Article

SM+S
social media + society

Social Media + Society
April-June 2026: 1–14

DOI: 10.1177/20563051261437497
journals.sagepub.com/home/sms
Sage

Bystanders and Reporters: Who Acts Against Illegal Online Content?

Friederike Quint[1], Yannis Theocharis[1], Spyros Kosmidis[2], and Margaret E. Roberts[3]

Abstract

Harmful and illegal content on social media is widespread, but what should be taken down is widely disputed, creating ongoing challenges for resolving the tension between free speech and user safety. User reporting is a key mechanism for addressing such content, yet little is known about who reports, what motivates them, and how they compare to the general population. We study these questions using two datasets: (1) a unique survey with individuals verified to have previously reported potentially illegal content to a third-party organization in Germany and (2) a quota-based sample approximating the German population. We show that individuals who have previously reported potentially illegal content via a third-party reporting service represent a distinct, civically engaged subset of users. They tend to be older, more often men than women, highly educated, highly politically active, and markedly left-leaning. They are not politically representative of the German population and take a distinctly different position when balancing free speech and protection from harm, putting more emphasis on protecting from harm. Reporting users' motivations appear primarily civic-minded rather than reactive, especially among those who do it frequently and those intervening on behalf of others. These insights highlight reporting as a form of digital civic participation and offer perspectives relevant for understanding political engagement online, platform governance, user agency, and trust and safety regulation.

Keywords

content moderation, social media, user reporting, trusted flagger, political behavior

Introduction

Harmful content remains widespread on social media; however, what content should be removed is disputed, posing challenges for resolving the tension between free speech and user safety. To combat harmful – and sometimes illegal – content, platforms have traditionally relied not only on automated detection and human moderators but also on user flagging and reporting (Singhal et al., 2023). User-centered moderation, one of the earliest forms of moderation on social media (York, 2022), has increasingly moved to the forefront – epitomized by X's (Twitter's) 2021 Birdwatch program, later rebranded on X as Community Notes. The shift accelerated in January 2025, when Meta's CEO Mark Zuckerberg announced that the company would end U.S. fact-checking and rely increasingly on users reporting content before action is taken by Meta (Isaac & Schleifer, 2025). Coupled with Meta's termination of third-party fact-checking – reversing years of prioritizing automation (Meta, 2020) – these moves underscore renewed reliance and higher responsibilities on users.

User reporting seems democratic in nature, but limited flagging options and opaque enforcement often leave users frustrated and doubtful that their reports matter (Crawford & Gillespie, 2016; Jhaver & Zhang, 2023). Many users abandon reports when categories are unclear or outcomes uncertain, reinforcing perceptions of arbitrariness (Blackwell et al., 2018; DiFranzo et al., 2018). In response to these limitations, some countries have established independent third-party reporting portals that allow citizens to report abusive or potentially illegal content outside of platform environments.

[1]Technical University of Munich, Germany
[2]Oxford University, UK
[3]University of California, San Diego, USA

Corresponding Author:
Friederike Quint, Department of Governance, Munich School of Politics and Public Policy, Technical University of Munich, Richard-Wagner-Straße 1, Munich 80333, Germany.
Email: friederike.quint@tum.de

Reporting has furthermore drawn at times significant criticism. For example, platform mechanisms are criticized for legitimizing takedowns based on single reports that may not reflect majority views (Crawford & Gillespie, 2016), fueling debates over unequal treatment and uneven moderation across groups (Kemp & Ekins, 2021). Users can also exploit reporting for strategic flagging to silence opponents (Crawford & Gillespie, 2016). These dynamics reveal a key tension: reporting can safeguard users but also risk being weaponized in ways that undermine free speech and democratic discourse.

Despite the central role reporting plays in platform accountability – and the costs and uncertainties tied to enforcement – we still know remarkably little about users who report content. Evidence from Germany suggests that just over half of politically active individuals have reported content directly to a platform (Koch et al., 2025), compared with 34% of the general population, while only a small minority – around 5% – report content via third-party organizations (Brennauer et al., 2024). Yet it remains unclear who these reporters are, how they differ from the broader public, and what motivates them to take action.

A key obstacle to answering these questions is the lack of reliable ground truth. Platforms do not systematically disclose data on reporting behavior, making it difficult to assess who reports content, how frequently, or under what conditions. As a result, existing research relies heavily on self-reported survey measures or platform-specific case studies, which limit our ability to characterize reporters and to link reporting behavior to broader patterns of political engagement, attitudes, and trust.

To address these gaps, we adopt an alternative empirical strategy that focuses on reporting through a third-party portal. We draw on two complementary surveys: a novel dataset of verified individuals ($n = 699$) who have previously reported potentially illegal content via *REspect!* (here defined as individuals identified through reporting to the third-party organization *REspect!*), a German civil society organization that also serves as the country's first "Trusted Flagger" under the Digital Services Act (DSA), and a sample drawn from the online panel of Bilendi ($n = 1353$) that used quotas to approximate Germany's general population.

Studying reporting via *REspect!* offers a rare opportunity to examine confirmed reporting behavior in a context where reports are institutionally recognized, externally verifiable, and embedded within an emerging regulatory framework. Using these data, we compare confirmed reporters (henceforth referred to as "reporting users") to the general population across a range of characteristics, including demographics, reporting behavior, political and civic engagement, attitudes, social media experiences, and levels of trust. Finally, we examine the motivations associated with reporting behavior among individuals who have actively chosen to intervene by submitting reports through a third-party mechanism.

Our findings reveal that reporting is performed by a distinct and civically engaged subset of the public. Reporting users tend to be older, more often men than women, and highly educated, and – relative to the general population – include a substantially overrepresented share of non-binary and "other" respondents. They are also highly politically active, more left-leaning, value protection from the harm speech may cause, and differ from the general population in their attitudes, digital behaviors, and motivations. While some respondents cite personal experiences with hate speech as their main motivation for reporting content in general, the majority report being motivated by a sense of civic responsibility or concern for the integrity of online spaces, especially among frequent reporters or when harm affects others. The primary motivation for turning to an independent reporting portal such as *REspect!* is the perceived ineffectiveness of platform-based reporting and enforcement of illegal content, pointing to deeper grievances, declining trust in social media platforms, and greater trust in third-party organizations.

Content Moderation and User Reporting

Platforms employ a range of content moderation mechanisms to address harmful and illegal content. While much of this content is removed through algorithmic detection, moderation also involves human review – either in combination with automated systems or independently (Crawford & Gillespie, 2016; Singhal et al., 2023). In addition, platforms rely on users to report or "flag" content they believe violates platform policies or national laws. To facilitate this, they generally provide reporting tools and publish guidelines to help users identify violative content. However, these guidelines are often excessively long and complex, especially on very large platforms, making them difficult to understand – with some platforms – usually smaller ones not providing guidelines at all (Nahrgang et al., 2025).

After years of emphasis and investment in automated content moderation, user reporting moved back to center stage in January 2025, when Mark Zuckerberg announced that Meta would end its partnership with third-party fact-checkers in the United States and, for lower-severity violations, would begin to rely more on "someone reporting an issue before [Meta] takes action," framing this shift as a return to free expression (Meta, 2025). While the European context differs, owing to stricter legal frameworks such as the DSA, user reporting remains a central pillar of enforcement. Its importance stems from the limitations of algorithmic and human moderation, which are prone to error, failing to detect harmful or illegal content.[1] In such cases, reporting becomes a crucial fallback mechanism, with platforms often relying on user reporting to enforce their rules and to improve algorithmic training and inform human review processes (Crawford & Gillespie, 2016). Yet this reliance is contested: systems are reactive rather than preventive (Crawford & Gillespie, 2016; Douek, 2022), vulnerable to coordinated "strategic" flagging, and for being far more complex than a simple binary user decision, as they constitute an "interaction between users, platforms,

humans, and algorithms, as well as broader political and regulatory forces" (Crawford & Gillespie, 2016, p. 411).

The European Union's (EU) DSA, adopted in 2022, is the EU's new regulatory framework for online platforms. It sets out obligations for risk management, transparency, and accountability in content moderation and pivots from strict takedown mandates to transparency duties scaled by platform size (European Union, 2022). Crucially, transparency here concerns not just published rules but their implementation through moderation mechanisms such as user reporting. However, as Busch notes, the DSA "abstains from the difficult task of drawing a line between legal and illegal content" (Busch, 2022, p. 53), leaving much discretion to private actors in defining such speech. To address this, the DSA created "Trusted Flaggers" – recognized organizations whose reports are prioritized (European Union, 2022). *REspect!* supports victims, forwards potentially criminal content to law enforcement, and campaigns against online hate. These frameworks shape reporting workflows, making flagging a legally and politically consequential act.

Viewing reporting as user agency within content moderation, its value for safeguarding public discourse hinges on whether – and how – people use these mechanisms. What motivates users to intervene and who chooses to report are therefore central questions: do reporters mirror the public, or are they skewed by attitudes and demographics? Answering this is essential to assess whether reporting strengthens democratic participation and accountability online or risks amplifying existing biases and inequalities.

Reporting, Participation, and Online Behavior

A useful starting point is to examine how users respond to potentially illegal and harmful content online. On large platforms, low identifiability and diffusion of responsibility reduce the perceived obligation and social pressure to act – people expect others to intervene or view the issue as outside their remit (Aleksandric et al., 2022; Blackwell et al., 2018; Obermaier et al., 2016). Such reluctance, and the sense that harmful or illegal content is inevitable, signals both normalization of harm and resignation about platforms' capacity or will to enforce rules (Theocharis et al., 2025).

Possible user responses to potentially illegal and harmful content span a spectrum: low engagement actions such as unfollowing, muting, or blocking; reporting of content or accounts (for Community Guidelines violations or alleged illegality); and counterspeech, either publicly or via private messages. Compared with reporting, counterspeech typically carries higher costs because it requires direct engagement – often in public, exposing users to potential confrontation or backlash. Nonetheless, empirical work documents modest but reliable effects of empathy-based appeals, especially when individuals connect their own experiences as targets to their behavior toward outgroups (Gennaro et al., 2025; Hangartner et al., 2021), and shows that identity-based counterspeech can significantly reduce online hate (Munger, 2017; Siegel & Badaan, 2020). Related evidence indicates that empathy increases public intervention against cyberbullying; however, the effects are more pronounced when interpersonal similarity is low (Wang, 2021).

Intervention hinges on perceived threat, moral judgment, and identity-linked norms. Moral outrage, often triggered by perceived victimhood, both fuels virality by increasing sharing and motivates corrective action (Gray & Kubin, 2024). Users are more likely to intervene when they believe enforcement is transparent and fair (Shim & Jhaver, 2024); among young German adults, personal responsibility and empathy raise direct and indirect intervention (Obermaier, 2024; Wang, 2021). Non-intervention commonly reflects limited knowledge of reporting, uncertainty about illegality, low interest, or doubts about impact (Böswald et al., 2025; Doseva et al., 2024).

Design choices further shape intervention. Making moderators who flagged content visible can deter bystanders and suppress peer enforcement (Bhandari et al., 2021), whereas exposure to bystander interventions and other norm-conforming responses can spur action (Blackwell et al., 2018). Users favor low-cost options like reporting over higher-cost counterspeech (Hansen et al., 2024); if reporting is ambiguous or opaque, perceived costs rise, and bystanding persists. Beyond psychology and platform design, reporting is political, moral, and normative; unlike visible engagement or counterspeech, it is private, platform-directed, and procedurally mediated (Crawford & Gillespie, 2016; Suzor, 2019).

Finally, political identity and moral priorities shape both action and inaction. Partisans often agree on which hate speech is most censorable but misperceive the other side's priorities, fueling conflict over moderation (Solomon et al., 2024). Debates over content moderation actions mirror these divides, with both left and right alleging bias (Appel et al., 2023; Corduneanu-Huci & Hamilton, 2022; Kemp & Ekins, 2021; Vogels et al., 2020). Left-leaning users generally advocate stricter platform rules, while others reject content moderation as censorship or hold back when the statements align with their identity or target lower-priority outgroups (Kozyreva et al., 2023; Munzert et al., 2025; Pradel et al., 2024; Rasmussen, 2022). Thus, intervention hinges on whether users expect their action to matter, at what personal and social cost, and in light of the norms and identities that make some speech, and some targets, feel more deserving of intervention than others.

Understanding who remains a bystander versus who reports helps illuminate how citizens engage, assume responsibility, and exercise agency online. Reporting operates as delegated governance, flagging content against platform norms, democratic values, and law (Gillespie, 2018). Framed as civic responsibility and personal agency, it is constrained by opaque processes, limited feedback, and strategic misuse,

exposing tensions between safety and free expression (Theocharis et al., 2025). As a core moderation mechanism, reporting can both enable meaningful user participation and safeguard speech through due-process safeguards (clear standards, transparency, accessible notice-and-appeal). Realizing this potential requires knowing who reports and why. We therefore ask: Who reports content, how do reporters differ from the general population, and what motivates them to report?

Data Collection

To identify who reports potentially illegal content and why, we use a unique dataset of confirmed reporters to the German third-party portal *REspect!*.[2] The survey link was sent via auto-reply and was accessible only through an anonymous URL shown to those who filed a report. Data were collected from May 9, 2024 to July 31, 2025.[3] We only included respondents 18 years or older.

Following public announcements on the "Trusted Flaggers," targeted attacks against the reporting portal occurred during our study, and the link to the survey became one target of these attacks. To safeguard data integrity and only include true reporters, we applied predefined exclusion rules (A2, Supplementary Material, hereinafter SM), removing 156 inauthentic responses and yielding a final sample of $n = 699$. For our analysis, we include both those who passed and failed the attention checks, and we examine our results by attention check status (A5.7, SM). While voluntary, uncompensated participation limits external validity, adjusted monthly opt-in rates of 5%–11% are stable over time (A3, SM).[4]

To contextualize reporting users relative to the general population, we draw on a quota-based survey of the German population aged 18 and over ($n = 1353$; October–November 2024), fielded as part of a larger cross-national project conducted in 10 countries. For this study, the German subsample serves as a reference population, with overlapping items enabling direct comparison to the survey of reporting users. Data collection was coordinated by the polling firm Bilendi & Respondi in collaboration with the authors, who developed the survey instrument.[5] To further contextualize sociodemographic differences, we compare our samples to official census data for the German population (SM A1.1).

While quotas and post-stratification weights were used to approximate the German population, the sample is not fully representative. It includes only respondents identifying as male or female, covers an age range of 18 to 69, and slightly underrepresents individuals with high levels of education; these deviations reflect adjustments made at the survey provider's request after the relevant quota could not be fully filled.

We first profile reporting users – their sociodemographics, social media experiences, and civic–political engagement, among others. We then compare them to the German survey and German census data to assess whether they form a distinct subgroup. Finally, we analyze motivations to report, including subgroup differences, and open-text answers to the question of what motivated respondents to report to *REspect!*.

Measurement

Sociodemographic Measures. Our main points of comparison are sociodemographic characteristics, a classic determinant of politically active citizens (Verba & Nie, 1987). To analyze the differences between reporting users and the general population, we look at the relative differences concerning the age, gender, and education of respondents.[6]

Reporting Behavior. We measure reporting users' reporting behavior by looking at the shares of previous reporting behavior to *REspect!* and in general to platforms or other websites. In addition, we look at the shares of platforms/websites the content was reported to, classifications of the content that was reported, and, finally, who the target was.

Political Measures. We include political behaviors and attitudes relevant to civic engagement and reporting: political participation, trust in institutions and media, ideology, satisfaction with democracy, political interest, and experiences with harmful content that may shape preferences and actions (Vogels et al., 2020).[7] We further examine attitudes toward freedom of speech, central for how users engage with platform content, and views on content moderation and engagement with its mechanisms (Howard, 2019; Theocharis & De Moor, 2021; Theocharis et al., 2025).

Motivation to Report. Prior work links online intervention to civic-mindedness and personal experience, among other motives (Brennauer et al., 2024; Delmas, 2018; Gennaro et al., 2025; Koch et al., 2025; Porten-Cheé et al., 2020). For respondents' motivations to report in general, we offered: "Concern for the online community," "Personal experience with hate speech," "Advocacy for a specific cause," "My civic duty," "It is a bad way to treat people," and "Other/I'm not sure/Prefer not to say."[8] We then recoded selections into four categories: Prosocial Motives ("Concern for the online community," "It is a bad way to treat people"), Personal Experiences ("Personal experience with hate speech"), Civic Motivation ("Advocacy for a specific cause," "My civic duty"), and Other.

To gain a more detailed view of respondents' motivations for reporting to *REspect!*, we asked an open-text question: "Even though there is a reporting mechanism on the website where you encountered the content, why did you report it via *REspect!*'s website?" The survey was bilingual (German/English). After removing NAs ($n = 699$), we analyze 694 responses; 22 English responses were translated into German, resulting in a unified German text variable ("respect

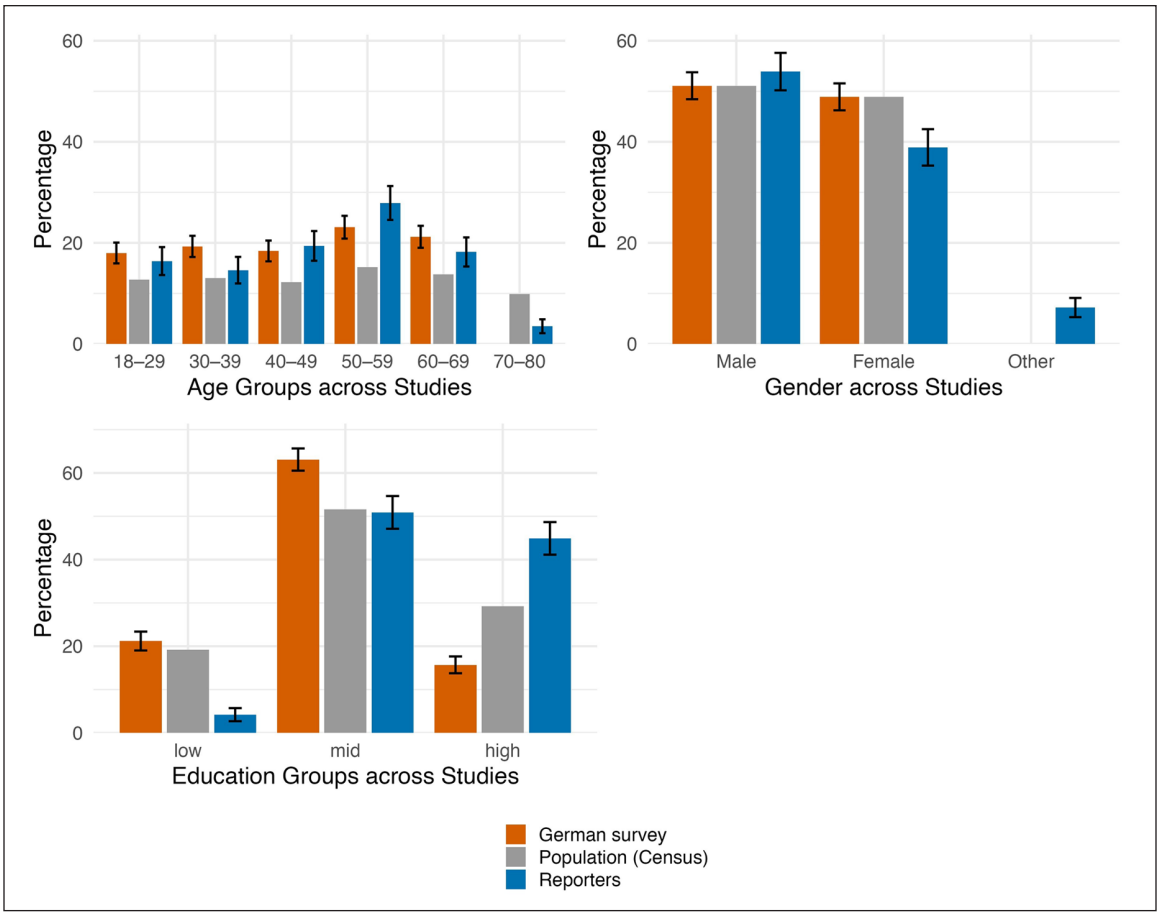

Figure 1. Bar plots with error bars showing the relative frequencies of demographic characteristics across the two samples are further benchmarked against official census data for the German population. The German comparison sample was fielded using predefined age quotas (with a top-coded upper age category) and weighted to approximate the general population within these categories.

motivation"). For measurement details, see A4 and A5, SM; key items are briefly restated in each subsection.

Results

Given the limited existing research on reporting users, our analytical strategy is descriptive and correlational. We characterize reporting users' political attitudes, behaviors, and motivations in comparison to the German public, using an inductive approach to identify empirical patterns that may inform future theory-building and causal research. We begin by descriptively characterizing reporting users and their reporting behavior in comparison to the German population.

Who Are the Reporting Users and What Do They Report?

Sociodemographic Profiles. Figure 1 compares the demographic profiles of reporting users, the German census, and the broader quota-based survey conducted in Germany. The majority of confirmed reporters fall within the 40 to 59 age brackets, with an average age of 47 years. This average closely matches the German survey for the same age range (46 years) and is comparable to the overall German population average (45 years). The majority of reporting users are male (54%), while 39% are female and 7% identify as "Other" – a category that includes the options "Other," "Transgender," "Non-binary," and "Prefer not to say." While men are the majority within the reporting users group, the most pronounced gender difference concerns the higher prevalence of non-binary or other gender identities among reporting users.

Reporting users are highly educated: 45% hold a high level of education (at least a bachelor's degree), and 51% fall into the mid-level category (see Table A1, SM). Compared to the general population, those who report content to *REspect!* tend to be slightly older on average, with the strongest representation among middle-aged cohorts, and are more likely to be male and highly educated. While broadly comparable to the German survey, reporting users appear more highly educated than in the German survey; this gap is partly driven by the underrepresentation of highly educated respondents in

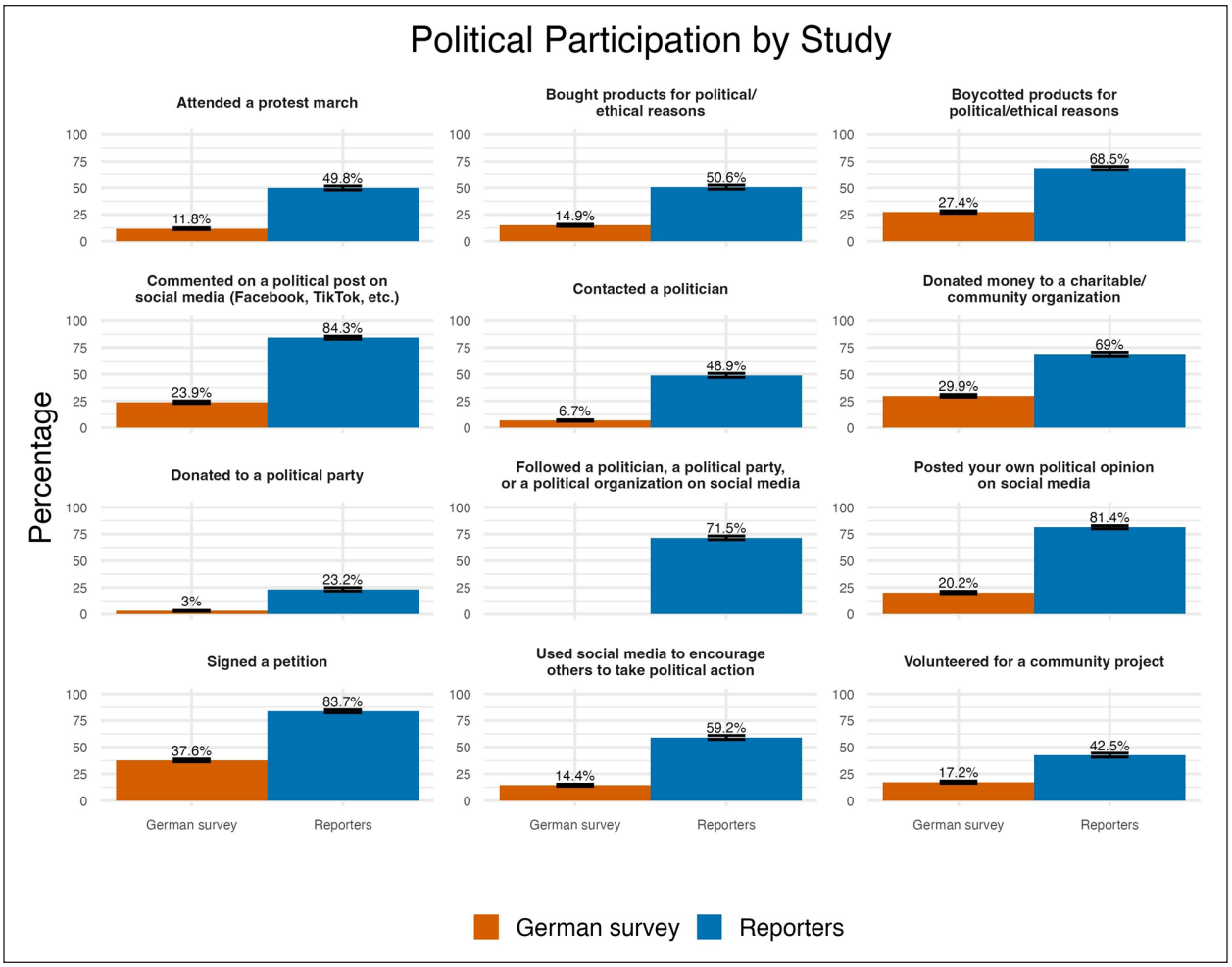

Figure 2. Bar plot with error bars showing the percentage of having done each of the options during the last 12 months, answering either with "Yes" or "No" for each sample. The percentages for the German survey are weighted to approximate the general population.

the German survey, which is potentially due to unfilled quotas during fieldwork.[9]

Users' Reporting Behavior. We examine reporting frequency, where content was encountered, its type, and target-relatedness. Most respondents reported to *REspect!* once, while platform/site reporting was more common – often 10 or more times – although a sizable minority reported only once (Figure A3, SM). Reported content was most often seen on Facebook (42%), X/Twitter (27%), and Instagram (12%) (Figure A4, SM). The most frequently reported type was intolerance (53%; e.g., discrimination/dehumanization), followed by threats of violence and offensive language (Figure A5, SM). Regarding the targets of the reported content, the largest share concerned outgroups to which respondents did not belong (39%), with fewer personal attacks (18%) or incidents involving acquaintances (2%) (Figure A6, SM).

Political Attitudes and Behavior. Although there is limited evidence on individuals who report content to social media platforms, existing research suggests that politically active individuals tend to report more frequently than the general public (Brennauer et al., 2024; Koch et al., 2025). It is therefore reasonable to assume that people in our sample may also exhibit higher levels of political engagement compared to the broader population (Figure 2).[10] Consistent with previous work, we find that reporters exhibit higher levels of political engagement compared to the broader population.

Reporting users are markedly more politically active than the German population. As Figure 2 shows, 81% have posted political opinions and 84% have commented on political posts; 72% follow politicians or organizations on social media. Furthermore, 84% have signed a petition, 43% volunteered for a community project, and 50% attended a protest. These gaps underscore that *REspect!* Reporters are highly active both online and offline.

Notably, reporting users are overwhelmingly left-leaning compared to the German population (Figure 3). Their ideological distribution is sharply concentrated at the left or liberal end of the spectrum, whereas the German reference sample is more centrally distributed. The magnitude and direction of this divergence point to a substantively meaningful ideological

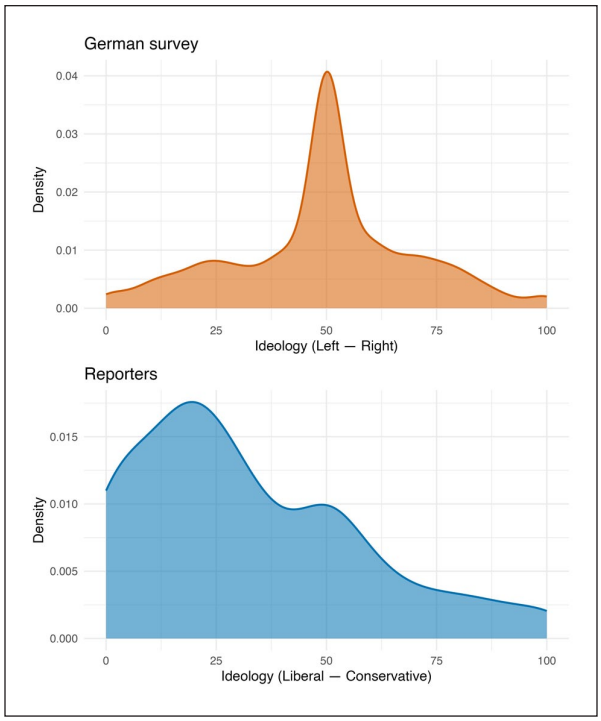

Figure 3. Density curves of ideology measured on a scale from 0 = "Left/Liberal" to 100 = "Right/Conservative." The German scale uses "Left/Right"; the reporters survey uses "Liberal/Conservative." The German survey is weighted to approximate the general population; reporters are unweighted.

imbalance among reporting users, complementing their high levels of political engagement.

Given their high political engagement, one might expect reporting users to also be more trusting of democratic institutions. Interestingly, our sample of reporting users displays similar levels of institutional trust to those of the general population (Figure A2, SM). Moreover, reporting users trust the police less and the national parliament more. We see a sharper contrast in media trust: reporters are less trusting of traditional media and show the lowest trust in social media platforms. At the same time, they report higher satisfaction with democracy and very high political interest. Overall, reporting users emerge as civically and politically engaged, ideologically skewed to the left, digitally active, and more critical of platforms as institutional actors.

Experiences Online. We examine whether reporters' behavior reflects harsher online experiences. Compared to the German public, reporting users report substantially more negative experiences (Figure A8, SM A5.4) – both witnessing and suffering attacks based on opinions or identity, as well as engaging in self-censorship due to feared reactions. They are also more likely to have had posts flagged, labeled, or removed, indicating greater exposure to moderation. An additional item fielded only to reporting users reveals that majorities report being called offensive names and encountering intolerance (both 76%), with over half reporting experiences of physical threats (Figure A10a, SM). As discussed in the SM, these elevated exposure rates reflect the highly selected nature of the reporting user sample rather than the prevalence in the general population. Finally, experiences of online hostility are patterned by gender, with male reporting users reporting higher exposure across most experience categories. Respondents identifying as "Other" also report exposure across several categories, though estimates are based on a smaller group (Figure A9, SM). Overall, reporting users face a markedly harsher online environment, which plausibly feeds into their intervention behavior.[11]

Freedom of Speech Values. Do reporting users, given their harsher online experiences, prioritize protection from harm over free speech? Prior harm and moderation can push attitudes either toward stronger protection or, conversely, toward heightened concern for expression. We find that reporting users tilt toward protection: on a 0–100 scale (0 = "strongly prefer freedom of speech," 100 = "strongly prefer protection from harm"), reporting users score higher than the German public, which leans toward unrestricted expression (Figure A11, SM). This pattern holds across items: reporters are more supportive of limiting harmful speech and misinformation, while the general population places relatively more weight on protecting expression, even when offensive or misleading (Figure A12, SM). In short, reporting users endorse free speech as a democratic value but set clearer limits when speech causes harm, consistent with norms likely motivating their proactive reporting.

Attention Checks. We examined differences between respondents who passed and failed the attention check. Despite age and gender differences in attention-check performance (Figure A18, SM), substantive outcomes are similar across groups (Figure A19, SM); as discussed in the SM, attention checks likely capture variation in digital literacy rather than inattentiveness per se (Guess & Munger, 2023; Munger et al., 2021) (SM A5.7).

Taken together, our findings reveal that, in comparison to the general public, users who reported content to *REspect!* are predominantly older, a majority are male, and highly educated. These individuals also differ markedly from the general population in their political engagement, attitudes, and digital behavior, suggesting that reporting behavior to a third-party organization can be associated with a specific, civically motivated subset of the public rather than a cross-section of society. However, to better understand what motivates individuals to report, we treat motivation as the key outcome variable in the following section.

What Motivates Users to Report Potentially Illegal Content?

To examine heterogeneity in reporting motivations, we classify respondents' stated reasons for reporting into distinct

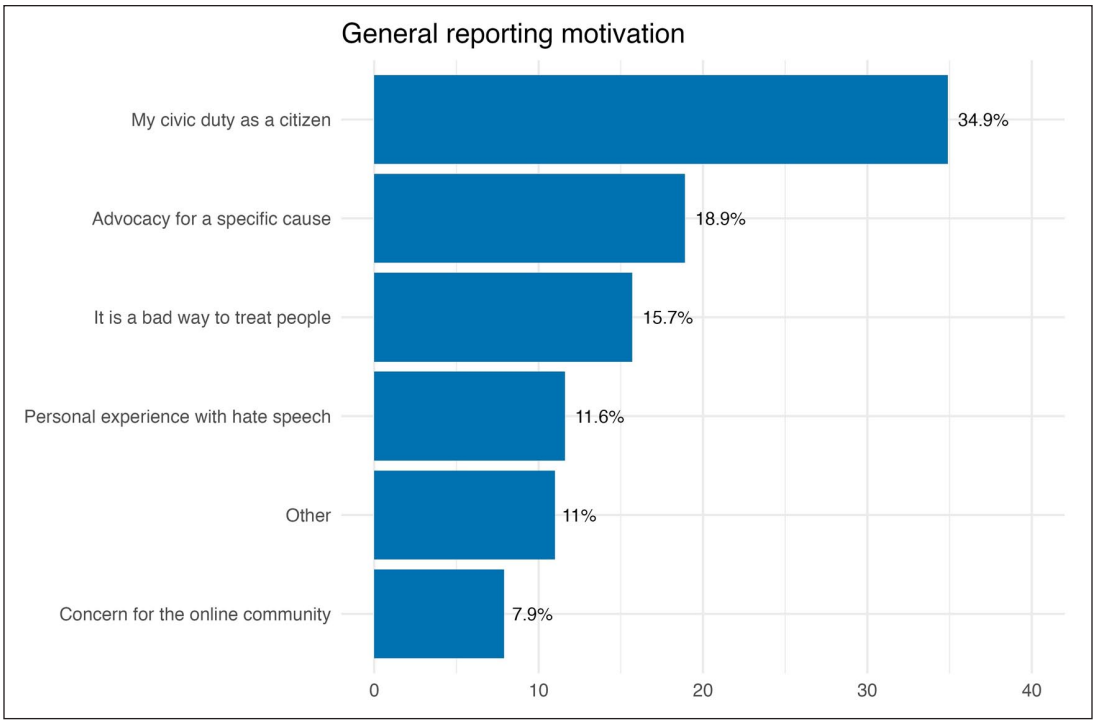

Figure 4. Bar plot with percentages of motivations to report content in general. Respondents were asked "What motivated you to report the content in general?"

outcome categories. Given the unordered categorical structure of this variable, we estimate multinomial logistic regression models. Predictors include demographics (age, gender, education), attitudes (democratic satisfaction, free expression), contextual factors (target, experience index), and reporting behaviors (social media reporting and submission volume).[12] We report model-based predicted probabilities with 95% confidence intervals, holding other covariates constant, and visualize these estimates to summarize associations.

General Motivation to Report. Respondents answered two motivation questions: (i) why they report in general and (ii) why they report to *REspect!*. Most cited civic duty (35%), followed by cause advocacy (19%); relatively few named personal experience (Figure 4). That many suggest they are motivated by advocacy for a specific cause supports arguments that flagging can be strategic (e.g., coordinated by aggrieved groups) (Crawford & Gillespie, 2016), suggesting motivations extend beyond individual grievance.

The fact that reporting users see reporting as their civic duty is consistent with respondents' reported flagging behavior. Most reporting users indicate that they have reported posts or comments multiple times, routinely engage in moderation actions, and regularly block or mute other users or accounts (Figure A7, SM). They are both highly exposed and responsive, actively managing their feeds as civic–protective engagement. In addition, a plurality reports content targeting groups they don't belong to (39%) versus personal victimhood (18%) (Figure A6, SM), signaling solidarity over self-defense. Third-party reporting is thus not merely reactive but proactive and value-driven.

Controlling for covariates, motives vary modestly by demographics. Civic motivation is common regardless of gender, with men only slightly more likely than women to cite prosocial reasons (Figure A16, SM). By age, civic motivation is as high as 60% among the youngest, but only 43% among the oldest (Figure A13, SM); prosocial motives are stable across age, while personal experience and "other" reasons rise slightly with age. Education does not seem correlated with motivation, although lower-educated respondents more often cite personal experience and less often prosocial reasons (Figure A17, SM).

We expected an index of negative social media experiences to predict personal experience motives. Respondents in our reporting user survey report a high frequency of negative experiences online. In total, 76% have been called offensive names, 76% faced intolerance/discrimination, 53% were purposefully embarrassed, and 52% were physically threatened (Figure A10a, SM). As negative experiences increase, personal victimization motives rise while civic motivations for reporting fall, from roughly 61% to just under 40%, suggesting exposure to negative experiences shifts motivations from civic-mindedness toward self-related concerns (Figure A14, SM).

Finally, we examine how reporting motivations vary with both the target of the reported content and users' reporting frequency, measured as the number of submissions to *REspect!*. As descriptive context, Figure A6 (SM)

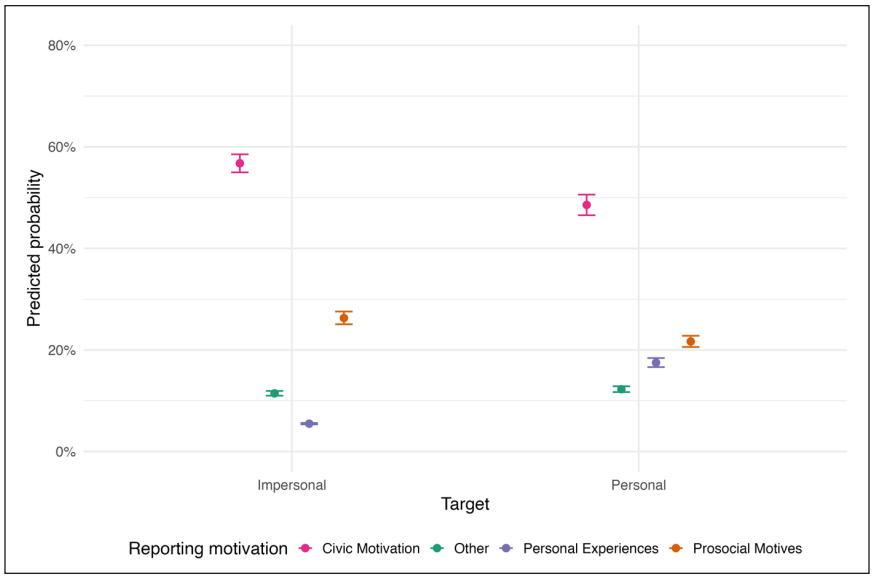

Figure 5. Predicted probabilities of the reported content's target on motivation to report based on multinomial models with "Other" as the reference category.

shows that most reporting users flagged content targeting individuals or groups unrelated to themselves or their in-groups, while Figure A3 (SM) documents substantial variation in how often users report content both to *REspect!* (upper panel) and in general (lower panel), with a small group of frequent reporters accounting for a disproportionate share of submissions and displaying more homogeneous ideological profiles (Table A4, SM). Building on this, Figure 5 presents model-based estimates of reporting motivations by target, and Figure A15 (SM) shows predicted probabilities across the full range of reporting frequency to *REspect!*. Users who reported content that targeted themselves or someone they know are more likely to cite personal motivations, whereas reports concerning unrelated or impersonal targets are more strongly associated with civic motivations.

Descriptively, most respondents reported content only once via *REspect!*, while a smaller share submitted reports multiple times (Figure A3, SM). This variation in reporting frequency provides additional context for the model-based results. As shown in Figure A15, civic motivation becomes increasingly prevalent as the number of submissions rises, from about 48% among infrequent reporters to just over 60% among those reporting at a high rate, while prosocial and personal experience motives decline from roughly 10 percentage points. Attention-check status has minimal influence on these patterns (Figure A20, SM).

Motivation to Report to REspect!. We measure motivation to report to *REspect!* using an open-ended question asking respondents to describe, in at least one sentence, why they used *REspect!* rather than (or in addition to) platform-based reporting. After excluding non-responses, we analyze 694 answers (mean length: 25.4 tokens, ≈ 1.8 sentences). As a first step, we conducted an initial qualitative assessment to identify recurring themes. To systematically analyze thematic structure at scale and complement this assessment, we estimate a Latent Dirichlet allocation (LDA) topic model (Blei et al., 2003). Following standard preprocessing, we compared models with K ∈ 3,4,5,7 and selected a five-topic solution based on coherence, diversity, and interpretability (Röder et al., 2015). Because two platform-focused topics substantially overlapped, we merged them, yielding four interpretable meta-topics. Full preprocessing details, model diagnostics, topic mappings, and illustrative quotes are provided in SM Section A4.

Both approaches converge (Figure 6): a primary driver is frustration with weak enforcement on major platforms – especially Facebook and X – where respondents perceive rule or even law-violating content as persisting without consequences. A second, complementary theme is trust in *REspect!* as a fair, thorough, and reliable channel, often paired with the hope that the office will act where platforms do not.

Beyond these pragmatic concerns, civic values and democratic responsibility also emerge. Respondents specifically mention that, for example, "*Facebook is not taking any action against the hate speech and incitement of AfD politicians and AfD voters, not to mention the bots, trolls, and fake accounts that amplify this incitement!*" Another argues that they reported "*[b]ecause nothing happens on Instagram and Facebook when you report something.*" Beyond these concerns on the lack of enforcement, civic values and attitudes toward democratic responsibility emerge. Respondents express worries that hateful or extremist content undermines democracy if left unchecked, with a majority framing reporting as their civic duty. Others stress that reported content

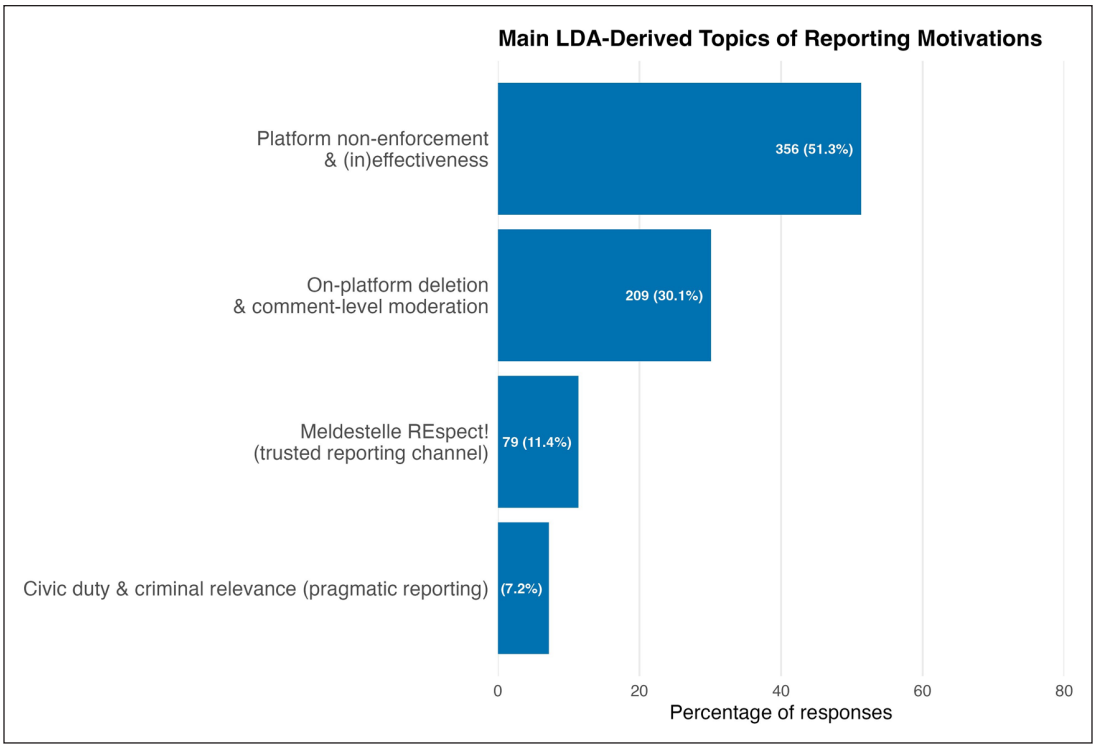

Figure 6. Bar chart with the absolute and relative frequencies of LDA-derived topics from open-ended responses on reporting motivations. Respondents were asked, "Even though there is a reporting mechanism on the website where you came across the content, why did you report the content on the *REspect!*'s website?"

is clearly criminally relevant – thereby listing concrete examples such as "*[a]t X, even the worst insults and clear cases of incitement to hatred are regularly not punished. If you report such things, you get throttled*" or that "*Holocaust denial is a criminal offense and must not go unpunished.*" Furthermore, personal safety concerns appear as an additional important motivator. A subset of reporting users cites fear, direct threats, or experiences of online harassment as reasons for turning to *REspect!*. Trust in *REspect!* as a reliable reporting channel also motivates use, with respondents emphasizing its thoroughness, fairness, and ability to act where platforms do not, and for example, that "*[. . .] the site [is] much more trustworthy. Most platforms don't do anything.*" Another mentions that "*[they] trust REspect to ensure effective prosecution.*"

Discussion and Conclusion

Content moderation has traditionally relied on algorithmic detection and human review, with user reports serving as a crucial backstop for content that slips through. User-centered moderation has recently regained attention following Meta's decision to relax certain moderation rules and end its U.S. fact-checking program – a move widely seen as accommodating the new U.S. administration (Isaac & Schleifer, 2025). Despite the growing importance of user reporting, there is still limited research on who reports potentially illegal content – whether to platforms or third-party organizations – and what motivates them.

This article set out to fill this gap by providing empirical evidence on who reports content, how do reporting users differ from the general population, and what are their key motivations to report content. We address these questions with two complementary surveys, comprising a novel and unique dataset with individuals verified to have previously reported content to a third-party organization in Germany and a quota-based sample designed to approximate the general population in Germany.

Reporting users tend to be older, more often men than women, and highly educated; compared to the general population, they are substantially more politically engaged, markedly left-leaning, and more digitally active. Rather than reflecting a broad cross-section of society, reporting is more common among individuals who view it as a civic duty, raising questions about representation, equity, and the effectiveness of moderation systems that rely on user participation. Respondents' dissatisfaction with the transparency, perceived effectiveness, and enforcement of platform-based moderation further helps explain why some turn to third-party organizations for support.

Our study uniquely surveys reporting users, offering novel insights into their motivations and behavior. A key limitation is that the opt-in design, while well-suited to reach this population, likely overrepresents highly motivated and civically

engaged individuals, as reflected in our results. To contextualize self-selection, we draw on monthly reporting data and observe a relatively high response rate. Some reporting may be strategic or aimed at gaming the system; while we cannot rule this out, it warrants future research. Ideological orientation was captured using different self-placement items across samples, limiting comparability across groups.

Despite its largely descriptive focus, the study makes a theoretical contribution by conceptualizing user reporting as a form of digital civic participation within broader debates on political behavior and platform governance. By identifying the social, attitudinal, and behavioral profiles of reporting users, we extend research on new and digitally enabled forms of political participation (Theocharis, 2015; Theocharis & De Moor, 2021). Our findings suggest that reporting to *REspect!* is linked to civic dispositions, positioning reporting as a practice that can express civic norms and moral judgment within existing content moderation systems.

While respondents were identified via a third-party reporting organization, they were asked about motivations for reporting both in general and to that organization specifically. Findings should be interpreted with caution, as they reflect users who actively engage in reporting, were recruited through a third-party service, and opted into the survey and may not generalize to all users or moderation contexts. The study informs debates on platform governance, user agency, and regulatory approaches to trust and safety, while also contributing to research on online political participation and the design of policies aimed at holding both platforms and perpetrators accountable.

By highlighting potential shortcomings in content moderation mechanisms and their effects on users, our analysis suggests that, in the context studied here, reliance on user reporting as a central pillar of moderation may disproportionately amplify the voices of a narrow, civically engaged group. This raises concerns about representational bias and potential blind spots in the identification of harm. To promote fairness, accountability, and effectiveness, platforms and regulators should consider how to broaden participation and transparency in reporting, and in content moderation more broadly, by making these mechanisms more accessible, transparent, and trusted across diverse user groups. Recognizing reporting as a form of digital civic engagement also opens new avenues for empowering users as active stewards of online communities.

Acknowledgements

The authors thank the editor and anonymous reviewers for their valuable feedback. The authors are grateful to Jan Zilinsky, Jesper Sommer Rasmussen, and Giuliano Formisano for helpful discussions and feedback and to Arjun Premkumar and Sebastián Aguilar for excellent research assistance. The authors also thank participants at the 2025 European Political Science Association (EPSA) Conference, the American Political Science Association (APSA) Annual Meeting, and other workshops for helpful comments, and REspect! for their continued collaboration.

ORCID iDs

Friederike Quint https://orcid.org/0009-0006-7158-1369
Yannis Theocharis https://orcid.org/0000-0001-7209-9669
Spyros Kosmidis https://orcid.org/0000-0002-9465-0789
Margaret E. Roberts https://orcid.org/0000-0001-6900-4366

Ethical Considerations

Both studies were reviewed and received ethical approval. The survey on confirmed reporters received approval by the Ethics Committee of the Technical University of Munich (Ethikkommission der Technischen Universität München, approval number: 2024-38_1-NM-BA). The survey with German respondents received ethical approval from the ethics board of the University of Oxford (approval number: SSH/DPIR_C1A_24_006) as part of a larger 10-country study.

Consent to Participate

All participants gave informed consent and were warned that some questions could be sensitive or distressing. Participation was voluntary, and respondents could skip any question or withdraw at any time without penalty. We provided contact details for queries to the respective institutions' complaints procedure.

Consent for Publication

Not applicable.

Author Contributions

The lead author developed the study; led the research design, data collection, and analysis; and drafted the manuscript. The co-authors contributed to study design and theory; measurement; provision of the collection environment; statistical modeling and robustness checks; and manuscript revision. All authors approved the final version.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article. The survey on reporting users was funded by the Council of Europe's impact study and coordinated by the Justice for Prosperity Foundation, in collaboration with *REspect!*, a reporting office by the Youth Foundation Baden-Wuerttemberg. The second survey was funded by the Technical University of Munich and the University of Oxford. This research was further supported by the 2022 Max Planck-Humboldt Research Award, which was awarded to Professor M.E.R.

Declaration of Conflicting Interests

The other authors declare no conflicts of interest.

Data Availability Statement

Given the sensitivity of the underlying data, the dataset will not be publicly released. Replication materials will only be shared upon reasonable request from university-affiliated researchers whose stated purpose is to reproduce this study. For the second survey approximating the German population, data can be shared at request for replication reasons.

Supplemental Material

Supplemental material for this article is available online.

Notes

1. As Frances Haugen testified to Congress, Facebook's internal research found it identifies only 3%–5% of hateful content on the platform (U.S. Senate, 2021).
2. *REspect!* is operated by the German state of Baden-Wuerttemberg's Youth Foundation and funded through federal and state programs, including the German Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (BMFSFJ) under the program "Demokratie leben!". While publicly funded, it operates as a civil society reporting portal rather than a government authority. Additional institutional context is provided in Supplementary Material (hereinafter SM) A1.3.
3. See SM A1 and A2 for procedures and invitation text.
4. We elaborate further on external validity and possible limitations in Section A3, SM.
5. More details on the sample can be seen in A1 and A1.1, SM.
6. In the representative survey, gender was determined using the response options "Male," "Female," and "Other." However, none of the respondents selected the "Other" category, whereas in the reporting user survey, more respondents fall into this category.
7. For a longer discussion of ideology, see A5.3, SM.
8. We refrain from further analyzing the open-text "Other" responses.
9. For more details, see Table A1 in SM A1.1.
10. The item "Followed a politician, a political party, or a political organization on social media" was not measured in the German survey.
11. For more details, see SM A5.4.
12. See SM A5.7 for discussion of demographic gradients and interpretation.

References

Aleksandric, A., Singhal, M., Groggel, A., & Nilizadeh, S. (2022). Understanding the bystander effect on toxic Twitter conversations. *arXiv*. https://arxiv.org/abs/2211.10764

Appel, R. E., Pan, J., & Roberts, M. E. (2023). *Partisan conflict over content moderation is more than disagreement about facts*. https://doi.org/10.2139/ssrn.4331868

Bhandari, A., Ozanne, M., Bazarova, N. N., & DiFranzo, D. (2021). Do you care who flagged this post? Effects of moderator visibility on bystander behavior. *Journal of Computer-mediated Communication*, *26*(5), 284–300. https://doi.org/10.1093/jcmc/zmab007

Blackwell, L., Chen, T., Schoenebeck, S., & Lampe, C. (2018). When online harassment is perceived as justified. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1). https://doi.org/10.1609/icwsm.v12i1.15036

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Böswald, L.-M., Dolezalek, C., Jost, P., & Schmid, U. K. (2025). Between click and consequence: An evaluation of platform reporting procedures under the Digital Services Act. *Das NETTZ*. https://www.das-nettz.de/sites/default/files/2025-10/ENG_Langfassung_DSA.pdf

Brennauer, J., Dander, V., Dolezalek, C., & Kompetenznetzwerk gegen Hass im Netz (Eds.). (2024). *Lauter Hass – leiser Rückzug: Wie Hass im Netz den demokratischen Diskurs bedroht: Ergebnisse einer repräsentativen Befragung* [Loud hate – quiet withdrawal: How online hate threatens democratic discourse: Results of a representative survey]. Kompetenznetzwerk gegen Hass im Netz. https://hateaid.org/wp-content/uploads/2024/04/Studie_Lauter-Hass-leiser-Rueckzug.pdf

Busch, C. (2022). *Regulating the expanding content moderation universe: A European perspective on infrastructure moderation*. UCLA Journal of Law & Technology. https://uclajolt.com/wp-content/uploads/2022/04/Busch_Final-4.14.pdf

Corduneanu-Huci, C., & Hamilton, A. (2022). Selective control: The political economy of censorship. *Political Communication*, *39*(4), 517–538. https://doi.org/10.1080/10584609.2022.2074587

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. https://doi.org/10.1177/1461444814543163

Delmas, C. (2018). *A duty to resist: When disobedience should be uncivil*. Oxford University Press.

DiFranzo, D., Taylor, S. H., Kazerooni, F., Wherry, O. D., & Bazarova, N. N. (2018). Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). ACM. https://doi.org/10.1145/3173574.3173785

Doseva, S., Carathanassis, F., Schmid-Petri, H., & Heckmann, D. (2024). Insults on social media: How affected are users on social media and how do they defend themselves? *bidt*. https://en.bidt.digital/publication/insults-on-social-media-how-affected-are-users-on-social-media-and-how-do-they-defend-themselves/

Douek, E. (2022). Content moderation as systems thinking. *Harvard Law Review*, *136*, 526–607. https://doi.org/10.2139/ssrn.4005326

European Union. (2022, October). *Regulation (EU) 2022/2065 of the European Parliament and of the Council (Digital Services Act)*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065

Gennaro, G., Derksen, L., Abdelrahman, A., Broggini, E., Green, M. A., Haerter, V. A., Heer, E., Heidler, I., Kauer, F., Kim, H.-N., Landry, B., Levis, A., Li, J., Şimşir, Ş., Srbinovska, I., Vital, R. A., Donnay, K., Gilardi, F., & Hangartner, D. (2025). Counterspeech encouraging users to adopt the perspective of minority groups reduces hate speech and its amplification on social media. *Scientific Reports*, *15*, 22018. https://doi.org/10.1038/s41598-025-05041-w

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press. https://doi.org/10.12987/9780300235029

Gray, K., & Kubin, E. (2024). Victimhood: The most powerful force in morality and politics. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 70, pp. 137–220). Elsevier. https://doi.org/10.1016/bs.aesp.2024.03.004

Guess, A. M., & Munger, K. (2023). Digital literacy and online political behavior. *Political Science Research and Methods*, *11*(1), 110–128. https://doi.org/10.1017/psrm.2022.17

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Munoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wuthrich, F., Gilardi, F., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, *118*(50), Article e2116310118. https://doi.org/10.1073/pnas.2116310118

Hansen, T. M., Lindekilde, L., Karg, S. T., Bang Petersen, M., & Rasmussen, S. H. R. (2024). Combatting online hate: Crowd moderation and the public goods problem. *Communications*, *49*(3), 444–467. https://doi.org/10.1515/commun-2023-0109

Howard, J. W. (2019). Free speech and hate speech. *Annual Review of Political Science*, *22*, 93–109. https://doi.org/10.1146/annurev-polisci-051517-012343

Isaac, M., & Schleifer, T. (2025, January 7). Meta to end fact-checking program in shift ahead of Trump term. *The New York Times*. https://www.nytimes.com/2025/01/07/technology/meta-fact-checking-facebook.html

Jhaver, S., & Zhang, A. (2023). Do users want platform moderation or individual control? *arXiv*. https://arxiv.org/abs/2301.02208

Kemp, D., & Ekins, E. (2021). *Poll: 75% don't trust social media to make fair content moderation decisions, 60% want more control over posts they see*. Cato Institute. https://www.cato.org/survey-reports/poll-75-dont-trust-social-media-make-fair-content-moderation-decisions-60-want-more

Koch, L., Voggenreiter, A., & Steinert, J. (2025). *Angegriffen & alleingelassen: Wie sich digitale Gewalt auf politisches Engagement auswirkt. Ein Lagebild* [Attacked and left alone: How digital violence affects political engagement. A situational report]. OSF. https://doi.org/10.17605/osf.io/j4stx

Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., & Reifler, J. (2023). Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, *120*(7), Article e2210666120. https://doi.org/10.1073/pnas.2210666120

Meta. (2020, November 19). *How AI is getting better at detecting hate speech*. https://ai.meta.com/blog/how-ai-is-getting-better-at-detecting-hate-speech/

Meta. (2025, January). *Meta to end fact-checking on its platforms* [Instagram post]. https://www.instagram.com/reel/DEhf2uTJUs0/

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629–649. https://doi.org/10.1007/s11109-016-9373-5

Munger, K., Gopal, I., Nagler, J., & Tucker, J. A. (2021). Accessibility and generalizability: Are social media effects moderated by age or digital literacy? *Research & Politics*, *8*(2), 20531680211016968. https://doi.org/10.1177/20531680211016968

Munzert, S., Traunmuller, R., Barberá, P., Guess, A., & Yang, J. (2025). Citizen preferences for online hate speech regulation (K. Ognyanova, Ed.). *PNAS Nexus*, *4*(2), pgaf032. https://doi.org/10.1093/pnasnexus/pgaf032

Nahrgang, M., Weidmann, N. B., Quint, F., Nagel, S., Theocharis, Y., & Roberts, M. E. (2025). Written for lawyers or users? Mapping the complexity of community guidelines. *Proceedings of the International AAAI Conference on Web and Social Media*, *19*, 1295–1314. https://doi.org/10.1609/icwsm.v19i1.35873

Obermaier, M. (2024). Youth on standby? Explaining adolescent and young adult bystanders' intervention against online hate speech. *New Media & Society*, *26*(8), 4785–4807. https://doi.org/10.1177/14614448221125417

Obermaier, M., Fawzi, N., & Koch, T. (2016). Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying. *New Media & Society*, *18*(8), 1491–1507. https://doi.org/10.1177/1461444814563519

Porten-Cheé, P., Kunst, M., & Emmer, M. (2020). Online civic intervention: A new form of political participation under conditions of a disruptive online discourse. *International Journal of Communication*, *14*, 1–21. https://ijoc.org/index.php/ijoc/article/view/10639

Pradel, F., Zilinsky, J., Kosmidis, S., & Theocharis, Y. (2024). Toxic speech and limited demand for content moderation on social media. *American Political Science Review*, *118*(4), 1895–1912. https://doi.org/10.1017/S000305542300134X

Rasmussen, J. (2022, June 8). *The (limited) effects of target characteristics on public opinion of hate speech laws*. OSF Preprints. https://doi.org/10.31234/osf.io/j4nuc

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 399–408). ACM. https://doi.org/10.1145/2684822.2685324

Shim, Y., & Jhaver, S. (2024). Incorporating procedural fairness in flag submissions on social media platforms. *arXiv*. https://doi.org/10.48550/arXiv.2409.08498

Siegel, A. A., & Badaan, V. (2020). #No2sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review*, *114*(3), 837–855. https://doi.org/10.1017/S0003055420000283

Singhal, M., Ling, C., Paudel, P., Thota, P., Kumarswamy, N., Stringhini, G., & Nilizadeh, S. (2023). SoK: Content moderation in social media, from guidelines to enforcement, and research to practice. In *Proceedings of the 2023 IEEE European Symposium on Security and Privacy* (pp. 868–895). IEEE. https://doi.org/10.1109/EuroSP57164.2023.00056

Solomon, B. C., Hall, M. E. K., Hemmen, A., & Druckman, J. N. (2024). Illusory interparty disagreement: Partisans agree on what hate speech to censor but do not know it. *Proceedings of the National Academy of Sciences*, *121*(39), Article e2402428121. https://doi.org/10.1073/pnas.2402428121

Suzor, N. P. (2019). *Lawless: The secret rules that govern our digital lives*. Cambridge University Press. https://doi.org/10.1017/9781108666428

Theocharis, Y. (2015). The conceptualization of digitally networked participation. *Social Media + Society*, *1*(2), Article 2056305115610140. https://doi.org/10.1177/2056305115610140

Theocharis, Y., & De Moor, J. (2021). Creative participation and the expansion of political engagement. In *Oxford research Encyclopedia of politics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228637.013.1972

Theocharis, Y., Kosmidis, S., Zilinsky, J., Quint, F., & Pradel, F. (2025). Content warning: Public attitudes on content moderation and freedom of expression. *Content Moderation Lab at TUM*. https://doi.org/10.17605/OSF.IO/F56BH

U.S. Senate. (2021, October 5). *Protecting kids online: Testimony from a Facebook whistleblower*. https://www.congress.gov/117/chrg/CHRG-117shrg54110/CHRG-117shrg54110.pdf

Verba, S., & Nie, N. H. (1987). *Participation in America: Political democracy and social equality*. University of Chicago Press.

Vogels, E. A., Perrin, A., & Anderson, M. (2020). *Most Americans think social media sites censor political viewpoints*. Pew Research Center. https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/

Wang, S. (2021). Standing up or standing by: Bystander intervention in cyberbullying on social media. *New Media & Society*, *23*(6), 1379–1397. https://doi.org/10.1177/1461444820902541

York, J. C. (2022). *Silicon values: The future of free speech under surveillance capitalism*. Verso.

Author Biographies

Friederike Quint (MA, University of Mannheim) is a doctoral candidate and researcher at the Chair of Digital Governance at the Munich School of Politics and Public Policy, Technical University of Munich. Her research interests include political communication and political behavior, with a focus on content moderation, platform regulation, and public attitudes toward social media and freedom of expression.

Yannis Theocharis (PhD, University College London) is Professor and Chair of Digital Governance at the Munich School of Politics and Public Policy, Technical University of Munich. His research interests include political behavior, political communication, content moderation, and computational social science.

Spyros Kosmidis (PhD, University of Essex) is Associate Professor in the Department of Politics and International Relations at the University of Oxford. His research interests include public opinion, political behavior, social attitudes, and party competition.

Margaret E. Roberts (PhD, Harvard University) is Professor at the Department of Political Science at the University of California, San Diego. Her research interests focus on the intersection of political methodology and digital politics.