Swiss Journal of Sociology, 50 (1), 2024, 105–124

Can Legal Sanctions Reduce Cyberviolence? How Changes in Cost-Benefit Calculations and Norm Neutralizations Affect Self-Censorship

Lea Stahel* and Sebastian Weingartner**

Abstract: This study examines whether and how legal sanctions help reduce cyberviolence. Interviews were conducted with offenders who were legally sanctioned for posting criminal online comments in Switzerland. The results of the thematic analysis indicate that offenders self-censor after facing legal sanctions. This is explained with reference to rational choice theory and neutralization theory. The study contributes to the hitherto lacking knowledge about the effectiveness of legal countermeasures against cyberviolence.

Keywords: Cyberviolence, social media, countermeasures, techniques of neutralization, rational choice theory

Können rechtliche Sanktionen Cybergewalt eindämmen? Wie sich Veränderungen von Kosten-Nutzen-Kalkulationen und Normneutralisierungen auf die Selbstzensur auswirken

Zusammenfassung: Die Studie untersucht, ob und wie rechtliche Sanktionen Cybergewalt reduzieren. Es wurden Interviews mit Straftäter:innen geführt, die in der Schweiz für das Hochladen von Online-Kommentaren rechtlich belangt wurden. Die Ergebnisse der thematischen Analyse zeigen, dass sich Täter:innen nach der Erfahrung rechtlicher Sanktionen zensieren. Dies wird mit Bezug auf die Rational-Choice-Theorie und die Neutralisierungstheorie erklärt. Die Studie trägt zum fehlenden Wissen über die Wirksamkeit rechtlicher Gegenmassnahmen gegen Cybergewalt bei.

Schlüsselwörter: Cybergewalt, Soziale Medien, Gegenmaßnahmen, Neutralisierungstechniken, Rational-Choice-Theorie

Les sanctions juridiques peuvent-elles réduire la cyberviolence ? Comment les changements dans les calculs coûts-avantages et les neutralisations de normes affectent l'autocensure

Résumé : L'étude examine si et comment les sanctions légales réduisent la cyberviolence. Des entretiens ont été menés avec des délinquants qui ont été punis pour avoir publié des commentaires en ligne en Suisse. Les résultats de l'analyse thématique indiquent que les délinquants s'autocensurent après avoir fait face à des sanctions légales. Nous l'expliquons à l'aide de la théorie du choix rationnel et à la théorie de la neutralisation. L'étude contribue au manque de connaissances sur l'efficacité des contre-mesures juridiques contre la cyber-violence.

Mots-clés : Cyberviolence, médias sociaux, contre-mesures, techniques de neutralisation, théorie du choix rationnel

* Department of Sociology, University of Zurich, Andreasstrasse 15, CH-8050 Zurich, stahel@soziologie.uzh.ch.

** Statistical Office of the Canton of Zurich, CH-8090 Zurich, sebastian.weingartner@statistik.ji.zh.ch.

DOI 10.26034/cm.sjs.2024.6032

1 Relevance and Goals of Study[1]

Activists are defamed on YouTube, women are harassed on Telegram, and religious minorities are insulted in tweets (e. g. Semenzin and Bainotti 2020; Park et al. 2021). On social media, *cyberviolence,* which is the harm and abuse inflicted through digital and technological means (Backe et al. 2018, 135), is inflationary. The internet allows anyone to publish and distribute corresponding online comments, videos, memes, and other formats instantly, regardless of physical proximity to the victim, to a potentially large audience, and usually with impunity. Such content is replicable and can escalate quickly (Sallavaci 2018, 17). This can result in devastating emotional, social, and economic consequences for victims. Moreover, the dissemination of violent content harms society as a whole by fostering a climate of prejudice and polarization.

Consequently, increasing attention is being paid to how to counter cyberviolence. Although various countermeasures have been called for and implemented, very little is known about how effective they actually are (Banks 2010; Blaya 2018). This applies to legal sanctions in particular (e. g., Bakalis 2018). Legal sanctions typically include being reported to the police, being involved in court proceedings, and being sentenced to fines and out-of-court financial settlements. In many countries, various acts of cyberviolence are criminal. However, enforcing laws online is difficult (Banks 2010; Bakalis 2018). And even when offenders are actually held accountable, we do not yet know how legal sanctions affect their perceptions and future online behaviour (e. g. El Asam and Samara 2016). This lack of knowledge about the deterrent effect of legal sanctions is not surprising because the population of convicted offenders is very difficult to access. However, studying this population helps to assess whether and under which circumstances legal sanctions may be a promising way to counter cyberviolence.

Therefore, the present study explores how legal measures affect offenders' perceptions, attitudes, and behaviours related to cyberviolence. We interviewed adult offenders who between 2016 and 2019 had faced legal sanctions in Switzerland for posting criminal comments on social media. We analyse the interview data thematically by drawing on theories of criminal behaviour: rational choice theories (Gibbs 1985; Opp 2020) and theories of neutralization of moral norms (Brewer et al. 2020; Sykes and Matza 1957). We retrospectively track changes in the cost-benefit calculations and neutralization techniques employed by the offenders at the time of the offence and during and after legal sanctions.

The remainder of the paper is organized as follows: First, we introduce the concept of cyberviolence and discuss the meagre state of knowledge on the effectiveness of legal countermeasures. Second, we present rational choice theory and neutralization theory as central reference points for explaining behavioural changes through legal

1 We thank Dr. Simon Milligan of Academic Language Services GmbH for linguistic proofreading of the manuscript.

sanctions. Third, we explain our empirical approach, involving data, recruitment, interviewing, transcription, analytical strategy, and the process of analysis, before fourth, discussing our findings. We conclude by highlighting contributions and discussing limitations.

2 Cyberviolence and the Effectiveness of Legal Countermeasures

Backe et al. (2018, 140) observe an "evident lack of definitional, theoretical, or methodological consensus within the scientific community" in the conceptualization of cyberviolence. Nevertheless, they broadly define cyberviolence as "harm and abuse facilitated by and perpetrated through digital and technological means" (Backe et al. 2018, 135), typically including online harassment, cyberbullying, cyber dating abuse, revenge porn, and cyberstalking. We adopt this definition and refer specifically to acts that violate criminal law. In Switzerland, where this study was conducted, name-calling, defamation, and degradation of both individuals and groups, for example, are all criminal acts.

Against this background, measures against cyberviolence are increasingly demanded and implemented by state actors, social media companies, and civil society organizations (Banks 2010; Blaya 2018; Sallavaci 2018). Although these measures are diverse, very little is known about how effective they actually are in preventing or reducing cyberviolence (Blaya 2018). This unsatisfactory state of affairs also applies to legal sanctions (Banks 2010; El Asam and Samara 2016; Bakalis 2018). A variety of legislative and enforcement issues have been highlighted, including insufficient consideration of the nature of digital harm in current legislation, the tension between country-specific legislation and jurisdictions and transnational social media companies in dealing with freedom of expression, the anonymity of online offenders, the infrequency of victim reporting, and the often-inadequate training of law enforcement personnel.

Moreover, even when offenders are actually held accountable, we do not yet know how legal measures affect offenders' behaviour. Several authors have demanded more knowledge about the preventive and deterrent effects of legal sanctions against cyberviolence and how offenders respond to legal sanctions (e. g. El Asam and Samara 2016, 127). This gap presents both theoretical and empirical challenges. At the theoretical level, we need to understand how the effects of legal sanctions on online behaviour could possibly be explained (Xu et al. 2016, 642; Holt et al. 2019, 1153). Empirically, the difficulty is to collect data on the tiny and hard-to-reach population of reported offenders.

3 Theoretical Approaches to the Effectiveness of Legal Sanctions

3.1 Rational Choice Approach

To examine whether and how legal sanctions can affect cyberviolence, we need to understand how criminal behaviour in general occurs. Perhaps the most influential approach to the explanation of criminal behaviour is derived from theories of rational choice (Opp 2020). These theories view criminal acts as the results of a utility maximization process that calculates the risk of being caught and punished (Becker 1968). A criminal act is executed if its expected benefits exceed its expected costs. Thus, the greater an individual estimates the benefits or costs and the probability of their occurrence to be, the more or less likely it is that the individual will commit a criminal act. In a wide version of rational choice theory (Opp 2020, Chapter 4), the contents of this cost-benefit calculation are hardly restricted and can be material (e. g. money, penalties), social (e. g. gain or loss of social status), and even emotional (e. g. pleasure, shame). Accordingly, formal legal sanctions aim at deterring individuals from committing crimes by increasing the costs of doing so (Gibbs 1985). However, this only holds if individuals actually perceive the punishment as severe *and* consider they are likely to be caught. Because this is often not the case in practice, the deterrent effect of legal measures has been questioned (McGuire 2002; Paternoster 2010, 765).

The digital environment can increase the likelihood of criminal behaviour by affecting the perception of costs and benefits. Most people are not clearly aware of which online actions are actually criminal and how severe the punishments are (e. g., Sallavaci 2018; Stalans and Donner 2018). Furthermore, opportunities for anonymization and low reporting rates online can reduce the subjective probability of getting caught.

3.2 Normative Approach and Techniques of Neutralization

Recent research suggests that cost-benefit calculations are only relevant causes of crime if people do not feel bound by normative rules or moral beliefs (Kroneberg et al. 2010). The normative approach to criminal behaviour argues that whether a crime is committed or not results from a more or less automatic actualization of internalized social and moral norms, triggered by situational stimuli (Wikström 2017). Most people do not even consider criminal behaviour an alternative because they have so completely internalized the norms and rules of the legal system. From this normative perspective, criminal behaviour can occur under two conditions:

First, some people might be at least partially socialized in an alternative social environment, leading to an internalization of some deviant norms. If these norms are activated by situational cues, they behave in a deviant or criminal way without much deliberation. In such settings, legal sanctions are ineffective because offenders do not calculate possible costs. However, this paper does not further address the

option that cyberviolence is caused by an automatic actualization of a proviolence norm. This is because all offenders interviewed can, by and large, be classified as members of mainstream society. Hence, they can be assumed to accept widely shared interpersonal norms and adhere to moral standards that are compatible with the Swiss legal system. This is particularly evident in their self-image as law-abiding citizens. For instance, our interviewees emphasized that they never "had debts" (Michael), that they had "never been unemployed" (Fritz), or that they were "good, tax-paying Swiss citizen[s]" (Laura) or even "sweetheart[s]" (Michael; see the empirical section for information on the sample and data).

Second, even if people are socialized in full compliance with the legal system, the binding effect of nonviolence norms can be neutralized under certain circumstances. That means that social and moral norms are temporarily inactivated, and people deviate from the corresponding prescriptions even though these prescriptions are still internalized (Goldsmith and Brewer 2015; Brewer et al. 2020, 548). The underlying cognitive activities that offenders use to avoid guilt and maintain a positive self-image are summarized as techniques of neutralization (Sykes and Matza 1957; Brewer et al. 2020). When applied prior to a crime, techniques of neutralization can explain criminal behaviour. But when applied retroactively, they can be used to rationalize or justify crimes (Stalans and Donner 2018, 35). In any case, the relationship between techniques of neutralization and legal sanctions is not clear a priori.

Sykes and Matza (1957) propose five key neutralization techniques. When offenders *deny responsibility* for their delinquent acts, they portray the crimes as accidental, not their choice, or beyond their control. By *denying the injury,* offenders deny inflicting direct harm, which makes the deviance seem more acceptable. *Denying the victim* acknowledges the harm but views it as justified retribution that the victim deserves. *Condemning the condemners* diverts attention from the offender's criminal acts to those who disapprove of their crimes, including authorities and opposing "others". For example, condemners' motives and actions are delegitimized as unjust, overly restrictive, and ineffective. Finally, by *appealing to higher loyalties,* offenders invoke higher goals and norms that serve their own group and are prioritized over societal demands.

Norm neutralizations are particularly likely in the digital sphere because it favours situational framings that override the usual norms. Social-technological contexts multiply opportunities for inactivating norms and justifying crimes, enabling "digital drifts" in which individuals can easily both engage in and disengage from crime (Goldsmith and Brewer 2015). For example, neutralizations in cyberbullying, flaming, and cyber-racism (e. g. Vysotsky and McCarthy 2017) illustrate the ease with which victims and injury can be denied because of the technology-induced distance that prevents empathy in the perpetrator and makes the harm invisible.

The following empirical analysis explores the mechanisms through which legal measures can help reduce crime-enhancing cost-benefit calculations and neutralizations.

4 Empirical Method

4.1 Data and Recruitment

Identifying cyberviolence offenders and sampling them for scientific research is very challenging because only a tiny fraction of offenders is actually reported, and even when they are reported, privacy regulations make it difficult to contact them. Therefore, we cooperated with a Swiss association that supports and legally advises victims of cyberviolence. According to our knowledge, at the time of data collection, the association possessed the largest pool of information about the cyberviolence offender population in Switzerland. Through the association, we gained access to the contact information of this population. Because the selection of individuals is based on accessibility, we here deal with a nonprobabilistic convenience sample. This is the usual procedure for highly exploratory studies with hard-to-reach groups (Raifman et al. 2022). From this overall sample, we invited all offenders whose criminal online comments were made no more than three years ago (about 70 persons) to participate in interviews. By this approach, we do not aim for representativeness for cyberviolent offenders in Switzerland and therefore do not claim it. In the invitation letters, the individuals were informed about the aim and relevance of the study, the receipt of contact data by the association, that the researchers otherwise act independently of it, and the interview conditions: voluntariness, anonymity, confidentiality, and a small expense allowance for time spent and travel. Reminder letters were sent three weeks later.

We were able to conduct interviews with four adults who faced legal sanctions in Switzerland between 2016 and 2019. This is remarkable considering the highly sensitive topic of criminal delinquency and the fact that cyberviolence offenders typically prefer to remain anonymous and shy away from talking about their deeds. The latter, in turn, explains the lack of evidence on convicted offenders in the current literature. All interviewed offenders (Table 1) were reported for posting offensive or

Table 1 Information on the Interviewees and Their Offences

Pseudonym	Gender	Age group at interview	Topic of criminal comment	Mode of interview
Fritz	Male	60–70	Target's weight	Personal
Michael	Male	70–80	Target's marriage	Telephone
Laura	Female	40–50	Target's intelligence	Personal
Ralph	Male	40–50	Target's private life	Personal

defamatory comments directed against female public figures, a director of an NGO and a politician, on the social media platform Facebook. The offenders received substantial fines ranging from $ 300 to $ 1200, two of which were negotiated in out-of-court settlements.

4.2 Interview Procedure and Transcription

The interviews were conducted between June and July 2019. Each participant was interviewed once, and both principal researchers were present for every interview. The interviews were held in German and Swiss German, and each lasted about an hour. Prior to each interview, consent was obtained for the audio recording, transcription, and scientific analysis of the interview (in writing for in-person interviews and verbally for telephone interviews). A second written consent was obtained a short time after the interviews. The interview questions were semi-structured. They followed the chronological order of events, starting at the time of the publication of the comment through the legal process to the time after the legal conviction. Participants were asked about their perceptions, attitudes, and behaviours in relation to media use, the specific criminal comment, the subsequent legal process, and their reactions to it. The interviews were transcribed by an experienced transcriptionist. Swiss German was translated into German, leaving specific dialect expressions in the original. Interviews were transcribed orthographically, with all spoken words and sounds reproduced, including hesitations and pauses indicated by ellipses.

4.3 Thematic Analysis

We applied thematic analysis, a process of systematically identifying, describing, analysing, and reporting patterns of shared meaning (Clarke and Braun 2017). The aim was to identify common discursive themes that were relevant to the research question and to investigate whether these themes occurred to different extents and in different forms before and after the legal sanctions. We combined an inductive, data-driven approach with a deductive, theory-based approach. This combination lends itself to the availability of theoretical frameworks explaining crime, which requires an experiential orientation, and the simultaneous lack of research on legal measures in cyberviolence, which requires an exploratory orientation. Thematic analysis is therefore optimal because it allows theoretical flexibility that takes into account pre-existing theoretical categories but is not bound by them.

First, the authors familiarized themselves with the data by reading and rereading the interview transcripts several times with the rational choice and neutralization theories in mind. Second, the authors coded the transcripts, and initial themes inspired by the research question and theories were generated with MAXQDA. The coders were open to possible new themes emerging. The themes were generated both at the manifest semantic level, where expressions are taken at face value as describing what is said, and at a latent level, where the codes reflect the researchers'

interpretations of what is meant. Third, the themes were reviewed and reformulated in an iterative process of discussing disagreements among coders. This ensured that the final thematic framework was applied consistently and coherently to the data. In addition, content that contradicted the identified themes was sought. Fourth, the themes were clearly defined and concisely named. Compelling quotations were selected to illustrate the themes. To present the results, the quotations were translated from German into English. Individual sounds were ignored for better readability. Provided that the sense of the overall statement remained unchanged, irrelevant content between successive relevant statements was ignored but marked with ellipses.

5 Results and Discussion

The overall results indicate that after facing legal sanctions all offenders heavily censored themselves in the frequency and content of online comments. We explain this by two key mechanisms: Offenders increasingly calculate the expected costs of offensive comments, a pattern we identified in all offenders, and partially disable previous norm neutralizations, as identified in one offender. The following sections present in detail the data on which these conclusions are based. The themes are discussed according to the chronological order of events, from the offence up to the legal process and afterwards. Despite seeking distinctions between the themes, overlaps cannot be ruled out, as in other qualitative analyses.

5.1 Cost-Benefit Calculation at the Time of the Offence

Generally, we can identify all the relevant elements of rational-choice theories of crime in our data: the costs and benefits of online comments and the corresponding expectations of their occurrence. However, at the time of the offence, we note hardly any conscious calculations of expected costs. However, social benefits seem to be relevant.

› *Little cost calculation.* When recalling the actual moment of the criminal comment, the offenders report acting predominantly in a spontaneous decision-making mode (Kroneberg et al. 2010). They describe the comment posting as "spontaneous" (Michael) and "relatively quick" (Ralph). Hardly any conscious reflection is discernible: Ralph admits to having thought "not for a second" about consequences. The inconceivability of any costly consequences is vividly illustrated by Michael's answer when questioned about the extent to which he had expected to ever be punished: "Absolutely not! Aaabsolutely not! Otherwise I wouldn't have written this! I do not have too much money!". Slight reflection was only reported by Laura – but in a rather unexpected direction. She points out that she deliberately did not formulate the comment too offensively

because she wanted to spare the victim too much public humiliation. Here again, avoiding costly sanctions does not seem to have been salient.

› *Expected social benefits.* In contrast to the largely absent cost calculation, the benefits of social recognition seem obvious to offenders. Fritz and Ralph report that their criminal comments were socially recognized by their online peers ("the normal people" according to Fritz). Ralph mentions the many likes he received. Hesitantly laughing, he said that "it [the criminal comment] was well received, of course". Such likes also seem to be valued in the context of respondents' more general posting activity. The importance of bystanders for aggressive online behaviour and hate-filled echo chambers that normalize hate has already been pointed out elsewhere (e. g. Harel et al. 2020).

5.2 Cost-Benefit Calculation During and After the Legal Measures

During and after facing legal sanctions, offenders started to weigh the costs and benefits of their online behaviour more deliberately. Legal sanctions not only changed their subjective expectation of the probability of being caught but also their perception of the severity of legal sanctions. In addition, fewer benefits were expected.

› *Experienced severity of sanctions.* Over time, the offenders increasingly experience the emotional, financial, and potential social costs caused by the criminal comment. During the police interrogation, offenders experience emotional costs. Laura reports being "shocked" and Michael "not happy". Laura panics: "Jesus Christ, if I am charged and have a criminal record with a child and … oh, help, help!". Fritz is bothered by the time and effort involved in court proceedings. Besides, financial costs are perceived as severe, and subsequent social costs are feared. Laura, Ralph, and Michael all signalled that they had limited amounts of money at their disposal and rated the fines determined later as high and burdensome (e. g. "It annoyed me to pay it."). In contrast, Fritz reports that the fine "didn't hurt" but that he feared his reputation would suffer. Indeed, one result of the court case was that he was scrutinized in his role as a volunteer custodian.

› *Increased cost expectation.* Only during police questioning did the offenders realize that their criminal online comment was the reason for the charges against them. This underscores how future costs were hardly expected when posting the comment. Fritz admits to "having already forgotten about it [the comment]" in the meantime. Laura similarly reports that she mistakenly thought that she had been reported for a recent insult to another road user.

Thinking of the time since the legal process, the offenders report more consciously calculating the costs of online commenting. Accordingly, they act in a rational-deliberate mode (Kroneberg et al. 2010). Even during police questioning, the consideration of long-term consequences becomes clear (Laura:

"I was already thinking about the future while I was sitting there"). The offenders start to estimate detection as more probable, from which an expansion of their imagined online audience may be inferred (Marwick and Boyd 2011). This now includes individuals who may be offended by their posted content and law enforcement agencies (Michael: "I know that the enemy is listening"; Laura: "Everyone can read it"). They become "more cautious" (Ralph) and more aware of their digital identifying traces (Laura: "Online, you have the name, the address"). Not all offenders see this as a bad thing: Laura admits to being "glad to have been stopped". She describes her steep learning curve as follows:

Yes, you have now been shown … it is actually dangerous in the sense that you act rashly and you don't reckon with the consequences at all. If you post "something shitty" and someone doesn't like it, then they can report you and then you will be punished. And then you know that it will never happen again.

› *Decreased expected benefits.* In parallel, the benefits expected from online commenting seem to decrease. Some offenders started to perceive online commenting as ineffective and shifted their focus to the offline space. To Fritz, online commenting "is no use" and to Michael, it "really doesn't change anything". Laura refrains from publishing impulsive thoughts as they represent "a useless fart comment … not serving anyone". In addition, she mentions the shift to the offline context: "My interests now are actually with my son and what concerns everyday life. I prefer to live life 'live' and not in front of the computer". Similarly, others seem to deprioritize their commenting activities. Fritz remarks that "if I don't have it [online exchange], I don't have it. I can exchange [with others] in other ways as well … I go out and I live".

5.3 Neutralization Techniques at the Time of the Offence

We identify all the neutralization techniques proposed by Sykes and Matza (1957) in the interview data. Looking back to the moment of the crime, all offenders deny responsibility, an injury, and a victim, and all appeal to higher loyalties. The subsequent legal sanctions additionally motivate condemning the condemners. However, the experience of the legal sanctions hardly seems to weaken these justifications.

› *Denial of responsibility.* In describing the actual moment of the criminal comment, the offenders deny responsibility primarily by depicting the act as an emotionally impulsive reaction to an overstraining, allegedly headliner-focused news media environment. In particular, the offenders recall that "she [the victim]" and "public figures" generally are "always present … in every channel" (Laura), where viewers are exposed to "the same thing over and over again" (Ralph). Laura perceives this as an intrusion into her personal space "because I have my life and I don't want to be part of her [the victim's] life". This is

consistent with other studies finding that an online audience can perceive other users' excessive publishing of media content to violate privacy norms (Solove 2006). Ralph explains that "I let my emotions out in full" and Michael argues that "you can't always keep everything bottled up". Similarly, Laura explains how she had problems at work, and when she came home, instead of relaxing, she had to see "this [victim's] face" on TV and on the computer. Such statements suggest that the criminal comment is excused by a necessary release of externally provoked frustrations.

The offenders also refer to the latently hostile social media environment. The platform architecture is blamed for retaining users. Fritz talks about the time prior to the criminal comment: "If I could have deleted it [Facebook], I would have deleted it and then I wouldn't have commented anymore". A hostile online climate is suggested by offenders' reports of being insulted and blocked by others and vice versa. This climate relativizes offenders' criminal comments: "People actually only hate each other on Facebook" (Laura); "The others don't keep their mouths shut either" (Michael). However, offenders also refer to online peers as a normative source to legitimize their criminal acts. For instance, Fritz describes other users' outrage at the event that triggered his criminal comment. By highlighting the uncontrollable and unpredictable factors of the online environment that "nudge" users into delinquency (Brewer et al. 2018, 119), offenders deny responsibility.

› *Denial of injury.* In the moment of the offence, all offenders deny an injury by portraying the criminal comment as harmless. They play down its seriousness as "just a silly expression", "nothing evil" (Fritz), and "harmless" (Laura). Michael does not "feel I have personally offended anyone", and Laura positively distances herself from other users' hostile comments: "'She is a witch' or much, much worse". Michael emphasizes the disproportionality between online expressions and offline legal consequences: "I can't believe it [having been reported to the police] – only because of such a sentence on Facebook". The purported lack of direct harm presents the offence as more acceptable.

› *Denial of victim.* To deny the victim, offenders claim that the person targeted deserves victimization due to personal or public misconduct. The targets are accused of violating norms, whether falsely accusing men of sexual abuse to "ruin" them (Ralph) or failing to conform to female attractiveness norms (not "aesthetically pleasing"; Fritz). Such expressions are consistent with the widespread misogyny on social media (see e. g., Semenzin and Bainotti 2020). Targets are additionally accused of pushing themselves into the media spotlight too much. Laura argues that "those who are always in the public eye … contribute to these hate comments". People in the public spotlight thus "have to expect" criticism: "Every action provokes a reaction" (Laura). Such

attributed misconduct deprives the victims of treatment according to human morals and values (Bandura 1999).

Offenders also deny victims by not expecting their presence in the online space of the crime, while paradoxically imagining this space very vaguely and at least semipublic. Some felt deliberately deceived because a victim had apparently entered the space with a fake profile: Laura explains that "of course" she knew nothing about the victim's presence. Thus, not perceiving a victim materially or digitally denies the victim's existence. Simultaneously, the offenders reflected very little about their imagined online audiences (Marwick and Boyd 2011). Michael answers who he thinks may have read his comment with "I can't imagine that at all". The others remain very vague about who they think their audience was: "many" (Fritz); "like-minded people" (Laura); "general" (Ralph). The three offenders describe the social space of the crime very broadly as "Facebook". More concretely, Laura describes a Facebook group consisting mainly of "patriotic" but also of "left-wing and green" people, thus admitting to having been aware of the comment's publicity. Offenders here ignore the fact that semipublic or public disparagement also produces victims despite the lack of personal contact with targets.

› *Appealing to higher loyalties.* Finally, when respondents justify their criminal comments, they appeal to truth, freedom of expression, and injustice. The ideologies behind these ideas neutralize violent behaviour by framing it as moral action (Vysotsky and McCarthy 2017). The resulting harassment serves to enforce a certain morally correct social order on the Internet (Marwick 2021). Two offenders rationalize their comment by claiming to know the truth. Michael says: "Personally, I'm just rock-solidly convinced that that didn't happen the way she [the victim] writes it. … I wrote this because I was firmly convinced of it". Similarly, Ralph explains that "to me, it was simply clear: She is lying". Further, offenders appeal to free expression in like-minded online communities. The need to speak freely in such a context is normalized, for example as "human" (Laura). Indirectly defending the idea of digital enclaves (Harel et al. 2020), Laura argues for a separation of ideological online communities without mutual influence ("the groups … are for those who have the same opinion"). Lastly, the offenders resort to perceived social injustice, and their higher loyalty to this is appealed indirectly. They bring up diverse, apparently outrageous events that they have read about online, in particular related to politics, violence, and paedophilia: "terrible" and "insane" events, according to Michael. They describe their opposition to it by regularly posting comments on these events. For example, several offenders refer to immigrants and the feeling of being disadvantaged compared to them. Fritz, for example, is outraged by the state of affairs in neighbouring Germany:

> *I don't call them refugees, I call that (um) (thinking) don't know anymore. Just not refugees. They are immigrants. Yes, they want to get as much as possible from the social system. They are always … most of them are … at the beginning they always said that they were well-educated people. They are mostly useless … 95 % … 99 % useless. Just wrecking the German state.*

Expressing such grievances could be interpreted as an implicit justification of criminal comments as merely negligible incidents in a much larger campaign against injustice. In the face of these apparent grievances, the offenders report feeling "powerless", "because nothing can be done" (Michael). Similarly, Ralph feels ill-informed by "mainstream media" and unable to offer sufficient criticism due online content moderation.

5.4 Neutralization Techniques During and Since Facing Legal Sanctions

The offenders report that they had to undergo an interrogation after being invited to the police station by a letter that did not state the reason for the invitation. The result of the process was either an out-of-court settlement or a legal conviction. During this process, another neutralization technique was introduced: condemning the condemners. At the same time, denial of injury, denial of victim, and appealing to higher loyalties partially ceased, but only in one offender.

- *Condemning the condemners.* To divert attention from their offence, the offenders condemn the person who reported the offence to the police: either the victim or the person supporting the victim in the court proceeding. Offenders depict this supporter of the victim as hypocritical. Laura accuses her of using fake profiles to "hunt haters": "She was searching for us; she was searching for us". This person is also accused of enriching herself financially through out-of-court settlements: "It was all about the money" (Fritz); "This way you can also earn money!" (Ralph). This apparent practice is delegitimized as "incomprehensible" (Michael), "amusing", and "theatre" (Fritz). The settlement offers are dubbed "blackmail" (Ralph) and "hush money" (Fritz). The reporting person is also indirectly delegitimized through alleged hypersensitivity. Ralph stresses that he is not "squeamish" and does "not dream" of reporting others for similar offences. Likewise, Michael does not at all consider suing the "people in Thailand" who apparently call him a "long-nose" whenever he goes there.

  In contrast, the police and the courts as the ultimately enforcing state bodies are not explicitly condemned and are even viewed somewhat positively. Fritz expresses contentment with the legal proceedings as such. Laura and Ralph feel supported by the "quite nice" policemen: "The policeman was fully on my side". Laura expresses sympathy for "all people who are involved" in the process as she considers them to be equally burdened by the person reporting the offence: "There are wiser things in life. … The policeman had to

interrogate me. … [He] might have done other things". We cannot therefore identify a sense of injustice towards authorities, as has been observed among young people who commit piracy and accuse law enforcement of being unjust and inconsistent (Holt et al. 2019; Matza 1964). One exception is Laura, who considers her penalty "disproportionate" and "unjust". The prevalent support for the established institutions demonstrates the offenders' digital drifts (Goldsmith and Brewer 2015).

› *Discarding previous neutralizations.* The neutralizations appeared to be largely unchanged during and since the legal sanctions in three offenders. Fritz vividly expresses the stable denial of injury and victim in front of the court:

> *There was absolutely no repentance, I did nothing wrong and I stick to it … in court I just wrote that I feel completely in the right. I had actually only paid for it so that I would have peace of mind.*

In contrast, one offender discards his violence-accepting neutralizations. Ralph reports having developed a positive personal relationship with the victim through written contact: "We have written to each other … and we had a really good time". This interaction led him to acknowledge the injury and the victim: "I have even defended her on Facebook in some discussions". He recalls feeling guilty and having apologized:

> *I then also realized my mistake … Because I was … no longer sure … I mean, I wasn't there. I don't know what happened. So that's why I can't form a judgement. And therefore, I apologize of course. Sincerely. … I was then really sorry afterwards for what I had written there.*

Accordingly, he no longer appeals to the higher loyalty of truth and instead accepts uncertainty. He also reports stronger perspective-taking: "Well, that [the legal measures] has simply had the effect that … I thought about it. … How it comes across to the other person."

Although no causal relationship between the positive contact and the attenuated neutralizations can be conclusively established with the data available, this association is consistent with the hypothesized prejudice-reducing effects of positive interactions (Allport et al. 1954; Paluck et al. 2019).

5.5 Behavioural Change

For all offenders, the typical behavioural response to facing legal sanctions seems to be self-censorship when commenting online. They report commenting less frequently or no longer at all, whether violently or in general. Michael now shares critical thoughts only with friends but "not publicly": not in online groups that are accessible to a broad public. Laura mentions her attempts "to keep quiet about [her thoughts]". Laura and Fritz explicitly report nevertheless continuing to passively

observe what others comment, even though they "no longer react". Those offenders who still speak out avoid particular formulations: Fritz reports having become "very reserved", and Ralph remains "factual" and "decent". Further, less risky forms of expression are used, such as emoticons. Michael remarks: "Now I … have not written anymore. The most I can do is click 'like'. Or you can also click 'heart' or 'sad'".

6 Conclusion

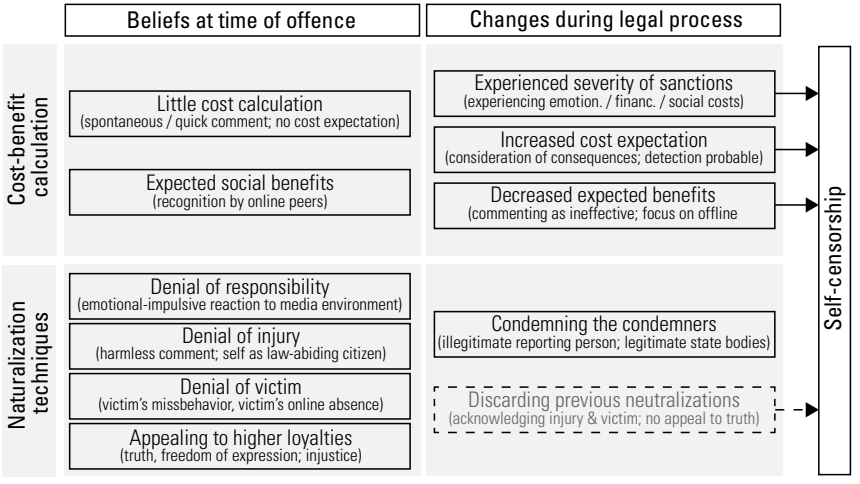
6.1 Summary

This study aimed to explore whether and how legal sanctions can prevent or reduce cyberviolence. Therefore, we investigated how offenders' experience of facing legal sanctions affected their reported perceptions, attitudes, and behaviours related to online commenting. Our findings (summarized in Figure 1) suggest that offenders typically respond to the legal process with self-censorship, resulting in a decrease in cyberviolence. We identify two key mechanisms that can explain this. The predominant mechanism, observed in all interviewed offenders, is a change from a barely considered cost calculation at the moment of the offence while still bearing some social benefits in mind to a deliberate inclusion of possible of sanctions in the weighing process. Following the legal conviction, sanctions are expected to be more severe and more likely. A less predominant mechanism, observed in only one offender, is the abandonment of neutralizing beliefs that hitherto had inactivated internalized normative and moral convictions, possibly triggered by positive interactions with the victim. Nevertheless, techniques of neutralizing moral norms are crucial to explaining criminal online comments in the first place. All offenders excuse the criminal comment as an emotional and impulsive reaction to an overwhelming media environment (denial of responsibility); portray the comment as harmless (denial of injury); claim that the victims deserve victimization because of their misbehaviour and because of the unkept online absence (denial of victim); and appeal to higher loyalties of truth, freedom of expression, and injustice. Finally, after having faced legal penalties, they condemn the people who reported them (condemning the condemner), but not the state bodies.

6.2 Contribution

This study contributes previously lacking knowledge to the literature on cybercrime about the effectiveness of legal sanctions on cyberviolence. By retrospectively tracking offenders' reported online behaviour from the act of cyberviolence to self-censorship, we provide first insights into the largely unexamined preventive and deterrent effects of legal sanctions against cyberviolence (El Asam and Samara 2016). By contrasting rational incentives with neutralization techniques before, during, and after

Figure 1 Cost-Benefit Calculations and Neutralization Techniques Regarding Criminal Online Comments Before and After Legal Process



Note: Dashed-grey box and arrow indicate observation in only one respondent.

the experience of legal measures, we also shed much-needed light on explanatory mechanisms (Xu et al. 2016; Stalans and Donner 2018). The current study suggests that legal penalties can effectively reduce cyberviolence due to the rational change they produce over time, as explained by rational choice theory (Gibbs 1985; Opp 2020). Conversely, reduced norm neutralizations are a less likely consequence of facing legal penalties. Nonetheless, our study suggests that the consequent change in perspective can be strengthened through direct positive interactions between offenders and victims.

More broadly, the suggested positive effect of legal sanctions can also be attributed to digital drift (Goldsmith and Brewer 2015). The offenders drifted easily into cyberviolence due to the effortless acceptance of justifications offered by the sociotechnical context and the nonsalience of costs. Legal measures then unexpectedly collapsed the online and offline space (Marwick and Boyd 2011). The consequent sharp increase in experienced and expected costs was accompanied by drifting out of cyberviolence, rendering this engagement "episodic and … trifling" without apparent enculturation in deviant cybercultures (Brewer et al. 2018, 115). Legal sanctions therefore seem to be particularly effective for previously law-abiding but "online-disinhibited" (Suler 2004) citizens. However, against the backdrop of offenders' persistent denial of harm from cyberviolence, there is a risk of collateral damage: groups that traditionally support law enforcement, such as the politically right-leaning individuals in the present sample, might lose trust in established legal institutions in the long

run if online speech is prosecuted more systematically. Moreover, enforcing laws that regulate violent expression in general and in the digital realm specifically may affect free expression in several ways: On the one hand, it benefits free expression by protecting the right of victims to express themselves without fear of violent reprisal. On the other hand, there is a risk of excessive self-censorship so that convicted offenders completely refrain from expressing themselves online, as our data suggest. However, given the current proliferation of unpunished cyberviolence, the latter concern seems still ill-founded (for more, see Bakalis 2018). Beyond this, however, there is also a risk that laws against cyberviolence will be misused by governments to censor noncompliant citizens, i.e., to criminalize activist online expression that does not in fact involve violence. This is particularly true in jurisdictions of less democratic societies.

6.3 Limitations

This study has several limitations that provide important avenues for future research. First, cyberviolent populations beyond the one examined here could be affected differently by neutralization beliefs, cost-benefit considerations, and legal sanctions. The present sample largely corresponds to the majority of disseminators of digital hate speech in the Swiss population (Stahel et al. 2022): they are largely low-income, male, right-wing conservatives. The last two characteristics are well-known predictors of digital hate in the literature, which may be related to the neutralization techniques compatible with violence-affirming masculinity norms and with advocating inequality between social groups. However, the sample differs in age, as disseminators in the Swiss population tend to be young. There is a possibility that among digital natives, the application of legal sanctions will lead to less self-censorship because their school-based education has informed them in advance about the laws against cyberviolence and the potential costs. In any case, the mechanisms studied here warrant further testing quantitatively, experimentally, and causally in larger and more diverse samples.

Secondly, we cannot exclude the possibility that the retrospective questions led to unintentional distortions and memory errors. Moreover, the participants could have answered untruthfully. However, information given by offenders in interviews is usually consistent with the official record (Wright and Bennett 1990). Furthermore, voluntarily agreeing to be interviewed makes lying meaningless, as the participants did not have to agree in the first place. We nevertheless addressed this problem by checking the consistency of our participants' statements. Future studies could advance this by triangulating data, including court documents, as far as privacy policies allow.

Third, our focus on the effect of legal sanctions intentionally ignores social mechanisms that may explain the development of the neutralization beliefs that allow cyberviolence in the first place. For example, whereas we point to social recognition

in online spaces as a form of expected benefits, future research could extend this to the differential association of offenders with potential online hate groups and the learning processes that might disengage such offenders from established norms and institutions (Akers and Jennings 2016).

Overall, this study offers innovative insights into the promising effect of legal countermeasures on engagement in cyberviolence among very hard-to-reach offenders. The results speak in favour of raising awareness of legal sanctions, actually enforcing laws in cyberspace, and promoting perspective-taking through positive victim–offender encounters.

7 References

Akers, Ronald L., and Wesley G. Jennings. 2016. Social Learning Theory. Pp. 230–240 in *The Handbook of Criminological Theory*, edited by Alex R. Piquero. West Sussex, UK: John Wiley & Sons, Inc.

Allport, Gordon Willard, Kenneth Clark, and Thomas Pettigrew. 1954. *The Nature of Prejudice*. Massachusetts: Addison-Wesley Publishing Company.

Backe, Emma Louise, Pamela Lilleston, and Jennifer McCleary-Sills. 2018. Networked Individuals, Gendered Violence: A Literature Review of Cyberviolence. *Violence and Gender* 5(3): 135–146. https://doi.org/10.1089/vio.2017.0056.

Bakalis, Chara. 2018. Rethinking Cyberhate Laws. *Information & Communications Technology Law* 27(1): 86–110. https://doi.org/10.1080/13600834.2017.1393934.

Bandura, Albert. 1999. Moral Disengagement in the Perpetration of Inhumanities. *Personality and Social Psychology Review* 3(3): 193–209.

Banks, James. 2010. Regulating Hate Speech Online. *International Review of Law, Computers & Technology* 24(3): 233–239. https://doi.org/10.1080/13600869.2010.522323.

Becker, Gary S. 1968. Crime and Punishment: An Economic Approach. *Journal of Political Economy* 76(2): 169–217.

Blaya, Catherine. 2018. Cyberhate: A Review and Content Analysis of Intervention Strategies. *Aggression and Violent Behavior* 45: 163–172. https://doi.org/10.1016/j.avb.2018.05.006.

Brewer, Russell Colin, Jesse Cale, Andrew John Goldsmith, and Thomas Holt. 2018. Young People, the Internet, and Emerging Pathways into Criminality: A Study of Australian Adolescents. *International Journal of Cyber Criminology* 12(1): 115–32. https://doi.org/10.5281/zenodo.1467853.

Brewer, Russell Colin, Sarah Fox, and Caitlan Miller. 2020. "Applying the Techniques of Neutralization to the Study of Cybercrime." Pp. 547–565 in *The Palgrave Handbook of International Cybercrime and Cyberdeviance*, T. J. Holt, and A. M. Bossler. Palgrave Macmillan.

Clarke, Victoria, and Virginia Braun. 2017. Thematic Analysis. *The Journal of Positive Psychology* 12(3): 297–298. https://doi.org/10.1080/17439760.2016.1262613.

El Asam, Aiman, and Muthanna Samara. 2016. Cyberbullying and the Law: A Review of Psychological and Legal Challenges. *Computers in Human Behavior* 65: 127–141. https://doi.org/10.1016/j.chb.2016.08.012.

Gibbs, Jack P. 1985. Deterrence Theory and Research. *Nebraska Symposium on Motivation* 33: 87–130.

Goldsmith, Andrew, and Russell Brewer. 2015. "Digital Drift and the Criminal Interaction Order." *Theoretical Criminology* 19(1): 112–130. https://doi.org/10.1177/1362480614538645.

Harel, Tal Orian, Jessica Katz Jameson, and Ifat Maoz. 2020. The Normalization of Hatred: Identity, Affective Polarization, and Dehumanization on Facebook in the Context of Intractable Political Conflict. *Social Media + Society* 6(2): 1–10.

Holt, Thomas J., Russell Brewer, and Andrew Goldsmith. 2019. Digital Drift and the "Sense of Injustice": Counter-Productive Policing of Youth Cybercrime. *Deviant Behavior* 40(9): 1144–1156. https://doi.org/10.1080/01639625.2018.1472927.

Kroneberg, Clemens, Isolde Heintze, and Guido Mehlkop. 2010. The Interplay of Moral Norms and Instrumental Incentives in Crime Causation. *Criminology* 48(1): 259–294. https://doi.org/10.1111/j.1745-9125.2010.00187.x.

Marwick, Alice E. 2021. Morally Motivated Networked Harassment as Normative Reinforcement. *Social Media + Society* 7(2): 1–15.

Marwick, Alice E., and Danah Boyd. 2011. I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience. *New Media & Society* 13(1): 114–133. https://doi.org/10.1177/1461444810365313.

Matza, David. 1964. *Delinquency and Drift.* New York: John Wiley & Sons.

McGuire, James. 2002. Criminal Sanctions versus Psychologically-Based Interventions with Offenders: A Comparative Empirical Analysis. *Psychology, Crime and Law* 8(2): 183–208. https://doi.org/10683160208415005.

Opp, Karl-Dieter. 2020. *Analytical Criminology: Integrating Explanations of Crime and Deviant Behavior*. London: Routledge.

Paluck, Elizabeth Levy, Seth A. Green, and Donald P. Green. 2019. The Contact Hypothesis Re-Evaluated. *Behavioural Public Polic* 3(2): 129–158. https://doi.org/10.1017/bpp.2018.25.

Park, Chang Sup, Qian Liu, and Barbara K. Kaye. 2021. Analysis of Ageism, Sexism, and Ableism in User Comments on YouTube Videos About Climate Activist Greta Thunberg. *Social Media + Society* 7(3): 1–14.

Paternoster, Raymond. 2010. How Much Do We Really Know about Criminal Deterrence? *The Journal of Criminal Law and Criminology* 100(3): 765–824.

Raifman, Sarah, Michelle A. DeVost, Jean C. Digitale, Yea-Hung Chen, and Meghan D. Morris. 2022. "Respondent-driven Sampling: a Sampling Method for Hard-to-reach Populations and Beyond." *Current Epidemiology Reports* 9(1): 38–47.

Sallavaci, Oriola. 2018. Crime and Social Media: Legal Responses to Offensive Online Communications and Abuse. Pp. 3–23 in *Cyber Criminology*, edited by Hamid Jahankhani. Cham: Springer.

Semenzin, Silvia, and Lucia Bainotti. 2020. The Use of Telegram for Non-Consensual Dissemination of Intimate Images: Gendered Affordances and the Construction of Masculinities. *Social Media + Society* 6(4): 1–12.

Solove, Daniel J. 2006. A Taxonomy of Privacy. *University of Pennsylvania Law Review* 154(3): 477–564.

Stahel, Lea, Sebastian Weingartner, Dirk Baier, and Katharina Lobinger. 2022. Digitale Hassrede in der Schweiz: Ausmass und sozialstrukturelle Einflussfaktoren. Biel: Office of Federal Communication (OFCOM).

Stalans, Loretta J., and Christopher M. Donner. 2018. Explaining Why Cybercrime Occurs: Criminological and Psychological Theories. Pp. 25–45 in *Cyber Criminology*, edited by Hamid Jahankhani. Cham: Springer.

Suler, John. 2004. The Online Disinhibition Effect. *Cyberpsychology & Behavior* 7(3): 321–26. https://doi.org/10.1089/1094931041291295.

Sykes, Gresham M., and David Matza. 1957. Techniques of Neutralization: A Theory of Delinquency. *American Sociological Review* 22(6): 664–670.

Vysotsky, Stanislav, and Adrienne L. McCarthy. 2017. Normalizing Cyberracism: A Neutralization Theory Analysis. *Journal of Crime and Justice* 40(4): 446–461. https://doi.org/10.1080/0735648X.2015.1133314.

Wikström, Per-Olof H. 2017. Character, Circumstances, and the Causes of Crime. Pp. 502–521 in *The Oxford Handbook of Criminology*, edited by Alison Liebling, Shadd Maruna, and Lesley McAra. Oxford: Oxford University Press.

Wright, Richard, and Trevor Bennett. 1990. Exploring the Offender's Perspective: Observing and Interviewing Criminals. Pp. 138–151 in *Measurement Issues in Criminology*, edited by Kimberly, L. Kempf. New York, NY: Springer.

Xu, Bo, Zhengchuan Xu, and Dahui Li. 2016. "Internet Aggression in Online Communities: A Contemporary Deterrence Perspective." *Information Systems Journal* 26 (6): 641–67. https://doi.org/10.1111/isj.12077.