only 3% completely or somewhat agree that this is a normal expression of opinion. Regarding punishability, this model case has an intermediate position in our case sample: 35% of the respondents (rather) agree that this posting is punishable, 55% (rather) agree that this posting should be punishable. This case prompted high ratings related to rather passive responses/strategies. 79% of the respondents (rather) agree that this posting should be deleted, 46% (rather) think that they would do nothing if they saw it online. If respondents were to take action themselves, they would most likely report the posting to the platform.

4.2 'Parasite scum'

Please rate the following posting. Comment on an article about religious holidays/festivities: "This parasite scum vanished 6-million-fold through chimneys without a trace."

4.2.1 The law

This posting does not explicitly name a group attribute ("*This* parasite scum […]"). It is still deducible from the context – religious festivities in the original posting – that it is directed against persons based on their religion. Without this context, a target group could not be determined. The terminology used within the posting against an identifiable, protected group of the population, is per se sufficient to fulfil the requirements of § 283 (1) no. 2 ACC. The statement constitutes verbal abuse infringing upon human dignity. This interpretation can be based on a general linguistic understanding. The term 'scum' is an intentional, disparaging designation of a part of an entirety (here: society as a whole) which is considered as inferior. The term 'parasite' signifies a lifeform which lives at the expense of another and was already used in National-Socialist propaganda in a stigmatising fashion against Jews* and other minorities (Musolff, 2011). By considering the posting as a whole, the semantic content of the incriminated expression becomes specifically obvious as a ridiculing trivialisation of the murder of Jews*. This constitutes not only a pejorative degradation, but also a statement that is able to induce feelings of hatred against persons based on their religion within the scope of § 283 (1) no. 1 second case ACC (see OGH 23.5.2018, 15 Os 33/18v = EvBl 2018/143).

4.2.2 The data

This case has a very high OHS rating: 95% of the respondents (rather) agree that this posting is OHS and only 3% (rather) agree that this is a normal expression of opinion. This case also has the highest punishment ratings: 78% of the respondents (rather) think this posting is already punishable, 85% think the posting should (rather) be punishable. This case has the highest ratings related to active responses/strategies. Many respondents would (rather) report the posting to the platform (73%), to the reporting app (34%) and to the police (18%). Furthermore, 33% of the respondents would (rather) answer to this posting publicly, 94% think it should be deleted. Only 23% of the respondents would (rather) do nothing.

4.3 'Dirty cunt'

Comment to a female user: "Dirty cunt, you sleazy wench. Yikes, you hate-consumed cunt."

4.3.1 The law

This posting fulfils the requirements of criminal defamation and insult under §§ 111, 115 ACC. As required by § 111 ACC, a user is accused of possessing a despicable characteristic/disposition as being "hateful" in a manner perceivable for third parties. This accusation is able to reduce the target's reputation and esteem she enjoys among her fellow human beings (Tipold 2016: § 111 no. 4). The term "sleazy wench" may not be punishable as an accusation of dishonourable conduct or of conduct against common decency, however, it signifies contempt against the targeted person. In the context of another abusive word used in the posting, namely "dirty cunt", "sleazy wench" further serves to humiliate a female person (Rami 2021: § 115 no. 8). The posting must be assessed in its entirety. By the cited misogynistic expressions, a female user is assailed based on her gender. The underlying goal of the posting is to articulate disdain against this user as required by § 115 ACC. The gender-based insults used in the posting are hence punishable under §§ 111, 115 ACC, however, they do not reach the threshold of incitement to hatred (against women) under § 283 ACC (Tipold 2016: § 115 no. 3).

Lately, international organisations, such as the Council of Europe and its independent expert body GREVIO, have been paying closer attention to gender-based violence and OHS. In its first general recommendation on the Istanbul Convention, GREVIO noted (2021: 19): "Sexist behaviour such as sexist hate speech, which often constitutes a first step in the process towards physical violence, may also escalate to or incite overtly offensive and threatening acts, including sexual abuse or violence or rape, thus falling within the remit of Article 40 of the Istanbul Convention." Article 40 stipulates state duties to combat sexual harassment – accordingly, further regulatory action in this area may be warranted, especially given the rising numbers of women* being targeted by OHS. Without such reform, it would be desirable for courts' interpretations of § 283 in cases of hate speech against women* to take this fact into account.

4.3.2 The data

This case has a very high OHS rating: 96% of the respondents (rather) agree that this posting is OHS and only 4% of the respondents (rather) agree that this is a normal expression of opinion. This case also has the second highest punishability ratings: 67% of the respondents (rather) think this posting is already punishable, 82% of the respondents stated the posting should (rather) be punishable. This case has high ratings related to active responses/strategies. Many respondents would (rather) report the posting to the platform (72%) or to the reporting app (31%), 35% of the respondents would (rather) answer to this posting publicly. Besides, 97% of the respondents think the posting should be deleted and only 25% would (rather) do nothing.

4.4 'Male refugees'

Comment on an article dealing with questions about refugee movements: "But one may still say that the many young men are more violent than families!?"

4.4.1 The law

The cited posting addresses the societal group of refugees and other displaced persons and reproduces the prejudice that the flight of predominantly male persons had led to rising numbers in “foreign crime”. The posting aims to instil fear and to reinforce antipathy against refugees. By claiming to ask a problematic question, the poster tries to trivialize the content and insinuates that this statement is at the margins of freedom of expression. If the contextualisation of criminality and origin/gender is intended to instil hostile sentiments against protected groups, § 283 (1) no. 1 second case ACC could be fulfilled. In the case at hand, however, the posting is not formulated sufficiently to deduce a tendentious incitement to hatred and contempt. Aversion, rejection, or contempt are not uttered to the extent legally required by § 283 ACC (Plöchl 2020: § 283 no. 19).

4.4.2 The data

This case has by far the lowest OHS ratings: 63 % of the respondents (rather) perceived this posting as OHS, 23% as a normal expression of opinion. This case also has the lowest punishment ratings: 13% of the respondents (rather) agree that this posting is punishable, 25% of the respondents (rather) agree that this posting should be punishable. This case has rather high ratings related to passive responses/strategies. 38% of the respondents would (rather) do nothing if they see the posting online, only 27% would report it to the platform. Interestingly, however, this posting has the highest value in terms of the respondents' response behaviour: 36% of the respondents would (rather) answer publicly to that posting.

4.5 'Siblings'

Comment on people with a migration history: "With most people from your area, the parents are also siblings!"

4.5.1 The law

The poster obviously intends to incentivise negative attitudes towards persons based on their origins and probably also their religion. Recipients could deduce from this statement that incest is common among persons with migration histories. A context between origin, and possibly also religion, and consanguinity is subtly alleged. The narrative that marriage between relatives is common in families with migration histories is commonly used by right-wing groups (Hödl, 2010; Deutscher Bundestag, 2018). These groups do not contribute to discussions on the historical roots of marriage within families or criticise such practices, but exclusively use this narrative for purposes of propaganda (Müller, 2012). This interpretation may show the underlying motivation of the poster, however, the statement per se is not sufficiently formulated to warrant for punishability under § 283 ACC (incitement to hatred) or any other provision of Austrian criminal law.

4.5.2 The data

This model case is clearly perceived as OHS: 85% of the respondents stated that they completely or somewhat agreed that this is OHS and 7% of the respondents completely or somewhat agreed that this is a normal expression of opinion. This case has the second lowest punishment ratings: 25% of the respondents (rather) agree that this posting is punishable, 41% of the respondents (rather) agree that this posting should be punishable. This case has rather high ratings related to passive responses/strategies. 36% of the respondents would (rather) do nothing if they see the posting online, 46% would report it to the platform and 20% would report it to the reporting app.

4.6 Cross-case analysis

Our cross-case comparison shows that, except for case 4 ‘male refugees’, all cases were predominantly perceived as OHS and not as ‘normal’ expression. It is noticeable that OHS against women* or anti-Semitic content were more likely to be perceived as OHS than attacks against people with a refugee, Muslim or Syrian background. This also affects the issues of punishability and deletion of postings: The perceived necessity of OHS regulation and removal is higher in case 2 ‘parasite scum’ and case 3 ‘dirty cunt’ than in all other cases of the sample. Case 4 ‘male refugees’ and case 5 ‘siblings‘ are, in contrast, not punishable under Austrian criminal law but respondents expressed a strong (ethical) desire for these content pieces to be removed. Interestingly, the punishable case 1 ’virgins‘ only displays marginally higher ratings in perceived and preferred punishability than the (unpunishable) cases 4 and 5. From these considerations, it seems possible to derive initial indications that topics that are closer to oneself are more likely to be perceived as OHS than topics that are more distant from one’s own reality of life. Most probably due to Austrian history, there is more awareness of and sensibility towards anti-Semitism than towards discrimination against Muslims to give just one example. Discrimination based on ethnicity or religion, in contrast to gender and political opinion, was hardly ever mentioned as a basis of respondents’ own experiences with online discrimination. This hints towards target-group dependence of OHS perception among users. It would be beneficial to conduct further research into the dependency of receptions of target groups in societies at large vis-à-vis users’ readiness to label content as illegal hate speech.

Furthermore, it is interesting to see who should regulate OHS from the user’s point of view. Respondents seem to perceive that Austrian criminal law lacks rigour, as punishability ratings under the – perceived – current criminal law regime are overall lower than those of *preferred* punishability. This can be interpreted as a desire for (more) state intervention in the regulation of OHS. Geschke et al (2019) have shown a similar desire for state intervention in their representative study on OHS in Germany: 75% of respondents (N=7,349) agreed with the demand that the state should consistently enforce existing laws against insults, hate speech and defamation on the internet. Community-driven responses to OHS, i.e., counter-speech and reporting, remain at low levels in all cases. In the few instances in which a posting is reported, it is more likely to be reported to the platform itself and not to the local hate speech reporting app offered by the Antidiscrimination Office Styria or the police. Thus, in addition to the state, platforms themselves are perceived as responsible for the regulation of OHS or helpful in the fight against OHS. In general, it can also be stated that more active strategies are applied when postings are strongly perceived as OHS, as shown by the analysis of case 2 ‘parasite scum’ and case 3 ‘dirty cunt’. We can hence conclude for our sample that counter-speech happens – if at all – rather after, instead of before the escalation of OHS into illegality. This is also shown by the recommendations for deleting postings: Except for case 4 ‘male refugees’, respondents show an overall tendency towards endorsing deletion of postings regardless of how they are perceived under criminal law. Thus, it is considered better to make content invisible than to become active and influence the discourse itself. This shows the idea of counter-speech as a mechanism against OHS lacks support among our respondents. The self-regulation of OHS among users, as in the *marketplace of ideas*, is not likely under such circumstances as also Knauder and Romanin (2021) indicated.

5. Conclusions, recommendations and outlook

Legally and ethically, we conclude that OHS will not be held at bay by states, platforms, or users alone, but by an interplay of these actors. Their different scopes of action can mutually reinforce each other. States can refine national norms and their application, platforms can reform their standards and reinforce moderation practices, and users can contribute with counter-speech, moderation, and, where applicable, by reporting to platforms, NGOs, and the police. We have shown that some of these responses are considered to be more effective than others by users.

Given the complexity of human expression – e.g., a multitude of languages, humour, sarcasm, and irony – especially the detection of emerging and grey area forms of OHS will need flexible, participatory models of governance. Societal climates can support or hinder the detection of OHS and illegality even among expert respondent groups, as we have shown above. Large-scale, integrated, and multi-/interdisciplinary studies on the interrelation of societal discourse and surges in OHS, as well as on the legal and ethical views of internet users, are hence needed. Conducting such studies will require transparency of platforms regarding their moderation practices. At the same time, by engaging professionals in the field as well as the public at large, media competence could be raised – and the perceived need for stricter hate speech laws, as well as expenditure in law enforcement lowered in the future. Human-rights-based, high-quality and human moderation by platforms will be key therein, as well as the participation of targets of OHS/advocacy groups to detect emerging surges in

OHS. Ethics, human rights law, expert knowledge from NGOs as well as state officials and, finally, user participation by means of low-threshold tools could help, in their interplay, to realise less hate-driven social media for all – an endeavour that has become even more important during COVID-19.

References

Literature

Anderson, A. et al (2014) "The 'Nasty Effect': Online Incivility and Risk Perceptions of Emerging Technologies", *Journal of Computer-Mediated Communication*, 19:3, pp. 373–387.

Brodnig, I. (2016) *Hass im Netz. Was wir gegen Hetze, Mobbing und Lügen tun können*, Brandstätter, Wien.

Brown, A. (2020) "Models of Governance of Online Hate Speech", [online], Council of Europe, https://rm.coe.int/models-of-governance-of-online-hate-speech/16809e671d (Retrieved: November 26, 2021).

Cole, C.J. et al (2021) "INSIGHT: Social Media Reform. Can the U.S. Learn From France?", [online], Bloomberglaw, https://news.bloomberglaw.com/us-law-week/insight-social-media-reform-can-the-u-s-learn-from-france (Retrieved: November 26, 2021).

Davidson, T. et al (2017) "Automated Hate Speech Detection and the Problem of Offensive Language", *Proceedings of the International AAAI Conference on Web and Social Media 11*, pp. 512–515.

Facebook (2021a) Community Standards, [online], Facebook, https://transparency.fb.com/policies/community-standards/hate-speech/ (Retrieved: November 26, 2021).

Facebook (2021b) KoPl-G Transparenzbericht, [online], Facebook, https://about.fb.com/de/wp-content/uploads/sites/10/2021/10/Facebook-KoPl-G-Transparenzbericht-Oktober-2021.pdf (Retrieved: November 26, 2021).

Gadringer, S. et al (2021) *Digital News Report 2021. Detailergebnisse für Österreich*, University of Salzburg, Salzburg.

Geschke et al (2019) *#HASS IM NETZ. Der schleichende Angriff auf unsere Demokratie*, IDZ, Jena.

Gleicher et al (2021) "Adverserial Threat Report", [online], Meta, https://about.fb.com/wp-content/uploads/2021/12/Metas-Adversarial-Threat-Report.pdf (Retrieved: December 2, 2021).

Haider, I. and Millner, C. (2021) "Hasspostings im Strafverfolgungssystem", *Online Hate Speech. Perspektiven aus Praxis, Rechts- und Medienwissenschaften* (ed. G. Fischer / C. Millner / S. Radkohl), NWV, Vienna, pp. 91–139.

Hödl, K. (2010) "Sarrazin und der Zeitgeist: Sarrazin argumentiert zweifellos rassistisch", [online], DerStandard, https://www.derstandard.at/story/1282978632580/sarrazin-und-der-zeitgeist (Retrieved: December 2, 2021).

Kang, C. and Isaac, M. (2019), "Defiant Zuckerberg Says Facebook Won't Police Political Speech", [online], New York Times, https://www.nytimes.com/2019/10/17/business/zuckerberg-facebook-free-speech.html (Retrieved: November 26, 2021).

Knauder, B. and Romanin, A. (2021) "Politikwissenschaftliche Perspektive aud die Regulierung von Online Hate Speech und deren Einflüsse auf die demokratische Gesellschaft", *Online Hate Speech. Perspektiven aus Praxis, Rechts- und Medienwissenschaften* (ed. G. Fischer / C. Millner / S. Radkohl), NWV, Vienna, pp. 57–66.

Lumsden, K. and Morgan, H. (2017) "Media framing of trolling and online abuse: Silencing strategies, symbolic violence, and victim blaming", *Feminist Media Studies*, 17:6, pp. 926–940.

Müller, T. (2012) "Damit alles in der Familie bleibt", [online], Wiener Zeitung, https://www.wienerzeitung.at/nachrichten/politik/oesterreich/430859-Damit-alles-in-der-Familie-bleibt.html?em_cnt_page=1 (Retrieved: December 2, 2021).

Musolff, A. (2011) "Metaphorische *Parasiten* und 'parasitäre' Metaphern: Semantische Wechselwirkungen zwischen politischem und naturwissenschaftlichem Vokabular", *Metaphern und Gesellschaft: Die Bedeutung der Orientierung durch Metaphern* (ed. M. Junge), Springer, Wiesbaden, pp. 105–119.

Oversight Board (2021) Governance, [online], Oversight Board, https://oversightboard.com/governance/ (Retrieved: November 26, 2021).

Plöchl, F. (2020) "§ 283 StGB", *Wiener Kommentar zum Strafgesetzbuch* (2nd ed., F. Höpfel / E. Ratz), Manz, Vienna, https://rdb.manz.at (Retrieved: December 12, 2021).

Pörksen, B. (2018) *Die große Gereiztheit. Wege aus der kollektiven Erregung*, Hanser, Munich.

Rami, M. (2021) "§§ 111, 115 StGB", *Wiener Kommentar zum Strafgesetzbuch* (2nd ed., F. Höpfel / E. Ratz), Manz, Vienna, https://rdb.manz.at (Retrieved: December 12, 2021).

Sponholz, L. (2020) *Hate Speech in den Massenmedien. Theoretische Grundlagen und empirische Umsetzung*, Springer VS, Berlin.

Tipold, A. (2016) "§§ 111, 115 StGB", *Leukauf/Steininger StGB* (4th ed., C. Aichinger et al), Linde, Vienna, https://rdb.manz.at, (Retrieved: December 12, 2021).

Waseem Z. and Hovy, D. (2016) "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter", *Proceedings of the NAACL Student Research Workshop*, pp. 88–93.

Zannettou, S. et al (2020) "Measuring and Characterizing Hate Speech on News Websites", *Proceedings of the 12th ACM Conference on Web Science*, pp. 125–134.

Legal Sources and Recommendations

Austrian Ministry of Justice, Decree on the Agreement with Facebook on Deleting Hate Speech, German: E*rlass vom über die Vereinbarung mit Facebook zur Löschung von Hasspostings und Informationserteilung*, 20 July 2016, BMJ-S884.024/0014-IV/2016.

Austrian National Council, Communication Platforms Act (CPA), German: *Bundesgesetz über Maßnahmen zum Schutz der Nutzer auf Kommunikationsplattformen, Kommunikationsplattformen-Gesetz, KoPl-G*, BGBl. I Nr. 151/2020.

Austrian Supreme Court of Justice (Oberster Gerichtshof/OGH), 23.5.2018, Case no 15 Os 33/18v = EvBl 2018/143.

Council Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law, 2008/913/JHA, 28 November 2008, OJ L 328.

Deutscher Bundestag, Drucksache 19/1444 – „Schwerbehinderte in Deutschland" – Kleine Anfrage der Abgeordneten Höchst/Gminder/Pohl/Hartmann und der Fraktion der AfD, Berlin (2018).

Directive of the European Parliament and of the Council on certain legal aspects of information society services, in particular electronic commerce, Internal Market (Directive on electronic *commerce)*, 2000/31/EC, 8 June 2000.

GREVIO (2021), *General Recommendation No. 1 on the digital dimension of violence against women*, Council of Europe, Strasbourg, https://rm.coe.int/grevio-rec-no-on-digital-violence-against-women/1680a49147 (Retrieved: November 26, 2021).

International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) 1965, (resolution 2106 (XX)), opened for signature 21 December 1965, entered into force 4 January 1969.

Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, 2020, COM/2020/825 final.

United Nations (2020) "United Nations Strategy and Plan of Action on Hate Speech", [online], United Nations, https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf (Retrieved: November 26, 2021).