Putting the Count Back Into Accountability

An Audit of Social Media Transparency Disclosures, Focusing on Sexual Exploitation of Minors

ROBERT GRIMM, Independent Investigator, United States

This paper explores a lightweight, quantitative audit methodology for transparency disclosures called *scrappy audits*. It amounts to little more than treating redundant and repeated disclosures as opportunities for validating quantities. The paper applies two concrete audits to social media disclosures about content moderation. The first compares legally mandated reports about the sexual exploitation of minors as disclosed by social media and the national clearinghouse receiving them. The second compares historical quantities included in platforms' CSV files across two subsequent disclosures of the data. Despite their simplicity, these scrappy audits are nonetheless effective. Out of 16 surveyed social media platforms, 11 make transparency disclosures about content moderation and 8 meet the prerequisites of one audit. Yet only 4 platforms pass their audits. The paper continues probing the limits of transparency data by presenting a data-driven overview of the online sexual exploitation of minors. Accordingly, the analysis is particularly careful to identify threats to validity as well as potentially helpful, but unavailable statistics. Likewise, it identifies major shortcomings of widely used technologies for the automated detection of images and videos depicting sexual abuse of minors. Overall, the data shows an alarming growth in such material over the last decade. However, there also are strong indicators that current statistics, which treat all such material the same, are large and unhelpful overcounts. Notably, many technical violations of the law, e.g., teenagers sexting, are not necessarily grounded in actual harm to minors but still reported as such.

CCS Concepts: • **Social and professional topics → Computing / technology policy**; • **Information systems → Social networking sites**.

Additional Key Words and Phrases: social media, content moderation, transparency reporting, audit, minor safety, sexual exploitation of minors, child sexual exploitation, child sexual abuse material, CSAM, CyberTipline, National Center for Missing and Exploited Children, NCMEC, teenage sexting

1 INTRODUCTION

The work described in this paper started with the question: Can an interested individual, without institutional backing and without privileged access to social media platforms, make an objective determination about the accuracy of transparency disclosures on content moderation? While the question's preconditions reflect my own status as independent researcher, the question has broader relevance, applying to people interested in performing open source intelligence or strengthening civil society just as well. By exploring the limits of individual agency, the question may have broader relevance still. But the posture of underdog fighting corporate behemoths would also be misleading, since I worked as software engineer for Facebook from 2018 to 2019. However, I did *not* work on content moderation and I do *not* retain any financial interest in the firm.

In looking for such *scrappy audits*, I soon realized that redundant and repeated disclosures of the same metric, especially when originating from different organizations, provide just the desired opportunities for validating disclosed statistics because every divergence is a true positive error. That immediately led to the first audit. Compare so-called CyberTipline report counts disclosed by social media with those by the national clearinghouse receiving the reports, the National Center for Missing and Exploited Children or NCMEC.

The second audit was the result of a series of fortunate events, even if they didn't look it at first. After refactoring my code for analyzing Meta's CSV files, I carefully compared outputs to ensure that the new version still worked as expected. When numbers didn't match, I first reviewed my code and, finding nothing wrong, next checked the programs' inputs.

Author's address: Robert Grimm, Independent Investigator, Brooklyn, New York, United States, rgrimm@alum.mit.edu.

After noticing that I was using the CSV file for Q3 2022 instead of Q2 2022, I compared all quantities between the two files besides values newly added in Q3 and discovered that 113 out of 2,633 or 4.3% of quantities differed, 77 of them dating back to Q4 2020. I had found the cause for the divergent outputs—and the second audit for this paper!

Neither of the two audits requires more than a computer capable of running basic data analysis tasks—datasets easily fit into memory and the most "complex" math in this paper is a least-squares-fit over a log function—as well as as some time and attention to detail while extracting statistics from HTML and PDF files to make them machine-readable. In case of disclosures about the sexual exploitation of minors, you wouldn't even need the latter, since I already did the work and publicly released all code and data.[1] In short, the two audits definitely qualify as scrappy and enable interested individuals to do their own research—without all the downsides so often associated with just that phrase. Yet despite the methodology's simplicity, this type of investigation still is novel. Besides a parallel investigation comparing transparency disclosures under the EU's Digital Services Act and the corresponding statements of reasons database statistics [132], I am not aware of similar studies.

The audits also are effective. Out of 16 surveyed social media platforms, 11 make transparency disclosures about content moderation and 8 meet the prerequisites of one audit each. Yet only 4 platforms pass their audits. That's pretty remarkable since all but one of the surveyed platforms have been operational for at least a decade and engage in surveillance capitalism [144]. In other words, they should excel at the data collection and analysis tasks necessary for transparency disclosures. Yet the failures are substantial. Notably, in the case of Meta's CSV files, differences for the previous quarter might, in theory, stem from corrections applied after the quarterly reporting deadline. But IRL, seven successive pairs of quarters have substantial differences across all kinds of metrics dating back up to two years. Hence Facebook and Instagram clearly fail that audit.

This paper makes three contributions. First, previous audits of social media tranparency disclosures focus either on overall scope and semantics of such disclosures including metrics or compare measured user experience with platforms' claims [1, 23, 44, 124, 138]. That also seems to be the case for the two audits Meta commissioned [12, 82, 107, 114]. As far as I know, the focus on auditing the data itself is unique to this work. It also is complementary to previous work and thereby contributes towards a more holistic understanding of social media transparency disclosures. Second, a critical preparatory aspect of the work was the curation of transparency data in machine-readable form and the development of the analysis code. Both have been released as open source.[2] Third, right-wing activists have been instrumentalizing the sexual exploitation of minors as a political cudgel [16, 39, 45, 111]. To counter their manipulative appeals to strong emotions, I provide an evidence-based overview of what transparency disclosures tell us about the online sexual exploitation of minors, with a focus on social media from 2019 to 2022. Along the way, I am careful to identify threats to validity and topics missing entirely from the data, while also validating previously published work.

2 BACKGROUND AND RELATED WORK

2.1 US Law

Chapter 110 of 18 US Code [135], the United States' primary criminal code, concerns the "sexual exploitation and other abuse of children" (the chapter's title) and prohibits grooming children (§2251), selling or buying children (§2251A), the production, distribution, and possession of sexually explicit imagery involving minors (§2252) including child pornography (§2252A) even in other countries (§2260), and tricking minors into accessing harmful materials with misleading domain names (§2252B) or page contents (§2252C). As long as providers of electronic services immediately

[1] https://github.com/apparebit/intransparent
[2] See footnote 1

report such activities and materials to NCMEC's CyberTipline (§2258A), it explicitly limits their civil and criminal liability (§2258B), which otherwise includes criminal (§2253) and civil forfeiture (§2254) as well as mandatory restitution (§2259) in addition to long prison sentences, e.g., 5–20 years for the first violation of §2252A and 15–40 years for subsequent violations.

Chapter 110 also establishes NCMEC as a clearinghouse for information about the sexual exploitation of minors. It requires NCMEC to forward CyberTipline reports to suitable law enforcement agencies (domestic and foreign alike, §2258A). It also authorizes NCMEC to collect hashes and other identifying information for reported imagery and share that information with service providers (§2258C). That puts NCMEC into a rather unique position. It works closely with both industry and law enforcement. Similarly, while its role is established through the United States criminal code, the organization itself is a *private* non-profit corporation. The Congressional Research Service has documented the relevant background and history [41]. As far as CyberTipline reports are concerned, the documentation for the bulk API used by electronic service providers is public [102]. Furthermore, individual reports have been disclosed in the course of legal proceedings, of course without any attachments [87, 88].

2.2 Transparency Reporting and Audits

Transparency disclosures about content moderation—and hence the need for auditing them—are a relatively recent phenomenon emerging from the confluence of three trends. First, news of Russia's covert disinformation campaign during the 2016 US presidential election [44], the spread of medical and public-health misinformation during the COVID-19 pandemic [20, 42, 46], misinformation during the 2020 US presidential election and the subsequent attempted coup, as well as the trove of Meta's internal documents released by Frances Haugen [19, 31] kept reminding the public of social media's corrosive influence and increasing demands for better accountability about content moderation [49, 74]. Increasing use of algorithmic decision systems, notably at the beginning of the pandemic, only intensified the pressure on social media [116].

Second, while researchers used to have broad API-based access to platform data including individual posts, social media firms severely restricted or phased out bulk access in the aftermath of the scandal caused by Cambridge Analytica siphoning huge amounts of user data from Facebook [14, 109, 139]. When AI startups leveraged, amongst many other sources, Reddit's and Twitter's content for training large language models, the affected platforms responded by instituting hefty charges for access to remaining APIs, making them unaffordable even for most commercial applications [65].

Third, internet platforms originally started making transparency reports to hold governments accountable after Edward Snowden's leak of secret government documents revealed extensive communications surveillance by the United States' clandestine services in cooperation with their Australian, British, and Canadian counterparts. After Google released its "Government Requests Tool" in 2010, other internet platforms started making similar disclosures [133]. By now, all but three of the 16 platforms surveyed for this article regularly disclose government requests for user information as well as content removal.

Etsy was the first internet platform to release a transparency report about content moderation in 2015. Two years later, Germany's NetzDG law required them for actions performed in response to its dedicated flagging mechanism [133]. YouTube was the first social media platform to release a transparency report about content moderation in 2018, followed by Facebook and Twitter that same year, Reddit in 2019, and Pinterest, Snap, as well as TikTok in 2021 [10, 143]. While some civil society organizations have been content at simply tracking firms that make any transparency disclosures [1, 124], the Santa Clara Principles, released in 2018 and revised in 2021, capture the civil society consensus on

best practices [2]. At least Reddit and Twitter acknowledged as much by paying lip service to the Santa Clara Principles in their transparency disclosures [110, 134]. Alas, no social media platform comes even close to complying with them IRL [136].

That is not surprising given that transparency disclosures about content moderation, unlike earlier disclosures about government requests, always played a more ambiguous role. While academics and civil society would like to treat them as an accountability mechanism, social media also treat them as a means for deflecting criticism. Hence Meta's and Twitter's disclosures prominently highlight seemingly random metrics that make them look good. Meanwhile Reddit can't commit to the metrics it discloses, changing some of them every reporting period. Reddit also managed to bamboozle the Electronic Frontier Foundation into believing that it provided notifications for all its content takedowns when it simply doesn't [23, 53].

The obvious means for addressing the low quality and disparity of transparency disclosures would seem to be legal mandates. However, so far, extreme polarization and the resulting political gridlock in Washington, DC have proven stronger. Judging by the Harvard Law Review debate about the straw man of a non-existent "standard picture" for content moderation [27, 69, 71, 85], American legal scholars are just as gridlocked.

Meanwhile the EU has been moving ahead with comprehensive regulation of the technology industry, leading to the Brussels effect [11]. For purposes of this paper, the relevant law is the Digital Services Act or DSA [34]. To achieve its goal of creating "a safer digital space," the law places comprehensive obligations on platforms with respect to identifying and managing systemic risks, flagging illegal content, fairness and transparency of content moderation, assured compliance with the act, and so on. At the same time, the DSA is designed to be proportionate and differentiates by service functionality—from all intermediary services to hosting services to online platforms—and service size—from small to regular to very large—with small services generally exempted and very large online platforms having the most obligations. Additionally, requirements were gradually phased in between November 2022 and February 2024.

Critically, the DSA mandates that online platforms including social media notify users of content moderation decisions and provide a *statement of reason* or SOR. Furthermore, it requires that online platforms make at least yearly transparency disclosures with aggregate information in machine-readable form in addition to submitting those same statements of reasons "without undue delay" to the EU Commission's database at https://transparency.dsa.ec.europa.eu. Very large online platforms also need to identify and mitigate systemic risks (arguably the DSA's most flexible and substantial instrument), provide accredited researchers with access to their data, and submit to external audits for their overall compliance with the DSA.

The inclusion of audits in the DSA's obligations prompted civil society organizations to explore audits and their methodologies [3, 4, 9, 21, 81]. Alas, the EU Commisssion sidestepped questions about the details of audit methodology in its final rules by largely focusing on the elaboration of platform and audit risks [32]. Interestingly, a first examination of the statements of reasons database shows distinct degrees of compliance and significant divergence from initial DSA transparency disclosures, thus pointing to more opportunities for scrappy audits [132].

Alas, Meta's two voluntary audits of its transparency disclosures nicely illustrate how scope determines effectiveness. The first audit questioned a panel of academic experts on suitable metrics and resulted in an informative, well-reasoned document [12, 107]. The second audit asked certified public accountants to review "Meta's management assertion" of the transparency disclosures for Q4 2021 [82, 114]. While Meta's blog post suggests a comprehensive and in-depth review, the actual management assertion and public accountants' letter do not include any concrete "controls" that were, in fact, investigated. Both documents are also dated on the same day. That raises significant doubts about

the extent of the audit. Hence, it isn't particularly surprising that the audit missed all short-comings identified in this paper.

2.3 Terminology

While the title of Chapter 110, 18 US Code references "children," section paragraphs are expressed in terms of *minor*, i.e., per §2256 "any person under the age of eighteen years." The same section also defines *child pornography* in terms of minor. In contrast, NCMEC uses *child sexual abuse material* or CSAM. Others use *child sexual exploitation and abuse imagery* or CSEAI. In my mind, the two terms are synonymous and either seems preferrable over child pornography because they more clearly delineate illegal imagery featuring minors from legal adult pornography. In contrast, *child sexual exploitation* or CSE is more general and covers, for example, grooming and trafficking as well.

Throughout this paper, I generally use minor instead of child—even if ECPAT's terminology guidelines make a rather convoluted argument for the opposite [47]. Not only is minor the legally accurate term, but informal usage of child tends to imply prepubescence and child sexual anything increasingly has inflammatory, politicized impact. Nonetheless, when referring to imagery depicting the sexual abuse of minors, I follow NCMEC's precendent and use the acronym CSAM. Furthermore, when discussing the transparency disclosures of a particular platform, I use the platform's terminology, including the acronyms defined in the previous paragraph. Finally, since CyberTipline reports are called just that, reports, I use the term *transparency disclosure* to avoid confusion between the two types of reports.

During my research, I encountered a number of news articles referring to the National Center on Sexual Exploitation (NCOSE) as if it was NCMEC. Previously known as Morality in Media, that organization is a right-wing advocacy group with a deeply homophobic and sexphobic history [141]. Since its re-branding in 2015, it has softened some of its more extreme positions. It even tried to ingratiate itself with the LGBTQ+ community [52]. Alas, disavowing one of its worst smears does not make a sincere apology. Furthermore, it doesn't help that the same blog post insists that NCOSE's sexphobic opposition to all pornography and sexwork is supportive of the community.

3 SCRAPPY AUDITS OF SOCIAL MEDIA TRANSPARENCY REPORTS

In selecting platforms for my scrappy audits, I started with Buffer's list of the 20 most popular platforms in 2022 [79], dropped the five platforms targeting China because they are unlikely to adhere to the same governance standards, dropped Facebook Messenger because it is part of the already included Facebook, replaced Microsoft Teams and Skype with Microsoft because the firm does not offer more granular transparency disclosures, and added Tumblr because it became a destination for users leaving Twitter as well as employees laid off by the firm in the wake of Elon Musk's takeover [104]. Finally, to cover all platforms owned by the same corporate parent, I added Google and Wordpress. When compared to the EU's list of very large online platforms and search engines, my list does not include five e-commerce firms, three adult-oriented platforms, and one not-for-profit. Like Microsoft, Google does not make more granular transparency disclosures. Hence my list contains Google only once, whereas the EU distinguishes between Search, Play, Maps, and Shopping in addition to YouTube, which is included in both lists.

Table 1 provides an overview of the selected platforms and their transparency disclosures about the sexual exploitation of minors. In addition to platform names, which link to the platform's transparency disclosures where available, the table comprises the following columns: *Terminology* lists the terms used to identify violative content as introduced in Section 2.3. *Metrics* indicates disclosed quantities, with *pieces* referring to individual photos and videos and *reports* referring to CyberTipline reports. *Coverage* lists the granularity of disclosures and the covered calendar range, using Q for quarterly, H for semiannually, and Y for yearly. Pinterest is designated Q/H because its statistics are calculated

Table 1. Social media platforms and their transparency disclosures

Platform	Terminology	Metrics	Coverage		Format	Audit
Facebook	CSE	rounded piece counts	Q	Q3 2018–	one HTML page, CSV	✘
Google	CSAM	piece & report counts	H	H1 2020–	one HTML page	✔
Instagram	CSE	rounded piece counts	Q	Q2 2019–	one HTML page, CSV	✘
LinkedIn	child exploitation	piece counts	H	H1 2019–	one HTML page	···
Microsoft	CSEAI	piece counts	H	H1 2020–	Excel file/period	···
Pinterest	CSE, CSAM	piece & report counts	Q/H	H1 2020–	two HTML pages	✘
Quora	···	···	···	···	···	···
Reddit	minor sexualization, CSAM	piece & report counts	H	2021–	HTML page/period	✔
Snap	CSEAI	piece & report counts	H	H2 2019–	HTML page/period	✔
Telegram	···	···	···	···	···	···
TikTok	sexual exploitation of minors	piece fractional share	Q	Q1 2022–	HTML page/period, CSV	✘
Tumblr	···	···	···	···	···	···
Twitter	CSE	unique piece counts	H	H2'18–H1'22	one tabbed HTML page	···
WhatsApp	···	···	···	···	···	···
Wordpress	···	···	···	···	···	···
YouTube	CSAM	piece & report counts	H	H1 2020–	one HTML page	✔
NCMEC	CSAM	report counts	Y	2019–	PDF/table	···

per quarter but released only semiannually. Reddit is designated H but lists a start year because the firm started out making annual disclosures but then switched to semiannual ones in 2022. *Format* indicates the format of disclosures, and *Audit* shows the platform's audit performance.

A quick glance at the table suffices for determining that five out of the 16 surveyed platforms make no transparency disclosures about the sexual exploitation of minors. In fact, Telegram makes no transparency disclosures at all, whereas Quora and WhatsApp have made only legally required, minimal privacy disclosures. Meanwhile, Automattic does make regular disclosures for Tumblr and Wordpress, but only covering government requests and intellectual property.

Only a little less obvious is the fact that only three out of 16 platforms make disclosures in machine-readable format. Meta has been releasing CSV files for Facebook and Instagram since Q2 2021. TikTok released Excel files from Q3 2021 to Q3 2022 and has been releasing CSV files since. The project's GitHub repository collects these CSV files. Since Meta simply updates the file accessible through the same link every quarter, I have been collecting quarterly snapshots since Q2 2022 and filled in previous releases through the Internet Archive.[3] Microsoft's use of Excel files is curious because they only contain the data points for the reporting period in a nicely formatted spreadsheet. Twitter's transparency reports include a menu item to download CSV data but I never could get that to work, even before Musk's takeover of the firm. The GitHub repository also contains machine-readable versions of all other transparency disclosures. I extracted them with ad-hoc scripts and advanced text editor features where possible and otherwise fell back onto manual data entry.

Out of the sixteen platforms, Telegram is the only platform based outside of the United States and hence not subject to 18 US Code §2258A's reporting requirements. The other 15 platforms including TikTok are headquartered in the US and hence *are* subject. According to NCMEC's transparency disclosures, the 15 US-based platforms *do* submit CyberTipline reports. Alas, only five platforms belonging to four corporations disclose the number of reports filed, enabling a comparison with NCMEC's disclosures. They are Google, Pinterest, Reddit, Snap, and YouTube. An additional three

[3] https://web.archive.org/web/20240000000000*/https://transparency.fb.com/sr/community-standards/

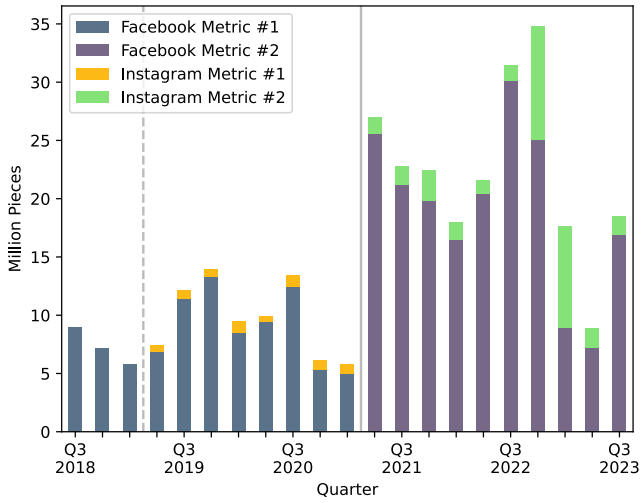

Fig. 1. Pieces maybe containing CSE on Facebook and Instagram (Meta)

platforms belonging to two corporations disclose their transparency data in machine-readable form, enabling a comparison with previous disclosures. They are Facebook, Instagram, and TikTok. The rest of this section discusses the detailed audit findings; it is organized by platform but combines those owned by the same corporation. I start with NCMEC.

3.1 The National Center for Missing and Exploited Children

In 2020, NCMEC started making transparency disclosures by releasing two multipage PDF documents with a table each listing CyberTipline reports by electronic service provider (ESP) [90, 92, 94, 97] and country of the suspected user [89, 91, 91, 93]. In 2022, the organization added a third PDF document with a table listing notifications sent to ESPs [95, 98]. It also began releasing an actual, written report [99, 100]. While the latter did contain some of the same material, NCMEC provided a more complete and meaningful accounting of its work only in 2023, when it released its report to the Department of Justice's Office of Juvenile Justice and Delinquency Prevention [101]. Section 4 makes extensive use of the report and its data.

While making the data in NCMEC's PDF files machine-readable again, I did encounter data quality issues with the organization's breakdown of CyberTipline reports by country. In particular, the data included two entries for the same country, labelled "French Guiana" and "Guiana, French;" entries for the Netherlands Antilles in addition to the three island states resulting from the Antilles' split in 2010; Bouvet Island even though this subantarctic Norwegian territory is an uninhabited nature preserve; and "Europe," presumably meaning the continent, in addition to pretty much all European countries. Taken together, problematic countries account for less than 250 CyberTipline reports per year from 2019 to 2022. Hence I dropped them.

3.2 Meta: Facebook, Instagram, and WhatsApp

Meta's transparency disclosures do not include CyberTipline reports but photos and videos, or *pieces*, that feature the sexual exploitation of children. As indicated by the dashed vertical line in Figure 1, Meta originally collected that data

for Facebook only and started to account for Instagram as well with Q2 2019. As indicated by the solid vertical line and different bar colors, Meta switched metrics from "Child Nudity & Sexual Exploitation" to "Child Endangerment: Sexual Exploitation" with Q2 2021. While the names suggest that the former includes the latter metric, the jump in piece counts that quarter suggests otherwise.

By searching for "Meta Community Standards Enforcement Report Q2 2021" with an external search engine, I eventually found a blog post [35] linking a PDF document with the explanation [36]. The latter metric also includes "Sexualization of children" and "Inappropriate interactions with Children." While these activities are related to minor safety as well, they don't seem obvious choices for combination into a single metric. The fact that Meta did so anyways suggests an ulterior motive, such as trying to impress readers by the sheer number of successfully intercepted pieces. However, that is speculation on my part.

Table 2. CyberTipline reports for Meta (Meta and NCMEC)

Year	Pieces	π	Reports	% Total
2019	39,368,400	2.48	15,884,511	94.3
2020	38,890,800	1.92	20,307,216	94.7
2021	78,012,400	2.90	26,885,302	92.2
2022	105,800,000	3.89	27,190,665	85.5

Table 2 provides a little more context by relating Meta's piece counts to CyberTipline report counts and then identifying Meta's share of *all* CyberTipline report counts (as disclosed by NCMEC). In this table (and several subsequent ones), the π stands for *product* and the column shows the multiplicative factor relating reports to pieces. Remarkably, Meta detects and reports roughly nine times as many incidents involving the sexual exploitation of minors *as everyone else combined*! Yet despite its industry-leading efforts, Meta still misses sexually exploitative activities targeted at minors on both Facebook and Instagram, with pedophiles having dedicated groups and CSAM producers advertising their wares [60, 61].

At the same time, there is evidence that piece counts severely overstate the problem. To improve prevention, Meta developed a taxonomy for capturing the intent behind shared CSAM [15]. It also started displaying automated warnings tailored according to intent [25]. While the firm did not report on longer-term outcomes of those warnings, it did report some striking statistics about intercepted CSAM. Notably, 90% of pieces reported to NCMEC during October and November 2020 had been reported before or were visually similar to previously reported content, with six videos accounting for more than half of the reported content. Further analysis of 150 accounts reported to NCMEC during July and August 2020 as well as January 2021 showed that 75% of them didn't share CSAM with malicious intent but "for other reasons, such as outrage or in poor humor (i.e. a child's genitals being bitten by an animal)." Unfortunately, Meta has not been disclosing any further statistics, though the firm recently announced its plans to do so [83].

As already discussed in Section 1, Meta's disclosures for Facebook and Instagram diverge for some fraction of historical quantities. Between Q3 2021 and Q4 2022, an average of 102 or 4.1% of such quantities differ every quarter, affecting a wide range of metrics and going back years. But in Q1 2023, the number of affected values dropped sharply, to 18 or 0.6% and the next two quarters see no such divergent quantities. These counts may very well be undercounts because of Meta's practice of rounding quantities. If I were to speculate, timing points to the EU's Digital Services Act, with Meta reigning in data fluctuations just before the EU Commission designated Facebook and Instagram as very large online platforms on April 25, 2023 (which started the four-month clock to come into compliance) [33]. Still, the verdict is clear: Facebook and Instagram fail the audit.

Meta does not make meaningful transparency disclosures for WhatsApp. However, as NCMEC's disclosures for 2021 and 2022 demonstrate, Meta does file CyberTipline reports for WhatsApp: 1,372,696 in 2021 and 1,017,555 in 2022.

3.3 Google and YouTube

Table 3. CyberTipline reports for Google and YouTube (Google and NCMEC)

Year	Pieces	π	Reports	Δ%	NCMEC
2019	...	...	...	...	449,283
2020	4,437,853	8.10	547,875	-0.2137	546,704
2021	6,696,497	7.69	870,319	0.6278	875,783
2022	13,402,885	6.16	2,174,319	0.0105	2,174,548

Table 3 shows the combined number of pieces and reports for Google and YouTube, tabulating Google's and NCMEC's counts. The difference is less than 0.7% per year, which seems acceptable. Hence Google and YouTube pass the audit.

3.4 LinkedIn and Microsoft

Even though Microsoft owns both LinkedIn and Skype, it treats LinkedIn as a separate entity, with independent branding, logins, and transparency reports, whereas Skype and Microsoft Teams are lumped in together with the rest of Microsoft's services and hence also transparency disclosures. Both Microsoft and LinkedIn report only pieces, not reports, and hence my audit methodology does not apply.

3.5 Pinterest

Table 4. CyberTipline reports for Pinterest (Pinterest and NCMEC)

Year	Pieces	π	Reports	Δ%	NCMEC
2019	...	...	...	...	7,360
2020	...	...	3,432	$\equiv$	3,432
2021	1,608	0.599	2,684	-14.94	2,283
2022	37,136	1.127	32,964	4.08	34,310

Table 4 shows the number of pieces and reports for Pinterest, tabulating Pinterest's and NCMEC's counts. The maximum difference is almost 15%, which seems too big an error. For divergent CSV data, the responsible party is clear. For divergent CyberTipline report counts, it is impossible to determine with certainty which side, Pinterest or NCMEC or both, made tabulation errors. However, Pinterest is the more likely party, given that NCMEC's counts agree with those for the other audited platforms. While both organizations would be well advised to validate disclosed quantities for 2021 and 2022, I attribute the failure to the likely culprit: Pinterest fails the audit.

3.6 Reddit

Table 5 shows the number of pieces and reports for Reddit, tabulating both Reddit's and NCMEC's counts. The counts aren't just close, they are identical and Reddit passes the audit.

In the Electronic Frontier Foundation's latest audit in 2019, Reddit fared best, receiving five out of five possible stars [23]. That ranking is undeserved. First, Reddit changes some number of disclosed metrics every report, which

Table 5. CyberTipline reports for Reddit (Reddit and NCMEC)

Year	Pieces	π	Reports	Δ%	NCMEC
2019	···	···	724	≡	724
2020	···	···	2,233	≡	2,233
2021	9,258	0.920	10,059	≡	10,059
2022	80,888	1.538	52,592	≡	52,592

makes tracking changes impossible. Second, Reddit got one star for providing meaningful notice upon taking down content. While it committed to doing so when it touted its adherence to the Santa Clara Principles in 2019 [110], the firm doesn't do so in practice. Instead, Reddit makes extensive use of shadow-banning, with content long deleted from Reddit still appearing as visible to the users who posted it. Changing "`reddit`" to "`reveddit`" for any Reddit URL surfaces such content for everyone [53].

3.7 Snap

Table 6. CyberTipline reports for Snap (Snap and NCMEC)

Year	Pieces	π	Reports	Δ%	NCMEC
2019	···	···	···	···	82,030
2020	···	···	···	···	144,095
2021	···	···	···	···	512,522
2022	1,273,838	2.31	550,755	0.0601	551,086

Table 6 shows the number of pieces and reports for Snap, tabulating both Snap's and NCMEC's counts. The difference is less than 0.07%, which is acceptable. Snap passes the audit, albeit based on only one pair of data points.

3.8 Telegram

Telegram does not make transparency disclosures. Furthermore, since it is headquartered in Dubai, it does not need to report CSAM to NCMEC. However, NCMEC reports that Telegram does respond, albeit slowly, to its CSAM notifications. In 2021, Telegram received 229 notifications and took 8.0 days to act on average. In 2022, it received 73 notifications, taking 5.1 days to act on average. By comparison, the industry-wide response time in 2021 was 1.22 days on average.

3.9 TikTok

TikTok is one of only two surveyed companies to release transparency data in machine-readable form. However, almost all data is of limited accuracy even in the absence of data quality issues, since TikTok mostly discloses percentages instead of counts. Furthermore, some of the data, including the *sexual exploitation of minors* subcategory of the *minor safety* category, is incomplete: TikTok reports the fractional share for the subcategory for human moderation only, making it impossible to calculate the full count.

For the second quarter of 2023, TikTok updated its community guidelines, including how it organizes violative categories. In its transparency report for Q2 2023, the firm declared [128]:

> This report reflects our updated Community Guidelines, which took effect in April and provide our community with more transparency about our rules and how we enforce them. After consulting with more than 100 organizations globally, we overhauled how we organize our policies, simplified the

> language we use, and added granularity to help everyone, from creators to researchers, easily access the information they need. We've also refreshed various data visualizations to make them easier to read and understand, including for people with color vision deficiency.

However, in reality, category names and organization became more confusing and less granular. The firm also continues to report fractional shares only.

To wit, before the update, the clearly named category *minor safety* included the sexual exploitation of minors, grooming behavior, physical and psychological harm of minors, harmful activities by minors, as well as nudity and sexual activity involving minors [129]. The five subcategories are largely self-explanatory; though past transparency reports helpfully added that the final subcategory largely comprises "minors in minimal clothing" and "sexually explicit dancing."

In contrast, consider the new category of *safety & civility*. Not only is it obviously coarser, but its only seven subcategories also are clearly coarser. Furthermore, their descriptions do not seem well differentiated [128]:

- Harassment & bullying: We do not allow language or behavior that harasses, humiliates, threatens, or doxxes anyone.
- Violent behaviors & criminal activities: We do not allow violent threats, incitement to violence, or promotion of crimes against people, animals, or property.
- Violent & hateful orgs & individuals: We do not allow violent and hateful organizations or individuals, or the promotion or material support of them.
- Hate speech & hateful behavior: We do not allow any hateful behavior, hate speech, or promotion of hateful ideologies.
- Human exploitation: We do not allow promoting or facilitating human exploitation, including trafficking and smuggling.
- Sexual exploitation & gender-based violence: We do not allow showing or promoting physical, sexual or image-based abuse, sextortion, or sexual harassment
- Youth exploitation & abuse: We do not allow any youth exploitation or abuse, including child sexual abuse material, pedophilia, or youth nudity.

To add insult to injury, subcategories and their explanations are only accessible through tooltips when pointing to slices of an otherwise unmarked pie chart. But even with the explanations, the distinctions between the first five subcategories can be subtle and border on arbitrary. Worse, what used to be an entire category with five subcategories, minor safety, now is just one subcategory.

TikTok released Excel files from Q3 2021 through Q3 2022 and CSV data thereafter. While data is already marred for the above stated reasons, we can still determine quarter-over-quarter stability of historical disclosures. Alas, the three comparisons between Q4 2022 and Q3 2023 (the latest available) yield over 900 non-duplicated rows, each specifying a single quantity, and that only after manually fixing column names to be consistent. Spotchecks reveal the use of different names for what should be identical categories and what appear to be unrounded floating point numbers. TikTok fails the audit.

3.10 Twitter

Unlike other surveyed platforms, Twitter discloses counts of *unique* pieces. The idiosyncratic choice of metric seems designed to downplay the problem and hence mislead readers. Related disclosures don't seem any better. Consider 2021:

Twitter disclosed 12,883 unique pieces that, per NCMEC, resulted in 86,666 CyberTipline reports. While that seems realistic, Twitter also claims to have suspended 1,050,751 accounts for child sexual exploitation. That's 12× more accounts than reports, even though a report describes a single incident and hence typically involves only a single user.

Since Elon Musk's takeover, X née Twitter made one more transparency disclosure, covering Twitter before the takeover, while also announcing the cessation of such disclosures. Its only DSA disclosure so far has the rather peculiar reporting period August 28 through October 20, features several tables with 28 columns (one for each EU member country and one for the sum) and around 30 rows as well as two tables with 28 columns and 83 rows, but does not provide the data in aggregate or machine-readable form. It almost seems like the disclosure is designed to be inscrutable on top of being meaningless thanks to its reporting period.

4 ONLINE SEXUAL EXPLOITATION OF MINORS

Having performed the scrappy audits, I now describe what those disclosures tell us about the sexual exploitation of minors online, with a focus on social media from 2019 to 2022. This section builds on the data contained in NCMEC's report to the Office of Juvenile Justice and Delinquency Prevention [101], but also draws on NCMEC's disclosures of reports per country [89, 91, 93, 96] and platforms' disclosures of CyberTipline reports and pieces, as presented in the previous section. To better contextualize my overview, I also draw on the computer-science literature about cryptographic as well as perceptual hashes and the legal as well as medical literature on teenage sexting.

Somewhat ironically, this section wouldn't be possible without the growth-obsessed, effectively neoimperialist and neocolonial business practices of US-based internet services and the resulting near-global reach of social media platforms [22, 127]. All surveyed platforms besides Telegram must report content and activities related to the sexual exploitation of minors to NCMEC, independent of users' locations, and hence produce data for most countries on Earth.

4.1 Caveats

Before discussing the data and its implications, I need to immediately register three important caveats that apply to almost all metrics used in this section. First, the primary goal for any effort in curtailing the sexual exploitation of minors should, of course, be minimizing harm to minors. One way of measuring potential harm is the *prevalence* of CSAM and violative activities on the internet. Alas, we have no data on prevalence. Even Meta, which is unique amongst surveyed platform in using random sampling to provide prevalence statistics for other violative content, has not done so for CSAM.

Second, CyberTipline reports reflect *detected incidents* of such materials and activities. They very likely are correlated with prevalence, but they nonetheless measure something different. Furthermore, I emphasized "incidents" because CyberTipline reports are not the same as *pieces* of CSAM. A report about child grooming may have no attachments or pieces, while a report about a distributor of CSAM may have a large number of attachments or pieces. While we have no data to characterize the relationship between prevalence and CyberTipline reports, we do have data on the relationship between pieces and such reports. I have already included some of that in the previous section and will discuss it below in Section 4.3.

Third, over the last few decades, people interested in CSAM have used various technologies to find each other and to exchange such material [120]. Even relatively recently, such modalities included peer-to-peer systems and the dark web. While mobile communication tools currently appear to be the dominant technology, people very likely continue to use peer-to-peer systems and the dark web. Alas, *none* of the data in this paper reflects activities on peer-to-peer systems or the dark web.

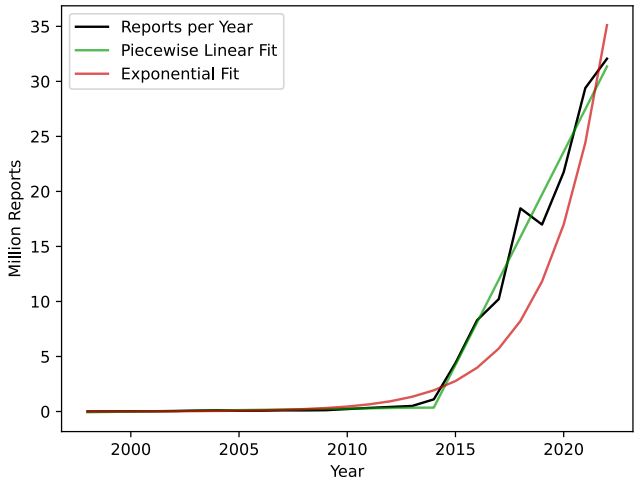

Fig. 2. Yearly CyberTipline reports (per NCMEC)

Table 7. CyberTipline violative activities in descending order for 2022 (NCMEC)

Activity	2020	%	2021	%	2022	%
Child pornography	21,669,264	99.62383	29,309,106	99.69870	31,901,234	99.50780
Online enticement for sexual acts	37,872	0.17412	44,155	0.15020	80,524	0.25117
Obscene material sent to child	3,547	0.01631	5,177	0.01761	35,624	0.11112
Child sex trafficking	15,879	0.07300	16,032	0.05453	18,336	0.05719
Child sexual molestation	11,770	0.05411	12,458	0.04238	12,906	0.04026
Misleading words/images online	8,689	0.03995	5,825	0.01981	7,517	0.02345
Misleading domain name	3,109	0.01429	3,304	0.01124	1,948	0.00608
Child sex tourism	955	0.00439	1,624	0.00552	940	0.00293

To put this differently, succinctly: I am only painting a partial picture in this section. Furthermore, I am painting a distorted picture in this section. But I can't fully tell which parts of the picture are distorted and to what degree.

4.2 CyberTipline Reports and Their Topics

Figure 2 plots yearly reports since inception of the CyberTipline in February 1999 as a black line. It also shows two least-squares-fit lines, with the piece-wise linear fit in green and the exponential fit in red. In its OJJDP report, NCMEC describes this growth as "exponential." While the sharp acceleration around 2014 does remind of exponential curves, the growth before 2014 and after 2014 didn't consistently accelerate. In fact, growth has been slowing of late and, given Meta's markedly smaller piece counts for the first three quarters of 2023 as shown in Figure 1, its average pieces to reports ratio as discussed in Section 4.3 below, and its over 80% share of all CyberTipline reports, that slowdown will very likely hold for 2023 as well.

Table 7 breaks down CyberTipline reports by violative activity from 2020 to 2022. Rows are sorted in descending order of counts for 2022. As the table makes ample clear, the vast majority of reports, over 99.5%, concerns child

Table 8. CyberTipline reports and attached pieces (NCMEC)

Year	Reports	π	Pieces	π	Unique Pieces	π	Similar Pieces
2020	21,751,085	3.00	65,344,724	2.39	27,333,171	4.62	14,141,118
2021	29,397,681	2.88	84,795,507	2.32	36,625,281	3.85	22,005,389
2022	32,059,029	2.72	87,179,813	2.01	43,399,901	3.21	27,136,862

pornography or CSAM. At the same time, the remaining fraction of a percent of reports still represents devastating harm to minors. For example, NCMEC separately discloses that in 2022 1.66 reports have the same perceptually similar attachments and hence are redundant. If we assume a similar relationship between reports and victims for sex trafficking, 18,336 reports for that activity in 2022 correspond to more than 11,000 minors being trafficked, which is an uncomfortably large number even in a global context.

4.3 CyberTipline Reports and Their Attachments, i.e., Pieces

Table 8 relates CyberTipline reports and attached pieces, that is, pictures, videos, and other documents (including PDF files). It provides a global view in addition to the platform-specific data already included in Table 2 for Facebook and Instagram, Table 3 for Google and YouTube, Table 4 for Pinterest, Table 5 for Reddit, and Table 6 for Snap. Columns labelled π for *product* show the multiplicative factor relating grouped counts to the number of pieces in the fourth column. Unlike previous tables, Table 8 counts pieces in three different ways:

- Column #4 contains the total number of pieces (as in other tables).
- Column #6 contains the number of unique pieces, filtered with MD5.
- Column #8 contains the number of perceptually similar pieces, filtered with PhotoDNA and Videntifier.

In other words, the sixth column doesn't count a photo or video if the MD5 hash function produces the same value as for a previously counted photo or video. Likewise, the eighth column doesn't count a photo or video if the PhotoDNA or Videntifier perceptual hash function produces a similar value. The critical difference between the two kinds of hash functions is that a small change in input to MD5 produces a completely different value, whereas a small change in input to PhotoDNA or Videntifier produces almost the same value, thus allowing for the detection of visually similar material [38]. Alas, the use of MD5, PhotoDNA, and Videntifier also is highly problematic.

First, cryptographically secure hash functions should make it practically impossible to construct an input for a given output. While that isn't practical for MD5 in the most general case, so-called preimage attacks, it *is* practical for more restricted cases, notably chosen prefix attacks [123]. But even a single known pair of colliding inputs, which have been published for MD5 and SHA1, suffices for constructing pairs of PDF documents with different visible contents and identical hashes [77, 122]. With some limitations, images in the JPG and PNG formats as well as videos in the MP4 format are vulnerable to such attacks as well.[4]

If internet platforms and NCMEC use MD5 hashes to identify CSAM without human intervention, as they apparently do already, that opens up the possibility of attacks that are designed to close down accounts or to trigger a law enforcement investigation akin to swatting. An attacker would need access to at least one known CSAM hash, construct a document that has the same hash, and socially engineer the targeted user into posting that document. Given the devastating consequences of a false positive for CSAM [57], the possibility of such attacks argues strongly against the continued use of MD5—and SHA1 for that matter. However, even though these vulnerabilities aren't particularly new, dating back to

[4] https://github.com/corkami/collisions

2012 and 2017, NCMEC was still asserting in 2023 that "[i]mages that share the same MD5 hash are identical," displaying no awareness of the problem [102].

Second, Microsoft's PhotoDNA [37] is widely used for CSAM detection, probably as a result of the firm making the technology freely available to organizations such as NCMEC. While Microsoft has kept the exact algorithm secret, it has not only been reverse-engineered [75], but practical second preimage attacks have been documented [108]. Furthermore, despite Microsoft's claims otherwise, it may be possible to recreate low resolution versions of the original images from PhotoDNA hashes, i.e., CSAM [8]. In other words, PhotoDNA is even more susceptible to adversarial abuse than MD5 [121].

Third, little is known about Videntifier's algorithm, which was developed by the eponymous Icelandic firm in collaboration with NCMEC. That makes it impossible to assess its strengths and weaknesses. While PhotoDNA's history suggests that such secrecy can delay but not prevent public scrutiny, the more troubling lesson from other perceptual hash algorithms is that they just aren't very robust. Notably, Meta's PDQ [26] doesn't perform all that well and is also vulnerable to second preimage attacks [76, 108]. Despite relying on a machine learning model to identify image features, Apple's NeuralHash suffers from similar limitations, including (again) susceptibility to second preimage attacks [126].

When comparing overall and platform-specific report and piece counts, Meta's and Snap's numbers are fairly close to the overall ratio. But Google's and YouTube's are significantly larger, whereas Pinterest's and Reddit's are significantly smaller. Since Google, Reddit, and YouTube passed audits of their report counts and Pinterest's report counts aren't that far off, these discrepancies most likely reflect real differences in platform usage and content moderation. However, only the platforms are in a position to identify what factors make the difference and none of them provide an explanation. Such divergences between platform-specific and industry-wide statistics are not covered in any best practices guidelines but also are just the phenomena that stick out when closely examining transparency disclosures.

4.4 Global Distribution of CyberTipline Reports

To consider the global spread of CSAM, I turn to NCMEC's reports per country covering 2019–2022 [89, 91, 91, 93]. As discussed above, the data does have minor quality issues, but they impact less than 0.001% of the reports, i.e., are negligible. However, making meaningful comparisons between countries also requires normalizing the data. While the number of countries' internet users would form the most suitable denominator, I could not find an up-to-date and accurate dataset with that information. Almost all datasets claiming to do so, including the Worldbank's, trace back to the International Telecommunication Union (ITU). But that dataset has not been updated for years for many countries and provides percentage fractions of countries' populations only. Instead of scaling population counts by outdated and hence unreliable fractions, I instead decided to use population counts by themselves. In particular, the United Nations provide up-to-date, yearly statistics, even if they often are (informed) estimates.

That normalization strategy also is the first of two major caveats about my use of reports per capita, country, and year in this section. If different countries have significantly different shares of internet users when compared to the entire population, then their reports per capita, country, and year will be skewed proportionally and hence their absolute magnitudes are incomparable. Though relative trends will still be meaningful. In practice, comparisons based on reports per capita favor Sub-Saharan African countries, which have much lower internet penetration than much of the rest of the world. That is reflected in the data: When ranking all subcontinental regions by absolute CyberTipline report counts, West-, East-, and Southern Africa place in the third quartile. But when ranking them by reports per capita, the three regions place in the fourth quartile.

Independent of dataset used for normalization, it's a good idea to drop localities with tiny populations. For instance, in 2021, the Cocos (Keeling) Islands, an Australian territory in the Indian Ocean south-westish of Sumatra, had a population of 593 and 168 CyberTipline reports. The resulting outlier of 0.283 reports per capita is seven times larger than the next largest quantity of 0.040 over all four years. On a linear, continuous color scale (which nicely avoids futzing with histogram binning strategy, one of the dark arts of data visualization), that effectively compresses the color range for all other countries to shades surprisingly similar to that for zero.

The second major caveat concerns the numerator, i.e., report counts. If countries' online populations significantly differ in what platforms they use, then cumulative report counts may be more reflective of differences in CSAM detection than differences in populations' CSAM usage. That is trivially the case for countries with a large share of platforms that are not US-based, notably the People's Republic of China. While the country appears in NCMEC's disclosures, with a maximum of 7,644 reports for 2021, the small number most certainly is a direct result of China banning many foreign websites and services including social media. In fact, out of the 16 surveyed platforms and firms, only Microsoft Bing and Skype, Snap, and Wordpress were accessible through Top10VPN's testing tool[5] in January 2024. North Korea's international rogue status is similarly reflected in the data: Its maximum was 9 reports in 2020. By comparison, South Korea, with about double the population, had 100,709 reports that same year.

I believe that a comparison based on reports per capita, country, and year still is meaningful thanks to Meta's astounding global reach and its overwhelming lead in reporting CSAM. Notably, over the four years covered by NCMEC's disclosures, the size of Meta's family monthly active people grew from 2.69 billion during Q1 2019 to 3.74 billion during Q4 2022. That is 35–47% of the world population including China—or 43–57% excluding China! Meta also was responsible for 94.3% of all CSAM reports in 2019, 94.7% in 2020, 92.2% in 2021, and 85.5% in 2022. In other words, the vast majority of CyberTipline reports reflect Meta's screening for CSAM amongst the roughly half of humanity counting as the firm's monthly users.

Figure 3 illustrates reports per capita, country, and year, using UN population data to normalize report counts, a colorscale capped by the maximum value over all four years, and the Equal Earth projection [115]. In addition to being aesthetically pleasing, that projection is equal-area and hence suitable for choropleths. Maps for 2019 and 2020 are largely consistent with the map included in in [18] based on earlier, non-public NCMEC data. Assuming that their analysis is subject to similar limitations as mine, which is reasonable given that they too start with CyberTipline report counts, the consistency suggests that CSAM sharing patterns did not much change between the two studies.

Alas, my analysis does surface one oddity: Members of the 22-country strong Arab League have noticeably high reports per capita over the covered four years; though they markedly and consistently decline over that time span as well. At the extreme, either the United Arab Emirates or Libya is the country with the most reports per capita for each of the four years. Meanwhile the decline becomes obvious if we treat the Arab League as a country: It would rank 15th, 18th, 19th, and 26th, respectively. Furthermore, per capita rates for those countries do not increase whereas those for the rest of the world do increase, with the total number of reports doubling from 2019 to 2022. While I cannot explain this oddity, the fact that the United Arab Emirates and Libya are highest ranked for two out of four years even though they are polar opposites when its comes to wealth and political stability does point to some shared cultural trait. Alas, that shared trait may very well be a bias by content moderation systems!

[5] https://www.top10vpn.com/tools/blocked-in-china/

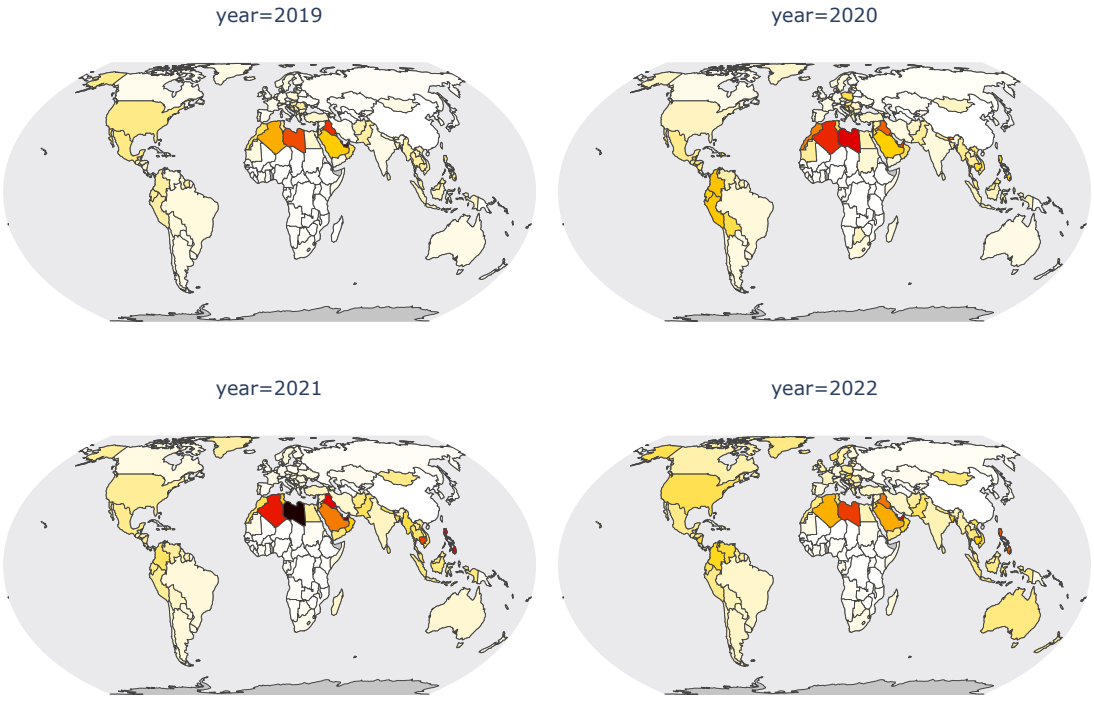

Fig. 3. Reports per capita, country, and year. The four choropleth panels plot CyberTipline reports per country and year over UN population size per country and year using an Equal Earth projection. Colors continuously range between white for 0.000 through yellow and red to black for 0.040.

4.5 Relationship between Victims and Offenders

In addition to providing additional data on CyberTipline reports, NCMEC's report to the Office of Juvenile Justice and Delinquency Prevention [101] also contains information about its Child Victim Identification Program, which maintains a separate database tracking the identities of minors appearing in CSAM. Apparently, many law enforcement agencies contribute CSAM found during investigations to the database and also provide, if known, information about the relationship between victims and suspected offenders. Unfortunately, NCMEC presents this information broken down by pieces and unique pieces but not by person. Furthermore, while yearly piece counts are just below or well over one million, the number of minors in the database appears to be much smaller. In particular, NCMEC added 773 minors in 2020, 1,477 minors in 2021, and 4,464 minors in 2022. Given the small number of victims, I also report equivalent statistics published by OJJDP based on the FBI's National Incident Reporting System Master Files for 2018 and 2019 [103].

The majority of suspects are members of the victim's extended family, accounting for 64.2% of all unique pieces in 2022. That includes fathers with 30.4%, step-fathers with 8.1%, uncles with 8.0%, and the guardian's partner with 6.3%. Another 24.6% are socially familiar, including family friends, neighbors, and teachers. The rest are online enticement

with 9.3%, trafficking with 1.4%, and strangers with 0.5%. By comparison, OJJDP's statistics attribute 53.1% to a victim's family, 44.7% to acquaintances, and 2.2% to strangers. If we treat online enticement as involving acquaintances as well, the statistics are roughly the same, with the offenders being family members in the clear majority of cases and strangers in less than one thirty-second.

It seems safe to conclude that drag queens, which count as strangers, do not figure prominently as threats. However, fathers and step-fathers most certainly do. Hence, moms for liberation would be well advised to reject the yoke of child-abusing patriarchy and instead rear their children within lesbian relationships!

4.6 Missing Data on Teenage Sexting

NCMEC provides statistics on age and gender only for minors newly added to its Child Victim Identification Program but not for CyberTipline reports. But even if it did, that wouldn't suffice for separating out reports about voluntary and consensual teenage sexting. I am highlighting sexting because it is very popular, with prevalences in a recent meta-analysis at 19.3% for sending, 34.8% for receiving, and 14.5% for forwarding without consent [86]. While the latter is ethically problematic, the harsh penalties of Chapter 110 hardly seem an appropriate tool to foster teenage learning. Yet according to federal law, teenage sexting qualifies as CSAM all the same.

Advocating an abstinence-only approach to sexting is bound to fail just like prior abstinence-only sex education did, despite the US spending over $2 billion on such programs since the 1990s [43]. While NCMEC doesn't quite advocate abstinence, its webpage on the topic[6] seems rather dismissive of teenagers who are sexting, featuring the above statistics under the technically correct but nonetheless misleading heading "Most Teens Are Not Texting," and then offers only potential negative consequences under "Protect yourself," none of them actionable advice that might help teenagers protect themselves. Furthermore, the page does recommend reporting any inappropriate incident to the CyberTipline, reassuring readers that "it is not the child who is at fault." Given US law, that seems positively misleading.

While federal prosecution of sexting teenagers is unlikely, the same cannot be said for US states. 23 out of 50 states have no exemption for consensual sexting by teenagers on the books. The other 27 do, but still punish even consensual sexting, typically as a misdemeanor [58, 59, 125]. Worse, lower penalties may have the pervese effect of making state prosecutors *less* reluctant to charge teenagers [51]. Even if that isn't the case, 62% of state prosecutors in one survey have handled such cases and 36% initiated prosecutions [140]. While they claim to have done so largely for cases where teenagers were abusive, it takes little effort to find examples where that was not the case, often with disastrous consequences including suicide [28, 40, 68, 78, 84, 142]. Against that background, NCMEC's apparent indifference to the topic and the lack of data on its share amongst CyberTipline reports is positively troubling.

4.7 Summary

Overall, the analysis of available transparency data on the sexual exploitation of minors shows an alarming growth of detected imagery and other activities over the last decade and hence is in line with similar observations by, say, the New York Times [24, 70]. In particular, since the vast majority of that growth is due to a single organization, Meta, much of that growth is very likely true growth and not attributable to improved awareness and detection technology. The analysis also suggests that existing detection technology, notably MD5 and PhotoDNA, is easily abused and hence should be phased out. At the same time, Meta's investigation of apparent intent and the lack of data on teenage sexting

[6] https://www.missingkids.org/netsmartz/topics/sexting.

also suggest that current statistics are not sufficiently differentiated and hence significant overcounts. Clearly, more robust detection technology and more fine-grained statistics are urgently needed!

5 DISCUSSION

Since audits and analysis have different overall thrusts, distinguishing between them for structuring this paper's exposition made eminent sense. But in practice, the distinction between audits and analysis was far less clear and their connection more organic: I simply kept poking at the data to make sense of available statistics. In my mind, the transition from just poking quantities to telling a coherent story grounded in the data is just what needs to happen for transparency disclosures to result in meaningful accountability. That suggests several takeaways about transparency disclosures in general and about their presentation. I start with higher-level observations.

First and not surprisingly, there is a bottom threshold for metrics and data to be useable in data-driven story telling. In particular, without having overall statistics on the relationship between CyberTipline reports and pieces, the varying platform-specific ratios were impossible to classify. NCMEC's publication of its report to the Office of Juvenile Justice and Delinquency Prevention (OJJDP) not only filled that hole but provided several more helpful statistics.

Second and more unexpectedly, there is value even in incomplete or buggy transparency disclosures *as long as* they can be combined with similar disclosures by other organizations. It helps when one organization's disclosures are general and robust enough to provide a foundation, as NCMEC's OJJDP report did for the previous section. But with that backbone in place, I could rely on platform-specific disclosures to add interesting detail, even if they, like Pinterest's, were not entirely accurate. That should serve as warning against simply dismissing existing disclosures as "transparency theater" [27].

Third, transparency disclosures run counter the scientific method. Quite literally: The latter requires the formulation of hypothesis and research questions before the selection of metrics and, if not already available, the collection of data. Yet here, civil society and other interested outsiders need to make do with the data already collected and disclosed by platforms. While actual practice isn't quite as lopsided as that (with civil society influencing platforms' practices to some degree), this *modus operandi* is reflective of data "science" in general and leads to predictably imprecise results. To stick to theatrical metaphors: If you were performing to an empty room every night, how much effort would you put into your performance?

The implication is that meaningful platform transparency requires *continuous* engagement with outside stakeholders. It also requires adjusting disclosed metrics in response to outside questions, e.g., about teenage sexting. That does complicate life for regulators, who would prefer to just pick from a menu of internationally agreed upon metrics [50]. However, keeping disclosure requirements adaptable does not mean that regulation is inadvisable or impossible [27, 71]. On the contrary, as I argued above, Meta sticking with its historical disclosures after two years of significant churn illustrates the beneficial impact of regulation. Similarly, the EU's statements of reasons database promises to enable more scrappy audits for all of content moderation [132]. At the same time, an effective regulator really needs to have their own research and data divisions to engage platforms [66, 67].

In addition to these more general observations, I do have concrete suggestions for improving transparency disclosures. Notably, fancy, irregular layouts and statistics embedded in text obfuscate the data. NCMEC actually did start reporting violative activities and attached pieces a year earlier, in its CyberTipline 2021 Report [99]. But that didn't register when I scanned the written report because of the layout and long passages of low-information text. Meanwhile, the more focused presentation of the OJJDP report with its simple, inline tables registered immediately.

Having said that, raw counts in machine-readable form, i.e., CSV, are *much preferrable*. Having to say so for data originating from some of the largest technology firms in the world feels a bit surreal. But the fact that many of them fail to do so even for DSA disclosures despite the act's explicit requirement only underlines the importance of this point. Also, do not round numbers. They are inaccurate and (literally) do not add up. Do not include (percentage) fractions. They almost always are rounded as well and hence inaccurate. Worse, if the original denominator isn't readily available, they get in the way of recovering counts. That is the case for TikTok's CSE disclosures as well as for the ITU's data on internet users per country.

6 OUTLOOK

To improve the accountability of social media platforms, I audited the transparency disclosures of 16 major platforms by comparing redundantly or repeatedly disclosed quantities. Even though the audit methodology is rather basic and I focused on an area with clear legal reporting mandates, the sexual exploitation of minors, audit results were not exactly encouraging. Out of the 16 surveyed platforms operated by 11 technology firms:

- ✘ 3 platforms do not make transparency disclosures;
- ✘ 2 platforms do not make transparency disclosures about content moderation;
- ✘ 3 platforms make such disclosures but do not meet audit requirements;
- ✘ 3 platforms failed the audit of repeatedly disclosed historical data;
- ✘ 1 platform failed the audit of redundantly disclosed CyberTipline reports;
- ✔ 4 platforms passed the audit of redundantly disclosed CyberTipline reports.

Despite these significant shortcomings, the existing data still helps characterize the sexual exploitation of minors across social media platforms and countries. At the same time, the overview would have been impossible without more comprehensive disclosures by the clearing house for reporting such activities, NCMEC.

Based on my experiences with audits and analysis, I argued that meaningful transparency disclosures require continuous engagement by platforms and civil society alike. But getting social media firms to fully embrace such an iterative transparency process and make good faith disclosures may turn out to be difficult. Meta illustrates why. The firm has been industry-leading in its efforts to suppress the spread of CSAM. Not only does it detect the vast majority of cases, but it has the longest track record of releasing statistics to the public, has consistently supported NCMEC over the last several years, with two employees serving as board members until recently, and performed helpful research to better characterize the sheer volume of CSAM.

But as my audit demonstrated, those transparency disclosures also suffer from serious data quality issues. Meta similarly misreported advertising metrics to customers [13, 62–64, 137], platform usage data to a Harvard-based research consortium [130], and adverts in its transparency database [113, 117, 119]. Human error is a plausible explanation in all these cases. Though one research group with consortium access to Meta's data also published a rather pointed critique about its quality well before the discovery that the dataset was incomplete and hence unusable [55, 56]. That experience prompted even the consortium head to call for industry-wide regulation [106].

Meanwhile, moving too fast does not suffice for explaining several other cases. Notably, Meta kept violating non-discrimination laws in advertising for years, despite having been repeatedly notified of its noncompliance [5–7, 72, 73, 80, 131]. Next, as I discovered during my employment, the firm also manipulated view impressions, only the most basic advertising metric, to its own economic benefit [48]. Despite its significant efforts on minor safety, the firm did run an advertising campaign (for TikTok nonetheless) that sexualized teenage girls for the benefit of middle-aged

straight men. When employees raised concerns, the firm cut off their access to information about the ad campaign but otherwise continued running the ads [118]. In the aftermath of the US presidential election and attempted coup and facing a barrage of criticism, Meta deplatformed the academics who had been collecting the data enabling many of the above news reports about the firm's ad database [29, 30]. At that time, the firm also disbanded the team responsible for its CrowdTangle transparency tool [112], with the project's founder leaving the firm a few months later [54] and Meta "pausing" sign-ups for new users a few more months later [105].

With the Digital Services Act, the EU created a powerful but measured, multipronged oversight and enforcement mechanism. On paper, the DSA more than suffices for reigning in a firm with a long track record of moving either too fast or too deviously. However, doing so in practice will likely require active intervention to prod Meta into compliance. Whether the commission and member countries' regulators can muster the necessary stamina and wherewithal is an open question. Unfortunately, their track record over six years with GDPR is less than impressive [17]. In other words, we are well-advised to grow the arsenal of scrappy audits and other means for holding social media accountable. Luckily, the EU Commission has already provided us with a new tool for doing so through the statements of reasons database. While other researchers have already performed a first comparison of database entries with aggregate DSA transparency disclosures, I am particularly interested in fully automating these audits to hold social media accountable for correct counts when it comes to transparency!

ACKNOWLEDGMENTS

An NCMEC employee helped complicated my thinking about CSAM—while somehow also avoiding to answer any of my questions. Meanwhile Karin Wolman did answer my questions about 18 US Code. Thank you both!

REFERENCES

[1] Access Now. 2021. Transparency Reporting Index. https://www.accessnow.org/transparency-reporting-index/

[2] Access Now, ACLU Foundation of Northern California, ACLU Foundation of Southern California, Article 19, Brennan Center for Justice, Center for Democracy & Technology, Electronic Frontier Foundation, Global Partners Digital, InternetLab, National Coalition Against Censorship, New America's Open Technology Institute, Ranking Digital Rights, Red en Defensa de los Derechos Digitales, and Witness. 2021. Santa Clara Principles on Transparency and Accountability in Content Moderation. https://santaclaraprinciples.org

[3] Action Coalition on Meaningful Transparency. 2023. *Implementing Risk Assessment Obligations under the Digital Services Act.* Technical Report. Action Coalition on Meaningful Transparency. https://4eec262c-28cb-48a0-a0c6-572b30649370.usrfiles.com/ugd/4eec26_09033519b6db46129e7924bbc460974f.pdf

[4] Ada Lovelace Institute. 2021. *Technical Methods for Regulatory Inspection of Algorithmic Systems.* Technical Report. Ada Lovelace Institute. https://www.adalovelaceinstitute.org/report/technical-methods-regulatory-inspection/

[5] Julia Angwin and Terry Parris Jr. 2016. Facebook Lets Advertisers Exclude Users by Race. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race

[6] Julia Angwin, Noam Scheiber, and Ariana Tobin. 2017. Dozens of Companies Are Using Facebook to Exclude Older Workers From Job Ads. *ProPublica* (Dec. 2017). https://www.propublica.org/article/facebook-ads-age-discrimination-targeting

[7] Julia Angwin, Ariana Tobin, and Madeleine Varner. 2017. Facebook (Still) Letting Housing Advertisers Exclude Users by Race. *ProPublica* (Nov. 2017). https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin

[8] Anish Athalye. 2021. Inverting PhotoDNA. https://anishathalye.com/inverting-photodna/

[9] Aliya Bhatia and Asha Allen. 2023. Auditing in the Dark: Guidance Is Needed to Ensure Maximum Impact of DSA Algorithmic Audits. https://cdt.org/insights/auditing-in-the-dark-guidance-is-needed-to-ensure-maximum-impact-of-dsa-algorithmic-audits/

[10] Matt Binder. 2018. Twitter Releases 2018 Transparency Report Including Policy Violation Stats for the First Time. *Mashable* (Dec. 2018). https://mashable.com/article/twitter-transparency-report-2018

[11] Anu Bradford. 2020. *The Brussels Effect.* Oxford University Press, Oxford, United Kingdom. https://www.brusselseffect.com/

[12] Ben Bradford, Florian Grisel, Tracey L. Meares, Emily Owens, Baron L. Pineda, Jacob N. Shapiro, Tom R. Tyler, and Danieli Evans Peterman. 2019. *Report of the Facebook Data Transparency Advisory Group.* DTAG Report. Yale Law School. https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf

[13] Alexandra Bruell and Sahil Patel. 2020. Facebook's Latest Error Shakes Advertisers' Confidence. *Wall Street Journal* (Nov. 2020). https://www.wsj.com/articles/facebooks-latest-error-shakes-advertisers-confidence-11606346927

[14] Axel Bruns. 2019. After the 'APIcalypse': Social Media Platforms and Their Fight against Critical Scholarly Research. *Information, Communication & Society* 22, 11 (Sept. 2019), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

[15] John Buckley, Malia Andrus, and Chris Williams. 2021. Understanding the Intentions of Child Sexual Abuse Material (CSAM) Sharers - Meta Research. https://research.facebook.com/blog/2021/02/understanding-the-intentions-of-child-sexual-abuse-material-csam-sharers/

[16] Cody Buntain, Monique Deal Barlow, Mia Bloom, and Mila A. Johns. 2022. Paved with Bad Intentions: QAnon's Save the Children Campaign. *Journal of Online Trust and Safety* 1, 2 (Feb. 2022). https://doi.org/10.54501/jots.v1i2.51

[17] Matt Burgess. 2022. How GDPR Is Failing. *Wired* (May 2022). https://www.wired.com/story/gdpr-2022/

[18] Elie Bursztein, Travis Bright, Michelle DeLaune, David M. Eliff, Nick Hsu, Lindsey Olson, John Shehan, Madhukar Thakur, and Kurt Thomas. 2019. Rethinking the Detection of Child Sexual Abuse Imagery on the Internet. In *The World Wide Web Conference*. International World Wide Web Conference Committee, San Francisco, CA, USA, 2601–2607. https://doi.org/10.1145/3308558.3313482

[19] Dell Cameron, Shoshana Wodinsky, Mack DeGeurin, and Thomas Germain. 2023. Read the Facebook Papers for Yourself. *Gizmodo* (June 2023). https://gizmodo.com/facebook-papers-how-to-read-1848702919

[20] Hichang Cho, Pengxiang Li, Annabel Ngien, Marion Grace Tan, Anfan Chen, and Elmie Nekmat. 2023. The Bright and Dark Sides of Social Media Use during COVID-19 Lockdown: Contrasting Social Media Effects through Social Liability vs. Social Support. *Computers in Human Behavior* 146 (Sept. 2023), 107795. https://doi.org/10.1016/j.chb.2023.107795

[21] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1571–1583. https://doi.org/10.1145/3531146.3533213

[22] Nick Couldry and Ulises A. Mejias. 2019. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford University Press, Stanford, California, USA. https://www.sup.org/books/title/?id=28816

[23] Andrew Crocker, Gennie Gebhart, Aaron Mackey, Kurt Opsahl, Hayley Tsukayama, Jamie Lee Williams, and Jillian C. York. 2019. Who Has Your Back? Censorship Edition 2019. https://www.eff.org/wp/who-has-your-back-2019

[24] Gabriel J. X. Dance. 2019. Fighting the Good Fight Against Online Child Sexual Abuse. *The New York Times* (Dec. 2019). https://www.nytimes.com/interactive/2019/12/22/us/child-sex-abuse-websites-shut-down.html,https://www.nytimes.com/interactive/2019/12/22/us/child-sex-abuse-websites-shut-down.html

[25] Antigone Davis. 2021. Preventing Child Exploitation on Our Apps. https://about.fb.com/news/2021/02/preventing-child-exploitation-on-our-apps/

[26] Antigone Davis and Guy Rosen. 2019. Open-Sourcing Photo- and Video-Matching Technology to Make the Internet Safer. https://about.fb.com/news/2019/08/open-source-photo-video-matching/

[27] Evelyn Douek. 2022. Content Moderation as Systems Thinking. *Harvard Law Review* 136, 2 (Dec. 2022), 526–607. https://harvardlawreview.org/2022/12/content-moderation-as-systems-thinking/

[28] Bridgette Dunlap. 2016. Why Prosecuting a Teen Girl for Sexting Is Absurd. *Rolling Stone* (Oct. 2016). https://www.rollingstone.com/culture/culture-news/why-prosecuting-a-teen-girl-for-sexting-is-absurd-127458/

[29] Laura Edelson and Damon McCoy. 2021. How Facebook Hinders Misinformation Research. *Scientific American* (Sept. 2021). https://www.scientificamerican.com/article/how-facebook-hinders-misinformation-research/

[30] Laura Edelson and Damon McCoy. 2021. We Research Misinformation on Facebook. It Just Disabled Our Accounts. *The New York Times* (Aug. 2021). https://www.nytimes.com/2021/08/10/opinion/facebook-misinformation.html

[31] Vittoria Elliott, Nilesh Christopher, Andrew Deck, and Leo Schwartz. 2021. The Facebook Papers Reveal Staggering Failures in the Global South. *Rest of World* (Oct. 2021). https://restofworld.org/2021/facebook-papers-reveal-staggering-failures-in-global-south/

[32] European Commission. 2023. Delegated Regulation on Independent Audits under the Digital Services Act. https://digital-strategy.ec.europa.eu/en/library/delegated-regulation-independent-audits-under-digital-services-act

[33] European Commission. 2023. Digital Services Act: Commission Designates First Set of Very Large Online Platforms and Search Engines. https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413

[34] European Parliament and Council. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act). https://eur-lex.europa.eu/eli/reg/2022/2065/oj

[35] Facebook. 2021. Safety and Integrity on Our Platforms: Progress in Q2. https://www.facebook.com/business/news/safety-and-integrity-on-our-platforms-progress-in-q2

[36] Facebook. 2021. *Safety and Integrity Quarterly Roundup: Q2 2021*. Roundup. Facebook. https://facebook.com/business/f/973284106800107

[37] Hany Farid. 2018. Reining in Online Abuses. *Technology & Innovation* 19, 3 (Feb. 2018), 593–599. https://doi.org/10.21300/19.3.2018.593

[38] Hany Farid. 2021. An Overview of Perceptual Hashing. *Journal of Online Trust and Safety* 1, 1 (Oct. 2021). https://doi.org/10.54501/jots.v1i1.24

[39] John Feffer. 2021. The Global Right Wing's Bizarre Obsession with Pedophilia. https://ips-dc.org/the-global-right-wings-bizarre-obsession-with-pedophilia/

[40] Amy E. Feldman. 2020. For Teens, Sexting Can Be a Crime. *Wall Street Journal* (Nov. 2020). https://www.wsj.com/articles/for-teens-sexting-can-be-a-crime-11605801722

[41] Adrienne L. Fernandes-Alcantara and Emily J. Hanson. 2021. *The Missing and Exploited Children's (MEC) Program: Background and Policies.* Technical Report RL34050. Congressional Research Service. https://crsreports.congress.gov/product/pdf/RL/RL34050

[42] Timothy Foley and Melda Gurakar. 2022. Backlash or Bullying? Online Harassment, Social Sanction, and the Challenge of COVID-19 Misinformation. *Journal of Online Trust and Safety* 1, 2 (Feb. 2022). https://doi.org/10.54501/jots.v1i2.31

[43] Ashley M. Fox, Georgia Himmelstein, Hina Khalid, and Elizabeth A. Howell. 2019. Funding for Abstinence-Only Education and Adolescent Pregnancy Prevention: Does State Ideology Affect Outcomes? *American Journal of Public Health* 109, 3 (March 2019), 497–504. https://doi.org/10.2105/AJPH.2018.304896

[44] Camille François and evelyn douek. 2021. The Accidental Origins, Underappreciated Limits, and Enduring Promises of Platform Transparency Reporting about Information Operations. *Journal of Online Trust and Safety* 1, 1 (Oct. 2021). https://doi.org/10.54501/jots.v1i1.17

[45] David Gilbert. 2023. A Far-Right Moms Group Is Terrorizing Schools in the Name of Protecting Kids. *Vice* (April 2023). https://www.vice.com/en/article/dy3gnq/what-is-moms-for-liberty

[46] Trisha Greenhalgh, Mustafa Ozbilgin, and David Tomlinson. 2022. How Covid-19 Spreads: Narratives, Counter Narratives, and Social Dramas. *BMJ* 378 (Aug. 2022), e069940. https://doi.org/10.1136/bmj-2022-069940

[47] Susanna Greijer and Jaap Doek. 2016. *Terminology Guidelines for the Protection of Children from Sexual Exploitation and Sexual Abuse.* Technical Report. ECPAT International, Bangkok, Thailand. https://www.unicef.org/media/66731/file/Terminology-guidelines.pdf

[48] Robert Grimm. 2022. Wrong Impression: "We Don't Reveal This Definition Externally". https://apparebit.com/blog/2022/wrong-impression

[49] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 466:1–466:35. https://doi.org/10.1145/3479610

[50] Anna-Sophie Harling, Declan Henesy, and Eleanor Simmance. 2023. Transparency Reporting: The UK Regulatory Perspective. *Journal of Online Trust and Safety* 1, 5 (Jan. 2023). https://doi.org/10.54501/jots.v1i5.108

[51] Amy Adele Hasinoff. 2016. Teenage Sexting Is Not Child Porn. *The New York Times* (April 2016). https://www.nytimes.com/2016/04/04/opinion/teenage-sexting-is-not-child-porn.html

[52] Dawn Hawkins. 2023. NCOSE's Commitment to the LGBTQ+ Community. https://endsexualexploitation.org/articles/ncoses-commitment-to-the-lgbtq-community/

[53] Rob Hawkins. 2023. About Reveddit: FAQ. https://www.reveddit.com/about/faq/

[54] Alex Heath. 2021. The Founder of Facebook's CrowdTangle Tool Is Leaving. *The Verge* (Oct. 2021). https://www.theverge.com/2021/10/6/22713109/facebook-crowdtangle-founder-brandon-silverman-leaves

[55] Simon Hegelich. 2020. Facebook Needs to Share More with Researchers. *Nature* 579, 7800 (March 2020), 473–473. https://doi.org/10.1038/d41586-020-00828-5

[56] Simon Hegelich, Fabienne Marco, Joana Bayraktar, and Morteza Shahrezaye. 2020. The Social Science One Facebook Cooperation: A Systemic Failure. https://politicaldatascience.blogspot.com/2020/03/the-social-science-one-facebook.html

[57] Kashmir Hill. 2022. A Dad Took Photos of His Naked Toddler for the Doctor. Google Flagged Him as a Criminal. *The New York Times* (Aug. 2022). https://www.nytimes.com/2022/08/21/technology/google-surveillance-toddler-photo.html

[58] Sameer Hinduja and Justin W. Patchin. 2022. *State Sexting Laws.* Technical Report. Cyberbullying Research Center. https://cyberbullying.org/pdfs/2022_Sexting_Laws.pdf

[59] Brian Holoyda, Jacqueline Landess, Renee Sorrentino, and Susan Hatters Friedman. 2018. Trouble at Teens' Fingertips: Youth Sexting and the Law. *Behavioral Sciences & the Law* 36, 2 (March 2018), 170–181. https://doi.org/10.1002/bsl.2335

[60] Jeff Horwitz and Katherine Blunt. 2023. Instagram Connects Vast Pedophile Network. *Wall Street Journal* (June 2023). https://www.wsj.com/articles/instagram-vast-pedophile-network-4ab7189

[61] Jeff Horwitz and Katherine Blunt. 2023. Meta Is Struggling to Boot Pedophiles Off Facebook and Instagram. *Wall Street Journal* (Dec. 2023). https://www.wsj.com/tech/meta-facebook-instagram-pedophiles-enforcement-struggles-dceb3548

[62] Andrew Hutchinson. 2016. A Complete List of Facebook's Misreported Metrics and What They Mean. *Social Media Today* (Dec. 2016). https://www.socialmediatoday.com/social-networks/complete-list-facebooks-misreported-metrics-and-what-they-mean

[63] Andrew Hutchinson. 2016. On Facebook's Inflated Video Metrics and What It Means for Marketers. *Social Media Today* (Sept. 2016). https://www.socialmediatoday.com/social-business/facebooks-inflated-video-metrics-and-what-it-means-marketers

[64] Andrew Hutchinson. 2017. Facebook Found Two New Errors in Their Ad Metrics, Issued Refunds. *Social Media Today* (Nov. 2017). https://www.socialmediatoday.com/news/facebooks-found-two-new-errors-in-their-ad-metrics-issued-refunds/510291/

[65] Mike Isaac. 2023. Reddit Wants to Get Paid for Helping to Teach Big A.I. Systems. *The New York Times* (April 2023). https://www.nytimes.com/2023/04/18/technology/reddit-ai-openai-google.html

[66] Julian Jaursch. 2022. *Platform Oversight: Here Is What a Strong Digital Services Coordinator Should Look Like.* Policy Brief. Stiftung Neue Verantwortung, Berlin, Germany. https://www.stiftung-nv.de/en/publication/platform-oversight-what-strong-digital-services-coordinator-should-look

[67] Julian Jaursch. 2023. Here Is Why Digital Services Coordinators Should Establish Strong Research and Data Units. https://dsa-observatory.eu/2023/03/10/here-is-why-digital-services-coordinators-should-establish-strong-research-and-data-units/

[68] Justin Jouvenal. 2023. Teen 'Sexting' Case Goes to Trial in Fairfax County. *Washington Post* (May 2023). https://www.washingtonpost.com/local/teen-sexting-case-goes-to-trial-in-fairfax-county/2013/04/17/4936b768-a6b7-11e2-b029-8fb7e977ef71_story.html

[69] Thomas E. Kadri. 2022. Juridical Discourse for Platforms. *Harvard Law Review* 136, 2 (Dec. 2022). https://harvardlawreview.org/forum/vol-136/juridical-discourse-for-platforms/

[70] Michael H. Keller and Gabriel J. X. Dance. 2019. The Internet Is Overrun With Images of Child Sexual Abuse. What Went Wrong? *The New York Times* (Sept. 2019). https://www.nytimes.com/interactive/2019/09/28/us/child-sex-abuse.html

[71] Kate Klonick. 2023. Of Systems Thinking and Straw Men. https://harvardlawreview.org/forum/vol-136/of-systems-thinking-and-straw-men/

[72] Ava Kofman and Ariana Tobin. 2019. Facebook Ads Can Still Discriminate Against Women and Older Workers, Despite a Civil Rights Settlement. *ProPublica* (Dec. 2019). https://www.propublica.org/article/facebook-ads-can-still-discriminate-against-women-and-older-workers-despite-a-civil-rights-settlement

[73] Ariana Tobin Kofman, Ava. 2022. Facebook Finally Agrees to Eliminate Tool That Enabled Discriminatory Advertising. *ProPublica* (June 2022). https://www.propublica.org/article/facebook-doj-advertising-discrimination-settlement

[74] Anastasia Kozyreva, Stefan M. Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser, and Jason Reifler. 2023. Resolving Content Moderation Dilemmas between Free Speech and Harmful Misinformation. *Proceedings of the National Academy of Sciences* 120, 7 (Feb. 2023), e2210666120. https://doi.org/10.1073/pnas.2210666120

[75] Neal Krawetz. 2021. PhotoDNA and Limitations. https://www.hackerfactor.com/blog/index.php?archives/931-PhotoDNA-and-Limitations.html

[76] Neal Krawetz. 2022. FB TMK PDQ WTF. https://www.hackerfactor.com/blog/index.php?/archives/971-FB-TMK-PDQ-WTF.html

[77] Gaëtan Leurent and Thomas Peyrin. 2020. SHA-1 Is a Shambles - First Chosen-Prefix Collision on SHA-1 and Application to the PGP Web of Trust. https://eprint.iacr.org/2020/014

[78] Melissa R. Lorang, Dale E. McNiel, and Renée L. Binder. 2016. Minors and Sexting: Legal Implications. *Journal of the American Academy of Psychiatry and the Law Online* 44, 1 (March 2016), 73–81. https://jaapl.org/content/44/1/73

[79] Alfred Lua. 2022. 20 Top Social Media Sites to Consider for Your Brand in 2023. https://buffer.com/library/social-media-sites/

[80] Jeremy B. Merrill. 2020. Does Facebook Still Sell Discriminatory Ads? *The Markup* (Aug. 2020). https://themarkup.org/the-breakdown/2020/08/25/does-facebook-still-sell-discriminatory-ads

[81] Anna-Katharina Meßmer and Martin Degeling. 2023. *Auditing Recommender Systems: Putting the DSA into practice with a risk-scenario-based approach.* Technical Report. Stiftung Neue Verantwortung, Berlin, Germany. https://www.stiftung-nv.de/de/publication/auditing-recommender-systems

[82] Meta. 2022. *Report of Management on the Internal Controls over the Calculation and Reporting of the Facebook and Instagram Community Standards Enforcement Report as of December 31, 2021 and the Calculation of the Metrics Reported within the Facebook and Instagram Community Standards Enforcement Report for the Period October 1, 2021 to December 31, 2021.* Technical Report. Meta. https://about.fb.com/wp-content/uploads/2022/05/EY-CSER-Independent-Assessment-Q4-2021.pdf

[83] Meta. 2023. Transparency into Meta's Reports To the National Center for Missing and Exploited Children | Transparency Center. https://transparency.fb.com/ncmec-q2-2023/

[84] Michael E. Miller. 2015. N.C. Just Prosecuted a Teenage Couple for Making Child Porn — of Themselves. *Washington Post* (Sept. 2015). https://www.washingtonpost.com/news/morning-mix/wp/2015/09/21/n-c-just-prosecuted-a-teenage-couple-for-making-child-porn-of-themselves/

[85] Martha Minow and Newton Minow. 2023. Social Media Companies Should Pursue Serious Self-Supervision — Soon: Response to Professors Douek and Kadri. *Harvard Law Review* 136, 8 (June 2023). https://harvardlawreview.org/forum/vol-136/social-media-companies-should-pursue-serious-self-supervision-soon-response-to-professors-douek-and-kadri/

[86] Camille Mori, Julianna Park, Jeff R. Temple, and Sheri Madigan. 2022. Are Youth Sexting Rates Still on the Rise? A Meta-analytic Update. *Journal of Adolescent Health* 70, 4 (April 2022), 531–539. https://doi.org/10.1016/j.jadohealth.2021.10.026

[87] National Center for Missing and Exploited Children. 2017. CyberTipline Report 5074778. https://epic.org/wp-content/uploads/amicus/algorithmic-transparency/wilson/62-3.pdf

[88] National Center for Missing and Exploited Children. 2017. CyberTipline Report 5778397. https://epic.org/wp-content/uploads/amicus/algorithmic-transparency/miller/US-Exhibits-Report.pdf

[89] National Center for Missing and Exploited Children. 2020. *2019 CyberTipline Reports by Country.* Transparency Report. National Center for Missing and Exploited Children. https://www.missingkids.org/content/dam/missingkids/pdfs/2019%20CyberTipline%20Reports%20by%20Country.pdf

[90] National Center for Missing and Exploited Children. 2020. *2019 CyberTipline Reports by Electronic Service Providers (ESP).* Transparency Report. National Center for Missing and Exploited Children. https://www.missingkids.org/content/dam/missingkids/pdfs/2019-reports-by-esp.pdf

[91] National Center for Missing and Exploited Children. 2021. *2020 CyberTipline Reports by Country.* Transparency Report. National Center for Missing and Exploited Children. https://www.missingkids.org/content/dam/missingkids/pdfs/2020-reports-by-country.pdf

[92] National Center for Missing and Exploited Children. 2021. *2020 CyberTipline Reports by Electronic Service Providers (ESP).* Transparency Report. National Center for Missing and Exploited Children. https://www.missingkids.org/content/dam/missingkids/pdfs/2020-reports-by-esp.pdf

[93] National Center for Missing and Exploited Children. 2022. *2021 CyberTipline Reports by Country.* Transparency Report. National Center for Missing and Exploited Children. https://www.missingkids.org/content/dam/missingkids/pdfs/2021-reports-by-country.pdf

[94] National Center for Missing and Exploited Children. 2022. *2021 CyberTipline Reports by Electronic Service Providers (ESP).* Transparency Report. National Center for Missing and Exploited Children. https://www.missingkids.org/content/dam/missingkids/pdfs/2021-reports-by-esp.pdf

[95] National Center for Missing and Exploited Children. 2022. *2021 Notifications Sent by NCMEC Per Electronic Service Providers (ESP).* Technical Report. National Center for Missing and Exploited Children. https://www.missingkids.org/content/dam/missingkids/pdfs/2021-notifications-by-

ncmec-per-esp.pdf

[96] National Center for Missing and Exploited Children. 2023. *2022 CyberTipline Reports by Country.* Technical Report. National Center for Missing and Exploited Children. https://www.missingkids.org/content/dam/missingkids/pdfs/2022-reports-by-country.pdf

[97] National Center for Missing and Exploited Children. 2023. *2022 CyberTipline Reports by Electronic Service Providers (ESP).* Technical Report. National Center for Missing and Exploited Children. https://www.missingkids.org/content/dam/missingkids/pdfs/2022-reports-by-esp.pdf

[98] National Center for Missing and Exploited Children. 2023. *2022 Notifications Sent by NCMEC per Electronic Service Providers (ESP).* Technical Report. National Center for Missing and Exploited Children. https://www.missingkids.org/content/dam/missingkids/pdfs/2022-notifications-by-ncmec-per-esp.pdf

[99] National Center for Missing and Exploited Children. 2023. *CyberTipline 2021 Report.* Technical Report. National Center for Missing and Exploited Children. https://www.missingkids.org/content/dam/missingkids/pdfs/2021-CyberTipline-Report.pdf

[100] National Center for Missing and Exploited Children. 2023. *CyberTipline 2022 Report.* Technical Report. National Center for Missing and Exploited Children. https://www.missingkids.org/content/dam/missingkids/pdfs/2022-CyberTipline-Report.pdf

[101] National Center for Missing and Exploited Children. 2023. *U.S. Department of Justice—CY 2022 Report to the Committees on Appropriations—National Center for Missing and Exploited Children (NCMEC) Transparency.* Technical Report. National Center for Missing and Exploited Children. https://www.missingkids.org/content/dam/missingkids/pdfs/OJJDP-NCMEC-Transparency_2022-Calendar-Year.pdf

[102] National Center for Missing and Exploited Children. 2024. CyberTipline Reporting API Technical Documentation. https://report.cybertip.org/ispws/documentation/

[103] Office of Juvenile Justice and Delinquency Prevention. 2022. OJJDP Statistical Briefing Book. https://www.ojjdp.gov/ojstatbb/victims/qa02403.asp?qaDate=2019

[104] Nilay Patel. 2022. How to Buy a Social Network, with Tumblr CEO Matt Mullenweg. *The Verge* (Dec. 2022). https://www.theverge.com/23506085/wordpress-twitter-tumblr-ceo-matt-mullenweg-elon-musk

[105] Shivam Patel and Elizabeth Culliford. 2022. Meta Pauses New Users from Joining Analytics Tool CrowdTangle. *Reuters* (Jan. 2022). https://www.reuters.com/technology/meta-pauses-new-users-joining-analytics-tool-crowdtangle-2022-01-29/

[106] Nathaniel Persily and Joshua A Tucker. 2021. *How to Fix Social Media? Start with Independent Research.* Technical Report. Brookings Institution. https://www.brookings.edu/research/how-to-fix-social-media-start-with-independent-research/

[107] Radha Iyengar Plumb. 2019. An Independent Report on How We Measure Content Moderation. https://about.fb.com/news/2019/05/dtag-report/

[108] Jonathan Prokos, Tushar M. Jois, Neil Fendley, Roei Schuster, Matthew Green, Eran Tromer, and Yinzhi Cao. 2021. Squint Hard Enough: Evaluating Perceptual Hashing with Machine Learning. https://eprint.iacr.org/2021/1531

[109] Cornelius Puschmann. 2019. An End to the Wild West of Social Media Research: A Response to Axel Bruns. *Information, Communication & Society* 22, 11 (Sept. 2019), 1582–1589. https://doi.org/10.1080/1369118X.2019.1646300

[110] Reddit. 2022. *Transparency Report 2021.* Technical Report. Reddit. https://www.redditinc.com/policies/transparency-report-2021-2/

[111] Aja Romano. 2022. The Right's Moral Panic over "Grooming" Invokes Age-Old Homophobia. *Vox* (April 2022). https://www.vox.com/culture/23025505/leftist-groomers-homophobia-satanic-panic-explained

[112] Kevin Roose. 2021. Inside Facebook's Data Wars. *The New York Times* (July 2021). https://www.nytimes.com/2021/07/14/technology/facebook-data.html

[113] Matthew Rosenberg. 2019. Ad Tool Facebook Built to Fight Disinformation Doesn't Work as Advertised. *The New York Times* (July 2019). https://www.nytimes.com/2019/07/25/technology/facebook-ad-library.html

[114] Vishwanath Sarang. 2022. Community Standards Enforcement Report Assessment Results. https://about.fb.com/news/2022/05/community-standards-enforcement-report-assessment-results/

[115] Bojan Šavrič, Tom Patterson, and Bernhard Jenny. 2019. The Equal Earth Map Projection. *International Journal of Geographical Information Science* 33, 3 (March 2019), 454–465. https://doi.org/10.1080/13658816.2018.1504949

[116] Mark Scott and Laura Kayali. 2020. What Happened When Humans Stopped Managing Social Media Content. *Politico* (Oct. 2020). https://www.politico.eu/article/facebook-content-moderation-automation/

[117] Mark Scott and Zach Montellaro. 2021. Scores of Political Groups Sidestepped Facebook's Ad Ban. *Politico* (March 2021). https://www.politico.com/news/2021/03/04/politcal-groups-facebook-ad-ban-473698

[118] Craig Silverman and Ryan Mac. 2020. "Facebook Get Paid". *BuzzFeed News* (Dec. 2020). https://www.buzzfeednews.com/article/craigsilverman/facebook-ad-scams-revenue-china-tiktok-vietnam

[119] Craig Silverman and Ryan Mac. 2020. Facebook Promised to Label Political Ads, but Ads for Biden, The Daily Wire, and Interest Groups Are Slipping Through. *BuzzFeed News* (Oct. 2020). https://www.buzzfeednews.com/article/craigsilverman/facebook-biden-election-ads

[120] Chad M. S. Steel, Emily Newman, Suzanne O'Rourke, and Ethel Quayle. 2020. An Integrative Review of Historical Technology and Countermeasure Usage Trends in Online Child Sexual Exploitation Material Offenders. *Forensic Science International: Digital Investigation* 33 (June 2020), 300971. https://doi.org/10.1016/j.fsidi.2020.300971

[121] Martin Steinebach. 2023. An Analysis of PhotoDNA. In *Proceedings of the 18th International Conference on Availability, Reliability and Security.* ACM, Benevento Italy, 1–8. https://doi.org/10.1145/3600160.3605048

[122] Marc Stevens, Elie Bursztein, Pierre Karpman, Ange Albertini, and Yarik Markov. 2017. The First Collision for Full SHA-1. In *Advances in Cryptology – CRYPTO 2017*, Jonathan Katz and Hovav Shacham (Eds.). Vol. 10401. Springer International Publishing, Cham, 570–596. https:

//doi.org/10.1007/978-3-319-63688-7_19

[123] Marc Stevens, Arjen K. Lenstra, and Benne De Weger. 2012. Chosen-Prefix Collisions for MD5 and Applications. *International Journal of Applied Cryptography* 2, 4 (2012), 322. https://doi.org/10.1504/IJACT.2012.048084

[124] Kathleen Stoughton and Paul Rosenzweig. 2022. Toward Greater Content Moderation Transparency Reporting. https://www.lawfareblog.com/toward-greater-content-moderation-transparency-reporting

[125] Victor C. Strasburger, Harry Zimmerman, Jeff R. Temple, and Sheri Madigan. 2019. Teenagers, Sexting, and the Law. *Pediatrics* 143, 5 (May 2019), e20183183. https://doi.org/10.1542/peds.2018-3183

[126] Lukas Struppek, Dominik Hintersdorf, Daniel Neider, and Kristian Kersting. 2022. Learning to Break Deep Perceptual Hashing: The Use Case NeuralHash. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 58–69. https://doi.org/10.1145/3531146.3533073 arXiv:2111.06628 [cs]

[127] Jim Thatcher, David O'Sullivan, and Dillon Mahmoudi. 2016. Data Colonialism through Accumulation by Dispossession: New Metaphors for Daily Data. *Environment and Planning D: Society and Space* 34, 6 (Dec. 2016), 990–1006. https://doi.org/10.1177/0263775816633195

[128] TikTok. 2023. *Community Guidelines Enforcement Report April 1, 2023–June 30, 2023.* Technical Report. TikTok. https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2023-2/

[129] TikTok. 2023. *Community Guidelines Enforcement Report January 1, 2023–March 31, 2023.* Technical Report. TikTok. https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2023-1/

[130] Craig Timberg. 2021. Facebook Made Big Mistake in Data It Provided to Researchers, Undermining Academic Work. *Washington Post* (Sept. 2021). https://www.washingtonpost.com/technology/2021/09/10/facebook-error-data-social-scientists/

[131] Ariana Tobin and Jeremy B. Merrill. 2018. Facebook Is Letting Job Advertisers Target Only Men. https://www.propublica.org/article/facebook-is-letting-job-advertisers-target-only-men

[132] Amaury Trujillo, Tiziano Fagni, and Stefano Cresci. 2024. The DSA Transparency Database: Auditing Self-reported Moderation Actions by Social Media. arXiv:2312.10269 [cs] http://arxiv.org/abs/2312.10269

[133] Trust & Safety Professional Association. 2022. History of Transparency Reports. https://www.tspa.org/curriculum/ts-fundamentals/transparency-report/history-transparency-reports/

[134] Twitter. 2022. *Rules Enforcement.* Transparency Report 20. Twitter. https://transparency.twitter.com/en/reports/rules-enforcement.html

[135] United States. 2018. Sexual Exploitation and Other Abuse of Children. https://www.law.cornell.edu/uscode/text/18/part-I/chapter-110

[136] Aleksandra Urman and Mykola Makhortykh. 2023. How Transparent Are Transparency Reports? Comparative Analysis of Transparency Reporting across Online Platforms. *Telecommunications Policy* 47, 3 (April 2023), 102477. https://doi.org/10.1016/j.telpol.2022.102477

[137] Suzanne Vranica and Jack Marshall. 2016. Facebook Overestimated Key Video Metric for Two Years. *Wall Street Journal* (Sept. 2016). http://www.wsj.com/articles/facebook-overestimated-key-video-metric-for-two-years-1474586951

[138] Ben Wagner, Krisztina Rozgonyi, Marie-Therese Sekwenz, Jennifer Cobbe, and Jatinder Singh. 2020. Regulating Transparency?: Facebook, Twitter and the German Network Enforcement Act. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 261–271. https://doi.org/10.1145/3351095.3372856

[139] Shawn Walker, Dan Mercea, and Marco Bastos. 2019. The Disinformation Landscape and the Lockdown of Social Platforms. *Information, Communication & Society* 22, 11 (Sept. 2019), 1531–1543. https://doi.org/10.1080/1369118X.2019.1648536

[140] Wendy Walsh, Janis Wolak, and David Finkelhor. 2013. *Sexting: When Are State Prosecutors Deciding to Prosecute? The Third National Juvenile Online Victimization Study (NJOV-3).* Technical Report. University of New Hampshire. https://scholars.unh.edu/ccrc/43/

[141] Wikipedia. 2024. National Center on Sexual Exploitation. *Wikipedia* (Jan. 2024). https://en.wikipedia.org/w/index.php?title=National_Center_on_Sexual_Exploitation&oldid=1194719506

[142] Janis Wolak, David Finkelhor, and Kimberly J. Mitchell. 2012. How Often Are Teens Arrested for Sexting? Data From a National Sample of Police Cases. *Pediatrics* 129, 1 (Jan. 2012), 4–12. https://doi.org/10.1542/peds.2011-2242

[143] Jillian C. York. 2018. Facebook Releases First-Ever Community Standards Enforcement Report. https://www.eff.org/deeplinks/2018/05/facebook-releases-first-ever-community-standards-enforcement-report

[144] Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* PublicAffairs, New York, NY, USA. https://www.hachettebookgroup.com/titles/shoshana-zuboff/the-age-of-surveillance-capitalism/9781610395694/?lens=publicaffairs