2021-05

SoK: hate, harassment, and the changing landscape of online abuse

SoK: Hate, Harassment, and the Changing Landscape of Online Abuse

Kurt Thomas[◇], Devdatta Akhawe[▽], Michael Bailey[§], Dan Boneh[□], Elie Bursztein[◇], Sunny Consolvo[◇], Nicola Dell[○], Zakir Durumeric[□], Patrick Gage Kelley[◇], Deepak Kumar[§], Damon McCoy[‡], Sarah Meiklejohn[△◇], Thomas Ristenpart[○], Gianluca Stringhini[*]

[◇]Google [*]Boston University [○]Cornell Tech [▽]Figma, Inc. [‡]New York University
[□]Stanford [△]University College London [§]University of Illinois, Urbana-Champaign

***Abstract*—We argue that existing security, privacy, and anti-abuse protections fail to address the growing threat of online hate and harassment. In order for our community to understand and address this gap, we propose a taxonomy for reasoning about online hate and harassment. Our taxonomy draws on over 150 interdisciplinary research papers that cover disparate threats ranging from intimate partner violence to coordinated mobs. In the process, we identify seven classes of attacks—such as toxic content and surveillance—that each stem from different attacker capabilities and intents. We also provide longitudinal evidence from a three-year survey that hate and harassment is a pervasive, growing experience for online users, particularly for at-risk communities like young adults and people who identify as LGBTQ+. Responding to each class of hate and harassment requires a unique strategy and we highlight five such potential research directions that ultimately empower individuals, communities, and platforms to do so.**

I. Introduction

Emerging threats like online hate and harassment are transforming the day-to-day experiences of Internet users. Abusive attacks include intimate partner violence [27], [65], [66], [108], anonymous peers breaking into a target's account to leak personal communication and photos [131], and coordinated bullying and sexual harassment campaigns that involve tens of thousands of attackers [1]. In a survey by Pew in 2017, 41% of Americans reported personally experiencing varying degrees of harassment and bullying online [118]. Globally, 40% of people reported similar experiences [110].

Despite this changing abuse landscape, existing security and anti-abuse protections continue to lag and focus almost exclusively on disrupting cybercrime. Such defenses take into account the profit incentives of attackers and their requirement to target as many victims as possible to scale and maximize returns [3], [136]. Hate and harassment does not adhere to this for-profit paradigm. Attackers are instead motivated by ideology, disaffection, and control: a landscape where interpersonal and geopolitical conflicts happen as much online as offline. Consequently, threats are highly personalized [108], vary across cultural contexts [127], and often exploit unintended applications of widely accessible technologies [27].

In this work, we explore how online hate and harassment has transformed alongside technology and make a case for why the security community needs to help address this threat. We collate over 150 research papers and prominent news stories related to hate and harassment and use them to create a taxonomy of seven distinct attack categories. These include—among others—*toxic content* like bullying and hate speech, and *surveillance* including device monitoring and account takeover.

We then provide in-depth, longitudinal statistics on the growth of hate and harassment and the at-risk communities currently being targeted. Our analysis draws on a three-year survey collated from 50,000 participants located in 22 different countries. We find that 48% of people globally report experiencing threats including sustained bullying (5%), stalking (7%), and account takeover by someone they know (6%). Over the past three years, the odds of users experiencing abuse have increased by 1.3 times. Young adults aged 18–24 and LGTBQ+ individuals in particular face heightened levels of risk. These observations requires that practitioners take into account regional variations and at-risk groups when designing interventions.

Based on our findings, we propose five directions for how our community can re-imagine security, privacy, and anti-abuse solutions to tackle hate and harassment. Our proposed interventions span technical, design, and policy changes that assist in identifying, preventing, mitigating, and recovering from attacks. Exploring these directions, however, requires resolving multiple social equities that are in conflict. Tensions include balancing notions of free speech with community or platform-based moderation, and the well-being of raters with the necessity of human review to interpret context. Resolutions to these tensions must come from researchers, practitioners, regulators, and policy experts at large in order to stem the threat posed by online hate and harassment.

II. What is online hate and harassment?

To appropriately ground our taxonomy and solutions, we first scope what abusive behaviors fall under the umbrella of online hate and harassment. We then discuss the interplay between these attacks and other emerging online threats, such as violent extremism and inaccurate information.

A. Hate and harassment background

Hate and harassment occurs when an aggressor (either an individual or group) specifically targets another person or group to inflict emotional harm, including coercive control or instilling a fear of sexual or physical violence [36]. Examples of highly publicized attacks include "Gamergate", a coordinated campaign where several women tied to the video game industry received tens of thousands of messages that threatened rape and death [1]. More recently, an attacker publicly leaked nude images of former Rep. Katie Hill, resulting in her resigning from office [54].

While hate and harassment have a long history in the social sciences and the ethos of the Internet [92], [130]—with common adages like "don't read the comments"—the public increasingly views hate and harassment as a threat that needs to be addressed. In a survey by Pew in 2017, 76% of Americans believed that platform operators have a duty to step in when hate and harassment occurs on their service [118]. This shift in public opinion is also reflected in the Terms of Service of online platforms. For example, in 2009, Twitter's rules covered only impersonation and spam [81]. As of 2020, Twitter's rules also cover violence, harassment, and hateful conduct, among a multitude of other abusive behaviors [138]. Hate and harassment is now explicitly prohibited by almost all major online social networks [116].

While the intent to cause emotional harm differs strongly from the profit incentives of cybercrime [3], some parallels exist between the underlying tools and techniques of both types of attacks: spamming, creating fake accounts, compromising devices, obtaining privileged access to sensitive information, and more. When protecting users from cybercrime, however, security practitioners can disrupt the profit-generating centers fueling abuse like scams, ransomware, or banking fraud [136], which in turn eliminates the incentives behind fake accounts, spam, or related abuse. There is no equivalent strategy for hate and harassment where attacks presently lack a dependency chain. Technical interventions can merely address the symptoms of conflicts rooted in politics and culture.

B. Related emerging threats

The types of attackers engaged in hate and harassment may also be involved in other emergent forms of abuse such as violent extremism or spreading inaccurate information (see Figure 1). We briefly discuss these related threats and how their goals at times overlap, while also proposing how to distinguish hate and harassment.

Violent extremism. Over the last two decades, violent extremists—actors that glorify or enact ideologically-motivated violence—have adopted emerging technologies for coordination, recruitment, and distributing propaganda [91]. In the early 2000s, online abuse by jihadists was isolated to forums like Al-Hesbah where readers could "follow links to attack videos from active jihad campaigns" [21]. In the 2010s, extremist content migrated to social media sites, where extremist media personalities developed a brand around violence

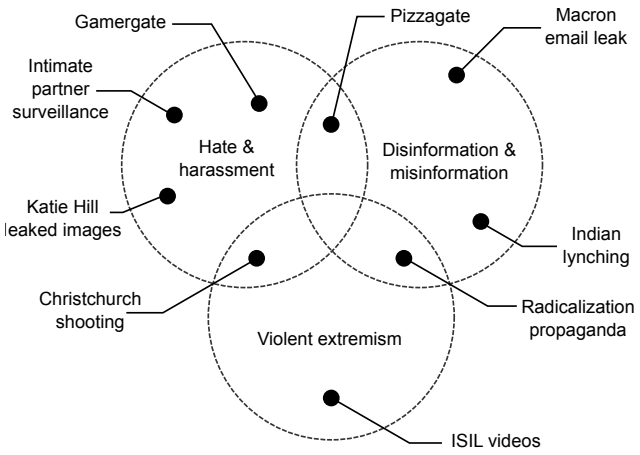

Fig. 1: Attackers engaged in hate and harassment may also engage in violent extremism and disinformation and misinformation. We argue there is not always a clear boundary between each class of abuse, but that the underlying intents differ.

while subverting hashtags and trending pages via search engine optimization tactics to reach a wider audience [9], [16], [109], [114]. Once in the public sphere, this content may be sought after as part of a path towards radicalization [55].

These same tactics have been adopted in recent years by far-right extremists, such as the 2019 live-streaming of a mosque shooting in New Zealand [122] and of a synagogue shooting in Germany [79]. Technology platforms have coordinated their response to this threat through groups like the Global Internet Forum to Counter Terrorism (GIFCT), for example sharing digital fingerprints of extremist content to enable automatic removal [68]. As with cybercrime, this has led to an adversarial arms race, with extremists moving to less moderated communities [100], [149], adopting encrypted messaging platforms [135], and testing new mediums such as short-lived videos [144].

In the context of hate and harassment, extremist propaganda may overlap or be amplified by communities actively involved in online hate. However, the aim of violent extremism is to radicalize as many people as possible via indiscriminate distribution, which falls outside our requirement that hateful or harassing content target a specific individual or group.

Disinformation and misinformation. Disinformation encompasses any efforts by "groups, including state and non-state actors, to manipulate public opinion and change how people perceive events" [133]. Tactically speaking, this entails deliberately spreading false or misleading information [78]. Misinformation involves the spread of unintentional inaccuracies [78], such as when Indian villagers used messaging apps to spread rumors of child abduction, ultimately resulting in the lynching of a man [8]. Both disinformation and misinformation abuse online platforms as a tool for dissemination.

There is an intrinsic link with the role that rumors and falsehoods can play in online hate and harassment. A prime example is "Pizzagate" [85]. Initially a conspiracy theory

that presidential candidate Hillary Clinton ran a pedophilia ring out of a Washington DC pizzeria, the event spiraled, leading to harassment of the restaurant's staff [85] and an armed man entering the premises to "self-investigate" the situation [22]. The motives and tactics behind inaccurate information campaigns can thus overlap with hate and harassment. However, while there may be emotional harm incurred by inaccurate information, the primary focus of attackers is to indiscriminately scale in order to change the perceptions of as many people as possible. For this reason, we treat inaccurate information campaigns as a separate class of abuse.

III. A Taxonomy of Online Hate and Harassment

Given the breadth of hate and harassment attacks, we propose a threat model and taxonomy to assist in reasoning about strategies for detection, prevention, mitigation, and recovery. Our taxonomy (Table I) identifies the criteria that differentiate attacks, the harms incurred, and the scale of abuse.

A. Literature review

We developed our threat model and taxonomy by manually examining the last five years of research from IEEE S&P, USENIX Security, CCS, CHI, CSCW, ICWSM, WWW, SOUPS, and IMC. Our team focused on topics related to hate speech, harassment, trolling, doxing, stalking, non-consensual image exposure, disruptive behavior, content moderation, and intimate partner violence. We then manually searched through the related works of these papers for relevant research, including findings from the social sciences and psychology communities (though restricted solely to online hate and harassment, rather than hate speech or bullying in general). Additionally, we relied on the domain expertise of the authors to identify related works and major recent news events. In total, we reviewed over 150 news articles and research papers on the topic of online hate and harassment.

B. Defining a threat model

We interpret hate and harassment through a threat model that consists of an *attacker* and a *target*.[1] The attacker's intent is to emotionally harm or coercively control the target, irrespective of other side effects. Attackers may include intimate partners such as a spouse, family and peers, anonymous individuals, public figures (such as media personalities or politicians), or coordinated groups and Internet mobs. Likewise, targets may include intimate partners, family and peers, individuals, public figures, or at-risk groups (e.g., LGBTQ+ people or minorities). Our threat model makes no assumptions about the capabilities of attackers, the existence of a direct communication channel between the attacker and target, or the protections available to targets.

[1]Similar to research in intimate partner violence, we intentionally avoid the term "victim" in favor of "target" to not disempower people facing abuse.

C. Identifying criteria that differentiate attacks

Through an iterative process reviewing the attacks described in the research papers we analyzed, three researchers arrived at seven criteria to differentiate attacks. These researchers annotated all papers for whether they partially or fully satisfied any of the criteria, with the rest of the research team validating the annotations. Our criteria include the AUDIENCE exposed to an attack (A1-2), the MEDIUM through which an attacker reaches a target (M1), and the CAPABILITIES required for the attack to succeed (C1-4). Each criterion is represented as a column of our taxonomy in Table I.

In selecting our criteria, we favored broad themes that we believe will remain stable despite the rapidly expanding nature of hate and harassment. As such, we opted for a taxonomy that is agnostic to the technology platform abused (e.g., messaging application, video application, or social network) or the exact type of information involved. At the same time, we developed our criteria to be granular enough to meaningfully differentiate threats, and thus assist us in identifying solutions that apply to only a single segment of hate and harassment, rather than all abuse as a whole. We detail our criteria below.

Is the attack intended to be seen by the target? (A1). We differentiate attacks that deliberately expose a target to harmful content—such as bullying—from potentially undetected attacks like covert stalking and monitoring. Awareness on the part of the target allows them to report abusive behavior or reach out to others for support. Attacks that may be visible to the target, but not directly sent to the target (e.g., negative reviews for a target's business), partially satisfy this criteria.

Is the attack intended to be seen by an audience? (A2). We consider whether attacks inherently require an audience to incur harm, such as intentionally leaking personal information like a target's sexuality. Fully public exposure can exacerbate the challenge of removing abusive content, but it opens up the possibility for bystanders to intervene. Attacks that do not require an audience, but that may be visible to bystanders (e.g., threats in a comment on a video), partially satisfy this criteria.

Does the attack use media such as images or text? (M1). We consider whether an attacker requires a communication channel to the target or other audience to disseminate text, images, or other media. The use of media opens up the possibility of moderation or filtering via the operator of the communication channel. Scenarios where an attacker can share predefined reactions (e.g., thumbs down on a video) only partially satisfy this criteria.

Does the attack require deception of an audience? (C1). In terms of capabilities, we examine whether or not the attack relies on some level of deception of an online audience in order to humiliate the target or otherwise damage their reputation.

Does the attack require deception of a third-party authority? (C2). This capability is more nuanced and considers whether or not the attacker leverages an authority to (inadvertently) take action on the attacker's behalf.

Does the attack require amplification? (C3). Some attacks inherently require coordinated action or amplification to succeed, such as mass down-voting a target's videos, or denial-of-servicing a target's website. In some cases, amplification may come from the platform itself (e.g., video conference calls which focus all viewers on a speaker). While all attacks likely benefit from some form of amplification, we limit this criterion only to when amplification is a necessity.

Does the attack require privileged access to information, an account, or a device? (C4). As our final criterion, we consider whether or not an attacker requires privileged access. Such access may come through coercion, misplaced trust (e.g., a spouse or peer), or a security or privacy vulnerability (e.g., a weak password). Scenarios where the information available to an attacker may be public, but not widely available—such as legal documents—partially satisfy this criteria.

D. Harms that result from attacks

Apart from the capabilities required to conduct an attack, our taxonomy also explores the specific harms that attackers likely intend as the outcome of online hate and harassment. In particular, we highlight whether an attack's intent is to silence a target; to damage a target's reputation; to reduce a target's sense of sexual or physical safety; or to coerce a target. As our threat model covers a gamut of complex relationships between attackers and targets, we argue it would be inappropriate for our taxonomy to specifically categorize attacks based on harms. Instead, we merely highlight potential harms to better explain the difference between threats like sexual harassment and bullying. These harms may play a role in policy development but do not impact technical solutions, which is the primary role of our taxonomy.

E. Scale of attacks

The last part of our taxonomy differentiates attacks targeting an individual—like the non-consensual exposure of intimate images—or an entire group. In some cases, both targeting strategies are equally possible. The targeted nature of online hate and harassment differs strongly from for-profit threats, and thus heavily influences the design of new solutions.

F. Categorization of attacks

By labeling the attacks in our literature review using our criteria, we identified seven distinct categories of hate and harassment. When discussing each, we also highlight the primary security principle—confidentiality, integrity, or availability—that the attacks in each category undermine. We make no claim our list of attacks is exhaustive. Instead, our goal is to illustrate how each class of attacks requires a different solution due to the capabilities and motives involved.

Toxic Content [Availability; A1-A2, M1 exclusively]. Toxic content covers a wide range of attacks involving media (M1) that attackers send to a target or audience (A1-A2), but without the necessity of more advanced capabilities (not C1-C4). Attacks in this category include bullying, trolling (e.g., intentionally provoking audiences with inflammatory remarks), threats of violence, and sexual harassment. A close equivalent in for-profit cybercrime is spam [99]. Repeated abuse may result in targets deleting their account to avoid toxic interactions, effectively silencing and marginalizing the targets [105], [127], [131]. This illustrates how toxic content can be used to violate availability, preventing victims from properly taking advantage of an online community and even forcing them to leave it.

Numerous studies have examined toxic content that attackers spread via social networks [30], [75], [119], [124], with a particular focus on toxic content targeting minorities [123] and women [28], [35], [141]. Other threats in this space include the viral distribution of hateful or racist memes and videos [60], [115], [125], [134], [150] and abuse carried out among online gaming players [11], [98], [137]. This content, which can originate in communities dedicated to hate and harassment, often makes its way into mainstream social networks [61], [150], in turn impacting a much broader audience. All of these attacks are aided in part by the anonymous nature of online communication, which hampers accountability [152].

Content Leakage [Confidentiality; A2 + M1 + C4]. Content leakage involves any scenario where an attacker leaks (or threatens to leak) sensitive, private information (M1 + C4) to a wider audience (A2). Often, the attacker's intent is either to embarrass, threaten, intimidate, or punish the target [53]. An attacker may obtain access to this information via privileged access, such as compromising the target's account or socially engineering a third party; via information requests, public legal records, and records exposed by data breaches; or by coercing the target through duress. Snyder et al. previously studied over 4,500 "doxing" attacks that exposed a target's personal information to a broad audience [131]. They found that 90% of incidents exposed physical mailing addresses, 60% exposed phone numbers, and 53% exposed personal email addresses. Specific to the LGTBQ+ community, attackers may also reveal an individual's sexual identity (e.g., "outing") or reject an individual's gender identity by using the former name of a transgender or non-binary person (e.g., "deadnaming"). In turn, Internet audiences may amplify the fallout of a target's personal information being exposed [82], [102].

Other forms of content leakage are rooted in sexual violence. For example, an attacker (e.g., former partner) can expose intimate images to the target's friends, family, colleagues, or even publicly. This is often referred to as non-consensual intimate imagery or, colloquially, as "revenge porn". Survivors of intimate partner violence report this as a frequent problem [66], [108], [145]. In a prior survey, Microsoft estimated as many as 2% of people have been recorded in an intimate situation without their consent [110]. Another survey found 4% of Americans have been threatened with or experienced non-consensual intimate image exposure [43]. Such threats can in turn lead to a vicious cycle of extortion (e.g., "sextortion"), where the target continues to supply intimate images to avoid exposure.

		Criteria							Harms					Scale	
Category	**Non-exhaustive list of attacks**	Intended to be seen by target? (A1)	Intended to be seen by audience? (A2)	Requires media such as images or text? (M1)	Requires deception of an audience? (C1)	Requires deception of a third-party authority? (C2)	Requires amplification? (C3)	Requires privileged access? (C4)	Intent to silence?	Intent to damage reputation?	Intent to reduce sexual safety?	Intent to reduce physical safety?	Intent to coerce?	Targets an individual?	Targets a group?
Toxic content	Bullying	●	◐	◐					◐	◐				●	
	Trolling	●	◐	◐					◐	◐				◐	◐
	Hate speech	●	◐	●					◐	◐				◐	◐
	Profane or offensive content	◐	◐	●					◐	◐				◐	◐
	Threats of violence	●	◐	●					◐	◐		●	◐	◐	◐
	Purposeful embarrassment	◐	◐	●					◐	◐			◐	●	
	Incitement	◐	◐	●					◐	◐				◐	◐
	Sexual harassment	●	◐	◐					◐	◐	●		◐	◐	◐
	Unwanted explicit content (“sexting”)	●	◐	●					◐	◐	●		◐	●	
Content leakage	Sextortion	●	◐	●				◐	◐	◐	●		●	●	
	Doxing	◐	●	●				◐	◐	◐	◐	◐		●	
	Outing and deadnaming	◐	●	●				◐	◐	●	◐	◐	◐	●	
	Non-consensual image exposure (“revenge porn”)	◐	●	●				●	◐	●	●	◐	◐	●	
	Leaked chats, profiles	◐	●	●				●	◐	●	◐	◐	◐	●	
Overloading	Comment spam	●	●	●			●		●					●	
	Dogpiling	●	●	●			●		●					●	
	Raiding or brigading	●	●	●			●		●					◐	◐
	Distributed denial of service (DDoS)	●					●		●					◐	◐
	Notification bombing	●		◐			●		●					●	
	Zoombombing	◐	●	●			●		●	◐	◐	◐		◐	◐
	Negative ratings & reviews	◐	●				●		●	●				●	
False reporting	SWATing	●				●		◐				●		●	
	Falsified abuse report	●	◐	●		●			●					●	
	Falsified abuse flag	●	◐			●			●					●	
Impersonation	Impersonated profiles		●	◐	●					●				●	
	Impersonated chats or images		●	●	●					●				●	
	Impersonated webpages (SEO)		●	◐	●		●			●				◐	◐
	Synthetic pornography		●	●	◐					●				◐	◐
	Hijacked communication	◐	●	●	●			●		●				●	
Surveillance	Stalking or tracking							●			◐	●	●	●	
	Account monitoring							●			◐	●	●	●	
	Device monitoring							●			◐	●	●	●	
	IoT monitoring (passive)							●	◐			●	●	●	
	Browser monitoring (passive)							●	◐				●	●	
Lockout and control	IoT manipulation (active)	●						●	◐			●	●	●	
	Browser manipulation (active)	●						●	◐				●	●	
	Account lockout	●						●	◐				●	●	
	Content deletion	◐						●	◐				●	●	

TABLE I: Taxonomy of online hate and harassment attacks, broken down by audience, communication channel, and capabilities involved. We annotate each attack with the most common intents of the attacker, though nuanced relationships between an attacker and target make such harms difficult to generalize. A ● indicates that a criterion always holds true, while a ◐ indicates that a criterion frequently holds true. No entry indicates a criteria does not hold. The stratification of attacks across our criteria result in seven distinct categories of threats.

Overloading [Availability; A1 + C3]. Overloading includes any scenario wherein an attacker forces a target (A1) to triage hundreds of notifications or comments via amplification (C3), or otherwise makes it technically infeasible for the target to participate online due to jamming a channel (potentially via a distributed denial of service attack) (C3). Examples include organized trolling activity orchestrated through Facebook [120], Reddit [95], and 4chan [75]; the use of "SMS bombers" to send thousands of text messages to a target [126]; or "zoombombing" which disrupts a video conference [101]. These attacks lead to frustration, fatigue, and a reduced sense of emotional safety. The content used may also be toxic or leaked, exacerbating the harm.

Noteworthy examples include "brigading", where a large group of people overwhelm the comment feed of a targeted group or individual (e.g., coordinated "raids" on YouTube channels by 4chan members [106]); or "dogpiling" where a person is targeted in order to recant an opinion or statement. DDoS attacks can also enable censoring by overloading an individual's network connection, preventing them from using the Internet or disabling a web site and thus making content unavailable [93]. Such attacks closely mirror for-profit DDoS attacks using botnets [87]. Finally, attacks may involve en masse negative comments and reviews, similar to Pizzagate and Gamergate [1], [85].

False Reporting [Integrity; C2]. False reporting broadly captures scenarios where an attacker deceives a reporting system or emergency service (C2)—originally intended to protect people—to falsely accuse a target of abusive behavior. Prominent examples include SWATing, where an attacker falsely claims a bomb threat, murder, or other serious crime in order to send emergency responders to the target's address. The FBI reported roughly 400 cases of SWATing in 2013 [80], and in 2017 there was one fatal incident [94]. Other forms of false reporting include when an attacker flags a piece of content or an account as abusive (for instance, on social media platforms), which we call "falsified abuse flagging". These markings may in turn trigger automated algorithms that remove the "offending" content or suspend the target's account. Past examples include a far-right group in Israel abusing Facebook's reporting tools to suspend a rival's account and to report images of his children [147]. Attackers may also file doctored evidence (e.g., "falsified abuse reports") with either platforms or police to convince an authority to take action on a target.

Impersonation [Integrity; A2 + M1 + C1]. Impersonation occurs when an attacker relies on deception of an audience (A2 + C1) to assume the online persona of a target in order to create content (M1) that will damage the target's reputation or inflict emotional harm. Satire does not meet this attack definition, as there is no intent to deceive. Attacks involving impersonation include setting up fake social media accounts purported to be associated with a target [66]; exploiting privileged access to a target's account to send emails or social media messages [66], [108]; spoofing the sender email address or phone number of a target to make it appear as if the target authored the message [66]; and setting up websites that appear to be authored by the target, often in conjunction with use of search-engine optimization (SEO) techniques to ensure impersonation websites appear in searches related to the target. For-profit equivalents of impersonation include phishing [38] and counterfeit online storefronts [46].

In addition to reputation harm and isolation, attackers may also use impersonation to physically and sexually threaten targets. In one case, an former intimate partner created dating profiles that impersonated the target to arrange for strangers to arrive at the target's house and place of work seeking intimate engagements [66], [69]. A related impersonation attack includes the synthetic generation of media depicting a target, such as "deep fakes" or "photoshopping". A study by Simonite et al. found that 96% of all deep fakes that they identified in the wild were pornographic in nature [129]. We distinguish this from disclosure of authentic but non-consensual intimate images (which falls under content leakage).

Surveillance [Confidentiality; C4 exclusively]. Surveillance involves an attacker leveraging privileged access to a target's devices or accounts (C4) to monitor the target's activities, location, or communication. In a for-profit cybercrime ecosystem, adjacent tools include keyloggers and remote access trojans that monitor a target's activities [59], [76]. Attackers can repurpose these off-the-shelf tools for hate and harassment, or alternatively subvert a target's devices such as their mobile phones [6], IoT devices [20], and GPS trackers [151] to surveil the target's activities. Indeed, Chatterjee et al. found an active ecosystem of attackers that develop "stalkerware" and who share techniques on how to subvert applications to monitor a target without their knowledge [27]. Havron et al. also reported on experiences of survivors of intimate partner violence who learned their abusers accessed remote backups (e.g., photos uploaded to iCloud) after the survivor had physically separated from the abuser [72]. Abusers may also surveil a target's finances and spending [42]. Threats in this space illustrate the challenges of practitioners designing secure software without considering hate and harassment as part of their threat model.

Lockout & Control [Integrity, Availability; A1+¬M1+C4]. Our final category of attacks includes scenarios where an attacker leverages privileged access to a target's account or device—including computers, phones, or IoT devices (C4)—to gaslight the target or interfere with how they engage with the world (A1). Such lockout and control excludes the creation of images or text (not M1); instead, attackers rely on actively subverting technology services. Passive monitoring via privileged access is covered instead by surveillance.

Examples of attacks in this category include an abusive party hijacking a smart home's microphone to broadcast profanity [15], or turning up a home's smart thermostat to 90°F [37]. Outside the IoT space, attacks include deleting communication with a target to prevent documenting and reporting abuse; controlling a target's access to online resources for help; or removing a target's access to their online accounts

(including financial resources [42])—a common threat in intimate partner violence [27], [108]. Ransomware represents the closest equivalent in a for-profit cybercrime context [90]. One survivor of intimate partner violence reported how her abuser would delete email responses to job applications in order to restrict the survivor's financial situation [108].

IV. Prevalence and At-Risk Populations

To demonstrate the global and growing threat posed by online hate and harassment, we conducted a three year survey spanning North and South America, Europe, Asia, Africa, and the Middle East to understand people's experiences with online abuse. Wherever possible, we compare our results to similar surveys conducted by Pew [118], Data and Society [44], the Anti-Defamation League [4], and Microsoft's Digital Civility Index [110].

A. Survey design

Instrument Design. Our survey asked participants "Have you ever personally experienced any of the following online?" and then listed a fixed set of experiences that participants could select from. We refer readers to the Appendix for our full survey question. We developed our survey to include five of the six experiences used by Pew in 2014 [118] to enable replication and comparison to their metrics.[2] To this end we also inherited some of their limitations, including asking if this behavior was experienced (prevalence only) and not measuring frequency or severity. We did expand the set to include eight other experiences related to lockout and control, surveillance, content leakage, impersonation, and a deeper treatment of toxic content beyond just name calling (as used by earlier works). However, as our survey precedes the construction of our final taxonomy, we lack a one-to-one mapping between the attacks outlined in our taxonomy and those appearing in our survey.

Country Selection. We selected countries for inclusion seeking diversity across a number of features: multiple regions of the world, measures of development (HDI), cultural and legal responses to online content, and through conversations with experts, as well as ability to survey using high-quality panels within a nation. To maximize the number of countries included, some countries do not appear in our sample every year. Table V in the Appendix contains our final sample size per country and the year it was collected, along with the unweighted demographic breakdown averaged across the entire survey period.

Survey Deployment. We conducted this survey in coordination with an industry leading market research firm, of which experiences with online abuse was just one segment in the context of a broader survey of privacy attitudes.[3] After completion of the entire survey instrument in English, the research team worked with in-country translation teams (through our research vendor partner) to then cover 22 countries. When the instrument was fully translated, it was then sent to a second in-country translation team for back-translation into English, which we reviewed and iterated on, as needed. Two earlier iterations of the survey were conducted in 2015 to validate and further refine the instrument, both within and outside of the US. In consultation with our research vendor partner, and their in-country teams, we aligned on the demographic traits that we could safely ask of participants in each country, using their standard demographic measures and survey items.[4]

With the exception of the US, all respondents were sourced directly from high quality, opt-in panels; that is previously created panels of volunteers willing to participate in surveys and market research. Across the 22 countries we used a combination of these panels from six different providers, all subcontracted through our research vendor partner). Consistent with the best panels available for online market research, such panels tend to be broadly representative of the general population in countries with high access to technology, but less representative of the general population in countries with more limited access to technology; for example, in developing countries they tend to skew urban. Respondents were recruited using stratified sampling with fixed quotas on country, age, and gender in each country. In the United States, we used a nationally representative panel that represents an accurate demographic probability-based sample, based on residential addresses. After data collection was completed, in each year, we followed standard procedures to apply a modest weighting adjustment to each respondent so that the samples in each country were more representative.

All participants were paid, with incentives differing by panel and by country, often through point systems which can be exchanged for products or gift cards through vendor partnerships, at an industry-standard amount within their market.

Respondent demographics. In aggregate, across all countries and years: 53% of participants identified as men; 47% as women; and due to the use of this question for stratification, we were unable to collect gender beyond binary. For age, overall, our sample is largely representative of the online populations in these countries. Aggregated: 15% 18–24; 29% 25–34; 23% 35–44; 15% 45–54; 11% 55–65; and 6% over 65. The sample is somewhat skewed to high educated participants, with 75% having some college education or higher; 25% have secondary education or lower. Regarding sexuality: 82% identified as heterosexual; 11% preferred not to say; and 7% were LGBTQ+. For full details by country, see the Appendix Table V.

B. Estimating Prevalence and Growth

We present a breakdown of prevalent online hate and harassment experiences reported by participants in Table II.

[2]We included a modified version of their sixth experience, shifting the item from general embarrassment to embarrassment caused by the posting of a private photo.

[3]The abuse experience question is just one of 60 items that was asked.

[4]For example, we do not ask about LGBTQ+ identification in China, Indonesia, the Kingdom of Saudi Arabia, or Russia; for a complete list of countries excluded, see the Appendix Table V.

Type	Abuse mechanism	2016–2018 Global	2016–2018 US-only	Pew 2017 (US-only)	DS 2016 (US-only)	ADL 2018 (US-only)	DCI 2018 (Global)
Moderate	Been exposed to unwanted explicit content	19%	16%	–	–	–	23%
	Been insulted or treated unkindly	16%	14%	–	–	–	–
	Had someone make hateful comments	16%	14%	–	–	–	–
	Been called offensive names†	14%	13%	27%	25%	41%	20%
	Been concerned because specific information about me appeared on the Internet	11%	8%	–	–	–	–
Severe	Been stalked†	7%	5%	7%	8%	18%	5%
	Had an account hacked by someone I know	6%	3%	–	–	–	
	Been sexually harassed†	6%	3%	6%	8%	18%	–
	Been harassed or bullied for a sustained period†	5%	4%	7%	5%	17%	4%
	Had someone post private photos of me to embarrass me	5%	3%	–	5%	–	3%
	Been impersonated by someone I know	5%	2%	–	6%	–	–
	Been physically threatened†	4%	2%	10%	11%	22%	5%
	Had someone I know use spyware to monitor my activities	4%	1%	–	–	–	4%
Aggregate	Been target of any online abuse	48%	35%	41%	36%	53%	40%
	Been target of any moderate online abuse	40%	32%	22%	–	–	–
	Been target of any severe online abuse	25%	13%	18%	–	37%	–

TABLE II: Frequency that participants reported experiencing hate and harassment online. We compare our results against previous surveys. We denote questions where the framing exactly matches a previous PEW survey with a dagger †. Our question framing differs from the other listed surveys, though the abuse mechanisms studied overlap.

Globally, an average of 48% of people across the 22 countries we surveyed reported experiencing some form of hate and harassment.[5] In line with a previous survey by Pew [118], we split experiences into a "moderate" category to indicate less severe forms of harassment with respect to harms or intensity, and "severe" to indicate extreme forms of harassment. Table II details the attacks that fall into each category. Of participants, 40% reported moderate experiences of hate and harassment, and 25% reported severe experiences, most frequently stalking (7%), account hijacking by someone the participant knew (6%), and sexual harassment (6%).

Many moderate hate and harassment experiences reported were brief and isolated incidents. Just 11% of participants who reported moderate harassment also reported being harassed or bullied for a sustained period. Another 58% never reported any form of severe hate or harassment. We also find that experiences with hate and harassment are often isolated to just one or two distinct experiences (e.g., stalking, sexual harassment). Of participants that encountered any hate and harassment, 43% reported experiencing only one type of attack, and 65% two or fewer. Restricting our analysis to those that reported severe hate and harassment, 85% of participants reported two or fewer types of attacks. This observation is critical when designing solutions for targets of hate and harassment as experiences are varied and non-overlapping.

Growth over time. For the 12 countries with data from both 2016 and 2018, participants reporting hate and harassment increased from 45% to 49% ($p < 0.0001$). The largest statistically significant growth ($p < 0.0001$) was in France (41% increase), Germany (41% increase), and the UK (38% increase). To discount other potential explanations for this growth (e.g., increasing social media usage, or changing demographics in the region), we modeled the outcome of experiencing any form of hate and harassment as a binomial distribution $Y_i \sim B(n_i, \pi_i)$ using a logarithmic link function. The model's parameters consist of categorical variables related to a participant's age, gender, and country of residence; and whether the participant self-identified as LGBTQ+, how frequently the participant reported using social media, and the year the survey was conducted.[6] Table III shows our results, with more detailed model parameters and significance testing available in the Appendix. Holding all variables other than time constant, we find that the odds of experiencing abuse in 2018 were 1.30 times higher than in 2016 ($p < 0.0001$). This shows that demographic shifts and changing social media usage alone cannot account for the increase in harassment year over year. Instead, incidents of hate and harassment continue to grow, suggesting that existing solutions are failing to stem its rise.

Comparison with other estimates. Of the other existing survey instruments that measure hate and harassment, only Microsoft's Digital Civility Index (DCI) tracks global experiences in a distinct set of 22 countries [110]. Their survey found 40% of participants reported "behavioral risks" such as bullying, stalking, and physical threats; and another 34% some form of "sexual risk" such as unwanted explicit content and unwanted attempts to form romantic partnerships. For attacks that overlap in both our survey and the DCI survey, we find similar rates as shown in Table II.

If we narrow our focus to only the US—the same as Pew [118], Data and Society (DS) [44], and the Anti-Defamation League (ADL) [4]—we find participants in our

[5]When calculating global averages, we first calculate the weighted mean per country over 2016–2018 to account for underrepresented demographics, and then calculate the mean across every country. This approach avoids underrepresenting any one country due to uneven sample sizes, such as when a country like Spain or Saudi Arabia is not present every year of the survey.

[6]When building this model, we excluded countries where we did not collect LGBTQ+ information.

Demographic	Treatment	Reference	Odds Any	Odds Moderate	Odds Severe
Gender	Male	Female	1.13	1.15	0.93*
LGBTQ+	LGBTQ+	non-LGBTQ+	1.86	1.55	2.12
Age	18-24	65+	3.99	3.46	5.41
	25-34	65+	3.39	2.91	4.86
	35-44	65+	2.36	2.16	3.34
	45-54	65+	1.71	1.66	2.26
	55-64	65+	1.16*	1.18*	1.36*
Social media usage	Daily	Never	2.48	2.38	2.71
	Weekly	Never	2.29	2.05	2.67
	Monthly	Never	1.89	1.65	2.44
Year	2017	2016	1.23	1.14*	1.30
	2018	2016	1.30	1.24	1.25

TABLE III: Increase in odds of experiencing hate and harassment online according to a binomial regression model. All values have significance $p < 0.0001$ unless otherwise noted with an asterisk.

survey reported roughly half the likelihood of specific attacks compared to these prior surveys. This holds even when the survey item terminology such as "been called offensive names" or "been physically threatened" were constant across survey instruments. Our results may be a conservative estimate, with rates that seem low and include a broader spectrum of the US population, while the ADL results should be read as an upper bound. Our rates of overall abuse compare similarly to Pew and DS, but more work is needed to better understand how the population interprets these concerns, questions of priming in the survey design, and the evolution of the threat over time to better estimate hate and harassment.

C. Identifying at-risk demographics

The self-reported demographics provided by participant suggest that LGBTQ+ populations, young adults, and active social media users are far more likely to report experiencing hate and harassment than other demographic groups (Table IV). These variations persist, even when holding all other explanatory variables constant as captured by our earlier model in Table III. We only discuss statistically significant variations where $p < 0.0001$.

Gender. Men in our survey reported slightly elevated rates of online hate and harassment compared to women (49% vs. 46%), though both genders reported similar rates of severe abuse. In particular, men were more likely to report being physically threatened (45% increase vs. women) and being called offensive names (26% increase). Women were more likely to report sexual harassment (114% increase vs. men) and stalking (41% increase). This stratification of experiences between men and women was also reported by other survey instruments [44], [106], [110], [118]. These results highlight that it is critical to avoid potential stereotypes of who faces harassment online, and that experiences differ across genders.

LGBTQ+. Of participants who self-identified as LGBTQ+ across the 15 countries where we collected such information, 60% reported experiencing some form of online hate and harassment, compared to 41% for non-LGBTQ+ participants (47% increase). Severe forms of harassment were especially pernicious among LGBTQ+ participants (85% increase vs. non-LGBTQ+). The top three heightened threats included sexual harassment (173% increase vs. non-LGBTQ+), the leakage of private photos to embarrass the participant (154% increase), and being harassed or bullied for a sustained period (118% increase). This elevated risk holds across ages, genders, and other factors as highlighted in Table III, where the odds of experiencing harassment increase by 1.86 times for LGBTQ+ people ($p < 0.0001$). Similar to our results, Data and Society found that LGBTQ+ people were more likely to face 18 of the 20 types of harassment they surveyed (the exceptions being account hijacking, which overlaps with cybercrime, and tracking, which overlaps with advertising) [44]. As such, LGBTQ+ populations represent a unique at-risk group that needs additional attention when designing potential solutions.

Age. We find that young adults aged 18–24 as well as participants aged 25–44 reported higher rates of any harassment (60% and 53%) compared to participants aged 45 and older (35%). In particular, participants aged 18–24 reported higher rates of sexual harassment (200% increase vs. ages 45+), the leakage of private photos to embarrass the participant (182% increase), and sustained harassment (162% increase). This heightened risk persists even when holding all other factors constant, with the odds of harassment increasing by 3.99 times ($p < 0.0001$) compared to participants aged 65 and older (Table III). These odds decrease steadily as a function of age, indicating a potential gap between the behaviors of young people online compared to older generations. Both Pew and Data and Society also found that people under 30 in the United States were far more likely to report hate and harassment [44], [118].

Social media usage. Social media usage was pervasive among the countries we surveyed: 73% of participants reported daily usage and another 12% weekly usage. Just 9% of participants self-reported never using social media. We find the most active social media users experience heightened levels of hate and harassment, with 50% of daily users experiencing harassment, compared to 25% of participants who never use social media. The largest increase in risk was associated with the leakage of private photos to embarrass the participant (251% increase vs. non-social media users). This higher incident rate of harassment holds across all levels of social media activity, even when taking into account all other explanatory variables (Table III). A related survey by the Anti-Defamation League found that, of daily social media users on different platforms, 47% of Twitch users (a video gaming community) reported experiencing harassment on the platform, compared to 38% of Reddit users, and 37% of Facebook users [4]. The lowest incident rate reported was for YouTube, at 15%. These results highlight that social media platforms in particular can potentially have an out-sized role in tackling hate and harassment online. Likewise, the design or audience of a platform can heavily influence interactions between people, as we discuss

Hate and harassment experiences	Gender		LGBTQ+		Age			Social Media Usage			
	Female	Male (Ref)	Non-LGBTQ+	LGBTQ+ (Ref)	18-24	25-44	45+ (Ref)	Daily	Weekly	Monthly	Never (Ref)
Been exposed to unwanted explicit content	19%***	19%	14%	20%	21%	20%	16%	20%	16%	15%	9%
Been insulted or treated unkindly	15%**	16%	15%	26%	24%	17%	10%	17%	13%	11%	6%
Had someone make hateful comments	15%	17%	14%	24%	21%	17%	12%	18%	13%	12%	6%
Been called offensive names	12%	15%	12%	22%	20%	15%	8%	15%	12%	11%	6%
Been concerned because specific information about me appeared on the Internet	11%***	11%	9%	15%	13%	13%	8%	12%	10%	9%	5%
Been stalked	8%	6%	5%	10%	10%	8%	4%	8%	7%	6%	4%
Had an account hacked by someone I know	6%	7%	5%	7%	9%	7%	4%	6%	6%	6%	3%
Been sexually harassed	8%	4%	5%	13%	9%	7%	3%	6%	5%	5%	2%
Been harassed or bullied for a sustained period	6%***	5%	4%	10%	8%	6%	3%	6%	5%	4%	3%
Had someone post private photos of me to embarrass me	5%	6%	5%	12%	8%	6%	3%	6%	5%	4%	2%
Been impersonated by someone I know	4%	5%	3%	7%	6%	5%	3%	5%	5%	4%	2%
Been physically threatened	3%	5%	4%	7%	6%	5%	3%	4%	4%	4%	2%
Had someone I know use spyware to monitor my activities	3%	4%	3%	6%	4%	5%	2%	4%	5%	5%	1%
Been target of any online abuse	46%	49%	41%	60%	60%	53%	35%	50%	48%	43%	25%
Been target of any moderate online abuse	39%	42%	34%	50%	50%	44%	31%	43%	38%	34%	21%
Been target of any severe online abuse	25%***	25%	20%	37%	34%	29%	15%	26%	26%	24%	11%

TABLE IV: Hate and harassment reported online across demographic subgroups. When calculating significance, we compare all values against a reference group (Ref). When reporting significance, no asterisk indicates $p < 0.0001$, * indicates $p < 0.001$, ** indicates $p < 0.01$, and *** indicates indicates $p >= 0.01$.

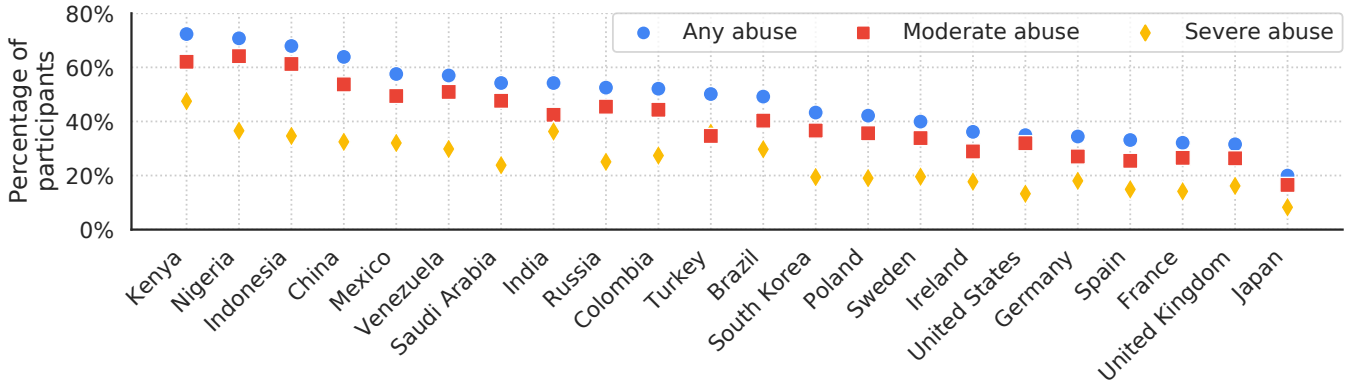

Fig. 2: Percentage of participants reporting any, moderate, or severe hate and harassment online per country, aggregated over 2016–2018.

in Section V.

Race and ethnicity. For the United States, we found no statistically significant difference between the prevalence of hate and harassment among White non-Hispanics (baseline), Black non-Hispanics ($p = 0.22$), and Hispanic peoples ($p = 0.31$). Data and Society also found no significant difference among hate and harassment experiences across these same ethnic groups, but did find Black non-Hispanics were more likely to report witnessing hate and harassment online [44]. Additionally, when participants were the target of harassment, Pew found that Black non-Hispanics and Hispanics were more likely to report the harassment was a result of their race or ethnicity (25% of Black non-Hispanic adults, 10% of Hispanic adults) compared to White non-Hispanics (just 3%) [118]. As such, it is also important for solutions to take into account the varied motivations for hate and harassment when designing interventions.

D. Variations around the world

We present a breakdown of the prevalence of hate and harassment across the 22 countries we surveyed in Figure 2. Participants from Kenya reported the highest prevalence of harassment (72%), while participants from Japan reported the lowest prevalence (20%). Our results match a previous finding on hate and harassment in South Asia, where the prevalence and severity of abuse was much greater than Western contexts [127]. When zooming in to severe issues, the relative ranking of attacks was not constant across countries. In the United Kingdom, physical threats were the most prevalent (5%), compared to sustained harassment and bullying in Ireland (6%), stalking in the United States (5%), or sexual harassment in Brazil (11%)). These variations highlight the need to tailor solutions to regional variations in hate and harassment experiences. Additionally, solutions must account for differing local interpretations of hate and harassment.

V. Towards Solutions & Interventions

We identify five directions for addressing online hate and harassment that either prevent or mitigate abuse. We synthesized these directions from the technical interventions identified during our literature review, existing approaches taken by platform operators, and potential expansions of for-profit security, privacy, and anti-abuse defenses. For a given solution, we examine which categories of abuse the solution addresses, provide evidence of early success (where possible), and suggest future directions for researchers to explore.

Solutions for hate and harassment face a unique combination of hurdles compared to other data-driven security defenses and threat models. Attackers may have *intimate access* to a target's data, devices, and social connections, or even physical access to their person. Likewise, a target's risk exposure can *span multiple platforms* (e.g., email, messaging, social media, search results), the totality of which may be targeted for attack, potentially by thousands of abusive parties. Risk can also be *highly dynamic*, with a single post or video triggering a deluge of hate and harassment, where previously the target may have been low-risk with no mitigations in place. Lastly, whereas abusive behaviors such as spam or malware have clear policy distinctions and filtering has broad support from platform users, hate and harassment is *ambiguously defined*, making it difficult to distinguish what behaviors cross the line.

A. Nudges, indicators, and warnings

Nudges and warnings provide valuable context to both abusers and targets about the risks of online hate and harassment. Strategies here hinge on prevention. For toxic content, a platform might prompt users with "Are you sure you want to post this comment?" [7], [19]. Similarly, a platform might warn abusers that posting a toxic comment will result in consequences, such as temporary disablement [19]. Bowler et al., through a design session with teens and college-age adults, synthesized such strategies into seven themes including allowing for reflection, empathy, and empowerment [18], [19]. Chang et al. found that by temporarily blocking abusive Wikipedia moderators to allow for reflection, 48% of users avoided any future incidents (but 10% left the platform) [26]. Likewise, after Reddit closed several offensive subreddits, researchers observed an 80% reduction in hate speech [23]. In terms of future directions, it remains to be determined whether such nudges deter behavior among dedicated attackers or throw-away accounts, and more generally to measure the effectiveness of any newly developed nudges.

Nudges or warnings need not be isolated to platform developers. Mathew et al. investigated the use of *counterspeech*, in which social network users countered hateful speech by directly responding to abusers [107]. Community feedback like this has previously been shown to shape user behavior [12], [33], [39], but intervention by bystanders may never manifest due to a belief that someone else will step in [48]. Difranzo et al. found 75% of participants in a user study did not intervene when they encountered another user being targeted by hate and harassment [47]. The other 25% of participants opted to flag the activity as abusive rather than engaging in any form of warning towards the abuser, or emotional support for the target [47]. A recent survey by Pew found similar results, where 70% of participants reported not intervening in any way—including flagging—after witnessing harassment [118]. Another challenge for community-based responses is that not all harassment is visible to an online audience. Finally, the subjective nature of hate and harassment may make instances difficult to identify, even when publicly visible [62].

Indicators and warnings can also surface proactive security advice. For example, two-factor authentication and security checkups can stem the risk of unauthorized access—similar to a for-profit abuse context [52]—reducing the risk of surveillance, lockout and control, and content leakage. Ensuring that visible notifications are always displayed whenever a resource (e.g., camera, GPS sensor) is being actively accessed can protect against covert access. Likewise, platforms can send users reminders about their sharing settings for sensitive content like location logs, photo backups, or delegated access to their online account to raise awareness of potential ongoing surveillance. Finally, indicators can also help to counteract impersonation, with visible indicators of trust (e.g., confirmed profiles) or influence (e.g., number of connections). In terms of future directions, research is needed to develop such indicators and identify which ones are most effective in enabling the rapid detection and prevention of harassment.

B. Human moderation, review, and delisting

The contextual nature of hate and harassment and lack of current automated solutions necessitate the use of manual review and moderation for both prevention and mitigation. Moderation is not limited to toxic content: it can also help address content leakage and impersonation via search delisting and removal, and overloading by triaging notification queues. At present, moderation is most often done at a platform level by human raters [58], [74].

We advocate for re-imagining the moderation ecosystem to one that empowers users, communities, and platforms to identify and act on hate and harassment. Such spheres of control implicitly provide more context in order to tackle the "gray areas" of hate and harassment. At a user level, this would be as simple as "I do not want to see this content", similar to existing flagging infrastructure. At a community level, the owners of a page, channel, or forum would be equipped with tools to set the tone and rules for user-generated content, and to potentially receive flag information from the community. Similar strategies are already in place for Reddit [24] and gaming platforms [104]. Such an approach enables communities to establish their own norms and rules. Finally, platform-level moderation would provide a baseline set of expectations for all user-generated content.

A multitude of systems have explored how to design collaborative moderation and reporting tools. Project Callisto allows victims of sexual assault to name their attacker, with the name revealed only if another victim (or some threshold number of victims) identify the same perpetrator [121]. Block Together

curates a crowd-sourced list of known abusive social networking accounts that can be filtered at a user level [14], [83]. HeartMob provided an interface to report hate and harassment to bystanders for confirmation and emotional support [13]. HateBase maintains a dictionary of hate speech across multiple languages that others can then integrate with for moderation or filtering [71]. Squadbox provides a tool for family and friends to step in and moderate toxic content on behalf of a target to spread the emotional burden and time required to review content [103]. Similarly, Kayes et al. explored strategies for having users directly report incidents of online hate [89]. As part of these design strategies, a common request from users is feedback—both in terms of accuracy and outcomes—to enable a sense of validation and meaningful results [7], [13]. In the absence of automated classifiers to produce moderation queues, such systems must instead rely on trusted raters that build a reputation over time as non-abusive users, in order to prevent false reporting [148]. As a future direction, research is needed to identify which tools would best enable community moderators to perform filtering or reporting. Alternatively, bug bounty programs can reward participants who identify applications that enable surveillance or lockout and control, or even entirely new vectors of hate and harassment.

C. Automated detection and curation

Another key area for development is the automated detection of hate and harassment in order to scale enforcement to billions of users and devices. Solutions in this space need not implicitly result in automated decisions like removing a post or suspending an account; instead, classifier scores can feed into moderation queues, content ranking algorithms, or warnings and nudges. Numerous studies have explored how to design classifiers to detect toxic content [32], [45], [49], [56], [77], [113], [128], [140], [142], [146] as well as word embeddings to identify toxic-adjacent content [25], [51], [113]. Other research has explored identifying abusive users and accounts, rather than individual instances of hate and harassment [28], [29], [41], [57]. Another strategy relies on predicting the targets of hate and harassment and at-risk users [31], [106]. With respect to content leakage, Facebook has explored the possibility of users providing hashes of non-consensual intimate images to enable automated detection and removal [132]. Beyond text and images, automated tools and reputation services can also play a role in detecting false reporting, surveillance, and lockout and control. Similar to a for-profit abuse context, future directions might include classifiers to identify instances of account or device takeover, or suspicious activity on an account or device.

All of the aforementioned strategies struggle with obtaining representative datasets of abusive content for training. Existing datasets of toxic content originate via crowdsourced labels of Wikipedia and news comments [84]; user-reported flags of harassment in gaming communities [11], [104]; content containing blacklisted keywords [67]; content that carries a negative sentiment score [62]; or content posted by suspended accounts (which may conflate various types of online abuse rather than solely harassment) [34]. Unlike a for-profit abuse context, bias in training data can result in classifiers incorrectly learning that terms for at-risk populations like “gay” or “black” are by default hate and harassment [5], [50]. Complexity here also stems from the fact that interpretations of hate and harassment vary across cultural contexts [97], [98] or even between the personal history of different targets [66]. Constructing unbiased and representative datasets—that either generalize or are tailored to users, communities, platforms, or regions—remains a core challenge for tackling online hate and harassment.

D. Conscious design

Designing platforms to combat hate and harassment also means consciously considering how systems and user interfaces can shape the nature of discourse in online spaces. The Anti-Defamation League has shown that experiences of hate and harassment can vary wildly by platform [4], potentially due to the communities, enforcement techniques, or design decisions involved. A fruitful area for future research may be an exploration of which design features seem to foster hate and harassment. Examples of conscious design that have recently garnered interest include whether social networks should have a “retweet” function or the ability to “subquote” other users [86]. Related considerations include how widely messages should be allowed to spread in WhatsApp [88], or whether users should have to reach a certain level of community trust—for example, subscribers on YouTube [143]—before being allowed to monetize content.

Potential design solutions for future exploration include providing targets with tools to control their audience, thus avoiding exposure to hostile parties and toxic content. Similarly, platforms might disallow sensitive material from being forwarded, preventing content leakage. Other examples include not taking automated action on user flags, or allowing people to pre-register as high risk targets to avoid false reporting, similar to anti-SWATing measures [10]. Technical measures such as cryptographic authentication on the origin of messages can also prevent spoofing and thus some forms of impersonation.

Design concepts from the privacy community can also protect users from surveillance or lockout and control. For example, delegated access to a user’s sensitive information (e.g., location, photos) might expire without that user’s explicit re-approval. This mirrors recent strategies such as automatically deleting a user’s location history after a set period [117]. Likewise, in the event of account takeover, sensitive actions such as exporting all of a user’s personal emails might require additional approval or multiple days before completing to enable detection and remediation by targets. Combined, these strategies reflect a concrete need for both platforms and the security community to re-evaluate their existing threat models when balancing utility versus safety. For new features or platforms, threat modeling for potential abuse scenarios can be just as important as modeling potential security risks.

E. Policies, education, and awareness

Apart from technical and design solutions, tackling hate and harassment also requires investing in better social structures, including policies, education resources, training, and support infrastructure [40], [64], [72], [111]. In 2016, Pater et al. found that 13 of 15 prominent social networks forbade hate and harassment, but none provided an actual definition [116]. Instead, the platforms listed potential types of abusive activities such as attacks, bullying, defamation, harm, hate, impersonation, racism, stalking, and threats [116]. The lack of well-crafted policies or definitions in turn can demoralize targets of hate and harassment [13].

VI. Tensions and Challenges

When considering the expansion of threat models and technical enforcement to address hate and harassment, there remain significant tensions around how best to balance the competing social equities at stake. Challenges also remain for how our community can safely conduct research in this space.

A. Tensions balancing social equities

Empowering vs. burdening targets. Strategies like nudges, indicators, and moderation can empower users to control their online environment. At the same time, they place much of the burden of staying safe on targets, which might exacerbate emotional tolls. Although studies have explored outsourcing this burden to family and peers [103], in general, platforms must balance between "paternalism" and overloading targets with safety decisions, while working to place burdens on attackers, not targets.

Moderation vs. filter bubbles and free speech. In providing users, communities, and platforms with technical mechanisms to filter abusive content, there is a risk that moderation turns into a form of censorship—either intentionally or unintentionally. A real example includes a prominent social network "suppress[ing] the reach of content created by users assumed to be 'vulnerable to cyberbullying"', including "disabled, queer, and fat creators" [17]. Even with well-defined policies, moderators may step beyond the bounds of their expected role to suppress content they dislike [148].

Manual review vs. well-being. Given the subjective nature of hate and harassment, manual review remains a crucial task in arriving at accurate decisions. However, this can place significant emotional burdens on reviewers and result in mental health issues without proper safeguards in place [73], [112].

Restricting content vs. research access. As platforms face pressure to remove or restrict hate and harassment in a timely fashion, researchers face an added challenge of identifying or retroactively studying attacks. A similar challenge exists presently for transparency around disinformation campaigns and preventing attacks from causing futher damage [70].

Privacy vs. accountability. Increasing user expectations around privacy add new challenges to combating abuse. Privacy-preserving technologies like end-to-end encryption provide many capabilities including secrecy, deniability, and untraceability. Secrecy precludes platforms from performing content-based analysis and filtering. Deniability compounds the difficulty of targets being able to collect and provide evidence of hate and harassment. Untraceability masks the source of hate and harassment, especially in the presence of viral distribution. Research has explored the possibility of providing weaker—but still meaningful—privacy guarantees; and anonymous tools for reporting abuse [96], [121], [139].

B. Challenges for researchers

Researcher safety and ethics. Currently, there are no best practices for how researchers can safely and ethically study online hate and harassment. Risks facing researchers include becoming a target of coordinated, hostile groups, as well as emotional harm stemming from reviewing toxic content (similar to risks for manual reviewers) [2]. Likewise, researchers must ensure they respect at-risk subjects and do not further endanger targets as they study hate and harassment.

Risks of greater harm. As platforms prevent and mitigate hate and harassment, there is a risk that the subsequent arms race escalates the severity of attacks. In particular, attackers may migrate to private, virulent communities that glorify hate and harassment. Other risks may include attackers resorting to physical or sexual violence against a target [63], [65], [66].

Defining success. When expanding threat models to include hate and harassment, it is natural to wonder what success would look like. Progress can be measured quantitatively for some areas, like measuring toxic content by the volume of people exposed to abusive messages, or qualitatively as a whole by the decrease in people reporting negative experiences online. At the same time, there are multiple other factors that underpin such metrics: the overhead and friction of solutions on platforms and users; the cost of maintaining defenses in an adversarial setting; and balancing false positives (e.g., incorrectly penalizing legitimate people) against false negatives (e.g., people exposed to hate and harassment).

VII. Conclusion

In this work, we argued that security, privacy, and anti-abuse protections are failing to address the growing threat of online hate and harassment. We proposed a taxonomy, built from over 150 research articles, to reason about these new threats. We also provided longitudinal evidence that hate and harassment has grown 4% over the last three years and now affects 48% of people globally. Young adults, LGBTQ+ individuals, and frequent social media users remain the communities most at risk of attack. We believe the computer security community must play a role in addressing this threat. To this end, we outlined five potential directions for improving protections that span technical, design, and policy changes to ultimately assist in identifying, preventing, mitigating, and recovering from hate and harassment attacks.

ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation under grants 1704527, 1748903, 1916096, 1916126, 1942610, 2016061, the Simons Foundation, and by a gift from Google. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] Sarah A. Aghazadeh, Alison Burns, Jun Chu, Hazel Feigenblatt, Elizabeth Laribee, Lucy Maynard, Amy L. M. Meyers, Jessica L. O'Brien, and Leah Rufus. *GamerGate: A Case Study in Online Harassment*, pages 179–207. Springer International Publishing, Cham, 2018.

[2] Hannah Allam. 'it gets to you.' extremism researchers confront the unseen toll of their work. https://www.npr.org/2019/09/20/762430305/it-gets-to-you-extremism-researchers-confront-the-unseen-toll-of-their-work, 2019.

[3] Ross Anderson, Chris Barton, Rainer Boehme, Richard Clayton, Michel J.G. van Eeten, Michael Levi, Tyler Moore, and Stefan Savage. Measuring the cost of cybercrime. In *Proceedings of the Workshop on Economics of Information Security*, 2012.

[4] Anti-Defamation League. Online hate and harassment: The american experience. https://www.adl.org/onlineharassment, 2019.

[5] Dennys Antonialli. Drag queen vs. david duke: Whose tweets are more 'toxic'? https://www.wired.com/story/drag-queens-vs-far-right-toxic-tweets/, 2019.

[6] Elle Armageddon. When technology takes hostages: The rise of 'stalkerware'. https://www.vice.com/en_us/article/nejmnz/when-technology-takes-hostages-the-rise-of-stalkerware, 2017.

[7] Zahra Ashktorab and Jessica Vitak. Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.

[8] BBC News. India lynchings: Whatsapp sets new rules after mob killings. https://www.bbc.com/news/world-asia-india-44897714, 2018.

[9] Zack Beauchamp. Isis captured and executed james foley and steven sotloff, two american journalists. https://www.vox.com/2018/11/20/17996042/isis-captured-and-executed-james-foley-and-steven-sotloff-two-american-journalists, 2015.

[10] Carmen Best. Protect yourself from swatting. https://www.seattle.gov/police/need-help/swatting, 2020.

[11] Jeremy Blackburn and Haewoon Kwak. Stfu noob!: predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the 23rd international conference on World wide web*, pages 877–888. ACM, 2014.

[12] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. When online harassment is perceived as justified. In *The International AAAI Conference on Web and Social Media*, 2018.

[13] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. Classification and its consequences for online harassment: Design insights from heartmob. In *Proceedings of the ACM on Human-Computer Interaction*, 2017.

[14] Block Together. A web app intended to help cope with harassment and abuse on twitter. https://blocktogether.org/, 2019.

[15] Sam Blum. Google warns Nest users to update security settings after uptick of hacked cameras. https://www.popularmechanics.com/technology/security/a26214078/google-nest-hack-warning/, 2019.

[16] Elizabeth Bodine-Baron, Todd C Helmus, Madeline Magnuson, and Zev Winkelman. Examining ISIS support and opposition networks on Twitter. Technical report, RAND Corporation Santa Monica United States, 2016.

[17] Elena Botella. TikTok admits it suppressed videos by disabled, queer, and fat creators. https://slate.com/technology/2019/12/tiktok-disabled-users-videos-suppressed.html, 2019.

[18] Leanne Bowler, Cory Knobel, and Eleanor Mattern. From cyberbullying to well-being: A narrative-based participatory approach to values-oriented design for social media. *Journal of the Association for Information Science and Technology*, 2015.

[19] Leanne Bowler, Eleanor Mattern, and Cory Knobel. Developing design interventions for cyberbullying: A narrative-based participatory approach. In *Proceedings of the iConference*, 2014.

[20] Nellie Bowles. Thermostats, locks and lights: Digital tools of domestic abuse. https://www.nytimes.com/2018/06/23/technology/smart-home-devices-domestic-abuse.html, 2018.

[21] Jarret M Brachman. High-tech terror: Al-qaeda's use of new technology. *Fletcher F. World Aff.*, 2006.

[22] Les Carpenter. Armed man charged after 'self-investigating' pizzagate conspiracy. https://www.theguardian.com/us-news/2016/dec/05/washington-pizza-child-sex-ring-fake-news-man-charged, 2016.

[23] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, (CSCW), 2017.

[24] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):32, 2018.

[25] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. The bag of communities: identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.

[26] Jonathan Chang and Cristian Danescu-Niculescu-Mizil. Trajectories of blocked community members: Redemption, recidivism and departure. In *Proceedings of the The World Wide Web Conference*, 2019.

[27] Rahul Chatterjee, Periwinkle Doerfler, Hadas Orgad, Sam Havron, Jackeline Palmer, Diana Freed, Karen Levy, Nicola Dell, Damon McCoy, and Thomas Ristenpart. The spyware used in intimate partner violence. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 441–458. IEEE, 2018.

[28] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Hate is not binary: Studying abusive behavior of #gamergate on Twitter. In *ACM Hypertext Conference*, 2017.

[29] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Mean birds: Detecting aggression and bullying on Twitter. In *ACM Web Science Conference*, 2017.

[30] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Measuring #GamerGate: A tale of hate, sexism, and bullying. In *International Conference on World Wide Web Companion*, 2017.

[31] Charalampos Chelmis and Mengfan Yao. Minority Report: Cyberbullying Prediction on Instagram. In *ACM Web Science Conference*, 2019.

[32] Hao Chen, Susan McKeever, and Sarah Jane Delany. The use of deep learning distributed representations in the identification of abusive text. In *AAAI International Conference On Web and Social Media*, 2019.

[33] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. How community feedback shapes user behavior. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[34] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Antisocial behavior in online discussion communities. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[35] Shira Chess and Adrienne Shaw. A conspiracy of fishes, or, how we learned to stop worrying about #GamerGate and embrace hegemonic masculinity. *Journal of Broadcasting & Electronic Media*, 2015.

[36] Danielle Keats Citron. Addressing cyber harassment: An overview of hate crimes in cyberspace. *Journal of Law, Technology & the Internet*, 2014.

[37] Courtney Copenhagen and Katie Kim. Homeowner's blood 'ran cold' as smart cameras, thermostat hacked, he says. https://www.nbcchicago.com/investigations/My-Blood-Ran-Cold-as-Smart-Cameras-Thermostat-Hacked-Homeowner-Says-505113061.html, 2019.

[38] Marco Cova, Christopher Kruegel, and Giovanni Vigna. There is no free phish: an analysis of "free" and live phishing kits. In *Proceedings of the Workshop on Offensive Technologies*, 2008.

[39] Tiago Oliveira Cunha, Ingmar Weber, Hamed Haddadi, and Gisele L Pappa. The effect of social feedback in a reddit weight loss community. In *Proceedings of the 6th International Conference on Digital Health Conference*, 2016.

[40] Dana Cuomo and Natalie Dolci. Gender-Based Violence and Technology-Enabled Coercive Control in Seattle: Challenges & Opportunities. https://teccworkinggroup.org/research-2, 2019.

[41] Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Canadian Conference on Artificial Intelligence*, 2014.

[42] Veronica Dagher. Financial abuse in the age of smartphones. https://www.wsj.com/articles/financial-abuse-in-the-age-of-smartphones-11575727200?mod=hp_lead_pos11, 2019.

[43] Data & Society. Nonconsensual image sharing: One in 25 americans has been a victim of "revenge porn". https://datasociety.net/pubs/oh/Nonconsensual_Image_Sharing_2016.pdf, 2016.

[44] Data & Society. Online harassment, digital abuse, and cyberstalking in america. https://datasociety.net/output/online-harassment-digital-abuse-cyberstalking/, 2016.

[45] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *AAAI International Conference On Web and Social Media*, 2017.

[46] Matthew Der, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. Knock it off: Profiling the online storefronts of counterfeit merchandise. In *Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.

[47] Dominic DiFranzo, Samuel Hardman Taylor, Franccesca Kazerooni, Olivia D Wherry, and Natalya N Bazarova. Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[48] Kelly P Dillon and Brad J Bushman. Unresponsive or un-noticed?: Cyberbystander intervention in an experimental cyberbullying context. *Computers in Human Behavior*, 2015.

[49] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. In *AAAI International Conference On Web and Social Media*, 2011.

[50] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

[51] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *The Web Conference*, 2015.

[52] Periwinkle Doerfler, Kurt Thomas, Maija Marincenko, Juri Ranieri, Yu Jiang, Angelika Moscicki, and Damon McCoy. Evaluating login challenges as adefense against account takeover. In *Proceedings of the The World Wide Web Conference*, 2019.

[53] David M. Douglas. Doxing: a conceptual analysis. *Ethics and Information Technology*, 18(3):199–210, Sep 2016.

[54] Yelena Dzhanova and Dan Mangan. Rep. katie hill's husband claimed his computer was 'hacked' before her private photos appeared online, report says. https://www.cnbc.com/2019/10/31/rep-katies-hills-husband-claimed-his-computer-was-hacked.html, 2019.

[55] Charlie Edwards and Luke Gribbon. Pathways to violent extremism in the digital era. *The RUSI Journal*, 2013.

[56] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *The International AAAI Conference on Web and Social Media*, 2018.

[57] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. Peer to peer hate: Hate speech instigators and their targets. In *AAAI International Conference On Web and Social Media*, 2018.

[58] Facebook. Community standards enforcement report. https://transparency.facebook.com/community-standards-enforcement, 2019.

[59] Brown Farinholt, Mohammad Rezaeirad, Paul Pearce, Hitesh Dharmdasani, Haikuo Yin, Stevens Le Blond, Damon McCoy, and Kirill Levchenko. To catch a ratter: Monitoring the behavior of amateur Darkcomet rat operators in the wild. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2017.

[60] Joel Finkelstein, Savvas Zannettou, Barry Bradlyn, and Jeremy Blackburn. A quantitative approach to understanding online antisemitism. In *Proceedings of the AAAI International Conference on Web and Social Media*, 2020.

[61] Claudia I Flores-Saviaga, Brian C Keegan, and Saiph Savage. Mobilizing the trump train: Understanding collective action in a political trolling community. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[62] Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *AAAI International Conference On Web and Social Media*, 2018.

[63] Cynthia Fraser, Erica Olsen, Kaofeng Lee, Cindy Southworth, and Sarah Tucker. The new age of stalking: Technological implications for stalking. *Juvenile and Family Court Journal*, 61:39 – 55, 11 2010.

[64] Diana Freed, Sam Havron, Emily Tseng, Andrea Gallardo, Rahul Chatterjee, Thomas Ristenpart, and Nicola Dell. "is my phone hacked?" analyzing clinical computer security interventions with survivors of intimate partner violence. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[65] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. Digital technologies and intimate partner violence: A qualitative analysis with multiple stakeholders. *PACM: Human-Computer Interaction: Computer-Supported Cooperative Work and Social Computing (CSCW)*, Vol. 1(No. 2):Article 46, 2017.

[66] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. "A Stalker's Paradise": How intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018.

[67] Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, et al. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, 2017.

[68] Google. Update on the Global Internet Forum to Counter Terrorism. https://www.blog.google/around-the-globe/google-europe/update-global-internet-forum-counter-terrorism/, 2017.

[69] Andy Greenberg. Spoofed Grindr Accounts Turned One Man's Life Into a 'Living Hell'. https://www.wired.com/2017/01/grinder-lawsuit-spoofed-accounts/, 2017.

[70] Sara Harrison. Twitter's disinformation data dumps are helpful—to a point. https://www.wired.com/story/twitters-disinformation-data-dumps-helpful/, 2019.

[71] Hatebase. The world's largest structured repository of regionalized, multilingual hate speech. https://hatebase.org/, 2019.

[72] Sam Havron, Diana Freed, Rahul Chatterjee, Damon McCoy, Nicola Dell, and Thomas Ristenpart. Clinical computer security for victims of intimate partner violence. In *Proceedings of the USENIX Security Symposium*, 2019.

[73] Alex Hern. Revealed: catastrophic effects of working as a Facebook moderator. The Guardian, September 2019. https://www.theguardian.com/technology/2019/sep/17/revealed-catastrophic-effects-working-facebook-moderator.

[74] Donald Hicks and David Gasca. A healthier twitter: Progress and more to do. https://blog.twitter.com/en_us/topics/company/2019/health-update.html, 2019.

[75] Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *Proceedings of the AAAI International Conference On Web and Social Media*, 2017.

[76] Thorsten Holz, Markus Engelberth, and Felix Freiling. Learning more about the underground economy: A case-study of keyloggers and dropzones. In *Proceedings of the European Symposium on Research in Computer Security*, 2009.

[77] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Detection of cyberbullying incidents on the instagram social network. In *AAAI International Conference On Web and Social Media*, 2015.

[78] Caroline Jack. Lexicon of lies: Terms for problematic information. https://datasociety.net/output/lexicon-of-lies/, 2017.

[79] Charlotte Jee. Germany's synagogue shooting was live-streamed on Twitch–but almost no one saw it. https://www.technologyreview.com/s/614529/germanys-synagogue-shooting-was-live-streamed-on-twitch-but-almost-no-one-saw-it/, 2019.

[80] Adrianne Jeffries. Meet 'swatting,' the dangerous prank that could get someone killed. https://www.theverge.com/2013/4/23/4253014/swatting-911-prank-wont-stop-hackers-celebrities, 2013.

[81] Sarah Jeong. The history of Twitter's rules. https://www.vice.com/en_us/article/z43xw3/the-history-of-twitters-rules, 2016.

[82] Sarah Jeong. *The Internet of Garbage*. Vox Media, Inc., 2018.

[83] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online harassment and content moderation: The case of blocklists. In *Proceedings of the ACM Transactions on Computer-Human Interaction*, 2018.

[84] Jigsaw. Toxic comment classification challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge, 2017.

[85] Cecilia Kang. Fake news onslaught targets pizzeria as nest of child-trafficking. https://www.nytimes.com/2016/11/21/technology/fact-check-this-pizzeria-is-not-a-child-trafficking-site.html, 2016.

[86] Alex Kantrowitz. The man who built the retweet: "we handed a loaded weapon to 4-year-olds". https://www.buzzfeednews.com/article/alexkantrowitz/how-the-retweet-ruined-the-internet, 2019.

[87] Mohammad Karami, Youngsam Park, and Damon McCoy. Stress testing the booters: Understanding and undermining the business of ddos services. In *Proceedings of the 25th International Conference on World Wide Web*, 2016.

[88] Jacob Kastrenakes. Whatsapp limits message forwarding in fight against misinformation. https://www.theverge.com/2019/1/21/18191455/whatsapp-forwarding-limit-five-messages-misinformation-battle, 2019.

[89] Imrul Kayes, Nicolas Kourtellis, Daniele Quercia, Adriana Iamnitchi, and Francesco Bonchi. Cultures in community question answering. In *ACM HyperText Conference*, 2015.

[90] Amin Kharraz, William Robertson, Davide Balzarotti, Leyla Bilge, and Engin Kirda. Cutting the gordian knot: A look under the hood of ransomware attacks. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, 2015.

[91] Jytte Klausen. Tweeting the jihad: Social media networks of western foreign fighters in syria and iraq. *Studies in Conflict & Terrorism*, 2015.

[92] Robin M Kowalski, Susan P Limber, Sue Limber, and Patricia W Agatston. *Cyberbullying: Bullying in the digital age*. Routledge, 2012.

[93] Brian Krebs. https://krebsonsecurity.com/2016/09/the-democratization-of-censorship/, 2016.

[94] Brian Krebs. Man behind fatal 'swatting' gets 20 years. https://krebsonsecurity.com/2019/03/man-behind-fatal-swatting-gets-20-years/, 2019.

[95] Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. Community interaction and conflict on the web. In *The Web Conference*, 2018.

[96] Benjamin Kuykendall, Hugo Krawczyk, and Tal Rabin. Cryptography for #metoo. *Proceedings on Privacy Enhancing Technologies*, 3:409–429, 2019.

[97] Haewoon Kwak and Jeremy Blackburn. Linguistic analysis of toxic behavior in an online video game. In *International Conference on Social Informatics*, pages 209–217. Springer, 2014.

[98] Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3739–3748. ACM, 2015.

[99] Kirill Levchenko, Andreas Pitsillidis, Neha Chachra, Brandon Enright, Márk Félegyházi, Chris Grier, Tristan Halvorson, Chris Kanich, Christian Kreibich, He Liu, Damon McCoy, Nicholas Weaver, Vern Paxson, Geoffrey M. Voelker, and Stefan Savage. Click trajectories: End-to-end analysis of the spam value chain. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2011.

[100] Lucas Lima, Julio CS Reis, Philipe Melo, Fabricio Murai, Leandro Araujo, Pantelis Vikatos, and Fabricio Benevenuto. Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 515–522. IEEE, 2018.

[101] Taylor Lorenz and Davey Alba. 'Zoombombing' becomes a dangerous organized effort. https://www.nytimes.com/2020/04/03/technology/zoom-harassment-abuse-racism-fbi-warning.html, 2020.

[102] Julia M MacAllister. The doxing dilemma: seeking a remedy for the malicious publication of personal information. *Fordham Law Review*, 2016.

[103] Kaitlin Mahar, David Karger, and Amy X. Zhang. Squadbox: A tool to combat online harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[104] Brendan Maher. Can a video game company tame toxic behaviour? *Nature News*, 2016.

[105] Enrico Mariconti, Jeremiah Onaolapo, Syed Sharique Ahmad, Nicolas Nikiforou, Manuel Egele, Nick Nikiforakis, and Gianluca Stringhini. What's in a name?: Understanding profile name reuse on twitter. In *International Conference on World Wide Web*, pages 1161–1170, 2017.

[106] Enrico Mariconti, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini. "you know what to do": Proactive detection of youtube videos targeted by coordinated hate attacks. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 2019.

[107] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. Thou shalt not hate: Countering online hate speech. In *AAAI International Conference On Web and Social Media*, 2019.

[108] Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F Churchill, and Sunny Consolvo. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2189–2201. ACM, 2017.

[109] Alexander Meleagrou-Hitchens, Shiraz Maher, and James Shaheen. *Lights, Camera, Jihad: Al-Shabaab's Western Media Strategy*. International Centre for the Study of Radicalisation and Political Violence, 2012.

[110] Microsoft. Civility, safety, and interaction online. https://www.microsoft.com/en-us/digital-skills/digital-civility, 2019.

[111] National Network to End Domestic Violence. Safety Net Project. https://nnedv.org/content/safety-net/.

[112] Casey Newton. The Trauma Floor: The secret lives of Facebook moderators in America. The Verge, February 2019. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

[113] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *The Web Conference*, 2016.

[114] Rod Nordland and Ranya Kadri. Jordanian pilot's death, shown in isis video, spurs jordan to execute prisoners. https://www.nytimes.com/2015/02/04/world/middleeast/isis-said-to-burn-captive-jordanian-pilot-to-death-in-new-video.html, 2015.

[115] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. Disturbed youtube for kids: Characterizing and detecting disturbing content on youtube. In *Proceedings of the AAAI International Conference on Web and Social Media*, 2020.

[116] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th International Conference on Supporting Group Work*, 2016.

[117] Sarah Perez. Google launches auto-delete controls for location history on iOS and Android. https://techcrunch.com/2019/06/26/google-launches-auto-delete-controls-for-location-history-on-ios-and-android/, 2019.

[118] PEW Research Center. Online harassment 2017. https://www.pewinternet.org/2017/07/11/online-harassment-2017/, 2017.

[119] Shruti Phadke and Tanushree Mitra. Many faced hate: A cross platform study of content framing and information sharing by online hate groups. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[120] Whitney Phillips. Loling at tragedy: Facebook trolls, memorial pages and resistance to grief online. *First Monday*, 2011.

[121] Anjana Rajan, Lucy Qin, David Archer, Dan Boneh, Tancrède Lepoint, and Mayank Varia. Callisto: A cryptographic approach to detect serial predators of sexual misconduct. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS)*, 2018.

[122] Helen Regan and Sandi Sidhu. 49 killed in mass shooting at two mosques in christchurch, new zealand. https://www.cnn.com/2019/03/14/asia/christchurch-mosque-shooting-intl/index.html, 2019.

[123] Kunal Relia, Zhengyi Li, Stephanie H Cook, and Rumi Chunara. Race, Ethnicity and National Origin-based Discrimination in Social Media and Hate Crimes Across 100 US Cities. In *AAAI International Conference On Web and Social Media*, 2019.

[124] Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. Characterizing and detecting hateful users on Twitter. In *AAAI International Conference on Web and Social Media*, 2018.

[125] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira. Auditing radicalization pathways on youtube. *arXiv preprint arXiv:1908.08313*, 2019.

[126] Kevin Roundy, Paula Barmaimon, Nicola Dell, Damon McCoy, Daniel Nissani, Thomas Ristenpart, and Acar Tamersoy. The many kinds of creepware used for interpersonal attacks. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2020.

[127] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. "they don't leave us alone anywhere we go": Gender and digital abuse in south asia. In *Proceedings of the Conference on Human Factors in Computing Systems*, 2019.

[128] A Saravanaraj, JI Sheeba, and S Pradeep Devaneyan. Automatic detection of Cyberbullying from Twitter. *International Journal of Computer Science and Information Technology & Security (IJCSITS)*, 2016.

[129] Tom Simonite. Most deepfakes are porn, and they're multiplying fast. https://www.wired.com/story/most-deepfakes-porn-multiplying-fast/, 2019.

[130] Robert Slonje and Peter K Smith. Cyberbullying: Another main type of bullying? *Scandinavian journal of psychology*, 2008.

[131] Peter Snyder, Periwinkle Doerfler, Chris Kanich, and Damon McCoy. Fifteen minutes of unwanted fame: Detecting and characterizing doxing. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2017.

[132] Olivia Solon. Inside Facebook's efforts to stop revenge porn before it spreads. https://www.nbcnews.com/tech/social-media/inside-facebook-s-efforts-stop-revenge-porn-it-spreads-n1083631, 2019.

[133] Kate Starbird, Ahmer Arif, and Tom Wilson. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. In *Proceedings of the Conference on Computer-Supported Cooperative Work*, 2019.

[134] Rashid Tahir, Faizan Ahmed, Hammas Saeed, Shiza Ali, Fareed Zaffar, and Christo Wilson. Bringing the kid back into youtube kids: Detecting inappropriate content on video streaming platforms. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, 2019.

[135] Rebecca Tan. Terrorists' love for telegram, explained. https://www.vox.com/world/2017/6/30/15886506/terrorism-isis-telegram-social-media-russia-pavel-durov-twitter, 2017.

[136] Kurt Thomas, Danny Yuxing Huang, David Wang, Elie Bursztein, Chris Grier, Tom Holt, Christopher Kruegel, Damon McCoy, Stefan Savage, and Giovanni Vigna. Framing Dependencies Introduced by Underground Commoditization. In *Proceedings of the Workshop on the Economics of Information Security*, 2015.

[137] Selen Türkay, Jessica Formosa, Sonam Adinolf, Robert Cuthbert, and Roger Altizer. See no evil, hear no evil, speak no evil: How collegiate players define, experience and cope with toxicity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2020.

[138] Twitter. The Twitter rules. https://help.twitter.com/en/rules-and-policies/twitter-rules, 2020.

[139] Nirvan Tyagi, Ian Miers, and Thomas Ristenpart. Traceback for end-to-end encrypted messaging. In *Proceedings of the Conference on Computer and Communications Security*, 2019.

[140] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. Automatic detection and prevention of cyberbullying. In *International Conference on Human and Social Analytics*, 2015.

[141] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2017.

[142] William Warner and Julia Hirschberg. Detecting hate speech on the World Wide Web. In *Proceedings of the second workshop on language in social media*, 2012.

[143] Chris Welch. YouTube tightens rules around what channels can be monetized. https://www.theverge.com/2018/1/16/16899068/youtube-new-monetization-rules-announced-4000-hours, 2018.

[144] Georgia Wells. Islamic state turns to teen-friendly tiktok, adorning posts with pink hearts. https://www.wsj.com/articles/islamic-state-turns-to-teen-friendly-tiktok-adorning-posts-with-pink-hearts-11571680389, 2019.

[145] Delanie Woodlock. The abuse of technology in domestic violence and stalking. *Violence against women*, 23(5):584–602, 2017.

[146] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[147] Oded Yaron. Another chapter in the facebook wars. https://www.haaretz.co.il/captain/net/1.1728851, 2012.

[148] Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, and Gianluca Stringhini. Understanding web archiving services and their (mis) use on social media. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[149] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, pages 1007–1014. International World Wide Web Conferences Steering Committee, 2018.

[150] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*. ACM, 2018.

[151] Kim Zetter. Murder suspect allegedly used gps tracker to find wife's lover. https://www.wired.com/2012/08/murder-suspect-allegedly-used-gps-tracker-to-find-wifes-lover/, 2012.

[152] Adam G Zimmerman and Gabriel J Ybarra. Online aggression: The influences of anonymity and social modeling. *Psychology of Popular Media Culture*, 2016.

APPENDIX

A. Survey instrument

Our complete survey instrument included 60 questions related to a respondent's experiences online with respect to security, privacy, and abuse. We report only the exact text of the survey components that we relied on for our study of hate and harassment. The survey instrument's questions about a respondent's age and gender were fixed by the panel provider to enable reaching a minimum stratified cross-section. As such, we were unable to ask specifically about non-binary participants at the time of our survey. Likewise, these questions did not have a "Prefer not to say" option in order to ensure sufficient samples per strata. When asking about whether a participant identified as LGBTQ+, we omitted the question in regions where such affiliations are heavily stigmatized or dangerous due to government policies, based on feedback from the panel provider. Finally, when asking about a participant's education level, we tailored the education categories per region. As such, when comparing demographics across countries, we normalized to two groups: "Secondary or less" and "Some college or more".

What is your gender? [Select one]
o Male
o Female

How old are you? [Select one]
o 18-24
o 25-34
o 35-44
o 45-54
o 55-64
o 65 or older

Please indicate the highest level of education you have attained [Select one, Europe version only]:
o Primary
o Junior secondary
o Senior secondary
o Technical
o Some college
o College graduate
o Postgraduate (Master's degree)
o Postgraduate (Doctorate degree)

Do you consider yourself to be... [Select all that apply]
o Heterosexual or straight
o Gay
o Lesbian
o Bisexual
o Transgender
o Transsexual
o Prefer not to say

How often do you use a social networking service online? [Select one]
o Multiple times per day
o About once per day
o 2-3 times per week
o About once per week
o 2-3 times per month
o About once per month
o Less than once per month
o Never

Have you ever personally experienced any of the following online? [Select all that apply]
o Been insulted or treated unkindly

o Had someone make hateful comments to me
o Been called offensive names
o Been stalked
o Been physically threatened
o Been harassed or bullied for a sustained period
o Been sexually harassed
o Had someone post private photos of me to embarrass me
o Been concerned because specific information about me appeared on the Internet
o Had an account hacked by a stranger
o Had an account hacked by someone I know
o Had someone I know use spyware to monitor my activities
o Been exposed to unwanted explicit content
o Been impersonated by someone I know

B. Unweighted survey demographics

We detail the unweighted, self-reported demographics of respondents from the 22 countries covered by our survey in Table V. In the event we surveyed the same country across multiple years, we report the aggregate, average breakdown. Our demographic profile includes a participant's gender, age, education, and whether they identify as LGBTQ+.

	Participants			**Gender**		**Age**						**Education**		**LGBTQ+**						
Country	**2016**	**2017**	**2018**	**Female**	**Male**	**18-24**	**25-34**	**35-44**	**45-54**	**55-64**	**65+**	**Secondary or less**	**Some college or more**	**Heterosexual**	**Gay**	**Lesbian**	**Bisexual**	**Transgender**	**Transexual**	**Prefer not to answer**
United States	1,221	1,100	1,101	51%	49%	6%	15%	14%	16%	24%	25%	36%	64%	90%	2%	1%	2%	0%	0%	4%
Brazil	1,207	1,114	1,113	52%	48%	24%	27%	24%	14%	9%	3%	33%	67%	84%	4%	1%	4%	0%	0%	6%
Colombia	0	1,124	0	50%	50%	22%	23%	31%	17%	6%	2%	30%	70%	84%	3%	1%	2%	0%	0%	10%
Mexico	1,113	1,140	1,107	54%	46%	20%	30%	27%	13%	7%	2%	27%	73%	84%	3%	1%	3%	0%	0%	9%
Venezuela	0	1,129	0	42%	58%	16%	20%	23%	24%	12%	5%	16%	84%	88%	3%	1%	2%	0%	0%	5%
France	1,119	1,183	1,108	53%	47%	10%	16%	21%	23%	21%	10%	24%	76%	86%	2%	1%	3%	0%	0%	8%
Germany	1,124	1,114	1,106	50%	50%	7%	18%	17%	23%	23%	12%	71%	29%	82%	2%	1%	3%	0%	0%	13%
Ireland	0	0	1,112	54%	46%	5%	19%	25%	21%	19%	10%	29%	71%	89%	2%	1%	4%	0%	1%	4%
Poland	1,240	0	1,119	53%	47%	8%	26%	29%	17%	15%	4%	12%	88%	85%	1%	0%	2%	0%	0%	11%
Russia	1,131	1,165	1,124	52%	48%	7%	36%	26%	18%	11%	1%	10%	90%	–	–	–	–	–	–	–
Spain	1,174	0	0	48%	52%	2%	21%	34%	27%	12%	4%	23%	77%	91%	3%	0%	2%	0%	0%	5%
Sweden	1,133	0	0	49%	51%	10%	25%	16%	15%	15%	19%	47%	53%	83%	2%	1%	5%	1%	0%	10%
United Kingdom	1,126	1,139	1,112	50%	50%	6%	15%	13%	19%	22%	25%	41%	59%	91%	3%	0%	2%	0%	0%	3%
Saudi Arabia	1,214	0	0	25%	75%	11%	53%	27%	7%	3%	0%	15%	85%	–	–	–	–	–	–	–
Turkey	0	0	1,150	49%	51%	15%	40%	32%	10%	3%	0%	18%	82%	–	–	–	–	–	–	–
China	1,194	1,172	1,123	46%	54%	9%	44%	32%	10%	4%	1%	8%	92%	–	–	–	–	–	–	–
India	1,660	1,819	1,803	41%	59%	35%	40%	17%	5%	3%	1%	9%	91%	66%	1%	1%	6%	1%	0%	25%
Indonesia	1,373	1,204	1,111	42%	58%	18%	41%	29%	9%	3%	0%	22%	78%	–	–	–	–	–	–	–
Japan	1,117	1,112	0	47%	53%	10%	24%	20%	22%	14%	10%	43%	57%	62%	0%	1%	2%	0%	0%	32%
South Korea	1,126	1,118	1,106	44%	56%	13%	27%	30%	21%	8%	1%	20%	80%	84%	1%	1%	4%	0%	0%	10%
Kenya	0	1,188	0	48%	52%	34%	39%	20%	5%	2%	1%	8%	92%	–	–	–	–	–	–	–
Nigeria	0	1,188	1,150	29%	71%	29%	45%	21%	4%	1%	0%	9%	91%	–	–	–	–	–	–	–

TABLE V: Unweighted survey demographics averaged across all three years of our survey. We include a participant's reported gender, age, education, and LGBTQ+ status. We note that due to limitations with our panel provider, we were unable to survey for non-binary participants. We specifically avoided asking whether participants identified as LGBTQ+ in regions where such affiliations are heavily stigmatized or dangerous due to government policies; we note these with a blank symbol (–).

C. Regression tables

We present the full parameters and resulting outputs for our logistic regression models that predict the likelihood of experiencing (1) or not experiencing (0) online hate and harassment. We report the independent variable (all categorical variables), our baseline of comparison, the model coefficient (β), standard error (SE), the z-score (z), p-value, and the resulting odds ratio (OR). Our models include: whether a participant would experience any form of online hate and harassment (Table VI); whether a participant would experience any form of severe hate and harassment (Table VII); and whether a participant would experience any form of moderate hate and harassment (Table VIII). Our definition of moderate or severe abuse were selected to enable comparison with previous survey results from Pew [118], and are not a value statement on the intensity of abuse.

Category	**Independent Variable**	**Baseline**	β	SE	z	$Pr(>\|z\|)$	OR
Country	Brazil	United States	0.0779	0.0550	1.4169	0.1565	1.0810
	Columbia	United States	0.3175	0.0795	3.9956	0.0001	1.3737
	Germany	United States	-0.0650	0.0564	-1.1537	0.2486	0.9370
	Spain	United States	-0.2020	0.0789	-2.5596	0.0105	0.8171
	France	United States	-0.3347	0.0561	-5.9719	$<$ 0.0001	0.7155
	Ireland	United States	-0.2060	0.0801	-2.5730	0.0101	0.8138
	India	United States	0.3851	0.0531	7.2513	$<$ 0.0001	1.4697
	Japan	United States	-0.6676	0.0771	-8.6630	$<$ 0.0001	0.5129
	South Korea	United States	0.1537	0.0798	1.9256	0.0542	1.1661
	Mexico	United States	0.4114	0.0558	7.3756	$<$ 0.0001	1.5089
	Poland	United States	0.0154	0.0615	0.2509	0.8019	1.0155
	Sweden	United States	0.1888	0.0798	2.3644	0.0181	1.2078
	United Kingdom	United States	-0.3314	0.0558	-5.9388	$<$ 0.0001	0.7179
	Venezuela	United States	0.5090	0.0786	6.4794	$<$ 0.0001	1.6636
Age	55-64	65+	0.1460	0.0608	2.4000	0.0164	1.1571
	45-54	65+	0.5366	0.0581	9.2392	$<$ 0.0001	1.7102
	35-44	65+	0.8574	0.0569	15.0751	$<$ 0.0001	2.3571
	25-34	65+	1.2219	0.0556	21.9843	$<$ 0.0001	3.3937
	18-24	65+	1.3838	0.0596	23.2171	$<$ 0.0001	3.9899
Gender	Male	Female	0.1193	0.0242	4.9214	$<$ 0.0001	1.1267
Social media usage	Monthly	Never	0.6357	0.0665	9.5626	$<$ 0.0001	1.8884
	Weekly	Never	0.8297	0.0571	14.5398	$<$ 0.0001	2.2927
	Daily	Never	0.9100	0.0491	18.5375	$<$ 0.0001	2.4842
LGBTQ+	LGBTQ+	Non-LGTBQ+	0.6182	0.0464	13.3348	$<$ 0.0001	1.8556
Year	2017	2016	0.2105	0.0338	6.2318	$<$ 0.0001	1.2343
	2018	2016	0.2638	0.0328	8.0384	$<$ 0.0001	1.3018

TABLE VI: Logistic regression predicting a participant experiencing any form of online hate and harassment (1) or never experiencing hate and harassment (0) based on their demographic profile.

Category	Independent Variable	Baseline	β	SE	z	$Pr(>\|z\|)$	OR
Country	Brazil	United States	0.5076	0.0678	7.4808	< 0.0001	1.6612
	Columbia	United States	0.4944	0.0937	5.2758	< 0.0001	1.6395
	Germany	United States	0.3198	0.0730	4.3822	< 0.0001	1.3768
	Spain	United States	0.0635	0.1041	0.6100	0.5418	1.0656
	France	United States	-0.1300	0.0753	-1.7261	0.0843	0.8781
	Ireland	United States	0.1399	0.1025	1.3655	0.1721	1.1502
	India	United States	0.9124	0.0646	14.1203	< 0.0001	2.4902
	Japan	United States	-0.4582	0.1083	-4.2288	< 0.0001	0.6324
	South Korea	United States	0.2332	0.1026	2.2729	0.0230	1.2626
	Mexico	United States	0.6077	0.0678	8.9586	< 0.0001	1.8363
	Poland	United States	0.1947	0.0793	2.4569	0.0140	1.2150
	Sweden	United States	0.4409	0.1014	4.3490	< 0.0001	1.5541
	United Kingdom	United States	0.0659	0.0736	0.8958	0.3703	1.0681
	Venezuela	United States	0.5935	0.0912	6.5063	< 0.0001	1.8103
Age	55-64	65+	0.3069	0.0979	3.1343	0.0017	1.3592
	45-54	65+	0.8133	0.0919	8.8453	< 0.0001	2.2553
	35-44	65+	1.2054	0.0891	13.5219	< 0.0001	3.3380
	25-34	65+	1.5807	0.0874	18.0789	< 0.0001	4.8584
	18-24	65+	1.6874	0.0899	18.7737	< 0.0001	5.4055
Gender	Male	Female	-0.0677	0.0285	-2.3798	0.0173	0.9345
Social media usage	Monthly	Never	0.8928	0.0945	9.4456	< 0.0001	2.4418
	Weekly	Never	0.9839	0.0833	11.8062	< 0.0001	2.6748
	Daily	Never	0.9968	0.0754	13.2210	< 0.0001	2.7096
LGBTQ+	LGBTQ+	Non-LGTBQ+	0.7507	0.0465	16.1335	< 0.0001	2.1186
Year	2017	2016	0.2646	0.0397	6.6555	< 0.0001	1.3029
	2018	2016	0.2209	0.0386	5.7190	< 0.0001	1.2472

TABLE VII: Logistic regression predicting a participant experiencing any severe form of online hate and harassment (1) or never experiencing severe hate and harassment (0) based on their demographic profile.

Category	Independent Variable	Baseline	β	SE	z	$Pr(>\|z\|)$	OR
Country	Brazil	United States	-0.1168	0.0555	-2.1037	0.0354	0.8898
	Columbia	United States	0.1880	0.0794	2.3666	0.0180	1.2069
	Germany	United States	-0.2804	0.0580	-4.8316	< 0.0001	0.7555
	Spain	United States	-0.4770	0.0831	-5.7395	< 0.0001	0.6206
	France	United States	-0.4545	0.0575	-7.9047	< 0.0001	0.6347
	Ireland	United States	-0.3792	0.0827	-4.5871	< 0.0001	0.6844
	India	United States	0.0606	0.0531	1.1414	0.2537	1.0625
	Japan	United States	-0.7373	0.0802	-9.1943	< 0.0001	0.4784
	South Korea	United States	-0.0093	0.0813	-0.1144	0.9089	0.9907
	Mexico	United States	0.2431	0.0555	4.3775	< 0.0001	1.2753
	Poland	United States	-0.1076	0.0624	-1.7244	0.0846	0.8980
	Sweden	United States	0.0432	0.0811	0.5324	0.5944	1.0441
	United Kingdom	United States	-0.4322	0.0572	-7.5548	< 0.0001	0.6491
	Venezuela	United States	0.4464	0.0779	5.7319	< 0.0001	1.5626
Age	55-64	65+	0.1693	0.0640	2.6458	0.0081	1.1844
	45-54	65+	0.5076	0.0611	8.3124	< 0.0001	1.6613
	35-44	65+	0.7716	0.0596	12.9367	< 0.0001	2.1633
	25-34	65+	1.0686	0.0582	18.3754	< 0.0001	2.9113
	18-24	65+	1.2399	0.0617	20.1039	< 0.0001	3.4552
Gender	Male	Female	0.1440	0.0246	5.8592	< 0.0001	1.1549
Social media usage	Monthly	Never	0.5028	0.0702	7.1653	< 0.0001	1.6533
	Weekly	Never	0.7190	0.0598	12.0289	< 0.0001	2.0524
	Daily	Never	0.8668	0.0516	16.7859	< 0.0001	2.3792
LGBTQ+	LGBTQ+	Non-LGTBQ+	0.4365	0.0447	9.7626	< 0.0001	1.5473
Year	2017	2016	0.1312	0.0343	3.8238	0.0001	1.1402
	2018	2016	0.2121	0.0332	6.3930	< 0.0001	1.2362

TABLE VIII: Logistic regression predicting a participant experiencing any moderate form of online hate and harassment (1) or never experiencing moderate hate and harassment (0) based on their demographic profile.