Handbook of Peace Psychology

Christopher Cohrs, Nadine Knab & Gert Sommer (Eds.)

ISBN 978-3-8185-0565-3

DOI: https://doi.org/10.17192/es2022.0074

Editing and formatting: Michaela Bölinger and Johanna Hoock

Cover picture and chapter design: Nadine Knab

Cover picture: Hope (Esperanza). Peace, gratitude, creativity and resilience are the symbols and elements that are harmonised in this artwork. In large format, it is part of the graffiti tour in Community 13 in Medellín, Colombia. The artwork conveys an important message of hope to both the local community and foreign visitors.
@medapolo.trece @fateone96 @radycalshoes @pemberproducciones

https://handbuch-friedenspsychologie.de
Website design: Tamino Konur, Iggy Pritzker, Nadine Knab

Forum Friedenspsychologie
https://www.friedenspsychologie.de

Online Moral Courage

Julia Sasse[1,2,3*], Niklas Cypris[1,4*] & Anna Baumert[1,4]

Abstract

Individuals and groups are frequently targets of bullying, sexual harassment, and hate speech on online platforms. Such norm violations can have detrimental negative consequences, for instance by causing psychological harm and damaging social cohesion. Finding ways to reduce and prevent online norm violations is hence crucial. Online users may play an important role in this context. We argue that it can be considered *morally courageous* if users decide to take a stand against perceived violations of their own moral beliefs and endorsed norms, as it may imply substantial risks for themselves. With this chapter, we aim to advance our understanding of online moral courage as a relatively new phenomenon. First, we provide an examination of critical characteristics of online environments that may facilitate or hinder moral courage. Second, we discuss consequences of online moral courage by considering its effects on perpetrators, further online users, and the general tonality of the online discourse. Last, we integrate insights on the facilitators and obstacles of online moral courage and its consequences to provide practical recommendations for the design and management of online platforms and user education and training.

Keywords: Moral courage, hate speech, harassment, anonymity, reach, bystander effect, prosociality

Zusammenfassung

Einzelpersonen und Gruppen sind auf Online-Plattformen häufig Ziel von Mobbing, sexueller Belästigung und Hassrede. Solche Normverstöße können schwerwiegende negative Folgen haben, indem sie beispielsweise psychisches Leid verursachen und den sozialen Zusammenhalt schädigen. Es ist daher zwingend notwendig, Wege zu finden, um Online-Normverletzungen zu reduzieren und zu verhindern. Nutzer*innen von Online-Plattformen können in diesem Zusammenhang eine wichtige Rolle spielen. Wir argumentieren, dass es als *zivilcouragiert* angesehen werden kann, wenn Nutzer*innen sich entscheiden, sich gegen Verletzungen ihrer eigenen moralischen Überzeugungen und gebilligter Normen zu stellen, da dies erhebliche Risiken für sie selbst bedeuten kann. Mit diesem Kapitel möchten wir das Verständnis von Online-Zivilcourage als relativ neues Phänomen weiterentwickeln. Zunächst untersuchen wir relevante Merkmale von Online-Umgebungen, die Zivilcourage fördern oder behindern können. Danach diskutieren wir die Konsequenzen von Online-Zivilcourage unter Berücksichtigung ihrer Auswirkungen auf Täter, weitere Online-Nutzer*innen („Bystander“)

[1]Max Planck Institute for Research on Collective Goods, Bonn
[2]Technical University of Munich
[3]Applied University Ansbach
[4]University of Wuppertal
*These authors contributed equally to this work

Handling editor for this chapter: N. Knab, accepted on 23.05.2022

und die allgemeine Tonalität des Online-Diskurses. Schließlich integrieren wir Erkenntnisse zu förderlichen und hinderlichen Faktoren für Online-Zivilcourage und ihren Folgen, um praktische Empfehlungen für die Gestaltung und die Handhabung von Online-Plattformen sowie für die Anleitung und das Training von Nutzer*innen zu geben.

Schlüsselwörter: Zivilcourage, Hassrede, Belästigung, Anonymität, Reichweite, Bystander-Effekt, Prosozialität

Online Moral Courage

Moral courage manifests in interventions intended to stop or redress others' transgressions of moral principles or social norms, despite the risk of incurring physical, financial, or social costs (Frey, Peus, Brandstätter, Winkler & Fischer, 2006; Greitemeyer, Fischer, Kastenmüller & Frey, 2006; Halmburger, Baumert & Schmitt, 2016; Niesta Kayser, Frey, Kirsch, Brandstätter & Agthe, 2016). Following this definition, a broad range of actions qualify as moral courage, for example interventions against bullying, discrimination, or oppression (Baumert, Li, Sasse & Skitka, 2020; Li, Sasse, Halmburger & Baumert, 2021). While such interventions may involve confrontation of and conflict with transgressors (see also Sasse, Li & Baumert, 2022), their ultimate goal is to uphold and defend moral principles or social norms that ensure the sound functioning of societies (Ellemers, van der Toorn, Paunov & van Leeuwen, 2019; Fehr & Gächter, 2002). In line with this, moral courage has also been considered as a behavior that is characterized by one's caring for others and that protects human and democratic values (Staub, 2015; Meyer, 2014). In the present chapter, we argue that moral courage manifests itself and plays an important role also in online contexts. We analyze the specific affordances and barriers posed by the online context, review empirical evidence on the positive effects of online moral courage, and propose practical recommendations for its enhancement.

Since our social interactions – spanning from friendships, dating, learning, to political debate – take place in considerable and increasing extent on social networking sites (SNS), also transgressions of moral principles and social norms occur in online contexts. According to a recent survey (Vogels, 2021), 41% of the participating US adults had personally experienced some form of online harassment. Within just six years (2014 to 2020), the share of people reporting severe forms of online harassment, such as physical threats or stalking, increased steeply, from 15% to 25%. Often, individuals and groups experience harassment or hate because of their political views, gender, ethnicity, religion, or sexual orientation (Vogels, 2021).

Citizens and policy makers alike have identified online norm transgressions as a major problem with a multitude of negative socio-psychological consequences (Vogels, 2021; van der Wilk, 2018). An Amnesty International survey (Dhrodia, 2017), investigating the effects

of online abuse and harassment on women, revealed that many targeted women subsequently experienced stress, anxiety, panic attacks, or lowered self-esteem as a consequence, and changed their own online behavior, up until the point of turning silent and withdrawing from online spaces altogether. Online norm transgressions can also aggravate social relations. Frequent exposure to hate speech against outgroups has been associated with increased prejudice towards those groups (Soral, Bilewicz & Winiewski 2018), and research from Germany has shown that increased anti-refugee sentiment on Facebook translated into higher crime rates against refugees (Müller & Schwarz, 2021), suggesting that online hate speech may spill over to physical violence offline.

The prevalence and ramifications of online norm transgressions call for effective countermeasures. Other online users can play an important role in this regard, just like bystanders in response to offline norm transgressions. If they perceive the actions of others as a violation of their moral convictions, or of social norms that they endorse, they may take steps to stop or redress these actions, for example by engaging in counterspeech or by reporting to authorities. While taking such steps is often socially desirable, it is not without risk to the person doing so. For instance, those who confront the norm transgressions of others might themselves quickly become the next target of harassment and hate. As such, taking action against online norm transgressions can be considered *morally courageous*.

The Need for Online Moral Courage

Similar to offline norm transgressions, the types of situations and contexts in which they occur are highly diverse and encompass, for example, cyberbullying, sexual harassment, and hate speech. Despite their differences, all these violations have in common that perpetrators violate fundamental social norms and moral values, such as fairness, and that they cause harm. In many cases, they also constitute transgressions of international legal agreements (such as the EU Framework Decision of 2008 (CFD, 2008)) and national law (such as the "incitement to hatred" paragraph in Germany, §130 StGB).

While policy-makers and citizens see online platform providers as responsible for detecting and dealing with violations (Dhrodia, 2021), doing so ex ante or proactively can prove difficult for them, often for technical (Ross, Rist, Carbonell, Cabrera, Kurowsky & Wojatzki, 2016), ethical (Post, 2009), or legal reasons (Zufall, Horsmann & Zesch, 2019). This highlights the need for community engagement, by which users who encounter content that violates social norms or their moral beliefs intervene in order to uphold and ensure civil discourse. Depending on the online environment and user rights, they can do so in various ways, for example by directly confronting the transgressor (e.g., through counterspeech), by banning transgressors from groups, or indirectly by reporting them. All these forms of interventions require at least some time and effort (e.g., the interruption of conversations or work, writing a reply or a report), and it is plausible that they bear risks for the person taking the action, ranging from receiving unwanted attention, harsh criticism, to backlash as a direct response

to actions, or to negative consequences that transpire into offline contexts and affect relationships, professional life, or physical well-being[1]. As such, intervening against online norm transgressions qualifies as online moral courage.

To date, research on moral courage has thus far mainly been conducted in offline environments. Here, theoretical and empirical work has pointed out that moral courage requires complex psychological processes, and whether or not individuals intervene against others' norm transgressions may depend on a range of individual and situational factors (Baumert, Halmburger & Schmitt, 2013; Halmburger, Baumert & Schmitt, 2016; Li, Sasse, Halmburger & Baumert, 2021; Niesta Kayser, Greitemeyer, Fischer & Frey, 2010; Toribio-Flórez, Sasse & Baumert, 2021). As offline and online environments differ in various ways, for example with regard to anonymity, situational factors in online contexts may shape the psychological processes of moral courage in unique ways.

In this chapter, drawing from a theoretical model of moral courage – the *integrative model of moral courage* by Halmburger and colleagues (2016) – we first identify several crucial situational characteristics of online environments and discuss how they may obstruct or facilitate the psychological processes underlying online moral courage. Second, we discuss both potential beneficial and adverse consequences of online moral courage. Third, we synthesize insights on the psychological processes and the consequences of online moral courage to derive practical recommendations that may inform platform policies and the work of practitioners.

Most evidence reviewed in this chapter stems from research on interventions against hate speech on SNS and we highlight whenever we draw from further research on further forms of online norm transgressions.

Obstacles and Facilitators of Online Moral Courage

What determines whether individuals show moral courage? According to the integrative model of moral courage (Halmburger et al., 2016, adapted from Latané & Darley, 1970), prior to acting, observers must *detect* the norm violation and *interpret* it as such, and they must then *assume responsibility* and the necessary *skills to intervene*, and finally *decide to intervene*. According to the model, only if each of these stages is passed successfully moral courage will be shown. For example, even if an observer interprets an instance of hate expressed against members of a minority as wrong, but do not feel responsible to address it, they will not do so.

Whether or not the stages of psychological processes are passed successfully should depend on characteristics of the individual person, as well as of the situation (Halmburger et

[1] While these risks may seem less apparent for reporting perpetrators to authorities, bystanders may still be concerned about them. Depending on the platform, the reporting process may be somewhat intransparent so that the own anonymity may not be seen as ensured or there may be concerns that perpetrators can infer who reported them.

al., 2016). Situation characteristics, in particular, may differ between online and offline contexts. In this chapter, we focus on five prominent characteristics of online contexts, which, we argue, can work as both facilitators and obstacles of moral courage, namely

(a) reach (Bor & Petersen, 2021; Brady, Crockett & Van Bavel, 2019; Obermaier, Fawzi & Koch, 2015; Ziegele, Naab & Jost, 2020),

(b) connectedness (Amichai-Hamburger, 2017),

(c) permanence (Barberá, Jost, Nagler, Tucker & Bonneau, 2015; Dillon & Bushman, 2015; Obermaier et al., 2015),

(d) asynchrony (Allison & Bussey, 2016; Obermaier et al., 2015; Suler, 2004), and

(e) anonymity (Obermaier et al., 2015; Postmes & Turner, 2015; Suler, 2004; Ziegele et al., 2020).

With reach, we refer to the fact that online environments provide the opportunity to communicate with large or distant audiences with little effort. Moreover, people cannot only unidirectionally reach out to other people across the world via the internet, but they can just as easily communicate multidirectionally and network with others, for example to mobilize and organize like-minded individuals. Especially SNS facilitate this *connectedness*. The reach of online communication is further enhanced through a temporal component. While statements made in face-to-face conversations are often of an ephemeral nature, those made online are rather *permanent,* as they remain accessible for a long time, providing the chance that more people will become aware of them or reproduce them at a later point. The permanence of online communication also allows for it to happen *asynchronously*. That is, interactions do not need to be temporally contingent. Instead, people can reply to messages months after they were originally posted. Another critical characteristic of online contexts is *anonymity*. In many online environments, users have – or can choose to have – no or few personal markers that make them identifiable. As such, communication partners can remain anonymous, rendering it uncertain who is making or reading a statement. We argue that these aspects of *reach*, *connectedness*, *permanence, asynchrony,* and *anonymity* can be both obstacles and facilitators for the psychological processes of online moral courage.

Detection and Interpretation of Online Norm Violations

For moral courage to occur, observers first need to detect the norm transgression and interpret it as such. While this may seem trivial, these processes are not always straightforward. For example, imagine coming across a comment on social media in which one user calls another '*bitch*'. From reading just this term, it is difficult to infer whether this is a sexist insult or whether a group of friends uses the term in a playful way to address each other. In other words, the intention for using the term is ambiguous and thus difficult to infer for observers. Consequently, ambiguity is a barrier to moral courage (Bowes-Sperry & O'Leary-Kelly, 2005;

Halmburger et al., 2016; Toribio-Flórez et al., 2021). In online communication, some factors can increase – and others reduce – ambiguity.

Often, individuals and groups who intentionally and frequently transgress norms online disguise their communication to make it particularly difficult for witnesses to detect and interpret transgressions. For instance, transgressors use ciphers to refer to specific marginalized groups without detection from outside witnesses and prosecution. For example, Black people are sometimes referred to with a capitalized "N" or Jews with three parentheses (e.g., commenting "(((they))) are behind everything"). Similarly, transgressors use codes to communicate hateful sentiments, such as '88' instead of 'Heil Hitler'. Plausibly, the connectedness in the online context facilitates the rapid development of hateful jargon, making it particularly difficult for users to detect and interpret transgressions.

Just as connectedness can contribute to norm transgressions, it may also facilitate their detection. Bystanders do not need to act alone, but instead may form groups to coordinate the detection of transgressions and initiate concerted interventions. For instance, groups such as Reconquista Internet (Garland, Ghazi-Zahedi, Young, Hébert-Dufresne & Galesic, 2020) and #ichbinhier (#iamhere) (Ley, 2018; Ziegele et al., 2020) inform their members about occurrences of hate and vitriolic language, so that members can seek them out and counter them collectively. That way, the detection of transgressions and their interpretation as such do not fall upon individuals, but are organized, thereby facilitating the passing of the first stages of the psychological processes in moral courage.

The interpretation of norm transgressions may also be affected by temporal asynchrony and permanence. On the one hand, if norm-transgressing posts remain visible for a long period of time without being visibly challenged, users might question whether any gut feelings of inappropriateness are in fact warranted. On the other hand, permanence and temporally asynchronous interaction provides users who suspect a norm transgression, for example behind jargon, with time to reflect and inform themselves. This way, ambiguity can be reduced which should facilitate subsequent psychological processes of online moral courage.

Assuming Responsibility

Once observers have interpreted a norm transgression as such, they need to determine whether intervening falls within their responsibility.

Here, the prevalent asynchrony of online contexts may pose a hurdle. In case of older hate speech, people may assume that the issue has been resolved outside of the visible communication channel (Allison & Bussey, 2016), or that the communication had moved on with no further need to circle back (Leonhard, Rueß, Obermaier & Reinemann, 2018).

In addition, the assumption of responsibility seems to depend on the number of bystanders present, and in online contexts with typically high reach, they are often many. For

example, in the context of cyberbullying, Obermaier and colleagues (2016; Study 2) found that students had lower intentions to intervene against cyberbullying when many bystanders were present, compared to very few. This effect was mediated by (lower) feelings of responsibility (see also Machackova, Dedkova & Mezulanikova, 2015; Song & Oh, 2018). Similarly, Leonhard et al. (2018) found that people were more likely to speak up against anti-immigrant hate speech when only four other SNS users saw the transgressive post as opposed to 4,000. These findings suggest that, with an increasing number of bystanders, *diffusion of responsibility* may occur (Darley & Latané, 1968; Fischer et al., 2011).

However, the negative association between number of bystanders and intervention behavior in computer-mediated communication does not always seem to be linear, as the actual and the perceived number of bystanders do not increase proportionately (Machackova et al., 2015; Obermaier et al., 2016). Instead, increases up to two dozen bystanders are perceived disproportionately larger than increases above that, and 24 bystanders are already considered rather many (Obermaier et al., 2016). This might lead to the finding that the bystander effect is more pronounced for increases in smaller groups of bystanders than for increases in bigger groups of bystanders (Machackova et al., 2015) and that there is no linear trend at all once hundreds of bystanders are involved (Obermaier et al., 2016). Potentially, this is because, at a certain point, the sheer number and heterogeneity of individuals in a large audience increase the chances that other factors facilitating interventions are present and outweigh the diffusion of responsibility. For example, Voelpel, Eckhoff and Förster (2008) proposed that the number of so-called “perpetual helpers”, individuals with a generally elevated disposition to help, increases with audience size. Hence, while findings suggest that diffusion of responsibility may be prevalent in the online context, the vast reach might at a certain point also serve to counter-act this effect by increasing the odds for the presence of more individuals who are generally disposed to act prosocially.

Subjective Intervention Skills

Beyond assuming responsibility, observers need to determine whether they dispose of the necessary – and effective – skills to intervene and have the opportunity to do so.

As mentioned earlier, norm transgressions are often committed by organized groups, facilitated by the connectedness in online contexts. For example, in the context of the 2017 German elections, the right-wing hate group *Reconquista Germanica*, which had only 1,500 to 3,000 members, published millions of vitriolic posts on Twitter in order to shift the online discourse in the direction of right-wing populism (Garland et al., 2020). In the face of concerted incivility and hate, it seems plausible that bystanders might feel unable to counter such attacks substantially and effectively.

At the same time, the majority of SNS provides guidelines for intervention, and many forms of intervention require little skill or effort, which might lower the threshold for inter-

vening. Due to temporal asynchrony, even users who may not be familiar with given standards have the opportunity to inform themselves about different intervention options. In general, interventions can be conducted either directly or indirectly (Latané & Darley, 1970). In the online context, direct interventions refer to actions such as writing counter-comments against group-based hate comments on a SNS. Also, other easy-to-implement measures can be taken in many online settings to express disagreement with norm transgressions, such as dislike functions to reject hate speech by others. An indirect way of intervening, instead, would be to notify the relevant authorities, for example by reporting the post to the SNS provider or a moderator. Indirect interventions are facilitated across most social media through functions like flagging and reporting of transgressive comments, which can normally be done with a few clicks (Naab, Kalch & Meitz, 2018). Plausibly, connectedness between users and providers or moderators enhances the knowledge of effective intervention options.

Decision to Intervene

According to the integrative model of moral courage, the final step of the psychological process is the decision to intervene. The model proposes that, at this point, individuals weigh the expected benefits against personal costs which they might suffer as a result of intervening. Those costs can range from the mere investment of time and effort to intervene to the loss of money, physical harm, or backlash from transgressors, as well as drawing unwanted attention to themselves or being evaluated by others (Latané & Darley, 1970; Schwartz & Gottlieb, 1976).

Plausibly, permanence that characterizes communication in many online environments may foster concerns about the costs of interventions. The fact that in online environments evidence of one's actions often prevails until long after the exchange has taken place (Slonje & Smith, 2008) might trigger fears that one's intervention might be perceived negatively by a wider audience (Dillon & Bushman, 2015; Fischer et al., 2011). Moreover, when engaging with users who use uncivil language, the fear of being associated with them for an unforeseeable amount of time could further raise perceived personal costs (Ziegele et al., 2020). The long-term documentation of one's direct intervention might also invite direct retaliation, such as online harassment or physical violence in the offline world, not only by the original transgressor, but also their sympathizers. Due to the broad reach of online environments, their number can be assumed to be high, but is often unknown to interveners, which may be perceived as particularly threatening.

However, with a large audience, potential interveners might not only fear backlash, but also anticipate support from like-minded individuals. A strong predictor for people speaking up against uncivil language online is expected positive social appraisal (Ziegele et al., 2020) and SNS offer various ways for bystanders to reward morally courageous comments (e.g., likes, following accounts, writing a supportive comment of one's own, retweeting, etc.).

Thus, if individuals anticipate support from others, large reach might also promote the decision to intervene.

Moreover, barriers to indirect means of intervention are often explicitly reduced in online environments. As mentioned above, most SNS offer low-effort ways to flag or report norm transgressions. Given that indirect interventions can often preserve the anonymity of interveners, such clear sets of indirect intervention measures should reduce the perceived riskiness of (indirect) intervention.

The anonymity of many online contexts, which emerges due to a scarcity of individualizing markers (e.g., a lack of visual representation of individuals), can also shape decisions to intervene by affecting the salience of group norms. A person's self-image is made up of individual characteristics as well as social group memberships (Tajfel, 1974; Tajfel & Turner, 1986). When a particular group membership becomes salient, people tend to see themselves more in terms of that group membership and consequently to act more in line with the respective group norms (Turner, Hogg, Oakes, Reicher & Wetherell, 1987). According to the *Social Identity Model of Deindividuation Effects* (SIDE; Reicher, Spears & Postmes, 1995), this salience is increased in environments with few individualizing markers. If individual characteristics of a person recede in a situation due to anonymity, their salient group membership becomes more influential and shapes attitudes and behavior. Thus, in contexts where norms of a salient group favor moral courage, members of that group can actually be more likely to engage in interventions (Levine & Crowther, 2008) and this effect can be especially strong in contexts of computer-mediated communication where reduced individuating cues trigger increased conformity with one's group (Lee, 2004; Postmes et al., 2001). In summary, group norms that support bystander interventions can shape the decision to intervene - in particular in an environment such as computer mediated communication as it does not contain many individuating components.

Consequences of Online Moral Courage

After having discussed how critical characteristics of online context may impact the processes driving moral courage, we now look at the consequences of interventions against online norm transgressions. Specifically, we will focus on the consequences of user-generated counterspeech as an intervention against hate speech that we consider morally courageous. We define counterspeech as any direct response to a transgression such as openly criticizing the hate comment or expressing solidarity with the hate target. Counterspeech is not limited to written comments but can also take the form of images, gifs, and other modes of communication. Consequences of user-generated counterspeech can be determined for different online agents, in particular transgressors and other online users, or through an inspection of the general tone of the online discourse. With regard to transgressors, it is necessary to establish whether counterspeech actually reduces the occurrence of future transgressions or, on the contrary, it might even stimulate backlash. Similarly, other online users

might react to interventions, either favorably or unfavorably. Lastly, on a superordinate level, courageous counterspeech might impact the general tonality of online discussion.

Effects on Transgressors

Potential effects of counterspeech on the original hate speaker are not well-understood and evidence for the effectiveness of interventions is mixed. For example, counterspeech that addressed people who posted anti-Roma comments on Facebook did not substantially change the sentiment of their subsequent comments in the conversation (Miškolci, Kováčová & Rigová, 2020). Still, there is initial evidence that, under the right circumstances, the behavior of norm transgressors may be influenced by counterspeech. In an experiment on Twitter, Munger (2017) showed that it mattered *who* spoke up against racial slurs. He confronted men who used the racial slur *"n****r"* to insult others on Twitter with counterspeech by an ostensible other Twitter user. Importantly, this user was either White (i.e., the participants' in-group) or a Person of Color (i.e., the participants' out-group) and had either few or many followers. Transgressors were less likely to use the slur subsequently, but only after counterspeech by a popular ingroup member (i.e., White with many followers), compared to a no-intervention control condition. These results suggest that counterspeech *can* be effective in reducing the occurrence of racial slurs, but further research is necessary to determine its boundary conditions comprehensively.

Effects on Bystanders

Compared to the effects of interventions on transgressors, those on further online users are much better understood. In general, there is ample evidence that users have a tendency to adjust their tone to the rhetoric they encounter in a given online discussion, both with regard to civil and constructive behavior (Berry & Taylor, 2017; Han & Brazeal, 2015; Han, Brazeal & Pennington, 2018; Molina & Jennings, 2018; Seering, Kraut & Dabbish, 2017) and uncivil, disruptive online behavior (Cheng, Bernstein, Danescu-Niculescu-Mizil & Leskovec, 2017; Garland et al., 2020; Gervais & Hillard, 2014; Seering et al., 2017). For example, in an experimental SNS setting, participants were either shown comments in favor of or against Chinese people. In the hostile condition, participants were substantially more likely to post comments against Chinese than in the pro-Chinese condition (Hsueh, Yogeeswaran & Malinen, 2015). Taken together, these results indicate a general and substantial malleability of online conversations in line with prior comments.

Indeed, this emulation effect has also been observed for speaking up against transgressions. For example, Han and colleagues (2018) showed that online users were more likely to call for more civil discourse in response to vitriol in comment sections if someone else had already done so, compared to when they only saw hateful comments. People were also more likely to engage in metacommunication, that is, talking about the general tone of a conversation and asking for it to improve, when they saw comments made by someone else arguing

in the same direction (Molina & Jennings, 2018). Also, if some people criticized hate speech on social media, other users were more likely to speak up in favor of the attacked marginalized group (Miškolci et al., 2020).

Different psychological mechanisms could be at play when it comes to the emulation of counterspeech. Comments could serve as primes, making certain kinds of commenting behavior more salient and accessible. Moreover, the comments could exert their influence on further behavior by informing people about social norms. Here, other commenters would serve as exemplars – people who are perceived as prototypical for the respective group as a whole, which can shape perceptions of group norms (Klein et al., 2007; Zillmann, 2002). As mentioned above, people seem especially prone to adhering to such ingroup norms in deindividuated online communication (Reicher et al., 1995). Indirect evidence for this idea stems from work which showed that the status of a commenter within a group determined whether other users copied their prosocial and antisocial behavior (Seering et al., 2017). If users with high status on the SNS (i.e., moderators as well as people with a paid subscription), as compared to users who did not hold such a status, displayed prosocial behavior in the stream chat, the next ten messages were substantially more likely to contain similar behavior. It thus seems plausible that also interventions from high-status group members are particularly effective.

However, some studies could not find evidence for the emulation of counterspeech. For instance, Leonhard and colleagues (2018) did not find counterspeech emulation in the domain of anti-refugee hate speech in Germany, and mixed effects have been found for counterspeech against hate speech that is directed at different kinds of marginalized groups such as women, Jews, welfare-recipients, and members of the LGBT community (Kunst, Porten-Cheé, Emmer & Eilders, 2021; Mathew et al., 2019). Thus, more research is needed to determine limits of and boundary conditions for the emulation effect.

In summary, there is tentative evidence that when people interact online, they tend to be substantially affected by the behavior of others, which might be due to the perception of social norms that is informed by their actions.

General Tonality

Counterspeech can break a spiral of aggressiveness, leading to an overall more respectful discourse. For example, Obermaier and colleagues (2021) demonstrated that counterspeech against Islamophobic hate speech reduced the willingness of Muslim participants (i.e., members of the targeted group) to react with hateful countercomments to the hate speech. There are also preliminary indications that organized counterspeech has a potentially positive effect on the overall discussion climate. A large-scale study investigated the interactions between Reconquista Germanica and Reconquista Internet on Twitter (Garland et al., 2020). The amount of counterspeech by Reconquista Internet was associated with decreased occurrence as well as decreased extremity of hate speech by transgressors. In addition, unlike

counterspeech by non-organized counterspeakers, which was sometimes associated with strong hate speech backlash, organized counterspeech under the Reconquista Internet umbrella was only associated with more counterspeech and neutral comments. While organized counterspeech does not always achieve the ideal of maintaining a purely rational discussion (Keller & Askanius, 2020), its overall effect thus tends to pacify online discourse.

Practical Recommendations

In this final section, we synthesize insights on the facilitators and obstacles of online moral courage and its effects in order to derive practical recommendations. To promote moral courage in online environments, misperceptions need to be addressed and modalities of SNS can be leveraged to facilitate interventions across the different steps of the intervention process, as laid out in the integrative model of moral courage (Halmburger et al., 2016).

Highlighting Responsibility

Through prevalent narratives like “Don’t feed the troll” and the focus on SNS providers and the government as responsible to combat hateful behavior through deletion and prosecution, people may not see themselves in the place to display moral courage on SNS and intervene against norm transgressions. However, just as in the offline sphere, civil society can play an important role in creating and protecting a tolerant and inclusive environment online. This role should be communicated more broadly, so that people are more disposed to assuming personal responsibility.

Competency

In order to facilitate counterspeech, online users can focus on its positive effects on other uninvolved witnesses. While trying to change a transgressor’s mind might be quite difficult and often depends on characteristics that lie outside the intervener’s control (e.g., shared group memberships, Munger, 2017), targeting an audience of neutral and sympathetic bystanders will often prove easier and more successful (Keller & Askanius, 2020). While there are some indications that different methods of counterspeech are favored depending on which kinds of hate speech and incivility are addressed (such as using humor to ridicule or compassion to connect with a norm transgressor; Mathew et al., 2019), the general rule is that even very short and simple counterspeech can be beneficial. SNS users can be influenced even by the simplest messages such as emojis (Seering et al., 2017), and simple statements of approval or disapproval of a message without further elaboration can already be sufficient to affect the discussion climate positively due to the disposition to copy behavior. Moreover, not even one’s own counterspeech is necessary to leave a positive impact. Bystander interventions by others can be supported at even lower effort levels by showing support and affirmation for counterspeech comments (e.g., by “liking” them). As mentioned above, positive

feedback is a driver of bystander interventions by more active interveners (Ziegele et al., 2020). Highlighting the positive effects of low-key online behavior can help to activate and empower online users.

Costs

To facilitate online moral courage, its potential costs need to be reduced. On the one hand, the danger of attracting unwanted attention due to the reach of the internet can be reduced individually by limiting the information that is displayed to others. Many SNS enable users to withhold certain profile information from strangers, and users should consider which of their information actually needs to be publicly available, and which should only be accessible to a more limited audience.

Also, a clear communication of easily and anonymously implementable measures such as flagging and reporting from platforms can reduce costs, due to evaluation apprehension and effort, and has been shown to be effective (Naab et al., 2018). Also increasing awareness of the often substantial support for counterspeech online (Keller & Askanius, 2020) could further serve to reduce fears of negative reactions by a wider audience to one's own intervention.

Organized Counterspeech

Bystander interventions can further be facilitated through the promotion of coordinated counterspeech efforts. Computational simulations found that the ratio of counterspeakers to hate speakers is strongly associated with the effectiveness of counterspeech on Facebook (Schieb & Preuss, 2016), and similar effects were observed on Twitter (Garland et al., 2020): Counterspeech becomes more effective the more people engage in it. Moreover, counterspeech as a concerted effort makes it easier for each individual to join in. Existing groups like #ichbinhier and Reconquista Germanica facilitate interventions on various steps of the *integrative model of moral courage* (Halmburger et al., 2016): Critical situations are pointed out (*detection*) and defined as requiring bystander action (*interpretation*). Group members are characterized as responsible agents (*assuming responsibility*) and are provided with the necessary skills and training to address norm transgressions competently (*skills to intervene*). Finally, mutual support by group members can increase personal benefits of interventions (*decision to intervene*). In the face of a threatening outgroup, visibility to a supportive ingroup makes people speak up even for norms that might be punished by the outgroup (Reicher & Levine, 1994), and others' support can be a substantial predictor for intervention behavior (Ziegele et al., 2020).

Conclusion and Outlook

Online norm transgressions such as hate speech and bullying have become a major issue, as they harm individuals, groups, and the societal discourse (Vogels, 2021). Moral courage could play an important role in the attempt to reduce the occurrence of norm transgressions online and attenuate their detrimental effects especially on social media; yet, to date, this role is not well understood. In this chapter, following the integrative model of moral courage (Halmburger et al., 2016), we highlighted how some defining features of online environments and online communication, namely *reach*, *connectedness*, *permanence, asynchrony,* and *anonymity*, might be both facilitators and obstacles at different stages of the psychological process of moral courage. These considerations are of theoretical relevance for our understanding of online moral courage and may provide a road map for its future comprehensive investigation.

Our analysis of the consequences of online moral courage showed that it can have the intended positive effects. Yet, especially with regard to effects on transgressors, the evidence is scarce and somewhat mixed so far. Further research is needed, especially to stake out under what conditions intervention exerts a positive effect on transgressors – preferably in realistic settings, as done by Munger (2017). Going beyond isolated investigations of effects of interventions on perpetrators, bystanders, and the general tonality, a joint analysis under consideration of reciprocal effects also seems advised. This allows for a more accurate approximation of reality and would thus advance a comprehensive understanding of online moral courage.

Our systematic approach to online moral courage, coupled with the analysis of its consequences, further allows us to derive practical recommendations. Though the research on online moral courage is still in its infancy, given the first indications that it can be effective, it is crucial for SNS to facilitate it. For instance, platforms should protect the anonymity of counterspeakers. Moreover, platforms can choose to make counterspeech pronouncedly visible to their communities. Such recommendations may help to develop and govern SNS in a way that lowers the threshold for interventions. Moreover, our recommendations can aid the development of intervention training and the organization of effective, collective interventions, thereby contributing to a more peaceful and respectful online environment, while simultaneously reducing risks for interveners.

References

Allison, K. R. & Bussey, K. (2016). Cyber-bystanding in context: A review of the literature on witnesses' responses to cyberbullying. *Children and Youth Services Review*, *65*, 183–194. https://doi.org/10.1016/j.childyouth.2016.03.026

Amichai-Hamburger, Y. (2017). *Internet psychology: the basics* (1st ed.). London: Routledge.

Dhrodia, A. (2017). *Unsocial media: the real toll of online abuse against women*. Amnesty Global Insights. Available at: https://medium.com/amnesty-insights/unsocial-media-the-real-toll-of-online-abuse-against-women-37134ddab3f4

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A. & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Bastiaensens, S., Vandebosch, H., Poels, K., Van Cleemput, K., DeSmet, A. & De Bourdeaudhuij, I. (2014). Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior*, *31*, 259–271. https://doi.org/10.1016/j.chb.2013.10.036

Baumert, A., Halmburger, A. & Schmitt, M. (2013). Interventions against norm violations: dispositional determinants of self-reported and real moral courage. *Personality and Social Psychology Bulletin*, *39*(8), 1053–1068. https://doi.org/10.1177/0146167213490032

Baumert, A., Li, M., Sasse, J. & Skitka, L. (2020). Standing up against moral violations: psychological processes of moral courage. *Journal of Experimental Social Psychology*, *88*, 103951. https://doi.org/10.1016/j.jesp.2020.103951

Berry, G. & Taylor, S. J. (2017). Discussion quality diffuses in the digital public square. *Proceedings of the 26th International Conference on World Wide Web - WWW '17* (pp. 1371–1380). Perth, Australia: ACM Press. https://doi.org/10.1145/3038912.3052666

Bor, A. & Petersen, M. B. (2021). The Psychology of online political hostility: a comprehensive, cross-national test of the mismatch hypothesis. *American Political Science Review*, 1–18. https://doi.org/10.1017/S0003055421000885

Bowes-Sperry, L. & O'Leary-Kelly, A. M. (2005). To act or not to act: the dilemma faced by sexual harassment observers. *Academy of Management Review*, *30*(2), 288–306. https://doi.org/10.5465/amr.2005.16387886

Brady, W. J., Crockett, M. & Van Bavel, J. J. (2019). The MAD Model of moral contagion: The role of motivation, attention and design in the spread of moralized content online. *PsyArXiv,* preprint. https://doi.org/10.31234/osf.io/pz9g6

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C. & Leskovec, J. (2017). Anyone can become a troll: causes of trolling behavior in online discussions. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17* (pp. 1217–1230). Portland, OR: ACM Press. https://doi.org/10.1145/2998181.2998213

Council Framework Decision 2008/913/JHA. (2008). Official Journal of the European Union, Series L, 328, 45-58. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008F0913&from=EN

Darley, J. M. & Latané, B. (1968). Bystander intervention in emergencies: diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*, 377–383. https://doi.org/10.1037/h0025589

Dhrodia, A. (2017). *Unsocial media: the real toll of online abuse against women*. Amnesty Global Insights. Available at: https://medium.com/amnesty-insights/unsocial-media-the-real-toll-of-online-abuse-against-women-37134ddab3f4

Dillon, K. P. & Bushman, B. J. (2015). Unresponsive or un-noticed?: Cyberbystander intervention in an experimental cyberbullying context. *Computers in Human Behavior*, *45*, 144–150. https://doi.org/10.1016/j.chb.2014.12.009

Ellemers, N., van der Toorn, J., Paunov, Y., & van Leeuwen, T. (2019). The psychology of morality: a review and analysis of empirical studies published from 1940 through 2017. *Personality and Social Psychology Review*, *23*(4), 332–366. https://doi.org/10.1177/1088868318811759

Fehr, E. & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–140. https://doi.org/10.1038/415137a

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., Heene, M. et al. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, *137*, 517-537. https://doi.org/10.1037/a0023304

Frey, D., Peus, C., Brandstätter, V., Winkler, M. & Fischer, P. (2006). Zivilcourage. In H.-W. Bierhoff & D. Frey (Eds.), *Handbuch der Sozialpsychologie und Kommunikationspsychologie* (pp. 180–186). Göttingen: Hogrefe. Available at: https://www.zora.uzh.ch/id/eprint/98092/

Garland, J., Ghazi-Zahedi, K., Young, J.-G., Hébert-Dufresne, L. & Galesic, M. (2020). Countering hate on social media: Large scale classification of hate and counter speech. In Association for Computational Linguistics (Ed.), *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp 102–112). Stroudsburg, PA: Association for Computational Linguistics. Available at: http://dx.doi.org/10.18653/v1/2020.alw-1.13

Gervais, S. J. & Hillard, A. L. (2014). Confronting sexism as persuasion: effects of a confrontation's recipient, source, message, and context. *Journal of Social Issues*, *70*, 653–667. https://doi.org/10.1111/josi.12084

Greitemeyer, T., Fischer, P., Kastenmüller, A. & Frey, D. (2006). Civil courage and helping behavior: differences and similarities. *European Psychologist*, *11*, 90–98. https://doi.org/10.1027/1016-9040.11.2.90

Halmburger, A., Baumert, A. & Schmitt, M. (2016). Everyday heroes: determinants of moral courage. In S.T. Allison, G.R. Goethals & R.M. Kramer (Eds.), *Handbook of Heroism and Heroic Leadership* (pp. 165–184). London: Routledge.

Han, S.-H. & Brazeal, L. M. (2015). Playing nice: modeling civility in online political discussions. *Communication Research Reports*, *32*, 20–28. https://doi.org/10.1080/08824096.2014.989971

Han, S.-H., Brazeal, L. M. & Pennington, N. (2018). Is civility contagious? Examining the impact of modeling in online political discussions. *Social Media + Society*, *4*(3). https://doi.org/10.1177/2056305118793404

Hsueh, M., Yogeeswaran, K. & Malinen, S. (2015). "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors?: Online comments on prejudice expression. *Human Communication Research*, *41*, 557–576. https://doi.org/10.1111/hcre.12059

Keller, N. & Askanius, T. (2020). Combatting hate and trolling with love and reason? A qualitative analysis of the discursive antagonisms between organized hate speech and counterspeech online. *Studies in Communication and Media*, *9*, 540–572. https://doi.org/10.5771/2192-4007-2020-4-540

Klein, O., Spears, R. & Reicher, S. (2007). Social identity performance: extending the strategic side of SIDE. *Personality and Social Psychology Review*, *11*, 28–45. https://doi.org/10.1177/1088868306294588

Kunst, M., Porten-Cheé, P., Emmer, M. & Eilders, C. (2021). Do "Good Citizens" fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments. *Journal of Information Technology & Politics*, *18*, 258–273. https://doi.org/10.1080/19331681.2020.1871149

Latané, B. & Darley, J. M. (1970). *The unresponsive bystander: why doesn't he help?* (Century psychology series). Englewood Cliffs, NJ: Prentice-Hall.

Leonhard, L., Rueß, C., Obermaier, M. & Reinemann, C. (2018). Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *Studies in Communication Media*, *7*, 555–579. https://doi.org/10.5771/2192-4007-2018-4-555

Levine, M. & Crowther, S. (2008). The responsive bystander: How social group membership and group size can encourage as well as inhibit bystander intervention. *Journal of Personality and Social Psychology*, *95*, 1429–1439. https://doi.org/10.1037/a0012634

Ley, H. (2018). #ichbinhier: Zusammen gegen Fake News und Hass im Netz [#iamhere: Together against fake news and hate in the net]. Köln: DuMont.

Li, M., Sasse, J., Halmburger, A. & Baumert, A. (2021). Standing up against moral transgressions: An integrative perspective on the socio-psychological antecedents and barriers to moral courage. *under review*.

Machackova, H., Dedkova, L. & Mezulanikova, K. (2015). Brief report: The bystander effect in cyberbullying incidents. *Journal of Adolescence*, *43*, 96–99. https://doi.org/10.1016/j.adolescence.2015.05.010

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P. et al. (2019). Thou Shalt Not Hate: Countering Online Hate Speech. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 369–380.

Meyer, G. (2014). Zivilcourage und ihr Kern. In G. Meyer (Ed.) *Mut und Zivilcourage: Grundlagen und gesellschaftliche Praxis* (pp. 19-36). Leverkusen: Verlag Barbara Budrich.

Miškolci, J., Kováčová, L. & Rigová, E. (2020). Countering hate speech on Facebook: the case of the roma minority in Slovakia. *Social Science Computer Review*, *38*, 128–146. https://doi.org/10.1177/0894439318791786

Molina, R. G. & Jennings, F. J. (2018). The role of civility and metacommunication in Facebook discussions. *Communication Studies*, *69*, 42–66. https://doi.org/10.1080/10510974.2017.1397038

Müller, K. & Schwarz, C. (2021). Fanning the flames of hate: social media and hate crime. *Journal of the European Economic Association*, *19*, 2131–2167. https://doi.org/10.1093/jeea/jvaa045

Munger, K. (2017). Tweetment effects on the tweeted: experimentally reducing racist harassment. *Political Behavior*, *39*, 629–649. https://doi.org/10.1007/s11109-016-9373-5

Naab, T. K., Kalch, A. & Meitz, T. G. (2018). Flagging uncivil user comments: effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society*, *20*, 777–795. https://doi.org/10.1177/1461444816670923

Niesta Kayser, D., Frey, D., Kirsch, F., Brandstätter, V. & Agthe, M. (2016). Zivilcourage. In H.-W. Bierhoff & D. Frey (Eds.), *Soziale Motive und soziale Einstellungen* (Vol. 2, pp. 255–275). Göttingen: Hogrefe.

Niesta Kayser, D., Greitemeyer, T., Fischer, P. & Frey, D. (2010). Why mood affects help giving, but not moral courage: comparing two types of prosocial behaviour. *European Journal of Social Psychology*, *40*, 1136–1157. https://doi.org/10.1002/ejsp.717

Obermaier, M., Fawzi, N. & Koch, T. (2015). Bystanderintervention bei Cybermobbing. *Studies in Communication Media*, *4*, 28–52. https://doi.org/10.5771/2192-4007-2015-1-28

Obermaier, M., Fawzi, N. & Koch, T. (2016). Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying. *New Media & Society*, *18*, 1491–1507. https://doi.org/10.1177/1461444814563519

Obermaier, M., Schmuck, D. & Saleem, M. (2021). I'll be there for you? Effects of Islamophobic online hate speech and counter speech on Muslim in-group bystanders' intention to intervene. *New Media & Society*. https://doi.org/10.1177/14614448211017527

Palasinski, M. (2012). The roles of monitoring and cyberbystanders in reducing sexual abuse. *Computers in Human Behavior*, *28*, 2014–2022. https://doi.org/10.1016/j.chb.2012.05.020

Post, R. (2009). Hate Speech. In I. Hare & J. Weinstein (Eds.), *Extreme Speech and Democracy* (pp. 123–138). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199548781.001.0001

Postmes, T. & Turner, F. M. (2015). Deindividuation, Psychology of. *International Encyclopedia of the Social & Behavioral Sciences*, 38–41. Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.24015-4

Reicher, S. D., Spears, R. & Postmes, T. (1995). A Social Identity Model of Deindividuation Phenomena. *European Review of Social Psychology*, *6*, 161–198. https://doi.org/10.1080/14792779443000049

Reicher, S. & Levine, M. (1994). Deindividuation, power relations between groups and the expression of social identity: the effects of visibility to the out-group. *British Journal of Social Psychology*, *33*, 145–163.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N. & Wojatzki, M. (2016). Measuring the reliability of hate speech annotations: the case of the European refugee crisis. In M. Beißwenger, M. Wojatzki & T. Zesch (Eds.), *NLP4CMC III: 3rdWorkshop on Natural Language Processing for Computer-Mediated Communication, 22 September 2016*. Bochum: Bochumer Linguistische Arbeitsberichte. https://doi.org/10.17185/duepublico/42132

Sasse, J., Li, M., & Baumert, A. (2022). How prosocial is moral courage? *Current Opinion in Psychology*, *44*, 146–150. https://doi.org/10.1016/j.copsyc.2021.09.004

Schieb, C. & Preuss, M. (2016). Governing hate speech by means of counter speech on Facebook [conference paper]. *66th ica annual conference, at Fukuoka, Japan* (pp. 1–23).

Schwartz, S. H. & Gottlieb, A. (1976). Bystander reactions to a violent theft: crime in Jerusalem. *Journal of Personality and Social Psychology*, *34*, 1188–1199.

Seering, J., Kraut, R. & Dabbish, L. (2017). Shaping pro and anti-social behavior on Twitch through moderation and example-setting. In C. P. Lee & S. Poltrock (Eds.) *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17* (pp. 111–125). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2998181.2998277

Slonje, R. & Smith, P. K. (2008). Cyberbullying: another main type of bullying? *Scandinavian Journal of Psychology*, *49*, 147–154. https://doi.org/10.1111/j.1467-9450.2007.00611.x

Song, J. & Oh, I. (2018). Factors influencing bystanders' behavioral reactions in cyberbullying situations. *Computers in Human Behavior*, *78*, 273–282. https://doi.org/10.1016/j.chb.2017.10.008

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136–146. https://doi.org/10.1002/ab.21737

Staub, E. (2015). *The roots of goodness and resistance to evil: inclusive caring, moral courage, altruism born of suffering, active bystandership, and heroism*. Oxford: Oxford University Press.

Strafgesetzbuch (2021). §130. Bundesanzeiger Verlag.

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior, 7*(3), 321–326. http://dx.doi.org/10.1089/1094931041291295

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social Science Information*, *13*(2), 65–93. https://doi.org/10.1177/053901847401300204

Tajfel, H. & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & G.A. William (Eds.), *Psychology of intergroup relations* (2nd ed., pp. 7–24). Chicago, IL: Nelson-Hall.

Toribio-Flórez, D., Sasse, J. & Baumert, A. (2021). "Proof under reasonable doubt": Ambiguity of the norm violation as boundary condition of third-party punishment. *Personality and Social Psychology Bulletin, 0*(0). https://doi.org/10.1177/01461672211067675

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. & Wetherell, M. S. (1987). *Rediscovering the social group: a self-categorization theory*. Oxford: Basil Blackwell.

van der Wilk, A. (2018). *Cyber violence and hate speech online against women*. Brussels: Policy Department for Citizen's Rights and Constitutional Affairs. Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2018/604979/IPOL_STU(2018)604979_EN.pdf

Voelpel, S. C., Eckhoff, R. A. & Förster, J. (2008). David against Goliath? Group size and bystander effects in virtual knowledge sharing. *Human Relations*, *61*, 271–295. https://doi.org/10.1177/0018726707087787

Vogels, E. A. (2021). *The State of online harassment*. Available at: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

Ziegele, M., Naab, T. K. & Jost, P. (2020). Lonely together? Identifying the determinants of collective corrective action against uncivil comments. *New Media & Society*, *22*, 731–751. https://doi.org/10.1177/1461444819870130

Zillmann, D. & Bryant, J. (2002). *Media Effects: Advances in Theory and Research.* New York, NY: Taylor and Francis.

Zufall, F., Horsmann, T. & Zesch, T. (2019). From legal to technical concept: towards an automated classification of German political Twitter postings as criminal offenses. *Proceedings of the 2019 Conference of the North*, 1337–1347. https://doi.org/10.18653/v1/N19-1135

Julia Sasse is Professor for General Psychology and Media Effects at the Applied University Ansbach and affiliated researcher at the Institute for Ethics in Artificial Intelligence at TUM. She studied at the University of Düsseldorf and received her PhD in Social Psychology from the University of Groningen in 2017. In her research, she investigates the functions of emotions in the context of norm transgressions and social conflicts in offline and online environments.

Correspondence concerning this article should be addressed to Julia Sasse, Applied University Ansbach, Residenzstrasse 8, 91522 Ansbach, Germany, e-mail: julia.sasse@hs-ansbach.de.

Niklas Cypris is a Research Fellow at the Bergische Universität Wuppertal and an affiliated researcher at the Max-Planck-Institute for Research on Collective Goods, Bonn, as well as the Institute for Ethics in Artificial Intelligence at the Technical University of Munich. He studied at the University of Cologne and received his M.Sc. in psychology in 2020. For his doctoral thesis, he investigates behavioral effects of personalized interventions against online norm violations.

Foto: @JasperBastian

Anna Baumert is Professor for Social and Personality Psychology at the University of Wuppertal and Leader of the Max Planck Research Group "Moral Courage" at the Max Planck Institute for Research on Collective Goods in Bonn. She obtained her PhD in 2009 from the University of Koblenz-Landau. In her research, she investigates the interplay of social and personality processes for the explanation of perceptions and reactions to social injustice.

Niklas Cypris (cypris@uni-wuppertal.de), Anna Baumert (baumert@uni-wuppertal.de), University of Wuppertal, School of Human and Social Sciences, Gaussstrasse 20, 42097 Wuppertal, Germany.