Tennessee v. Meta
Case No. 23-1364-IV
PX-06506

TEEN ACCOUNTS, BROKEN PROMISES

HOW INSTAGRAM IS FAILING TO PROTECT MINORS



REPORT BY

ARTURO BÉJAR

CYBERSECURITY FOR DEMOCRACY

fairplay childhood beyond brands

MOLLY ROSE FOUNDATION

ParentsSOS PARENTS FOR SAFE ONLINE SPACES

WITH SUPPORT FROM

HEAT INITIATIVE

TABLE OF CONTENTS

4 FOREWORD

6 INTRODUCTION

11 A REPORT ABOUT BROKEN PROMISES AND DANGEROUS DESIGN, NOT CONTENT

13 METHODOLOGY FOR TESTING USER-FACING SAFETY FEATURES ON INSTAGRAM

18 HOW TO BUILD EFFECTIVE SAFETY AND USER REPORTING FEATURES FOR CHILDREN

22 OUR FINDINGS

24 Inappropriate Contact and Conduct

25 Meta's Broken Promises: Improper Contact and Conduct
26 What Meta Promised
27 Key Findings
31 Recommendations for Meta
32 Questions for Regulatory Inquiry

33 Sensitive Content

34 Meta's Broken Promises: Sensitive Content
35 What Meta Promised
37 Key Findings
41 Recommendations for Meta
42 Questions for Regulatory Inquiry

43 Time Spent and Compulsive Use

44 Meta's Broken Promises: Time Spent and Compulsive Use
45 What Meta Promised
46 Key Findings
48 Recommendations for Meta
49 Questions for Regulatory Inquiry

50 Age Verification, Minors and Sexualized Content

51 Meta's Broken Promises: Age Verification, Minors and Sexualized Content
52 What Meta Promised
53 Key Findings
56 Recommendations for Meta
57 Questions for Regulatory Inquiry

58 CONCLUSIONS

61 APPENDIX

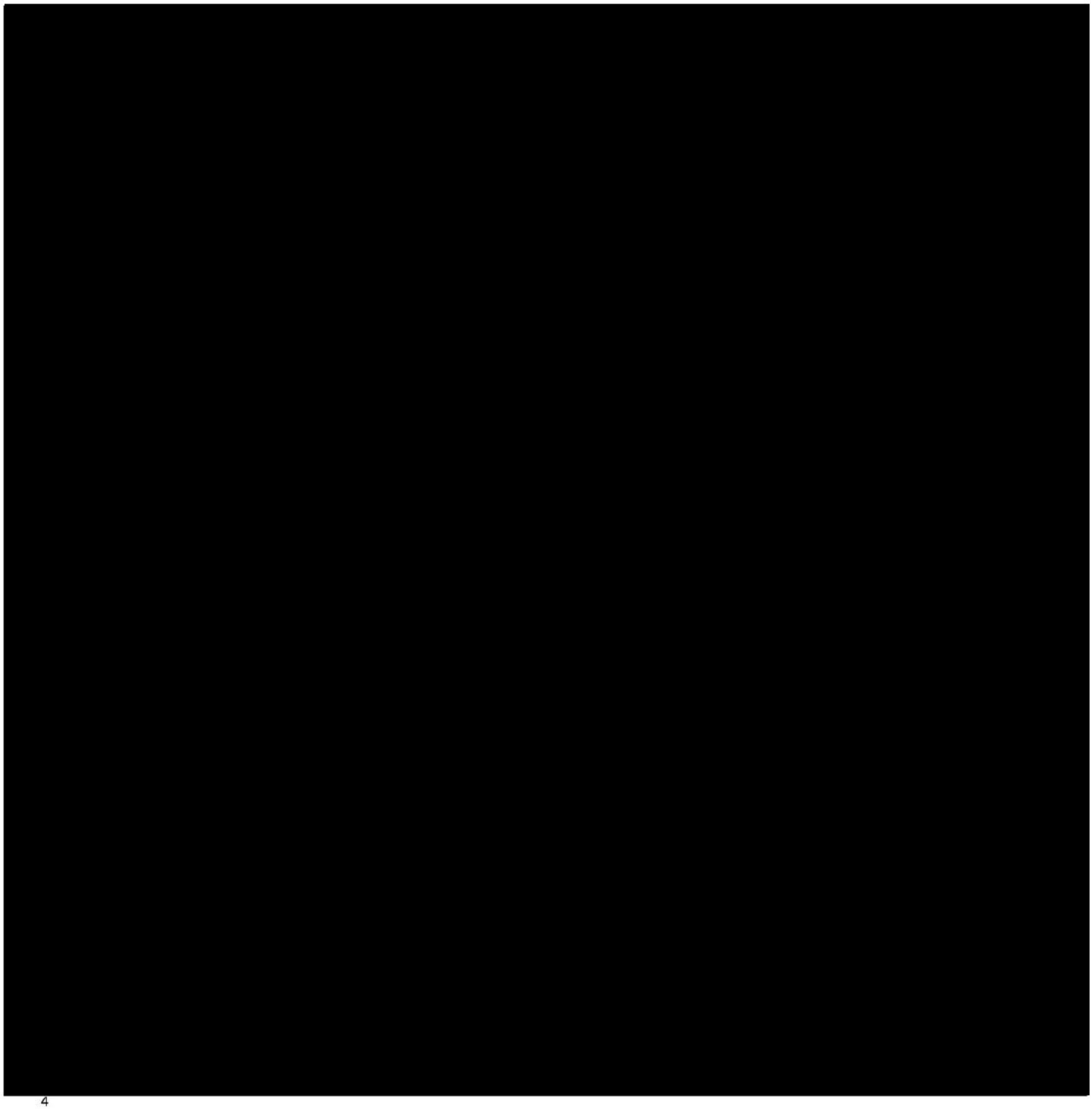
62 Appendix 1: Summary of Detailed Findings

72 Appendix 2: Scoring Rubric Applied During Safety Testing

73 Appendix 3: Detailed Findings

76 ABOUT THE AUTHORS

FOREWORD



It [redacted]. Thanks to a slew of whistleblowers, lawsuits, academic research, and regulatory inquiries, we know a lot more about Instagram and Meta. [redacted] We understand that courageous individuals within Meta tried on many occasions to sound the alarm on how Instagram's design was contributing to mental health harms for teens, but were rebuffed by senior leadership. We understand there are real steps that companies like Meta could take now to make their platforms safer and less addictive. Steps that would save lives.

[redacted]. As we have demonstrated, Meta's new safety measures are woefully ineffective. If anyone still believes that the company will willingly change on its own and prioritize youth well-being over engagement and profits, we hope this report will put that to rest once and for all.

Time and time again, Meta has proven they simply cannot be trusted. [redacted] edies, we need real regulation. In the US, that means passing new legislation like the Kids Online Safety Act, which would require social media companies to prevent and mitigate the harms to young people caused by platform design. In the UK, that means strengthening the existing Online Safety Act to compel companies to systematically reduce the harm their platforms cause by compelling their services to be safe by design.

[redacted]

INTRODUCTIONS



In September 2024, Meta announced that it was introducing Instagram Teen Accounts and that all teenagers would automatically be enrolled in the new program. According to the announcement, the launch was meant to "reassure parents that teens are having safe experiences." Added Meta: "This new experience is designed to better support parents, and give them peace of mind that their teens are safe with the right protections in place."

Meta's announcement came right before an important hearing in the U.S. House of Representatives on the Kids Online Safety Act (KOSA). This timing was almost certainly not coincidental. KOSA is the most important legislation of its kind to make significant progress in Congress in over 25 years. Meta has lobbied extensively to stop KOSA — and to influence in its favor legislation in the United Kingdom, European Union, and elsewhere that would hold the company responsible for its design choices that harm teens.



to make changes to make the platform safer. While the ostensible audience for Meta's announcement was parents who were increasingly concerned about how much time their children spent on Instagram and what they experienced there, Meta also was seeking to reassure lawmakers that the company was addressing those concerns and did not need to be regulated.

According to Meta, Teen Accounts have a number of safety features that differ from regular adult accounts: They are private by default, teens need to accept any new followers, and anyone who doesn't follow them can't see the teen's content or interact with their accounts. Teen accounts are also given the strictest messaging settings and the most restrictive Sensitive Content Controls, Meta says. In addition, if teens are on Instagram for 60 minutes in a given day, they will receive a notification telling them to leave the app. According to Meta, with Instagram's "sleep mode," teens' notifications will be muted, and their direct messages will receive autoreplies, between 10 pm and 7 am.

The launch of Instagram Teen Accounts promoted not just new safety features for teens but a number of new promises for parents. According to Meta, teen account holders under the age of 16 need a parent's permission to change the settings of any of the built-in protections for Teen Accounts. If your teenager is over 16, you can simply turn on parental supervision, which allows you to approve or deny their requests to change their account settings.

According to Meta, the supervision feature also gives parents other ways to get involved with their child's experience on Instagram. This includes parents being able to see whom their teen has messaged in the past seven days; set daily time limits for Instagram usage; block teens' access to Instagram for a specific time period; and see what topics their teens are looking at.

To parents and other interested observers who did not rigorously validate these claims, it would appear from these statements that Meta had fully addressed these hazards and children were again safe. But children are not safe on Instagram.

We undertook a comprehensive review of Meta's Teen Accounts and all of the safety tools in Instagram listed on Meta's website. Many of these safety tools preceded the announcement of Teen Accounts, while others were introduced when Teen Accounts launched. Our review included both these longstanding tools, which have been aggregated under Teen Accounts, as well as the new tools.

Meta's list includes 53 entries about Instagram's safety features. [redacted] given that it includes tools that have been discontinued (for example "Take a Break"), or fundamentally changed so as to not serve their original purpose (hiding view counts on posts). Some of the announcements are ostensibly improvements on existing safety tools. Our comprehensive reviews of Teen Accounts and safety tools included testing 47 of Instagram's 53 listed safety features (the reasons are outlined in the Research Note by Cybersecurity for Democracy).

47

SAFETY FEATURES TESTED



Using a three-tier framework — red, yellow, and green – our researchers systematically evaluated each of the 47 Instagram safety features. Each safety feature was assessed according to its effectiveness, usability, and visibility. **We rated 64% of the safety tools as "red" (30 tools) because they were either no longer available or ineffective. Another 19% of safety tools (9 tools) reduced harm, but came with notable limitations. Only 17% of the safety features (8 tools) worked as advertised, with no limitations.** Many of the safety tools that were ineffective are the foundation of Teen Accounts, including Sensitive Content Controls, inappropriate contact safety, and tools for kids to manage the time they spend on the platform. The minority of tools that worked address limited use cases and hazards. Given that several ineffective tools were announced years ago, we cannot estimate the harm that teens have experienced as a result. These findings are discussed at length in the report and our "red, yellow, green" rubric is discussed at length in the appendix.

We hope this report serves as a wake-up call to parents who may think recent high-profile safety announcements from Meta mean that children are safe on Instagram. [redacted] But we also urge regulators and lawmakers to consider the substantial evidence that the majority of Meta's safety initiatives have been little more than PR efforts. We cannot waste any more time, or allow more children to be harmed, by Meta's self-regulation.

In the United States, those interventions include passing KOSA, which would create a duty of care for social media companies to ensure that the design of their products is not contributing to serious harms for minors, including addictive use of platforms. It also means that the Federal Trade Commission and state attorneys general hold Meta accountable for violating the Children's Online Privacy Protection Act and Section V of the FTC Act.

In the UK, the Government should strengthen the Online Safety Act to ensure that regulation is more effectively focused on achieving measurable harm reduction, alongside other structural remedies that will put the onus more directly on platforms to identify and take effective steps to mitigate reasonably foreseeable harm. The regulator, Ofcom, must also become bolder and more assertive in enforcing its regulatory scheme.

Note: Cybersecurity for Democracy did not participate in writing this section and by policy does not endorse any legislation.

RESEARCH NOTE FROM CYBERSECURITY FOR DEMOCRACY

For many years, Meta and other companies have responded to growing concerns about youth safety by rolling out new user-facing safety tools. These announcements often arrive at moments of public scrutiny or looming regulation. As researchers, we had a simple question: **Do these tools actually work?**

To answer it, we borrowed from established practices in cybersecurity. Our team partnered with Arturo Béjar to apply "red team" style scenario testing to user safety tools on Instagram, taking a well understood security methodology into a new domain. To do that, we systematically identified every announced Instagram safety feature, designed controlled test scenarios for each to reflect real teen, parent, and adversary behaviors, and ran those scenarios with realistically configured test accounts. We also developed a taxonomy of the key dimensions of user-facing safety tools, to allow us to analyze these tools and features in a more systematic way.

In March 2025 Arturo Béjar undertook an initial round of testing, and in June and July 2025, we worked with Arturo to independently perform this scenario testing, performing a comprehensive review of Meta's Teen Accounts and all 53 descriptions of Instagram safety tools listed on Meta's website under "Our tools, features, and resources to help support teens and parents". It's important to note that this is a list of press releases, not a list of currently active tools, however. For example, several of the features listed have since been discontinued, a fact which isn't noted on this page. Additionally, several announcements describe changes to existing safety tools, rather than separate tools.

Ultimately, our scenario testing of Teen Accounts and safety tools included 47 of these 53 listed items. Two safety features were not analyzed for methodological reasons, and four others on the list did not primarily relate to safety objectives and so were not included in our analysis.

More research into social media user safety tools is urgently needed. Our findings show that many protections are ineffective, easy to circumvent, or have been quietly abandoned. User safety tools can be so much better than they are, and Meta's users deserve a better, safer product than Meta is currently delivering to them. Yet rigorous testing of user safety tools can tell us not only what is broken, but also point the way to solutions. The same methods that reveal failures can also show us the way forward: toward safety tools that are default-on, resilient against evasion, and genuinely useful for teens and parents.

Going forward, we believe independent scenario testing should become a standard practice, carried out not just by researchers but also by regulators and civil society to answer questions about platform functionality. Treating safety tools with the same rigor that cybersecurity applies to other critical technologies is the only way to know whether platforms are keeping their promises — and to ensure that future tools are designed to deliver real protection, which is verified through independent testing, as are most safety features in society today.

A REPORT ABOUT BROKEN PROMISES AND DANGEROUS DESIGN, NOT CONTENT

This report aims to assess safety features relative to Meta's promises to parents and regulators. Our focus is on product design, not on content or how it is moderated. Even when we discuss content-related features, such as Sensitive Content Controls, our focus is on the effectiveness of the promised feature, rather than the content itself.

This distinction is critical because social media platforms and their defenders often conflate efforts to improve platform design with censorship. However, assessing safety tools, and calling out Meta when these tools do not work as promised, has nothing to do with free speech. [redacted]

Recommendation-based features like Home, Reels, Discover, and Search should be fundamentally safe and age-appropriate by design. When Meta recommends content to a young person on these product features, typically through its personalized recommender algorithms, the choices that inform these recommendations are a product design issue. Meta has promised to ensure teens are "seeing content that's appropriate for their age." Meta should keep that promise, and it's entirely possible for the company to do so without limiting the ability of adults to share sensitive content with other adults.

Our testing of Teen Accounts and Meta's safety tools found a combination of tools that were no longer functional, tools that were buggy, and tools that by their own design would not prevent the harm they claimed to address. In several cases, we found that Meta's own design circumvented its own safety tools. These are all product design issues.

The lack of effective Time Spent tools to deal with issues including problematic time spent on Instagram or usage late at night that may interfere with sleep is a product design issue.

The delivery of rabbit holes of self-harm, suicide-related accounts, and Meta's search features recommending these kinds of content and accounts even when they are not what the teen is searching for, is fundamentally a product design issue.

Meta measures everything it does. The company knows how many users it has, how much time they spend on its products, and how often they interact with every one of its product features.

In contrast, Meta measures its safety tools in terms of the sheer number of tools the company has rolled out. Meta should measure and share how effective the tools are, the extent to which teenagers are adequately protected when the tools are in force, and whether specific tools are particularly helpful to address the safety and well-being risks that children face when using Instagram.

Meta could, of course, easily measure and publicly report on the impact and efficacy of its Teen Accounts measures if it wanted to. Questions may legitimately be asked about why the company consistently chooses not to do so.

METHODOLOGY FOR TESTING USER-FACING SAFETY FEATURES ON INSTAGRAM

In March 2025 and again in June and July 2025, we assessed Instagram's safety tools for teens using testing scenarios featuring test accounts.

Testing scenarios are an approach widely used by engineering and security professionals to understand and evaluate how a system behaves. This kind of testing is also called "red team" or "black-" or "gray-box" testing, depending on test conditions. Conceptually, it is similar to crash testing a car, where a tester sees how a car's crash protections actually behave under different controlled conditions. This kind of scenario testing is a standard systems engineering security practice.

A secondary goal of this report is to demonstrate the utility of scenario testing for understanding the efficacy of user-facing safety systems and user experiences more generally. We encourage academic institutions, independent organizations, and regulators to develop their own avatar test scenarios of all social media products and their safety features, and we encourage independent security auditors to develop a series of test scenarios to evaluate the efficacy of safety tools. We believe that testing safety features is essential, as these are products used by hundreds of millions of teenagers across the world. As such, independent red team testing should be a core element of their development process.

STEP 1

FEATURE IDENTIFICATION

We began by systematically reviewing Meta's public safety press releases to identify user-visible features implemented on Instagram. Each press release was examined for references to tools, interface elements, or settings that were claimed or implied to directly affect end-user safety. To ensure comprehensiveness, we included both new feature launches and updates to existing safety mechanisms. Features were included if they (a) were described as safety-related either explicitly or implicitly, (b) were accessible to end-users through the Instagram interface, and (c) had measurable or observable functionality that could be tested empirically.

STEP 2

TEST SCENARIO DEVELOPMENT

For each identified feature, we developed a structured testing scenario designed to simulate realistic user behavior. Scenarios specified the conditions under which the feature should activate, the type of account used (e.g. teen vs. adult, follower vs. non-follower), and the expected safety intervention. The design of scenarios was guided by threat-modeling principles, with a focus on three perspectives: (1) usage, search, and commenting without seeking to intentionally circumvent protective measures, (2) self-directed circumvention attempts, and (3) external adversaries seeking to bypass protective measures. This step ensured that tests reflected plausible use cases rather than artificial or contrived interactions.

STEP 3

TEST ACCOUNT CONFIGURATION

To accurately evaluate the behavior of safety features, we established a set of controlled test accounts or "avatars." Accounts were configured to reflect the adversary models defined in our framework:

Teen user model: Accounts were created with an age designation under 18 and configured with exploratory but non-malicious use patterns to simulate a teen attempting to circumvent restrictions placed on their own account. The account creation process mirrored the process of a parent giving their teen a new phone: Accounts were created using all privacy and safety defaults. To ensure testing captured regular conditions, the accounts were tested right after creation, and then tested after two weeks.

Supervising user model: Accounts were created as a user with a parental, supervisory relationship to the teen user model.

Targeting user model: Adult and teen accounts were established to simulate malicious actors attempting to interact with or contact teen users, with follower/non-follower and public/private account variations included to capture different relationship dynamics.

These controlled accounts provided a consistent and repeatable basis for executing the testing scenarios. By aligning account characteristics directly with user models, we ensured that observed outcomes could be clearly interpreted within the threat-model framework.

STEP 4

FEATURE TESTING

Initial testing was performed in March 2025, with a second round of follow-up testing in June and July 2025. Testing for all scenarios was done on iPhones installed with iOS with versions 16.6.1 (March) and 16.7.8 (June and July) and the most recent version of Instagram available at the time of testing. Each testing scenario was executed using the test accounts relevant to the test scenario. During testing, we carefully observed whether the safety feature functioned as described in Meta's public materials. All findings were documented through contemporaneous screen recordings and screenshots, which provide verifiable evidence of the product's behavior. Images and screen captures were not altered in any way. Where relevant, repeated trials were conducted to confirm consistency of outcomes.

STEP 5

EVALUATION

During testing, each feature or tool was evaluated to determine if it was currently functioning as described in Meta's public materials and if it was resistant to circumvention by accidental or trivial efforts. Additionally, each tool or feature was classified along five dimensions: the user target, the harm approach, the safety scope, the risk category, and the implementation style. Definitions and rubrics for each of these dimensions are in Appendix 3.

After testing, each tool or feature was graded using a three-tier rubric — red, yellow, and green — to classify the overall effectiveness and usability of safety features visible to users on Instagram. A red rating was given when a safety feature was found to be no longer available or, in a realistic testing scenario, was trivially easy to circumvent or evade with less than three minutes of effort. A yellow rating was made when a safety feature was functional and offered some level of protection but came with one or more serious limitations along a classified dimension. For example, if a tool was functional, but only reduced instead of prevented harm (see Harm Approach in Appendix 3), it would receive a yellow rating. A green rating was given when a safety tool worked effectively and as described. The full rubric used in the scoring process is described in Appendix 2.

LIMITATIONS

Please note this report makes no measurements of the frequency of harmful experiences or content being shown through algorithmic feeds. For example, we make no claims about the frequency of harmful content being algorithmically targeted to teens through Reels or Discover, or how often teens message with adults they would not know if not for Instagram's recommendations of whom to follow.

The focus of this report is to better understand the effectiveness of the safeguards that Meta claims to have put in place, rather than to measure the broader prevalence of harm on its platforms. Our goal was to better understand the effectiveness of the safeguards that Meta claims to have put in place relative to the kinds of harms that young people and parents are concerned about, and to test whether the Teen Account safeguards adequately prevent exposure to them.

Scenario testing is a well-established methodology for understanding foreseeable risks and to assess the efficacy of user-facing systems and tools. However, it cannot tell us about how frequently risks are exploited in practice. The frequency with which teens experience different kinds of harm can be determined by large-scale surveys of users, such as Meta's own Bad Experiences and Encounters Framework (BEEF) survey. Independent academic efforts like the USC Neely Social Media Index can measure this as well. We encourage Meta to undertake a comprehensive analysis of the frequency of exploitation of the risks we identify and circumvention of user safety tools with critical vulnerabilities, and to release the questions, methodology, anonymized data, and results publicly so that parents, regulators, and the general public can decide for themselves how safe Instagram is for children and teens.

HOW TO BUILD EFFECTIVE SAFETY AND USER REPORTING FEATURES FOR CHILDREN

Before detailing our findings of Meta's existing safety features, it is important to describe how social media platforms can achieve truly impactful safety-by-design by building safety measures and reporting mechanisms that are thoughtfully designed, built in an age-appropriate fashion, and driven by an overarching emphasis on achieving harm reduction.

SAFETY TOOLS

Truly effective safety tools should have four primary attributes: prevention, protection, resiliency, and ease of use.

1. **Prevention:** Safety tools must strive to effectively prevent teens from being exposed to harm in the first place.
2. **Protection:** If and when harm does take place, an effective safety tool should immediately provide help and support — for example, by allowing a teen to easily indicate when they have experienced unwanted sexual advances. The tool should automatically block users to prevent a further recurrence of harm, and should proactively record information that will help find criminals and other bad actors. Capturing this information also supports the development of stronger protections for other teens.
3. **Resiliency:** Safety tools must be resilient to active manipulation. If there is an easy workaround that a teen can figure out, the feature is essentially a safety tool in name only.
4. **Ease of use:** Safety tools should either be switched on by default or be used with a single click. Meta knows that default settings and ease of use directly determine the likelihood that a user will go on to use a feature. That is precisely why features that actively increase engagement are usually turned on by default, or users are proactively incentivized to turn them on. Safety tools should be designed in exactly the same way.

A good way to think about Safety Tools is to think of them like safety features in a car. It doesn't matter if a car has 50 airbags if they don't effectively protect the people inside when an accident happens.

Under robust regulatory oversight, safety features built into a car are independently tested to make sure the safety mechanisms work as intended. Likewise, Meta's safety features need to be resilient under duress, and the efficacy of every measure should be robustly and independently evaluated to ensure they offer the highest possible standards of protection to children.

EFFECTIVE REPORTING

It's easy to tell when a social media platform wants you to use one of its features. A platform will typically turn the feature on by default or proactively encourage you to turn it on, for example, through in-app prompts. In contrast, platforms typically seek to frustrate users from turning on features and settings that may not be in their commercial interests.

In such cases, platforms may deliberately choose to make certain settings and features hard to set up, confusing, or difficult to find. Over recent years, extensive research has catalogued the way Meta and other companies use these techniques, relying on so-called "dark patterns" and "friction" in the user experience to make it harder, and therefore less likely, for users to adopt certain features.

Meta's reporting tools for Instagram are an excellent example of intentional friction-by-design. Many of Meta's safety features, including some of the features analyzed in this report, typically require many steps to open and lodge a report. Users may be asked to go through multiple fields, or even to leave the app they are using to make their complaint.

For products used by teens, it is critical to have effective reporting. There are a number of criteria that reporting tools must meet in order to be helpful for teens.

REPORTING TOOLS SHOULD

- Be easy and rewarding to use: The user should feel that the tool helped them with the issue that they were experiencing.
- Use language that teens relate to the harm they are experiencing.
- Capture what happened (harm), where it happened (context), and how bad it was (i.e., intensity and severity). These steps should not be required, but they should be available and rewarding to use.
- Provide immediate support and protection to the user, independent of any content moderation considerations.

Providing immediate relief and support at the time a teen asks for help is critical to reducing harm in cases of harassment, bullying, grooming, self-harm, and other areas. In these cases, the messages sent or comments made are likely not going to be found to violate rules about content. In reviewing comments and content reported for bullying, more than 90% did not violate any policies, and in 50% of the most severe cases of bullying, the content looked benign or positive to the reviewer, who did not have the context.

Meta also continues to use language that is age-inappropriate and that the company understands may be actively likely to deter young people from making reports. Meta understood as early as 2012 that labeling tools as "report" had a significant negative impact on young users' confidence and willingness to share bad experiences, often because they worry that they, or the other person, may get in trouble. At the time, Meta also found the importance of using language that matches the teen's experience. Without the correct language, the majority of teens would not submit reports, even though they were having harmful experiences.

Meta continues to design its Instagram reporting features in ways that will not promote real world adoption. As a result, it was estimated by Meta that less than 1% of users report harmful experiences, and only 2% of those who submit reports get help. In other words, only two out of 10,000 people who have a harmful experience on Instagram actually get help from the platform.

We call on Meta to provide much needed transparency about the rate at which teens use the reporting tools relative to the harms they experience, the action rates (i.e. if a teen submits a report, what is the likelihood that it will be acted on), and whether the tool helped the teen with the issue they were experiencing.

0.02%

OF PEOPLE WHO HAVE HARMFUL EXPERIENCES ON INSTAGRAM GET HELP FROM THE PLATFORM

OUR FINDINGS

30 OF THE 47

SAFETY FEATURES WERE EITHER NO LONGER AVAILABLE OR WERE SUBSTANTIALLY INEFFECTIVE

Our analysis of 47 safety tools for teens found that the overwhelming majority were woefully ineffective, with over 60% receiving our worst red rating.

30 of the 47 safety features were either no longer available or were substantially ineffective. As a result, these safety features received a red rating. Nine of these safety features could not be triggered during our research and appear to have been discontinued. We also found that a further 20 safety measures could either be trivially circumvented or evaded, whether accidentally or with less than three minutes of effort.

9 of the 47 safety features offered some level of functionality, but came with notable limitations or flaws. As a result, these safety features received a yellow rating. Safety features were classified as yellow if they were not enabled by default and required the user to take steps to proactively find, activate, use, or configure them; or if they reduced harm rather than effectively preventing it.

8 of the 47 safety features analyzed were found to be fully functional and offered proactively or as a default. These measures received a green rating. The green rating was given to safety features that were wholly and demonstrably functional; offered proactively so that users didn't have to locate and set them up; and were found to be capable of improving user safety at both the individual and community levels.

Meta's promises around safety features and Teen Accounts are clustered around four main areas: Inappropriate Contact and Conduct, Sensitive Content, Time Spent and Compulsive Use, and Age Verification.

In the following sections, we explore and test Meta's safety features for each of these categories. In our analysis we:

- Detail the promises Meta has made to parents and regulators.
- Provide an overview of the findings from our testing.
- Outline a series of questions that regulators with discovery powers could pose to Meta about its safety tools and their efficacy.
- Provide recommendations for how each of these areas could be addressed by Meta and similar companies.

Our recommendations come from an understanding of the company's capabilities and the specific safety measures that would be straightforward to implement. All of the issues found in this report can be addressed by Meta — it has the technology and people to develop features that would effectively reduce the harms experienced by young people on its platforms.

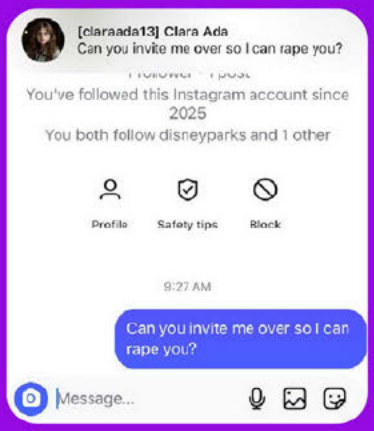
INAPPROPRIATE CONTACT AND CONDUCT

Inappropriate Contact and Conduct includes unwanted sexual advances, bullying and harassment, contact with strangers, etc.

Adults can communicate with minors through many features that are inherent in Instagram's design. In many of these cases, the adult strangers were recommended to the minor by Instagram's features: Reels, People to Follow, etc.

Most significantly when a minor experiences unwanted sexual advances or inappropriate contact, Meta's own product design inexplicably does not include any effective way for the teen to let the company know of the unwanted advance. [redacted] and means it is effectively impossible to manage or reduce this harm.

In another troubling design choice, Meta has implemented an animated reward or incentive for underage users to activate Disappearing Messages, and rewards their use with an animation. Disappearing Messages can be used for grooming, drug sales, etc., and leave the minor account with no recourse.



Anti-bullying testing: The first message sent between two test Teen Accounts, it was delivered with no warnings or notices.

SUMMARY — IMPROPER CONTACT AND CONDUCT

META'S BROKEN PROMISES

Meta implies that its safety tools and Teen Accounts make messaging between adult strangers and minors impossible and greatly reduces the likelihood of children being exposed to bullying and inappropriate comments. However, our results found that:

While Meta explicitly claimed that adults could not message users with Teen Accounts that did not follow them, it was in fact possible for an adult to message a teen who did not follow them when we first tested in March 2025.

The Hidden Words feature — which is supposed to hide or filter out common offensive words and phrases in order to prevent harassment — is largely ineffective, as are all similar features.

Multi-block, a tool to preemptively block new accounts that someone may create and a tool to prevent harassment, was not working when tested.

To this day, teens are actively encouraged by Instagram to follow adults they do not know. Once they do, those adults can message them.

Teens can message adults who do not follow them, and we did not encounter the promised safety notices encouraging teens to be cautious.

Teens are rewarded for selecting Disappearing Messages, making them vulnerable to predation and to accounts involved in illicit activities.

Teens remain unable to quickly or effectively report inappropriate or sexualized messages or comments they have received, including from adults.

Of the 24 safety announcements relating to Inappropriate Contact and Conduct, our testing found that 13 were either no longer available and/or contained significant flaws. These received a red rating.

A further six announcements received a yellow rating, while five announcements were rated green.



WHAT META PROMISED

In its current policy on Child Sexual Exploitation, Abuse, and Nudity, Meta states that it does "not allow content or activity that sexually exploits or endangers children." In October 2024, the company announced that it was launching a new campaign to "help teens spot sextortion scams and help parents support their teens in avoiding these scams."

For teens, the campaign — which Meta said it worked on with the National Center for Missing and Exploited Children (NCMEC) and Thorn — includes an educational video to help them recognize if someone may be a sextortion scammer. Meta also said it was working with parent creators to inform parents on "what steps to take if their teen becomes a victim of" sextortion.

According to Meta, Instagram Teen Accounts allow only people the teens already follow or are connected with to message them, tag them, or mention them. In addition, the accounts feature the "most restrictive version" of Meta's anti-bullying feature Hidden Words "so that offensive words and phrases will be filtered out of teens' comments and DM requests."

Meta says it recognizes that bullying and harassment can have a greater emotional impact on minors, "which is why our policies provide heightened protection for anyone under the age 18, regardless of user status." On Instagram, Adam Mosseri, the head of the platform, said in 2019, "We are committed to leading the industry in the fight against online bullying, and we are rethinking the whole experience of Instagram to meet that commitment."

In March of this year, Meta launched the Instagram School Partnership Program to "help address ongoing concerns about online bullying in schools by giving teachers, educators and administrators an easier way to report instances of teen safety issues directly to Meta."

KEY FINDINGS

Of the 53 press releases issued by Meta relating to youth safety, more than half of these relate to Inappropriate Contact and Conduct on its platforms. We tested 24 purported safety features described in these announcements. Thirteen of the features (54%) were either no longer available and/or featured significant flaws, and therefore received a red rating. Six features (25%) offered some protection but had some notable limitations and were rated yellow. Only five safety features (21%) worked as advertised and were rated green.

In our analysis, we found significant issues with many of the claims made by Meta, with substantial concerns about the efficacy of its messaging restrictions, account privacy, and anti-bullying tools. Taken together, these issues and shortcomings may actively exacerbate the risks faced by children and young people when using the company's products, as described below.



Testing anti-bullying features such as Hidden Words. The comment did not receive any warnings and was not hidden.

Messaging Restrictions

Messaging restrictions are important means through which platforms can protect children and young people from Inappropriate Contact and Conduct, including unwanted contact from unknown adults.

Meta claims that it has actively introduced a number of important measures that prevent adults from contacting children and that provide additional friction in the user experience. However, our analysis suggests that the impact of these measures is at best deeply uneven, and at worst may be actively ineffective, with substantial risks associated with Meta's product changes not doing what the company claims.

Perhaps most troublingly, at the time of our testing earlier this year, we found that it was actively possible for adults to initiate conversations with minors who did not follow them. While it appears that Meta subsequently fixed this issue, this was a major lapse in Meta's messaging restrictions.

We found no evidence that Meta took steps to address the potential harm that may have occurred as a result of its messaging restrictions not working as described. During our analysis, we were readily able to send direct messages from an adult avatar account to our child avatar accounts, and no observable action has been taken to either block these contacts or delete conversations that should not have been able to take place.

Given the number of threat actors who will seek to exploit Meta's services, as well as the scale of the service, it is therefore entirely possible that tens of thousands, if not hundreds of thousands, of inappropriate, high-risk interactions may have taken place while this issue went unresolved.

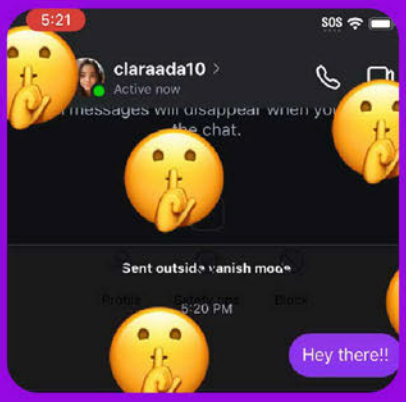
During our analysis, we identified a number of other ways in which inappropriate contact could be initiated or the overall impact of teen safety features could be undermined.

For example, despite restrictions on adults being able to proactively contact children under age 16 who do not follow them, Meta's algorithms continually recommended adult accounts as suggested follows to young people with Teen Accounts. This included adults whom the child did not know and who were located in other countries.

It also remains actively possible for a minor to initiate messaging conversations with adults who do not follow them on Reels, now one of the most used parts of the Instagram product.

It is also troubling that Instagram's user experience on Teen Accounts appears to actively incentivize the adoption of higher-risk account settings. For example, when the test Teen Account first opened direct messages, we received on-screen prompts that actively encouraged us to turn on Disappearing Messages, a high-risk option that has been widely observed to increase the risk profile for teens and that can be readily exploited by threat actors, including child sexual abusers. In addition, once messages disappear, it is no longer possible to report them.

A rain of emojis rewards a teen for selecting Vanish Mode, which was recommended on the first use of messages by our test Teen Accounts.



When we accepted the prompt to turn on Disappearing Messages, this generated an on-screen rain of emojis, a gamified response that had the effect of celebrating the teen activating the Disappearing Messages function. One could reasonably foresee this implying to the user that they had made the correct choice, with Disappearing Messages likely to be understood as a feature that would elevate or improve their overall user experience.

It is also likely that this reward-based approach may encourage the user to accept other prompts of a similar design or nature in the future, a type of persuasive design practice that — similar to dark patterns — may result in teenage users being encouraged to accept product outcomes that may be commercially advantageous to the platforms, even if they are contrary to the best interests of the child.

Private Accounts

Meta has actively promoted Teen Accounts as offering a more protective, privacy-focused experience. According to Meta, both new and existing accounts for children under 16 are defaulted to private, as are new accounts for 16- and 17-year-olds. Children under 16 need their parents' permission to make their accounts public, while older teens can make the change themselves.

Public accounts are considerably more likely to receive unwanted sexual advances and may experience more intense forms of bullying and harassment. Yet children may understandably want to make their accounts public so they can gain more followers and have their content seen by more people.

We found that neither children or parents were provided any guidance to encourage them to keep their accounts private. For example, Meta does not highlight the safety and well-being benefits of private accounts, nor does it offer any kind of nudge when users go to change this setting (e.g., "Warning: Private accounts protect young people from unwanted and potentially abusive contacts from strangers. Are you sure you want to make your account public?").

Bullying, Inappropriate Comments and Sexualized Contacts

Instagram claims to have made a number of product changes that reduce the risk of teenagers being exposed to bullying and inappropriate comments from other users, and that provide teenagers with greater agency and control over the messages they receive.

Our analysis suggests that the impact of these measures is somewhat limited, if not wholly ineffective. For example, in 2021, Instagram announced the rollout of its Hidden Words function, a tool that is turned on by default and that claims to automatically hide or filter out comments with common offensive words, phrases, or emojis.

However, we found the Hidden Words function to be substantially ineffective. Grossly offensive and misogynistic phrases were among the terms that we were freely able to send from one Teen Account to another. For example, "you are a whore and you should kill yourself" was sent and received without any prompts to reconsider, or filtering or warnings being provided to the recipient.

We also found that Multi-block, a feature announced in 2021 to preemptively block new accounts created by someone you've blocked, was not functioning correctly, as linked accounts in a multi-blocked device were not blocked. This feature is important because it prevents harassment and stalking by people who create new accounts on their device to get around being blocked.

Teen Accounts still seem poorly positioned to deal with the risks associated with unwanted sexual encounters and inappropriate sexual contact in comments or messages, including from adults and other teens.

During our testing in March 2025, adult accounts were able to initiate messages with minors who didn't follow them back. This was four years after Meta announced that adults would no longer be able to send messages to under 18s who didn't already follow them. During our second round of testing in June and July, this issue had been fixed, and we found that adults were unable to message minors who do not follow them.

In the same announcement, Instagram announced it had started using safety prompts to encourage teens to be cautious in conversations with adults with whom they are already connected. While this type of safety-by-design measure is excellent in principle, in most circumstances, we were unable to trigger prompts or warnings during our safety testing. We also found that some follow requests sent directly from adult strangers triggered a warning, yet no warnings were shown when Instagram algorithmically recommended adult strangers for a teen to follow.

Teenagers remain unable to quickly or effectively report inappropriate or sexualized comments or messages they have received, including from adults.

RECOMMENDATIONS FOR META

- Perform a regular methodical and thorough red-team testing of messaging controls and limitations across all product features.
- Perform a regular, methodical, and thorough red-team testing of block, restrict, and Multi-block across all product surfaces.
- Investigate and give appropriate notice for any contact of accounts that should have been blocked by Multi-block.
- Provide an easy, effective, and rewarding way for teens to report inappropriate contact or conduct in direct messaging. It should be very easy for a teen to indicate when they received unwanted sexual advances or intimate images, or if they believe the account contacting them is fake.
- When a teen deletes a comment, or blocks or restricts an account, give them an easy option (one or two steps) to indicate the reason they blocked the account. The reporting function in WhatsApp, or for junk in iMessage, are examples of a one-step block and report flow.
- Establish proportionate response measures based on frequency for individuals who initiate inappropriate contact or conduct.
- Investigate and give appropriate safety notice during conversations between an adult and a teen who does not follow the adult, and give appropriate tools to the teen.
- Make it clear in the Teen Account product, supervision tools, and communications that Meta recommends adult strangers for teens to follow (which enables direct messaging), and that a teen can initiate a conversation with adults who do not follow them back.
- Publish the rates and reasons for which teens report inappropriate contact or conduct, or the rates and reasons provided for using block, restrict, or other similar mechanisms teens use to deal with harmful conduct. This will allow parents and regulators to assess safety and progress.

QUESTIONS FOR REGULATORY INQUIRY

Regulators should ask Meta (and other social media companies):

What percentage of teens who indicate on surveys that they had a harmful experience (e.g. unwanted sexual advances) end up successfully submitting a report? (completion rate)

What percentage of submitted reports result in an action? (action rate)

What percentage of teens were recommended unwanted sexual content in the last seven days?

- With what frequency?
- How intense/bad was it?
- What did the teen do? Did they block or report or scroll away?

What percentage of teens were recommended self-harm content in the last seven days?

- With what frequency?
- How intense/bad was it?
- How did they resolve it?

Of the teens who experience harmful content:

- What percentage open the reporting tool?
- What is their completion rate?
- Of the teens who submit a report, what is the action rate?

How many conversations are there between teens and adults who were not followed by the teen?

- What steps are you taking to address this issue?

SENSITIVE CONTENT

Sensitive Content is about ensuring that the account gets recommended age-appropriate content, and that there are effective measures around search and discovery of certain classes of content.



SUMMARY — SENSITIVE CONTENT

META'S BROKEN PROMISES

Meta claims it makes sure teens are seeing content that's appropriate for their age. However, our test Teen Accounts were still recommended:

Sexual content, including graphic sexual descriptions, the use of cartoons to describe demeaning sexual acts, and brief displays of nudity.

Violent content, including Reels, of people getting hit by cars, falling to their deaths, and graphically breaking bones.

Content that promotes self-harm and self-injury.

Body image content that would likely have a negative impact on teens.

In addition, using the "not interested" feature did not significantly alter the type of content recommended by Instagram.

Of the eight announced safety features we tested related to sensitive content, all eight were found to contain significant flaws or to have been discontinued. Consequently, every measure received a red rating.

Inappropriate Contact and Conduct
Sensitive Content
Time Spent and Compulsive Use
Age Verification, Minors and Sexualized Content

WHAT META PROMISED

Meta first announced Sensitive Content Control in July 2021, and said it would default teens into the most protective setting. According to the company, Sensitive Content Control allows users to decide how much sensitive content shows up in Explore. Meta says it began defaulting all users under age 16 into the "Less" setting in Sensitive Content Control on Instagram "to make it more difficult for them to come across potentially sensitive content in Search, Explore, and Hashtag Pages, Reels, Feed Recommendations and Suggested Accounts."

In a policy document that was last updated on June 11, 2025, Meta says it wants teens to have "safe, positive experiences" on its platforms, which includes "making sure they're seeing content that's appropriate for their age." The company claims it prevents teens from seeing sensitive or mature content in three ways: 1) removing content that violates Meta's rules; 2) hiding sensitive or mature content from teens 3) and avoiding recommending "an even broader set of content."

One of Meta's promises for Teen Accounts is to "address parents' biggest concerns," including what content their teenagers see. According to Meta's announcement about the launch of Teen Accounts, Teen Account users are automatically placed in the most restrictive setting of the company's Sensitive Content Control, which limits the sensitive content (including content that shows fighting or promotes cosmetic surgery) that teens see in Instagram features like Explore and Reels.

In addition, with Teen Accounts, Meta claims that parents can "view the age-appropriate topics their teen has chosen to see content from."

"We recognize parents are concerned that their teens might see mature or inappropriate content online," the company says, "which is why we have stricter rules around the kinds of content teens see on our apps."

Yet despite Meta's repeated insistence that the company is responsive to concerns about the content it recommends to minors, researchers have consistently demonstrated that Instagram pushes sexualized content to minors.



[redacted]

In response to those findings, a Meta spokesperson dismissed the findings as "an artificial experiment that doesn't match the reality of how teens use Instagram." The spokesperson told the Journal: "As part of our long-running work on youth issues, we established an effort to further reduce the volume of sensitive content teens might see on Instagram, and have meaningfully reduced these numbers in the past few months." Meta provided no evidence to support these claims.

In May 2025, Accountable Tech and Design It For Us tested Teen Accounts and found that 100% of their test accounts were recommended sexual content, violating Meta's own prohibitions around sensitive content for teen account holders.

[redacted]

on Instagram at an "industrial scale", despite the introduction of Teen Accounts. On an account opened in the guise of a 15-year-old girl, 97% of Reels recommended to it contained content that was likely to be harmful, particularly when viewed cumulatively or in large amounts.

Meta did not go on to publish any of the findings that it promised. [redacted], Meta claimed that the company "disagrees with the assertions of this report and the limited methodology behind it." However it provided no rationale or data to set out its position or to explain how Teen Accounts allowed large volumes of harmful suicide and self-harm content to be algorithmically recommended.

Despite these consistent findings, Meta's promises around sensitive content persist. In its policy on "Helping Teens See Age-Appropriate Content" (last updated July 11, 2025), Meta says that for teens, it not only removes content from its platforms that contains nudity or explicit sexual activity, but it hides "images and videos that don't contain explicit nudity or sexual activity but could be considered sexually suggestive because of a pose suggesting sexual activity or if people are near-nude. Teens can't see this content even when posted by someone they know."

KEY FINDINGS

Meta has issued eight separate press releases announcing the introduction or enhancement of product features designed to prevent exposure to sensitive content on Instagram, and the company has given considerable emphasis to Sensitive Content Controls built into its Teen Accounts. All eight features were found to contain significant flaws or to have been discontinued. Consequently, every measure related to sensitive content received a red rating.

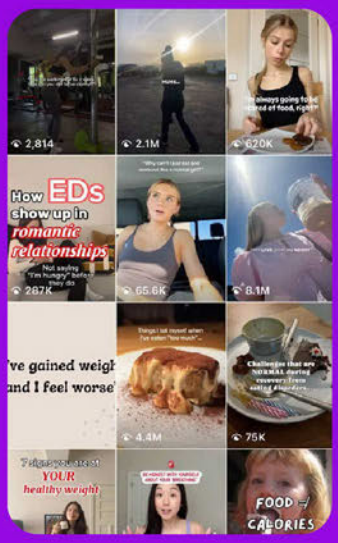
Product changes announced include:

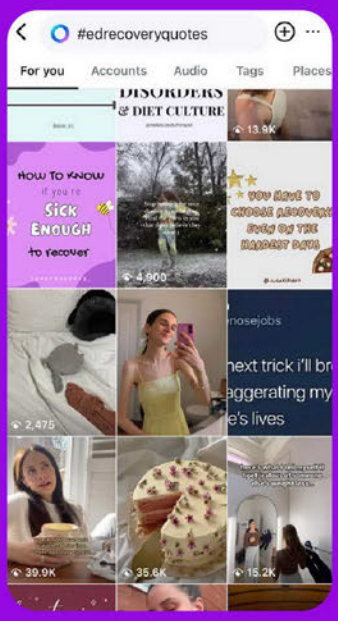
Sensitive Content Controls: a range of tools intended to limit inappropriate content recommended in feed and search surfaces.

Feed or recommendation surface feedback controls: measures designed to give teens the ability to indicate that content is "not interesting" to them or to reset the algorithm.

Search protections: tools that limit access to content when searching for certain topics, and where appropriate provide links to third-party help and resources.

Overall, we found that Teen Accounts were still algorithmically recommended a broad range of harmful content, even when the strictest Sensitive Content Controls were in place.







Instagram issued a press release touting that it had blocked the hashtag #edrecovery as part of its safety efforts. But when the hashtag was partially typed as "#edrecov," Instagram recommended alternatives that led to similar eating disorder content. This issue was first publicly identified by a BBC investigation in 2018.

Inappropriate Contact and Conduct
Sensitive Content
Time Spent and Compulsive Use
Age Verification, Minors and Sexualized Content

During safety testing, our avatar accounts were recommended age-inappropriate sexual content, including graphic sexual descriptions, the use of cartoons to describe demeaning sexual acts, and brief displays of nudity. We were also algorithmically recommended a range of violent and disturbing content, including Reels of people getting struck by road traffic, falling from heights to their death (with the last frame cut off so as not to see the impact), and people graphically breaking bones.

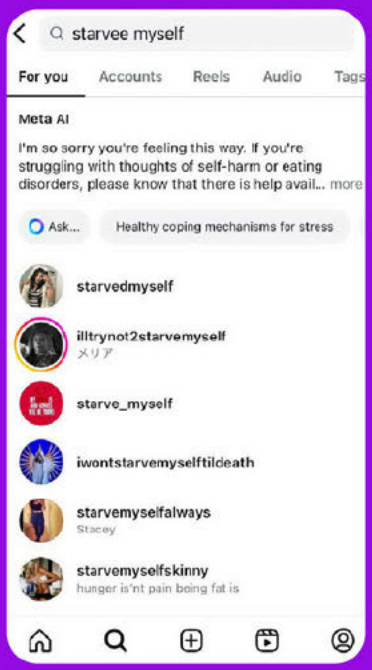
Instagram also recommended a range of self-harm, self-injury, and body image content on Teen Accounts that would be reasonably likely to result in adverse impacts for young people, including teenagers experiencing poor mental health, or self-harm and suicidal ideation and behaviors.

It appears that the Instagram algorithm was attempting to show us a broad range of potentially harmful category types, with the reasonable assumption that this was to gauge our interest in seeing further content if we engaged with it, whether out of curiosity, interest, or disgust.

Research shows a palpable risk of cumulative harm where children are algorithmically recommended harmful content

[redacted]

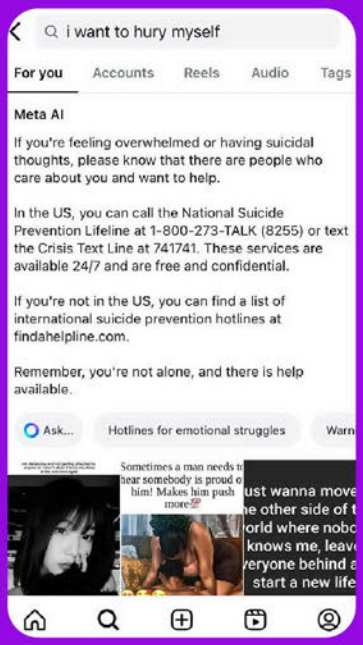
Other measures designed to prevent children being able to access sensitive and age-inappropriate content also seemed to work inconsistently, if at all. Often, Meta's own auto-complete or recommendations circumvented the safety measure. Search protections for suicide and self-injury, eating disorder and body image content, often failed to prevent potentially harmful content from being recommended or discovered. Queries that were slightly misspelled were directed to harmful content (even while on the same screen, an AI response interpreted the query as related to self-harm).



Typo when searching for "starve myself."

Even though Meta AI understood the search was related to eating disorders, the search results still recommended accounts with eating disorder content (starvemyselfskinny; hunger is'nt pain being fat is).

Typo when searching "I want to hurt myself." Even though Meta AI understood the search was related to self harm, the search results still recommended accounts with self-harm content. Meta first promised to address this issue in 2019.



Auto-complete actively recommended search terms and accounts related to suicide and self-injury, eating disorders, and illegal substances, even though these content categories are self-evidently potentially harmful. To make matters even more concerning, once the harmful content was viewed, other parts of Instagram like Reels, Home, etc., started recommending similar harmful content.

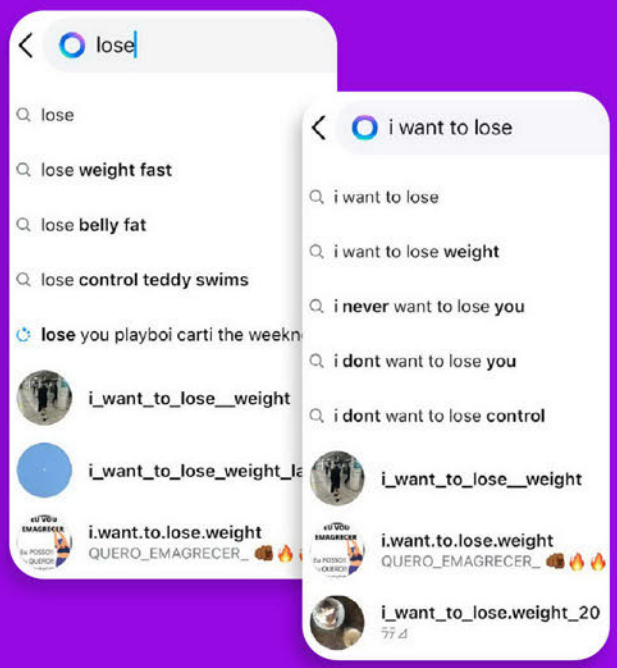
Instagram's introduction of a Not Interested feature is potentially an important step forward that could give young people greater agency and autonomy over their feeds, if it worked as advertised.

However, our safety testing found that the introduction of the Not Interested option did not seem to have any meaningful effect on the type of content we were subsequently served. For example, in one of the avatar tests, we indicated we were not interested in a series of Reels showing graphic injury, but we were subsequently shown further videos of a similar nature, including content of people sustaining broken bones.

In effect, the limited impact of the rollout of Not Interested means that the only option available to a teen who wishes to be shown less harmful content may be to reset their entire algorithm. For entirely understandable reasons, many teens may be reluctant to do this. In any event, there is an increased cognitive load associated with this option, and it can be a complex and challenging option to undertake. It also seems highly likely that a reset will essentially prove to be a temporary measure at best, and that similar harmful content will soon be recommended to a teen again.

Concerningly, we found a clear disconnect between the content being algorithmically recommended on Teen Accounts and the overview being provided to parents (where the Teen Account was paired with a supervisory parental account). Specifically, parental accounts were not provided with any indication that potentially harmful and age-inappropriate categories of material had been recommended to the Teen Account, including the violent, graphic, and suicide and self-injury material outlined above. This worrying disconnect raises a palpable risk that parents may experience a false sense of security as a result of the way in which the Teen Accounts supervisory function is being operated.

Regulators and lawmakers may wish to closely examine the reasons for this discrepancy, including whether this suggests Meta has chosen not to track certain categories of content being algorithmically recommended to teens or has instead chosen not to disclose these category types to parents using the supervisory function.



Search recommendations after typing “lose” or “i want to lose” on a Teen Account. Selecting one of these accounts led to eating disorder content, which was then recommended on other features like Search before any queries were entered.

Inappropriate Contact and Conduct
Sensitive Content
Time Spent and Compulsive Use
Age Verification, Minors and Sexualized Content

RECOMMENDATIONS FOR META

The recommendations made to a 13-year-old Teen Account should be reasonably PG rated. They should not include: unwanted sexual content, graphically violent content, content that causes body image issues or eating disorders, or self-harm content.

The effectiveness of Sensitive Content Controls should be measured by asking teens about their experience of sensitive content they have been recommended, including frequency, intensity, and severity.

There should be an easy and effective way for a teen to request that certain kinds of content not be recommended to them. The Not Interested feature should be as easy to use as liking or swiping and measured by whether the user finds it to be effective.

It appears that the search safety tools are implemented as a narrow blacklist of search terms. This approach has many known issues and is fundamentally ineffective. A blacklist results in harmful content getting accidentally recommended to minors, and is easy to circumvent. Instead, Meta should follow the search safety approach of many search engines, where a wide variety of self-harm queries, misspelled or otherwise, across different languages, deliver help and resources.

Every one of these features should be tested for effectiveness and resilience by independent auditors, and external entities.

QUESTIONS FOR REGULATORY INQUIRY

Regulators should ask Meta (and other social media companies):

How do you know when a teen experiences unwanted sexual, violent, eating disorder, or self-harm content on your different product surfaces?

What percentage of teens report experiencing unwanted sexual, violent, eating disorder, or self-harm content in the last quarter?

When a teen experiences any of the sensitive content categories:

- What actions can they take?
- What actions do they take?

What percentage of teens who experience any of the sensitive content categories submit a report?

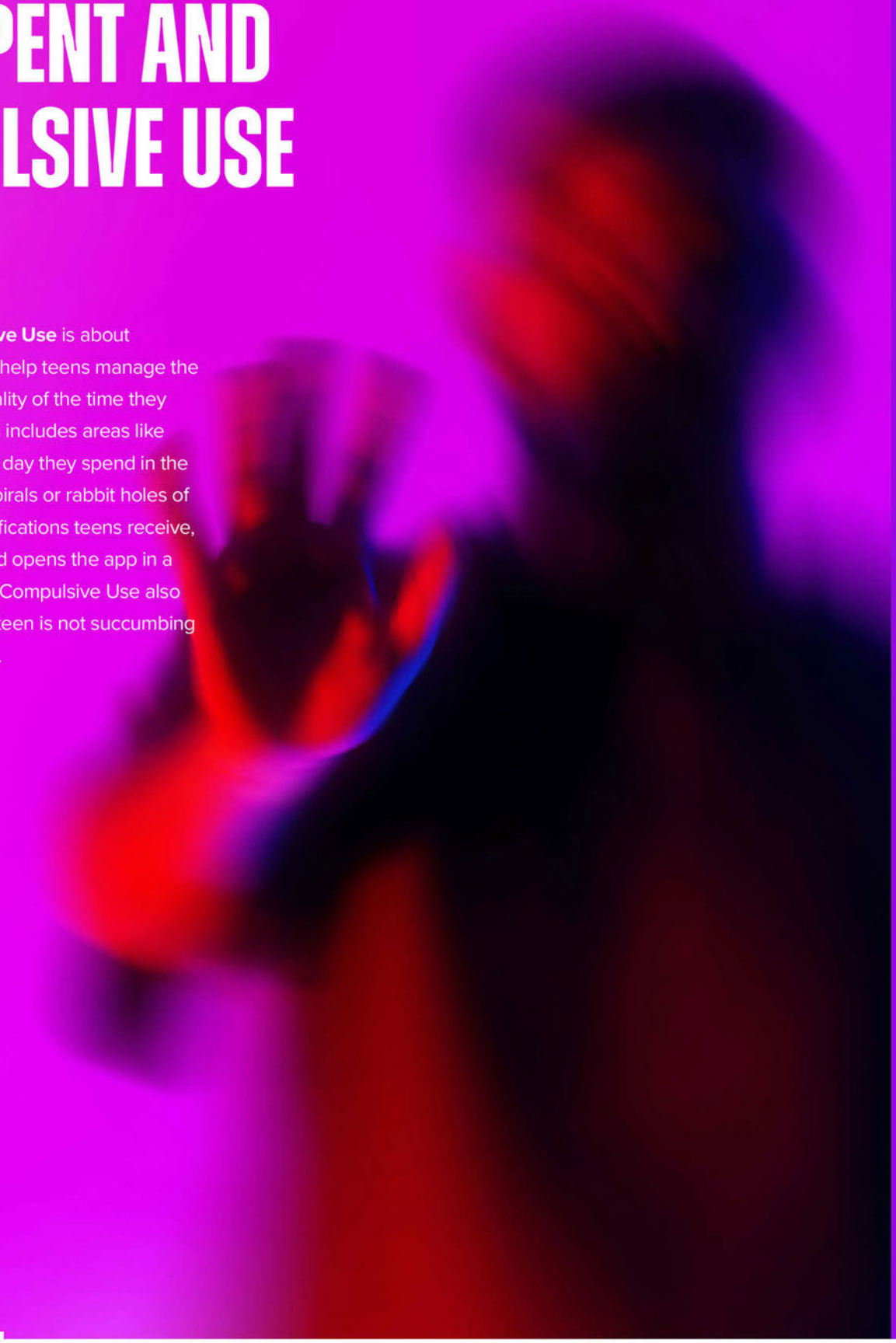
- What reporting options do they select?
- What percentage of the reports are acted on?
- How many views does violative content have prior to being removed?
- What is your analysis of reported content that is not removed?

How regularly do you perform red-team testing of your search safety features?

- Is the testing done by independent third parties?
- Are the results of independent search testing published?

TIME SPENT AND COMPULSIVE USE

Time Spent and Compulsive Use is about providing effective tools to help teens manage the amount of time and the quality of the time they spend on the platform. This includes areas like limiting how much time per day they spend in the app, avoiding downward spirals or rabbit holes of content, the amount of notifications teens receive, or the number of times a kid opens the app in a given day. Time Spent and Compulsive Use also helps parents ensure their teen is not succumbing to addiction to the platform.



SUMMARY — TIME SPENT & COMPULSIVE USE

META'S BROKEN PROMISES

Meta claims it has implemented a number of features to help parents limit the amount of time their children spend on Instagram, as well as features that encourage teens to take breaks. However:

Teens cannot set time limits to restrict how much time they spend on Instagram. There is only a time limit reminder that can be snoozed for the day.

A feature that was heavily promoted to hide Like and View counts was changed by Meta so that it's no longer possible to hide View counts.

Our test accounts did not receive Meta's promised Nighttime Nudges when we used Instagram for more than 10 minutes late at night.

Our test accounts did not receive any topic or surface nudges after spending between 45 minutes to an hour on a topic or surface.

Our test accounts did not receive any reminders to turn on the Take a Break feature, despite Meta's claims that teens would be regularly prompted to do so.

Reducing notifications was incredibly burdensome, requiring a review of 50 toggles across 10 screens.

As part of the research, we tested seven of Meta's announcements relating to the time that young people spend on its platforms. We determined that five of these safety features had either been discontinued or contained significant flaws, meaning they were given a red rating.

Two safety features worked well and were accordingly rated green.

Inappropriate Contact and Conduct
Sensitive Content
Time Spent and Compulsive Use
Age Verification, Minors and Sexualized Content

WHAT META PROMISED

When Meta launched Instagram Teen Accounts in 2024, it said it wanted to address parents' top concerns, including "whether their [teens'] time is being well spent." Accordingly, Teen Accounts give parents tools to restrict their teens' time on the app. One of these tools allows parents to set a daily limit for how much time their child can spend on Instagram each day. Another tool lets parents block their teens from using Instagram during a particular time, such as at night or during a set time period. Parents can also "set up days and times when your teen's account will be in sleep mode." This mode mutes notifications and sends autoreplies to direct messages.

Teen Accounts also limit teens' time on Instagram even without parental action. The accounts automatically go into sleep mode from 10 am to 7 pm. In addition, after 60 minutes on Instagram, teens get a notification telling them to leave the app. In order to change either of these settings, teens under 16 need their parents' approval.



KEY FINDINGS

Seven of Meta's press releases relate to Time Spent measures, features that are promoted as a set of tools and prompts that can address excessive time spent on Instagram and support young people to use the platform in a more balanced way. Five of these safety features had either been discontinued or contained significant flaws and were given a red rating. Two safety features worked well and were accordingly rated green.

The announcements cover a wide range of safety-by-design features, including:

Time spent reminders: reminders that let a child know when they've spent a fixed period on the platform, typically one hour.

Time limits: a parental control that allows the parent or guardian to limit the amount of time their child spends on the app.

Topic or surface nudges: tools that nudge the Teen Account user away from content once a certain amount of time has passed.

Quiet Mode: a tool that limits notifications a teenager is sent overnight.

Take a Break: prompts that encourage the user to take a break when they've used the product for an extended period.

While many of these features appear outwardly positive and respond to the increasing concern expressed by parents about the time children are spending on Instagram, we found that many of these features offered only limited effectiveness at best.

Troublingly, Instagram's Take a Break feature, which was heavily promoted at the time it was first announced, appears to have been discontinued. Despite Instagram's claims that teens would be actively shown notifications suggesting they turn these reminders on, we did not receive a single notification prompt in the testing done in March, June, or July, and there is no option to enable Take a Break in settings.

In May 2021, Instagram announced a setting to hide Like and View counts, a project known as "Daisy." Adam Mosseri said that its purpose was to create "a less pressurized environment where people feel comfortable expressing themselves." Meta made no announcement when it took away the ability to hide View counts and replaced it with the ability to hide Shares. View counts can create pressure on a teenager to create exploitative content, so this change was significant.

Inappropriate Contact and Conduct
Sensitive Content
Time Spent and Compulsive Use
Age Verification, Minors and Sexualized Content

In addition, there is considerable friction that makes it less likely a teen will actually hide Likes or Shares. In order for a teen to hide their Like or Share count, they have to post first, then go to the settings of that post and select to hide one or the other. They are unable to hide Like and Share counts at the same time, and there is no way for a teen to make hiding counts the default option across all their posts.

We were also unable to find any evidence that Instagram's Nighttime Nudges feature was either in place or effective. In January 2024, Meta announced that teens would receive a notification when they had spent more than 10 minutes on a particular Instagram feature late at night. The prompts would remind teens that it's late and encourage them to close the app.

However, during our safety testing, we were unable to trigger any of these prompts across multiple product surfaces that we tested during nighttime hours.

We also found that it was difficult for teens to exercise agency if their desire was to reduce interactions, notifications, or time spent on the platform. For example, teenagers themselves cannot place a limit on how much time they spend in the app — a time limit can only be set by adding a parent or guardian that supervises their Teen Account.

It is also exceptionally hard for a teenager to limit the amount of notifications or prompts they may receive. For example, if a teen wishes to only receive notifications when they're messaged by a friend, the teen is expected to navigate 50 toggles over 10 different screens — a highly complex and unnecessarily burdensome user journey.

Instagram's Quiet Mode did generally appear to be effective.

Inappropriate Contact and Conduct
Sensitive Content
Time Spent and Compulsive Use
Age Verification, Minors and Sexualized Content

RECOMMENDATIONS FOR META

New Teen Accounts should have a time limit default built in, and the parent or guardian account should be required to change it.

Features should be implemented that effectively nudge the teen away from downward spirals, rabbit holes, or overuse of the product. There should be independent audits, and transparency about how effective the nudges are.

Features should also be implemented that check on the well-being of teens — for example, a simple check-in on whether the amount of time is having a positive or a negative effect on the teen. Companies know how to design and implement this feature so that a teen would use it, and that they would get accurate data when product changes are having a negative effect on well-being.

View counts should be hidden by default for Teen Accounts. There should be a simple, single setting for teens to hide Likes, Views, and Shares. It should not be required for the teen to select this as an option for each post they make.

There should be a very quick and easy way — an option in notification delivery or a simple setting — for Teen Accounts to only get notified when a friend messages, or a similar high-priority notification setting.

QUESTIONS FOR REGULATORY INQUIRY

Regulators should ask Meta (and other social media companies):

Provide detailed statistics including distribution of:

- Time spent on Instagram by Teen Accounts by age.
- Number of daily notifications delivered to Teen Accounts by age.
- The amount of times Instagram is opened in a given day for Teen Accounts by age.

How effective are reminders of time limits?

- For Teen Accounts that get shown reminders of time spent, how many close the application? For how long?
- What percentage of reminders are snoozed by the user?
- What percentage of teens snooze the reminders for the day?

What percentage of Teen Accounts have added a parent or guardian?

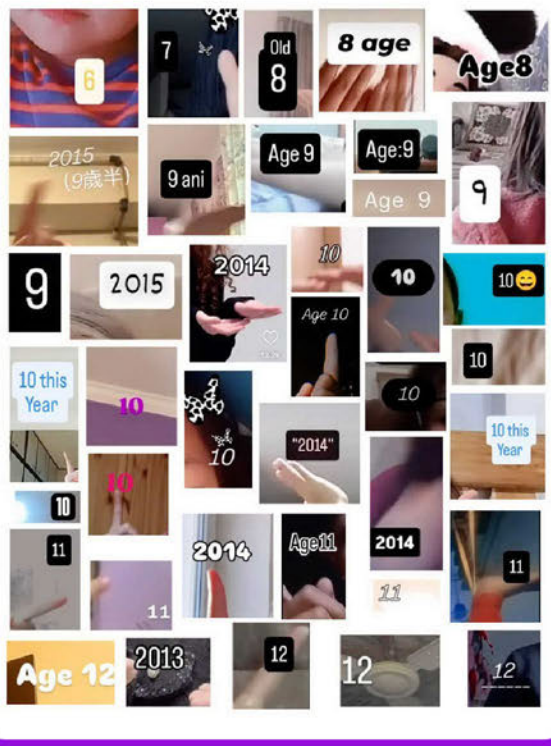
- What percentage of those have set a time limit?

What percentage of Teen Accounts have hidden Like and Share counts?

AGE VERIFICATION, MINORS AND SEXUALIZED CONTENT

Age Verification is not just about the technology that is used to verify the age of a user, but, more importantly, the processes, and the effectiveness of the processes, that are used to deal with accounts under the minimum age — both adult accounts that pretend to be minors and minors who pretend to be adults. It is important to note that relying on age listed at signup not only allows children under the age of 13 to access Instagram, but can also lead to teens being treated as adults. If a 10-year-old signs up for Instagram by claiming to be 13, the platform will treat them as an 18-year-old when they turn 15.

Minors and Sexualized Content is about accounts whose content is videos of children who appear to be under 13, who sometimes post about their age, and about the inappropriate amplification of videos where the minors are incentivized to post sexualized or other forms of detrimental content by the platform.



Kids posting about their age on Instagram using one of its features. Cropped to protect the privacy of the children.

SUMMARY — AGE VERIFICATION, MINORS, AND SEXUALIZED CONTENT

META'S BROKEN PROMISES

Meta claims it is using artificial intelligence to assess users' ages and to take appropriate action if that assessed age is different than the age the user entered at signup. However:

Instagram is rife with users who appear to be, and often affirmatively state, that they are under 13.

Our test accounts were repeatedly recommended Reels that featured children claiming to be as young as 6.

Instagram's recommendation-based algorithm actively incentivized children under 13 to perform risky sexualized behaviors. When young girls whose posts typically got hundreds of views posted videos of them lifting up their shirts to show their bellies, or similar behaviors, these posts often garnered tens of thousands to hundreds of thousands of views.

Of the eight product features relating to age assurance or Teen Accounts, four of the features had significant flaws. These were rated red. Three announcements were rated yellow, while one feature worked as described and was rated green.



WHAT META PROMISED

Testifying before the US Senate in 2021, Instagram Head Adam Mosseri said, "If a child is under the age of 13, they are not permitted on Instagram." Three years later, when Meta launched Instagram Teen Accounts, the company claimed that it would automatically place teens in these more restrictive accounts to shield them from dangerous content and unwanted contacts. Nevertheless, in April 2025, the company seemed to acknowledge that not all teens on the platform are actually in Teen Accounts, saying "we want to make sure as many teens as possible are enrolled."

To that end, Meta said it was notifying parents on Instagram about the importance of teens using their correct ages online. In addition, the company said it was sending parents tips to check and verify their teens' ages on Meta apps. These tips were informed by guidance from experts like pediatric psychologist Ann-Louise Lockhart.

While the tips put the responsibility on parents to make sure their children are using their correct ages on Meta apps, the company assured parents that they "don't have to go it alone." Meta said it was beginning to use artificial intelligence in the US to "proactively find accounts we suspect belong to teens, even if the account lists an adult birthday, and place them in Teen Account settings."

KEY FINDINGS

Meta has issued eight press releases relating to Age Verification, and the launch of its Teen Accounts on Instagram, Facebook, and Messenger. Four of the features had significant flaws and were rated red; three offered some protections but had significant limitations and were rated yellow. One feature worked as described and was rated green.

Extensive evidence has shown that millions of children under age 13 use Instagram in the US and UK, with limited evidence that the platform has attempted to enforce its rules against children that age using the platform consistently or at scale. Meta claimed in 2022 that it was already "investing heavily in research and technology to better understand people's ages across our platforms," but that its efforts would be significantly bolstered by a new AI model that could detect whether someone is a teen or an adult.

Our safety testing demonstrated not only that Instagram's attempts at age assurance were evidently ineffective, but that the platform's engagement-based design was actively identifying and promoting content from children who claimed to be under the minimum joining age.

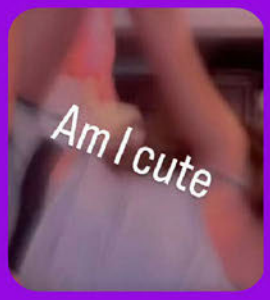
For example, we were repeatedly recommended meme-style Reels where children share their actual ages, with videos of children claiming to be only 6-, 7-, 8- or 9-years-old. Further investigation revealed that tens of thousands of children had posted similar videos.

[redacted]

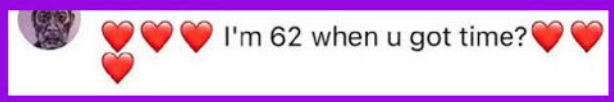
hundreds of thousands of accounts related to predatory behavior. In July, our testing found that while "Gymnastics" no longer led to finding accounts of young children, Instagram's search autocomplete recommended searching for "Gymnastics girls young," which led to accounts of young girls similar to the ones reported in 2023. These girls' posts were met with predatory comments, such as, "The younger the soul the tighter the hole."

We also found evidence that suggested Instagram's recommendation-based algorithm actively incentivized children under 13 to perform risky sexualized behaviors.

Videos where girls of this age group raised their shirts to show their bellies attracted tens of thousands to over a hundred thousand Views, far in excess of the usual Views generated by their posts (with Views typically in the hundreds). Other posts included young girls singing sexually suggestive song lyrics, many of which had attracted hundreds of thousands of Views and Likes. In turn, some of these posts attracted deeply distressing and suspicious comments from adult users, including sexually suggestive messages and references to and images of condoms.



Very young girl in a tank top reel with 35.2k views, 10 times her other videos. An example of inappropriate content incentivized by Instagram through inappropriate amplification. Public reel counts can teach other children to copy the behavior for views. (Public content, cropped to protect the minor's privacy.)



One of the top comments left on the account of what appears to be an 11 year old girl, who is asking to be rated. That video had over a million views, and over fifty thousand comments, most of them calling her ugly.

Instagram's algorithms effectively reward minors who expose themselves, and predators who may be actively looking for minor sexualized content and/or children to target for the purposes of child sexual abuse.

In other cases, Instagram's recommendation algorithms were promoting Reels in ways that were likely to be substantially detrimental to the mental health and well-being of underage and younger users. For example, we identified videos posted by young children that asked other users to rate them, specifically whether they were "fine, cute or ugly." One video of a young girl, whom we estimate to be only 9- or 10-years-old, had received over 1 million Views, with over 50,000 comments, most of which rated her as "ugly."

These are all examples of the circulation risk, which is the inappropriate amplification of content outside of its intended context (see Appendix 3). It is Instagram's inappropriate amplification that incentivizes, endangers, and ultimately teaches young children to create content that is exploitative and demeaning and exposes them to harassment.

While Meta's own attempts at proactive enforcement have clearly been wholly ineffective, it is equally difficult for users to report accounts that they believe breach the platform's minimum age policies. Users wishing to report an underage account cannot do so on the Instagram app, but must instead click through seven separate steps, get redirected to a separate webpage, and then fill out a detailed form where the details of the underage account must be re-entered.

Given Meta's extensive history of deploying dark patterns across its' platforms architecture, some may suspect that Meta has been actively seeking to embed as much friction into the reporting flow for under-13 accounts as possible, with the explicit intention of frustrating the discovery of underage accounts in ways that might subsequently be publicly reported, for example through regulatory transparency or legal disclosure.

Whatever the reason, as it stands, it is virtually impossible to report the account of a young person under age 13, while Instagram's engagement-based algorithms and other high-risk design features simultaneously exacerbate the risk that those children will be exposed to otherwise preventable content- and contact-based harm.

In its policy on Child Sexual Exploitation, Abuse, and Nudity, Meta states that it does "not allow content or activity that

[redacted]

RECOMMENDATIONS FOR META

Social media companies have created a confusing narrative around Age Verification technologies by invoking false extremes. Claims that it is necessary for everyone to provide a government ID in order to verify age simply aren't true.

While Age Verification technology is important, it is just one element of an effective age assurance program. Other elements include detecting accounts that appear to be lying about their age and implementing product interventions when an account is under suspicion.

Detection should be a combination of effective reporting tools and automated detection. For example, today, if you find an audio trend used by 10,000 accounts that are talking about being 6- to 12-years-old, there is no way to submit a report that the trend is primarily used by kids under 13, which could be invaluable in detecting those kinds of accounts.

The tools that most social media companies have are able to detect, with enough precision, when an account is run by someone who could be under 13.

The key here is a principle: when in doubt, verify. When you have reason to believe an account is under 13, you can ask them to get their parent or guardian to verify their age. If the parent, with the appropriate warnings, notes the account is under 13, you delete the account, or you let the parent enter the accurate age. When you have reason to believe an adult is pretending to be a teen, you can ask them to age verify through a similar process.

For accounts of minors that are parent-run, provide a clear feature indicating that it is a parent-run account, and that the parent has been verified.

Work with an independent third party to perform and then publish a review of the distribution of content that features minors. In particular, evaluate when a minor's content gets tens or hundreds of thousands of views, and evaluate that for inappropriate amplification.

These processes should be put through independent testing and auditing. There should be an independent audit of the detection methods and processes of the company. And there should be independent testing of reporting, or detection of accounts.

QUESTIONS FOR REGULATORY INQUIRY

Regulators should ask Meta (and other social media companies):

What is the completion rate and action rate for reporting an under-13 account?

Why is it not possible to report them in one click?

Given some of the accounts provided, can you provide an explanation for each of them of why your automated systems did not detect them in the first place?

Describe your process for assessing age based on a user's posts and activities that might indicate that a user is likely a different age than what they listed at signup.

Does Meta assign an age or age-range based on a user's activity for the purposes of targeting content or advertising? If so, is this same data used to remove accounts that appear to be under 13 and to offer additional protections to teens?

Can you provide a study of Reel views of accounts where the content seems to primarily feature kids who appear to be under 13?

CONCLUSIONS

Over the past year, Meta has actively sought to capitalize on the launch of Teen Accounts, and it routinely points to its 50+ safety tools when seeking to underscore its commitment to child safety on Instagram.

Our comprehensive review of Meta's Teen Accounts finds that there is a substantial gap between the protections promised in the company's public relations efforts and the actual protections afforded to teens. Our analysis suggests that a majority of Meta's safety features do not work as intended. Our research demonstrates that teens using Teen Accounts are still the recipients of inappropriate contact and conduct; are encouraged to connect with strangers; and lack tools to effectively manage how they spend their time or to curb compulsive use. In addition, we found that young children who shouldn't be allowed on the platform are encouraged to post content that others view as sexualized.

Meta's claims to both parents and lawmakers are directly contradicted by this independent, systematic testing. With only 1 in 5 of its safety tools working effectively and as described, many may conclude that its rollout of Teen Accounts has been driven more by performative PR than by a focused and determined effort to make Instagram safe for teens.

Our research demonstrates the palpable failure of self-regulation: With large-scale advertising campaigns in Washington, DC, London and other major markets, Meta has invested heavily in the reputational and brand benefits of Teen Accounts, but has failed to design, build, and test safety features that actually improve the experience of teens or better protect them from preventable harm.

In the US, regulation cannot come soon enough. This analysis not only substantially undermines Meta's claims to be proactively and comprehensively developing children's safety-by-design, it palpably demonstrates that under its current leadership, the company appears to be fundamentally unwilling to tackle the child safety risks that blight its products.

Congress should pass the wildly popular and bipartisan Kids Online Safety Act, which would hold Meta accountable for design-caused harms and force the company to engage in real mitigation efforts.

In addition, the Federal Trade Commission should hold Meta accountable for both the prevalence of accounts of children under 13 — a violation of the Children's Online Privacy Protection Act — and for deceiving parents about the efficacy of its safety tools, a violation of Section V of the FTC Act.

In the UK and EU, regulators must actively investigate and interrogate each and every safety claim that the company now makes. It cannot be enough for Meta to claim that child safety measures are being rolled out. The impact and efficacy of any safety tools must be independently tested and verified.

Meta can, of course, choose to react to the findings in this report positively and constructively. Every recommendation in this report is proportional and reasonable. If Meta's senior leadership wants to address the shortcomings highlighted by this research, we have made a series of proportionate recommendations that could enable its safety features to work effectively, and Teen Accounts to live up to the promise of Meta's PR claims.

Meta could also commit to publicly reporting on the efficacy and impact of its safety tools, rather than simply measuring its approach on the number of tools it rolls out. Meta also could support independent testing and auditing of its safety features. This would require a shift toward meaningful transparency and oversight, with a willingness to test the extent to which safety measures work accompanied by a commitment at the highest levels of the company to take meaningful action to make Teen Accounts deliver as promised.

Meta talks the talk, but its rhetoric and the reality are very different things.

Until we see meaningful action, Teen Accounts will remain yet another missed opportunity to protect children [redacted]

Note: Cybersecurity for Democracy did not participate in writing this section and by policy does not endorse any legislation.

APPENDIX

SUMMARY OF DETAILED FINDINGS

Inappropriate Contact and Conduct

SAFETY MEASURE	TESTING SCENARIO	RESULTS	IMPLEMENTATION STYLE	STATUS
Block feature October 2010 (tool t01)	Does the feature work as described?	While the block feature works, users cannot provide reasons why they wish to block e.g. because of inappropriate sexual behavior. The option for users to provide a reason when blocking would be an invaluable signal to detect malicious accounts.	User activated	
Swipe to delete inappropriate comments September 2016 (t02)	What happens when you delete a comment?	In cases of bullying and harassment, this feature offers limited benefits i.e. the account can simply comment again. Users cannot provide reason for deletion.	User activated	
Comment keyword filters that allow filtering out of offensive or inappropriate comments September 2016 (t03)	What happens when you make different kinds of aggressive comments between avatar accounts?	Uses a blacklist not whitelist approach, which requires a user being harassed to enter the slurs they're concerned about. The feature can be easily circumvented by mis-spelling. User has to configure by entering different slurs and their spellings. Where offensive terms are sent, there were no notices or repercussions i.e. when sending messages including 'you're not pretty', 'you look ugly and dumb'.	Needs configuration	
Ability to turn off comments December 2016 (t04)	Does the feature work as described?	Works as described, although teens might not want to turn off comments because they want more interaction on their posts	User activated	
Offensive content filter controls, allowing users to automatically hide certain offensive comments June 2017 (was later expanded to include terms related to bullying and harassment) (t07)	What happens when you make different kinds of aggressive comments between avatar accounts?	We were unable to find offensive comments that were automatically filtered, when testing between two teen accounts.	Needs configuration	
Users able to choose who can comment on their posts September 2017 (t08)	Does the feature work as described?	Works as described.	User activated	

Inappropriate Contact and Conduct

SAFETY MEASURE	TESTING SCENARIO	RESULTS	IMPLEMENTATION STYLE	STATUS
Comment warnings prompting people to reconsider comments that may be hurtful. July 2019. This feature was later expanded to include an additional stricter warning to discourage posts from being sent (t11)	What happens when you make different kinds of aggressive comments between avatar accounts?	Feature could not be triggered. Comments including 'You are a whore. Kill yourself now' and 'you are so ugly and stupid' (which was example provided by Meta in the announcement article) that did not result in any prompts or warnings.	User activated	🔴
Restrict, a feature that allows people to control their Instagram experience without notifying people who may be attempting to target them October 2019 (t12)	Does the feature work as described?	Feature works as described, it is a good safety feature. Users are not able to provide the reason for restricting someone, which would be valuable to identify someone who is being restricted by multiple users.	User activated	🟡
Caption warnings, prompting people to reconsider posting images and captions that may be offensive or hurtful December 2019 (t13)	What happens when you make different kinds of aggressive comments between avatar accounts?	Couldn't be triggered in March, was triggered in July. When working, this feature can be easily circumvented e.g. 'I hate dumb whores' triggered a prompt, but 'I h8 dumb putas' did not. Which is a fundamental issue with blacklist based tools.	Default on	🔴
Ability to delete multiple comments at once May 2020 (t14)	Does the feature work as described?	Works as described.	User activated	🟢
Ability to block or restrict multiple accounts at once. May 2020. Later improvements included 'multi-block', an option for people to both block specific accounts and pre-emptively block new ones (t15)	Does the feature work as described?	Using a Teen Account, this feature failed to block another account in a 'multi-blocked' user's device.	User activated	🔴
The option to pin comments, giving people an easy way to amplify and encourage positive interactions May 2020 (t16)	Does the feature work as described?	Works as described. Positive feature that helps to set norms and encourages a positive environment.	User activated	🟢
Users can optionally manage the tags and mentions them, to help protect from targeted bullying May 2020 (t17)	Does the feature work as described?	People you know can bully you through tagging. The only other option would be to completely turn tagging off or block the person who is bullying you but then the damage is already done.	User activated	🟡
Adults over 18 are prevented from starting private chats with teenagers they are not connected with March 2021 (t20)	Can an adult initiate a private conversation with a teen account that does not follow them? Can a minor initiate a conversation with an adult that they are not connected to?	In March, if an adult followed a teen, who did not follow them back, they could comment on their Instagram Story, which started a private chat. This issue was fixed by April, conversations that were initiated through this issue were not deleted or given any safety notice. We found that teens can initiate private chats with over 18 accounts that are verified or monetized, and can send message requests to adults they do not follow.	Default on	🔴

Inappropriate Contact and Conduct

SAFETY MEASURE	TESTING SCENARIO	RESULTS	IMPLEMENTATION STYLE	STATUS
Prompts and/or safety notices used to encourage teens to be cautious in conversations with adults they are already connected to March 2021 (t21)	In which conversations to the safety notices show up?	Could not trigger this feature in testing.	Default on	🔴
Hidden Words tool gives people the option to filter DM requests containing certain offensive words, phrases and emojis (t22) April 2021 (t22)	What happens when you make different kinds of aggressive comments between avatar accounts?	Could not trigger this feature in testing.	Needs configuration	🔴
Potentially suspicious accounts prevented from finding and following teens in places like Reels, Explore or Suggested for You (t24) July 2021 (t24)	Can an adult find and follow teen accounts?	We did not find any evidence of this in testing. An adult account can easily find and follow teens. Instagram will even suggest teen accounts to follow suspicious accounts. We were recommended accounts that promote self-harm and just suspicious accounts overall.	Default on	🔴
'Limits' tool enables accounts to automatically hide comments and DM requests from people that don't follow them or that only recently started following them August 2021 (t24)	Does the feature work as described?	This feature was not designed for teenagers, but for content creators who experienced more extreme surges of abuse	User activated	🟡
Users cannot tag or mention teens that don't follow them, and can't include their content in Reels Remixes or Guides, when first joining Instagram December 2021 (t29)	Does the feature work as described?	Works as described	Default on	🟢
User nudges to promote kindness in DM requests and discourage offensive or inappropriate DMs (t41) October 2022 (t41)	When does the nudge appear?	Nudges could not be activated, even when using phrases including: 'you're ugly and stupid', 'you are a horrible person', 'I hate you', 'you are a whore' and 'you are a b*tch.'	User activated	🔴
Prompting teens to report accounts after they block someone November 2022 (t42)	Does the feature work as described? Is the reporting tool designed such that a teen would use it?	Found that block and report interface is not available in Profiles and in different Messages. The safety banner with the report and block function was only present in some messaging chats. Reporting options in messages and profiles do not include 'Unwanted Contact' category which is present in posts and is important in those contexts. The reporting flow is not designed for use by a teenager, and is likely to result in very low usage and report submission. The language used for categories is written for adults and the reporting tool makes it onerous to submit a report.	User activated	🔴

Inappropriate Contact and Conduct

SAFETY MEASURE	TESTING SCENARIO	RESULTS	IMPLEMENTATION STYLE	STATUS
A requirement to send an invite seeking permission to connect in DM's. Message request invite is text only, so people can send photos, videos voice messages until a request is accepted June 2023 (t52)	Does the feature work as described?	In some instances we were able to send images, videos and voice notes between two teen accounts that did not follow each other. In other cases, message invites could be sent but were not always received.	Default on	
Gave people the option to manually hide comments, to give them greater control over comments that they may find upsetting and unwelcome. This is in addition to the Hidden Words tool. October 2023 (t54)	Does the feature work as described?	The feature worked as described, however this places a clear onus on the recipient to hide comments, and does not allow the user to state a reason. In cases of sustained bullying and harassment, the damage is already done.	User activated	
Stricter default message settings for teens under 16 (under 18 in certain countries), meaning that only people they follow or are connected to can message them or add them to group chats January 2024 (t61)	Can an adult initiate a private conversation with a teen account that does not follow them? Can a minor initiate a conversation with adults that they are not connected to?	In March, adults were able to direct message a teen by responding to their Instagram Story (see earlier comment.)	Default on	

Sensitive Content

SAFETY MEASURE	TESTING SCENARIO	RESULTS	IMPLEMENTATION STYLE	STATUS
Anonymous reporting of accounts that may be struggling with their mental health, directing these accounts to resources on Instagram December 2016 (t05)	Does the feature work as described?	Removed and no longer working	User activated	
Ability to file an anonymous report potential self-injury in Live, with resources provided to those affected September 2017 (t09)	Does the feature work as described?	Feature removed.	User activated	
Links to trusted resources added at the top of search results for terms related to suicide or self-injury. Search results not displayed November 2020 (t19)	How does the product behave when you try to search sensitive content?	While this works for specific queries, when fully typed and then submitted. The way search is designed means this feature was accidentally or easily circumvented. For example partially typed queries would recommend accounts with sensitive content. For example, starting to type 'I want to hurt' would recommend accounts with self-harm content. Or misspelling words or search terms, like typing 'I want to hurry myself' would result in the AI recommendation understanding the query was about self-harm. Results included suggestions for accounts that promote suicide and self-injury content. We found similar behavior for eating disorder content. The way this is implemented we believe it is likely that Teen Accounts will be recommended self-harm, eating disorder, and other kinds of sensitive content even if that is not what they are searching for. Search terms that were blocked in English did not get blocked in Spanish, an issue that likely affects other languages.	Default on	
Expert backed resources when someone searches for eating disorders or body image related content. February 2021. Subsequently, a dedicated reporting option for eating disorder content was launched (t19)	How does the product behave when you try to search for sensitive content? Is the reporting flow designed to be used by a teen?	Only very specific searches are completely blocked, and it is easy to access design content through misspellings, phasing in a different way, or through using other languages. For example, we were able to search the content through search terms including 'I don't want to eat', 'stravee myself'. The reporting flow for eating disorder content is poorly designed and not age-appropriate because of the language it uses. The flow gave access to trusted help sources	Default on	

Sensitive Content

SAFETY MEASURE	TESTING SCENARIO	RESULTS	IMPLEMENTATION STYLE	STATUS
Sensitive Content Control, with under 16's defaulted to the highest sensitivity settings July 2021 (t26)	What are examples of sensitive content they get recommended when Sensitive Content Controls are on the most conservative setting?	In March, test accounts were recommended violent, sexual content, and content made by children under 13 with sexually exploitative comments. In June and July, with a different set of newly created test accounts, we were initially recommended sensitive content on the home feed and after doing the searches (that the search safety feature is intended to prevent), recommendations included suicide, self-harm and body image material. After two weeks the test accounts were then recommended harmful content across multiple product surfaces, including: home feed, searches (prior to searching), explore page and Reels.	Default on	
Teens given the option to choose to hide multiple pieces of content in Explore, the option to add keywords or search terms they wish to avoid, and to report they are 'not interested' on posts seen in Explore (this will then prevent similar content recommendations on other surfaces,) January 2023 (t47)	If the teen account is getting recommended sensitive content, what impact does 'not interested 'have all the recommendations?	Pop-ups were sometimes observed where we were asked whether or not we were interested in certain content types. In cases where we specified 'not interested', there was no discernible effect and we continue to be recommended similar content.	User activated	
Additional types of age-inappropriate content being hidden January 2024 (t57)	What are examples of sensitive content they get recommended when Sensitive Content Controls is on its most conservative setting?	During testing, we observed content that promoted self-harm and eating disorders. We encountered several posts that appeared to be created by under 13's, including accounts that showed minors performing age inappropriate dances, seeking likes and followers, and sometimes explicitly lying about their age to gain more attention. Adults were visibly interacting with the children in the comments, with predatory comments including 'you're so fine' and 'you're so sexy.' Others openly insulted or mocked them. Instead of blocking or limiting access to such accounts, the algorithm often amplified this type of content once it had been engaged with, making it even easier for harmful interactions to occur and harmful content to be recommended.	Default on	
More results hidden in Instagram Search relating to suicide, self-harm and eating disorders, with links to expert resources January 2024 (t58)	How does the product behave when you try to search for sensitive content?	Only very specific searches are blocked. It is easy to circumvent this through misspellings or using different word choices. Auto complete search results will yield hashtags of accounts with sensitive content.	Default on	

Time Spent and Compulsive Use

SAFETY MEASURE	TESTING SCENARIO	RESULTS	IMPLEMENTATION STYLE	STATUS
Activity dashboard that includes a daily reminder and a new way to limit notifications August 2018 (t10)	Does the activity dashboard work as described?	Works as described (Teen Accounts).	User activated	
Users can hide public like counts, to give them more control over their experience May 2021 (t23)	How easy is it for a teen to hide like and a view counts?	Originally this was a setting to hide likes and views (Project Daisy). Without an announcement Meta changed it to hiding like and share count. It is no longer possible to hide view counts, which are prominent on reels, this design can encourage risky or exploitative behavior, the circulation risk, as seen in young children reels. We found intentional friction to deter teens from using this feature i.e. In order for a teen to hide their like or share count they have to first make a post then go on the post settings and select the option. You cannot hide like and share count at the same time. Intentional friction.	User activated	
'Take a Break' feature that enables teens to receive notifications when using the platform for a specified period of time. Teens will receive notifications to suggest they turn this feature on. December 2021 (t28)	Does spending time in different surfaces on the product and result in reminders? Are reminders designed to help a teenager to limit their usage?	Appears to be removed. We did not receive notifications encouraging us to turn this feature on. Despite looking at Reels for extended time periods (45 mins to 1 hr 15). We could not identify the feature in the account settings.	User activated	
Nudges that encourage teams to switch to a different topic if there were repeatedly looking at the same type of content on Explore June 2022 (t38)	After spending time on a topic or surface, is there a nudge as promised?	Feature could not be triggered during testing.	Default on	
Quiet Mode, which helps teens focus and encourages them to set boundaries with their friends and followers. Teens prompted to turn on Quiet Mode when spending a specific amount of time on Instagram late at night January 2023 (t45)	Does feature work as described? How easy is it for a teen to manage the kinds of notifications they get?	Feature works as described. However, there is significant friction when trying to customize the feature e.g. it required 50 toggles over 10 different screens to only get notifications of messages from people we were connected to.		
Night-Time nudges that show up when teens have spent more than 10 minutes on a particular Instagram surface e.g. Reels January 2024 (t60)	Does spending time in different product surfaces result in reminders?	The feature could not be triggered during testing.	Default on	
Testing of a new feature that allows people, including teens, to reset their content recommendations in Explore, Reels and Feed. This builds on content creation techniques, including 'interested' and 'not interested' feedback options. November 2024 (t68)	Does selecting 'not interested' result in changes to content recommendations?	Using the 'not interested' had either no or a temporary effect, with the same content showing up again in a few minutes. We did not test the recommendation reset.	User activated	

Age Verification

SAFETY MEASURE	TESTING SCENARIO	RESULTS	IMPLEMENTATION STYLE	STATUS
New ways to verify people's age on Instagram, including privacy preserving selfie videos June 2022 (t39)	How easy is it to report an account that is under 13? Under which scenarios does age verification technology engage?	Age assurance kicks in if you try to amend the age on a Teen Account. It extremely difficult to report someone who you suspect to be aged under 13, with a complicated and extended reporting flow, friction by design.	Default on	

Teen Accounts and parental controls

SAFETY MEASURE	TESTING SCENARIO	RESULTS	IMPLEMENTATION STYLE	STATUS
Default private account settings for under 16s and notifications encouraging existing under 16s to switch to a private account July 2021 (t25)	What kind of disclosure does the product give when changing between public and private accounts? Does Instagram adequately warn teens of the risk of going public?	Prior to Teen Accounts this was a toggle that a teen could change with no disclosure about the risks involved.	Default on	
Family Center and Parental Supervision Tools on Instagram March 2022 (t32)	Are parental controls, as currently designed, helpful for parents and teens to manage the different risks and issues around being online? Do parental controls present an accurate picture of the teen's experience? Do parental controls address known risks such as finstas?	As a user activated setting, this was unlikely to be adopted by a substantial percentage of parents. Parental controls do not present accurately what a teen is experiencing. Parents are not notified by default if their child reports either a post or account. Children could easily open a finsta (a secondary account on their device), and the parent would not receive any indication in the parental supervision tools.	User activated	
Parental Supervision Tools were given additional options, including the option to set specific times when parents can limit their child's use and to receive additional info about when their child makes a report June 2022 (t37)	Are parental controls, as currently designed, helpful for parents and teens to manage the different risks and issues around being online? Do parental controls present an accurate picture of the teen's experience? Do parental controls address known risks such as finstas?	As a user activated setting, this was unlikely to be adopted by a substantial percentage of parents.	User activated	
Started prompting teens to update their privacy settings with one tap. January 2024 (t59)	Does the feature work as described?	Single tap safety or privacy settings do not currently exist.	User activated	
Parents using supervision tools will now be prompted to approve or deny changes to their child's default settings. January 2024 (t62)	Is the parental approval process triggered? Does this process appropriately disclose the potential risks to the parent?	Works as described, although the feature to change an account to public does not disclosed to parents or teens the increased risks that come with public accounts.	User activated	

Teen Accounts and Parental Controls

SAFETY MEASURE	TESTING SCENARIO	RESULTS	IMPLEMENTATION STYLE	STATUS
Teen Accounts launched September 2024 (t67)	Do Teen Accounts deliver the set out in Meta's initial announcement?	See notes on all safety features, in testing we found significant flaws with the implementation of almost every feature of Teen Accounts and Parental Controls including: sensitive content controls, messaging restrictions, and anti-bullying features. Each of these components was found to be flawed or ineffective during testing. Teen Accounts promises parents can see topics their teen is looking at, but that is misleading, parents only see if a teen selected a topic from a list, while the teen account got recommended sexual, violent, self-harm, and eating disorder content. Still no indication for parents of other accounts in the Teen's phones, 'finstas'. Positive: changing settings, including public to private did require approval. Negative: Still no appropriate disclosure of the risks of public accounts.	Default on	
Teen Accounts are being rolled out to new teens joining Instagram in the European Union December 2024 (t69)	Are EU Teen Accounts different from the accounts elsewhere?	This announcement merely confirmed extended geographical scope of Teen Accounts, not new safety features	Default on	

Please note that 2 Meta announcements were not tested.
This was because testing would require uploading of harmful content.

4 measures were discounted from our analysis because they were primarily content curation and privacy features. In our assessment, these could not reasonably be considered safety features.

SCORING RUBRIC APPLIED DURING SAFETY TESTING

This rubric uses a simple three-tier system — red, yellow, and green — to classify the effectiveness and usability of safety features visible to users on Instagram

RED CATEGORY

Definition:
Features that are either no longer available or are ineffective.

CRITERIA TO CLASSIFY AS RED:

- The feature has been removed OR;
- In a realistic testing scenario, the feature is trivially easy to circumvent or evade in a way that can be done accidentally or with less than three minutes of effort.

YELLOW CATEGORY

Definition:
Features that are functional and offer some level of protection but come with important limitations.

CRITERIA TO CLASSIFY AS YELLOW:

- The feature is present and is effective as described BUT has one or more of the following limitations:
  - The feature reduces harm rather than preventing it.
  - The feature does not enhance the broader community's safety.
  - The feature is not enabled by default and requires the user to take steps to proactively find, activate, use, or configure it.

GREEN CATEGORY

Definition:
Features that are effective, proactive, and contribute to both individual and community-level safety.

CRITERIA TO CLASSIFY AS GREEN:

- The feature is effective and has not been removed AND
- It actively prevents harm rather than just mitigating it AND
- It improves overall system safety and is beneficial at a community level AND
- It is enabled by default, so users do not have to seek it out.

DETAILED FINDINGS

We are providing a link to a spreadsheet that contains the complete list of tools and the questions used for the test scenarios. We also analyzed the tools relative to the taxonomy below in order to help further independent study. For our research please visit https://fairplayforkids.org/resources/ or https://[redacted]/ or https://cybersecurityfordemocracy.org/research.

TAXONOMY OF USER-FACING SAFETY TOOLS ON SOCIAL MEDIA

This taxonomy provides a structured framework for categorizing safety tools that are visible and actionable in the user interface (UI) or user experience (UX) of social media platforms. We do not include broader network security or safety efforts that have no user-visible component.

DIMENSION 1: USER TARGET

This dimension describes the user model of the feature or tool.

CATEGORY	DESCRIPTION
Individual	Designed for the individual user's self-management or protection.
Interpersonal	Manages interactions between users; often used to mitigate abuse or harm.
Supervisory	Enables oversight or controls by a third party (e.g. parents or guardians).

Examples:

- **Individual:** Screen time reminders, blocking undesired content, warnings about potentially violating actions
- **Interpersonal:** Blocking, muting, restricting contact
- **Supervisory:** Parental dashboards, approval workflows

DIMENSION 2: HARM APPROACH

This dimension describes the intended impact of the intervention.

CATEGORY	DESCRIPTION
Harm Prevention	Aims to stop harm before it occurs.
Harm Reduction	Focuses on reducing or mitigating harm after it has begun.

This category does not apply to tools with a 'Supervisory' user target, as these tools are not inherently harm-oriented.

Examples:

- **Harm Prevention:** Setting private profiles by default, nudges during risky actions
- **Harm Reduction:** In-app reporting, safety alerts after exposure to harmful content

DIMENSION 3: SAFETY SCOPE

This dimension describes the target of focus for the feature or tool.

CATEGORY	DESCRIPTION
Individual Safety	Primarily benefits the user enabling the tool.
Community Safety	Contributes to the overall safety of the platform and community.

This category does not apply to tools with a 'Supervisory' user target, as these tools are not inherently harm-oriented.

Examples:

- **Individual:** Blocking DMs, turning off comments
- **Community:** Reporting or flagging tools, feedback that reduces viral spread of harmful trends

DIMENSION 4: RISK CATEGORY (4 + 1 CS)

This dimension categorized the type of risk that the tool or feature aims to prevent or reduce. The 4 C's are a commonly used youth online risk framework developed by Livingstone et al.[1] A forthcoming work by Renkai Ma, Dominique Geissler, Stefan Feuerriegel, Tobias Lauinger, Damon McCoy, Pamela J. Wisniewski extends this framework to add Circulation risk, which we adopt as well. Finally, we also add Compulsivity risk, to describe tools intended to limit overall excessive or mis-timed usage.

RISK TYPE	DESCRIPTION
Content	Exposure to harmful content (e.g. graphic violence, misinformation)
Contact	Harmful interactions with other users (e.g. grooming, harassment)
Conduct	Risk from behavior (e.g. oversharing, bullying)
Contract	Commercial exploitation (e.g. manipulative ads, hidden purchases)
Compulsivity	Excessive or mis-timed use that is problematic or for which a user doesn't feel agency
Circulation	Inappropriate amplification of content outside of its user-intended context.

DIMENSION 5: IMPLEMENTATION STYLE

This dimension describes the default activity the tool is implemented with.

CATEGORY	DESCRIPTION
Default-On	Enabled automatically; requires no user action.
Prompted	Suggested or surfaced in context, but user must opt in.
User-Activated	Available for user to enable, but not suggested by default.
Needs Configuration	Available for user to enable, but requires non-trivial effort to configure, use, or maintain once activated.

Examples:

- **Default-On:** DM filters for unknown contacts.
- **Prompted:** In-app prompts to turn on a privacy setting.
- **User-Activated:** App usage dashboards in settings.
- **Needs Configuration:** Comment blocking based on user-configured block lists.

[1] Livingstone, S., & Stoilova, M. (2021). The 4Cs: Classifying Online Risk to Children. (CO:RE Short Report Series on Key Topics). Hamburg: Leibniz-Institut für Medienforschung | Hans-Bredow-Institut (HBI); CO:RE - Children Online: Research and Evidence. https://doi.org/10.21241/ssoar.71817

ABOUT THE AUTHORS

ARTURO BÉJAR



Arturo Béjar was the senior leader at Facebook responsible for engineering, product design, and research for security, safety, child safety, and customer care from 2009 to 2015. From 2019 to 2021, Arturo returned to Meta as a consultant to work on well-being issues at Instagram. During his second stint, Arturo led research, documented, and briefed the executive team on material harms experienced by teens on Instagram. It was the willful disregard of those harms that led Arturo to blow the whistle on Meta and advocate for the safety of young people online.

CYBERSECURITY FOR DEMOCRACY

Cybersecurity for Democracy is a research-based, nonpartisan, and independent effort to expose online threats to our safety, health, democracy and social fabric — and recommend how to counter those threats. We are a multi-university research project of the Center for Cybersecurity at the NYU Tandon School of Engineering and the Cybersecurity and Privacy Institute at the Northeastern University Khoury College of Computer Sciences. Learn more at cybersecurityfordemocracy.org.

ParentsSOS

PARENTS FOR SAFE ONLINE SPACES

Parents for Safe Online Spaces (ParentsSOS) is an educational initiative created by families who have lost children as a result of online harms. The initiative's goal is to raise awareness about the importance of the Kids Online Safety Act (KOSA), a piece of legislation addressing growing concern about the impact of online and social media platforms on children and teens. The initiative can be found on X as @Parents4SOS, on Facebook as Parents for Safe Online Spaces, and online at parentssos.org.

fairplay

childhood beyond brands

Fairplay is the leading US nonprofit committed to helping children thrive in an increasingly commercialized, screen-obsessed culture. Since our founding 25 years ago, we have grown from a small group of concerned parents, health professionals, and educators into a powerful force for children and families. Fairplay is truly an independent voice — we do not accept donations from Big Tech or any corporation. Our unique approach helps put kids' well-being first at home, in communities, and in corporate boardrooms. Learn more at fairplayforkids.org.

HEAT INITIATIVE

This report was produced with support from **Heat Initiative**, a nonprofit working to hold the world's most valuable and powerful tech companies accountable for failing to protect kids from online sexual exploitation.

DESIGNED BY OX CREATIVE