Empty Promises Tracker

NCOSE downloaded this document and made additions. All prior entries (non-NCOSE entries) are in BLUE font*

A project of Issue One's Council for Responsible Social Media, the Empty Promises Tracker catalogs the history of public proclamations and policy changes announced by the largest technology companies that purported to protect users (including and especially minors), prioritize vulnerable communities, or safeguard the broader information ecosystem in which democracies operate. Each of these changes was announced publicly, only to be later retracted, significantly altered, marginalized, or never come to fruition. Many are half-truths or deflections that hide a different reality. We hope this document will inform lawmakers, advocates, and platform users as they seek to apply new responsibilities, standards, and oversight measures to these companies.

Note: Contributions to this tracker also came from the Institute for Strategic Dialogue and Design It For Us. In coming months, we will be building out this document to cover other areas of broken promises.

Meta: Instagram, Facebook, and WhatsApp

Additions Made by the National Center on Sexual Exploitation, last updated March 2024

Promise	Reality
Automated Harm Detection Tools: Over the last five years, Meta has increasingly shifted resources away from human detection of harmful content and toward automated systems. Under this process, engineers compile data sets of unacceptable content — such as terrorism, pornography, bullying, or "excessive gore" — and then	**2017** An internal email from 2017 notes that a Facebook executive **opposed scanning Messenger for "harmful content,"** because it would be a "competitive disadvantage vs other apps who might offer more privacy."

train machine-learning models to screen future content for similar material. Meta claims that these tools have been tremendously successful. The company's 2023 Community Standards Enforcement report determined that Meta's proactive detection technology removed 87.8% of bullying and harassment content, 99% of child exploitation content, and 95% of hate speech before users reported it. Meta spokespeople reiterated these figures to Insider, stating, "Using industry-leading technology, over 99% of child exploitation content we remove from Facebook and Instagram is found and taken down before it's reported to us."	**PRESENT DAY:** Meta continued their opposition to proactive scanning of messages up until they defaulted E2EE, consequently refusing to proactively scan direct messages on Messenger and Instagram Direct for grooming behavior or predatory adults interacting with minors for more than a decade. **Meta only scanned messages if they receive a report.** **2020** NM Attorney General Investigators also reported that a July 2020 document titled "Child Safety — State of Play (7/20)," **Meta referred to disappearing messages as one of the "immediate product vulnerabilities"** that could harm children, because of the **difficulty reporting disappearing videos and confirmation that safeguards available on Facebook were not always present on Instagram**." The same document the list of **"immediate product vulnerabilities"** also acknowledges "**livestreaming abuses**," negative impact on the ability to report content that disappears on Stories, and "interop" presumably referencing cross-platform conduct. The same 7/20 Child Safety - State of Play document further confirmed that safeguards on Facebook were not available on Instagram and "work is not planned to prevent FB and IG adults from initiating messages to IG minors," even though Meta was working to prevent contact between adults and unconnected "minors on FB." The document continues to say that efforts to "Block reach of unconnected adults to IG minors" was work that "has not been prioritized and more resources are needed." The NM State Attorney General also found that Meta's internal documents make clear that, despite all of these internal studies demonstrating harm from its platforms, Meta refused to devote sufficient resources in order to address the problems, despite its public statements to the contrary. **An August 2020 email cited a "severe lack of capacity for restricted content," including "nudity, graphic violence, child safety and SSI content review:"** **2021** The effectiveness of Facebook's automated tools were first called into question in 2021, when the Wall Street Journal's review of internal documents revealed that the company only removes a sliver of the posts that violate its own hate speech rules. **Multiple internal teams estimated the**

real figure to be lower than 5%.

In a blog post updating users about their child safety policies and measures, as well as in her testimony to US Congress in October 2021, Antigone Davis said "To understand how and why people share child exploitative content on Facebook and Instagram, we conducted an in-depth analysis of the illegal child exploitative content we reported to NCMEC in October and November of 2020. **We found that more than** 90 percent of this content was the same as or visually similar to previously reported content. **And copies of just six videos were responsible for more than half of the child exploitative content we reported in that time period.** However, one victim of this horrible crime is one too many."

- This is wildly deceptive considering that Meta's automated detection systems, which were newly introduced at the time and also rely on hashing, meaning that CSAM has to be first identified and entered into the hash database to be identified by AI detection systems such as those readily employed by Meta and 2. They only conducted this research over a span of 30 days, less than a year after their new hashing technology was deployed.

2023

Meta whistleblower Arturo Béjar built on these findings by uncovering two major flaws in the company's assessment of its own success.

1. **Meta's reported figures only apply to the content that the company ultimately removed, which is very different from the totality of violative content.** This is a major sleight of hand.
2. To grade its own homework, Meta used a measurement called prevalence: the percentage of content viewed worldwide that explicitly violates a Meta rule. This measurement found glaringly low rates of self-harm and eating disorders, bullying and harassment, and child exploitation content that were easily disproved by Béjar's survey of Instagram users and by countless other studies listed below in this document.

Bejar brought these revelations to the attention of Meta leadership, including Mark Zuckerberg, Chief Operating Officer Sheryl Sandberg, and Instagram head Adam Mosseri. Instead of acting on Bejar's findings, Meta ignored proposed design-focused fixes, shut down Bejar's research, and fired most of the team behind it.

Facebook's Messenger Kids: Facebook's Messenger Kids app was built with the promise that children wouldn't be able to talk to users who haven't been approved by their parents. CEO Mark Zuckerberg referred to the chat platform as "industry-leading work" and "better and safer than alternatives."	Despite Facebook's promises, a flaw in Messenger Kids allowed thousands of children to be in group chats with users who hadn't been approved by their parents. Facebook tried to quietly address the problem by closing violent group chats and notifying individual parents. The problems with Messenger Kids were only made public when they were covered by The Verge.
Underage Users: In order to comply with the Children's Online Privacy Protection Act, Meta's own Codes of Conduct prohibit users under the age of 13 from signing up for an Instagram or Facebook account. In his 2021 testimony before the Senate Commerce Committee, Instagram head Adam Mosseri reiterated, "If a child is under the age of 13, they are not permitted on Instagram."	**2021** In January 2021, Meta internally admitted to knowing that Instagram users under the age of 13 lie about their age to gain access to the platform. Meta then went on to acknowledged the danger of relying on a user's stated age, noting that "[s]tated age only identifies 47% of minors on [Instagram]. The majority of minors lie when stating their age." **Moreover, "99% of disabled groomers do not state their age . . . The lack of stated age is strongly correlated to IIC [inappropriate interactions with children] violators."** **Forty-five percent of US children aged 9 to 12 report using Facebook every day.** According to recent research, over a quarter of 9- to 12-year-olds report having experienced sexual solicitation online. **2022** According to the unsealed legal complaint brought by 33 state attorneys general against Meta, the **company has received more than 1.1 million reports of users under the age of 13 on its Instagram platform since early 2019 yet it "disabled only a fraction" of those accounts. Instagram, in particular, actively courted these users.** A study found that Instagram and Facebook made $801.1 million and $137.2 million, respectively, in ad revenue from users under 12 years old in 2022.
Bully, Harassment, and Hate Speech: Instagram's Codes of Conduct prohibit bullying and offensive comments, and the platform makes a strong showing of its commitment.	**2021** **Newly unredacted documents in New Mexico's lawsuit against Meta show that a 2021 internal Meta estimate found as many as 100,000 children every day received sexual harassment**. This finding came as the company "dragged its feet" on implementing new safeguards for minors and showed a "historical reluctance" to keep children safe, despite being aware of adults' ability to message minors or leave sexualizing comments on their photos or posts. The lawsuit alleges that

	Meta only began to take these threats seriously after an Apple executive's 12-year-old child was solicited on the platform, prompting concern that Apple would remove Meta's apps from its store. **2022** In 2022, the child safety company, Bark, analyzed more than 4.5 billion messages across texts, email, YouTube, and 30+ apps and social media platforms for its annual surveys of online harms to children. Instagram was the only platform to rank in the "top 5 worst" for every category of harm: severe sexual content, severe suicidal ideation, depression, body image concerns, severe bullying, hate speech, and severe violence. In a 2022 McAfee study of 11,687 parents and children in 10 countries, nearly 80% of respondents reported cyberbullying on Instagram, compared to 50% on TikTok and Snapchat. This has increased from 2017 numbers, which showed that Instagram (42%) followed by Facebook (37%) are the two platforms where people experienced cyberbullying the most. During his time at Meta, whistleblower Arturo Béjar conducted surveys of user experiences on Instagram which found that more than 1 in 4 users under age 16 witnessed "hostility against someone based on their race, religion or identity" in a single week.
Transparency: In a 2021 blog post, CEO Mark Zuckerberg rebuked claims that Meta was operating in secrecy by saying that the company had "established an industry-leading standard for transparency and reporting."	**2021** An internal study BEEF (June 27, 2021, to July 8, 2021) shed light on some concerning practices at Meta, particularly regarding Instagram's younger user base. Here are some key takeaways: ○ **High Rates of Negative Experiences Among Teens:** The study highlights that over half of all Instagram users experienced "bad experiences" within a week, with rates soaring among teenagers. **Specifically, 54.1% of users aged 13-15 and 57.3% of those aged 16-17 reported such experiences**. ○ **Concerning Incidents of Nudity and Unwanted Advances:** Alarmingly, **19.2% of users aged 13-15 encountered nudity, and 13% faced unwanted advances on Instagram in just one week**. The figure for unwanted advances doubles when considering a broader timeframe. ○ **Discrepancy in Reported Prevalence:** Meta's own Community Standards Enforcement Report starkly contrasts these findings. It reported much lower

	"prevalence" rates for issues like bullying, hate speech, and adult nudity/sexual activity, with figures ranging between 0.02% and 0.06% of views. **2022** Facebook did operate Crowdtangle, a leading data analytics and social monitoring tool that allowed academics, watchdog organizations, and journalists to identify harmful content on the platform, including CSAM. But in 2022, Facebook shut down Crowdtangle. It did so by quietly reassigning or removing team members, including the tool's former CEO and co-founder, Brandon Silverman. Facebook then stopped accepting any new user applicants, citing "staffing constraints," while the tool became buggy and broken for users who still had access.
Child Trafficking and Grooming: In response to a 2023 report by the Guardian, a Meta spokesperson said, "The exploitation of children is a horrific crime – we don't allow it and we work aggressively to fight it on and off our platforms."	**2017** According to the NM Stat Attorney's lawsuit against Meta, in 2017, Meta's subcontracted content moderators spoke out about their experiences and the type of content that Meta not only hosted but that supervisors permitted to stay on their platforms. One moderator said that she "**always saw cases of adults grooming children and then making plans to meet them for sex, as well as discussions about payment in exchange for sex.**" Reports often **took months to resolve and would often end in an automated email saying the content didn't violate their policies**. Another moderator described a similar experience, "On one post I reviewed, there was a picture of this girl that looked about 12, wearing the smallest lingerie you could imagine. . . . I**t listed prices for different things explicitly, like, a blowjob is this much.** It was obvious that it was trafficking." **Her supervisor said no further action was needed.** **2019** In 2019, the National Society for the Prevention of Cruelty to Children found that Instagram was the #1 platform for child grooming in the UK; they identified more than 5,000 crimes of sexual communication with children and a 200% increase in how Instagram was used to abuse children, all in an 18 month period (Forbes, 2019). A 2019 investigation by the NSPCC, a leading UK-based children's advocacy organization, determined that Instagram had become the leading platform for child grooming in the country.

The research was based on freedom of information requests covering an 18-month period.

2020

According to a 2020 report by the Human Trafficking Institute (HTI), Facebook was the platform most used to groom and recruit children by sex traffickers (65%), based on an analysis of 105 federal child sex trafficking cases that year. The HTI analysis ranked Instagram second most prevalent (14%).

Detailed in the NM State attorney's lawsuit against Meta, in July 2020 an internal Meta chat revealed that one employee asked "**What specifically are we doing for child grooming** (something I just heard about that is happening a lot on TikTok)?" He received a response—"somewhere between zero and negligible"

2021

A Child Safety Presentation from March 2021 reported on the "IG-specific challenges," confirming that inappropriate interactions with children, which the presentation refers to as **"IIC, a.k.a. 'grooming'" exists on Meta's platforms.** It continued to say that that Meta "underinvested in minor sexualization on IG, notable on sexualized comments on content posted by minors. Not only is this a terrible experience for creators and bystanders, it's also a vector for bad actors to identify and connect with one another."

Another 2021 internal document, reflected Meta's awareness that its recommendation algorithm, **People You May Know, or PYMK, had a direct link to trafficking**: In a string of comments, under the heading **"IIC/Grooming," a Facebook employee wrote: "in the past, PYMK contributed up to 75% of all inappropriate adult-minor contact,"** prompting another employee to pose the question: **"How on earth have we not just turned off PYMK between adults and children? . . . It's really, really upsetting."**

In 2021, a Facebook whistleblower reported to the SEC that Facebook has never devoted "adequate assets" to addressing CSAM. According to this former employee, Facebook broke up and redeployed an internal team that was supposed to develop software which could detect indecent videos of children because this project was seen as "too complex".

	2023 The Guardian's 2023 investigation confirmed that Facebook and Instagram were still operating as major marketplaces for child sex trafficking. According to the Guardian, "many of those we interviewed said they felt powerless to get the company to act."
Harassment, Predation, and Sexually Explicit Content: Meta explicitly prohibits material that sexually exploits or endangers children, including any transactions or content that involves trafficking, coercion, sexually explicit language, and non-consensual acts.	**2011** According to the Tech Transparency Report, Mexican human rights activists and researchers **discovered 1,400 Facebook profiles linked to alleged child predators who were taking and selling images of infants to prepubescent children ages 7-10**. **After reporting these profiles, only a few were removed, the researchers' personal Facebook profiles were removed**, and when they asked Facebook about this issue at a public event in New York in late 2011, a spokesperson said they [Facebook] don't disable accounts "simply because people are discussing controversial topics." (Tech Transparency Report, 2019) **2016** In February 2016, the BBC reported on Facebook groups used by pedophiles to exchange photos and videos of children. Undercover journalists joined these groups, reported 20 images using Facebook's reporting system, but the company removed only four. In 2016, a DOJ working group on child exploitation citing the results of a survey of over 1,000 investigators across federal, state, and local government, reported that Facebook was "the most commonly mentioned platforms...being used by offenders to contact children for sexual purposes." **2017** The BBC conducted a follow-up investigation after Facebook told reporters they improved their reporting and detection systems. **Out of the 100 images reported 18 were removed. Researchers**

	also reported profiles of five convicted pedophiles (Facebook prohibits sex offenders from having accounts) none of which were removed. **2018** Facebook launched a survey, they later called a "mistake" asking **how users would handle grooming behavior on Facebook** and if they thought users or Facebook should decide the rules for whether **adult man should be allowed to ask for sexually explicit images (CSAM was still illegal then) from "14-year-old girls" on Facebook**. **2019** In December 2019, **Meta employees circulated a Sunday Times article (detailing NCOSE's joint campaign #wakeupinstagram) that focused on Instagram's algorithm recommending objectionable content.** One individual in the article "specifically calls for [Instagram] to stop recommending children's accounts to anyone. They have a case study of a parent who set up an account to showcase her daughter at gymnastics and was horrified that this was then potentially promoted to other people." The email [internally] further notes the existence of Instagram users under the age of 13 and "[t]he ease of which a stranger can direct message a child on Instagram and the lack of proactive protections in place." **2020** A September 2020 internal analysis revealed that the **prevalence of inappropriate sexual communication ("sex talk") directed at minors is 38 times higher on Instagram Direct compared to Facebook Messenger in the U.S.** • **Over one-quarter (28%) of minors on Instagram received message requests from adults they did not follow**, indicating a significant risk of unwanted or potentially harmful contact.

- **Internal reports characterized Instagram's child safety protections as "minimal," with policies regarding minor sexualization described as "immature**," and a minimal focus on issues like trafficking.

A September 2020 email reveals Meta's pervasive problems with respect to CSAM and other sexually explicit content. "[W]hen you search for these terms, there are no results under the 'hashtag' tab, but there are endless results under the 'Top Accounts' and 'Accounts" tab, and almost all are violating."

2021

An internal presentation from 2021 estimated that **100,000 children per day received sexually explicit content like photos of adult genitalia.**

2022

Major layoffs to the Trust & Safety teams dedicated to combating CSAM and child trafficking, among other integrity issues, have compounded these problems. **Since November 2022, Meta has laid off around 21,000 people, or 25 percent of its workforce.**

A 2022 study by the National Center On Sexual Exploitation found that 22% of minors who used Instagram reported experiencing a sexually explicit interaction.

Arturo Béjar's disclosed user surveys found that 13% of Instagram users had experienced unwanted sexual advances in the past seven days.

A 2022 study published by Thorn, a leading resource on online child exploitation, found that Instagram tied with Kik and Tumblr as the platform where minors reported the second highest rates of online sexual interactions with people they thought were adults.

2023

According to 2023 investigations by the Wall Street Journal and researchers at Stanford University and the University of Massachusetts Amherst, **Instagram's recommendation system and hashtags help promote a vast network of pedophiles and guide them to content sellers.** Since receiving

	inquiries from the Journal, the platform said it has blocked thousands of hashtags that sexualize children, some with millions of posts, and restricted its systems from recommending users search for terms known to be associated with sex abuse. On March 8, 2023, the Center for Countering Digital Hate published a report about bullying, sexual harassment of minors, and harmful content on Horizon Worlds. As of April 2023, Meta made Horizon Worlds available to young users between the ages of 13 to 17. The UK Children's Commissioner report, released in January 2023, found that 33% of children who had seen pornography saw it on Instagram.
The Mental Health of Teen Girls: Meta has long purported to value the mental health of its young users, including and especially teenage girls. In a 2021 blog post, Zuckerberg wrote that in "serious areas like loneliness, anxiety, sadness, and eating issues -- more teenage girls who said they struggled with that issue also said Instagram made those difficult times better rather than worse."	**2021** The disclosures revealed by Meta whistleblower and Council for Responsible Social Media member Frances Haugen painted a very different picture of the effect Meta employees knew their platforms were having on teenage girls. **Internal researchers at Meta warned that Instagram's monetization of "face and body," the pressure to look a certain way, and an algorithmic feed that encourages constant engagement "exacerbate each other to create a perfect storm," for teenage girls.** The study linked this trifecta to eating disorders, body dysmorphia and dissatisfaction, loneliness, and depression, including the alarming finding that **32% of U.K. teen girls said that Instagram made them feel worse about their bodies**. Two years later, many of these warnings were backed up by another Meta whistleblower, Arturo Béjar, and the survey of user experiences he led.
Suicide and Self-Harm: Meta has clearly stated that it removes content that depicts or encourages suicide or self-injury, including graphic imagery and real-time depictions. This includes promising to place a sensitivity screen over content that doesn't violate its policies but may still be upsetting to some users.	**2021** **Instagram's internal research found that "13% of UK teenagers and 6% of US users" traced a desire to kill themselves back to Instagram.** The BBC found that Instagram "removed almost 80% less graphic content about suicide and self-harm" during the height of the COVID-19 pandemic. Despite these findings in 2020 and 2021, it was not until 2024 that Meta announced they would hide posts about suicide and eating disorders from all teens' Instagram and Facebook feeds, even if it is shared by an account they follow.

Parental Burden: Amid criticisms of its platforms, Meta has rolled out some 30 parental controls to manage who their kids can talk to or how much time they spend on Facebook and Instagram.	Most of the parental controls require both the parent and the minor to opt-in. While parents can supervise some of their teen's activities and time spent on the app, they have no ability to limit the time spent on the apps.
User Reporting Mechanism: Both Instagram and Facebook promise to allow users to report harmful or upsetting content.	After 2019, **internal Meta documents show that the company added steps to the reporting process to discourage users from filing reports.** And while users could still flag things that upset them, Meta shifted resources away from reviewing them. Meta said the changes were meant to discourage frivolous reports and educate users about platform rules.