Association for Information Systems
AIS Electronic Library (AISeL)

ICIS 2022 Proceedings | Social Media and Digital Collaboration

Dec 12th, 12:00 AM

Chances and Limits of Community-Based Hate Speech Detection – Results from a Combined Behavioral-NeuroIS Study

Vita Eva Maria Zimmermann-Janssen
Heinrich Heine University, vita.zimmermann@hhu.de

Nadine R. Gier
Heinrich Heine University, nadine.gier@hhu.de

Follow this and additional works at: https://aisel.aisnet.org/icis2022

Recommended Citation

Zimmermann-Janssen, Vita Eva Maria and Gier, Nadine R., "Chances and Limits of Community-Based Hate Speech Detection – Results from a Combined Behavioral-NeuroIS Study" (2022). *ICIS 2022 Proceedings*. 14.
https://aisel.aisnet.org/icis2022/social/social/14

This material is brought to you by the International Conference on Information Systems (ICIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in ICIS 2022 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Chances and Limits of Community-Based Hate Speech Detection – Results from a Combined Behavioral-NeuroIS Study

Completed Research Paper

Vita E. M. Zimmermann-Janssen
Heinrich Heine University, Germany
Universitätsstr. 1, 40225 Düsseldorf
vita.zimmermann@hhu.de

Nadine R. Gier
Heinrich Heine University, Germany
Universitätsstr. 1, 40225 Düsseldorf
nadine.gier@hhu.de

Abstract

Communication via social media is characterized by immediacy and anonymity, enabling free expression and sharing of opinions, but also the abuse of language in form of hate speech. Given the volume of online content, IS research offers approaches to efficiently detect hate speech. However, research and politics call for more independent, transparent, and social approaches to increase credibility and acceptance. In response, this two-part behavioral and neural study investigates flagging as a community-based solution to hate speech detection. By experimentally varying the displayed shares of flagging users and testing behavioral responses, results reveal opposing behavioral patterns as a function of the valuation of hate speech prevention. Moreover, by framing the display of the user community's flagging behavior as a sort of social normative information and hate speech prevention as a public good, the theoretical model might help explain (seemingly) conflicting results in social norm and public goods research.

Keywords: Hate speech, social media, social norms, collective action, NeuroIS, fNIRS, crowd-based solutions, flagging, altruism theory, crowding-out, public goods

Introduction

With the advent of Web 2.0 services, the Internet took on a central role in the exercise of freedom of speech (United Nations Human Rights Council [UNHRC], 2011). It transformed formerly passive users into senders of information, enabling them to freely express themselves, discuss and share information. Maybe ironically, it seems that it is also the Web 2.0 characteristics of immediacy and anonymity contributing to the growing abuse of speech harming the rights of others (Ullmann & Tomalin, 2020). This phenomenon of online hate speech (HS) can be observed especially in social media, and affects not only the victims but also societal structures (Meske & Bunde, 2022).

HS can be defined as any form of communication "that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, (…) based on identity factors" (United Nations [UN], 2020, p. 10). Accordingly, HS should neither be confused with communication that contains unacceptable expressions without the intent to offend, nor with offensive language directed against a person based on individual characteristics (Plaza-del-Arco et al., 2021). HS involves a generalization of negative qualities and a target group or target person who is considered representative of that group, to whom negative characteristics are attributed because of identity factors, such as ethnicity or sex (UN, 2020). Depending on HS's severity, HS must or can be lawfully restricted, though, specific forms of HS are to date not lawfully prohibited. Especially, this latter form of 'bottom level HS' rather calls for non-legal measures, according to international human rights law. Thus, in this bottom area of HS, it is up to society to decide whether offensive content is threatening the social fabric

and should thus be prevented. Social media platforms such as Facebook, YouTube, and Twitter have already implemented alike procedures, allowing the removal of as HS classified content (Ullmann & Tomalin, 2020). But deciding which content is bottom level HS is challenging not only because of the enormous content posted daily in various forms (Meske & Bunde, 2022; Ullmann & Tomalin, 2020). Also, the line between HS and appropriate free expression is blurred (Kapil & Ekbal, 2020). Moreover, this decision is still largely in the hands of human moderators of social media platforms (Plaza-del-Arco et al., 2021; Ullmann & Tomalin, 2020). Considering the sheer volume of content posted daily, this requires a tremendous amount of content moderators, numbering in the thousands on Facebook alone (Newton, 2019). In addition, this human factor can be viewed critically, not only for reasons of efficiency and accuracy but also because it risks overemphasizing the interests or capabilities of technology companies.

The efficiency and accuracy problem has recently gained attention in IS research. Machine learning approaches for HS detection (i.a., MacAvaney et al., 2019), human decision supporting AI systems (i.a., Meske & Bunde, 2022), and new concepts of dealing with offensive comments, such as treating it as malware with quarantining mechanisms (Ullmann & Tomalin, 2020), have been proposed, discussed and partially tested. The danger of overemphasizing platform interests, and the balancing of platforms', governmental, and user interests, still seems unresolved. In that light, there are raising voices from research and politics not only calling for more social but particularly independent and transparent measures of HS detection (MacCarthy, 2020; UN, 2020; UNHRC 2011). The question arises as to whether the common socio-technical mechanism of user flagging combined with a transparent display of the user community's flagging behavior might not be an alternative way toward a more independent, discoursive, and thus a more democratic form of HS detection. The provision of such information would not only transparently map the opinion of the user community, but also give users the opportunity to co-negotiate the collective's opinion by adjusting their own flagging behavior in response. In addition, it would also reduce the number of human moderators required, resulting in corporate efficiency gains.

The transparent presentation of the user community's flagging behavior is basically a specific form of normative information, called descriptive normative information (DNI). Numerous experiments have already investigated the impact of DNI on consumer decisions in various domains, in some cases even using it as an intervention to consciously guide behavior (Melnyk et al., 2010). DNI interventions have also attracted research interest for many years in the field of private provision of public goods, which, as will be explained, includes hate speech prevention. Typically, DNI experiments use stimuli such as "70% of people behave in a certain way". Initially, experiments converged on a uniform pattern of behavioral response, which regularly coincided with the communicated majority (i.e., Agerström et al., 2016; Bartke et al., 2017). However, in recent years, more and more studies have 'surprisingly' led to non-significant or even opposite results (e.g., Neumann, 2019; Zimmermann-Janssen, 2020). The unconstrained efficacy of DNI has since been debated and studies have been criticized for insufficient theorizing, ultimately preventing research from explaining these disparate findings (van Teunenbroek et al., 2020). Accordingly, this study not only addresses the question of whether a transparent display of the user community's flagging behavior could be a sufficient alternative for detecting HS online. It also proposes a theoretical solution to overcome the conflicting theories and evidence. To this end, a bipartite study consisting of an experimental behavioral part and a neural part was conducted. Both experiments provide convincing evidence that the theoretical solution may eventually be able to unify the theories.

Theoretical Framework and Hypotheses

Displaying Other Users' Behavior as a Sort of Normative Information

Social norms are rules of behavior indicating what is (in)appropriate in social contexts (Cialdini et al., 1991). Thereby, they can take two forms (e.g., Cialdini et al., 1991), addressing different motivational structures in decision-making. While injunctive social norms describe what is socially (not) approved of and therefore should be done, descriptive social norms rather specify behavioral rules by indicating common behavior. A typical descriptive normative information (DNI) thus refers to the behavior of others, showing either majorities that exhibit a certain behavior or minorities that refrain from it. As such, DNI provides an efficient mental shortcut in decision-making by triggering the reasoning that if most behave this way, it must be a good action (ibid.). In that sense, the transparent display of user community behavior itself not only contains DNI, but also recommends a behavior, simply put: 100% of users have flagged the comment, so flag the comment – or – 0% of users have flagged the comment, so please don't.

Thus, by framing the transparent display of user community behavior as DNI, further research on DNI interventions seems useful to derive behavioral expectations.

Hate Speech Prevention as Public Good

The reasoning for the either lawful or unlawful prevention of hate speech (HS) is based on the idea that HS "constitutes incitement to discrimination, hostility or violence" and threatens the "right to freedom of expression" as stated in Articles 19(2) and 20(2) of the International Covenant on Civil and Political Rights (1966). Thus, hate speech prevention (HSP) can be understood as ensuring freedom of speech, equality, and inviolability. As such, it can be interpreted as a public good in fulfilling the characteristics of non-rivalry and non-excludability (Duncan, 2004). Continuing this attribution, a user's contribution to HSP can be seen as an act of private provision of a public good. Accordingly, theoretical approaches from this research area, among others, will be used during hypotheses derivation.

Conflicting Research on the Effect of Descriptive Normative Information

Reflections on the Negative Impact Argument

By classifying HSP as a public good, theories on the private provision of public goods, explaining why individuals voluntarily contribute, represent a suitable conceptual starting point. Originally, pure public good models (i.e., Olson, 1971) assumed that the motivational structure behind the private provision of public goods consists purely of utility calculations arising from the public good itself. They postulate the utility of an individual i to be the sum of the utilities derived from private consumption x_i and the aggregated levels of public goods provided, G. The utility function, $U_i = U_i(x_i, G)$, predicts that the private provision of public goods must fall short in Pareto-efficiency since the consumption of G is not limited to those having contributed to G (Olson, 1971). Thus, a self-interested i would be motivated to 'free-ride', that is, to enjoy the benefits of the public good without having directly provided G. Moreover, as i tries to maximize the marginal utilities from x_i and G, a rising contribution of others to G should crowd-out i's contribution, since the marginal value of i's contribution to G drops (e.g., Roberts, 1984). However, not only was the private contribution found to be above the theoretically predicted levels but also the crowding-out effect was found to be less than dollar-by-dollar (e.g., Andreoni, 1989). Theorists argued that these discrepancies stem from the fact that i not only derives utility from G, but also from the pure act of contributing, g_i, extending the utility function to $U_i = U_i(x_i, G, g_i)$. Among others, Andreoni (1989) suggested in his impure public goods model g_i to be a positive feeling of 'warm glow' resulting from the satisfaction associated with the contribution. The warm glow effect, interpreted somewhat modified as moral satisfaction (Kahneman & Knetsch, 1992), stimulated further research. While Margolis (1982) assumed warm glow to be positively related to the size of the contribution, impact philanthropy theory suggests warm glow to be additionally dependent on the impact of i's contribution to G (Duncan, 2004).

What unites these theoretical approaches is the assumption that an increase in the contributions of others to G decreases the contribution of i, either due to a decreasing marginal value or decreasing relative impact of the contribution. As argued earlier, displaying the user community's flagging behavior makes the contributions of others salient. Consequently, the display of higher shares of flagging users should minimize i's contribution, which translates into lower probabilities of flagging, and vice versa.

Reflections on the Positive Impact Argument

Perhaps the most obvious argument for a positive influence of DNI on norm-compliant behavior can be directly derived from the motivational source of descriptive norms. As depicted before, descriptive norms allow individuals to reach a decision quickly by applying the simplest heuristic 'what many do will already be right'. Considering that people tend to prefer less effortful decision-making to solve problems, as implied by cognitive misers theory (Fiske & Taylor, 1984), deciding to behave as the DNI recommends represents a comparatively effortless option. Another argument is rather related to other-regarding preferences. Many alternative models – i. e. fairness, conformity, or reciprocity models – have been proposed that incorporate these other-regarding preferences but diverge in preferences' nature and motivation (for an overview see Fehr & Schmidt, 2003). Experimental research on social norm interventions, like the integration of DNI in the choice architecture, oftentimes attributes the found

positive effects to these models (Hysenbelli et al., 2013; Shang & Croson, 2009). Also, in extending pure public goods models, theorists introduced other-regarding preferences for the utility g_i derived from the pure act of contribution, for example, Holländer (1990) who model-theoretically applied the motivational factor of pursuit of social approval, following a social exchange approach. Relevant research on public goods confirms this influence of social approval, too. For instance, a field experiment on voluntary donations (DellaVigna et al., 2012) showed that the proportion of doors opened to solicitors decreases when the group of solicitees was preinformed (vs. not informed) about the upcoming visit. This decrease was interpreted as avoidance of face-to-face interaction and thus avoidance of social pressure, which in turn reflects one motivational source of social approval steaming from avoidance of social sanctions and exclusion. Conceptually, social sanctions for norm-deviant behavior are associated with injunctive, but not descriptive norms (Lapinski & Rimal, 2005). However, these conceptual clearly definable norm types are only hard to separate when processed and become blurred in cognition, as bidirectional associations are regularly made (Eriksson et al., 2015). Thus, one might expect that indicating the descriptive norm by displaying the user community's flagging behavior, also makes injunctive norms more salient. Consequently, given the perceived social pressure or the expectation of sanctions for noncompliant behavior, the display of higher shares of flagging users should lead to norm-compliant behavior, which translates into higher probabilities to flag, and vice versa.

Resolving the Conflict

One of the fundamentals of decision-making is to pursue the decision that increases benefits while minimizing costs (Shapiro & Grafton, 2020). The theories discussed above are based on such a cost-benefit calculus, although they focus on different benefits or costs, which could either be attributed more to cognitive biases or more to an idealized homo economicus. Considered in isolation, they lead to different behavioral assumptions, and have sometimes been empirically confirmed and sometimes not. In the following, it is argued, that this divergence is less caused by fundamentally flawed theories, but rather by undifferentiated reflections that ignore individual differences in the valuation of the public good.

First, only uniting the cited theories and evidence, the utility of contributing (here: flagging) to a public good (here: HSP), can be described as a utility function incorporating the prevalent social norm SN, the aggregated level of the public good G, feelings of warm glow g, the mere size of contribution C, the relative impact of the contribution IC, and the contribution's effort E. Furthermore, theories assume feelings of warm glow g to be positively related to the contribution's mere size C (impure public goods models) and its relative impact IC (impact philanthropy theory). Moreover, C and IC are supposed to be conditioned by the contribution of others to G in that either the marginal value of i's contribution (pure public goods models) or its relative impact drops (impact philanthropy theory) the more others contribute. Now suppose that this contribution of others is indicated by DNI (here: displayed shares of flagging users; SFU) from which the prevailing SN can be derived. The behavior of others can then be approximately equated with SN and the utility function can be written as follows:

$$U_i = U_i(SN, G, g(C, IC), C(SN), IC(SN), E)$$

Next, assume interindividual differences in the valuation of a public good. As an argument, Kahneman and Knetsch (1992) point out that not every public good is of equal relevance, among other things because of different expectations to enjoy it personally (use value). To be sure, this argument is applied to distinguish the relevance of different public goods. However, the use value can also differ across individuals, especially in the case of HSP, not least because of different social media usage patterns and intensity (Kepios Pte. Ltd., 2022), resulting in varying exposure to HS. G is therefore supplemented by an additional valuation factor, denoted γ_i. Furthermore, if the valuation of G is variable, so should the relevance of the relative impact IC on G as well as the marginal values derived from the contribution C to G be conditioned by its subjective value. For simplicity, assuming a perfect positive correlation of the valuation factors, γ_i can be applied equally to G, C, and IC. It follows that the lower the valuation of HSP, the less influential are not only the weighted components G, C, and IC, but also g, as it is determined by C, and IC. Thus, individuals placing extreme low value on HSP, are supposed to primarily include SN and E. In case of flagging, E can be assumed as an externally specified factor given by the platform's flagging interface and process. In addition, it is unaffected by the level of SFU as the formula specifies. Hence, SN becomes the only variable determining behavior. Recalling that SN is derived from DNI given by SFU,

different displayed levels of SFU should result in a user behavior pattern oriented to these SFU levels. Briefly, the higher the displayed level of SFU the higher the likelihood that the next user will also flag it.

For individuals placing an extreme high value on HSP, further aspects are additionally considered in the decision-making process: the value of the public good itself $\gamma_i G$, the feelings of warm glow $g(C, IC)$, the marginal value $\gamma_i C(SN)$ as well as the relative impact $\gamma_i IC(SN)$ of the contribution. As the value of the public good itself $\gamma_i G$ is unaffected by *SN,* and thus SFU, no alterations are assumed across SFU levels in that aspect. But as the formula specifies, the latter three components are conditioned by SN, hence SFU. Replicating the (imperfect) crowding-out hypothesis from (im)pure public goods models, C should decrease when higher levels of SFU are displayed due to diminishing marginal values, and vice versa. Similarly, IC should vary due to declining relative impacts as stated in impact philanthropy theory. Lastly, also feelings of warm glow g should vary like C and IC due to dependence.[3] As a result, users' flagging behavior should be diametrically opposed to the displayed levels of SFU. Simply put, the higher the displayed level of SFU the lower the likelihood that the next user will also flag it.

In the expectation that the extreme values used for modulation do not or only very rarely occur empirically, following trend processes are assumed to be observed at the behavioral level:

> **H1$_{\text{behavioral}}$**: *The displayed share of flagging users affects the individual flagging behavior but depending on the users' valuation of hate speech prevention in opposite directions.*
>
> > **H1a$_{\text{behavioral}}$**: *For users with low valuation of hate speech prevention, individual flagging behavior increases as displayed shares of flagging users increase.*
> >
> > **H1b$_{\text{behavioral}}$**: *For users with high valuation of hate speech prevention, individual flagging behavior decreases as displayed shares of flagging users increase.*

To further test the reasoning behind these behavioral main effects, an additional neural investigation is executed. Neuroscientific methods are advantageous in that they allow decision-making processes to be tested and observed directly in the organism (Plassmann et al., 2015). Thereby, they also minimize the risk of response biases that can occur with self-reported data in the studied context, e.g., socially desirability (Holtgraves, 2004), or sorts of recall bias (Colombo et al., 2020). To this end, the process assumptions are translated into hypotheses about correlated neural activation structures in the following.

In the hypotheses derivation of the behavioral main effects, it is argued that the decision to (not) flag a comment is based on a utility function comprising multiple components. The calculation of the assumed utility function describes a process of goal-directed decision-making, in which the relationships between response and outcome are evaluated against current goals in order to choose the response with the best possible outcome (Gęsiarz & Crockett, 2015). Generally, decision-making is associated with brain areas that evolved later in evolution, enabling higher cognition and complex behavior in primates (Carlén, 2017). These areas are mainly located in the frontal part of the brain – the prefrontal cortex (PFC). Within the PFC, goal-directed decision-making is primarily associated with lateral brain areas of the PFC (lPFC) (Gęsiarz & Crockett, 2015). The lPFC is located sidewards at the front end of the PFC (for precise localization see Carlén, 2017). It is associated i.a. with processes of planning, and working memory, where accessible information from long-term memory is organized and updated against current situational inputs (Gęsiarz & Crockett, 2015). Hence, if activation is identified in the lPFC, it is likely that more aspects need to be kept in working memory to be organized and updated to evaluate the behavioral response. Consequently, when information is displayed in the form of SFU to comments, it is assumed that this information is included in the evaluation process. Thus, initially, higher activity in the lPFC can be generally expected, independent of the level of the displayed SFU.

> **H2$_{\text{neural}}$:** *The lPFC is more activated when shares of flagging users are displayed to the comment compared to the display of the comment only.*

Moreover, in deriving the behavioral hypotheses, the argument proceeds that two components of the utility function are directly and negatively determined by the behavior of others: the contribution's marginal utilities and its relative impact. Both, in turn, should ultimately positively influence the feelings of warm glow. Neural areas that could be associated with such emotional value are brain regions in the

[3] One could certainly argue that DNI might not only influence behavior, but also opinions on the public good. However, the potential influence of SFU on HSP was controlled for in study 1 and confirmed to be not present.

medial PFC (mPFC), located in the middle at the front end of the cortex (for precise localization see Carlén, 2017). Areas in the mPFC are considered a crucial part of the valuation system in the brain (Bartra et al., 2013). In the context of SFU display the mPFC could be associated with the emotional value of flagging (warm glow), as it has already been demonstrated in the voluntary private provision of public goods (i.e., charity, Hare et al., 2010). Given that the behavior of others is actively induced by displayed SFU levels and presumably negatively influences positive affect in form of feelings of warm glow via diminishing contribution's marginal utility and relative impact, it can be expected that the display of low SFU levels should result in higher mPFC activation. Conversely, the display of high SFU levels should cause no neural response.

H3a$_{neural}$: *The mPFC is more activated when low shares of flagging users are displayed to the comment compared to the display of the comment only.*

H3b$_{neural}$: *The mPFC is not activated when high shares of flagging users are displayed to the comment compared to the display of the comment only.*

Lastly, the utility function suggests that the weight of the aforementioned components – the contribution's relative impact and marginal utility that ultimately manifest in feelings of warm glow – depends on the valuation of HSP. More precisely it is assumed that the higher the valuation of HSP, the more relevant are these components for utility calculus, and vice versa. Therefore, users with a high (vs. low) valuation of HSP are expected to have greater activation in the mPFC when SFU is displayed.

H4$_{neural}$: *When shares of flagging users are displayed to the comment, the mPFC is more activated among individuals with high (vs. low) valuation of hate speech prevention.*

Study Overview and Stimuli Generation

The study has been approved by the university's ethics committee. Since both the neural and behavioral part of the study required the presentation of offensive comments as potentially upsetting and triggering stimuli, a content warning was given in the beginning, for ethical and responsibility reasons. The assumed countervailing trend effects were in the focus of the first behavioral part, which examined the main effects of different displayed shares of flagging users (SFU) on the likelihood of flagging and interaction effects with different valuations of hate speech prevention (HSP). The second part provides further insight into the procedural assumptions underlying the behavioral hypotheses by revealing neural activation structures. In both studies, three SFU levels, displaying shares of 30%, 60%, or 90% of users having flagged the comment, were intentionally chosen to (1.) cover minority, small, and clear majority shares and (2.) achieve equal percentage point intervals to allow for linear trend analysis (Thompson, 2006).

Each study part first required suitable stimuli to be created. One comment would have sufficed for the between-subject design of the behavioral study. Though, the within-subject design of the neural study necessitated repeated measurements per SFU condition to resolve the neural response, to overcome physiological confounds in signals, and to increase the reliability of brain activity measurements (Yücel et al., 2021). An initial set of 165 comments was produced by combining frequently used hate expressions (Silva et al., 2016) and hate categories (sourced from https://hatebase.org) in social networks. In a pretest, 173 German natives (M_{age}=42.71(12.24), 59.4% male) rated 30 randomly chosen comments on *message credibility* (Appelman & Sundar, 2016), *offensiveness* (attributes used: insulting, discriminatory, offensive), and *comprehensibility*. Thereby, 60 comments were identified that were perceived as equally credible, were understandable by a broad public, and allowed for different offense ratings, reflecting cases of potential hate speech that are not necessarily prohibited by law but rather require societal responses.

Behavioral Study 1

Study Design

Sample

Participants with at least one active profile on social media platforms were recruited from the crowd working platform Clickworker (https://www.clickworker.de). To ensure high data quality, two methodologically different attention checks and a control for social desirability bias were applied

(Aguinis et al., 2021). First, an item was administered where only those participants who abstained from answering were allowed to continue the survey. Second, the time spent reading the task instruction (150 words) was measured. The threshold of at least 15 seconds[4] excluded an additional 97 participants from those 361 who passed the first attention check. Finally, excluding the participants with the most extreme scores on the gamma factor scale of socially desirable response behavior (Kemper et al., 2012) reduced the sample by an additional 25 participants, who had extreme tendencies to deny their negative qualities and to exaggerate their positive qualities. This results in an analysis sample[5] of 239 participants (M_{age}=40.76(12.36), 61.9% male).

Experimental Setup and Questionnaire

To test how different levels of share of flagging users (SFU) affect an individual's flagging behavior, a one-factor between-subjects experiment was conducted. Therefore, a flagging decision on an ethnicity-based offensive comment selected from the pretest set was integrated into an online survey, covered as a study on social media behavior. In the task instructions, it was explained that a comment from a social media platform would be displayed on the next page, without explicitly mentioning that this comment was accompanied by supplemental information about the flagging behavior of others. This was to ensure that participants are not additionally encouraged to draw preoccupations about the behavior of others due to the mere research design. It was further explained that upon displaying the comment, participants would be asked whether they like to flag the comment, and therewith report it to the platform operator. Thereby, it was emphasized that the decision to flag will be anonymously reported to the platform operator and would cause additional effort, as the decision had to be refined afterward. This should attempt to replicate the effort of flagging and at the same time avoid a hypothetical generosity bias (Clark, 2020).

Participants were randomly assigned to one SFU condition (SFU-30%, -60%, -90%) displaying that "about X% of the readers of the comment flagged it" or the control condition with no such information. Those who decided to flag were then asked to state their main reason for flagging from a given choice set and to elaborate further on this in a free text field. For all non-flaggers these questions were omitted. Thereafter, participants were asked to rate different scales on 7-point (exception: harm of hate speech on 5-point) Likert-type scales ranging from "strongly disagree" to "strongly agree". They were asked to indicate their *offensiveness rating* (as in pretest), *decision satisfaction* (single-item statement "Overall, I am satisfied with my decision"), experienced *warm glow* (Ferguson & Flynn, 2016), and *anticipated guilt* (Theotokis & Manganari, 2015). Thereafter, trait-based covariates like *reactance to compliance* and *resisting the influence of others* (Hong & Faedda, 1996), *racism* (Alba & Johnson, 2000), as well as the focal variable *harm of hate speech* were queried in random order. *Harm of hate speech* was assessed with the eponymous 16-item scale (Cowan et al., 2002), six items of which were selected for feasibility reasons after comparative analysis with the full scale during pretesting. The scale served as a social-desirability-avoiding proxy for participants' valuation of hate speech prevention (HSP). *Personal experience* with hate speech, the *gamma factor of socially desirable response behavior* (Kemper et al., 2012), and sociodemographic variables were finally polled for sample description and as potential covariates.

Data Preparation and Analysis

As manipulation check, deviations between the conditions were examined in terms of both the comment's offensiveness and credibility evaluation. No significant differences were found for offensiveness, $F(3, 235)=1.006$, $p=.559$, with means ratings ranging from 5.24 to 5.54 ($1.05 \leq SD_{offensive} \leq 1.97$). Similarly, the credibility was rated equally across conditions, $F(3, 235)=0.538$, $p=.657$. For both, no differences were also found between the conditions when testing for those with high and low valuation of HSP separately.

According to hypothesis H1 and its sub-hypotheses H1a and H1b, trend patterns are assumed along the SFU conditions, precisely a positive linear trend for individuals with low valuations of HSP and a negative linear trend for individuals with high valuations of HSP. To test these opposing trends, the SFU conditions were first combined into a single variable coded as trend contrast, indicating a linear positive trend from SFU-30% (-1), via SFU-60% (0), to SFU-90% (1). To test for the hypothesized opposing trends

[4] As Clickworkers are assumed to be proficient in study participation, the fastest normal reading speed (600 standard words/minute) in German (Carver, 1990) was set as maximum.

[5] Equivalence across conditions was checked. Tests revealed no significant differences for latent measures and sociodemographics.

according to the valuation of HSP, an interaction term of SFU_{linear} and HSP was calculated. Important to note, HSP has been applied in the product term as a *z*-standardized variable. The trend variable, SFU_{linear}, the interaction SFU_{linear} × HSP, as well as HSP as direct effect, to separate the direct effect from the hypothesized moderation effect, were then applied in logistic regression analysis on the dependent variable of flagging decision coded as a binary dummy (1: yes; 0: no). Via a stepwise approach, including more and more potential covariates[7], it was further controlled for the focal effects' robustness. For visualization and further deep-dive, the conditional effects of the focal trend contrast SFU_{linear} were finally calculated at the values -1 *SD*, 0, and +1 *SD* of the *z*-standardized variable HSP.

Results

The logistic regression reveals a significant direct effect of HSP, b=1.090, p<.001, and a nonsignificant direct effect of SFU_{linear}, b=0.018, p=.929. Importantly, the focal interaction of SFU_{linear} × HSP indicates a negative moderating effect, confirming the hypothesized opposing effect of SFU on flagging behavior as a function of HSP, b=-0.665, p=.016. Further controlling for sociodemographics and personal experience with HS, the interaction term stays robust in effect size with a significance level increasing, b=-0.749, p=.009. Additionally, controlling for the latent control variables as stated in the questionnaire section (full model), the effect size slightly rises at the same significance level, b=-1.228, p=.006.

The negative sign of the interaction term's coefficient already indicates the hypothesized behavioral patterns, that is, a diminishing (increasing) likelihood to flag for HSP_{high} (HSP_{low}). These are further confirmed by conditional effects calculated on the full model. The conditional effect of SFU_{linear} at HSP_{low} (-1 *SD*) proves positive and significant, indicating a norm-compliant behavioral response pattern, b=0.887, p=.041. The conditional effect of SFU_{linear} at $HSP_{average}$ (level 0) is not significant, indicating no effect of SFU_{linear}, b=-0.194, p=.570. In turn, the conditional effect of SFU_{linear} at HSP_{high} (+1 *SD*) is found negative and significant, indicating a norm-deviant response pattern, b=-1.248, p=.035.

Neural Study 2

Study Design

Sample

Individuals were recruited from the university participant pool. Again, they should have at least one active profile on social media platforms to meet the inclusion criteria of study 1. In addition, all participants met the requirements for neural measurement and were, among others, not pregnant or breastfeeding, not taking medication, and did not suffer from severe mental or neurological disorders. Replicating the data cleansing conducted in study 1, participants with extreme degrees of social desirability bias were excluded. This left the final analysis sample of 28 records (M_{age}=27.07(9.63), 64.3% female), which is above the average sample size in neuroscientific studies (n≈18; Lieberman et al., 2009). It also exceeds the common thresholds of 20, or 24, respectively, at which 80% power is assumed (Murphy & Garavan, 2004).

Experimental Procedure

An event-related functional near-infrared spectroscopy (fNIRS) experiment was conducted using a within-subject design with the share of flagging users (SFU) as a factor with three levels (30%, 60%, 90%). First, participants were informed about the experimental procedure as well as the attachment and operation of the fNIRS device both in writing and verbally. Once any ambiguities were clarified, informed consent was obtained according to the Declaration of Helsinki. Thereafter, the participants were asked to sit in a comfortable position in front of the computer, which would present the stimulus material. The fNIRS device was then attached to the participant's forehead. Once the calibration phase was successfully completed, a further darkening hood was placed over the fNIRS device to minimize measurement artifacts from external light sources. After this, the experimenter left the room, still being callable auditorily upon request. Participants could then start the task at their convenience by pressing the Enter key.

[7] Latent constructs used here and during the neural study were first assessed for applicability, using Cronbach's alpha to measure internal consistency, univariate exploratory factor analysis to test one-dimensionality, and structural equation modeling for the overall test of the measurement model. Constructs were judged to be reliable and discriminantly valid.

Out of the 60 qualified comments, 12 comments each were randomly assigned to the three SFU conditions, resulting in twelve trials per condition. The trials were presented in a fully randomized order in the experimental task. Each trial began with the display of a comment without additional information (comment-only; 7 s). After an interstimulus interval (ISI) in form of a jitter was displayed with a random duration (1-5 s), the same comment was displayed again with supplemental SFU (comment-SFU; 7 s). Since the order of the comments and the order of the SFU levels were completely randomized, SFU levels were not predictable for the participants by any rule of thumb or other schemes. After another ISI (1-5 s), the question of whether the participants wanted to flag the comment or not was displayed without answer options (2 s). Hereafter, the answers (flag comment/do nothing) were presented below the question for another two seconds and participants could indicate their choice by pressing the appropriately designated keys. If participants decided to flag the comment, they were pleased to answer two follow-up questions to increase the effort of flagging. First, they had to choose between six possible options (racism, violence, discrimination, incitement, insult, other reason) to state their main reason for flagging. Second, they had to indicate which group the comment was directed against (nationality, gender, sexuality, other). As before, answers could be entered using appropriately assigned keys on the keyboard. In this phase, there were no time restrictions. Subsequently, a new trial started after an intertrial interval (ITI) in form of a jitter (2-6 s). This ITI also describes the trial progress in case of the decision to not flag the comment.

The primary intent of ISI and ITI is to increase the density of fNIRS data points (temporal resolution) to achieve greater measurement accuracy (Watanabe et al., 2013). However, as participants' choices affected the task's duration, both ISI and ITI further anonymized their choices by adding additional variance to task duration. Participants were told that their decisions could not be deduced by the task duration and that they should feel and act freely as they would normally do in social media. After experimental task completion, the device was taken off and a questionnaire was given with measures used in study 1. The total participation time, including experimental task and questionnaire, was about one hour on average (max. 1 ½ hours). Participants received a compensation of €30 for participation, meeting the national minimum wage measured against the maximum participation time.

Neural Measurement

To replicate the results of study 1 and disclose the underlying neural processes, the non-invasive mobile neuroimaging method of fNIRS was used, applying the most commonly used continuous wave technique (Scholkmann et al., 2014). In this method, near-infrared light sources with different wavelengths of 760 and 850 nm penetrate the human tissue. As oxygenated and deoxygenated blood differently absorb these wavelengths, measuring the reflected wavelengths allows analyzing the levels of hemoglobin concentration (HbO and HbR) which is associated with neural activity across different brain regions (Ferrari & Quaresima, 2012). Thereby, increasing HbO and decreasing HbR levels are associated with increased neural activation, and vice versa (Quaresima & Ferrari, 2019). This method is superior to alternative methods like functional magnetic resonance imaging (fMRI) or electroencephalography (EEG) in terms of user-friendliness, robustness against movement artifacts, and feasibility of the research design (Pinti et al., 2020). fNIRS is suitable for this study as the expected neural activity is located in cortical brain regions. Moreover, as two types of neural signals can be quantified in parallel with fNIRS, the HbO and HbR level, the measured signals are stronger and can be evaluated as robust, requiring fewer repetitions in the experimental paradigm compared to EEG (Pinti et al., 2020; Quaresima & Ferrari, 2019). These fewer repetitions are also advantageous to avoid fatigue effects and emotional blunting towards the offensive comments. Moreover, considering the risk of a potentially high social desirability bias, the laboratory situation should create an atmosphere in which participants feel unobserved and as anonymous as possible. In typical fMRI experiments, such a situation is difficult to establish as the experimenter must have visual contact with the subject for technical and safety reasons. In turn, the use of mobile fNIRS allows the experimenter to leave the situation and have no visual contact while the task is performed by participants. This reduces the (perceived) social presence, creating the desired situation. In addition, other measures were taken to increase perceived anonymity in the experimental paradigm design (i.e., ISI/ITI variance). Besides these research design aspects, fNIRS is not only more robust against motion or external electrical artifacts than EEG, but also reveals brain activity patterns comparable to fMRI results (Pinti et al., 2020), making fNIRS preferable for this research study.

The applied fNIRS device requires participants to wear a fitted headband. It features 22 channels arising from 8 light source emitters and 7 long-distance light detectors (average distance 30 mm) at a 7.81 Hz

sampling frequency. Therewith, it covers brain regions of interest within the PFC which are demonstrated to successfully measure functional neural activations, also in NeuroIS research (i.e., Gefen et al., 2014; Nissen & Gier, 2021). The 22 channels were categorized accordingly into medial (3, 4, 5, 10, 13, 18, 19, 20) and lateral PFC (1, 2, 6, 7, 9, 14, 16, 17, 21, 22). For the remaining channels (8, 11, 12, 15), precise categorization was not viable (see Figure 1). To ensure consistent neural data retrieval, the craniometric point of the nasal bone was used as a reference during headband application (Krampe et al., 2018). Potential signal disturbances were checked in an a-priori calibrating process. In case of poor signal quality the fit of the critical optodes was inspected and interfering hair was pushed aside with a cotton bud. As mentioned, fNIRS is relatively robust to movement and external electrical artifacts, but however is sensitive to external light sources. Therefore, a darkening hood was placed over the fNIRS device after successful calibration. The NIRS-Star software package (v14.2) was used for recording and calibration.

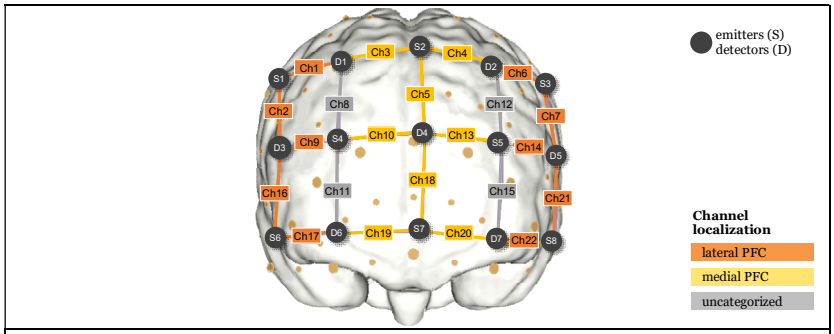

Figure 1. Optode Montage and Channel Locations

Data Preprocessing and Analysis

Behavioral Level

Split plot repeated measures ANOVA (SP-RM-ANOVA) was used to test and compare flagging patterns across SFU conditions (within-subjects factor) between HSP_{high} and HSP_{low} groups (between-subjects factor). The two levels of hate speech prevention (HSP) were calculated from the post-task questionnaire responses on the harm of hate speech scale, using the sample median as cutting value. As each participant had to decide on flagging 12 times per SFU condition, the proportion of flagged comments (propFlag) was calculated per condition relative to the total number of active decisions. By choosing active decisions as the baseline, it was possible to directly control for missing decisions. Missing decisions were possible since time limits for the decision to flag or not were set in the experimental paradigm for reasons of data accuracy. In addition, questionnaire scores served as covariates replicating the analysis strategy of study 1. Both data from the post-task questionnaire and the fNIRS task were matched using pseudonymized IDs.

Prerequisites of SP-RM-ANOVA were confirmed by statistical analyses. Residuals were normally distributed in each of the six cells, as implied by nonsignificant Shapiro-Wilk tests. Mauchly's test of sphericity indicates equality of variances of the differences between the SFU conditions, W=0.974, $x^2(2)$=0.617, p=.735. Box's M of 14.243, $F(2, 4603.01)$=2.072, p=.053, indicates equality of the covariance matrices of propFlag across the cells formed by HSP groups, albeit narrowly. Also homogeneous variances of propFlag could be assumed between HSP groups as assessed using Levene's test per condition, $F_{30\%}(1, 26)$=0.513, p=.480, $F_{60\%}(1, 26)$=1.455, p=.239, $F_{90\%}(1, 26)$=0.081, p=.778. Three outliers were detected. However, since measures of outliers' influence on the estimate suggest a moderate impact, and a test calculation without the outlier-producing cases yields comparable results in effect size and p values, the analysis of all cases will be reported to increase power and comparability to neural results.

Neural Level

Raw fNIRS data got preprocessed using the NIRS AnalyzIR toolbox in MATLAB (Santosa et al., 2018). Thereby, the raw fNIRS signal of 7.81 Hz was first downsampled to 4 Hz to control for the high autocorrelation in the fNIRS signal (Huppert, 2016). Furthermore, fNIRS data time series were smoothed and artifacts (e.g., heart rate or drifts in the optical signal) were removed. Baseline correction was

performed to remove statistical outliers. For motion correction a spatial principal component filter was implemented using principal component analysis (Zhang et al., 2005). Thereafter, the optical density was calculated (Huppert, 2016). In a last step, the modified Beer-Lambert law with a partial pathlength factor of 0.1 was applied (Kocsis et al., 2006) to convert the raw optical density signals into hemoglobin values.

A general linear model (GLM) was set up per participant (within-subjects level) to model neural activity during the experimental task. The time periods of comment-only, comment-SFU, and flagging decision were modeled separately for the experimental conditions. The regressors of the SFU conditions during the comment-only and the comment-SFU period were defined as the six regressors of interest. The AR-IRLS algorithm was used for GLM estimation, allowing residual motion artifacts in the fNIRS signal to be filtered out (Barker et al., 2013). Each time course was convolved by the canonical hemodynamic response function. For group analysis, a mixed-effects model was calculated using SFU conditions per HSP group as fixed effects and subjects as random effects. To evaluate the hypotheses, neural activation was analyzed, computing different second-level group contrasts. To test H2 and H3a/b, three contrasts were taken between the comment-only and the comment-SFU period for each SFU condition. To test H4, the SFU conditions during the comment-SFU period were contrasted between the two HSP groups. Two criteria were applied to determine significance and robustness within the fNIRS channels, avoiding false positive results. First, significant channels were identified for HbO or HbR signals using the family-wise error corrected threshold of $q<.05$. Second, significant channels were considered as robust only if the parallel HbO or HbR levels in the alike channels suggested the same neural, not necessarily significant, activation. If the same direction (both HbO and HbR increase or decrease) was identified, the signal might potentially have been a false positive and was not further considered. To visualize the neural activation, the resulting *t* contrast activation maps were rendered into a standardized brain template.

Results

Behavioral Level

Results of SP-RM-ANOVA reveal a significant main effect of SFU on user flagging behavior, $F(2, 48)=3.852$, $p=.028$, with a partial eta-squared (η_p^2) of 0.138 indicating a large effect. The between-subject factor of HSP group was only marginally significant, $F(1, 24)=3.192$, $p=.087$. Most importantly, the interaction SFU × HSP group indicates a significant large effect, $F(2, 48)=4.718$, $p=.013$, $\eta_p^2=0.164$.[8] Moreover, within-subject contrasts indicate a significant overall linear trend pattern for SFU, $F(1, 24)=7.739$, $p=.010$, $\eta_p^2=.244$.[9] The separate plot of curves for each HSP group show the expected opposite trends (see Figure 2), replicating the flagging patterns found in study 1.

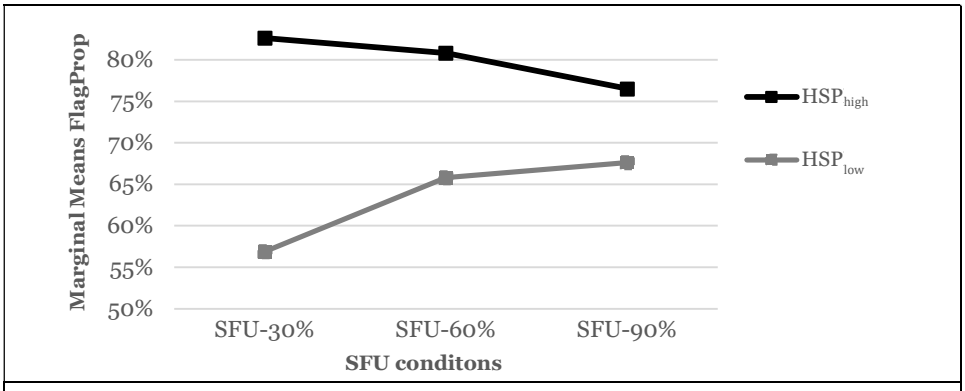

Figure 2. Marginal Means of Flagging Proportion

Neural Level

To test H2 and H3a/b, the results from contrasting the comment-SFU (vs. comment-only) period for each

[8] Results for outlier-corrected sample with n=26: SFU ($F(2, 44)=3.493$, $p=.039$, $\eta_p^2=0.137$); HSP group ($F(1, 22)=8.655$, $p=.008$, $\eta_p^2=0.282$); SFU × HSP group ($F(2, 44)=5.236$, $p=.009$, $\eta_p^2= 0.192$)

[9] Outlier-corrected sample: SFU ($F(1, 22)=6.696$, $p=.017$, $\eta_p^2=0.233$)

SFU (see Figure 3; top row) are presented first. To begin with lPFC, significant and robust[10] increased activation can be identified in all SFU conditions confirming H2. For SFU-30%, increased bilateral activation in the lPFC can be detected in channels located in the left (6, 14, 22) and right hemisphere (9). Bilateral lPFC activation can also be found for SFU-60% as indicated by accordingly categorized left (14) and right channels (1, 2, 9, 17). Similarly, for SFU-90%, an increased activation can be identified for the left (14) and right lPFC (2, 9). Contrasting the comment-SFU (vs. comment-only) period also indicates neural activation in the mPFC. For SFU-30%, the channels 4, 5, 10, 18, and 22 show increased mPFC activation. For SFU-60%, an increased mPFC activation can be identified in channels 10, 19, and 20. In turn, for SFU-90%, two channels (18, 19) indicate decreased activation in the mPFC. Hence, H3a can be confirmed, but H3b has to be rejected as non-significant activation was expected. However, the opposing activation pattern in the form of decreased activation indicates somehow different brain mechanisms involved in SFU-90% compared to SFU-30% and SFU-60%. For completeness, increased activation was found in the two uncategorized channels 11 and 15 for both SFU-60% and SFU-90%.

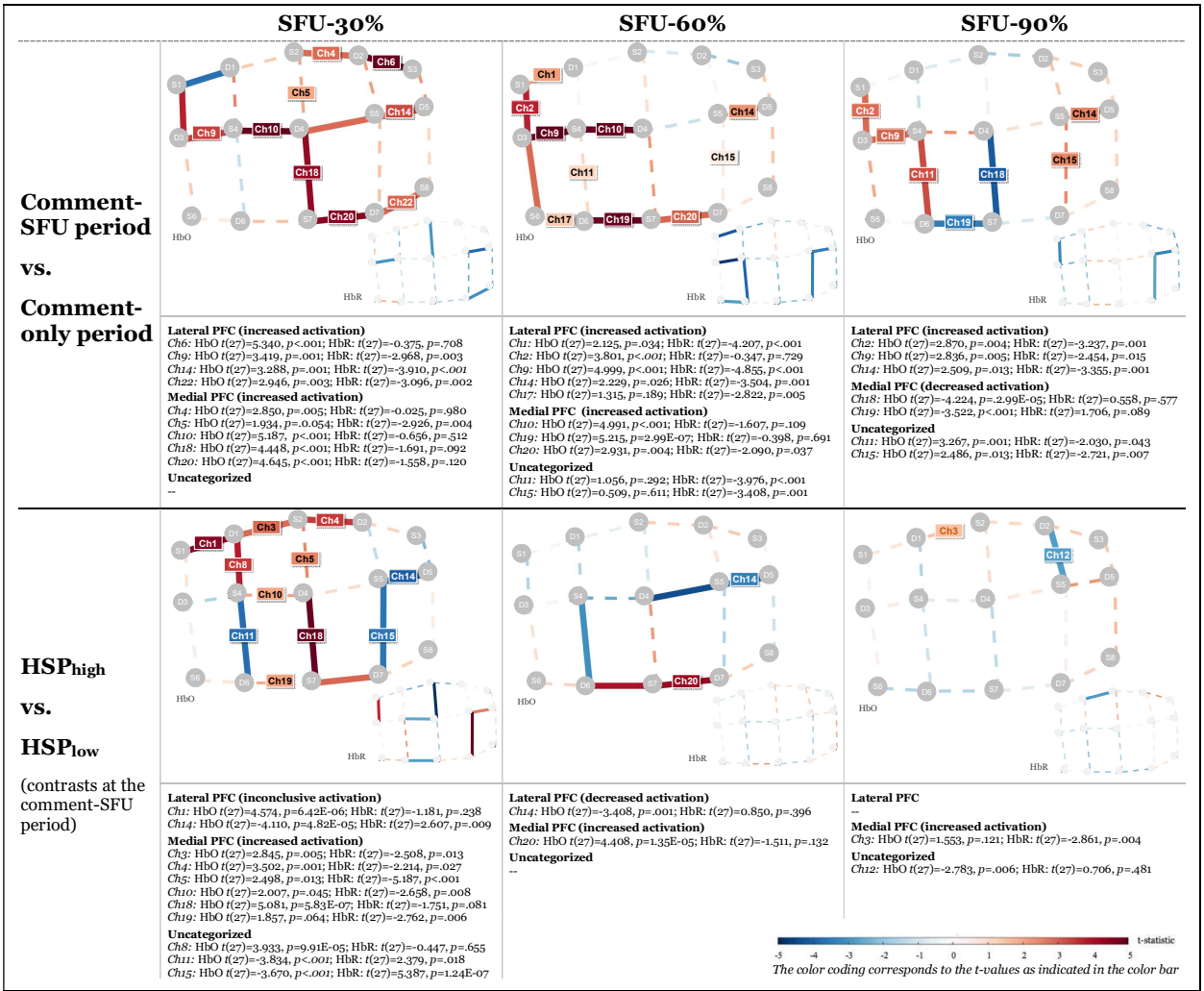

Figure 3. Significant Channels at the Contrasts for SFU-30%, SFU-60%, and SFU-90%

Contrasting HSP_{high} against HSP_{low} in the comment-SFU period (see Figure 3; bottom row), further evaluates the hypothesized groupwise effects of SFU conditions in the mPFC. The results indicate that H4 can be supported since a constantly higher activation for HSP_{high} (vs. HSP_{low}) could be identified across all

[10] Please note that only significant and robust channels are reported hereafter (for explanation see Data Preprocessing and Analysis)

SFU conditions. For SFU-30%, a strong neural increase for HSP_{high} compared to HSP_{low} was identified in the mPFC in channels 3, 4, 5, 10, 18, and 19. For the other two conditions, one channel each shows increased neural mPFC activation for HSP_{high} compared to HSP_{low}, specifically channel 20 for SFU-60% and channel 3 for SFU-90%. For completeness, additional activation was found in the lateral and uncategorized regions. For SFU-30%, lPFC activation was not conclusive with increased activation in channel 1 but decreased activation in channel 14. Additionally, inconclusive activation was found in uncategorized regions with increased activation in channels 11 and 15 but decreased activation in channel 8. For SFU-60%, a decreased activation was found in the lPFC (14). For SFU-90%, no activation could be identified in the lPFC, though the uncategorized channel 12 showed decreased activation.

Discussion

Theoretical Contribution

The twice-collected behavioral data largely support the focal hypothesis of divergent behavioral responses to descriptive normative information (DNI) depending on the valuation of a public good. This was demonstrated by the use case of hate speech prevention (HSP) with DNI being induced by the display of shares of flagging users (SFU). It was found that those to whom HSP seemed comparatively less relevant tended to behave in accordance with the descriptive norm, but not those who rated HSP as of greater relevance. Indeed, they behaved exactly opposite to the norm.

The neural results support the basic neural hypothesis that DNI cause higher activations in the lateral prefrontal cortex (PFC), indicating increased processing of cost-benefit trade-offs that urge higher-order cognitive efforts, calculations of strategic motivations, and working memory (Figner et al., 2010; Heinze et al., 2014). More interesting, the neural activation patterns in the medial PFC also support the assumption that those components associated with feelings of warm glow – the contribution's impact, its marginal value, and the warm glow in the form of immanent moral satisfaction itself – recede into the background for users with lower valuations of HSP (HSP_{low}) compared to those with higher alike valuations (HSP_{high}). However, whether all, some, or only one of these components affected the found differences in medial PFC activation between HSP_{low} and HSP_{high}, cannot be differentiated from the neural data.

In summary, the results give reason to be confident that the proposed theoretical solution could help explain the (apparently) contradictory behavioral patterns found in experimental research. The opposing behavioral patterns suggest that different overall behavioral outcomes are possible in different samples. This outcome should not only depend on the overall average valuation of the public good under study but much more on the proportions of HSP_{low} and $HSP_{high,}$ as well as the trajectories of the opposing curves, which could sometimes resemble more, i.e., a (reversed) straight linear, exponential, or J-shaped form.

Managerial Contribution

In contrast to offline discussions, there have hardly been any means in social media of expressing disagreement through small but publicly noticeable reactions. The use of ignorance as a form of deliberate non-communication or other effortless options such as rejecting gestures to voice dissent do not exist on platforms that only allow positive or inconclusive emoji reaction buttons[12]. While public agreement is seemingly effortless, public disagreement can only be stated through more laborious participation in discussions in the comment sections. Moreover, public comments and even emoji reactions are often not anonymous. Users who actually disagree with posted content and perceive it as hate speech (HS) thus face the adverse choice of voicing their dissent through tedious, non-anonymous commenting on content or reporting content to the platform's moderators in an anonymous but nonpublic, thus non-discoursive manner. As a result, the broad masses presumably prefer to remain silent, which is problematic, especially considering HS detection. Not only is it impossible to interpret the opinion of the silent crowd, but also 'bottom level HS' comments remain on the platform as manifest and seemingly collectively accepted statements without any (strong) counter-speech. The socio-technical approach of flagging combined with

[12] On social media platforms (≥ 1bn active users; except instant messengers; Kepios Pte. Ltd., 2022), only on YouTube users can clearly express disagreement with a "dislike" button. Facebook offers six emojis and a "like" button, but the emojis that appear to express negative reactions (astonishment, sadness, and anger) are ambiguous, as their use can also merely express agreement with the emotion of the initial comment. On Instagram and TikTok, it is only possible to express agreement using a heart emoji.

the transparent display of the user community's behavior could hereby provide a tool to overcome the silence of the crowd by combining the advantages of anonymity and discoursive public participation.

This socio-technical approach could also prove practical for companies for other reasons. First, it offers companies an opportunity to respond to the demand for more independent, democratic decision-making processes. Second, it also allows companies to take greater responsibility for their employees and contract labors involved in HS detection. The shift of responsibility for the ultimate decision on HS classification could relieve them not only physically but also psychologically by reducing the intensity of their involvement with HS, reducing the accountability for potentially ambiguous decisions, and finally, minimizing the workload itself (Newton, 2019). Third, the inclusion of the user community might also increase the accuracy of the analysis, given that content moderators and also (supportive) automatic HS detection systems face a myriad of different languages, language styles, and dialects for which sufficient skills or training data are not (yet) equally available (Canales, 2021; MacAvaney et al., 2019).

It should be borne in mind, however, that displaying SFU on every comment would most likely dilute the signaling effect that content is worthy of and in need of discussion. Rather, the display of SFU should be more targeted. One approach could be to first set a threshold based on the users' nonpublic and presumably more hesitant flagging behavior. Only when this threshold is reached will the SFU be displayed and the comment's HS potential gets discussed in the public sphere. From this point on, the behavior patterns identified in this study are assumed to emerge and the displayed SFU needs to be recalculated with every new user engaging in the discussion. As a result, the displayed SFU dynamically changes and, at some point, transitions to a stable state at which the probability of the next user's flagging and the recalculated displayed SFU level coincide. This stable final SFU level which has developed through a community-based transparent democratic discourse could then serve as the basis for further comment processing. While SFU levels that have settled in a minority range might not necessitate countermeasures, SFU levels that have settled at higher majorities may provide the impetus for the platform to take corrective action, such as comment removal.

However, such final-instance decisions like comment removal, and thus discussions about the legitimacy of content banning, could eventually be bypassed with the display of the stable final SFU level. Since its transparent and easily accessible display already signals that the no longer silent crowd does not share the opinion of the hater, this information could, in the end, be even more valuable for the victims themselves and the social fabric.

Limitations and Future Research

In both studies, the attempt to create an anonymous setting in the scenario itself (anonymous flagging) and in the overall study design cannot completely exclude the influence of social desirability. The relevance of the valuation factor γ_i might thus have been reduced in relation to motives like social approval or avoidance of social sanctions. Nonetheless, study results demonstrate behavioral patterns that are not purely norm driven. At the same time, this also gives rise to further research potentials, such as the investigation of the influence of anonymity on the opposing curves (Zimmermann-Janssen, 2020). In addition, future research could also seek to further elucidate the variability of behavioral responses within HSP_{high} and HSP_{low} groups. For example, research on guess norms (Bartke et al., 2017) could provide promising impetus by suggesting that previously formed expectations about the behavior of others may influence behavioral responses to SFU in that they determine the strength of the found trend effects.

Future research could also add further robustness to the identified behavioral patterns by examining reactions to more finely cut SFU levels. Thereby, it would be possible to identify different progressions and special geometric properties of not only the total behavioral response but also the specific curves for HSP_{high} and HSP_{low}. In addition, simulations could help track the dynamic trajectories and identify resulting stable points by calculating different severity levels of comments and different community compositions in terms of HSP valuation.

For the behavioral part of the neural study, it can be remarked that usually users are not repeatedly confronted with flagging decision situations as offensive, potentially harmful content constitutes only a fraction of all social media content. The first behavioral study more closely approximates this real-world context. However, the applied single decision may have reduced the impact of the effort associated with flagging. While this should not determine flagging patterns since effort was constant across conditions,

it may have potentially led to consistently higher shares of flagging users. However, since the objective was not to examine the absolute flagging shares, but the flagging patterns, this seems less crucial.

For the neural part, further contrasts can be tested beyond the tested main effects. I.e., considering that SFU conditions vary in explicitness of the user community behavior, this might cause different degrees of working memory needed to evaluate the cost-benefit ratio, being associated with lateral PFC activation. The first applied contrast set suggests such an activation structure quite cautiously, but contrasts between SFU conditions, possibly nested with HSP groups, are rather indicated here. Such contrasts, where SFU conditions are directly compared, should also add robustness to the detected activation differences in medial PFC that were so far separately assessed for the SFU conditions (comment-SFU vs. comment-only periods). Lastly, the non-expected deactivation in the medial PFC in the SFU-90% condition for users with high (vs. low) valuations of HSP opens up avenues for further (neural) research and hypotheses.

References

Agerström, J., Carlsson, R., Nicklasson, L., & Guntell, L. (2016). Using descriptive social norms to increase charitable giving: The power of local norms. *Journal of Economic Psychology*, *52*, 147–153.

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management*, *47*(4), 823–837.

Alba, R., & Johnson, M. (2000). Zur Messung aktueller Einstellungsmuster gegenüber Ausländern in Deutschland [Measuring current patterns of attitudes towards foreigners in Germany]. In R. D. Alba, P. Schmidt, & M. Wasmer (Eds.), *Blickpunkt Gesellschaft 5. Deutsche und Ausländer: Freunde, Fremde oder Feinde? Empirische Befunde und theoretische Erklärungen* (pp. 229–254). Westdeutscher Verlag.

Andreoni, J. (1989). Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy*, *97*(6), 1447–1458.

Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism and Mass Communication Quarterly*, *93*(1), 59–79.

Barker, J. W., Aarabi, A., & Huppert, T. J. (2013). Autoregressive model based algorithm for correcting motion and serially correlated errors in fNIRS. *Biomedical Optics Express*, *4*(8), 1366–1379.

Bartke, S., Friedl, A., Gelhaar, F., & Reh, L. (2017). Social comparison nudges—Guessing the norm increases charitable giving. *Economics Letters*, *152*, 73–75.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, *76*, 412–427.

Canales, K. (2021, September 16). Facebook's AI moderation reportedly can't interpret many languages, leaving users in some countries more susceptible to harmful posts. *Business Insider*. https://www.businessinsider.com/facebook-content-moderation-ai-cant-speak-all-languages-2021-9

Carlén, M. (2017). What constitutes the prefrontal cortex? *Science*, *358*(6362), 478–482.

Carver, R. P. (1990). *Reading rate: A review of research and theory*. Academic Press.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. *Advances in Experimental Social Psychology*, *24*, 201–234.

Clark, K. R. (2020). Consumer neuroscience: Evolution and commercial applications. In D. Atli (Ed.), *Analyzing the Strategic Role of Neuromarketing and Consumer Neuroscience* (pp. 185–207). IGI Global.

Colombo, D., Suso-Ribera, C., Fernández-Álvarez, J., Cipresso, P., Garcia-Palacios, A., Riva, G., & Botella, C. (2020). Affect recall bias: Being resilient by distorting reality. *Cognitive Therapy and Research*, *44*(5), 906–918.

Cowan, G., Resendez, M., Marshall, E., & Quist, R. (2002). Hate speech and constitutional protection: Priming values of equality and freedom. *Journal of Social Issues*, *58*(2), 247–263.

DellaVigna, S., List, J. A., & Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *Quarterly Journal of Economics*, *127*(1), 1–56.

Duncan, B. (2004). A theory of impact philanthropy. *Journal of Public Economics*, *88*(9–10), 2159–2180.

Eriksson, K., Strimling, P., & Coultas, J. C. (2015). Bidirectional associations between descriptive and injunctive norms. *Organizational Behavior and Human Decision Processes*, *129*, 59–69.

Fehr, E., & Schmidt, K. M. (2003). Theories of fairness and reciprocity: Evidence and economic applications. *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, *1*, 208–257.

Ferguson, E., & Flynn, N. (2016). Moral relativism as a disconnect between behavioural and experienced warm glow. *Journal of Economic Psychology*, *56*(June), 163–175.

Ferrari, M., & Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *NeuroImage*, *63*(2), 921–935.
Figner, B., Knoch, D., Johnson, E. J., Krosch, A. R., Lisanby, S. H., Fehr, E., & Weber, E. U. (2010). Lateral prefrontal cortex and self-control in intertemporal choice. *Nature Neuroscience*, *13*(5), 538–539.
Fiske, S. E., & Taylor, S. T. (1984). *Social cognition*. Random House.
Gefen, D., Ayaz, H., & Onaral, B. (2014). Applying functional near infrared (fNIR) spectroscopy to enhance MIS research. *AIS Transactions on Human-Computer Interaction*, *6*(3), 55–73.
Gęsiarz, F., & Crockett, M. J. (2015). Goal-directed, habitual and pavlovian prosocial behavior. *Frontiers in Behavioral Neuroscience*, *9*, Article 135.
Hare, T. A., Camerer, C. F., Knoepfle, D. T., & Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *Journal of Neuroscience*, *30*(2), 583–590.
Heinze, K., Ruh, N., Nitschke, K., Reis, J., Fritsch, B., Unterrainer, J. M., Rahm, B., Weiller, C., & Kaller, C. P. (2014). Transcranial direct current stimulation over left and right DLPFC: Lateralized effects on planning performance and related eye movements. *Biological Psychology*, *102*(1), 130–140.
Holländer, H. (1990). A social exchange approach to voluntary cooperation. *The American Economic Review*, *80*(5), 1157–1167.
Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin*, *30*(2), 161–172.
Hong, S. M., & Faedda, S. (1996). Refinement of the Hong Psychological Reactance Scale. *Educational and Psychological Measurement*, *56*(1), 173–182.
Huppert, T. J. (2016). Commentary on the statistical properties of noise and its implication on general linear models in functional near-infrared spectroscopy. *Neurophotonics*, *3*(1), Article 010401.
Hysenbelli, D., Rubaltelli, E., & Rumiati, R. (2013). Others' opinions count, but not all of them: Anchoring to ingroup versus outgroup members' behavior in charitable giving. *Judgment and Decision Making*, *8*(6), 678–690.
International Covenant on Civil and Political Rights, December 16, 1966, https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights
Kahneman, D., & Knetsch, J. L. (1992). Valuing public goods: The purchase of moral satisfaction. *Journal of Environmental Economics and Management*, *22*(1), 57–70.
Kapil, P., & Ekbal, A. (2020). A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, *210*, Article 106458.
Kemper, C. J., Beierlein, C., Bensch, D., Kovaleva, A., & Rammstedt, B. (2012). Eine Kurzskala zur Erfassung des Gamma-Faktors sozial erwünschten Antwortverhaltens [A short scale to assess the gamma factor of socially desirable response rehavior]. *GESIS-Working Papers*, *25*, 1–27.
Kepios Pte. Ltd. (2022). *Digital around the world*. DataReportal. https://datareportal.com/global-digital-overview
Kocsis, L., Herman, P., & Eke, A. (2006). The modified Beer-Lambert law revisited. *Physics in Medicine and Biology*, *51*(5), 91–98.
Krampe, C., Gier, N. R., & Kenning, P. (2018). The application of mobile fNIRS in marketing research – Detecting the 'First-Choice-Brand' effect. *Frontiers in Human Neuroscience*, *12*, Article 433.
Lapinski, M. K., & Rimal, R. N. (2005). An explication of social norms. *Communication Theory*, *15*(2), 127–147.
Lieberman, M. D., Berkman, E. T., & Wager, T. D. (2009). Correlations in social neuroscience aren't voodoo: Commentary on Vul et al. (2009). *Perspectives on Psychological Science*, *4*(3), 299–307.
MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLoS ONE*, *14*(8), Article e0221152.
MacCarthy, M. (2020). *Transparency requirements for digital social media platforms: Recommendations for policy makers and industry*. Transatlantic Working Group.
Margolis, H. (1982). *Selfishness, altruism, and rationality: A theory of social choice*. Cambridge University Press.
Melnyk, V., van Herpen, E., & van Trijp, J. C. M. (2010). The influence of social norms in consumer decision making: A meta-analysis. In M. C. Campbell, J. Inman, & R. Pieters (Eds.), *Advances in Consumer Research* (Vol. 37, pp. 463–464). Association for Consumer Research.
Meske, C., & Bunde, E. (2022). Design principles for user interfaces in AI-based decision support systems: The case of explainable hate speech detection. *Information Systems Frontiers*.
Murphy, K., & Garavan, H. (2004). An empirical investigation into the number of subjects required for an event-related fMRI study. *NeuroImage*, *22*(2), 879–885.
Neumann, R. (2019). The framing of charitable giving: A field experiment at bottle refund machines in

Germany. *Rationality and Society*, *31*(1), 98–126.
Newton, C. (2019, February 25). The trauma floor: The secret lives of Facebook moderators in America. *The Verge*. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona
Nissen, A., & Gier, N. R. (2021). A neural investigation of eLoyalty on eCommerce websites. *ICIS 2021 Proceedings*, *6*.
Olson, M. (1971). *The logic of collective action: Public goods and the theory of groups*. Harvard University Press.
Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., & Burgess, P. W. (2020). The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Annals of the New York Academy of Sciences*, *1464*(1), 5–29.
Plassmann, H., Venkatraman, V., Huettel, S., & Yoon, C. (2015). Consumer neuroscience: Applications, challenges, and possible solutions. *Journal of Marketing Research*, *52*(4), 427–435.
Plaza-del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications*, *166*, Article 114120.
Quaresima, V., & Ferrari, M. (2019). Functional near-infrared spectroscopy (fNIRS) for assessing cerebral cortex function during human behavior in natural/social situations: A concise review. *Organizational Research Methods*, *22*(1), 46–68.
Roberts, R. D. (1984). A positive model of private charity and public transfers. *Journal of Political Economy*, *92*(1), 136–148.
Santosa, H., Zhai, X., Fishburn, F., & Huppert, T. (2018). The NIRS Brain AnalyzIR toolbox. *Algorithms*, *11*(5), Article 73.
Scholkmann, F., Kleiser, S., Metz, A. J., Zimmermann, R., Mata Pavia, J., Wolf, U., & Wolf, M. (2014). A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. *NeuroImage*, *85*, 6–27.
Shang, J., & Croson, R. (2009). A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. *Economic Journal*, *119*(540), 1422–1439.
Shapiro, A. D., & Grafton, S. T. (2020). Subjective value then confidence in human ventromedial prefrontal cortex. *PLoS ONE*, *15*(2), Article e0225617.
Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*, 687–690.
Theotokis, A., & Manganari, E. (2015). The impact of choice architecture on sustainable consumer behavior: The role of guilt. *Journal of Business Ethics*, *131*(2), 423–437.
Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. The Guilford Press.
Ullmann, S., & Tomalin, M. (2020). Quarantining online hate speech: Technical and ethical perspectives. *Ethics and Information Technology*, *22*(1), 69–80.
United Nations. (2020). United Nations strategy and plan of action on hate speech: Detailed guidance on implementation for United Nations field presences. *United Nations Report*, *September*, 1–52.
United Nations Human Rights Council. (2011). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue. UN Doc A/HRC/17/27*.
van Teunenbroek, C., Bekkers, R., & Beersma, B. (2020). Look to others before you leap: A systematic literature review of social information effects on donation amounts. *Nonprofit and Voluntary Sector Quarterly*, *49*(1), 53–73.
Watanabe, M., Bartels, A., Macke, J., Murayama, Y., & Logothetis, N. (2013). Temporal jitter of the BOLD signal reveals a reliable initial dip and improved spatial resolution. *Current Biology*, *23*(21), 2146–2150.
Yücel, M. A., von Lühmann, A., Scholkmann, F., Gervain, J., Dan, I., Ayaz, H., Boas, D., Cooper, R. J., Culver, J., Elwell, C. E., Eggebrecht, A., Franceschini, M. A., Grova, C., Homae, F., Lesage, F., Obrig, H., Tachtsidis, I., Tak, S., Tong, Y., … Wolf, M. (2021). Best practices for fNIRS publications. *Neurophotonics*, *8*(1), Article 012101.
Zhang, Y., Brooks, D. H., Franceschini, M. A., & Boas, D. A. (2005). Eigenvector-based spatial filtering for reduction of physiological interference in diffuse optical imaging. *Journal of Biomedical Optics*, *10*(1), Article 011014.
Zimmermann-Janssen, V. E. M. (2020). The more, the worse – When the communication of high majority sizes decreases the impact of descriptive normative information on consumers' willingness to donate. In J. Argo, T. M. Lowrey, & H. J. Schau (Eds.), *NA - Advances in Consumer Research* (Vol. 48, pp. 1242–1242). Association for Consumer Research.