Content-Oblivious Trust and Safety Techniques: Results from a Survey of Online Service Providers

Riana Pfefferkorn

Abstract. We present the results of a survey about the trust and safety techniques of a group of online service providers that collectively serve billions of users. We classify techniques that require the provider to be able to access the contents of users' files and communications at will as *content dependent*, and *content oblivious* otherwise. We find that more providers use abuse-reporting features (which are content oblivious) than other abuse-detection techniques, but that participants' abuse-reporting tools do not consistently cover the types of abuse that users may encounter. We also find that, despite strong consensus among participating providers that automated content scanning (which is content dependent) is the most useful means of detecting child sex abuse imagery, they do not consider it to be nearly as useful for other kinds of abuse. These results indicate that content-dependent techniques do not constitute a silver bullet to protect users against abuse. They also demonstrate that the impact of end-to-end encryption (which, controversially, impedes outside access to user content) on abuse detection may vary by abuse type. These findings have implications for policy debates over the regulation of online service providers' anti-abuse obligations and their use of end-to-end encryption.

1 Introduction

With billions of people worldwide now using the internet (DataReportal 2021), some online activity will inevitably be abusive. To contend with abusive user behavior, the providers of many online services adopt a set of business practices and policies that are collectively known as their "trust and safety" (T&S) program (Feerst 2019). This paper investigates the T&S practices of a group of online service providers that collectively serve a significant proportion of the world's internet users. This research seeks to highlight some lesser-known approaches to T&S that are pertinent to ongoing policy debates about regulatory responses to online abuse. Contrary to what policymakers may believe, our findings indicate that effective abuse detection does not necessarily require providers to have access to user content; nor does end-to-end encryption (E2EE) impede abuse detection efforts as much as is usually assumed.

10.54501/jots.v1i2.14

How do providers carry out their T&S mission? It varies by service. An abusive public post on a social media platform, for example, can be spotted by the platform's provider. By contrast, the provider of an E2EE messaging app cannot read the user communications passing through its system (Internet Society, n.d.); other approaches to fighting abuse are thus needed.

We define T&S strategies in which the provider cannot (or does not) access user content on its own as *content oblivious*. *Content-dependent* techniques require the provider to have the technical ability to access the contents of users' files and communications at will (i.e., without needing the user or another third party to actively share particular content with the provider). This is typically possible on non-E2EE services. Examples of content-dependent T&S techniques include early spam filters that scanned the contents of email messages (Cranor and LaMacchia 1998) and human moderators who monitor activity on a service (e.g., Reddit "mods"; Pierce (2020)).

A provider may also use content-oblivious strategies to fight abuse, for example if it lacks the technical capability to access user content (i.e., due to E2EE), if it chooses not to access content as a matter of policy (e.g., bulletproof hosting services, CyberBunker.com (2016)), or if it also employs content-dependent techniques.

To investigate the prevalence and utility of content-oblivious T&S techniques, we conducted a survey of online service providers—which we define as "providers of apps and services for online communications and/or data storage"—from April to June 2021.[1] Our analysis includes the responses of 13 individuals from 11 companies and organizations, ranging from a website run by a single individual with about 2,000 monthly active users (MAU) to a private messaging service with 2 billion MAU.

The survey focused primarily on two T&S techniques that we deem content oblivious: metadata[2] and user reports. User reporting may not seem "content oblivious" at first glance, since the provider will see the contents of a message or file that a user (or outside third party) has reported as abusive. Nevertheless, we categorize it in this way because the provider either cannot access the abusive content on its own (in E2EE systems)[3] or does not discover the abuse using its own capability to access the content (in non-E2EE systems); in both cases, the provider obtains knowledge of the abusive content because someone else brought it to their attention.

Our survey posed several major questions: Which techniques (content oblivious or content dependent) do providers use to detect, prevent, and mitigate abuse? Which types of abuse do providers try to detect using content-oblivious techniques, specifically metadata analysis and user reports? Do providers consider content-oblivious techniques (metadata or user reports) or content-dependent ones (automated content scanning) to be most useful for detecting each type of abuse?

All of the providers surveyed reported using both techniques. More employ user reporting than any other method (whether content oblivious or content dependent). They all offer

1. This definition of "online service" is narrower than the definition of "interactive computer service" in Section 230 of the Communications Decency Act of 1996, the U.S. federal law that protects the providers of such services from legal liability for most kinds of abusive user material. Our definition is more akin to a combination of the definitions of "electronic communication service" and "remote computing service" found in U.S. federal electronic privacy law (Stored Communications Act, Stat. [1988], https://www.law.cornell.edu/uscode/text/18/2711; Wiretap Act, Stat. [1986], https://www.law.cornell.edu/uscode/text/18/2510).

2. "Content" and "metadata" are not stable categories; the same data can be characterized as either depending on the context (Bellovin et al. 2016). We consider metadata to be *information about a message*, file, or user, as distinguished from the *information in the message* or file, which we consider content. Thus a picture transmitted in the body of a message is content, while a picture used as the avatar for a user or user group would be considered metadata.

3. Systems for enabling and authenticating user reports can be implemented even in E2EE services (Mayer 2019).

some kind of abuse reporting; almost all include an in-app abuse reporting feature. Except for a few categories such as child sex abuse imagery (CSAI) where automated content scanning either predominates or ties with other methods, we find that providers consider user reporting to be the most useful means of detecting most types of abuse; yet their user-reporting tools do not consistently enable reporting of all the relevant types of abuse. Furthermore, despite a strong consensus among the providers that automated content scanning is the most useful means of detecting CSAI, we find the consensus is weaker in relation to other child safety offenses such as grooming and enticement, which we group under the term "child sexual exploitation" (CSE).

These findings have ramifications for salient global policy debates over the regulation of online service providers' anti-abuse obligations and over whether to regulate E2EE given its supposed effect on the detection of harmful or criminal conduct. These findings also illuminate changes that providers should make to improve the tools they supply to users for reporting abuse. Our goal is to contribute findings that help improve understanding among both policymakers and the general public of the prevalence and perceived usefulness of non-content-dependent T&S tools in fighting different types of abuse.

2 Motivation and Related Work

Online service providers are currently under intense scrutiny from regulators all around the world over how they deal with users' abusive or criminal conduct. Governments have pressured providers to better police their services for content that violates the law or the provider's terms of service (Morar and Santos 2020). In the United States, where providers are legally immune from liability for most types of abusive user content, legislators have repeatedly proposed to pare back those protections and increase providers' accountability (D. Keller 2020; Pfefferkorn 2020b). Providers have also come under fire from governments worldwide for using E2EE to protect their users' privacy and security because it makes it more difficult for providers and law enforcement to monitor users' communications (Perlroth 2019).

In these distinct but interrelated policy debates, government officials tend to focus on one particular approach to countering abuse or crime on a service: providers' mass automated scanning of files and communications to proactively detect abusive content (Llansó 2020). We use "automated content scanning" as an umbrella term that covers multiple techniques with varying degrees of technical sophistication that providers have deployed against different kinds of online abuse (Llansó et al. 2020). These include fingerprinting and hashing technologies, which compare an item's characteristics to those in a database of known, verboten content (commonly used for CSAI, copyright infringement, and terrorist content);[4] naïve Bayesian filters, which predict the likelihood that an item is abusive (commonly used for spam) (Bloch-Wehba 2020); and more sophisticated machine-learning classifiers that predict whether new, previously unseen content is abusive (increasingly used for hate speech and harassment, as well as nudity detection and multiple other purposes) (Duarte, Llansó, and Loup 2017; Gillespie 2018; Siegel 2020).

These automated systems have significant limitations (see Section 5.2), yet they are indispensable to abuse detection at the massive scale at which new content appears online (Bloch-Wehba 2020; Llansó et al. 2020). Online service providers such as YouTube

4. For a discussion of whether hash-based techniques should be classified as "content oblivious" or "content dependent," see the note to Appendix A.1.

and Facebook[5] have proudly touted their algorithmic abuse-detection capabilities as the solution to a range of abuses—and governments take those promises of wizardry at face value (Bloch-Wehba 2020; Gillespie 2018). This fixation on automated content analysis has the effect of making any impediment to that capability appear to be an existential threat to a provider's T&S mission. Hence the governmental handwringing over E2EE. Driven by the common but mistaken belief that "[c]ontent moderation is fundamentally incompatible with end-to-end encrypted messaging" (Mayer 2019, p. 1), government officials assert that because it thwarts automated content scans, E2EE "wholly preclud[es]" online service providers "from being able to act against" abuse on their services (T. U. S. D. o. Justice 2020, paras 2, 5) and "completely hinder[s]" law enforcement agencies' mission (Lomas 2021, para 4).

These hyperbolic claims are premised on the misconception that content-dependent techniques such as automated scanning are the only way that providers can counter abuse online. This misplaced assumption ignores the content-oblivious options available to providers. Our findings show that online service providers (including providers of E2EE services) can—and do—prevent, detect, and mitigate abuse using techniques that do *not* rely on the technical capability to access user content at will.

Related work on online service providers' T&S programs has often relied on outside analyses of their abuse-reporting functions (C3P 2020; Crawford and Gillespie 2016) and public-facing policy documentation (Gillespie 2018; Henry and Witt 2021; Klonick 2018; Pater et al. 2016; Perkins, Cryst, and Grossman 2021; Venturini et al. 2016; York 2010). Some of the survey- and interview-based research about providers' T&S tools has examined the issue from the vantage point of the providers' users (Goulds et al. 2020; Thorn 2021; Van Royen, Poels, and Vandebosch 2016) or others who interact with their abuse-reporting tools (*Comments of the Copyright Alliance* 2016). Other such research has, like the current article, queried providers about their T&S programs.

That work (some of which is collected in Keller and Leerssen (2020)) has tended to focus on either a few specific major platforms or a specific type of abuse. For example, Gillespie (2018) interviewed content policy representatives from YouTube, Facebook, and Twitter; Klonick (2018) interviewed numerous current and former policy executives as well as content moderators (many of whom are unnamed or referred to by their initials or first name and last initial) at Google/YouTube and Facebook; and MacKinnon (2012) interviewed Facebook policy representatives and anonymous Google employees. Milosevic (2016) and Milosevic (2018), which focused on cyberbullying, combined an analysis of social media services' published policies with interviews of representatives of a range of social media and messaging companies (conditioned upon prior approval of quotes included in the study) who either held leadership roles or worked on the company's safety or public policy teams. Urban, Karaganis, and Schofield (2016), which explored copyright infringement, conducted confidential surveys and interviews with online service providers about their practices and systems for receiving notice of (and taking down) content that allegedly violated copyright law. Thorn (2021), which investigated children's potentially harmful online sexual interactions, surveyed minors about their experiences using blocking and reporting tools; preparatory work for the survey included testing various services' tools as well as interviewing representatives from five unnamed social networking service providers' T&S teams about their T&S tools and policies.

These related works indicate that user samples and providers' public-facing tools and documentation are readily accessible to researchers, whereas researchers' surveys and interviews of provider representatives are often contingent upon confidentiality and may

5. In October 2021, Facebook (the parent company of Instagram and WhatsApp) officially changed its name to Meta (Zuckerberg 2021). For ease of reference, this article refers to "Facebook" rather than "Meta" throughout, except where required for citation purposes.

not happen at all unless the researcher has an inside connection to the provider and agrees to certain conditions for publication (Milosevic 2016). While we experienced similar issues (discussed in Section 3.1), the scope of our survey is much broader than that featured in past research, much of which is based on a handful of major social media platforms. The current study covers providers that serve from a few thousand to several billion users, ranging from small link-aggregation sites to a massively popular E2EE messaging app. In addition, our survey asked about multiple kinds of abuse rather than focusing on just one. In sum, our contribution complements and extends prior work by surveying a range of online service providers of widely varying sizes about their use of several T&S techniques with respect to 12 distinct categories of abuse.

3 Methodology

We briefly summarize our process of selecting survey recipients and administering the survey. We then describe the set of participating providers and services included in our analysis, and list the relevant survey questions. The section concludes with a discussion of the limitations of our methodology and results.

3.1 Survey Administration

We administered the survey via the Qualtrics survey platform from April 2 to June 11, 2021. We distributed the survey to a total of 58 recipients who are currently or formerly affiliated with over 50 companies or organizations (some of which are parents or subsidiaries of each other, making it difficult to pinpoint an exact number).

The survey instructions informed recipients that the survey was aimed at "providers of apps and services for online communications and/or data storage." The services offered by the companies and organizations we reached out to are broadly either (1) online communications services, such as messaging and chat apps (including E2EE apps), webmail, and videoconferencing or (2) online services driven primarily by user-generated content (UGC) such as social media platforms, blogging and wiki platforms, online review sites, dating apps, and video-sharing services. We also reached out to cloud storage services, electronic payment processors, and ride-sharing apps, among others. We chose this heterogeneous group of online services because their business models all entail contending with abusive, fraudulent, and criminal behavior by their users.

To recruit survey respondents, we used purposive sampling to identify and select individuals with jobs that would have given them experience with the T&S practices we explore who could provide in-depth information about their usage within their organizations (Creswell and Plano Clark 2018). These included members of the organization's T&S team (if it has one) or the legal, public policy, or security teams. In the cover email sent to recipients that contained the Qualtrics survey link, we requested that the recipient forward the email internally to the appropriate person if the initial recipient did not have a relevant role (or the authority to respond on behalf of the organization). To recruit additional respondents, we also used snowball sampling (Wasserman, Pattison, and Steinley 2005) by asking recipients, in the same cover email, to refer us to people they knew in relevant roles at other companies.

To increase survey participation, we leveraged our inside connections at providers (as discussed in Section 3.4) and offered respondents anonymity. While several participating entities wished to remain anonymous, some respondents consented to the publication of their entity, service, and survey responses. Some respondents allowed the entity to be named as a survey participant as long as no specific responses were attributed to it;

in some cases, we have included specific quotes attributed to those respondents with their approval.

3.2 Sample Population

Our analysis includes the survey responses of 13 individuals representing 11 companies and organizations. Note that there is not a strict 1:1 correspondence of respondent:entity. For one company, we received responses from two people; we kept both in the analysis because they relate to two distinct services. For another company, Facebook, we received two responses about three different services under its umbrella: one from one person about WhatsApp, and the other from a different individual who answered about both Facebook's and Instagram's messaging services.

This lack of 1:1 correspondence is unsurprising in the context of the tech industry, where a single entity may offer a number of different products and services over time that are either developed in-house or due to various corporate mergers and acquisitions. It does, however, complicate how we quantify and describe the survey data. For simplicity, we use 13, the number of individuals whose responses are included, as our *n*, and count 13 as the number of providers. We use the term "respondents" when referring to the individuals who answered the survey, and "providers" or "entities" when referring to the online service providers.

Two responses are from former employees of the respective entities; the rest were current employees at the time of the survey. All but one response pertained to the currently offered version of the service in question (as of April–June 2021); the other related to a service that is no longer offered.[6]

Many of the 13 respondents included in the analysis chose to keep their provider and the app or service at issue anonymous, and provided us with a generic description to use in publication. The providers and services that consented to publicly disclose their participation in the survey are:

- Facebook Messenger
- Instagram Messaging
- Lobste.rs
- MetaFilter Network Inc.
- WhatsApp
- Wikimedia Foundation Inc., answering as to Wikidata
- A former employee of Yahoo!, answering about the now-shuttered Groups service

The anonymous companies (and, where it does not risk confidentiality to state all or part of the response, the service they addressed in the survey) are, in the respondents' words:

- A global software company
- A major consumer electronics company
- A major U.S.-based media and technology company that provides an email service
- A major U.S.-based social media platform

6. According to the two former employees' LinkedIn profiles, one had left their employer over 2 years prior to the survey's administration. The individual stated on the survey that they were answering in relation to the current version of the service in question. The other, a former employee of Yahoo!, had left the company over 14 years earlier; the service in question, Yahoo! Groups, shut down in 2020 (Kan 2020).

- A provider of a chat app (author's note: app is not E2EE)
- A provider of encrypted messaging services
- A social media + tech company that provides a video streaming service

These services are diverse enough that there is no specific term (such as "social media platform" or "E2EE chat app") that would encompass them all. Thus we use the umbrella term "online service providers" as defined in the Introduction.

Taken together, the 13 respondents represent services that are used by billions of internet users worldwide. WhatsApp alone has around 2 billion MAU, and Facebook Messenger and Instagram each have about 1.3 billion MAU (DataReportal 2021). Among the anonymous services in the sample, the median user base is well over 200 million; the smallest is over 70 million. After that, there is a very steep drop-off to the three smallest services (anonymous or not): Lobste.rs, with 2,000 MAU (Appendix A.33); MetaFilter, with 2,700 active users (as of 2019) (dmh, n.d.); and Wikidata, which (unlike the far more popular Wikipedia) has only about 24,000 active users ("Wikidata:Statistics" 2021). We refer to those three as "small providers" in the graphs and tables in Section 4, and the other 10 as "large providers." We cannot provide a precise tally of the total number of unique users of the services covered, since many people have accounts with more than one of the entities that offer them. (Facebook, which owns Instagram and WhatsApp as well as Messenger, tries to account for overlap in its metrics; it estimates that about 3.51 billion unique individuals log into one of its services every month (*Form 10-Q: Facebook, Inc.* 2021)).

In short, while our analysis includes only 13 individual respondents, they represent entities that collectively serve a large proportion, perhaps a majority, of the world's 4.7 billion internet users (DataReportal 2021).

3.3 Survey Questions

We created the survey using the Qualtrics platform. The question formats included forced-choice (yes/no), multiple-choice, check-all-that-apply (CATA), matrix, and open-ended questions. The multiple-choice, CATA, and matrix questions generally included the options to respond "other" (with a fill-in text entry box provided), "N/A," "don't know," or "refuse to answer." The survey used open-ended questions primarily to collect basic information (e.g., company name, respondent's email address) and to request additional details not captured by the closed-ended response options. We used conditional branching to display additional survey questions depending on respondents' answers about their service's design and the T&S techniques it used.

In addition to being asked to describe their services' abuse reporting features and how they use metadata, respondents were asked to provide basic information about the service in question and to provide links to documentation of the provider's content policies and transparency reports. Below are the survey questions that we will discuss in our Findings.

Question 1. We asked all respondents the CATA question, "Which of the following does your company or organization use to detect, prevent, and/or mitigate abuse?" The options given were as follows:

Content-dependent techniques:

- "automated monitoring or scanning of contents of user communications" (ACS)

- "human review of contents of user communications" [7]

Content-oblivious techniques:

- "metadata"
- "in-app reports of abuse by users"[8]
- "reports of abuse from outside the service" ("outside" or "off-app" reports)[9]
- "numerical limits on sharing or forwarding content"
- "numerical limits on group size"

Respondents could also select none of the above, other, N/A, don't know, or refuse to answer.

Question 2. If a respondent selected "metadata" in response to Question 1, we displayed the CATA question, "Which of the following categories of metadata does your company or organization use to detect, prevent, and/or mitigate abuse?" We listed 16 options:

- IP address
- other location data (e.g. GPS)
- user-provided location
- user-provided age (or date/year of birth)
- inferred user age
- user-provided gender
- inferred user gender
- account age
- frequency/volume of account actions (e.g. x messages sent per minute)
- user's email address (as provided)
- user's name (as provided)
- account's username
- account's avatar picture
- user's social graph
- previous reports of the account for abuse
- previous reports of the named individual for abuse

Respondents could also select none of the above, other, N/A, don't know, or refuse to answer.

Question 3. Next, we displayed (to the same subset shown Question 2) the matrix question, "Does your company or organization do metadata-based abuse detection for

7. The survey intended for this phrase to mean only reviews undertaken on their own initiative by content moderators acting on behalf of the entity, and to exclude moderator reviews that are prompted by user notice. However, the survey question did not spell out this distinction, so we do not know whether the respondents' interpretation matched our own.

8. These include reports submitted through a reporting flow within the service, whether in the context of a mobile app, desktop app, or web browser; the phrase "in-app" is not limited solely to apps.

9. These include reports submitted through a different channel that is extraneous to the service itself, such as an email that reports a post on a social media platform.

any of the following types of abuse?" The 12 categories of abuse (plus "other") were as follows:[10]

- Intellectual property (IP) infringement
- Spam
- Phishing or malware
- Child sexual abuse imagery (CSAI)
- Child sexual exploitation (e.g., grooming, enticement) (CSE)
- Terrorism or violent extremism ("terrorism" for short)
- Pornography, sexual content, or obscenity (non-child) ("porn" for short)
- Dis-/misinformation ("mis/disinfo" for short)
- Harassment, threats, (s)extortion, or intimidation ("harassment" for short)
- Hate speech
- Self-harm
- Bots or inauthentic behavior ("bots" for short)

The matrix response options were yes, no, N/A, don't know, and refuse to answer. After this question, respondents were prompted to describe, in a free-text entry box, additional details about the provider's use of metadata for abuse detection.

Question 4. If a respondent selected "in-app reports of abuse by users" in response to Question 1, we displayed the matrix question, "Does your company's or organization's app, product, or service enable in-app user reporting for any of the following types of abuse?" The 12 categories of abuse (plus "other") and matrix response options were the same as in Question 3. After this question, respondents were prompted to describe, in a free-text entry box, additional details about how the provider enabled in-app user reporting.

Question 5. We asked all respondents the matrix question, "For each of the following categories of abuse on the app, product, or service, which does your company or organization find most useful for detection: automated monitoring or scanning of content, metadata, or user reporting?"[11] The 12 categories of abuse (plus "other") were the same as above. The matrix response options were "automated monitoring/scanning of content," "metadata," "user reporting," N/A, don't know, and refuse to answer.

Final wrap-up question. Near the end of the survey, we gave all respondents an open-ended optional question with the prompt, "Please write in any additional information you would like us to know about your company's or organization's trust & safety efforts."

10. These categories were chosen through consultation between the author and her supervisor, drawing on the latter's expertise in the domain of T&S.

11. The wording of this question did not distinguish between in-app and off-app reporting. For the lone respondent who stated that their service does not employ in-app reporting, we followed up and confirmed that where they answered "user reporting," they were referring to off-app reporting. For the remaining 12 respondents, the data does not reveal whether they had either in-app or off-app reporting (or both) in mind when responding. Also, the survey question did not define "most useful." We intended "most useful" to indicate both volume (the technique identifies more abusive items than the other options do) and accuracy (there is a low false positive rate in the items identified), but we do not know whether the respondents interpreted the term in this way, or whether their definition entailed other factors (e.g., false negative rates, speed in detecting new items, cost of building and deploying the technique, etc.).

3.4 Limitations

We recognize that there are at least eight limitations to our survey methodology and results. First, our survey relies on self-reported information about the participant entities' T&S practices, which are provided without any guarantee of accuracy. Second, there is a risk that the individual respondents' survey answers may reflect their own opinions rather than the official position of their employer, even though (as seen above) we intentionally phrased the questions to clearly ask for the latter. Third, as a result of our purposive sampling strategy, the entities surveyed are skewed toward organizations where the author and her supervisor have points of contact thanks to our extended professional and personal networks. Those organizations greatly outnumber those we "cold called" among the entities included in the sample. Our sampling approach also skewed the sample toward individuals in more senior roles: Nine of the 11 current employees who responded are director level or equivalent; the other two are the owners of small hobbyist sites. Of the two former employees, one had a manager-level role, the other a senior managerial role just below director level. The seniority skew could also be partially attributable to the survey's requirement that respondents affirm (at the start of the survey) that they were authorized to disclose the information they provided, a requirement we intended primarily for quality control and legal purposes.

Fourth, the ability for providers to respond to the survey while withholding the name of the provider and/or service from publication also affected the sample. Fifth, while some small entities responded to the survey (such as Lobste.rs and MetaFilter), the data is biased toward bigger organizations with large user bases. Sixth, there is a geographic skew toward the U.S.: most of the organizations that received the survey are based in the U.S., as are most of the entities that responded. Seventh, during the data analysis we discovered some flaws in the survey design, such as using confusing, inconsistent, or ill-defined wording. We know these ambiguities affected some of the survey answers, and they may have also affected other responses that went unnoticed (see the note to Appendix A.1 for further discussion). Finally, the respondents are not representative of the many and varied kinds of online services currently available. In short, the survey results cannot be generalized to all online service providers, some of which likely have abuse-detection programs that diverge from the trends seen in the results.

4 Findings

In this section, we describe the results of analyzing the responses to the survey questions listed above.

4.1 What Techniques Do Providers Most Commonly Use To Detect, Prevent, and/or Mitigate Abuse?

Figure 1 on the next page displays which techniques respondents said their providers use to detect, prevent, and/or mitigate abuse. Abuse reports were most common: all 13 respondents stated that their providers use either in-app or off-app reports (all but one use in-app reports and 10 use off-app reports). The next-most popular method was human review, followed by a tie between metadata and ACS. Least prevalent were numerical limits on shares/forwards and on group size.[12]

Appendix A.1 reports selected write-in responses about "other" techniques used to fight

12. These answer options were most relevant to the seven respondents who said their providers enable group text messaging; six of those seven said their providers employed limits on group size, and three, limits on shares/forwards.

abuse. Selected write-in responses from the 10 respondents who selected the off-app reports answer appear in Appendix Table A.2. Appendix A.3 contains selected responses to the open-ended question at the end of the survey that gave respondents the option to write in additional information about their providers' T&S efforts.[13]

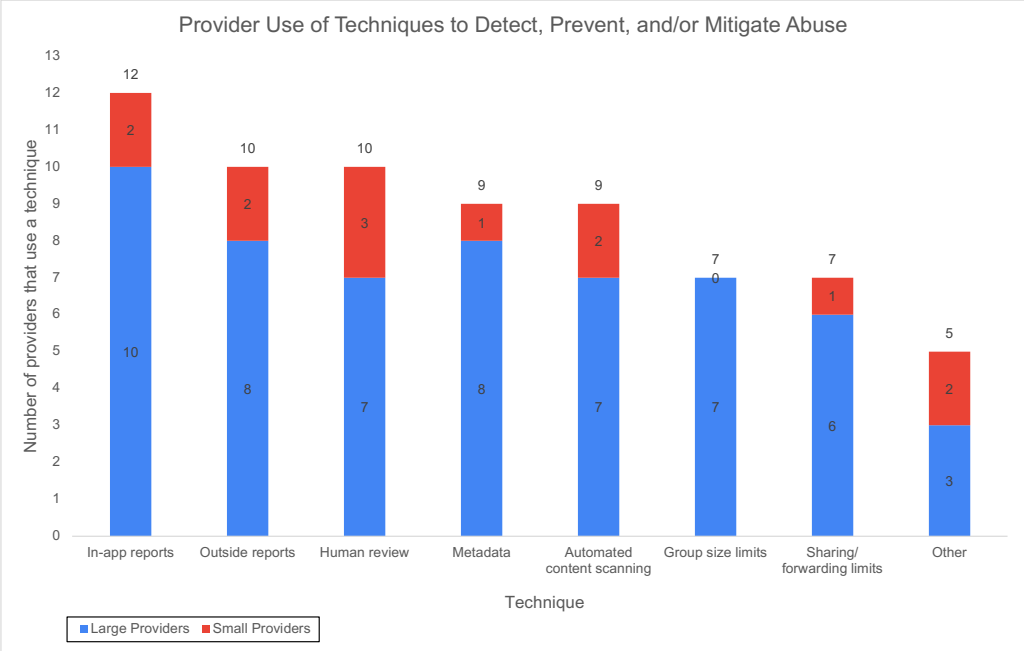

Figure 1: Number of providers that use a given technique for detecting, preventing, and/or mitigating abuse (n=13). Respondents could also answer "none of the above," "N/A," "don't know," or "refuse to answer" (not shown). Small providers are Lobste.rs, MetaFilter, and Wikidata; large providers are the other 10 organizations represented in the survey.

4.2 What Kinds of Metadata Do Providers Use To Detect Abuse, and What Kinds of Abuse Do They Use it To Detect?

Figure 1 illustrates that nine respondents reported that their providers use metadata to detect, prevent, and/or mitigate abuse. Figure 2 on the following page details *which kinds* of metadata they use. According to the responses, all nine utilize IP address, account username, and the frequency/volume of account actions (e.g., x messages sent per minute). Most use the account's age and previous reports of the account for abuse, and almost as many use the user's email address and the account's avatar picture. Five of the nine (about half) employ the user's age, name, social graph, and previous reports of the individual for abuse.[14] A minority use location information, gender information, or the user's inferred age.

13. We also asked all respondents to provide links to their services' content policy documentation and transparency reports (if any). Selected links, together with additional links to relevant material on the providers' websites that some respondents provided in the fill-in responses to other survey questions, are in Appendix B. We have not compared respondents' survey responses to the documentation they linked, as doing so would not verify or falsify our survey data because policy documents do not provide the level of information elicited by the survey questions (e.g., which T&S technique does a provider find most useful to combat each of 12 types of abuse); nor do the findings address the kinds of information typically contained in transparency reports (e.g., how many pieces of offending content were removed in a particular time period). Future research could analyze these materials and other survey response data not included in this paper.

14. *Individual* is distinguished from *account* because the same malicious individual can create multiple accounts on the same service under different fake identities, known as a *Sybil attack* (Xu et al. 2021).

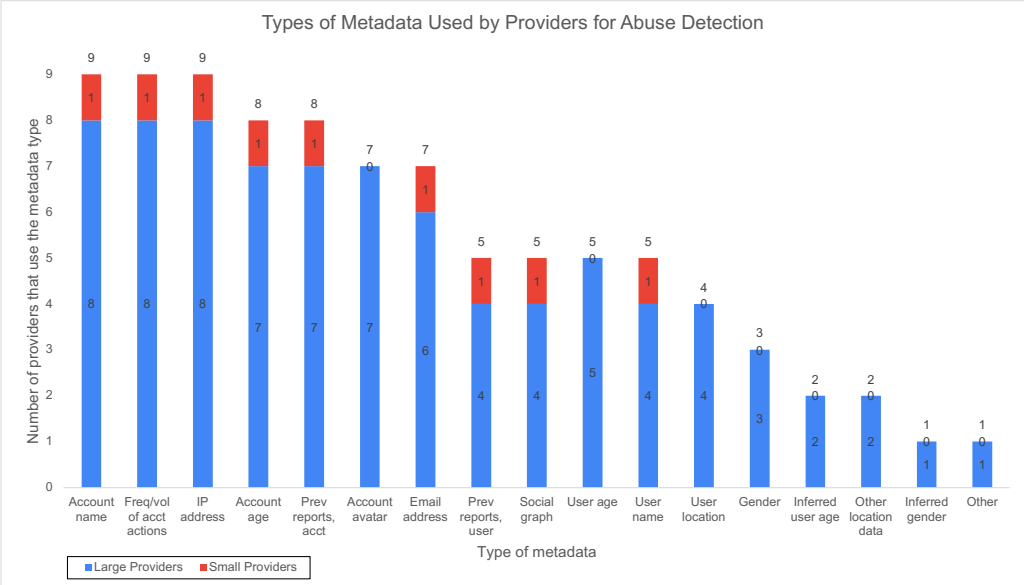

Figure 2: For providers that use metadata for abuse detection (n=9), the number that use a given type of metadata. Respondents could also answer "none of the above," "N/A," "don't know," or "refuse to answer" (not shown). Small providers are Lobste.rs, MetaFilter, and Wikidata; large providers are the other 10 organizations represented in the survey.

Figure 3 on the next page shows which kinds of abuse this same subset uses (or does not use) metadata to detect. According to the responses, all nine providers use metadata to detect spam and bots. Almost all use it to detect phishing/malware and CSAI; most use it to detect CSE. Five (about half) use metadata to detect terrorism, mis/disinfo, and harassment. A minority use it to detect porn, hate speech, IP infringement, or self-harm.

This data does not explain providers' rationales for (not) using metadata to monitor particular abuse categories; it only tells us how many providers do (or do not) use metadata to detect a given category of abuse. Nor can we use this data to determine whether a provider tried using metadata for a certain category in the past but then stopped, and if so, why; or whether they have never tried using metadata for that category, and if not, why not. Future research could explore these questions.

Appendix A.4 displays selected write-in responses with additional information about how providers use metadata to detect abuse. Several respondents described approaches that use metadata as an indicator of potential abuse that prompts human moderator review. This combination of content-oblivious and content-dependent techniques illustrates what Goldman (2021), in his taxonomy of providers' remedies for users' abusive acts on their services, calls "remedy combinations" (p. 52).

4.3 For What Kinds of Abuse Do Providers Enable User Reporting?

As seen in Figure 1 on the preceding page, almost all respondents surveyed (12 of 13) answered that their provider employs in-app user reports to detect, prevent, and/or mitigate abuse. (For comparison, only nine use metadata, all of which also use in-app reporting.) Figure 4 on page 14 shows which kinds of abuse this subset of providers reportedly enables user reporting to detect. Almost all (11) do so for spam; nine do so for phishing/malware, harassment, and bots; eight, for CSAI, CSE, and hate speech. After

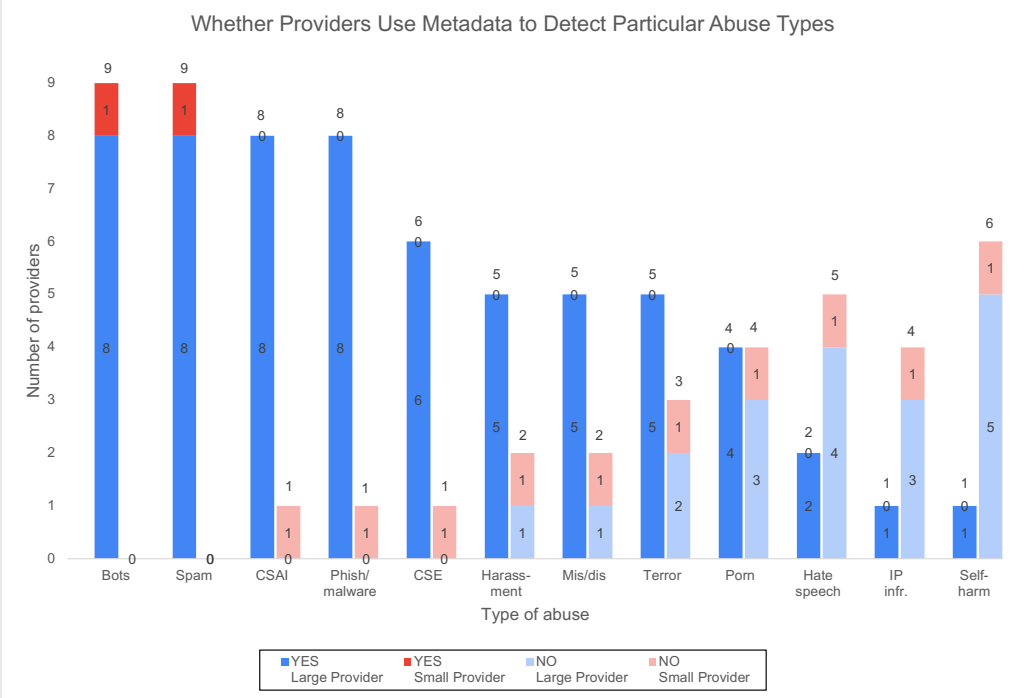

Figure 3: Among providers that use metadata for abuse detection (n=9), how many use or do not use metadata to detect a given abuse type. Other answer options besides "yes" and "no" were "N/A," "don't know," and "refuse to answer," which are not shown (thus the answer totals do not all sum up to 9). One large provider refused to answer for CSE. Another large provider left blank the following categories: IP infringement, porn, mis/disinfo, harassment, hate speech, and self-harm. Two large providers answered "don't know" for IP infringement; one former large-provider employee answered "don't know" for IP infringement, CSE, terrorism, porn, mis/disinfo, harassment, hate speech, and self-harm. No "N/A" responses were given. In addition to the listed abuse categories, respondents could also write in an "Other" category, but none did so. Small providers are Lobste.rs, MetaFilter, and Wikidata; large providers are the other 10 organizations represented in the survey.

that, in descending order, come porn and self-harm (tied), terrorism and mis/disinfo (tied), and IP infringement. Two providers answered "Other." One replied that user reporting is enabled for "[a]ny violation of ToS" (terms of service). The other stated that the service's in-app reporting tools do not distinguish among abuse types.

As with metadata, we did not ask why these 12 providers did (or did not) enable user reporting for different categories (or whether they had ever done so in the past), so their answers do not reveal the reasoning behind their practices. As noted, we gave these 12 respondents the option to write in additional information about how users can report abuse in-app. Selected responses are in Appendix A.5.

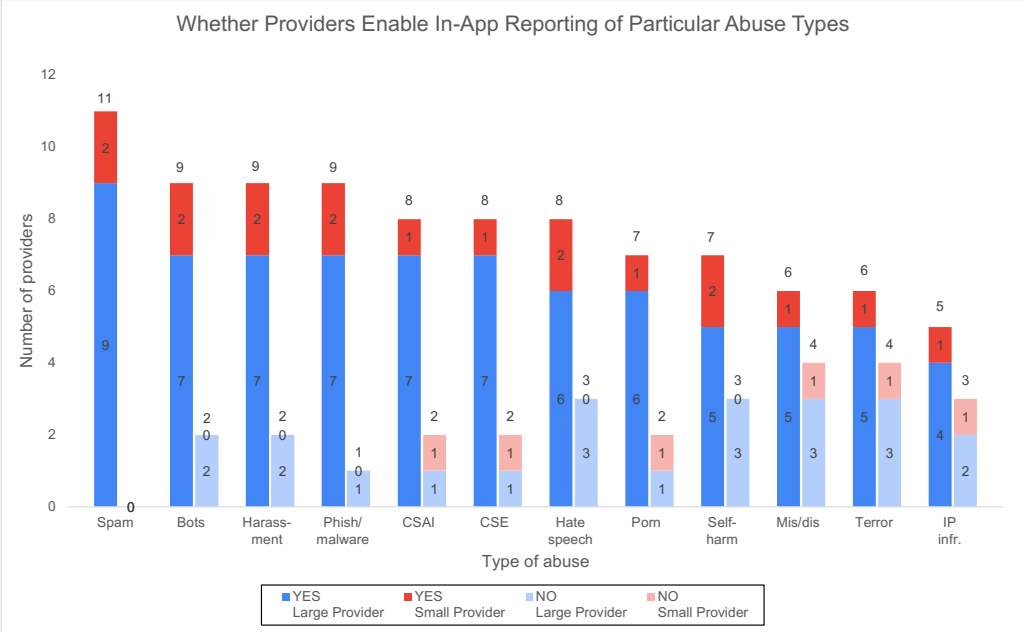

Figure 4: For providers that use in-app reports for abuse detection (n=12), how many enable in-app reports for a given abuse type. Other answer options besides "yes" and "no" were "N/A," "don't know," and "refuse to answer," which are not shown (thus the answer totals do not all sum up to 12). One large provider left the phishing/malware category blank. One large provider answered "don't know" for IP infringement; one former large-provider employee answered "don't know" for CSAI, CSE, terrorism, and porn; another former large-provider employee answered "don't know" for IP infringement, porn, mis/disinfo, and self-harm. One large provider answered "N/A" for all listed categories, and only filled in "Other." In total, two respondents filled in the write-in "Other" category as described above. Small providers are Lobste.rs, MetaFilter, and Wikidata; large providers are the other 10 organizations represented in the survey.

4.4 Which Technique Do Providers Find Most Useful for Detecting Each Type of Abuse?

All respondents were asked to select the technique their provider finds most useful for detecting each kind of abuse: ACS (content dependent), metadata, or user reporting (both of which are considered content oblivious). Table 1 on page 16 displays the responses to this "usefulness ranking" question. With the sole exception of CSAI (for which ACS far outranks the other options), user reporting either beats or ties with the other two techniques as the most useful for detecting each type of abuse. User reporting is ranked the most useful for nine of the 12 abuse categories: harassment, hate speech, self-harm, IP infringement, phishing/malware, mis/disinfo, terrorism, porn, and bots. ACS is ranked by far the most useful technique for detecting CSAI.[15] For the remaining two categories,

no one technique wins out: for CSE, there was a two-way tie between ACS and user reporting, and for spam, a three-way tie among the three techniques. (Metadata is nowhere deemed more useful than the other options.) Again, the survey did not ask providers to explain the rationale behind their rankings. As noted, the answer matrix for this question included "Other" along with the 12 listed abuse categories. Several respondents used this option to provide commentary (see Appendix A.6 for selected responses).

Next, we analyze a subset of "usefulness ranking" responses—those from the nine respondents who reported that their provider uses ACS to detect abuse (see Figure 1 on page 11). All of these respondents also said their providers use in-app user reporting, and all but one said they also use metadata. We single out this subset for examination because it allows us to compare the usefulness of the *content-dependent* technique this survey question asked about (ACS) to that of the *content-oblivious* ones (metadata and user reporting) from the perspective of the online service providers that use both.[16]

The responses for this subset are shown in Table 2 on the following page. ACS outranks the other options in three of the 12 categories: CSAI (where, again, there is strong consensus that ACS is the most useful), phishing/malware, and CSE. ACS ties with metadata in a fourth category, spam. In the others, one of the content-oblivious techniques either beats or ties the content-dependent technique as the most useful for abuse detection. User reporting prevails in seven categories: harassment, hate speech, self-harm, IP infringement, mis/disinfo, terrorism, and porn. Metadata prevails in the remaining category, bots. Again, the data does not disclose the providers' reasoning for their responses.

5 Discussion

Our findings reveal that all of the participating providers use both content-dependent and content-oblivious techniques, and that they consider content-oblivious ones (user reports and/or metadata) more useful than ACS (which we deem content dependent) for detecting most of the categories of abuse we asked about, with a few significant exceptions. These results indicate the current limitations of using ACS to detect abusive content, the power and unfulfilled potential of in-app abuse reporting tools for users, and the need for nuance in crafting policy responses to the multifaceted problem of online abuse, especially with respect to different types of online child safety offenses.

5.1 Importance of Strong Tools to Report Abuse

One clear take-away from the survey results is that user reporting is not only widely used for T&S; many participating providers also consider it to be more useful than metadata or ACS for detecting most kinds of abuse. This finding is important for two reasons. First, for policymaking purposes, it undercuts the assumption (described in Section 2) that at-will content analysis is the only means of fighting abuse in an increasingly end-to-end encrypted world. We discuss this implication more in Section 5.2.

Second, for the providers of online services, our results highlight the importance of investing in tools that empower users to report abuse. For providers considering how to improve their T&S practices, we recommend taking advantage of the "low-hanging fruit"

15. Seven respondents named ACS the most useful, including one who had not said their service used ACS to detect abuse. We followed up with the respondent about this anomalous answer. They clarified that while their provider does not use ACS on the particular service about which they filled out the survey, the entity does use ACS elsewhere in its services. Based on that, they said the provider deems ACS the most useful for CSAI.

16. This subset of nine respondents does not include the respondent who, per page 18 supra, clarified in follow-up correspondence that their provider uses ACS elsewhere but not in the service at issue in the survey.

Table 1: How many providers find a given technique most useful for detecting a given type of abuse (n=13), broken down by large and small providers. Small providers are Lobste.rs, MetaFilter, and Wikidata; large providers are the other 10 organizations represented in the survey. Totals do not sum up to 13 because the answer options "N/A" and "refuse to answer" are not shown in this table. At least one and no more than three respondents answered "N/A" for each of the 12 listed categories. One large provider answered "refuse to answer" for all listed categories, and only filled in "Other." A total of three respondents filled in the write-in "Other" category as described above.

	ACS			Metadata			User reports			Don't know		
Abuse Type	Large	Small	Total	Large	Small	Total	Large	Small	Total	Large	Small	Total
Bots	2	0	2	3	0	3	1	3	4	2	0	2
CSAI	6	1	7	0	0	0	1	1	2	2	0	2
CSE	3	0	3	2	0	2	1	2	3	2	0	2
Harassment	1	0	1	1	0	1	4	3	7	2	0	2
Hate speech	1	0	1	1	0	1	3	3	6	2	0	2
IP infringement	1	0	1	0	0	0	3	2	5	3	0	3
Mis/disinfo	0	0	0	3	0	3	2	3	5	2	0	2
Phish/malware	4	0	4	1	0	1	3	2	5	1	0	1
Porn	2	0	2	1	0	1	3	1	4	2	0	2
Self-harm	2	0	2	0	0	0	3	3	6	2	0	2
Spam	2	1	3	3	0	3	2	1	3	2	0	2
Terrorism	2	0	2	1	0	1	2	2	4	2	0	2

Table 2: For providers that use automated content scanning for abuse detection (n=9), how many find a given technique most useful for detecting a certain type of abuse. Responses are broken down by large and small providers. Small providers are Lobste.rs, MetaFilter, and Wikidata; large providers are the other 10 organizations represented in the survey. Totals do not sum up to 9 because the answer options "N/A" and "refuse to answer" are not shown in this table. At least one and no more than two respondents in this subset answered "N/A" for 10 of the 12 listed categories (except spam and bots). None of the respondents in this subset answered "refuse to answer" or "Other."

	ACS			Metadata			User reports			Don't know		
Abuse Type	Large	Small	Total	Large	Small	Total	Large	Small	Total	Large	Small	Total
Bots	2	0	2	3	0	3	0	2	2	2	0	2
CSAI	6	0	6	0	0	0	0	1	1	1	0	1
CSE	3	0	3	2	0	2	0	1	1	1	0	1
Harassment	1	0	1	1	0	1	3	2	5	1	0	1
Hate speech	1	0	1	1	0	1	3	2	5	1	0	1
IP infringement	1	0	1	0	0	0	2	2	4	3	0	3
Mis/disinfo	0	0	0	3	0	3	2	2	4	1	0	1
Phish/malware	4	0	4	1	0	1	1	1	2	1	0	1
Porn	2	0	2	1	0	1	2	1	3	1	0	1
Self-harm	2	0	2	0	0	0	3	2	5	1	0	1
Spam	2	1	3	3	0	3	0	1	1	2	0	2
Terrorism	2	0	2	1	0	1	2	1	3	1	0	1

we identify: Building full-featured user-reporting in-app tools that are granular enough to account for the kinds of abuse that tend to happen on a given service.

Almost all of the services assessed in this survey enable in-app user reporting (see Figure 1 on page 11), and user reporting is generally considered the most useful detection technique for the majority of abuse categories (see Table 1 on the facing page). Yet there are many gaps in the coverage of participating services' user-reporting tools (see Figure 4 on page 14). For example, while 11 of the 13 respondents answered that their services enable users to report spam, only eight said they enable them to report CSAI or CSE. That result is comparable to C3P (2020), which found numerous shortcomings in providers' CSAI reporting functions. Moreover, six respondents answered that their services enable user reporting for terrorist content and mis-/disinformation, and five, IP infringement. While not every abuse type will necessarily be a problem for every type of service,[17] these results are somewhat surprising. In the U.S., IP infringement and CSAI/CSE are specifically excluded from Section 230's legal protections for providers (Communications Decency Act 2018). Reporting mechanisms for supposedly infringing content are a key feature of the "notice and takedown" regimes that many providers have employed for decades pursuant to U.S. and European Union (EU) law (Kuczerawy and Ausloos 2016, pp. 233–35). Terrorist content and mis-/disinformation have lately been focal points for regulatory action globally (*The Christchurch call to action: to eliminate terrorist and violent extremist content online* 2019; Fried 2021; Funke and Flamini 2018; Sang-Hun 2021). And CSAI and CSE are a lightning rod for regulators in the U.S. and elsewhere, as discussed further in Section 5.3.

Based on our findings, we recommend that providers' in-app tools should add more options for users to report various kinds of abuse, to the extent that those types of abuse are pertinent to that service and taking into consideration which types there is regulatory pressure to address. We also recommend improvements in user-reporting tools because they may help defray the impact (discussed more below) of E2EE on a provider's abuse detection capabilities.

Providers should focus on users' needs and experiences when building out their abuse-reporting functionality. This recommendation is consistent with Kamara et al. (2021, 29)'s review of approaches to content moderation in E2EE systems, which concluded that "user-reporting and metadata analysis provide effective tools" for abuse detection, but that "more work is still needed." The report recommended "emphasiz[ing] user agency" in user reporting and conducting research on "the most effective techniques for encouraging user reporting of content such as user interface design" Kamara et al. (2021, 28).

A provider's user-centric abuse-reporting design should especially consider child users' needs. While the providers we surveyed did not deem user reporting very useful against CSAI, the survey did not ask them why.[18] Prior research suggests that missing or insufficient reporting tools could be part of the reason. In Thorn (2021), a majority of minors surveyed were confident in their knowledge of how to use reporting tools effectively, but less than half actually used those tools to report potentially harmful online interactions, and most wanted platforms to provide more information on the reporting process. For certain CSE-type encounters, many felt that none of the options commonly available in reporting menus accurately fit the situation. "[A]mbiguous [or missing] reporting functions" for CSAI and CSE can especially affect survivors of child sex abuse, "many of whom attempt to self-monitor the spread of their abuse imagery" even into adulthood

17. For example, terrorist content may be less of a worry on Wikimedia's Wikidata service than on, say, a social media platform or an encrypted chat app.

(C3P 2020, p. 4).

An insufficiently detailed reporting menu will not necessarily stop a user from reporting. One of our survey participants, who had responded that their provider found user reports most useful for "business misuse" (see Appendix Table 6), observed in follow-up correspondence that although business misuse was not included in the service's user-reporting prompt (unlike, say, spam), users nevertheless reported it through the service's abuse reporting flow anyway. Users' repurposing of a product feature to report a problem the feature did not account for brings to mind the saying by science fiction author William Gibson that "the street finds its own uses for things" (Gibson 1987, p. 187). Nevertheless, some users' determination to report abuse is not a substitute for enhancing the reporting tools available to them. Thorn (2021)'s findings accord with our own: abuse-reporting tools and resources are important, but they are not consistently responsive enough to the variety of situations a provider's users may encounter on a particular service.

We recognize that tools for reporting abuse are themselves susceptible to being misused, which highlights the need for further revision, testing, and fine tuning. For example, when Twitter ran a pilot program for users in three countries to report misinformation, it found that much of the reported content did not contain misinformation; only about 10% of the reports were "actionable" (Roth 2022). The challenge of building reporting tools that are accurate and efficient against different types of abuse, work well for different groups' needs, and are robust against misuse is beyond the scope of this paper.

If providers collect "before and after" data when they make changes to their user-reporting tools and use it to measure the changes' impact, they should release those findings publicly. We acknowledge, however, that such transparency runs the risk (among others) that a regulator might, upon seeing a provider's success with a particular technique, codify it into law without regard to whether it is workable across various types of online services or might one day become outmoded and ineffective.

5.2 Limitations of Automated Content Scanning to Detect Abuse

We find that, even among providers that can (and do) run automated scans of the content on their services, there are only three categories in which more providers said ACS was more useful for detecting abuse than metadata or user reports: CSAI (by far) and, by a smaller margin, phishing/malware and CSE. For the other types of abuse we asked about, content-oblivious strategies, especially user reporting, are deemed more useful than ACS (or, in the case of spam, as useful as ACS).[19]

18. While some CSAI recipients are surely willing ones who have no interest in reporting the content or the sender to the provider, it is not clear to what extent that explains user reporting's low perceived usefulness against CSAI in our survey results. Small-sample research conducted by a team at Facebook applied a taxonomy of CSAI sharers to conclude that "non-malicious" sharers outnumbered "malicious" ones 3:1 (Andrus, Buckley, and Williams 2021). This implies the theoretical existence of a corresponding taxonomy of CSAI *recipients*, including unwilling recipients who, unlike willing ones, might report the encounter if given the tools to do so. Examples drawn from the Facebook research include adults or children being solicited for CSAI, users who receive CSAI from a sender who finds it funny or outrageous, and users who accidentally encounter CSAI when seeking out adult sexual content.

19. It bears noting that our survey data cannot be presumed to be consistent with major online service providers' on-the-record statements elsewhere. In an April 2021 report involving a different set of providers than those included in this survey, some providers reported that their automated tools were responsible for detecting a huge percentage of certain types of abuse, or even all types of abusive content, on their platforms (GARM 2021). For example, YouTube claimed that in the second half of 2020, automated systems flagged the vast majority of spam comments, and Pinterest claimed that in Q4 2020, automated systems were responsible for about two-thirds of the "Pins" removed for hate speech. Although it is not always clear from the providers' language whether they are referring to content-based automated tools, metadata-based automated tools, or some combination thereof, it is nevertheless difficult to square those providers' claims about automated tools' effectiveness with our participating providers' responses that ACS is most useful against only a few types of

We do not have a clear picture of why that is. Perhaps the state of the art for CSAI and phishing ACS systems is more mature than that of other abuse types. Or perhaps those abuse categories are inherently the most amenable to high-accuracy automated scanning. Both possibilities are speculation. The survey did not ask which categories of abuse the providers use ACS to detect, how long they have done so for each, which ACS tools the providers use, or why the provider did or did not rate ACS the most useful technique for each category. Nevertheless, to the extent that we can draw inferences from the answers of the nine respondents who said their providers use ACS, we offer the following observations.

First, the survey data indicates that those three categories—CSAI, and to a lesser extent phishing/malware and CSE—could be expected to experience the biggest impact on a provider's abuse detection rates if a service started using E2EE by default. That is because E2EE takes away what the participating providers consider the most useful detection tool for them: ACS. Indeed, one major online service provider, Facebook (a longtime user of Microsoft's PhotoDNA program for detecting CSAI (Hill 2012; Microsoft, n.d.)), has outraged child safety advocates with its plan to end-to-end encrypt all its messaging services by default in the future: they predict that without ACS, Facebook will detect vastly less CSAI and CSE material than it does today (Benner and Isaac 2020; Burgess 2021; *Global threat assessment* 2019).[20]

However, we can predict that aside from those top three categories, if participating providers' views of usefulness reflect actual effectiveness, losing the ability to perform automated scanning due to E2EE may have a much smaller marginal impact on the detection of all other types of abuse because, according to the providers surveyed, it has not been very useful so far. For detecting the majority of abuse types, providers consider content-oblivious techniques to be more useful than ACS. Plus, unlike ACS, they would not be affected by the introduction of E2EE (assuming that, after implementing E2EE, the service's metadata remains unencrypted and the provider continues offering abuse-reporting tools that enable users to display the offending content to the provider).

Anticipating E2EE's varying effects on a provider's anti-abuse efforts should help providers such as Facebook plan how to adapt their T&S practices accordingly. We ask providers that implement default E2EE to publicly share as much information as they safely can about any changes they made to their T&S practices as part of that switch, as well as "before and after" data about detection rates of different categories of abuse.

These findings have at least two important implications for policymakers seeking to regulate providers' anti-abuse obligations. First, a T&S program cannot be dictated solely by the exigencies of fighting CSAI, which is distinct from other kinds of abuse. The providers we surveyed use an assortment of data and a suite of T&S tools to respond to the different challenges that various types of abuse present. We find that what providers say works best against CSAI does not, in their estimation, work nearly as well for other kinds of abuse, and vice versa; at present, ACS's superlative usefulness for CSAI is *sui generis*. We discuss the CSAI and CSE findings in more detail in Section 5.3.

---

abuse (and that user reports are generally more useful).

20. Facebook has attempted to quell concerns about child safety on its services in multiple ways, such as by testifying to Congress in late 2019 about the company's T&S strategy for E2EE services (*Testimony of Jay Sullivan: Hearing Before the United States Senate Committee on the Judiciary,* 116th Cong. [Dec. 10, 2019], https://www.judiciary.senate.gov/imo/media/doc/Sullivan%20Testimony1.pdf), while continuing to move forward gradually with its plans (Davis 2021; Wong 2021). By contrast, Apple prompted such an outcry among advocates for privacy, security, and human rights by unveiling a new child-safety feature for scanning photos uploaded to its iCloud service for CSAI that it put the plan on hold just a few weeks after announcing it in early August 2021 (Barrett and Newman 2021). A similar system to Apple's for content detection in E2EE environments (Kulshrestha and Mayer 2021) was deemed dangerous by the very researchers who designed it (Mayer and Kulshrestha 2021).

The second crucial implication of our survey findings is that scanning all content is not a silver bullet that will protect against every kind of abuse. The limitations of automated abuse-detection systems have long been recognized, for various types of abuse. For instance, hash-matching systems only detect known CSAI, not new imagery that has yet to be added to the hash database (a process that requires human review) (Gillespie 2018). Likewise, fingerprinting systems for detecting copyright-protected content generally cannot recognize non-infringing uses, such as fair use, that rely heavily on context (Bloch-Wehba 2020; Engstrom and Feamster 2017). In consultations with European regulators, representatives from YouTube, Facebook, and other companies reportedly acknowledged as much; the representative of one automated content-recognition service conceded the necessity of human judgment (P. Keller 2020). Hash-based systems for terrorist content encounter similar difficulties, since context is what distinguishes terrorist propaganda or recruitment from counter-programming or news coverage (Duarte, Llansó, and Loup 2017; Gillespie 2018). Likewise, machine-learning systems that can automatically detect hate speech against any group, in any language, on any service, remain a long way off (Siegel 2020). This is partly because hate speech is difficult to define, and research and tooling are under-developed in languages other than English (Duarte, Llansó, and Loup 2017; Siegel 2020). As a result, automated systems need human oversight ("Removals under the Network Enforcement Law: YouTube," n.d.) and tend to over-remove content that does not run afoul of any laws (Heldt 2019). Finally, a nudity detector is just a nudity detector; unlike U.S. Supreme Court Justice Potter Stewart, a machine-learning tool does not know obscenity (or pornography, or non-consensual intimate imagery) when it sees it (Gillespie 2018; *Jacobellis v. Ohio* 1964).

In short, subjectivity, context, and fluidity of meaning all influence whether content is considered abuse, creating a possibly insurmountable challenge to developing an effective automated abuse detection system (Gillespie 2018). The findings from our survey confirm those of prior studies: ACS alone cannot slay the many-headed beast of online abuse. "The dream of [automated] content moderation" that perfectly and fully replaces human discernment remains just that—a dream (pp. 101–2).

It is urgent that regulators understand the shortcomings of automated abuse detection, which have significant policymaking implications in light of recent global trends in the regulation of online service providers. Mandatory ACS (or "filtering" (Reda, Selinger, and Servatius 2020)) has been called "the third rail of intermediary liability law" (D. Keller 2020, para 15) due to serious concerns that it violates constitutional (Kosseff 2021; Pfefferkorn 2020a; *The future of intermediary liability in India* 2020) and fundamental rights (Geiger and Jütte 2021; Heldt 2019; Reda, Selinger, and Servatius 2020). Nevertheless, in recent years multiple jurisdictions have passed or proposed regulations that either require proactive automated filtering or effectively leave providers with little other choice despite purporting not to require it (Bloch-Wehba 2020). These regulations target the same kinds of abuse for which, as described above, automated systems' shortcomings are already well documented, and for which we find ACS is mostly not considered useful for detection even by providers that currently use it (see Table 2 on page 16). They include:

- Terrorism: regulations passed in Australia in 2019 (evelyn douek 2019) and the EU in 2021 (Chee 2021; *New rules adopted for quick and smooth removal of terrorist content online* 2021) and proposed in Canada (Canada 2021; Geist 2021);
- Hate speech: regulation proposed in Canada (Canada 2021; Geist 2021);
- IP infringement: regulations passed in the EU in 2021 (Reda, Selinger, and Servatius 2020; Vella Cardona 2021) and currently proposed in the UK (Goldman 2020);
- Porn/sexual content/obscenity: regulations passed in India in 2021 (for rape de-

pictions) (Electronics and Technology 2021) and proposed in Canada (for non-consensual intimate imagery) (Canada 2021); and

- "Harmful" content generally: currently proposed in the UK (Harbinja 2021; Voge and Wilton 2022) and passed in Australia (*Digital Rights Watch* 2021). [21]

Irrespective of their dubious legality, our results indicate that, if our participating providers' views are borne out in practice, automated scanning mandates may not fix the problems that governments intend them to solve. This assessment comes from those in the best position to know: online service providers that already engage in automated scanning of their services. Judging from the survey answers of our respondents—who collectively serve billions of internet users—governments seeking to reduce the online prevalence of terrorist content, hate speech, and so forth (without degrading the rights of their citizens) should start by incentivizing more providers to implement strong reporting tools before requiring ACS.[22]

5.3 Implications for Child Safety Policymaking

The survey response data about CSAI and CSE deserves special attention given the importance of these two categories of abuse in the policy sphere. Both CSAI and CSE fall under the umbrella of child safety offenses. In the ongoing policy debate over encryption, child safety has emerged as the most compelling rationale for policymakers around the globe to call for limits on online service providers' ability to offer E2EE to their users. These demands have taken various forms, including:

- New Indian rules that cover certain specified offenses (including CSAI) (Electronics and Technology 2021);
- Coordinated pressure by multiple governments on Facebook to drop its plan to encrypt all of its messaging services due to the anticipated impact on CSAI and CSE detection (U. S. D. o. Justice 2019);
- Legislative proposals in the U.S. (Pfefferkorn 2020b) and UK (*Internet Society: UK Online Public Safety Bill* 2021) that critics assert would, if passed, effectively ban E2EE in the name of child safety; and
- A draft European Commission document discussing proposed technical solutions for detecting CSAI and CSE that entail working around E2EE (Koomen 2021).

Our survey shows that child safety offenses are not a uniform problem demanding a uniform response. As noted above, we find that ACS's perceived usefulness for CSAI is *so* distinct from that of every other category of abuse that even the other category of child safety offenses, CSE, does not come close (see Figures 5 and 6).

Looking at the responses from all providers, ACS prevailed by far with regard to CSAI, but providers were split as to CSE. Seven providers deemed ACS the most useful detection technique for CSAI, whereas only two said user reporting was most useful and none said metadata was. For CSE, three named ACS as the most useful, and another three said user reporting; two said metadata.

21. Some of these—such as India's new rules and proposals in Canada, the UK, and the EU—also cover CSAI, for which ACS's utility outflanks its utility for the other types of abuse in our survey findings (*Inception Impact Assessment* 2020; Canada 2021; Harbinja 2021; Electronics and Technology 2021).

22. This is not to say that such regimes would not introduce their own problems. For example, Germany's "NetzDG" law was widely criticized for incentivizing over-censorship of user speech by pairing an inducement for large social networks to create easy-to-use reporting tools with stiff punishments for failing to remove reported content swiftly enough (evelyn douek 2017).

Among the subset of providers that use ACS to detect abuse, it again handily prevailed for CSAI, but only narrowly edged out the other options for CSE. Six providers said ACS was the most useful for CSAI, one said user reporting, and none said metadata. For CSE, three said ACS, one said user reporting, and two said metadata.

This data reveals a strong consensus among the participating providers that ACS is the most useful technique for detecting CSAI. However, respondents were more ambivalent about its ability to detect CSE. These findings have two important implications for online child safety policymaking, particularly with respect to E2EE.

First, we can expect the marginal impact of E2EE on providers' CSE detection rates to be less than the impact on detecting CSAI. That is due to both ACS's lesser and content-oblivious techniques' greater perceived utility for detecting CSE compared to CSAI: the former technique is affected by E2EE, while the latter techniques are not. This is noteworthy because encryption policy discussions often talk in circles about "whether the benefits of default encryption outweigh the costs" ("Understanding and demystifying," n.d., para. 2), including the costs and benefits to children in particular ("Wiretapping children's private communications" 2021; Kardefelt-Winther et al. 2020). While our survey results do not quantify those costs or benefits, they suggest that encryption's cost to CSE investigations, and therefore its overall cost to online child safety efforts, is less than these circular discussions might assume. CSE and CSAI thus cannot be treated as though they are the same problem. Since the shift to near-ubiquitous E2EE is now "a given reality" ("Understanding and demystifying," n.d., para. 3), there is a need for nuance in crafting technical and policy responses to its impact on child safety efforts.[23]

The second implication for online child safety policymaking is that the lack of a strong consensus among providers about how best to fight CSE demonstrates the importance of allowing them to be flexible in their T&S practices. Providers must have breathing room to test out ideas, pursue promising new tools, and abandon methods that no longer work well; "the best answers will come only from experimentation and empirical data" (Goldman 2021, p. 52). A heavy-handed regulatory regime that locks providers into using a particular technique would not only be premature, it would risk backfiring by enshrining a method that is already known to be of questionable efficacy, such as the automated scanning requirements found in India's new rule regarding rape depictions and Canada's contemplated rules regarding terrorist content and hate speech.

This observation is not limited to CSE. Even if one T&S technique is widely agreed to be the most useful against a particular abuse type, as with ACS for CSAI, regulatory mandates risk becoming outmoded because abusive users continually adapt their behavior to evade detection (Lowd 2018) and adopt new mediums and technologies for distributing abusive content (Bursztein et al. 2019).

The compliance costs of a regulatory mandate to fight a particular type of online abuse also risk leaving a provider with inadequate resources to fight other kinds of abuse that are not subject to a mandate (Miller 2021). When a provider is "only focused on legal compliance," one respondent observed, its T&S program suffers (see Appendix A.3). A mandate covering one abuse type may thus undermine efforts to combat others.

6 Future Work

Our survey results point to three potential directions for future research. The first is to ask online service providers how well the various techniques they use work in combina-

23. For more on the technical consensus against undermining E2EE, see (Abelson et al. 2015)

tion. The survey data shows that the participating providers all use some combination of content-dependent and content-oblivious techniques. While the survey asked the providers which *individual* technique (out of three given options) was most useful for detecting various types of abuse, it did not ask whether the provider finds any particular *combination* of techniques to be a superior means of fighting any given kind of abuse compared to any stand-alone technique (and if so, what that winning combination is). Goldman (2021, p. 52) asserts that combinations sometimes outperform individual techniques and anticipates that providers will have data about what works best. If so, the providers may be willing to share what they have learned.

Second, a future survey could ask providers for more information about their strategies for abuse *prevention* rather than *detection*, such as proactive user education about particular topics. While this survey did ask about providers' methods to "detect, prevent, and/or mitigate abuse," the questions focused mostly on detection. Understanding how providers' abuse prevention strategies fit in with their other anti-abuse efforts could help inform best practices for T&S programs. Likewise, future work could attempt to track changes over time in adoption rates of, and improvements in, prevention and detection technologies for various types of abuse (or, potentially, the lack of such a trajectory).

Finally, a tantalizing question to ask providers that at some point switched their service from unencrypted to E2EE by default is whether (and to what extent) they find content-oblivious techniques make up for the loss of content-dependent techniques, both overall and for specific types of abuse such as CSAI. Due to the political sensitivity of this question, however (not to mention the modest response rate for the present survey), we are skeptical that a significant number of providers would agree to give meaningful answers, even anonymously.

This skepticism is a byproduct of the frustrating lack of transparency that providers have historically afforded to their users, researchers, and the general public (with rare exceptions such as mandated reporting under German law) (D. Keller 2021; Miller 2021). As the new gatekeepers of speech and access to information, the details of online service providers' T&S practices can have far-reaching ramifications (Klonick 2018). Yet they tend to hold that information close while keeping outside researchers at arm's length (Persily 2021). Obtaining responses from online service providers to research questions can be "difficult," comes with strings attached, and may not happen at all unless the researcher has inside connections (as we did) (Milosevic 2016, pp. 5170, 5177). The "dramatic asymmetry of information between the platforms … and the rest of the world" (François 2019, p. 5) enables them to "reinforce their own ability to control political and public narratives regarding [their] accountability" (Bloch-Wehba 2020, p. 82). Although the large and unwieldy topic of online service provider transparency is beyond the scope of this paper, it will unavoidably continue to affect future research that relies on voluntary information sharing by providers.

7 Conclusion

We surveyed the trust and safety practices of a group of online service providers that collectively serve billions of internet users. The providers we surveyed employ an assortment of data and a suite of content-dependent and content-oblivious T&S strategies to fight abuse on their services.

We found that not only are content-oblivious T&S strategies widely used among the providers that participated in the survey; they are considered more useful than a content-dependent strategy (ACS) for fighting most kinds of abuse, with a few notable exceptions.

The participating providers considered user reporting to be particularly useful in anti-abuse efforts, although their answers revealed gaps in the coverage of their services' abuse-reporting tools that we recommend be rectified (to the extent a given abuse type is pertinent to the service).

While the participating providers are strongly in consensus that ACS is the most useful strategy for detecting CSAI, the same is not true for any other kind of abuse, not even other child safety offenses such as grooming and enticement. Consequently, we infer that the impact of implementing E2EE (which impedes ACS) on a provider's T&S efforts will vary significantly depending on the type of abuse.

Our findings demonstrate that when crafting policy responses to the different challenges posed by various types of abuse online, policymakers should take a nuanced approach and allow providers flexibility. As the varying problems of online abuse evolve, providers' T&S programs must be allowed to adapt as well. Our research indicates that incorporating content-oblivious techniques into those programs will be crucial to their success in the ongoing fight against online abuse.

References

Abelson, Harold, Ross Anderson, Steven M. Bellovin, Josh Benaloh, Matt Blaze, Whitfield Diffie, John Gilmore, et al. 2015. "Keys under doormats." :mandating insecurity by requiring government access to all data and communications. *Journal of Cybersecurity* 1, no. 1 (July): 69–79. http://hdl.handle.net/1721.1/97690.

Andrus, Malia, John Buckley, and Chris Williams. 2021. "Understanding the intentions of child sexual abuse material (CSAM) sharers." Meta Research, February 23, 2021. https://research.facebook.com/blog/2021/02/understanding-the-intentions-of-child-sexual-abuse-material-csam-sharers/.

Barrett, Brian, and Lily Hay Newman. 2021. "Apple backs down on its controversial photo-scanning plans." *Wired* (September 3, 2021). https://www.wired.com/story/apple-icloud-photo-scan-csam-pause-backlash/.

Bellovin, Steven M., Matt Blaze, Susan Landau, and Stephanie K. Pell. 2016. "It's too complicated: How the internet upends Katz, Smith, and electronic surveillance law." *Harvard Journal of Law & Technology* 30:1–101. https://jolt.law.harvard.edu/assets/articlePDFs/v30/30HarvJLTech1.pdf.

Benner, Katie, and Mike Isaac. 2020. "Child-welfare activists attack Facebook over encryption plans." *The New York Times* (February 5, 2020). https://www.nytimes.com/2020/02/05/technology/facebook-encryption-child-exploitation.html.

Bloch-Wehba, Hannah. 2020. "Automation in moderation." *Cornell International Law Journal* 53 (1): 41–96. https://community.lawschool.cornell.edu/wp-content/uploads/2021/03/Bloch-Wehba-final.pdf.

Burgess, Matt. 2021. "Police caught." One of the web's most dangerous paedophiles. Then everything went dark. *Wired* (May 12, 2021). https://www.wired.co.uk/article/whatsapp-encryption-child-abuse.

Bursztein, Elie, Einat Clarke, Michelle DeLaune, David M. Eliff, Nick Hsu, Lindsey Olson, John Shehan, Madhukar Thakur, Kurt Thomas, and Travis Bright. 2019. "Rethinking the Detection of Child Sexual Abuse Imagery on the Internet." In *The World Wide Web Conference,* 2601–7. WWW '19. New York, NY, USA: Association for Computing Machinery, May 13, 2019. Accessed February 15, 2022. https://doi.org/10.1145/3308558.3313482.

Canada, Government. 2021. "Technical paper." https://www.canada.ca/en/canadian-heritage/campaigns/harmful-online-content/technical-paper.html.

Canadian Centre for Child Protection Inc. 2020. *Reviewing Child.* Sexual Abuse Material Reporting Functions on Popular Platforms. https://protectchildren.ca/pdfs/C3P_ReviewingCSAMMaterialReporting_en.pdf.

Chee, Foo Yun. 2021. "Civil rights groups urge EU lawmakers to rebuff online terrorist content law." *Reuters* (March 24, 2021). https://www.reuters.com/article/us-eu-tech-security/civil-rights-groups-urge-eu-lawmakers-to-rebuff-online-terrorist-content-law-idUSKBN2BH0G3.

Comments of the Copyright Alliance. Before the U.S. Copyright Office, Docket No. 2015-7. 2016. https://copyrightalliance.org/wp-content/uploads/2016/11/Copyright-Alliance-Section-512-Comments1.pdf.

Communications Decency Act (2018), https://www.law.cornell.edu/uscode/text/47/230.

Testimony of Jay Sullivan: Hearing Before the United States Senate Committee on the Judiciary, 116th Cong. (Dec. 10, 2019), https://www.judiciary.senate.gov/imo/media/doc/Sullivan%20Testimony1.pdf.

Cranor, Lorrie Faith, and Brain A. LaMacchia. 1998. "Spam!" *Communications of the ACM* 41 (8): 74–83. https://doi.acm.org/10.1145/280324.280336.

Crawford, Kate, and Tarleton Gillespie. 2016. "What is a flag for? Social media reporting tools and the vocabulary of complaint." *New Media & Society* 18, no. 3 (March): 410–28. https://doi.org/10.1177/1461444814543163.

Creswell, John W., and Vicki L. Plano Clark. 2018. *Designing and conducting mixed methods research.* 3rd. Sage Publications.

CyberBunker.com. 2016. "Stay Online Policy." CyberBunker.com. https://web.archive.org/web/20160928055012mp_/http://www.cyberbunker.com/web/stay-onlinepolicy.php.

DataReportal. 2021. "Digital 2021 April global statshot report," April 21, 2021. https://datareportal.com/reports/digital-2021-april-global-statshot.

Davis, Antigone. 2021. "We'll protect privacy and prevent harm, writes Facebook safety boss." *The Telegraph* (November 20, 2021). https://www.telegraph.co.uk/business/2021/11/20/people-shouldnt-have-choose-privacy-safety-says-facebook-safety/.

Digital Rights Watch. 2021, November 1, 2021. https://digitalrightswatch.org.au/2021/11/01/bose/.

dmh. n.d. "I think beside and perhaps before the qualitative questions relating to community values and moderation/management culture it needs to." Metafilter. Accessed February 16, 2021. https://metatalk.metafilter.com/25774/State-of-the-Site-Feb-2021#1383942.

Duarte, Natasha, Emma Llansó, and Anna Loup. 2017. *Mixed Messages? The Limits of Automated Social Media Content Analysis.* Center for Democracy & Technology, November. https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf.

Electronics, Ministry of, and Information Technology. 2021. *Notification,* February 25, 2021. https://mib.gov.in/sites/default/files/IT%28Intermediary%20Guidelines%20and%20Digital%20Media%20Ethics%20Code%29%20Rules%2C%202021%20English.pdf.

Engstrom, Evan, and Nick Feamster. 2017. "The limits of filtering: A look at the functionality & shortcomings of content detection tools." *Engine* (March). https://static1.squarespace.com/static/571681753c44d835a440c8b5/t/58d058712994ca536bbfa47a/1490049138881/FilteringPaperWebsite.pdf.

"Wiretapping children's private communications." : Four sets of fundamental rights problems for children (and everyone else. 2021, February 10, 2021. https://edri.org/our-work/children-private-communications-csam-fundamental-rights-issues/.

evelyn douek. 2017. "Germany's bold gambit to prevent online hate crimes and fake news takes effect." Lawfare, October 31, 2017. https://www.lawfareblog.com/germanys-bold-gambit-prevent-online-hate-crimes-and-fake-news-takes-effect.

———. 2019. "Australia's 'abhorrent violent material' law: Shouting 'nerd harder' and drowning out speech." *Australian Law Journal* 94 (August 16, 2019): 41–60. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3443220.

Feerst, Alex. 2019. "Your speech, their rules: Meet the people who guard the internet." *OneZero* (February 27, 2019). https://onezero.medium.com/your-speech-their-rules-meet-the-people-who-guard-the-internet-ab58fe6b9231.

Form 10-Q: Facebook, Inc. 2021. https://www.sec.gov/Archives/edgar/data/0001326801/000132680121000049/fb-20210630.htm.

François, Camille. 2019. *Actors, behaviors, content: A disinformation ABC.* Transatlantic Working Group, September 20, 2019. http://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/05/ABC_Framework_TWG_Francois_Sept_2019.pdf.

Fried, Ina. 2021. "Exclusive: Coalition calls on Biden to form disinformation task force." *Axios* (April 29, 2021). https://www.axios.com/biden-disinformation-task-force-call-f522ba07-b047-4475-ba82-11d4d0cf1119.html.

Funke, Daniel, and Daniela Flamini. 2018. "A guide to anti-misinformation actions around the world." Poynter International Fact-Checking Network. https://www.poynter.org/ifcn/anti-misinformation-actions/.

Geiger, Christophe, and Bernd Justin Jütte. 2021. *Platform liability under Article 17 of the Copyright in the Digital Single Market Directive, automated filtering and fundamental rights: An impossible match.* PIJIP/TLS Research Paper Series. American University Washington College of Law. https://digitalcommons.wcl.american.edu/research/64.

Geist, Michael. 2021. "Picking up where Bill C-10 left off: The Canadian government's non-consultation on online harms legislation." Michael Geist, July 30, 2021. https://www.michaelgeist.ca/2021/07/onlineharmsnonconsult/.

Gibson, William. 1987. "Burning chrome." In *Burning Chrome,* 168–91. Ace Books.

Gillespie, Tarleton. 2018. *Custodians of the internet.* Yale University Press.

Global Alliance for Responsible Media. 2021. *GARM aggregated measurement report.* April 21, 2021. https://wfanet.org/l/library/download/urn:uuid:c90e9e05-826d-44f3-8ba2-31f91e457236/garm+aggregated+measurement+report+21apr21.pdf.

Global threat assessment. 2019: Working together to end the sexual exploitation of children online. 2019. WePROTECT Global Alliance. https://static1.squarespace.com/static/5630f48de4b00a75476ecf0a/t/5deecb0fc4c5ef23016423cf/1575930642519/FINAL+-+Global+Threat+Assessment.pdf.

Goldman, Eric. 2020. "The UK online harms white paper and the internet's cable-ized future." *Ohio State Technology Law Journal* 16 (2): 351–62. https://kb.osu.edu/bitstream/handle/1811/92278/OSTLJ_V16N2_351.pdf?sequence=1&isAllowed=y.

———. 2021. "Content moderation remedies." *Michigan Technology Law Review* (March 24, 2021). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3810580.

Goulds, Sharon, Miriam Gauer, Aisling Corr, and Jacqui Gallinetti. 2020. *Free to be online? Girls' and young women's experiences of online harassment.* https://plan-international.org/file/46061/download?token=pH3r4scC.

Harbinja, Edina. 2021. "U.K.'s Online Safety Bill: Not that safe, after all?" *Lawfare* (July 8, 2021). https://www.lawfareblog.com/uks-online-safety-bill-not-safe-after-all.

Heldt, Amélie Pia. 2019. "Upload-filters: Bypassing classical concepts of censorship?" *Journal of Intellectual Property Information Technology and Electronic Commerce Law* 10:56–64. https://www.jipitec.eu/issues/jipitec-10-1-2019/4877.

Henry, Nicola, and Alice Witt. 2021. "Governing image-based sexual abuse: Digital platform policies, tools, and practices." In *The Emerald international handbook of technology facilitated violence and abuse (Emerald Studies in Digital Crime, Technology and Social Harms,* edited by Jane Bailey, Asher Flynn, and Nicola Henry, 749–68. Emerald Publishing Limited. https://doi.org/10.1108/9781839828485.

Hill, Kashmir. 2012. *Facebook's top cop: Joe Sullivan.* Forbes, February 22, 2012. https://www.forbes.com/sites/kashmirhill/2012/02/22/facebooks-top-cop-joe-sullivan/?sh=7776335a54f4.

Inception Impact Assessment. : Regulation of the European Parliament and of the Council on the detection, removal and reporting of child sexual abuse online, and establishing the EU centre to prevent and counter child sexual abuse. 2020, December 30, 2020. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12726-Fighting-child-sexual-abuse-detection-removal-and-reporting-of-illegal-content-online_en.

Internet Society. n.d. "What is encryption." Internet Society. https://www.internetsociety.org/issues/encryption/what-is/.

Internet Society: UK Online Public Safety Bill. Is trying to legislate the impossible – a safe Internet without strong encryption. 2021. Internet Society, May 12, 2021. https://www.internetsociety.org/news/statements/2021/internet-society-uk-online-public-safety-bill-is-trying-to-legislate-the-impossible-a-safe-internet-without-strong-encryption/.

Jacobellis v. Ohio, 378 U.S. 184, 197 (1964), https://www.oyez.org/cases/1963/11.

Justice, The United States Dept. of. 2020. "International statement." : end-to-end encryption and public safety, October 11, 2020. https://www.justice.gov/opa/pr/international-statement-end-end-encryption-and-public-safety.

Justice, United States Dept. of. 2019. "Attorney General Barr." Signs letter to Facebook from US, UK, and Australian leaders regarding use of end-to-end encryption, October 3, 2019. https://www.justice.gov/opa/pr/attorney-general-barr-signs-letter-facebook-us-uk-and-australian-leaders-regarding-use-end.

Kamara, Seny, Mallory Knodel, Emma Llansó, Greg Nojeim, Lucy Qin, Dhanaraj Thakur, and Caitlin Vogus. 2021. *Outside looking in: Approaches to content moderation in end-to-end encrypted systems.* Center for Democracy and Technology, August. https://cdt.org/wp-content/uploads/2021/08/CDT-Outside-Looking-In-Approaches-to-Content-Moderation-in-End-to-End-Encrypted-Systems.pdf.

Kan, Michael. 2020. "Final nail in the coffin for Yahoo Groups lands Dec. 15." *PCMag* (October 12, 2020). https://www.pcmag.com/news/final-nail-in-the-coffin-for-yahoo-groups-lands-dec-15.

Kardefelt-Winther, Daniel, Emma Day, Gabrielle Berman, Sabine Witting, and Anjan Bose. 2020. *Encryption, Privacy and Children's Right to Protection from Harm.* UNICEF Office of Research - Innocenti, October. https://www.unicef-irc.org/publications/pdf/Encryption_privacy_and_children%E2%80%99s_right_to_protection_from_harm.pdf.

Keller, Daphne. 2020. "CDA 230 reform grows up: The PACT Act has problems, but it's talking about the right things." Center for Internet and Society at Stanford Law School, July 16, 2020. http://cyberlaw.stanford.edu/blog/2020/07/cda-230-reform-grows-pact-act-has-problems-it%E2%80%99s-talking-about-right-things.

———. 2021. "Some humility about transparency." Center for Internet and Society at Stanford Law School, March 19, 2021. http://cyberlaw.stanford.edu/blog/2021/03/some-humility-about-transparency.

Keller, Daphne, and Paddy Leerssen. 2020. "Facts and where to find them: Empirical research on internet platforms and content moderation." In *Social media and democracy: The state of the field, prospects for reform,* edited by Nathaniel Persily and Joshua A. Tucker, 220–51. SSRC Anxieties of Democracy. Cambridge University Press.

Keller, Paul. 2020. "Article 17 stakeholder dialogue: What have we learned so far." Communia, January 16, 2020. https://www.communia-association.org/2020/01/16/article-17-stakeholder-dialogue-learned-far/#filters.

Klonick, Kate. 2018. "The new governors: The people, rules, and processes governing online speech." *Harvard Law Review* 131 (6): 1598–670. https://harvardlawreview.org/wp-content/uploads/2018/04/1598-1670_Online.pdf.

Koomen, Maria. 2021. *The encryption debate in the European Union: 2021 update.* Carnegie Endowment for International Peace, March 31, 2021. https://carnegieendowment.org/2021/03/31/encryption-debate-in-european-union-2021-update-pub-84217.

Kosseff, Jeff. 2021. *Online service providers and the fight against child sexual exploitation: The Fourth Amendment agency dilemma.* The Digital Social Contract Lawfare Paper Series. Lawfare, January. https://s3.documentcloud.org/documents/20458337/online-service-providers-and-child-exploitation.pdf.

Kuczerawy, Aleksandra, and Jef Ausloos. 2016. "From notice-and-takedown to notice-and-delist: implementing Google Spain." *Colorado Technology Law Journal* 14:219–58. http://ctlj.colorado.edu/wp-content/uploads/2021/02/14.2_3_v2.final-Kuczerawy-and-Ausloos-4.5.16-JRD.pdf.

Kulshrestha, Anunay, and Jonathan Mayer. 2021. "Identifying harmful media in end-to-end encrypted communication: Efficient private membership computation," https://www.usenix.org/system/files/sec21summer_kulshrestha.pdf.

Llansó, Emma, Joris van Hoboken, Paddy Leerssen, and Jaron Harambam. 2020. *Artificial intelligence, content moderation, and freedom of expression.* Transatlantic Working Group, February 26, 2020. https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf.

Llansó, Emma J. 2020. "No amount of "AI" in content moderation will solve filtering's prior-restraint problem." *Big Data & Society* 7, no. 1 (April 23, 2020). https://doi.org/205395172092068.

Lomas, Natasha. 2021. "UK offers cash for CSAM detection tech targeted at e2e encryption." *TechCrunch* (September 8, 2021). https://techcrunch.com/2021/09/08/uk-offers-cash-for-csam-detection-tech-targeted-at-e2e-encryption/.

Lowd, Daniel. 2018. "Can Facebook use AI to fight online abuse?" The Conversation, June 12, 2018. https://theconversation.com/can-facebook-use-ai-to-fight-online-abuse-95203.

MacKinnon, Rebecca. 2012. *Consent of the networked: The worldwide struggle for internet freedom.* Basic Books.

Mayer, Jonathan. 2019. *Content moderation for end-to-end encrypted messaging,* October 6, 2019. https://www.cs.princeton.edu/~jrmayer/papers/Content_Moderation_for_End-to-End_Encrypted_Messaging.pdf.

Mayer, Jonathan, and Anunay Kulshrestha. 2021. "We built a system like Apple's to flag child sexual abuse material — and concluded the tech was dangerous." *The Washington Post* (August 19, 2021). https://www.washingtonpost.com/opinions/2021/08/19/apple-csam-abuse-encryption-security-privacy-dangerous/.

Microsoft. n.d. "PhotoDNA: Help stop the spread of child exploitation." https://www.microsoft.com/en-us/photodna.

Miller, Carly. 2021. "Can Congress mandate meaningful transparency for tech platforms?" TechStream, June 1, 2021. https://www.brookings.edu/techstream/can-congress-mandate-meaningful-transparency-for-tech-platforms/.

Milosevic, Tijana. 2016. "Social media companies' cyberbullying policies." *International Journal of Communication* 10:5164–85. https://ijoc.org/index.php/ijoc/article/viewFile/5320/1818.

———. 2018. *Protecting children online?: Cyberbullying policies of social media companies.* The MIT Press. https://library.oapen.org/bitstream/handle/20.500.12657/30535/645372.pdf?sequence=1.

Morar, David, and Bruna Martins dos Santos. 2020. "The push for content moderation legislation around the world." *Brookings Institution* (September 21, 2020). https://www.brookings.edu/blog/techtank/2020/09/21/the-push-for-content-moderation-legislation-around-the-world/.

New rules adopted for quick and smooth removal of terrorist content online. 2021, April 28, 2021. https://www.europarl.europa.eu/news/en/press-room/20210422IPR02621/new-rules-adopted-for-quick-and-smooth-removal-of-terrorist-content-online.

Pater, Jessica A., Moon K. Kim, Elizaveth D. Mynatt, and Casey Fiesler. 2016. "Characterizations of online harassment: Comparing policies across social media platforms." In *Proceedings of the 19th international conference on supporting group work,* 369–74. November 13, 2016. https://doi.org/10.1145/2957276.2957297.

Perkins, Shelby, Elena Cryst, and Shelby Grossman. 2021. *Self-harm policies and internet platforms.* Stanford Internet Observatory, April 13, 2021. https://github.com/stanfordio/publications/raw/main/20210408-Self-Harm-Policies-And-Internet-Platforms.pdf.

Perlroth, Nicole. 2019. "What is end-to-end encryption? Another bull's-eye on big tech." *The New York Times* (November 19, 2019). https://www.nytimes.com/2019/11/19/technology/end-to-end-encryption.html.

Persily, Nathaniel. 2021. "A proposal for researcher access to platform data: The platform transparency and accountability act." *Journal of Online Trust and Safety* 1 (1). https://doi.org/10.54501/jots.v1i1.22.

Pfefferkorn, Riana. 2020a. "The EARN IT Act is unconstitutional: Fourth Amendment." Center for Internet and Society at Stanford Law School, March 10, 2020. http://cyberlaw.stanford.edu/blog/2020/03/earn-it-act-unconstitutional-fourth-amendment.

———. 2020b. "The EARN IT Act: How to ban end-to-end encryption without actually banning it." Center for Internet and Society at Stanford Law School, January 30, 2020. http://cyberlaw.stanford.edu/blog/2020/01/earn-it-act-how-ban-end-end-encryption-without-actually-banning-it.

Pierce, David. 2020. "How a screenshot started a fight that took over Reddit." *Protocol* (May 27, 2020). https://www.protocol.com/reddit-powermods-war.

Reda, Julia, Joschka Selinger, and Michael Servatius. 2020. *Article 17 of the Directive on Copyright in the Digital Single Market: a fundamental rights assessment.* Gesellschaft für Freiheitsrechte, November 16, 2020. https://freiheitsrechte.org/home/wp-content/uploads/2020/11/GFF_Article17_Fundamental_Rights.pdf.

"Removals under the Network Enforcement Law: YouTube." Google Transparency Report. n.d. https://transparencyreport.google.com/netzdg/youtube.

Roth, Yoel. 2022. "We've found that reports are helpful, but noisy, for the first use (individual reporting). On average, only about 10% of misinfo reports were actionable — compared to 20-30% for other policy areas. A key driver of this was "off-topic" reports that don't contain misinfo at all." Twitter, January 17, 2022. https://twitter.com/yoyoel/status/1483094059023429635.

Sang-Hun, Choe. 2021. "Historical distortions" test South Korea's commitment to free speech." *The New York Times* (July 18, 2021). https://www.nytimes.com/2021/07/18/world/asia/korea-misinformation-youtube.html.

Siegel, Alexandra A. 2020. "Online hate speech." In *Social media and democracy: The state of the field, prospects for reform,* edited by Nathaniel Persily and Joshua A. Tucker, 56–88. SSRC Anxieties of Democracy. Cambridge University Press.

The Christchurch call to action: to eliminate terrorist and violent extremist content online. 2019. https://www.christchurchcall.com/christchurch-call.pdf.

The future of intermediary liability in India. 2020. Software Freedom Law Center, January. https://sflc.in/sites/default/files/2020-01/SFLC.in%20-%20Intermediary_Liability_Report_%282020%29_1.pdf.

Thorn. 2021. *Responding to Online Threats: Minors' Perspectives on Disclosing, Reporting, and Blocking.* May. https://info.thorn.org/hubfs/Research/Responding%20to%20Online%20Threats_2021-Full-Report.pdf.

"Understanding and demystifying." The fight against child sexual exploitation. n.d., https://www.wcl.american.edu/impact/initiatives-programs/techlaw/projects/understanding-and-demystifying-the-fight-against-child-sexual-exploitation/.

Wiretap Act, Stat. (1986), https://www.law.cornell.edu/uscode/text/18/2510.

Stored Communications Act, Stat. (1988), https://www.law.cornell.edu/uscode/text/18/2711.

Urban, Jennifer M., Joe Karaganis, and Brianna Schofield. 2016. *Notice and takedown in everyday practice.* UC Berkeley Public Law Research Paper, March 30, 2016. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2755628.

Van Royen, Kathleen, Karolien Poels, and Heidi Vandebosch. 2016. "Help, I am losing control! Examining the reporting of sexual harassment by adolescents to social networking sites." *Cyberpsychology, Behavior, and Social Networking* 19 (1): 16–22. https://doi.org/10.1089/cyber.2015.0168.

Vella Cardona, Mariosa. 2021. "Legality of article 17 of new copyright directive affirmed." *The Times of Malta* (August 10, 2021). https://timesofmalta.com/articles/view/legality-of-article-17-of-new-copyright-directive-affirmed.892702.

Venturini, Jamila, Luiza Louzada, Marilia Ferreira Maciel, Nicolo Zingales, Konstantinos Stylianou, and Luca Belli. 2016. *Terms of service and human rights: An analysis of online platform contracts.* http://hdl.handle.net/10438/18231.

Voge, Callum, and Robin Wilton. 2022. *Internet impact brief.* : end-to-end encryption under the UK's draft Online Safety Bill. Internet Society, January. https://www.internetsociety.org/wp-content/uploads/2022/01/IIB_Encryption_UK_Online_Safety_Bill_EN-1.pdf.

Wasserman, Stanley, Philippa Pattison, and Douglas Steinley. 2005. "Social networks." In *Encyclopedia of statistics in behavioral science,* edited by Brian Everitt and David C. Howell, 4:1866–71. John Wiley & Sons.

"Wikidata:Statistics." 2021, October 21, 2021. https://www.wikidata.org/wiki/Wikidata:Statistics.

Wong, Queenie. 2021. "Facebook Messenger adds more features, continues focus on encryption." *CNET* (August 25, 2021). https://www.cnet.com/tech/mobile/facebook-messenger-adds-more-features-continues-focus-on-encryption/.

Xu, Teng, Gerard Goossen, Huseyin Kerem Cevahir, Sara Khodeir, Yingyezhe Jin, Frank Li, Shawn Shan, Sagar Patel, David Freeman, and Paul Pearce. 2021. "Deep entity classification: Abusive account detection for online social networks." In *30th USENIX Security Symposium (USENIX) Security 21,* 1–18. https://www.usenix.org/system/files/sec21summer_xu.pdf.

York, Jillian C. 2010. *Policing content in the quasi-public sphere.* OpenNet Initiative. https://opennet.net/policing-content-quasi-public-sphere.

Zuckerberg, Mark. 2021. "Founder's letter, 2021," October 28, 2021. https://about.fb.com/news/2021/10/founders-letter/.

Author

Riana Pfefferkorn is a Research Scholar at the Stanford Internet Observatory and a non-residential fellow at the Stanford Center for Internet and Society.

Acknowledgements

Thank you to everyone who took the time to participate in this survey or to help us find someone who would. Thanks to the Stanford Internet Observatory's Josh Goldstein, Shelby Grossman, and Alex Stamos for their assistance and advice in designing and testing the survey; Sumana Harihareswara for referring several survey recipients to us; the Trust & Safety Professional Association for helping us get the word out about the survey; and Josh, Shelby, and our Stanford colleague Daphne Keller for their thoughtful feedback on an earlier draft of this paper. Any remaining errors are those of the author.

Data Availability Statement

Due to confidentiality agreements, neither supporting data nor the sources of the data can be made available.

Funding Statement

Thank you to Omidyar Network for providing funding and other support for the Stanford Internet Observatory's project studying trust and safety on end-to-end encrypted platforms. More information about this project can be found at https://cyber.fsi.stanford.edu/io/content/e2ee-workshops.

Ethical Standards

This research was deemed exempt from IRB review by Stanford Social & Behavioral (Non-Medical) IRB personnel.

Keywords

End-to-end encryption; online abuse; online safety; trust and safety

Appendices

Appendix A: Selected survey responses

To protect confidentiality, not all of the free-text responses received are quoted in these tables. In these tables and throughout the paper, we avoid quoting directly from survey responses where the response might, alone or in combination with other responses, allow the provider to be identified if they had requested anonymity. The responses that have attributions are from providers that consented to disclosure.

Appendix A.1 Selected responses about techniques used to detect, prevent, and/or mitigate abuse.

Table 3: Selected responses about techniques used to detect, prevent, and/or mitigate abuse.

(From Lobste.rs) “Rate limiting; content limiting; [making service] invite-only [to join]”; “Cross-checking bad content to detect other content/accounts. Cross-checking bad content for promotion by sockpuppets/coworkers.”*
(From MetaFilter) “strong up-front emphasis on site guidelines and acceptable conduct”; (regarding the “numerical limits on sharing or forwarding content” answer option) “daily limits on outgoing on-site mail messages, ≈30 a day”
For groups of users, the group’s image and description
(From a former Yahoo! employee regarding Groups) “MD5 hash matching of child [sex abuse] images”; “some form of MD5 image filtering based on image hashes provided by NCMEC”*
“We also scan content using [P]hotoDNA on upload.”*

*These two responses, which concern hash-matching systems, were given in response to the question about metadata discussed in Table 4, but are included in this table as a better fit. Answer options for this question included “metadata” and the undefined term “automated monitoring or scanning of contents of user communications” (see supra Section 4.1 and Figure 1 on page 11). There was no separate option for hash matching, and the survey respondents were not instructed whether to categorize hash matching as the former option (i.e., content oblivious) or the latter (i.e., content dependent). We did not call out hash matching specifically because we assumed the respondents would classify it as content dependent; however, if hash values are considered metadata about a file rather than the contents of the file, then hash-matching systems for conducting “automated monitoring/scanning of content” arguably count as metadata based and thus content oblivious. Indeed, the asterisked responses in this table about hash matching, in response to a question about metadata, could imply that those (and possibly other) respondents considered hash matching to be metadata based. However, that inference is contraindicated by the fact that 7 of 13 respondents selected “automated monitoring/scanning of content” when asked to pick the most useful technique for detecting CSAI, whereas no respondents selected metadata. Given their domain expertise in T&S, the survey respondents can reasonably be expected to know that the prevailing industry standard for CSAI detection is PhotoDNA, and that it is a hash-matching system. Thus we can reasonably infer from their answers that most respondents considered hash-matching systems under “automated monitoring/scanning of content” (and thus content dependent) rather than “metadata,” at least for the purposes of filling out the survey. We acknowledge that the proper classification of automated hash-matching systems as content dependent versus content oblivious remains open to debate and the answer might vary depending on the context.

Appendix A.2 Selected responses describing the use of reports of abuse from outside the service.

Table 4: Selected responses describing the use of reports of abuse from outside the service.

"Partner (NGO, researcher, civil society, etc) reporting options"
"Report form, reports from organizations"
(From Lobste.rs) "IRC, Twitter"
Reports accepted from law enforcement and partners
"We accept email report[s], including from non-users, regarding content on the platform"
(From the former Yahoo! employee regarding Groups) For CSAI, "user reports and FBI requests"†
(From Wikimedia Foundation) "If external sources surface issues with the Foundation, it either provides support to identify the right community venue for self-governance review (significant majority of cases) or, if the issue falls outside of community self-governance, reviews the issue."

†This response was given in response to the question noted in Table 4 (which concerns metadata, not abuse reports), but is included in this table as a better fit.

Appendix A.3 Selected responses: additional

Table 5: Selected responses to the question "Optional: Please write in any additional information you would like us to know about your company's or organization's trust & safety efforts."

(From Lobste.rs) "We're a hobby site, it's really just me running the show alone. I've been trying to bring on more volunteer mods ... [T]he site sees ≈250k unique visitors and ≈2k active users per month. We're not especially large, not trying to grow, not a business, pretty narrowly focused on a professional interest, and lean heavily on being invite-only to mitigate abuse."
(From MetaFilter) "We're a human-scale, fairly tight-knit online community, with a total active userbase numbering in the thousands and a primary emphasis on public thread-based group discussion. We have a 24/7 staff of paid moderators available to monitor and be directly responsive to site issues and we heavily emphasize the community's guidelines and moderator presence/availability on the site. So many of the trust & safety issues that come with larger-scale and/or automation-based and/or more one-to-one focused communication platforms translate poorly to our community structure and practices. We're able to put a proportionally great deal of human attention promptly on any given issue and to work directly with users when there's a report of inappropriate or abusive behavior."
"One of the biggest tradeoffs in the trust & safety space is quality - meaning, companies aren't willing to invest in quality interactions that could discourage some of the worst toxicity because they are more focused on scale and engagement.... This extends to all facets of infrastructure like good customer facing help pages and resources, reporting infrastructure, and most definitely the cost of providing adequate content moderation and support. ... Companies treat this stuff like toxic waste and have very little motivation to grapple with the consequences. ... I think this is especially true in the areas of context which exploit women and children the most - really, any areas where people are vulnerable T&S efforts often fail because they are only focused on legal compliance, not on doing the right thing, or trying to improve any outcomes or experiences for anyone. ... I realize that these [T&S] efforts can't necessarily change these human behaviors, but many of them are allowed to flourish with no accountability."

Appendix A.4 "Please describe how your company or organization uses metadata to detect abuse."

Table 6: Selected responses to the question "Please describe how your company or organization uses metadata to detect abuse."

"We have internal scripts and tooling that analyze the metadata to detect abuse."
"We have trained models that run on various forms of metadata, including whether or not a human took action on similar data in the past to surface content for review"‡
(From the former Yahoo! employee regarding Groups) "[W]e used membership in confirmed [CSAI] groups to work through the members' group networks in order to recommend further groups for content moderator review. ... [CSAI] groups also had distinctive usage patterns that differed from adult pornography groups."‡
(From Lobste.rs) Restrict account actions based on account age: new accounts (up to 70 days old) have "[m]any limitations," are "unable to invite users, post links to unseen domains, or suggest story edits"
(Also from Lobste.rs) Restrict account actions based on volume of reports of the account for abuse: "Account automatically prohibited from [posting content] if heavily flagged by users"

‡Note that these strategies use metadata (content oblivious) as an indicator that prompts human moderator review (content dependent), illustrating how content-oblivious techniques can be used in conjunction with content-dependent ones.

Appendix A.5 "Please describe how users can report abuse in-app."

Table 7: Selected responses to the question "Please describe how users can report abuse in-app."

(From the former Yahoo! employee regarding Groups) During employee's tenure in the early aughts, tooling was fairly rudimentary, "mostly reports to customer service agents that had to be reviewed manually"
"There's a report feature primarily for spam/phishing, but [it] could be used for [other kinds of] harmful content."
On mobile: in-app reporting flow; on desktop: form submission
"Users can block senders in the app and also report abuse by accessing our help page"
"We have an abuse report form along with in-app 'report' buttons for users and/or specific" content
"long press the message which users want to report"
(From Lobste.rs) "Users can flag [an item of content] with a list of pre-selected reasons" (not coextensive with the 12 abuse categories listed in the survey); for problems not covered in that list, "the in-app private message feature is used to message moderators as a catch-all"
(From MetaFilter) "Users can report issues and concerns about site content via an inline flagging mechanism on every comment and post, via a web-based contact form linked prominently on every page, via on-site mail to members of the moderation staff, via off-site email to individual or group company addresses for the moderation staff, and via an onsite posting queue."

Appendix A.6 Most useful: Other

Table 8: Selected responses entered under the "Other" option in response to the question "For each of the following categories of abuse on the app, product, or service, which does your company or organization find most useful for detection: automated monitoring or scanning of content, metadata, or user reporting?"

"Often we also use a combination [of methods,] ie scanning of user reporting."
"[F]ocusing on prevention of harm rather than waiting for harm to happen, before we detect it."
(From Wikimedia Foundation) "Most of the [12 categories of abuse] listed above fall under community self-governance"
"business misuse"§

§See Section 5.1 on page 15 supra for discussion of this response.

Appendix B: Platform Policies

Table 9: Selected links to participant services' content policies, transparency reports, and other documentation.

Facebook Messenger / Instagram Messaging	**Content policy documentation:** https://transparency.fb.com/policies/ **Transparency report:** https://transparency.fb.com/data/ **Other documentation:** https://www.messenger.com/privacy https://transparency.fb.com/data/community-standards-enforcement/fake-accounts/facebook https://www.facebook.com/notes/2420600258234172/ https://about.fb.com/news/2021/04/messenger-policy-workshop-future-of-private-messaging/ https://messengernews.fb.com/2021/02/09/stay-in-control-with-messenger-safety-features/ https://research.fb.com/blog/2021/02/understanding-the-intentions-of-child-sexual-abuse-material-csam-sharers/
Lobste.rs	**Content policy documentation:** https://lobste.rs/about **Transparency report:** https://lobste.rs/moderations
MetaFilter	**Content policy documentation:** https://www.metafilter.com/guidelines.mefi https://www.metafilter.com/content_policy.mefi https://www.metafilter.com/microaggressions.mefi **Transparency report:** "Discussions and updates on moderation actions, policy changes, and user questio[ns] take place on a specific subsite dedicated to intra-community topics": https://metatalk.metafilter.com/
WhatsApp	**Content policy documentation** https://www.whatsapp.com/legal/updates/terms-of-service/ **Other documentation:** https://faq.whatsapp.com/general/security-and-privacy/unauthorized-use-of-automated-or-bulk-messaging-on-whatsapp https://faq.whatsapp.com/general/how-whatsapp-helps-fight-child-exploitation https://faq.whatsapp.com/general/security-and-privacy/about-whatsapp-and-elections https://faq.whatsapp.com/general/security-and-privacy/staying-safe-on-whatsapp/
Wikimedia Foundation	**Content policy documentation:** https://foundation.wikimedia.org/wiki/Terms_of_Use/en "Within that framework, local, self-governing communities set their own refining content policies based on educational knowledge project purpose. In the case of our biggest project by contribution volume, Wikidata:" https://www.wikidata.org/wiki/Wikidata:Introduction https://www.wikidata.org/wiki/Wikidata:Notability https://www.wikidata.org/wiki/Wikidata:Deletion_policy https://www.wikidata.org/wiki/Wikidata:No_personal_attacks https://www.wikidata.org/wiki/Wikidata:Living_people **Transparency report:** https://wikimediafoundation.org/about/transparency/2020-2/