The Devil is in the Detail: Reconceptualizing Bystander Reactions to Online Political Hostility

Tanja Marie Hansen*
Aarhus University, Denmark

Lasse Lindekilde
Aarhus University, Denmark

Simon Tobias Schulz Karg
Aarhus University, Denmark

Abstract

Online content moderation is often equated with platform-initiated take-downs and account suspensions. Counterspeech and other types of pro-social bystander behaviors present alternative approaches. They rely on the reactions of online bystanders to crowd-moderate hostile content. However, the current conceptual understanding of bystander reactions to hostility on social media is inadequate. What constitutes a reaction; what functions do specific reactions serve; and what audiences are reactions aimed at? Conceptual clarity is a prerequisite for studying the prevalence and consequences of pro-social bystander reactions. We therefore offer a novel unifying framework for conceptualizing bystander reactions to online hostility: The Online Bystander Repertoire of Action (OBRA) Framework. This provides a foundation for future research on crowd-moderation, as it 1) combines insights across multiple research fields, 2) explicates the complex possible reactions beyond simplistic 'reaction/no reaction´ distinctions, and 3) draws attention to the degree of pro-sociality of bystander reactions – something that is often taken for granted.

Keywords
Counterspeech, hate speech, conceptualization, bystanding

Introduction

Though platforms come and go, and users rather seamlessly migrate from one to the other, it has become clear that a Social Media presence has become a staple of modern life for many people. Originally heralded as a hitherto unseen opportunity for widespread citizen-driven democratic dialogue (Mutz, 2008), a glance at a random comment thread on Social Media, as it looks today, can be a sobering experience. Despite immense effort from Social Media companies in moderating the content visible on their platforms – reflecting their desire to maintain a hospitable environment to platform users whose engagement pays the bills (Gillespie, 2018) – hostile content is widespread. The early vision of flourishing online deliberation is thus rarely found in practice. In response to the widespread nature of online hostile content, a growing body of research argues that citizen-generated responses (crowd-moderation) may assist in tackling the issue. By encouraging online bystanders, who encounter the hostile content as they move about their online lives, to react, one could potentially boost existing moderation efforts in the fight against hostile content online. This strategy of crowd-moderation promises great unrealized potential. On the one hand, research suggests that a large proportion of social media users report being exposed to hostile content (Duggan, 2017; Siegel and Badaan, 2020), i.e. they recognize content as hostile and, thus, are present and potentially able to react. On the other hand, research also tells us that the individual likelihood of bystander intervention is low (Vogels, 2021). Therefore, any strategy that would get more people to stand up against online political hostility would be highly important. However, effectively studying the potential of bystander reactions to online political hostility requires a better understanding of what reaction options online bystanders have when faced with hostile content, how these options differ in terms of costs and benefits, and whether these reactions are equally likely to have a positive impact on the deliberative environment on social media.

In this article, we therefore aim to provide a deeper understanding of online bystander reactions to online political hostility. We do so pursuing two main objectives. 1) We review and synthesize existing conceptualizations of online bystander reactions used within communication studies, political science and sociology in order to identify the main conceptual distinctions. 2) We introduce a unifying conceptual framework of online bystander reactions to online political hostility - The Online Bystander Repertoire of Action (OBRA) Framework - which constitutes a resource for reflecting on what constitutes online bystander reactions and how they are best subcategorized. The OBRA framework builds upon existing conceptual distinctions (e.g. the distinction between direct/visible and indirect/invisible forms of reactions as well as the idea of a varying threshold (low/high) for different types of reactions), but expands on these and introduces new ones to build a more comprehensive and unifying framework. As such, the OBRA framework considers the visibility of a reaction, the specific medium through which one communicates disagreement (*e.g.* reporting via clicking a button or typing counterspeech) and the intended audience (victim, perpetrator, onlooker). Likewise, the framework incorporates the different thresholds to react in various ways and highlights the degree to which a reaction, especially counterspeech, should be considered pro-social or can carry unwanted anti-social characteristics.

There are two main reasons why the existing literature would benefit from a shared framework for the study of online bystander reactions in response to online political hostility. As already hinted at, the topic of how

bystanders respond to online hostility has gained traction across various fields of study the last approximately fifteen years, and has been studied under the guise of multiple terms. Among the most used are 'bystander intervention' and 'online civic intervention', yet subcategories of the phenomenon have also gained traction with terms like 'flagging', 'reporting', 'counterspeech' and 'counterarguing' appearing en masse. While the increased interest on the one hand reflects a vibrant subfield of research, on the other hand the sheer number of terms covering similar – if not identical – phenomena, reflects some underlying problems within the subfield.

Firstly, the apparent conceptual disagreement reflects a tendency towards non-cumulative science. The multitude of terms appearing across fields of study within the social sciences, hinders the production of shared knowledge and understanding of the topic. In short, we need to all be on the same page about *what* we are studying (terms and definitions) before we go ahead and carry out plans on *how* to meaningfully study it (empirical studies). Given that a shared understanding starts at familiarizing oneself with the work already done by others – something made much harder by the topic appearing across different fields of research and under many names – a need exists for a shared naming protocol and framework for the study of online bystander reactions.

Second, the many terms used in extant literature indirectly reflect an underdeveloped understanding of the subject under investigation and a misleading perception that an online bystander intervention is something *we know when we see it*. Despite the tendency for scholars to coin new terms for the phenomenon, few delve into what the terms cover and what they leave out, and limited space is spent contemplating the many subcategories of behavior subsumed under the title of 'intervention'. The limited attention paid to the different reaction options available to online bystanders is unfortunate, given that for our shared knowledge to be turned into actionable advice on overcoming the democratic problem of online hostility, we must be sure we understand and measure the topic in all its guises. To give a few examples of why this conceptual inattention might pose a problem: How do we know that when no (publicly) visible reaction occurs this reflects bystander passivity and thus should be perceived as a challenge to democratic discourse online? What if bystanders merely chose non-visible strategies of reaction like sending a direct message to the victim, or deliberately chose to ignore an online troll to rob them of the satisfaction? And when does a statement cross the border from desirable pro-social counterspeech, calling out hostile content, over into anti-social mockery and hostility towards the initially hostile user? Again, it appears this research agenda would benefit from a shared understanding of what constitutes online bystander interventions, and which interventions are pro-social in nature.

Following this line of thought, it also becomes evident that a conceptual disentanglement from the term 'intervention' may be fruitful.[1] We believe that broadening our focus from merely the behaviors that constitute interventive action, to considering the full repertoire of actions available to an online bystander, would lower the risk of overestimating bystander passivity online (Moxey and Bussey, 2020). This would also better equip us to make conscious decisions about a) what we specifically mean when using the term 'intervention' in a given setting, and what bystander reactions this choice overlooks, and b) which response options carry the potential to make a positive and constructive difference in the face of an increasingly hostile online debate culture.

The remainder of this article is organized as follows: In the next section we define key terms before providing a summary and overview of extant conceptions of bystander reactions to online political hostility. In the

following section we present our unifying conceptual framework - The Online Bystander Repertoire of Action (OBRA) Framework. This section introduces the key conceptual distinctions of the framework regarding visibility, medium, intended audience, threshold, and degree of pro-sociality of bystander reactions. In doing so, we discuss the meaning of no visible bystander reactions and silence as well as the degree to which specific types of counterspeech are to be considered pro-social reactions. The conclusion summarizes our argument, and we contend that the OBRA Framework provides a comprehensive conceptualization which is better equipped for handling the many different forms of bystander reactions to online political hostility, and which provides important nuances to our understanding of the phenomenon that can help future research steer clear of both the tendency to overestimate bystander apathy and underestimate the degree of anti-social sentiments in e.g. counterspeech.

Summarizing Extant Conceptions of Bystander Reactions to Online Political Hostility

We use the term 'online political hostility' as an umbrella term for behaviors on social media that undermine democratic norms of public deliberation (Rasmussen 2023: 18). Online political hostility, thus, refers to publicly displayed incivility that violates deliberative norms and, thus, hinders "public discussion and carefully weighing a comprehensive set of ideas" (Muddiman 2017). This includes incivility (Papacharissi 2004), intentional spreading of misinformation (Guess & Lyons 2020) and hate speech, which can be defined as public degradation or incitement to violence or hatred against a group of persons or a member of such group defined on the basis of for instance race, descent, religion or belief, or national or ethnic origin (Sellars, 2016: 20). Such behavior undermines people's free and equal access to public discussions (Stromer-Galley & Wichowski 2011). Importantly, by applying this definition we exclude forms of cyber-bullying taking place in non-public forums concerning private matters and characteristics (e.g., class-mate bullying based on looks etc.). Building on this, we consider an online 'bystander' to be someone who is exposed to online political hostility as a witness and has the opportunity to react in some form or other. 'Bystander reactions' is then understood as any direct/visible or indirect/invisible reactions to online political hostility by a bystander. As we shall elaborate on below, such online bystander reactions can vary along a number of dimensions, including its pro-sociality, *i.e.*, the degree to which it uses a non-hostile framing and has a constructive aim of trying to restore democratic norms of public deliberation by showing disapproval of hostile content.

In our attempt to summarize conceptions of bystander reactions to online political hostility in existing academic research we made several delimiting choices. First, as the purpose of this summary mainly is to describe and illustrate the diversity of conceptions of bystander reactions to online political hostility in extant literature, we have not conducted a full systematic review (for a partial overview see Rudnicki et al. 2022), but rather a review akin to what Leidner (2018) calls an 'organizing review', which does "not claim to be comprehensive". Second, we mostly delimited ourselves to academic literature published within the social sciences, particularly communication studies, political science and sociology. Thirdly, we focused on the period from 2010 onwards (2010 – 2023). These decisions were based on these fields being the most preoccupied with

Table 1. A non-exhaustive list of terms covering bystander reactions to online political hostility used in academic research within communications studies, political science, and sociology

General term	Clicking	Typing
Bystander Intervention - Moxey and Bussey 2020• - Naab et al. 2018$^{\Delta}$ - Watson and Lewis 2019$^{\Delta}$ - Jost et al. 2020^{+} - Nelson et al. 2011^{+} - Lu and Luqiu 2023$^{\Delta}$ - Costello, Hawdon, and Cross 2016• - Celuch et al. 2023$^{\Delta}$ *Online civic intervention* - Kunst et al. 2021^{+} - Porten-Cheé et al. 2020$^{\Delta}$ *Audience intervention* - Lu and Luqiu 2023$^{\Delta}$ *Informal social control* - Watson and Lewis 2019$^{\Delta}$ - Costello, Hawdon, and Cross 2016• *Online collective efficacy* - Ozalp et al. 2020$^{\Delta}$ - Costello, Hawdon, and Cross 2016• *Civic engagement* - Jost et al. 2020^{+} *Social sanctioning* - Munger 2017^{+} *Pro-social behaviors* - Watson and Lewis 2019$^{\Delta}$ *Digital civil courage* - Jost et al. 2020^{+} *Opinion expressions* - Pang et al. 2016• *User engagement* - Kalch and Naab 2017$^{\Delta}$ *Intervention* - Álvarez-Benjumea and Winter, 2018• *Netiquette* - Costello, Hawdon, and Cross 2016• *Countermeasures* - Celuch et al. 2023$^{\Delta}$	*Flagging* - Wilhelm et al. 2020$^{\Delta}$ - Kunst et al. 2021^{+} - Porten-Cheé et al. 2020$^{\Delta}$ - Kalch and Naab 2017$^{\Delta}$ - Crawford and Gillespie 2016$^{\Delta}$ - Álvarez-Benjumea and Winter, 2018• - Ozalp et al. 2020$^{\Delta}$ *Reporting* - Wilhelm et al. 2020$^{\Delta}$ - Watson and Lewis 2019$^{\Delta}$ - Porten-Cheé et al. 2020$^{\Delta}$ - Crawford and Gillespie 2016$^{\Delta}$ - Álvarez-Benjumea and Winter, 2018• - Ozalp et al. 2020$^{\Delta}$ - Celuch et al. 2023$^{\Delta}$ *Click speech* - Jost et al. 2020^{+} - Pang et al. 2016• - Sklan 2013• *Social buttons* - Kunst et al. 2021^{+} - Sklan 2013• *Rating* Watson and Lewis 2019$^{\Delta}$ *Evaluation buttons* - Kalch and Naab 2017$^{\Delta}$ *Reactions* - Kalch and Naab 2017$^{\Delta}$ *Ranking tools* - Crawford and Gillespie 2016$^{\Delta}$ *Technical countermeasure* - Celuch et al. 2023$^{\Delta}$	*Counter-speech* - Siegel and Badaan 2020^{+} - Garland et al. Preprint• - Mathew et al. 2019• - Hangartner et al. 2021• - Garland et al. Preprint2• - Buerger 2020$^{\Delta}$ - Leonhard et al. 2018$^{\Delta}$ - Obermaier et al. 2021$^{\Delta}$ - Kunst et al. 2021^{+} - Porten-Cheé et al. 2020$^{\Delta}$ - Álvarez-Benjumea and Winter, 2018• - Ozalp et al. 2020$^{\Delta}$ *Counterarguing* - Leonhard et al. 2018$^{\Delta}$ - Obermaier et al. 2021$^{\Delta}$ - Naab et al. 2018$^{\Delta}$ - Porten-Cheé et al. 2020$^{\Delta}$ *Sanctioning message* - Siegel and Badaan 2020^{+} *Opposing speech* - Garland et al. Preprint2• *Replying* - Kalch and Naab 2017$^{\Delta}$ *Assertive response* - Nelson et al. 2011^{+} - Celuch et al. 2023$^{\Delta}$ *Discursive exchange of opinions* - Kalch and Naab 2017$^{\Delta}$ *Verbal sanction* - Álvarez-Benjumea and Winter 2018• *Statement of disapproval* - Costello, Hawdon, and Cross 2016•

Notes: The indicated scientific field of origin reflects the scientific outlet within which the manuscript was published. The category 'other' includes general scientific outputs, computational sciences, law studies and unpublished preprints.
Annotation: $^{+}$Political Science and sociology, $^{\Delta}$Communication research, and •Other.

the phenomenon during this period. Concretely, we started by searching the most generic terms such as 'bystander intervention' and 'bystander reaction' in combination with 'online' and/or 'social media' on Google Scholar, Scopus, ResearchGate and CORE. Then we branched out from these identified publications by looking at lists of references. In total we identified 26 academic publications that fit our criteria.[2] These titles were read, and their conception of online bystander reactions noted down. Table 1 summarizes these conceptions. The first column lists general terms used to capture the overall phenomenon of online 'bystander reactions', while column two and three list terms used to capture subcategories of reactions, organized according to whether they refer to reactions that involve either 'clicking' buttons or 'typing' a response (we borrow this overall typology from Jost et al. 2020). The list of terms is non-exhaustive as there may be conceptions we have missed, but the table includes the main conceptions used in extant literature and suffice for our purpose here of showcasing the diversity of conceptualization and identifying key conceptual distinctions to build upon when developing the OBRA Framework below.

Introducing the Online Bystander Repertoire of Action Framework

At its core, any bystander to hostile content online is faced with a choice: to act or not to act. This binary understanding of online bystander reactions has long dominated the literature on online bystander behavior, perhaps as a remnant of its ties to the literature on offline bystander reactions (Fischer et al., 2011). However, adhering to such a dichotomous understanding risks impeding our ability to gain a nuanced understanding of what constitutes a bystander reaction.[3] For instance, it raises questions like: How do we know that no bystander reaction occurred, simply because it was not visible to onlookers? And can we rightly assume that inaction reflects indifference or even agreement with the statement put forward by the offending party? And is it reasonable to assume that all online bystander interventions reflect an altruistic, pro-social intention to help?

While this widespread dichotomous operationalization of bystander reactions, and its inherent assumptions, raises issues as to measurement validity, research into the phenomenon indirectly encompasses a more multifaceted understanding, as visible in the subcategories presented in table 1. As the topic is currently treated in the literature, these types of bystander behavior fit neatly into one of two categories. Either the intervention involves 'clicking', that is pressing buttons unfolding pre-determined response-categories (*e.g.* reaction-buttons,[4] rating-buttons, or reporting-buttons), or 'typing', *i.e.* writing a message aimed at countering the hostile content.

Common to these categories is the implicit assumption that when no action is visible, no action has occurred, and that inaction can be interpreted as indirect approval of the hostile message. An article by Leonhard et al. directly puts words to this stating that "passive behavior of bystanders may be perceived as implicit approval of hatred" (2018: 559) while other contributions merely imply it by in various ways highlighting *action* as something "ordinary users [take] to fight disruptive online behavior with the aim of restoring civil and rational public discourse" (Kunst et al., 2021: 260) or by stating that "the open, unopposed expression of racism in a public forum can legitimize racist viewpoints" (Munger, 2017: abstract). The focus on action-prescribing terms like 'intervention' (Jost et al., 2020; Kunst et al., 2021; Moxey and Bussey, 2020; Naab et al., 2018; Nelson et al., 2011; Obermaier et al., 2021; Porten-Cheé et al., 2020; Siegel and Badaan, 2020; Watson et al., 2019) and

'engagement' (Kalch and Naab, 2017) further highlight this. There are, however, good reasons why going beyond this crude understanding of bystander reactions to look at the many forms reactions to online political hostility can take, may be preferable – not least with regards to dismantling the myth of silent approval.

The conceptual subdivision of 'clicking' and 'typing' that has taken place within the research field is commendable and reflects a research field in motion. Still, further subdivisions are necessary. In figure 1 we respond to this need as we build on existing subdivisions and take them a step further. The figure contains a visualization of The Online Bystander Repertoire of Action Framework featuring an overview of the many different response options available to an online bystander to hostile content on social media.

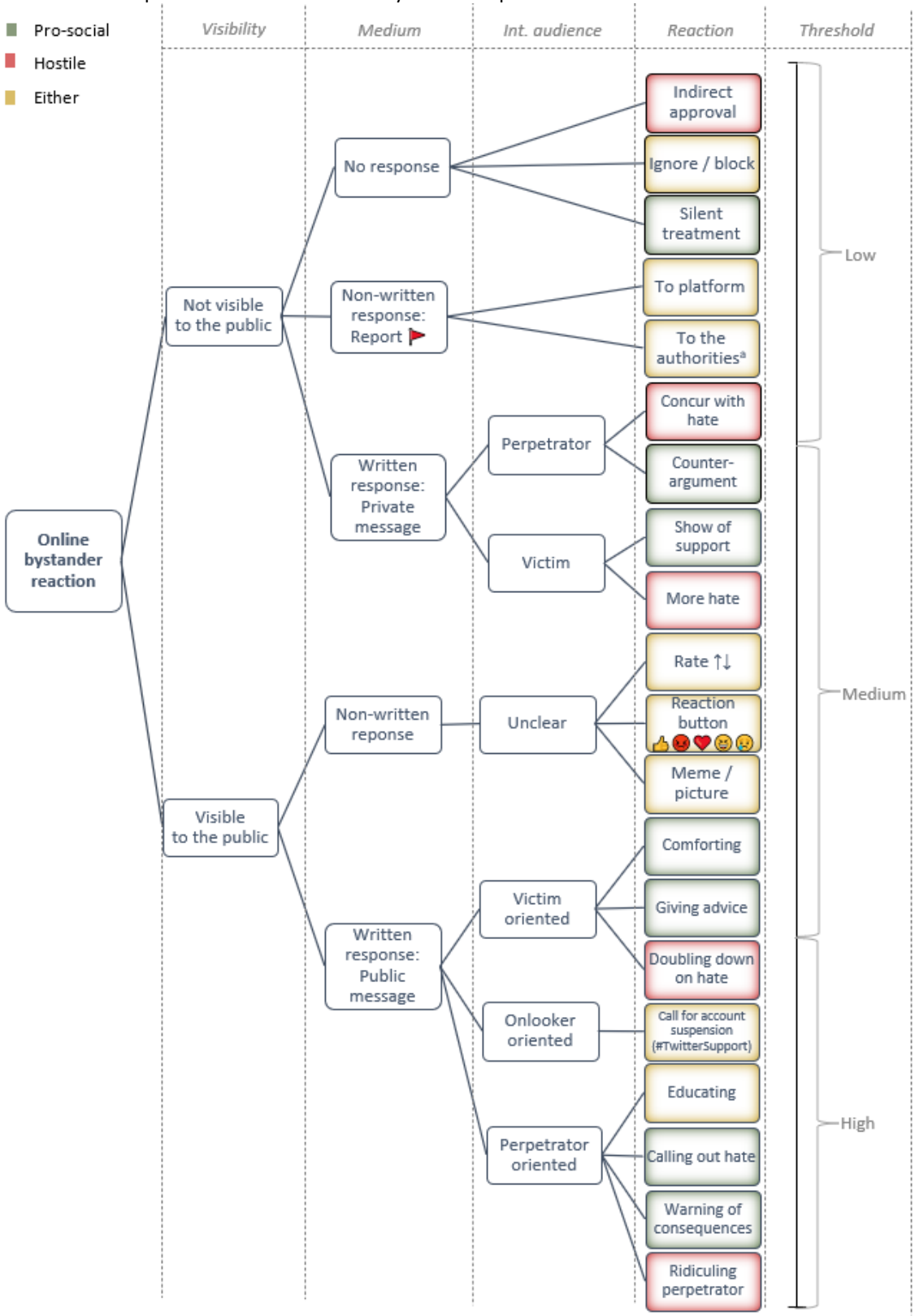
When no visible action occurs

There are a few plausible ways in which what may look like bystander inaction might in fact reflect non-visible strategies in response to witnessing online hostility. Just because action is not visible to other onlookers,[5] this does not necessarily entail inaction behind the scenes. Although focusing on visible action makes sense when researching offline bystander reactions, in the online setting a bystander can also choose to make use of the complex structure of social media platforms to act stealthily. Some of these reactions have received attention in existing studies, while others, to our knowledge, have yet to be explored empirically.

The clearest example of an action that is not displayed to observers of an online debate is the act of *reporting,* or "flagging" as it is commonly known. This involves what Watson et al. call "informal social control" (Watson et al., 2019: 1846) as users pass on information about the inappropriateness of content to either authorities, like law enforcement, or social media moderators. With regards to hostile content, reporting to moderators is by far the most utilized strategy, perhaps owing to the content's often offensive but legal nature (Kunst et al., 2021). Reporting is an integral part of the structure and inbuilt features of social media platforms that allow users to press a button to indicate that user-created content violates platform rules have become commonplace (Crawford and Gillespie, 2016: 411; Gillespie, 2018). In line with the weight put on reporting as a tool to counter online hostility, reporting has received a fair amount of attention within existing research as a subcategory of 'online bystander intervention' (Crawford and Gillespie, 2016; Kalch and Naab, 2017; Kunst et al., 2021; Porten-Cheé et al., 2020; Watson et al., 2019; Wilhelm et al., 2020).

Crucially, to report content only *indirectly improves* other users' perceptions of the civility level of the debate, within which the hostile content was posted. Given that only site moderators receive the information when content is flagged (Crawford and Gillespie, 2016: 421; Kalch and Naab, 2017: 402), reporting relays limited feedback about the appropriateness of the content to other bystanders. It, thus, can appear as if no reaction occurred. Only if content is removed, or a hostile user is sanctioned, in response to reports of violations of platform rules of engagement, does the bystander reaction indirectly become visible through some form of 'content has been removed' message. Given that this may take hours or days, and the original hostile content is no longer visible, the lesson to be learned by other users encountering the removal-message remains negligible.

Figure 1. A visual representation of the Online Bystander Repertoire of Action Framework



Note: Placement of reactions according to thresholds in figure 1 are approximate and depend on the specific situation (e.g., anonymity) and procedures (e.g., degree of detail necessary to report to authorities).

In the literature, reporting is mostly discussed as a pro-social tool for moderation, yet it carries a risk of misuse. Examples of this are out there, an infamous one being a hostile flagging campaign against the LGBTQ+ community on Facebook in 2012 (Gillespie, 2018: 92), and the LGBTQ+ community's later countereffortsto return the favor to the anti-gay community in 2020 by high jacking the Proud Boys' hashtag and flooding the zone with messages of love (Elassar, 2020). This highlights the anti-social potential inherent in reporting as a moderation tool. However, such misuse is far from the norm, as it requires a group effort consisting of multiple reports, and many platforms have additional layers to their moderation efforts and are thus not "obligated to honor the flags" (Crawford and Gillespie, 2016: 419) if signs of misuse appear.

Another example of a reaction that is not displayed to observers of an online debate, one which is so far overlooked within the literature, regards the act of sending a private message responding to the hostile content. At a theoretical level, this can take multiple forms and serve a variety of purposes, depending on the content and receiver of the private message. The user responsible for producing the hostile content may receive private messages either concurring with the hateful statement or countering it. Because of the concealed nature of this type of reaction, what may in fact represent a desirable practice of citizen deliberation, may be interpreted as silence by onlookers. Likewise, the already victimized user or group of users may receive further hateful messages or, alternatively, receive pro-social messages indicating sympathy or advice for action in response to experienced victimization.[6] The invisibility of private messages also impedes our knowledge of how widespread of a phenomenon this is. While social media sites can disclose levels of reporting and content-removal (Reynders, 2021: 3), and web scraping allows for collection and count of counterspeech in comment sections (Garland et al., 2020a; Mathew et al., 2019), these approaches are not available for private online conversations discussing the content of public conversations. Nevertheless, we occasionally see remnants of the approach, as commenters indicate that either the hostile user or the victim should "see DMs [direct messages]", and interviews with counterspeakers also reveal a recurring role for both pro-social and hostile exchanges via private messages (Buerger, 2021; Hansen et al., 2023).

Silence may not always reflect approval

Even when no visible response *correctly* reflects inaction, there is no obvious reason to assume that this silence (non-response) equals approval. In fact, doing so would paint a grim picture of the attitudes of social media users. The few numbers we do have indicate that both reporting and counter speaking constitute relatively rare events (Garland et al., 2020a: 8, 2020b: 7).[7] Also, when directly asked about their online behavior the majority of online bystanders respond that they do not react, or intend to act, when faced with hostile content online (Andresen et al., 2022: 20; Leonhard et al., 2018: 567). Approximately seven-in-ten Americans (Pew Research Center, 2021: 3) and more than half of the world's population (DataReportal, 2022: 9) regularly use some form of social media. Does this mean that almost every second of us is silently hiding behind our screen, approving of hostile content?

Kalch and Naab (2017) likewise expressed criticism of this simplistic interpretation of silence, stating that: "[…] not flagging inappropriate content (and the same may hold true for not using evaluation buttons and not

replying) is not necessarily a signal of agreement but may indicate ambivalence toward the content, inertia, a lack of knowledge, or a lack of perceived self-efficacy" (Kalch and Naab, 2017: 402; see also, Crawford and Gillespie, 2016). This view is in line with the finding that willingness to react pro-socially is lower when one does not feel personally affiliated with the targeted individual or group (Bastiaensens et al., 2014; Brewer and Kerslake, 2015; Liebst et al., 2018), and reflects the understanding that inaction might simply reflect inattentiveness and a lacking willingness to engage. Alternatively, inaction could reflect a belief that all speech is valuable, despite its hostile content, fearing its removal in response to action (Atske, 2021; Kozyreva et al., 2023). Then again, it could simply reflect the sense that any action taken would not make a positive difference (Crawford and Gillespie, 2016: 420).

Less discussed by researchers studying online bystander interventions is the idea that silence could be a pro-social weapon in its own right. As pointed out by Evita March (2016) the common phrase echoed across the internet: "Don't feed the trolls" may contain a kernel of truth. With the discovery that online trolls[8] are characterized by a negative potency type of social reward structure in which individuals are motivated by gaining negative attention (Craker and March, 2016: 80), interactions and conflict may be exactly what they desire. In this light, inaction may be a deliberate – and effective – strategy employed by bystanders when faced with online hostility. Silence could thus, at least for some, reflect purposeful silent treatment rather than approval.

Either of these alternative interpretations of online bystander inactivity – whether they react in a way we cannot see or remain silent for reasons of their own – require further empirical investigation to determine the underlying motivations driving online bystander behavior. For now, " […] the population of non-flaggers [and generally inactive bystanders] is a murky mix of some of all of these" (Crawford and Gillespie, 2016: 420). Perhaps a straightforward approach could involve asking questions about the considerations that lead people to remain silent, rather than only inquiring about motivations and consequences of action.[9]

Thresholds to reacting

Relative to remaining silent – no matter why one chooses to do so – acting when exposed to online hostility requires greater effort of the bystander. In the literature, this has been discussed applying the terminology of "thresholds" (Kunst et al., 2021; Porten-Cheé et al., 2020). Of the non-visible response options discussed above, reporting to platform moderators constitutes a low threshold strategy.[10] It requires little effort, as no formulation of argumentation or putting oneself on the line in the eyes of neither onlookers nor the hostile user is necessary. Correspondingly, contacting the hostile user through direct messaging – though not visible to others – poses a higher threshold strategy.

Generally, the perceived costliness of acting is likely to rely on two core features. First, how much effort (*e.g.,* time, money, cognitive effort) will I have to exude? And second, how severe repercussions, if any, am I likely to face if I chose to act? Reactions that are high on these features can be said to pose higher thresholds to action, while reactions that are low on one or both features pose lower thresholds. Theoretically, we would therefore expect low threshold bystander reactions to occur with higher frequencies and to be easier to promote than high threshold actions.

The visible response options, presented in figure 1, likewise differ with regards to the effort required by the online bystander. Generally, written responses make up the most demanding reactions, as they not only require effort and writing literacy to formulate, they also require the bystander to take a public stance, on what often constitute contentious political issues, unlike for instance upstanding behavior in response to online bullying of a non-political nature (Jost et al., 2020). This, therefore, also entails sticking one's neck out in an online setting already characterized by a hostile tone. As one publicly stands up to a hostile user, one risks becoming the new target of hostility. Although physical danger is not imminent, these costs may be perceived as especially high in the online setting, where written expressions of disapproval are visible to a larger audience, remain visible for longer, and require a more direct signal (*e.g.* words, memes, links) than in a typical offline bystander setting during which one can momentarily signal disapproval to the victim or other bystanders through, for instance, a shrug of the shoulders or shake of the head. The fear of becoming the new victim of hostilities may also be why some platforms have introduced less costly response options that mimic the response options available in offline settings, such as snooze buttons and the option of disabling further comments (Constine, 2017; Taylor, 2021). At the same time, however, costs may also be perceived as lower online than offline, given that the risk of physical harm is dramatically reduced for online activities.

Following this logic of costs, non-written, but visible, responses like rating, pushing social buttons or commenting using graphics (*e.g*., gifs, emojis, Memes) constitute midrange response strategies. It takes limited effort to, for instance, push a social button in the shape of a distraught emoji, downrate a hateful post or compose a comment only containing a humoristic gif. Yet, given the fact that the individual user's expression of disapproval is visible to others, including the hostile party, these responses carry many of the same costs affiliated with written counterspeech.

As regards *ratings and social buttons*, the logic of these non-written responses differs from that of counterspeech. These approaches rely less on a deliberative logic of persuasion – *e.g.,* "I disagree for the following *n* reasons" – and instead are designed to influence the tone of the online debate by impacting the placement and visibility of hostile content or awarding encouragement to other active bystanders. While "'Likes do not automatically imply support" (Pang et al., 2016: 899) – something that makes the specific motivations for pushing social buttons difficult to gauge – they frequently feature as a response option available to online bystanders and thus deserve more attention within research on countering online hostility.

Evidence from interviews with the organized Swedish counterspeech movement on Facebook, *Jagärhär* [I am here], hints at specific strategic uses of social buttons and rating systems to counter online hostility (Buerger, 2021). By up-voting or liking positive contributions to the online debate, or alternatively down-voting hostile inputs, a user can pro-actively influence the impression of the debate gained by online 'passers-by', who automatically view the content algorithmically deemed to be 'most relevant'. In the interviews, members of the movement described this strategy as a way to "lift them [counterspeech] up in the fields and poke down hateful comments" (Buerger, 2021: 5).

Much like reporting, this approach can, in theory, be coopted for purposes of a less pro-social nature, as the algorithms can be swayed in similar ways if hostile content gains deliberate likes and up-votes. Likewise, the fact that the level of interpretation necessary on the part of victims and bystanders in understanding the signal

intended by the press of a button, to an extend complicates its use as a pro-social reaction strategy. Facebook's reaction option, featuring a laughing emoji, illustrates this well. In interviews with Danish online counterspeakers, multiple interviewees indicated their frustration with Facebooks' laughing emoji, as they expressed difficulty in assessing whether people were laughing with them or at them (Hansen et al., 2023). It is, therefore, important to keep in mind the level of inbuilt pro-sociality of a response option when considering ways to encourage the strategic use of social buttons and rating systems in the future.

A feature likely to impact how the individual social media user experiences the threshold of reacting to online hate, yet only briefly discussed within the literature (Wong et al., 2021; You and Lee, 2019) relates to the impact of anonymity. Much like public visibility, online bystander reactions that clearly affiliate users with reactions can be assumed to carry higher costs, given that the element of anonymity curbs the risk of becoming the new victim of hate or other potential social repercussions. While reactions such as rating content, or reporting content to platform moderators, does not disclose involvement to a wider audience, nor to the perpetrator of hate speech (the reaction is anonymous except to platform owners), pressing most reaction-buttons (e.g., emojis) and all written responses involve disclosing the individual user's involvement. Currently this is not reflected in the OBRA framework, yet the logic is expected to transfer easily. Generally speaking, although the visualization of Figure 1 presents threshold size as continuous, placement of individual reaction options in terms of threshold size (i.e., costliness of reacting) is approximate. This is also reflected in differences among reaction options unrelated to user anonymity. For instance, to write a private message to a perpetrator of hate speech counter-arguing, could feasibly carry higher individual costs to that of writing a private message to the victim with additional hateful content. Distinguishing thresholds by broader categories like 'low', 'medium', or 'high' is therefore advised.

Not all counterspeech is pro-social

Publicly *visible written expressions of disapproval*, or as they are often known, counterspeech messages, constitute the online bystander response option with the highest threshold to action. Not only are they visible to the public[11] – hostile users included – they also take time and effort to formulate, making them the costliest bystander reaction covered by the OBRA-framework.

The existing literature is preoccupied with counterspeech (see for instance Buerger, 2021; Garland et al., 2020a, 2020b; Hangartner et al., 2021; Jost et al., 2020; Kalch and Naab, 2017; Kunst et al., 2021; Leonhard et al., 2018; Mathew et al., 2019; Nelson et al., 2011; Obermaier et al., 2021; Pang et al., 2016; Porten-Cheé et al., 2020; Siegel and Badaan, 2020), and a multitude of definitions of counterspeech exist. Across studies, agreement exists on the fact that the key defining feature of the concept is its responsive and verbal nature.[12] Counterspeech thus necessitates a direct response to a hateful utterance (Buerger, 2021; Garland et al., 2020a, 2020b; Leonhard et al., 2018: 559; Mathew et al., 2019; Obermaier et al., 2021: 3; Siegel and Badaan, 2020: 838). Many further highlight the crowd-sourced nature of the phenomenon (Leonhard et al., 2018: 559; Obermaier et al., 2021: 3). It is "a citizen generated response" (Garland et al., 2020a: 2) of a "civic" nature (Kunst et al. 2021), unlike the moderation carried out by employees hired by social media companies to remove or limit content, or journalists tasked with keeping utterances civil in the comment sections following news articles on social media.

Within this agreement, however, lies tension as to the intended mechanism through which counterspeech is expected to have its positive effect on online deliberation. This ranges from a focus on primarily limiting future utterances of online hate speech (prevention), by persuading the hateful user of their wrongdoings (Garland et al., 2020a; Hangartner et al., 2021: 1), to a focus on reducing the negative consequences of online hate speech for victims and onlookers, thus dampening its negative influence on online discourse (mitigation) (Buerger, 2021; Garland et al., 2020b; Siegel and Badaan, 2020: 838). Though rarely mentioned in concert, at a theoretical level nothing precludes both suggested causal mechanisms from occurring in unison, *i.e.*, both a preventive and mitigating function of counterspeech. Delving into the underlying mechanisms could consequently comprise a fruitful avenue of future research.

Despite the agreement on the positive societal impact of counterspeech presented above, surprisingly little research in the area includes a discussion of the ways in which online counterspeech can differ in terms of orientation and pro-sociality. In most instances, counterspeech is treated as innately constructive, only varying as regards its formulation (for an example see Hangartner et al., 2021). However, upon inspection examples of counterspeech on social media differ widely based on who the sender attempts to communicate with – victims, perpetrators, or onlookers – and whether they adhere to non-hostile or hostile forms of communication.[13]

On the one hand, some types of counterspeech are clearly pro-social, as they aim to shield a victim of online hostility by offering comforting statements or providing advice on how to respond to hostilities (*e.g.,* report, ignore, block, etc.). Alternatively, pro-social counterspeech can be directed at the perpetrator, calling out hate or warning the hostile user of the consequences of such utterances, be they exclusion from the platform, or the hurt feelings of those the hostile message targets. Pro-social counterspeech may also entail meta-communicative messaging where the message is steered towards neither victim nor perpetrator, but rather provides general commentary on pro-social behavior online, *e.g.* "Let's try to do something rare and have a civil and honest conversation" oriented at onlookers to the debate, or what Buerger calls "silent readers" (2021: 5).

On the other hand, other types of counterspeech are of a less pro-social nature, perpetuating the tendency towards hostile communication. This is known elsewhere in the literature as hateful (Obermaier et al., 2021: 3) or aggressive counterspeech (Young et al., 2018: 9), and can take many forms. The bystander may choose to counter the hostile message, but do so in a hostile manner, ridiculing or degrading the perpetrator of the initial hostile message, thus adding to the stream of online hate. Hostile bystander reactions may also target onlookers to the debate. This could, for instance, take the form of disrupting the online debate by intimidating onlookers from stating opposing views. Alternatively, victims may be targeted by further hostility from bystanders choosing to double down on the hate – not unlike when schoolchildren join the bully rather than stand up to defend the victim (Rudnicki et al., 2022). While disrupting the debate or doubling down on the hate are reactions available to bystanders to online hostility, they, however, cannot be categorized as counterspeech per se, given that they do not express an intention to *counter* the hostile content of the original message. They are, nevertheless, important potential written reactions to conceptualize, because of their clear negative consequences to the quality of online deliberation.

The ease with which these the level of pro-sociality of a counterspeech message can be deduced from the content and choice of wording of the message varies greatly. This interpretation not only hinges on whether we

try to infer the *intended* audience – i.e., who the sender had in mind when composing the message – but also on our interpretation of what it means to have a 'pro-social' effect, and a pro-social effect on what specifically. If our aim is to understand what drives online bystanders to engage in counterspeech, we should ideally aim to gauge the intended effect of the message. Was the message formulated with the intention of causing harm? However, if the aim is rather to understand how counterspeech messages affect the quality of online deliberation, negatively or positively, regardless of the aims of the message sender, this intentions-distinction becomes less relevant. Instead, empirical exploration of the experienced effects by all relevant audience types takes priority. Again, and importantly, this includes general exploration of whether the level of pro-social impact differs depending on whether evaluated in relation to the experience of the specific victim of a hostile comment,[14] or the wider network of onlookers to debates on social media. In other words, what aspect of online counterspeech one studies, matters for how one best conceptualizes written online bystander reactions as either hostile or pro-social.

Aside from the differences in level of pro-sociality related to the intention or impact-distinction discussed above, some types of online written bystander reactions may be hard to identify as either pro-social or hostile for other reasons. For instance, a person may attempt to educate another user who spews hate online. While providing facts can be a force for good (when correct) in guiding a debate back on track, a thin line exists between providing information and showing off in a show of moral grandstanding. At other times, counterspeech messages may be framed humorously to the extent that the readers struggle to disentangle the pro-social message from its entertainment value, which in turn can include a tendency to ridicule the author of the hostile message. This inbuilt tension could also be related to Hangartner et al.'s finding that humorous counterspeech showed limited potential in "reducing the production of xenophobic hate speech" (2021: 3). Likewise difficult to classify are Instances in which the producer of counterspeech attempts to provide advice to the victim of hate speech. This can, for instance, take the form of public, written calls for account suspensions. Counterspeech messages may tag platform moderators, *e.g.,* @TwitterSupport, to put a stop to hostilities. At first sight, this appears pro-social through and through. Nevertheless, this strategy can be misused as content risks classification as rule-breaking merely through association with the trending hashtag, regardless of whether the moderator concurs with the evaluation of content as in conflict with the platform's rules of engagement.

At the moment, the research community mostly tends to treat all types of counterspeech as pro-social and thus worth mobilizing in the fight against hostile content online. Notable exceptions are Obermaier et al.'s (2021) work on in-group bystander intentions to intervene. However, as discussed above, digging into examples of counterspeech on social media raises questions as to whether this view of counterspeech as pro-social by default is fruitful. While simply classifying all written comments that respond to online hate speech as pro-social counterspeech reflects and understandable pragmatic choice, it carries a risk of missing the trees for the forest. We instead advise cautious consideration when designing studies and behavioral interventions aimed at fostering a better online forum for deliberation, to make sure that the effort is focused on promoting the pro-social versions of counterspeech rather than the larger universe of online written bystander reactions – the hostile type included.

Conclusion

The recent surge in interest among scholars and practitioners in online phenomena such as content moderation, flagging, counterspeech, and crowd-moderation, and the accompanying increase in publications on the topic across multiple fields of research (communication, criminology, political science, sociology, etc.), is indicative of a field ripe for re-focusing. The surge has hitherto primarily informed foundational aspects of the phenomenon, asking questions like: When do online bystanders report content? What is counterspeech, and what motivates people to engage in it? And what, if any, are the effects on online deliberation of online citizen interventions like flagging and counterarguing?

The research field now stands to enter a more mature phase. Upon entering this second phase, the aim of this article has been to provide researchers with a conceptual framework to guide their thinking while designing and construing empirical studies of online bystander reactions to online hostility. With our framework, we have sought to make explicit the many implicit choices inherent to a study of something so complex as online bystander reactions. Our hope is that our framework may bring together research across multiple fields of research, and guide the production of future research towards conceptualizations and operationalizations that allow for a more nuanced and useful understanding of online bystander reactions.

We believe the core value of our framework relates to the way in which it encourages researchers to think through and explicitly discuss the conceptual choices that shape the consistency and generalizability of their findings. We have sought to do so with the OBRA framework by synthesizing conceptualizations scattered throughout the literature on the subject, as it relates to thresholds to action, orientation of communication, and related levels of pro-sociality. This critical reorganization of the literature has highlighted two features of online bystander reactions on social media that warrant special attention.

First, the OBRA framework makes explicit that online bystander inactivity is a potential reaction in its own right. *Bystander passivity* can, at a theoretical level, reflect a range of motivations, and assuming that the absence of a visible online bystander reaction equals absence of engagement with or even acceptance of the hateful content, is likely to overlook important aspects, drivers, and potentially pro-social effects of the strategy. Utilizing our framework thus encourages future research designs that are better able to pick up on diverse motivations for, and effects of, online bystander inactivity. Alternatively, it highlights potential limitations to the external validity of research findings that rest on dichotomous conceptions of bystander inactivity. In addition, the framework points to research questions for further study related to motivations and effects of deliberate bystander passivity and non-public bystander reactions on social media: What motivates online bystanders to abstain from publicly visible reactions to hate speech? Do online bystanders react in non-visible ways, e.g., through private messaging or in offline fora? What are the effects of bystander passivity, and is it always undesirable (not pro-social)? And is the level of pro-sociality of bystander inactivity perhaps perceived differently depending on the receiver (perpetrator, victim, onlookers)?

Second, the OBRA framework promotes explicit and detailed consideration of the level of *pro-sociality* inherent to different online bystander reactions, especially as regards publicly visible written bystander reactions (*e.g.*, counterspeech). To treat all online bystander reactions as inherently pro-social may be especially problematic for studies seeking to foster crowd-moderation through, for instance, promoting counterspeech or

user-reporting as an alternative strategy to top-down content moderation. Given that the stated objective of such interventions is to improve online deliberation by either reducing the experienced harm or the amount of hostile content online, it is of special importance to avoid inadvertently adding fuel to the fire by unwittingly promoting hostile counterspeech. From this also follows a set of separate research questions for future research: Are some types of counterspeech generally perceived as more pro-social? How prevalent are hostile counterspeech messages and other hostile bystander reactions? Does the level of perceived pro-sociality depend on who the bystander reaction targets (perpetrator, victim, onlookers)? Specifically in relation to studies focused on fostering crowd-moderation through mobilizing further bystander reactions: Do some bystander interventions, e.g., training and awareness campaigns, unwittingly foster hostile counterspeech, and if so at what ratio to pro-social counterspeech?

In sum, we contend that if we are to harvest the potential of crowd-sourced moderation, research within the field first needs to find common conceptual ground. Only then are we able to explore how we can successfully encourage more people to react pro-socially, and not least assess whether such reactions bring about a better and less hostile online debate environment.

References

Álvarez-Benjumea A and Winter F (2018) Normative Change and Culture of Hate: An Experiment in Online Environments. *European Sociological Review* 34(3): 223–237.

Andresen MJ, Karg STS, Rasmussen SHR, et al. (2022) *Danskernes oplevelse af had på sociale medier*. 9 June. Available at: https://pure.au.dk/portal/files/271291115/Danskernes_oplevelse_af_had_pa_de_sociale_medier_Rapport_Aarhus_Universitet_.pdf.

Atske S (2021) Americans and 'Cancel Culture': Where Some See Calls for Accountability, Others See Censorship, Punishment. In: *Pew Research Center: Internet, Science & Tech*. Available at: https://www.pewresearch.org/internet/2021/05/19/americans-and-cancel-culture-where-some-see-calls-for-accountability-others-see-censorship-punishment/ (accessed 24 January 2023).

Bastiaensens S, Vandebosch H, Poels K, et al. (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.

Brewer G and Kerslake J (2015) Cyberbullying, self-esteem, empathy and loneliness. *Computers in Human Behavior* 48: 255–260.

Buerger C (2021) #iamhere: Collective Counterspeech and the Quest to Improve Online Discourse. *Social Media + Society* 7(4): 205630512110638.

Carlson TN and Settle JE (2022) *What Goes Without Saying: Navigating Political Discussion in America*. 1st ed. Cambridge University Press. Available at: https://www.cambridge.org/core/product/identifier/9781108912495/type/book (accessed 1 December 2022).

Celuch M, Latikka R, Oksa R, et al. (2023) Online Harassment and Hate Among Media Professionals: Reactions to One's Own and Others' Victimization. *Journalism & Mass Communication Quarterly* 100(3). SAGE Publications Inc: 619–645.

Constine J (2017) Facebook 'Snooze' button temporarily hides people in your feed. In: *TechCrunch*. Available at: https://techcrunch.com/2017/09/14/facebook-snooze/ (accessed 18 October 2022).

Costello M, Hawdon J and Cross A (2016) Virtually Standing Up or Standing By? Correlates of Enacting Social Control Online. *International Journal of Criminology and Sociology* 6: 16–28.

Craker N and March E (2016) The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences* 102: 79–84.

Crawford K and Gillespie T (2016) What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18(3). SAGE Publications: 410–428.

DataReportal (2022) *Digital 2022 Global Overview Report*. Available at: https://datareportal.com/reports/digital-2022-global-overview-report (accessed 30 May 2022).

Duggan M (2017) 3. Witnessing online harassment. In: *Pew Research Center: Internet, Science & Tech*. Available at: https://www.pewresearch.org/internet/2017/07/11/witnessing-online-harassment/ (accessed 1 December 2022).

Elassar A (2020) Gay men have taken over the Proud Boys Twitter hashtag. Available at: https://www.cnn.com/2020/10/04/us/proud-boys-twitter-hashtag-gay-men-trnd/index.html (accessed 16 August 2022).

Fischer P, Krueger JI, Greitemeyer T, et al. (2011) The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin* 137(4). American Psychological Association: 517–537.

Garland J, Ghazi-Zahedi K, Young J-G, et al. (2020a) Countering hate on social media: Large scale classification of hate and counter speech. Epub ahead of print 2 June 2020.

Garland J, Ghazi-Zahedi K, Young J-G, et al. (2020b) Impact and dynamics of hate and counter speech online. Epub ahead of print 16 September 2020.

Gillespie T (2018) *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press.

Hangartner D, Gennaro G, Alasiri S, et al. (2021) Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences* 118(50). National Academy of Sciences.

Hansen T, Lindekilde L and Rasmussen J (2023) Standing Up to Hate(rs): Exploring the Motivations of Online Counter Speakers. *Working paper, Aarhus University*. Epub ahead of print 2023.

Jost P, Ziegele M and Naab TK (2020) Klicken oder tippen? Eine Analyse verschiedener Interventionsstrategien in unzivilen Online-Diskussionen auf Facebook. *Zeitschrift für Politikwissenschaft* 30(2): 193–217.

Kalch A and Naab TK (2017) Replying, disliking, flagging: How users engage with uncivil and impolite comments on news sites. *Studies in Communication | Media* 6(4): 395–419.

Kozyreva A, Herzog SM, Lewandowsky S, et al. (2023) Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences* 120(7). Proceedings of the National Academy of Sciences: e2210666120.

Kunst M, Porten-Cheé P, Emmer M, et al. (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(3): 258–273.

Leidner DE (2018) Review and Theory Symbiosis: An Introspective Retrospective. *Journal of the Association for Information Systems* 19(6). Atlanta, United States: Association for Information Systems: 552–567.

Leonhard L, Rueß C, Obermaier M, et al. (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(4): 555–579.

Liebst LS, Ejbye-Ernst P, Dausal KL, et al. (2018) Bystanders in Real-Life Dangerous Emergencies: Group Relationships Predict Intervention. In: *Dansk Sociologikongres*, Esbjerg, Denmark, 2018.

Lu S and Luqiu LR (2023) When Will one Help? Understanding Audience Intervention in Online Harassment of Women Journalists. *Journalism Practice* 0(0). Routledge: 1–19.

March E (2016) 'Don't feed the trolls' really is good advice – here's the evidence. Available at: http://theconversation.com/dont-feed-the-trolls-really-is-good-advice-heres-the-evidence-63657 (accessed 24 May 2022).

Mathew B, Saha P, Tharad H, et al. (2019) Thou shalt not hate: Countering Online Hate Speech. In: *arXiv:1808.04409 [cs]*, Munich, 4 April 2019, pp. 369–380. Available at: http://arxiv.org/abs/1808.04409 (accessed 28 January 2022).

Moxey N and Bussey K (2020) Styles of Bystander Intervention in Cyberbullying Incidents. *International Journal of Bullying Prevention* 2(1): 6–15.

Munger K (2017) Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior* 39(3): 629–649.

Mutz DC (2008) Is Deliberative Democracy a Falsifiable Theory? *Annual Review of Political Science* 11(1): 521–538.

Naab TK, Kalch A and Meitz TG (2018) Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society* 20(2). SAGE Publications: 777–795.

Nelson JK, Dunn KM and Paradies Y (2011) Bystander Anti-Racism: A Review of the Literature. *Analyses of Social Issues and Public Policy* 11(1): 263–284.

Obermaier M, Schmuck D and 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*. SAGE Publications: 14614448211017527.

Ozalp S, Williams ML, Burnap P, et al. (2020) Antisemitism on Twitter: Collective Efficacy and the Role of Community Organisations in Challenging Online Hate Speech. *Social Media + Society* 6(2). SAGE Publications Ltd: 2056305120916850.

Pang N, Ho SS, Zhang AMR, et al. (2016) Can spiral of silence and civility predict click speech on Facebook? *Computers in Human Behavior* 64: 898–905.

Pew Research Center (2021) *Social Media Use in 2021*. April. Pew Research Center.

Porten-Cheé P, Kunst M and Emmer M (2020) Online Civic Intervention: A New Form of Political Participation Under Conditions of a Disruptive Online Discourse. *International Journal of Communication* 14(0). 0: 21.

Reynders D (2021) 6th evaluation of the Code of Conduct. European Commission. Available at: file:///C:/Users/au702709/Downloads/2021_10_07_Factsheet_Code_of_Conduct_SPP.pdf.pdf (accessed 30 May 2022).

Rudnicki K, Vandebosch H, Voué P, et al. (2022) Systematic review of determinants and consequences of bystander interventions in online hate and cyberbullying among adults. *Behaviour & Information Technology* 0(0). Taylor & Francis: 1–18.

Sellars A (2016) Defining Hate Speech. *SSRN Electronic Journal*. Epub ahead of print 2016. DOI: 10.2139/ssrn.2882244.

Siegel AA and Badaan V (2020) #No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online. *American Political Science Review* 114(3). Cambridge University Press: 837–855.

Taylor J (2021) Facebook now lets users and pages turn off comments on their posts. *The Guardian*, 31 March. Available at: https://www.theguardian.com/media/2021/mar/31/facebook-turn-off-comments-on-post-limit-restrict-disable-comment-posts-moderation-control-tool (accessed 18 October 2022).

Vogels EA (2021) The State of Online Harassment. In: *Pew Research Center: Internet, Science & Tech*. Available at: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/ (accessed 1 December 2022).

Watson BR, Peng Z and Lewis SC (2019) Who will intervene to save news comments? Deviance and social control in communities of news commenters. *New Media & Society* 21(8). SAGE Publications: 1840–1858.

Wilhelm C, Joeckel S and Ziegler I (2020) Reporting Hate Comments: Investigating the Effects of Deviance Characteristics, Neutralization Strategies, and Users' Moral Orientation. *Communication Research* 47(6): 921–944.

Wong RYM, Cheung CMK, Xiao B, et al. (2021) Standing up or standing by: Understanding bystanders' proactive reporting responses to social media harassment. *Information Systems Research* 32(2): 561–581.

You L and Lee Y-H (2019) The bystander effect in cyberbullying on social network sites: Anonymity, group size, and intervention intentions. *Telematics and Informatics* 45: 101284.

Young R, Miles S and Alhabash S (2018) Attacks by Anons: A Content Analysis of Aggressive Posts, Victim Responses, and Bystander Interventions on a Social Media Site. *Social Media + Society* 4(1). SAGE Publications Ltd: 2056305118762444.

Notes

[1] Intervention understood in the dictionary sense as "the action of getting involved in a situation in order to improve it or stop it from getting worse" (Oxford Learner's Dictionary of Academic English)
[2] Five of these references were added based on reviewer input.
[3] For a discussion of the potential beneficial impacts of silence in the face of online hostility see page 8-9.
[4] The term reaction-button refers to inbuilt icons or buttons, e.g., emojis, that enable social media users to signal their opinion on a post to other social media users.
[5] *E.g.*, other readers of a post and comment section.
[6] Similarly, perpetrators and victims may also receive such messages offline or through other online messaging sites, *e.g.,* via email.
[7] Mathew et al. (2019: 3), however find a surprisingly high rate of counterspeech (49.5%) when scraping Youtube comments to videos including hateful content targeting Jews, Blacks and LGBT communities, very unlike the finding of only 13% by Garland et al. on Twitter (2020b: 7).
[8] Individuals engaged in "a form of online bullying and harassment (Pew Research Centre, 2014), common trolling behaviour includes starting aggressive arguments (Klempka & Stimson, 2013) and posting inflammatory malicious messages in online comment sections to deliberately provoke, disrupt, and upset others (Gammon, 2014)." (in Craker and March, 2016: 79).
[9] For a general account of silence in political discussions see Carlson and Settle 2022.
[10] Reporting content to the authorities is likely to present a higher threshold to action than many other reaction options. Such reports require extensive documentation and carries a potential for disclosure of action, when reports are relayed to the perpetrator during police investigations.
[11] Theoretically, counterspeech can also occur in non-visible private forums, yet the existing literature primarily focuses on the public manifestations.
[10] 'Verbal' as understood in its dictionary sense: "relating to or in the form of words" (Oxford's English Dictionary).
[11] In theory, the same reflections are relevant when thinking about comments that contain graphic elements (*e.g.,* gifs, emojis, Memes)
[12] Counterspeech may, for instance, contribute to further experiences of trauma on the part of victims, as the messages at times inadvertently bring further attention (algorithmically and otherwise) to the original hostile message.