Broken, Buried, or Missing

ANATOMIES OF FAILURE (AND SUCCESS) OF SOCIAL MEDIA CHILD SAFETY FEATURES

AUTHORED BY:

Lexie Matsumoto, Arturo Béjar, Abdulraheem Arar, Damon McCoy, & Laura Edelson

PUBLISHED BY:

HEAT INITIATIVE

CRC CYBERSAFETY RESEARCH CENTER

Table of Contents

Introduction

Is social media safe enough for children? Today, parents must primarily rely on claims that social media companies make about their products to answer that question. But how accurate are these claims? To answer that question, we undertook a methodical independent analysis of features claimed to be provided by: Instagram, TikTok, Snapchat, and YouTube. We found that most features do not live up to the promises that companies make about the safety they provide. Of the 86 features that we tested, 51 failed, or ~60%. The rest of our report explains in detail how we conducted our analysis and detailed findings. Our goal is to provide a robust understanding of the gap between the promise and reality of social media safety features.

PLATFORM	TESTED	SUCCESS	FAILURE	FAILURE RATE
INSTAGRAM	29	10	19	**66%**
SNAPCHAT	11	3	8	**73%**
TIKTOK	24	12	12	**50%**
YOUTUBE	22	10	12	**55%**

Table 1: Summary of Evaluations by Platform

Some of the safety failures we discovered were critical and pervasive across multiple products. For example, every product we tested promised that children cannot search for dangerous content, and that such queries will be intercepted, blocked, and the child redirected to crisis resources. What we found was very different. On a TikTok account registered to a minor, once the account had searched for material about disordered eating and self-harm, TikTok's search stopped blocking such content and instead started suggesting it. It recommended that our teen account look up "anna food tips" and "how to pretend to eat your food," both drawn from pro-anorexia communities, alongside "mentally suffering," "losing yourself to mental health," and, most disturbing of all, "razor blade skin." These were the product's own recommendations, served to a child, not phrases we went looking for. Instagram's version of the same feature failed differently but no less plainly: as our tester began typing "eating disorder," the autocomplete offered back the deliberate misspellings that pro-eating-disorder communities use to evade the very blocklist the feature depends on. In each case, the bypass took under three minutes to find. On Snapchat, critical safeguards meant

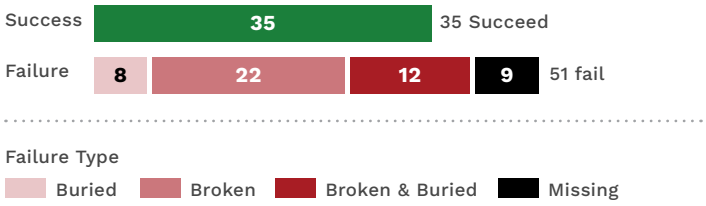
> ***All of the 10 conduct tools we tested (the safeguards meant to govern how users treat one another, such as detecting and preventing cyberbullying) failed, on all four products.***

to prevent adults from messaging children they are not connected to failed during our testing. We responsibly disclosed this and other critical vulnerabilities to social media companies prior to the publication of this report.

We audited features relative to the claims that social media companies make about them. Specifically, we evaluated each feature along two dimensions: whether it actually functions as described, and whether a child can realistically reach it. We count a feature as a success only if it does both. We also measured the ways features fail, which fall into three groups. First, some of these tools (nine total, or about one in ten) could not be triggered at all. Despite following the steps the company describes, we found no evidence they existed, and we refer to these as missing. Second, a further 34 were not functional: the tool existed, but it either failed in a way that defeated its purpose or could be circumvented with trivial, non-technical effort. Third, a further eight were effectively unreachable, buried under layers of settings, or so cumbersome to configure that it is unreasonable to expect a child to use them. In all, 35 of the 86 features both functioned as claimed and reached the child they were meant to protect.

SAFETY FEATURES ACROSS ALL PLATFORMS

Figure 1: Overall Success and Failure by Type



Children's safety online is an issue of Children's safety online is an issue of broad public concern, and as cybersecurity measurement researchers, we see providing the public with better information about the scope and scale of online risks as our responsibility. In addition, several of this report's authors are, ourselves, parents. If you are a parent, you should know that we have found systemic issues with the design and implementation of many of these features. For example, all of the 10 conduct tools we tested (the safeguards meant to govern how users treat one another, such as detecting and preventing cyberbullying) failed on all four products. Tools meant to curb compulsive use, like screen-time limits and break reminders, fared little better, with fewer than one in three succeeding. There were some bright spots, which we discuss in detail in this report, because they show what good safety design and implementation can look like.

> Safety features, in order to be effective, need to be on by default or easy to activate, be resilient to normal teenager use, and should demonstrably protect against harm.

Understanding the implementation quality or design effectiveness of current features is the first step. We do not assess the likelihood that a specific harm will happen, as this is best measured by survey instruments (Béjar 2025). Instead our testing focuses on 'Does this tool function in normal use?', the equivalent of 'Does the airbag go off when the car hits the wall?' Future research is still needed to assess questions like, 'Would this airbag design prevent harm?' or 'Does the way the airbag deploys protect the driver?'

We believe that from a regulatory and academic perspective, online safety should be treated with the same rigor as we treat cybersecurity, privacy, and effective safety processes that are used in other fields. Independent testing, and ensuring accuracy in how companies represent the effectiveness of their tools are part of the foundation needed to begin the work to define what an effective safety-by-design framework will need. We hope that this is the first of many studies of this kind. There remain many important unanswered questions about both risks to users and the safety of social media products that this auditing approach can answer. We also hope that social media companies will both learn from these findings and use our results to make their products safer for children. Finally, we hope our findings will inform the public conversation around social media safety and accountability. While it is clear to us that many of the safety features we studied were deeply flawed and fell far short of the promises that social media companies have made to their users, deciding what to do about that is a conversation we all need to have together.

While it is clear to us that many of the safety features we studied were deeply flawed and fell far short of the promises that social media companies have made to their users, deciding what to do about that is a conversation we all need to have together.

WHAT HARMS EXIST?

We claim that safety tools are necessary, but this makes an assumption that risk exists in the social media products we are testing. Research shows overwhelmingly that children using social media face harms associated with use (Livingstone and Stoilova 2021; Ma et al. 2025; Riehm et al. 2019; Staksrud et al. 2013), and the US Surgeon General released a report (Office of the Surgeon General 2023) about the potential risks stemming from use of these products.

Even during insulated testing where we interact only with new profiles and other researcher accounts, we experienced harms on child accounts. We cannot say from our own experiences in the course of this research how often harms occur, but it is clear that they are not hypothetical, and that when they do occur the consequences can be irreversible. This is why the basic effectiveness of these tools matters so much. When a company assures a parent that a feature will block harmful searches, keep strangers from messaging their child, or stop an image from spreading, the parent has little choice but to take that assurance on trust. A safeguard that quietly fails to do what it promises leaves a real risk in place, and the family that relied on it has no way of knowing.

> ***Even during insulated testing where we interact only with new profiles and other researcher accounts, we experienced harms on child accounts.***

How We Analyzed Safety Features

2.1 SAFETY FEATURE AUDITING

To evaluate whether companies are actually delivering on their child safety promises, we performed an audit of advertised features on Instagram, Snapchat, TikTok, and YouTube. We evaluated how features behave under realistic conditions rather than rely solely on how the companies describe them. Similar to crash-testing a car to assess real-world safety performance, we engaged in user behaviors meant to trigger these features and assessed their performance.

2.1.1 FEATURE IDENTIFICATION

We began by systematically reviewing press release communications on the newsroom and safety websites of each tested product. This allowed us to compile a comprehensive list of advertised safety features. We examined each article for references to child safety features, and from the relevant text, extracted features. Features were included in our dataset if they (a) describe a specific tool, feature, or restriction and (b) mention youth safety or target a youth-associated problem. This included features such as time limits, search restrictions for harmful content, and default private accounts. This did not include more generic safety features not specifically related to children, such as two-factor authentication or data security protocols.

We also only considered tools that are available to every teen on the platform–tools that require an additional supervisory account to activate were excluded, though there is some overlap with features a teen can opt into and ones supervisors can impose. We tested only the teen-activated aspects of these features and did not evaluate restrictions imposed by supervisory accounts.

We classified features across five distinct dimensions: Risk Category, Product Surface, Feature Oversight, Implementation Style, and UI Bound (where on the app the feature is); we describe this process in detail in the Appendix A.2. Importantly, we assigned each feature an associated harm it is meant to address. Livingstone and Stoilova (2021) developed the “4C’s,” widely used by researchers to define online risks to children: Content, Contract, Contact, and Conduct. We further extend this framework with two additional categories to become the “6C’s.” Circulation risk captures harm associated with the distribution and dissemination of a minor’s content beyond the intended audience and was identified by Ma et al (2025). We further add Compulsivity risk to address mis-timed (such as at night or during school hours) or excessive use of the product by minors.

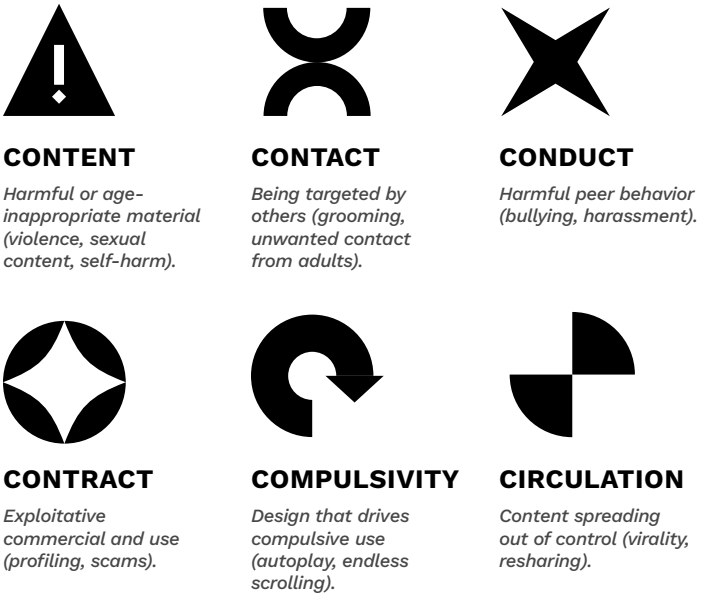
We show these 6C's in Figure 2 and further define them in Table C in the Appendix. *Content* is the risk associated with exposure to age-inappropriate content; *Contact* covers any harm as a result of interactions with other users (commenting, DMing, harassment); *Conduct* risk is the result of harmful behavior by the user; *Contract* is any commercial, monetary, or data exploitation of a minor user; *Circulation* is harm associated with distribution of an underage user's content beyond the intended audience; and finally, *Compulsivity* risk is use that is excessive or disruptive to the life of the user.

During testing, we encountered risks during testing that we did not specifically perform actions to encounter. These can be used as examples to illustrate what harm children have the ability to be exposed to.

Instagram has a mechanism that suggests accounts for the user to follow, which can be found in multiple points in the app. While performing a test that required us to be in the "Find Followers" portion of the user interface (UI) for a teen account, we were recommended only profiles of obviously adult (non-peer) men, with the exception of one account that did not have a profile picture, as shown in Figure 2 (names anonymized for privacy). This is a clear example of a *Contact* risk in which a child is suggested connections with adults they do not know, in our case, overwhelmingly so.

THE 6C'S OF ONLINE RISK FOR CHILDREN

Figure 2a



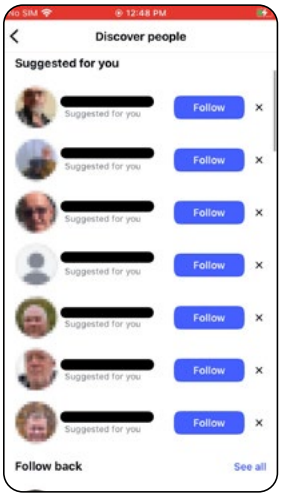


Figure 2b

> Follows suggested by Instagram to a 13 year old girl test account, illustrating Contact risk.

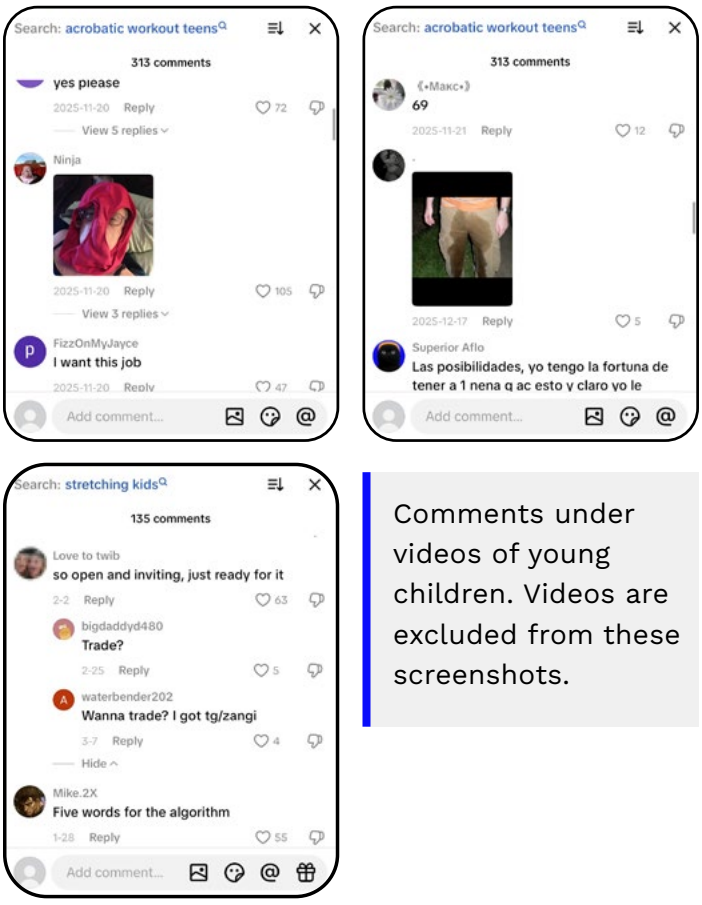
Additionally, during a test on TikTok, we were served a TikTok of young girls participating in a gymnastics lesson. The video itself did not contain graphic content and showed a class of young girls, clearly below 18, stretching and using standard gymnastics equipment. We were using an underage male account at the time.

In the comments of this video, we found several disturbing comments that prompted us to investigate similar content. We performed a follow-up search for benign gymnastics content and found that harmful comments were prolific on videos showing young girls performing routines or exercises. Some of these comments are shown in Figure 3, all under videos of young girls. We show an interaction where an account commenting a sexual allusion about the children in the video was replied to by other users asking to "trade," likely in reference to exchanging illicit underage material.

On TikTok, users are able to comment with images and short videos. We viewed images and videos of exposed male genitalia, allusions to sex, and photos of known pedophiles (such as Jeffrey Epstein) under these TikToks. We show two examples of these images, one being an adult man with undergarments across his face and another with a man wearing wet pants, referencing a sex act. This is an example of a *Circulation* risk: by the dozens of TikTok comments sections we observed, we know that people interested in watching, commenting, and leaving sexual imagery under young girls performing gymnastics were able to somehow source these TikToks and interact with them.

CIRCULATION RISK

Figure 3



Comments under videos of young children. Videos are excluded from these screenshots.

Assigning the associated 6C's risk with each feature allows us to perform analysis beyond simply pass/fail criteria. Rather than just asking "does this feature function as claimed," we are able to answer "does this feature function as claimed, and what harm does it leave unaddressed if it doesn't?" At a product level, this also allows us to understand structural patterns in how safety features are implemented across their safety infrastructure as a whole. For example, we can assess whether tools addressing Contact are more reliably implemented than those addressing other risks. This additional step of identifying harm risks moves our evaluation from a feature-by-feature accounting to analysis based on actual user risk.

2.1.2 TEST ACCOUNTS

To test these features, we created researcher accounts on each product. These profiles were either child or adult accounts, depending on the requirements of the test case.

Child accounts were mostly created with birthdays between the ages of 13 and 17. These accounts were created using default privacy and safety settings to represent how a teen would typically begin using the product. Additionally, one tested feature was intended for users under 13. To test this feature, we created test accounts with birthdays under 13. We used exploratory but non-malicious use patterns, simulating reasonable ways a teen would use the product.

Adult accounts were created with birthdays of 25 years or older. Generally, adult accounts were created to test restrictions on interactions with child accounts, so our use patterns here simulated malicious actors. However, we did not employ any specialized "attacks" or use resources outside the product. Any failure of a feature was through utilization of the product's native UI.

2.1.3 TEST DEVELOPMENT

For each identified feature, we developed a structured testing scenario. Each scenario described the conditions under which the feature should activate, how many accounts were involved, and the expected result. Scenarios were designed around three different perspectives: (1) a child using the product normally and encountering the feature, (2) a teen attempting to work around a restriction on their account, and (3) a malicious adult actor attempting to bypass protections on a teen account. This ensured testing reflected plausible real-world cases of use by actual users. Tests were performed using established auditing methodologies (Sandvig et al. 2014; Boeker and Urman 2022). All findings were documented through screen recordings and screenshots.

2.2 EVALUATION FRAMEWORK

To assess the safety features advertised by Instagram, TikTok, Snapchat, and YouTube, we developed a structured framework that is applied consistently across all products. Each feature we identified was evaluated along two dimensions: whether it functions as claimed, and whether it is accessible to users. We derived an answer to each question from a set of sub-questions, described below.

Before evaluating a tool, we determined **is this feature triggerable?** That is, can this feature actually be activated or observed through product testing? Some features that are described in company press releases were not able to be triggered despite following criteria to trigger the product as described in press releases. A typical user would not encounter them, so in a realistic scenario, they cannot be said to work. If we are unable to trigger a feature, we evaluate it to be a failure, but do not assess it for functionality or accessibility since we never observe it during testing.

MISSING FEATURE CASE STUDY

INSTAGRAM: PROMPT TO RETHINK

Instagram advertises that in an effort to prevent bullying, users who post comments with harmful or offensive language will be prompted to rethink posting the comment with a pop-up interactive screen. Despite commenting explicit words, clear bullying and intolerant phrases (shown in Figure C1), we were never shown a prompt to reconsider posting. Comments were posted from a teen account under a post from another teen account.

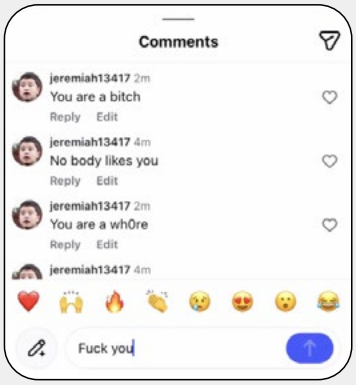


Figure C1

2.2.1 DOES THE FEATURE FUNCTION AS CLAIMED?

To determine if a feature is actually functional in preventing or mitigating harm, we test it for basic operability and resistance to trivial accidental or intentional circumvention. When a feature fails to operate or can be trivially circumvented, we determine where in the process first accounts for the failure. A feature can both be designed poorly and have a bad technical implementation, but in these cases we attribute this to a *design* failure since no implementation can compensate for a bad design. A feature that is well designed but is technically fragile fails at implementation, and a feature sound in both design and implementation but can be bypassed fails at circumvention.

1. Is this feature designed well?
Some features' failures occur because their basic design is such that the failure was inevitable. In effect, we evaluate whether the feature was *designed to fail* as such design decisions completely remove its ability to function as a safety tool.

2. Is this feature implemented well?
A theoretically well-designed feature can still fail if it is not technically implemented correctly. This question is evaluating the robustness and reliability of features in practice. Implementation failures include features that work in some UI bounds but not others, or that are easily broken by normal usage patterns.

3. Is this feature resistant to circumvention?
Safety features, especially in the context of child safety, must be resilient against bypass attempts, whether it be by the child themselves or a bad actor. A feature that can be trivially bypassed through non-skilled circumvention strategies provides only the illusion of protection. This question considers if the feature can be evaded with minimal effort, and features that fail this test were considered non-functional.

2.2.2 IS THIS FEATURE ACCESSIBLE?

A feature is only useful when it is activated; if a tool is strenuous to configure or difficult to find, it is inaccessible and thus offers limited real-world protection. This is particularly true for safety features intended for use by children. To evaluate accessibility, we ask two questions, both of which must be satisfied for a feature to be considered accessible.

Is this feature easy to turn on or set up?
We assess whether or not a typical user (in our case, a child) could be reasonably expected to activate and configure the feature without significant difficulty. Features buried far into setting menus or that require multiple non-trivial steps were considered inaccessible due to friction.

Is this feature on by default, or does the product prompt users to enable it?
Even a simple, useful, functional feature can go unused if users do not know it exists. We asked whether or not companies are enabling relevant features automatically for children, or whether the product prompts users to turn it on. Features that rely on users to discover and activate them independently, with no proactive enrollment or prompting, are considered inaccessible.

2.3 EVALUATION RUBRIC

Combining the two evaluation dimensions of functionality and accessibility, together with the triggerability check above, each feature is assigned one of five outcomes.

➔ **Successful.**
A feature is *successful* if it is functional and accessible. It can be triggered, is well designed, is well implemented, and is resistant to circumvention. It is easy to set up and surfaced to users.

➔ **Buried.**
A feature is *buried* if it is functional but fails one or both accessibility criteria. This feature works when activated, but the product does not surface it in a way that makes real-world use likely.

➔ **Broken.**
A feature is *broken* if it is not functional but is surfaced to users. It may be default on or prompted to users for visibility, but does not deliver meaningful protection in practice.

➔ **Broken and Buried.**
A feature is nonfunctional and hidden if it fails to pass functionality testing and is not surfaced to the user.

➔ **Missing.**
A feature is *missing* if it could not be triggered during testing. The product advertises the feature but we found no evidence of it on the product under the conditions the company describes.

A feature is considered a **Failure** if it falls into any of the four non-working categories: Buried, Broken, Broken and Buried, or Missing. Features that are functional and accessible are considered **Successful.**

> ***A feature is considered a Failure if it falls into any of the four non-working categories: Buried, Broken, Broken and Buried, or Missing. Features that are functional and accessible are considered Successful.***

BURIED FEATURE CASE STUDY

INSTAGRAM: AD PREFERENCES

Instagram provides teens a mechanism to indicate which topics they would prefer to see fewer advertisements about. This feature is never prompted to be used, and can only be found after navigating through five screens: Profile → Settings → Account Center → Ad Preferences → View All Topics, at which point the user arrives at a screen with a search bar. The teen must now enter the name of topics they prefer to see less advertising on; however, there is no browseable list of topics or indication of what they may be named. Instead, the teen must independently decide what phrase to search for and hope the search returns a matching category.

The search results themselves add another layer of friction. In our test, we emulated a teen who wanted to see less advertisements about food. She searched “Food” and was returned numerous overlapping categories such as “Foods,” “Food & Drink,” “Food Festivals,” “Food Photography,” “Food Humor,” and “Food & Drink Topics,” among others seen in Figure C2. It is not clear if opting out of the general “Foods” category has any effect on others. A teen navigating this interface may have to exclude themselves from every subtopic for full coverage, and multi-select is not an option.

It is also worth noting that this feature promises only to reduce advertisements on a given topic, not fully eliminate them. This feature has a complicated navigation path to find and an unintuitive search interface, along with ambiguous categories to choose from. This renders the feature effectively inaccessible to the average user, despite being available and functional.

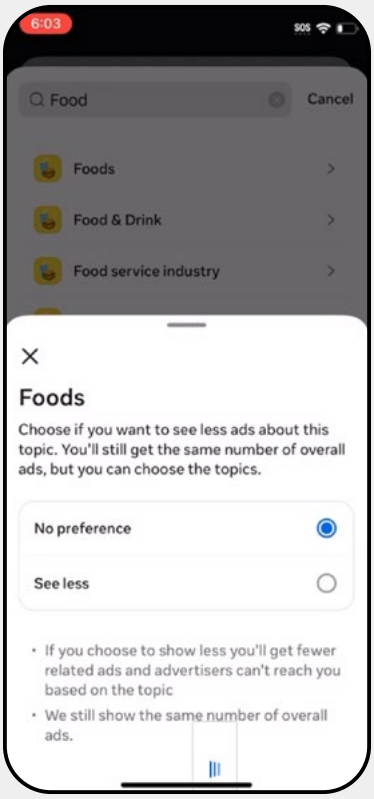


Figure C2

2.4 RESULTS

Given our evaluation strategy, we document the results of our audit in Figure 4.

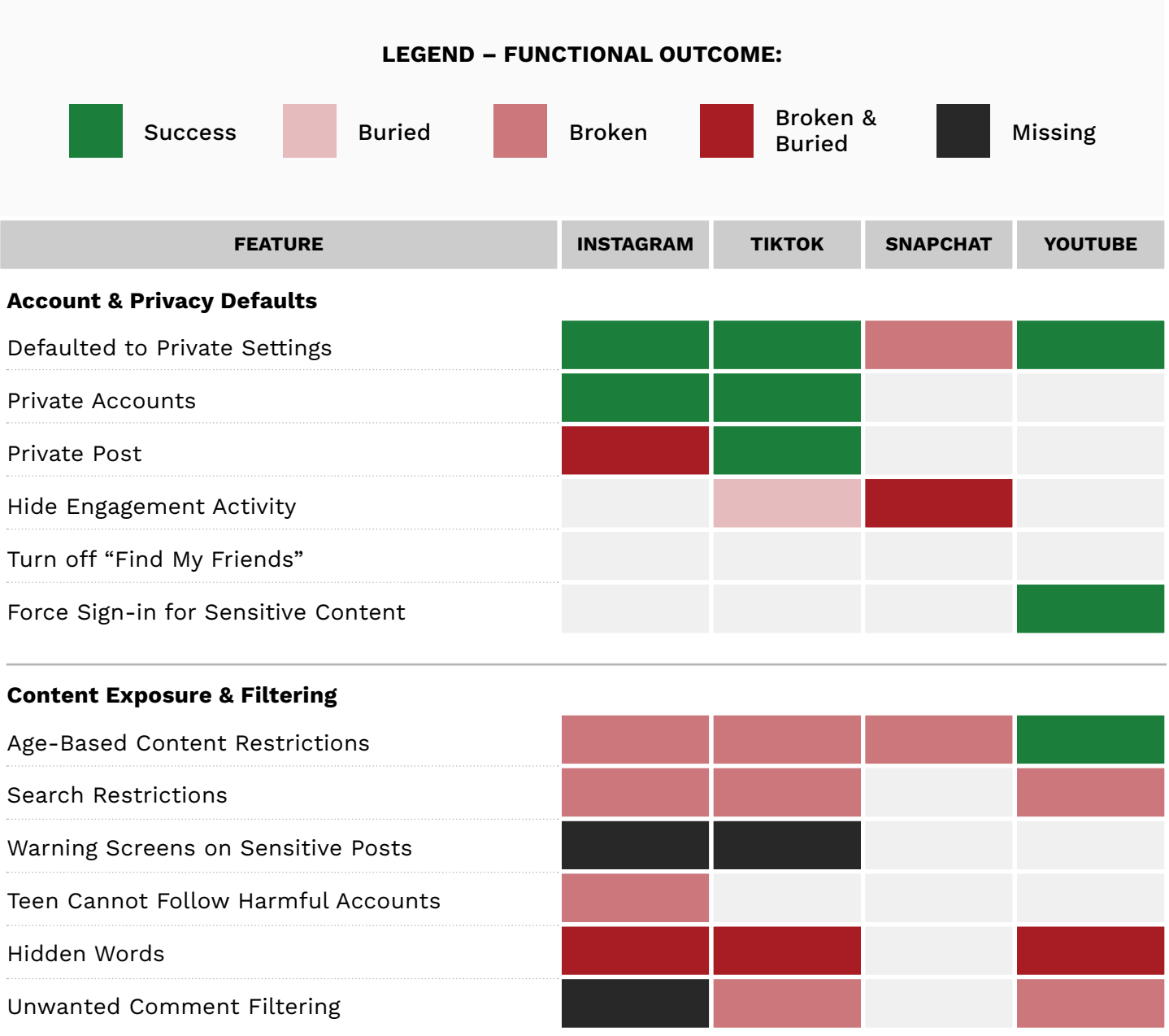


Figure 4

LEGEND – FUNCTIONAL OUTCOME:

Success | Buried | Broken | Broken & Buried | Missing

FEATURE	INSTAGRAM	TIKTOK	SNAPCHAT	YOUTUBE
Interpersonal Controls				
Blocking / Unfollowing	■	■	■	■
Block All Accounts	■		■	
Mute User	■	■		
Remove Follower		■		
Restrict User	■			
Disable Comments	■			■
Bulk Delete Comments	■	■		
Suspicious Account Warnings	■		■	
Prompt to Review Connections			■	
Messaging Restrictions				
Message Initiation Restrictions	■		■	
Disable Messaging		■		
No Images in Comments/DMs		■		
Cannot Screenshot Short-Term Messages	■			
Content Creation & Circulation				
Unable to Go Live	■	■		■
Restrict Downloads		■		
Cannot Save to Watch Later				■
Self-Tag Content as Inappropriate				■
TikTok for Younger Users		■		

Figure 4

LEGEND – FUNCTIONAL OUTCOME:

Success | Buried | Broken | Broken & Buried | Missing

FEATURE	INSTAGRAM	TIKTOK	SNAPCHAT	YOUTUBE
Screen Time & Compulsivity				
Time Limits	Broken	Broken		
Prompts to Take a Break	Broken	Broken		Broken
Do Not Disturb	Success	Broken		
Bedtime Reminders				Broken
Time Activity Dashboard	Success			Buried
Limit on Shorts				Broken & Buried
Turn Off Autoplay				Success
Reduce Notifications (Digest)				Buried
Turn Off Watch History				Buried
Advertising & Monetization				
Advertising Restrictions	Buried	Broken & Buried		Success
Virtual Gifting / Payments		Success		
Conduct & Reporting				
Resource Redirection	Broken	Broken	Broken	Broken
Prompt to Reconsider Comment/DM	Missing			
Report Underage Account	Broken & Buried			
Hide Like Counts	Broken & Buried		Buried	
Location Sharing Restrictions			Success	

Figure 4

Instagram

> We identified 29 features that were claimed to be implemented on Instagram and were testable. This makes it the largest single-product dataset in this audit.

Of Instagram's 29 features, we were unable to trigger 5 during testing; despite performing the conditions needed to activate the feature, we found no evidence of their existence on the product and are considered nonfunctional by default. These were Warning Screens on Sensitive Posts, Unwanted Comment Filtering, Suspicious Account Warnings, Prompt to Report, and Prompt to Reconsider. The remaining 24 features were triggerable and subject to evaluation.

3.1 FUNCTIONALITY PERFORMANCE

We first wanted to understand how many of these advertised features are Successful. A feature is successful if it is functional and resistant to basic circumvention. Of Instagram's testable 29 features, only 12 met the standard for functionality. The remaining 17 failed at some point during our evaluation due to design, implementation, or circumvention issues.

Instagram's non-functional features were distributed roughly evenly across the stages of our evaluation: 5 could not be triggered at all, 5 failed at the design stage, 4 failed at implementation, and 3 could be circumvented through trivial workarounds.

Five features failed at the design stage itself; these were designed poorly before implementation was even considered. Prompts to Take a Break, Bulk Delete Comments, Hide Like Counts, and Time Limits all failed here. When a teen user is prompted to take a break after some period of time, the design of the feature includes an option that lets the teen snooze the feature indefinitely through a button displayed prominently on the screen. In each case, the underlying concept of the feature reflects an inadequate design choice.

A further four features were designed well in principle but broke down in practice under normal use. Sensitive Content Control, and Search Restrictions all fall into this category. For example, restricting harmful searches in practice is good design, but Instagram's implementation of this tool failed dramatically. It appears to operate on a blocklist system where specific phrases are matched and blocked when submitted; however, Instagram's own generated search recommendations provide bypasses that suggest workarounds to the user as they begin to type harmful phrases.

Three features could be easily circumvented. Hidden Words, Resource Redirection, and Screenshotting Short-Term Messages could all be circumvented through basic strategies. Hidden Words and Resource Redirection could be bypassed by misspellings or alternative phrases to the intended blocked term. We found that Android devices could begin screen recording and document messages sent in Vanish Mode without notifying both parties in the chat.

SUCCESSUL FEATURE CASE STUDY

INSTAGRAM: MESSAGE INITIATION RESTRICTIONS

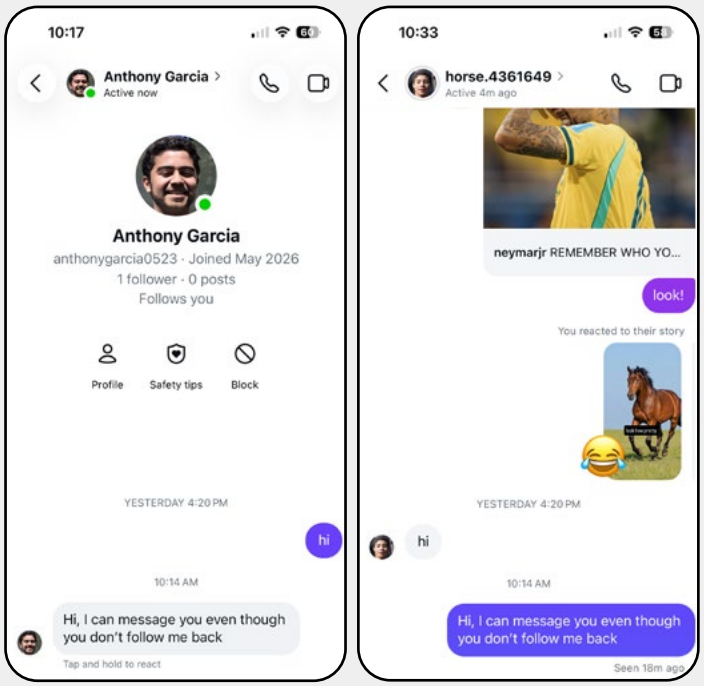
Instagram's press releases promise specific restrictions between adults messaging teens, which we evaluated as Successful but compromised. One press release states teen accounts can "only receive messages from accounts they follow or have previously connected with." However, if a child initiates a conversation with an adult they don't follow, the unconnected adult is then able to message the child completely unrestricted. That is, an adult can message a child that does not follow them back. In testing, our teen account was able to message an adult account they did not follow without any warnings. Despite Instagram's claims that the accounts needed to be "previously connected" for the child to receive messages, this was possible without previous interactions beyond the adult viewing and liking the child's Story. In another press release, Meta states that "we restrict people over 19 years old from sending private messages to teens who don't follow them." Again, we find that this is not strictly true; the adult is able to send messages to a child unrestricted after contact has been initiated, even if the child does not follow them back.

Despite these two scenarios, we define this feature Successful because it did prevent an adult from initiating contact with a minor account. We include this case study to make it clear that our evaluation relies on language in press releases, which in this case, is not clear.

We do not know what Meta considers a "previous connection," nor do they define it in their press releases. This is problematic for parents who rely on published company claims to make informed decisions for their child and to set their expectations for what harms they could encounter. We also use this as a warning that while we may determine features to be Successful, this does not mean that they are fully able to prevent harm on a product.

ADULT-TO-CHILD MESSAGING ON INSTAGRAM

Figure C3



C3a) Child POV: Chat from the perspective of the child, who can only see messages after they have initiated the chat.

C3b) Adult POV: Chat from the perspective of the adult, who is able to see previous attempts at reaching the child before the child initiates the message.

3.2 ACCESSIBILITY PERFORMANCE

Accessibility asks whether a feature is easy to set up and whether it is surfaced to the user, either by being defaulted on or actively prompted. Of Instagram's 24 triggerable features, 17 were accessible, and 7 were not.

3.3 EVALUATION

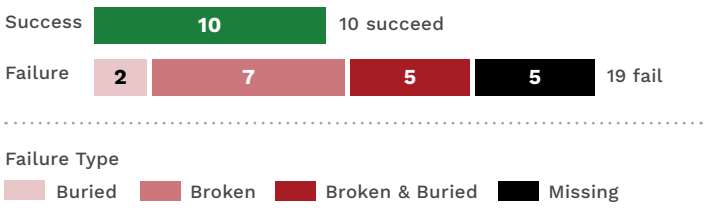
As described in Section 2, we consider successful features to be ones that are both functional and accessible. Of the 29 Instagram safety features we evaluated, ten met this standard. These features are: Default to Private Settings, Private Accounts, Inability to Go Live, Do Not Disturb Mode, Blocking/Unfollowing, Restrict Mode, Mute Mode, Time Activity Dashboard, Message Initiation Restrictions, and Block All Accounts. The remaining 19 failed, and they did so in distinct ways.

Five are *missing:* we could not trigger them under realistic testing, so a teen using Instagram would never encounter them at all. The rest failed among the 24 features we could observe and therefore scored on both axes.

The largest single failure group, seven features, were *broken*: a teen could find them and switch them on, but they did not do what they claimed. A further five were *broken and buried,* failing on both axes at once, and two were *buried:* functional, but not surfaced or simple enough that a teen would realistically enable them. Taken together, 12 of the 29 features were functional and 17 were not, while only ten delivered protection that a teen both received and could reach. Figure 5 shows the distribution of features across the success and various failure cases.

INSTAGRAM SAFETY FEATURES

Figure 5

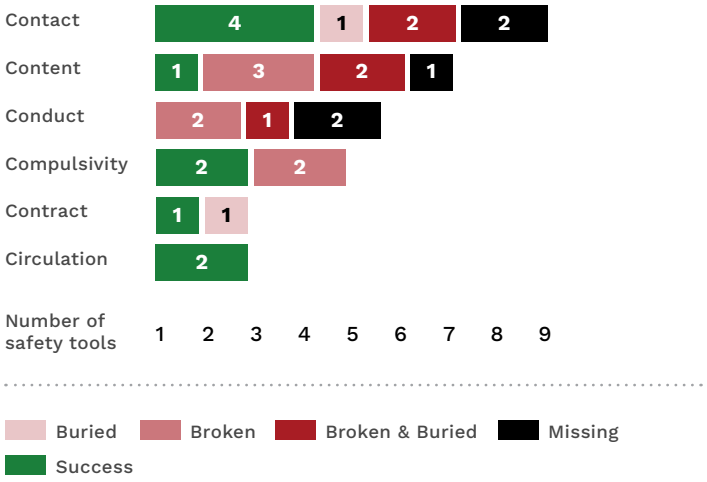


On Instagram: one pattern stands out among the features that fail on accessibility: the problem is more often one of discoverability than configuration. Of the 24 triggerable features, six were never surfaced to the user, against four that were hard to set up once found. Safeguards largely do exist and are not difficult to configure. Instead, they were buried behind many pages of settings and teens were simply never shown them.

Grouped by the six risk categories, success was distributed unevenly. Circulation features were the most successful(both of its 2), while Compulsivity (2 of 4), Contract (1 of 2) and Contact (3 of 9) had more middling success rates. The starkest gap between promise and reality was in Conduct. These are the features that govern how users treat one another, such as prompts to reconsider before sending an abusive message, and none of the 5 features in this category succeeded, either because they were broken, buried, or missing. Content performed almost as poorly, with only one of 7 succeeding. Evaluation results for each risk category are shown in Figure 6.

INSTAGRAM SAFETY FEATURES BY RISK CATEGORY

Figure 6



Snapchat

Snapchat has the smallest feature set of the companies in this audit with 11 features that we tested. Of those 11 features, three could not be triggered during testing (Message Initiation Restrictions, Suspicious Account Warnings, and Review Connections) and are considered *missing;* all failed to trigger under conditions they are advertised to activate. The remaining eight features were triggerable and we were able to further evaluate their results.

MISSING FEATURE CASE STUDY

ADULT CHILD MESSAGING

Snapchat promises that adults cannot find or initiate messages with underage accounts. During our testing, we created both an adult and child profile. From the adult profile, we were able to directly search for, find, and then message the child account with zero restrictions. The child account received the friend request and upon accepting it was able to view the history of messages that the adult had sent them with no warnings. This can be seen in Figure C4.

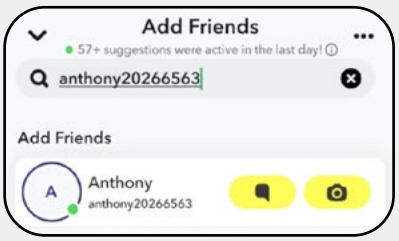


Figure C4

4.1 FUNCTIONALITY PERFORMANCE

> Of Snapchat's 11 testable features, four were functional and seven were not. We labeled every feature as well-designed, but they broke down at either the implementation or circumvention stage.

Three features failed on technical implementation: Defaulted to Private Settings, Age-Based Content Restrictions, and Turn Off "Find my Friends." In each case, the feature activated but was not robust enough to follow through on the intention of the feature. For example, Turn Off "Find my Friends" promises that your account will not be shown in the recommended friends section on other profiles. In our testing, however, the account with this setting turned on was recommended to a user who had their number saved in the phone's contacts.

The sole feature that fails via circumvention is Resource Redirection which promises to respond to dangerous search queries with the relevant help lines and resources (disordered eating help, suicide hotline). This feature is well implemented in narrow conditions (e.g., searching for "suicide"), but is easily bypassed in a realistic scenario of misspelling or using numbers instead of letters when typing the term (i.e., searching for "su1c1de").

BROKEN AND BURIED FEATURE CASE STUDY

SNAPCHAT: TURN OFF "FIND MY FRIENDS"

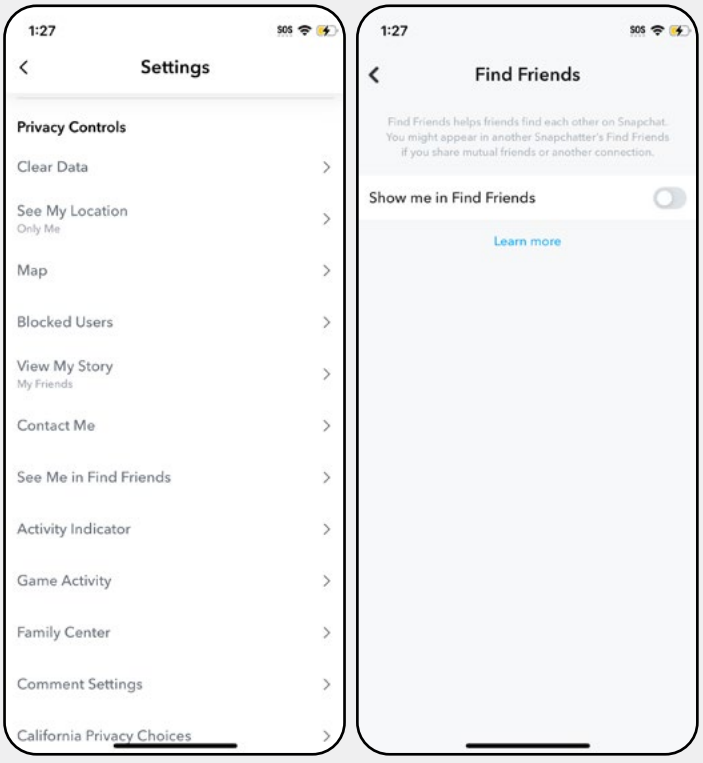
Snapchat offers a setting that allows a user to choose to not be recommended to others via the product's "Find My Friend" suggestion system. This toggle is located deep in the Settings page, past dozens of other toggles, and is not prompted as an option during account setup or any other time during our hours of testing.

In our testing, this feature did not work as described. We configured one test account to opt out of showing in "Find My Friends." We set up another account that had the original account's associated phone number in its Contacts, then allowed Snapchat to view Contacts. We monitored this account for incoming friend suggestions. Within 24 hours, the opted-out account appeared in the second account's Suggested Friends as a top recommended friend. These accounts had not otherwise been connected, yet the recommendation was made regardless. This is shown in Figure C5.

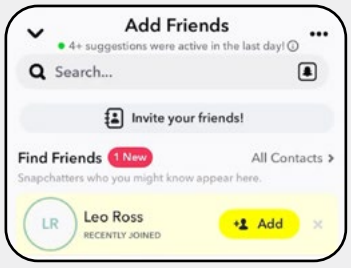
This failure is particularly relevant to child safety. Friend recommendation systems serve as a primary vector through which adults can connect and establish contact with minors on social media. A feature that promises to remove a user from such recommendations but does not do so reliably provides a false sense of security and protection. A minor who has taken a deliberate step of enabling this feature by navigating a dense settings page would reasonably believe they are no longer discoverable. The combination of inaccessible placement and unreliable implementation makes this a consequential failure in the scope of child safety.

SNAPCHAT FIND MY FRIENDS FEATURE

Figure C5



C5a) Settings Page *C5b) Setting Turned Off*



C5c) Friend Recommendations

4.2 ACCESSIBILITY PERFORMANCE

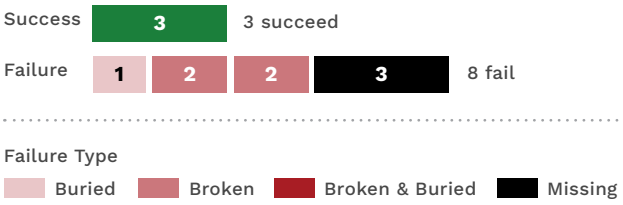
We find that 5 of Snapchat's 8 triggerable features are accessible, and 3 are not. All three of the inaccessible features are never surfaced to the user, meaning there is no prompt, default enrollment, or in-app indication that the feature exists.

4.3 EVALUATION

Of the 11 features evaluated, only three were successful, being both functional and accessible: Blocking/Unfollowing, Block All Accounts, and Location Sharing Restrictions. The remaining eight failed.

SNAPCHAT SAFETY FEATURES

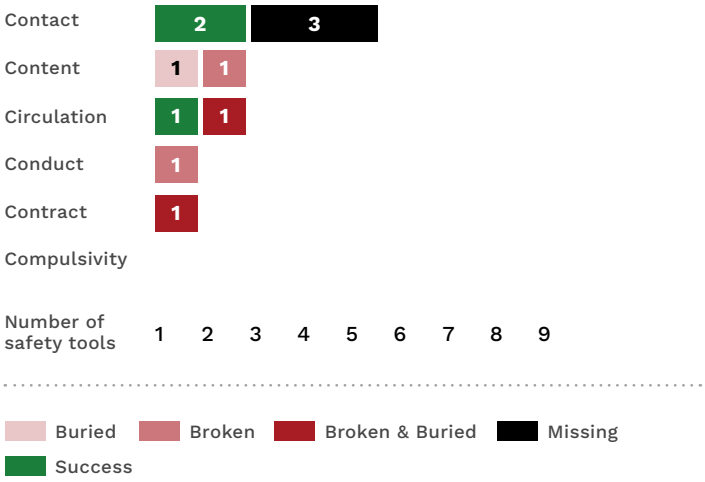
Figure 7



Three of Snapchat's youth safety features were missing and could not be triggered at all. Among the features we observed, 2 were *broken* (accessible but nonfunctional), 2 were *broken and buried* (failing on both dimensions), and 1 was *buried* (functional but never surfaced, namely Hide Like Counts). Taken together, four of the 11 features were functional and seven were not, while only three delivered protection that a teen both received and could reach. Figure 7 shows the distribution of features across the success and failure cases.

SNAPCHAT SAFETY FEATURES BY RISK CATEGORY

Figure 8



Grouped by the six risk categories, Snapchat's few successes were confined to two categories. Contact accounts for most of them (2 of 5), with one further success in Circulation (1 of 2). Conduct, Content, and Contract contained no working features, and Snapchat offered no compulsivity-related safety tools at all, so screen-time and overuse harms were unaddressed by Snapchat's safety toolkit. Given how few features sit in each category, we read these as illustrative rather than as reliable per-category rates. Evaluation results for each risk category are shown in Figure 8.

TikTok

We identified 24 features on TikTok that were testable. This is the second-largest set of features on one product, behind Instagram.

Of these 24 features, one was missing: Warning Screens on Sensitive Posts. Despite searching for (and finding) harmful content, we never encountered a warning screen. The remaining 23 features were triggerable and evaluated.

5.1 FUNCTIONALITY PERFORMANCE

We evaluated which of TikTok's features were functional; of the 23 triggerable features we audited, 13 met the standard for functionality of good design and robust technical implementionn.

Poor design was the reason for failure of four features on TikTok. Time Limits, Prompts to Take a Break, Advertising Preferences, and Bulk Delete Comments all fell into this category. Bulk Delete Comments allows a user to delete multiple comments at once; however, when using this feature, the user had to select each comment one at a time. This exposes potential victims to multiple occurrences of content they don't want to see, and in the case of hundreds or thousands of comments, puts too much onus on the user to filter through. A simple design change would be a "Select All" option, or the ability to search for specific phrases in comments to select them that way.

Three additional features had good design but failed in implementation. Age-Based Content Restrictions, Search Restrictions, and Do not Disturb are well-conceived, but all broke down in practice. Do Not Disturb is advertised to block notifications during specific time periods that can be set by the user. In our testing, despite being in the Do not Disturb time window, a child account was sent a notification. We were able to repeatedly verify this test by dismissing notifications and receiving a new one.

BROKEN FEATURE CASE STUDY

INSTAGRAM, SNAPCHAT, TIKTOK: SEARCH RESTRICTIONS

Instagram, Snapchat, TikTok, and YouTube all promise that underage users cannot search for harmful content on their products. They claim that these searches will be intercepted and content will be blocked from users, then they will be redirected to help resources if needed. We found that YouTube did not surface harmful content in our testing, but would allow harmful search terms to be bypassed by clicking a button to confirm you may be exposed to sensitive material.

On Instagram, TikTok, and Snapchat, simply misspelling or not finishing the search query

Continues...

BROKEN FEATURE CASE STUDY

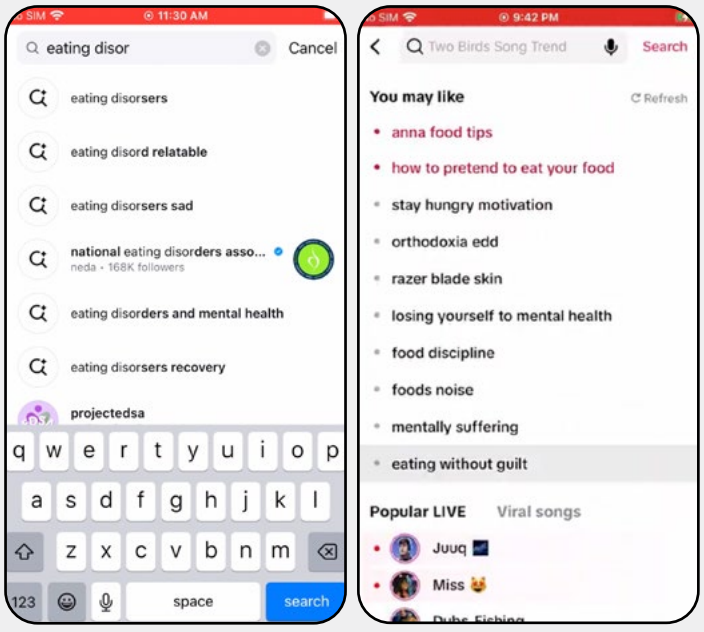
query (e.g., typing “eating dis” instead of “eating disorder”) was enough to bypass the restrictions. Snapchat’s search was especially brittle in this regard, with “eating disorde” surfacing content but “eating disorder” not showing results (additionally, no resources). On Instagram and TikTok, we found that as we typed in search terms, the built-in search recommendations would suggest misspellings, alternative words, and phrases that would bypass their own restrictions. Examples of these are shown in Figure C6. Every time we clicked on a suggested search, we were shown content relevant to the query; not a single one was restricted. For example, on Instagram, the suggested autocomplete for “eating disor” was a misspelling of “eating disorder,” “eating disorsers,” which is used by the disordered eating community to bypass blacklists and access harmful content that would be restricted. Additional recommended autocompletes are “eating disord relatable” and “eating disorsers sad.”

On TikTok, we saw similarly shocking results. After performing some of our tests, TikTok’s search “learned” what type of content we were searching for (disordered eating, self-harm, and suicide content). It suggested we search for “anna food tips” and “how to pretend to eat your food” as top searches, both relating to the anorexia eating disorder. They also recommended depression terms like “mentally suffering” and “losing yourself to mental health.” Most shockingly, TikTok recommended for our teen account to search for “razor blade skin,” referencing the self-harm action of cutting or a suicide method. Every single recommended search, shown in Figure C6, is related to either disordered eating, self harm, or depression.

This feature is considered broken because of how technically incomplete it is on Instagram, TikTok, and Snapchat. It seemed that the products were filtering based on keywords, not what material surfaced as a result of the query; additionally, it did not seem that their search recommendations have restrictions against the suggestions they generate. Our testing identified this problem and lack of coverage within minutes; for each test, it took under three minutes to find and be recommended harmful searches. For a child using the product and attempting to find harmful content, it takes little effort to be able to bypass search restrictions and encounter sensitive material related to disordered eating, self harm, and even suicide.

BYPASSING PROMISED SEARCH RESTRICTIONS

Figure C6



20a) Instagram *20b) TikTok*

Three features were circumventable, and thus failed. Hidden Words, Unwanted Comment Filtering, and Resource Redirection were all well-designed and technically sound but were vulnerable to non-skilled workarounds that occurred in ordinary use. Hidden Words and Comment Filtering shared the same vulnerability; both were bypassed by using leetspeak substitutions (using numbers in place of letters or symbols) which is a common and low-effort tactic.

5.2 ACCESSIBILITY PERFORMANCE

Accessibility asks whether a feature is easy to set up and whether it is surfaced to the user. Of TikTok's 24 tested features, 19 were accessible.

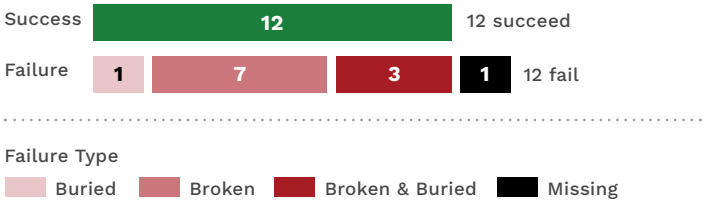
Advertising Preferences, Hidden Words, Bulk Delete Comments, and Hide Engagement Activity were not default-on or prompted and required users to independently discover them. This is particularly troubling for Hidden Words, which additionally required users to manually curate a personal blocklist of harmful terms.

5.3 EVALUATION

Considering functionality and accessibility together, 12 of TikTok's 24 features succeeded. That is the largest number of working safeguards on any product we examined, exactly half of those tested.

TIKTOK SAFETY FEATURES

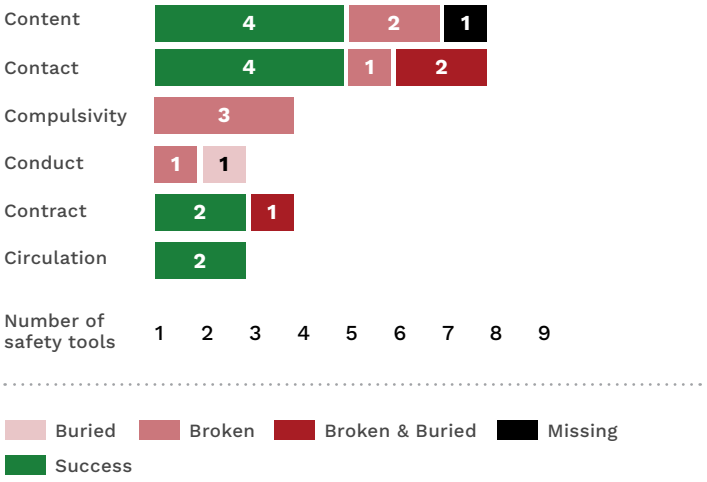
Figure 9



Beyond the one *missing* feature, seven were *broken,* one was *buried,* and three were *broken and buried.* In total, 13 of the 24 features were functional and 11 were not. Figure 9 shows the overall evaluation.

TIKTOK SAFETY FEATURES BY RISK CATEGORY

Figure 10



The risk breakdown echoes the other products, with successes split across Content (4 of 7), Contact (4 of 7), Contract (2 of 3), and Circulation (both of its 2). The conspicuous gaps are Conduct, where none of its 2 features succeeded, and Compulsivity, where all 3 failed. The latter is the same screen-time story seen above: the tools exist on paper but deliver nothing in practice. Figure 10 gives the per-category results.

YouTube

We identified 22 features that were testable on YouTube. Unlike the other companies in this audit, all of YouTube's promised features were triggerable, and were all evaluated fully.

6.1 FUNCTIONALITY PERFORMANCE

> Of YouTube's 22 features, 14 were considered functional. The remaining eight failed at some point in the evaluation.

Three of these features failed in their design. Prompt to Take a Break shared the same conceptual flaw as Instagram and TikTok with a built-in snooze mechanism that allowed the user to dismiss the prompt immediately. Limit on Shorts and Bedtime Reminders had the same problem, with the Shorts Limit prompt even displaying a link directly to settings where the user can turn it off.

Two features failed at implementation. Search Restrictions did not fully block the user from searching harmful queries and allowed the child user to bypass the restricted screen to view potentially harmful content. Prompt to Reconsider Comment did show with an explicit comment, but did not recur upon posting future harmful comments if dismissed once.

Three features were technically well implemented but were circumventable. Unwanted Comment Filtering, Hidden Words, and Resource Redirection were vulnerable to basic workarounds. Like TikTok and Instagram, Comment Filtering and Hidden Words were bypassed through leetspeak substitutions and Resource Redirection failed against misspellings or alternate phrasing of crisis-related searches.

6.2 ACCESSIBILITY PERFORMANCE

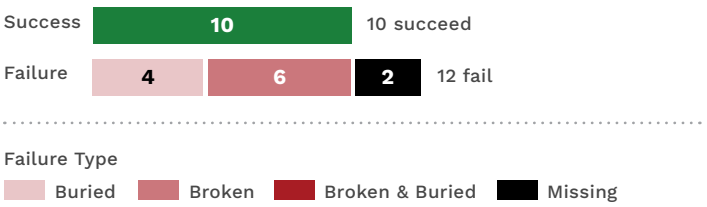
Of YouTube's 22 features, 16 were accessible and six were not. YouTube's accessibility failures followed a consistent pattern where the feature worked but was not surfaced to the user; all six of the inaccessible features (Hidden Words, Time Activity Dashboard, Disable Comments, Reduce Notifications, Turn Off Watch History, and Limit on Shorts) were not prompted or defaulted to on. One feature (Hidden Words) was additionally hard to set up once enabled.

6.3 EVALUATION

When considering overall performance, 10 of YouTube's 22 features succeed.

YOUTUBE SAFETY FEATURES

Figure 11

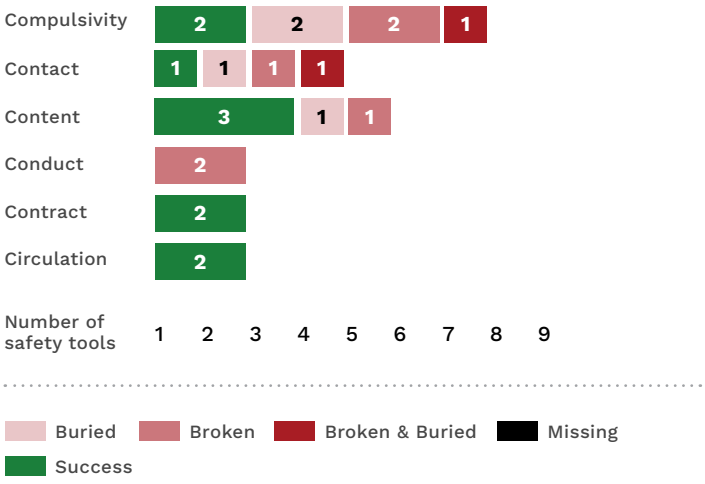


With no missing features, the failures were sorted into six broken, four *buried,* and two *broken and buried*. Broken tools were still the single largest group, but burial was YouTube's distinctive failure mode: the gap between the 14 features that worked and the 10 that also reached the user was four, the widest such gap in the audit, and every feature in it was buried. Figure 11 shows the distribution.

By risk category, YouTube's successes were spread more evenly than elsewhere, appearing in Content (3), Compulsivity (2), Circulation (2), and Contract (2), with one in Contact. Notably, its compulsivity successes, such as turning off autoplay and disabling save-to-playlist, were among the few working screen-time tools anywhere in the audit, matched in number only by Instagram. Conduct was again the gap, with zero successes. Figure 12 shows the per-category results.

YOUTUBE SAFETY FEATURES BY RISK CATEGORY

Figure 12



Discussion

Parents, doctors, legislators, and even the US Surgeon General have expressed concern about the safety of social media for children for years (Office of the Surgeon General 2023). Over the same period, the companies responded with safety tools, and they advertised them heavily: of the 2,574 press releases we collected to conduct this study, roughly one in eight describes a child-safety feature. As researchers, such strong public messaging left us with one question: **do the safety features companies advertise to families do what they say they do? For most of the tools, the answer is no.** Nearly three in five of the advertised features we tested are missing, broken, buried, (51 of 86), with only 35 that both functioned and reached the user.

Our findings document a significant disparity between advertised claims and the reality of usage on social media products for underage users. Press releases and advertising are the primary channel through which companies make promises to the public, including parents, regulators, and teens themselves about what protections exist. A parent who takes the company at their word about restricted adult-to-minor messaging, screen time limits, or bullying comment countermeasures will have a false sense of security.

They will know the promises of the product but not the real experience of their children as they interact with broken and buried safety tools. The press releases, in essence, function to protect or improve public reputation rather than to effectively protect minors. However, while the current state of child safety tools is poor overall, bright spots in the overall landscape do exist, and we will explore them below.

A parent who takes the company at their word about restricted adult-to-minor messaging, screen time limits, or bullying comment countermeasures will have a false sense of security.

7.1 FAILURE BY DESIGN

Frequently in our testing, features failed by design rather than by faulty implementation, meaning that the company did not design it in a way that it could be effective even if it was implemented perfectly. Examples include the screen-time limits that every product lets a teen dismiss with a tap, and the "Hidden Words" feature Instagram, TikTok, and YouTube rely on a child inputting all instances of harmful language to function. These are not bugs that slipped through testing. They are design decisions that give the appearance of a safeguard while leaving the underlying behavior available, and a teen or an adult encounters no real obstacle in avoiding them.

The same conclusion follows from how often the identical failure recurs across products. When a comment filter can be evaded by spelling a word with a zero in place of an "o," and that same trivial bypass works on Instagram, on TikTok, and on YouTube, the problem is not that robust filtering is unsolved. It is that none of the three companies invested the minimal effort that would be required to solve it.

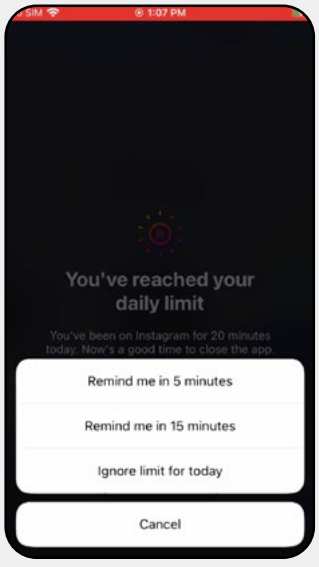
BROKEN FEATURE CASE STUDY

INSTAGRAM, TIKTOK, AND YOUTUBE: TIME LIMITS

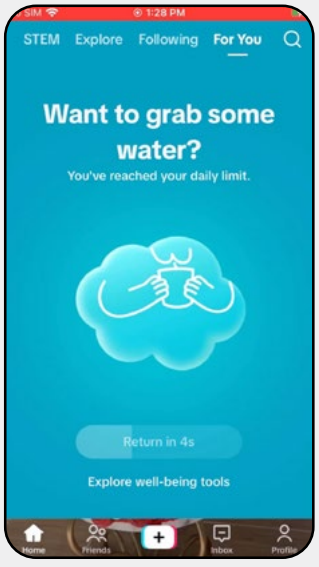
Three products (Instagram, TikTok, and YouTube) had a daily screen time limit of one hour after which users were prompted to leave the app; Snapchat did not have any feature like this. On all three platforms, this feature was default-on and accessible, and the actual functionality of the feature was implemented correctly. In a sense, the feature worked as described.

TIME LIMITS ON INSTAGRAM & TIKTOK

Figure C7



C7a) Instagram



C7b) TikTok

BROKEN FEATURE CASE STUDY

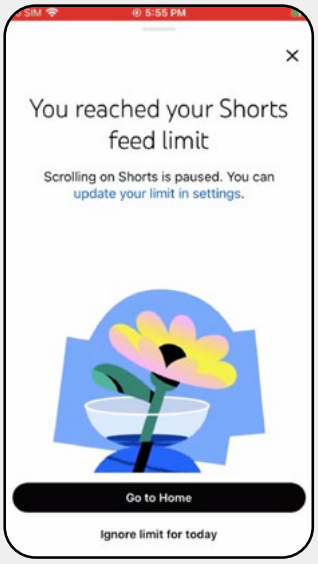
The failure of this tool was not in its implementation but rather in its inherent design. Each feature's prompt informing the user that their time limit was up prominently featured a mechanism through which the teen could ignore the limit and bypass the warning; Instagram and YouTube allowed this immediately while TikTok had an arbitrary five second waiting period before the user could continue. When the limit is reached, teens were presented with a large button on the UI and/or an X in the top corner, which dismissed the snooze setting. YouTube had a similar feature specific to YouTube Shorts, where a user was notified after they had reached some time limit specifically watching short-form content; this prompt is especially direct, including a link to the settings page where the time limit can be disabled entirely. These prompts can be seen in Figure C8.

This pattern of failure is consistent on all three products and points to an intentional design choice. A time limit that a teen is prompted to can dismiss with no hard shut-out does not meaningfully constrain usage. It functions instead simply as a notification of elapsed time. This feature is representative of a failure mode where a tool functions correctly but is structured in a way that undermines its own protective function.

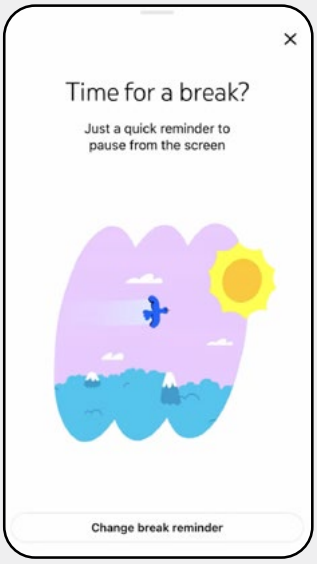
An effective way to limit usage on minor accounts would be through required parental authorization to override time limits, a functionality that each of the three products have demonstrated capability of applying in other features, but have not applied here.

TIME LIMITS ON YOUTUBE

Figure C8



C8c) YouTube Shorts



C8d) YouTube

7.2 SUCCESS BY DESIGN

The purpose of any audit is not only to understand the anatomy of failure, but also to understand the shape of success. The tools that worked are as instructive as the ones that failed, because they can show us concretely what success looks like and how we might improve other features that fall short. The successful features are not cleverer than the failed ones; instead they typically share a simpler profilee.

Two design patterns accounted for most of what worked, and both could be applied broadly to safety tools across all risk categories. The first is defaulting to the safest path. Instagram setting teen accounts to private at creation, or YouTube requiring a signed-in, age-bearing account before known sensitive content loads, protect the user before any choice is made, and so clear the discoverability hurdle that buries so many other tools. The second, and the more instructive, is removing risks rather than ineffectively reducing them. Most of the failures we documented are failures of containment: a filter that misses a spelling, a search restriction that is recommended away, a friend recommendation that slips in, each an attempt to permit an action that can expose the user to risk and then attempts to catch the bad cases after the fact.

TikTok for Younger Users takes the opposite approach, replacing the standard product with a constrained environment in which open search, messaging, and the algorithmic feed simply do not exist. A restriction cannot be circumvented if the capability it guards was never offered, so this design sidesteps the entire class of circumvention failures that impacts the patch-based safety features.

Another way to look at these features is that most of them are not safety features, but harm reduction features. There are harms that are intrinsic to the core of these products when used by minors: compulsivity, contact from strangers, harmful conduct, etc. These harms could be addressed by designing the product for the intended audience, rather than trying to reduce harms intrinsic to products designed for adults.

The lesson for policy is that safety by design requires a different approach. Safer systems can be achieved by reducing the risk surface rather than filtering a dangerous experience. It is also worth noting that the patterns that work appear to be the ones that tend to reduce engagement, while the patterns that fail tend to preserve it.

SUCCESSFUL FEATURE CASE STUDY

TIKTOK: TIKTOK FOR YOUNGER USERS

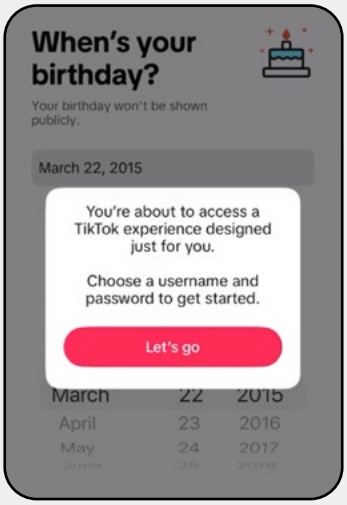
TikTok for Younger Users is a separate in-app experience that any user under 13 years old is automatically enrolled in, as shown in Figure C9. This mode restructures the standard TikTok interface in ways that address multiple categories of risk. Content that is shown is manually curated rather than algorithmically sourced, which reduces the risk of exposure to age-inappropriate material; some of these categories are shown in Figure 16. The entire experience is view-only, so users cannot comment, message other accounts, or share content in ways that would expose them to contact or conduct risks from other peers. Users are able to create their own TikTok videos that are stored locally on the device, not published or stored by the company. Additionally, a mandatory one-hour screen time daily limit is enforced; users are able to extend this time by 30 minutes by using a parent-configured passcode.

This experience also constrains content discovery. Users are not able to perform open ended searches and are instead allowed to browse different topic categories that are defined by TikTok. This design constricts the ability for users to encounter content outside the curated environment, a common failure mode of search functionality that can serve as a vector for accessing age-inappropriate material.

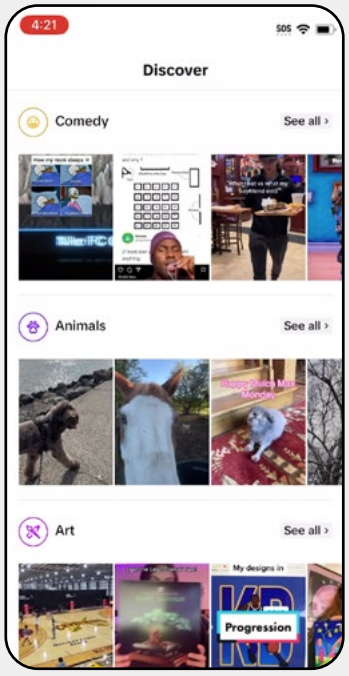
TikTok for Younger Users represents a structurally-distinct approach to child safety, one that does not use automated filtering on a product that hosts harmful material, but rather by curating and limiting the experience through ground-up constraints.

TIKTOK FOR YOUNGER USERS ON NEW ACCOUNT

Figure C9



C9a



C9b

SUCCESSFUL FEATURE CASE STUDY

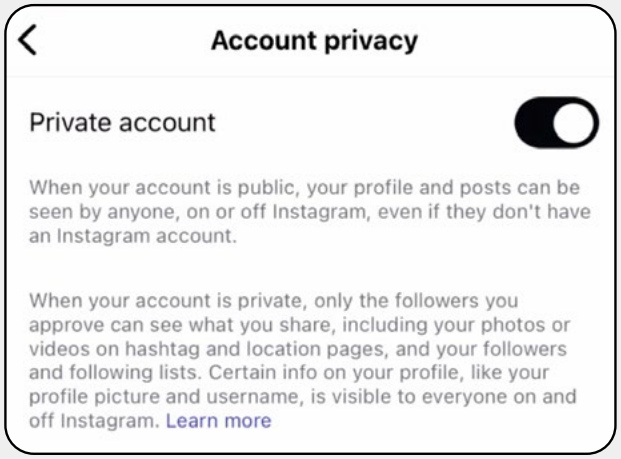
INSTAGRAM: DEFAULTED TO PRIVATE ACCOUNTS

When a minor creates an account on Instagram, it automatically configures the account as private. This restricts the content on the profile to approved followers. This default implementation represents one of the more straightforward implementations of privacy-by-default design, where users are proactively set to the most private configuration at account creation. In our testing, this feature performed as intended across all evaluation criteria. New accounts were consistently defaulted to private accounts, and profile content remained hidden unless the teen approved the follower, as seen in Figure C10.

Default settings are known to be sticky, meaning that users rarely change them once set; research has shown this effect with users either keeping default settings or choosing options proximal to them (Cho et al. 2019). For teen users in particular who may lack contextual awareness to evaluate the privacy and safety concerns of a public account, the default setting could encourage users to maintain private accounts without requiring the user to make an informed decision.

INSTAGRAM DEFAULT PRIVATE ACCOUNT SETTING ON NEW ACCOUNT

Figure C10



7.3 FAILURE BY BRITTLENESS

While some failures are conceptual, it is also clear that some safety features failed simply because they were poorly implemented. We discuss a clear example of this in Section 9.2: TikTok's Do Not Disturb feature allowed a single notification to stay active on the phone even if received during the DnD time frame. It blocked subsequent notifications, but if the user dismissed the sole notification, another took its place. This seemed to be an implementation bug that could be caught and fixed in a quality assurance phase of production. Search restrictions on Instagram and TikTok were bypassed through the product's own recommendation tools; simple in-app testing of common harmful search phrases (such as "eating disorder") could reveal suggested searches that were immediate and pervasive, like we discovered in our testing.

To be clear: we believe that the development teams behind these features are highly skilled, our goal in sharing examples like this to ensure those teams have the support they need to ship and maintain high quality features. However, poor quality of implementation of some safety features clearly impacts their effectiveness.

7.4 FAILURES BY RISK CATEGORY

The failures are not evenly distributed across the kinds of harm the tools address. Not one Conduct feature–the tools governing how users treat one another–both worked and was easily available to child users on any product in our audit. Compulsivity, the category covering screen time and compulsive use, is barely better served: only four such tools succeeded across all four products, and two of them carried no working compulsivity tool at all.

These are perhaps the harms often discussed in public debate about youth and social media, and they are where the market has most conspicuously failed to protect children.

The distribution of failures across harm categories is worth examining not just as a record of what was broken, but to ask the question of why? By success rate, Conduct and Compulsivity stand alone: they are the only two categories in which fewer than one in three tools both worked and could reach a child, at zero of 10 and four of 14 respectively. Unfettered connection building between users drives growth, and time-on-product is the main economic incentive. We do not claim to know the decisions that influenced the design and implementation of these safety tools, but the lack of effective tools for Conduct and Compulsivity are consistent.

7.5 SAFETY AS A STANDARD

The foundation of our online lives relies on three pillars: security, privacy, and safety. Security is about preventing unauthorized access. Privacy is about reliable clarity about who has access. Safety is about what happens once people have access to each other in an online environment. Youth in particular are a critical user group, as we rely on safety features to ensure that the products we give our youth are "safe enough."

The companies we audited invest considerably in cybersecurity and data privacy with dedicated security teams and commission third-party security testing and support for ethical hackers. They face legal liability when a user's data is compromised and take strides to prevent that from happening. There is an implicit standard for these practices, one that is taken more seriously than what we have observed in our audit of child safety features.

No equivalent standard currently exists for child safety features on an industry or legal level. Our findings document what this oversight means for children who use these social media products. Approximately three in five features failed, despite being advertised on public press releases–yet, there is no accountability. A broken child safety feature can exist alongside state-of-the-art recommendation algorithms (Meta 2025; Snapchat, n.d.; TikTok, n.d.; YouTube, n.d.).

> Companies that can make products that are sophisticated enough to predict a teen's every preference in real time are also capable of producing robust safety features for those products. Implementing robust tools is not a matter of technical feasibility but rather technical priority.

Online safety is as important as security and privacy, and as such requires a rigorous and independent evaluation approach. The processes that have been developed to test the claims companies make about security and privacy provide the foundation for an online safety discipline: testing scenarios, penetration testing, and adversarial attempts that document how these products actually behave. In many other fields, effective processes to regulate and measure safety are in place, and they are applicable to online safety. Safety, and especially child safety, deserves to be treated as a third pillar of company responsibility alongside security and privacy, and held to the same standard of rigor.

This means safety features should undergo independent quality assurance testing. The vast majority of our tests failed within minutes or even seconds, not after intense and complicated testing procedures. This standard is not unreasonable to impose. Crash tests for vehicles do not ask whether the automaker "promised" that the airbag would deploy, they verify that it will and measure the impact on the occupant in a real-world setting. When a data breach exposes private user information, companies face regulatory measures and investigations, mandatory disclosures, and legal liability, regardless of if the vulnerability was a design or implementation flaw. Child safety features on products with tens of millions of underage users should face no less scrutiny. There needs to be ongoing testing and accountability processes that go beyond the initial press cycle and announcement. The methods outlined in the audit equip independent auditors and other researchers with replicable, structured scenario testing done relatively inexpensively. For academics and companies alike, the ability to test these features exists, but the expectation doesn't. We are hoping to change that.

7.6 LIMITATIONS

There are several limitations to our findings that readers should understand. First, we did not measure which product is "safest", and the counts in this report should not be read as a leaderboard. Second, we measured whether tools do what companies claim, not whether children are measurably safer. A feature we marked successful is one that functions as advertised and is reachable, which is a necessary condition for protection but not proof of it. Third, each social media product does not have equal risk surfaces, which affects coverage, but not whether advertised tools work. A product built around ephemeral messaging, location, and friend-finding presents a different contact surface than one built around video consumption, and the appropriate set of tools would differ accordingly. Fourth, this is a point-in-time audit, conducted between December 2025 and June 2026. Products change, and a feature we found broken may be repaired while one we found working may break.

7.7 WHAT'S NEXT

Taken together, the picture of child safety tools on social media is not one of an industry struggling with an intractable problem. Rather, it is one of an industry that has learned to announce safety features but not deliver or maintain them. These safety feature announcements are typically in response to public scrutiny around the latest demonstration of the potential harms to young users of its products. Industry does this in a setting where companies are often not held accountable for the actual success or failure of those safety tools, or indeed the underlying risks of its products. However, risks to children on social media are very real (Twenge et al. 2022; Orben and Przybylski 2019), and so when safety tools fail, the harm is also very real. As researchers, we have now had the experience multiple times of reporting on failures of current tools, only to be immediately met with promises of new tools that reportedly protect users against the same risks as those current, failing, tools. What we have not seen is an honest grappling with why risks to children exist in the first place and why existing tools may be failing.

It is our hope for this research that by identifying overall patterns of failure (and success) and by proposing robust frameworks for evaluation, social media companies can move beyond the current cycle of critical safety failures being addressed with safety theater that quickly degrades until the next critical failure. All users, but especially young users, need social media products that are safe, and safety tools that are both functional and accessible. Better, safer social media for young users is possible. But much work remains to be done in order to achieve such a goal. As independent researchers, we are committed to this work and hope that social media companies will join us in this effort to ensure that their products are safe for all users.

ACKNOWLEDGEMENTS

The authors are grateful to Yaël Eisenstat for her feedback and editing on drafts of this work. We are also grateful to Thaddeus Kopczynski for assistance and support.

REFERENCES

[1] Béjar, Arturo. n.d. *Teen Accounts, Broken Promises: How Instagram Is Failing to Protect Minors.* https://fairplayforkids.org/wp-content/uploads/2025/09/Teen-Accounts-Broken-Promises-How-Instagram-is-failing-to-protect-minors.pdf.

[2] Béjar, Arturo. 2025. "Understanding the Harm Teens Experience on Social Media: A Systematic Approach to Mitigating Negative Experiences Online." *Queue* 23 (4): 24–46.

[3] Boeker, Maximilian, and Aleksandra Urman. 2022. "An Empirical Investigation of Personalization Factors on TikTok." *Proceedings of the ACM Web Conference 2022* (New York, NY, USA), WWW '22, 2298–309. https://doi.org/10.1145/3485447.3512102.

[4] Cho, Hichang, Sungjong Roh, and Byungho Park. 2019. "Of Promoting Networking and Protecting Privacy: Effects of Defaults and Regulatory Focus on Social Media Users' Preference Settings." *Computers in Human Behavior* 101: 1–13. https://doi.org/https://doi.org/10.1016/j.chb.2019.07.001.

[5] Livingstone, Sonia, and Mariya Stoilova. 2021. The 4Cs: *Classifying Online Risk to Children.* CO:RE Short Report Series on Key Topics. Leibniz-Institut für Medienforschung | Hans-Bredow-Institut (HBI). https://doi.org/https://doi.org/10.21241/ssoar.71817.

[6] Ma, Renkai, Dominique Geissler, Stefan Feuerriegel, Tobias Lauinger, Damon McCoy, and Pamela Wisniewski. 2025. *Analyzing Social Media Claims Regarding Youth Online Safety Features to Identify Problem Areas and Communication Gaps.* https://arxiv.org/abs/2512.14965

[7] Meta. 2025. "Instagram Feed Recommendations AI System." In *Instagram* Feed Recommendations. https://transparency.meta.com/features/explaining-ranking/ig-feed-recommendations/; Meta.

[8] Office of the Surgeon General. 2023. *Social Media and Youth Mental Health: The U.S. Surgeon General's Advisory.* U.S. Department of Health; Human Services. https://www.hhs.gov/sites/default/files/sg-youth-mental-health-social-media-advisory.pdf.

[9] Orben, Amy, and Andrew K Przybylski. 2019. "The Association Between Adolescent Well-Being and Digital Technology Use." *Nature Human Behaviour 3* (2): 173–82. recommendations/.

[10] Riehm, Kira E., Kenneth A. Feder, Kayla N. Tormohlen, et al. 2019. "Associations Between Time Spent Using Social Media and Internalizing and Externalizing Problems Among US Youth." *JAMA Psychiatry* 76 (12): 1266–73. https://doi.org/10.1001/jamapsychiatry.2019.2325.

[11] Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms." D*ata and Discrimination: Converting Critical Concerns into Productive Inquiry.*

[12] Snapchat. n.d. "How We Rank Content on Discover." In *Snapchat Support.* Https://help.snapchat.com/hc/en-us/articles/8961631424020-How-We-Rank-Content-on-Discover; Snapchat.

[13] Staksrud, Elisabeth, Kjartan Ólafsson, and Sonia Livingstone. 2013. "Does the Use of Social Networking Sites Increase Children's Risk of Harm?" *Computers in Human Behavior* 29 (1): 40–50. https://doi.org/https://doi.org/10.1016/j.chb.2012.05.026.

[14] TikTok. n.d. "How TikTok Recommends Content." In *Support.tiktok.com*. Https://support.tiktok.com/en/using-tiktok/exploring-videos/how-tiktok-recommends-content; TikTok.

[15] Twenge, Jean M, Jonathan Haidt, Jimmy Lozano, and Kevin M Cummins. 2022. "Specification Curve Analysis Shows That Social Media Use Is Linked to Poor Mental Health, Especially Among Girls." *Acta Psychologica* 224: 103512.

[16] YouTube. n.d. "Recommendations on YouTube." In *Recommendations.* Https://www.youtube.com/howyoutubeworks/recommendations/; YouTube.

Appendix

A.1 DATA COLLECTION

For Instagram, Snapchat, TikTok, and YouTube, we need to understand what features they claim to have available for children. To do so, we collect and extract features from all press releases on the official newsroom websites of Instagram, Snapchat, TikTok, and YouTube from 2019 to the present (n=2,574).

After collection, two independent coders annotated 100 releases per product (400 total) as irrelevant or relevant to teen safety features. We achieved a Cohen's Kappa of .81, indicating strong agreement. Given this, one coder completes the remaining articles.

Of all articles scraped, we find that 12.3% of all articles are relevant to child safety features. Table A contains the breakdown of articles by product; we find that Instagram has the highest concentration of relevant articles with over 30% of all press releases being relevant to child safety tools.

RELEVANT ARTICLES PER PRODUCT

Table A

PRODUCT	RELEVANT	IRRELEVANT	TOTAL
INSTAGRAM	150	347	496
SNAPCHAT	59	938	997
TIKTOK	45	423	468
YOUTUBE	64	549	613
TOTAL	**318**	**2257**	**2574**

We then analyze relevant press releases and extract a list of over 88 features across 4 products. Two coders independently extract the name of the feature, its general description based on article language, and the product for which it was advertised. Features were included in the final dataset if they described a specific tool, restriction, or setting and mentioned youth safety or a youth-associated problem. Generic safety measures, like two-factor authentication or data privacy protocols, were not included.

A.2 FEATURE CLASSIFICATION TAXONOMY

To enable consistent comparison across features and products, we developed a taxonomy that classifies features across five dimensions: Risk Category, Product Surface, Feature Oversight, Implementation Style, and UI Bound. These dimensions were applied to each feature before testing began so evaluations could be grounded in what the feature was designed to do. The taxonomy also informed how test scenarios were constructed, with Product Surface and UI Bound dimensions determining where on the product the test should be constructed and Feature Oversight influencing how many accounts were needed in the test.

Each dimension was used to determine a question about a feature's design: *where* it is on the product, *who* are the users it affects, *what* risk does it address, *when* does it intervene, and how it is activated. By assigning a value to each dimension a feature can be characterized in a way that informs the construction of a test case. We discuss this mapping in Section A.4.

A.2.1 PRODUCT SURFACE

Product Surface defines what part of the product design the feature is designed to operate. Product Surfaces are not exclusive in practice as some features may expand across multiple surfaces (e.g., time limits should work across *all* surfaces); as such a single feature may be assigned multiple Product Surfaces for which multiple tests may need to be performed.

We identify seven Product Surfaces, as described in Table B, what users see (Content Exposure), how users create and share (Content Creation), interactions between users (Interpersonal Communication), time and usage patterns on the product (Usage Controls), how visible users are to others (Account Visibility), how money is exchanged on the product (Monetization), and how supervisory roles influence a teen's experience (Parental Supervision). Parental Supervision is considered a cross-cutting area rather than a discrete surface since, generally, supervisory features do not operate in isolation.

CATEGORY	DESCRIPTION
Content Exposure	Viewing, searching, or recommendations of content; includes feeds and search results.
Content Creation	Creating content and the ability to post or share it; includes circulation and discoverability of content for child users.
Interpersonal Communication	Direct and indirect interactions between users; includes direct messaging, comments, likes, follows, and other engagement methods.
Usage Controls	Frequency, duration, and time limits of product use; includes compulsive use, do not disturb, and quiet hours.
Account Visibility	Profile and content visibility; includes profile discoverability, location sharing, friend requests, and live-streaming.
Monetization	Any usage that requires payment or results in receiving monetary gifts; includes advertising, virtual gifting, and creator profits.
Parental Supervision	Features that require parental oversight to activate or configure; some features may work with or without parental oversight, such as time limits.

Table B

A.2.3 RISK CATEGORY

To classify the type of harm a feature is designed to address, we adopt a 6C's framework. The foundational 4C's (Content, Contact, Conduct, and Contract) were developed by Livingstone and Stoilova (2021) and have been widely used by researchers to characterize online risks to youth. We extend this framework by adding two additional "C's", Circulation and Compulsivity. The Circulation risk captures harm from amplification and distribution of content involving minors, identified by Ma et al. (Ma et al. 2025). Compulsivity accounts for the class of features that address excessive or mis-timed use (such as at night or during school hours), primarily designed to contend against minors' inability to self-regulate their engagement. We define all categories in Table C.

CATEGORY	DESCRIPTION	Examples
Content	Exposure to harmful, age-inappropriate, or graphic material	Self harm content, violence, sexual content, eating disordered content.
Contact	Interactions from peers or other adults	Bullying, grooming, harassment, and unsolicited contact
Conduct	Risk that is the result of a minor's behavior	Oversharing, bullying others
Contract	Commercial and data exploitation	Targeted advertising, in-app purchases, undisclosed data collection, and virtual gifting
Circulation	Amplification or distribution of content beyond intended audience	Sharing or downloading of minor's content, viral spread of harmful content, screenshotting, and take-down tools (PhotoDNA, Take it Down)
Compulsivity	Excessive and/or mistimed use	Screen time controls, "do not disturb" mode, nudges to take a break

Table C

A.2.4 IMPLEMENTATION STYLE

The Implementation Style dimension describes how a feature is triggered. This lends an understanding for the burden placed on the user; features that are enabled by default require no activation by the user, while those that require configuration require significant user effort.

The four categories described in Table D form a spectrum from least to most user burden. Default On features are active without any user intervention and do not rely on user discovery or enabling to provide protection. Prompted features introduce a choice directly to the user, generally in response to another action. These features rely on the product to introduce the choice at the right time; testing must verify that the prompt appears and that declining the prompt does not incur harm. User Enabled features are available but passive and require the user to have knowledge of and active them independently. Configurable features are also user enabled, but secondarily require a nontrivial setup process before becoming useful.

CATEGORY	DESCRIPTION
Default On	Enabled automatically with no user action required
Prompted	Not enabled by default, but is suggested by the product in context
User Enabled	User must enable this by seeking it out themselves
Configurable	Requires a setup that is nontrivial, such as entering custom keywords, before it is functional, multi-step

Table D

A.3 TEST ACCOUNT CREATION

Our methodological approach is taken from literature on algorithmic auditing, particularly sock puppet audit methods. Sandvig et al. (2014) established principles for algorithm auditing, distinguishing between audit studies using researcher-controlled accounts versus observational studies of user experiences. Sockpuppet-based audits allow researchers to understand inputs and actual observed usage patterns better than observation can. This technique has been applied to other auditing research; Boeker and Urman (2022) used test accounts to study how user behavior influences TikTok's recommended feed.

Controlled accounts with known ages and usage streams are the only reliable way to test whether promised features activate and function as advertised. Our child test accounts are all created to be ages between 13 and 17, with a single account on TikTok being under 13 to activate TikTok For Younger Users. Adult accounts are created with ages over 25. We use randomly generated names taken from the most popular first and surnames in the United States as defined by census data. We also randomly generate specific birthdays and profile pictures for each account.

A.4 TEST SCENARIO DEVELOPMENT

Testing was conducted on iPhone and Android devices with the most recent updated version of Instagram, Snapchat, TikTok, and YouTube available at the time of testing. We began testing in December 2025 and finished in June of 2026. Findings are documented through screen recordings and screen shots.

For each feature, we developed a structured testing scenario based on the advertised functionality of the tool. Scenarios were performed from three perspectives: (1) a child using the product in an ordinary, non-adversarial way to demonstrate a natural use case of the feature; (2) a teen attempting to circumvent a feature or restriction placed on their account; and (3) a malicious adult user who is attempting to bypass restrictions and reach a child on the product. All testing was performed by interacting with the standard UI of the tested product with no advanced technical exploits or external resources.

A test case consists of four components: *Context, Setup, Test, and Evaluation.* Context and Setup are used to define the feature and prime the test scenario, while Test guides the execution and Evaluation records results.

A.4.1 CONTEXT

Context gives the researcher an understanding of what the risk is that the feature intends to mitigate. In the test case, this serves as both the motivation and intended outcome for the feature. It is informed by two taxonomy dimensions.

Risk Category (6C's) (Table C) determines the type of harm the feature is designed to address. Features vary widely in the harms they intend to mitigate from unsolicited adult contact to compulsive use; defining a risk category anchors the test to a specific harm that can be directly evaluated rather than an obscure notion of "safety." For example, a test of unsolicited contact should be a Contact risk, not a Content risk.

Product Surface (Table B) determines where on the product the defined harm manifests. The same risk may require different tests dependent on where the harm manifests; a Content risk that arises from Account Visibility requires different observations than one from Content Exposure.

The Context component also asks the tester to record the company's own claim about the feature. This established a standard against which the feature is later evaluated in terms of functionality. Finally, the potential harm outcome is documented. This is the concrete harm that could occur if the feature does not work as expected, such as unwanted contact or exposure to harmful content; during testing, this could influence the decisions the tester makes to observe if the harm can be realized despite the safety tool.

Context gives the researcher an understanding of what the risk is that the feature intends to mitigate. In the test case, this serves as both the motivation and intended outcome for the feature.

A.4.2 SETUP

Setup defines any preconditions that are necessary before a test can be executed. It is organized by three taxonomy categories.

Different tools are designed for different scopes of users; as such, we must have some structure for how many accounts need to be created and how they connect to each other on the product. Specifically, User Specific features require a single teen account as these features are managed by the user, for the user; Interaction features require, at minimum, two accounts since the feature is only activated through contact between users; Supervision requires both a teen account and a linked supervisory account; and Product features require a single account for access since they operate independent of any user's configuration.

Perspectives (Table E) is a new contribution from the test case framework. This specifies the stance of each account and how the tester should operate it to reflect realistic threat models. We define four perspectives: Child Intended (teen using product as normal), Child Circumvention (teen attempting to bypass restrictions), Adult Intended (adult using product normally), and Adult Circumvention (malicious adult attempting to reach a child against restrictions). Each test case specifies a primary perspective on which the feature will be activated and tested; a secondary perspective, if needed, can be used to interact with the primary account to test a feature reliant on interactions. Assigning perspectives before testing ensures the accounts are created with ages and behaviors that reflect real-world actors.

CATEGORY	DESCRIPTION
Child Intended	Child interacting with product with no intention to circumvent; realistic use
Child Circumvention	Child attempting to bypass restrictions or features in place on their account
Adult Intended	Adult interacting with the product with no intention to bypass restrictions
Adult Circumvention	Adult attempting to reach or interact with a child, against restrictions

Table E

Implementation Style (Table D) informs whether any configuration steps are required before testing begins. A Default On feature requires no user action and can be tested immediately upon account creation (though a tester should verify that default settings have not been altered). A Prompted feature is not active by default but should appear contextually, usually during some specific set of actions or after an allotted time period; this trigger should be noted by the tester via the company's advertised claim. A User Enabled feature requires explicit activation after account creation before a test can be conducted, usually through the product settings. A Configurable feature requires a non-trivial setup procedure (such as entering a custom list of keywords) before it is functional; the tester should document steps that are needed to activate this feature. Finally, Setup includes documentation of any steps taken to establish necessary account conditions for testing, such as posting or bot mitigation techniques.

A.4.3 TEST

The Test section outlines how the test should be performed. It is informed by three taxonomy dimensions.

UI Bound determines the surface where the feature is expected to operate and where the test should be focused. Table F contains a non-comprehensive list of UI components; as companies develop new features, this list could expand. Features may operate across one or several surfaces, such as direct messaging, content feed, search, profile pages, stories, or settings. A feature designed to restrict content in search may not apply to feed so it is critical to identify the correct surface, otherwise the tester risks a false negative.

COMPONENT
Messaging / DMs
Following Feed
Recommended Feed
Reels / Short Form
Content Creation Interface
Profile
Settings
Search
Stories
Post

Table F

Implementation Style (Table D) informs when the feature is expected to deploy. A Default On feature should be active throughout the test; Prompted should appear as a reaction to certain input. The tester should record the expected intervention timing (e.g., before a message is sent or upon viewing a profile) and document this against actual observations.

A.5 EVALUATION FRAMEWORK

We evaluate each feature across two independent dimensions: functionality and accessibility, using a decision chain for each. The logic is described in detail in Section 2, but we include a summary here for reference.

Triggerability. We first determine if the feature in question could be activated on the product. Features that could not be activated at all under conditions described by press releases, despite our best attempts, were marked as *Missing* and excluded from further analysis.

Functionality. We determine if a feature is functional by satisfying three different questions:

- → Is the feature well-designed?
- → Is the feature well-implemented?
- → Is the feature resistant to circumvention?

A single "no" at any stage was sufficient to classify the feature as nonfunctional, and therefore, **broken.**

Accessibility. We consider a feature accessible if it satisfies the following two questions:

- → Is the feature easy to turn on or set up?
- → Is the feature on by default, or does the product prompt users to enable it?

If either question is a "no," then the feature is not considered accessible, and therefore is **buried.**

If a feature is both functional and accessible, it is **Successful.** A feature that is nonfunctional but accessible is Broken. A functional but inaccessible feature is Buried. Finally, a feature that is both nonfunctional and inaccessible is Broken and Buried. Either of the three non-successful categories is considered a **Failure.**

TABLE 3: PRODUCT SURFACE

CATEGORY	DESCRIPTION
Content Exposure	Viewing, searching, or recommendations of content; includes feeds and search results.
Content Creation	Creating content and the ability to post or share it; includes circulation and discoverability of content for child users.
Interpersonal Communication	Direct and indirect interactions between users; includes direct messaging, comments, likes, follows, and other engagement methods.
Usage Controls	Frequency, duration, and time limits of product use; includes compulsive use, do not disturb, and quiet hours.
Account Visibility	Profile and content visibility; includes profile discoverability, location sharing, friend requests, and live-streaming.
Monetization	Any usage that requires payment or results in receiving monetary gifts; includes advertising, virtual gifting, and creator profits.
Parental Supervision	Features that require parental oversight to activate or configure; some features may work with or without parental oversight, such as time limits.

TABLE 4: FEATURE OVERSIGHT

CATEGORY	DESCRIPTION	Examples
User-specific	Designed at an individual level; can be managed by the user for their own protection.	Screen time limits, block, mute, sensitive content control.
Interactions	Deployed as a result of interactions between users. Typically reactionary.	Message initiation warnings, prompts for reconsidering comments, ability to follow/tag another user.
Supervision	Requires a third party; enables supervisor to monitor and/or control experience of another user.	Family Center, linked accounts, time limits.
Product	Implemented at a product level and not operated/controlled by the user.	Content removal, redirection to resources, account bans.

TABLE 5: 6 C's

CATEGORY	DESCRIPTION	Examples
Content	Exposure to harmful, age-inappropriate, or graphic material	Self harm content, violence, sexual content, eating disordered content
Contact	Interactions from peers or other adults	Bullying, grooming, harassment, and unsolicited contact
Conduct	Risk that is the result of a minor's behavior	Oversharing, bullying others, or adjusting privacy settings
Contract	Commercial and data exploitation	Targeted advertising, in-app purchases, undisclosed data collection, and virtual gifting
Circulation	Amplification or distribution of content beyond intended audience	Sharing or downloading of minor's content, viral spread of harmful content, screen-shotting, and take-down tools (PhotoDNA, Take It Down)
Compulsivity	Excessive and/or mistimed use	Screen time controls, "do not disturb" mode, nudges to take a break

TABLE 6: Harm Response

CATEGORY	DESCRIPTION
Prevention	Goal is to prevent harm; given functionality the user should not be exposed to harm at all.
Mitigation	Goal is to address or mitigate harm after it has occurred; assumes user has had some exposure.

TABLE 7: IMPLEMENTATION STYLE

CATEGORY	DESCRIPTION
Default On	Enabled automatically with no user action required
Prompted	Not enabled by default, but is suggested by the product in context
User Enabled	User must enable this by seeking it out themselves
Configurable	Requires a setup that is nontrivial such as entering custom keywords, before it is functional; multi-step

TABLE 8: PERSPECTIVES

CATEGORY	DESCRIPTION
Child Intended	Child interacting with product with no intention to circumvent realistic use
Child Circumvention	Child attempting to bypass restrictions or features in place on their account
Adult Intended	Adult interacting with the product with no intention to bypass restrictions
Adult Circumvention	Adult attempting to reach or interact with a child, against restrictions

TABLE 9: UI BOUND

COMPONENT
Messaging / DMs
Following Feed
Recommended Feed
Reels / Short Form
Content Creation Interface
Profile
Settings
Search
Stories
Post

About the Publishers

CRC CYBERSAFETY RESEARCH CENTER

CYPERSAFETY RESEARCH CENTER

Cybersafety Research Center (formerly Cybersecurity for Democracy) is a multi-university center for problem-driven research and research-driven policy. We conduct cutting-edge computer science research to better understand the distorting effects of algorithms and AI tools on large online networks and how they affect safety, public health, and democracy. We work with platforms and regulators to help all parties understand the implications of our findings and develop solutions.

HEAT INITIATIVE

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 FOR GOOD