Tennessee v. Meta
Case No. 23-1364-IV
PX-00128

Addressing online harms for children at Meta

Prepared for The Oversight Board
September 26, 2024
by Arturo Béjar

Public sources for this presentation

The documentation published by the Senate:

https://www.blumenthal.senate.gov/newsroom/press/release/blumenthal-releases-documents-provided-by-meta-whistleblower

The BEEF presentation, published by the New Mexico AG:

https://storage.courtlistener.com/recap/gov.uscourts.nmd.496039/gov.uscourts.nmd.496039.36.2.pdf

The problem

CONTEXT · **WHY BAD EXPERIENCES**

100%

of Community Standards policy focuses on enforcement (delete, MAD, or resources)

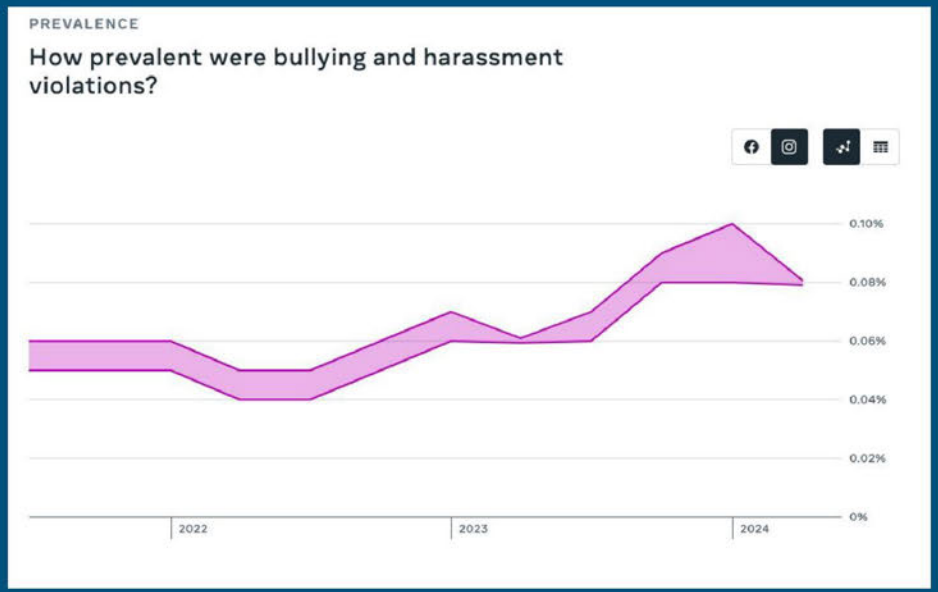
~100x

Incremental reach for B&H witness in TRIPS compared with Community Standards B&H policy violating impressions

Instagram bullying & harassment **prevalence** in 2021

Meta's transparency reports for Instagram in 2021:

Of 10,000 views 5-6 contain bullying & harassment.



Internal bullying & harassment data shows a completely different picture

Bad Experiences and Encounters Framework (BEEF)

Have you seen someone do any of these things to someone else on Instagram? Contact them in an inappropriate way, damage their reputation, threaten them, exclude them or leave them out, insult or disrespect them.

28% of 13-15 year olds said yes, in the last 7 days.

Has someone done any of these things to you?

11% of 13–15 year olds said yes, in the last 7 days.

Frequency: 64% more than once 12% more than 10 times.

Why the gap?

Of **10,000** people who have a bad experience:

- **100** submit something through the reporting tools.
- **2** have the report acted on.
- **49.2%** of people who experience bullying do not feel supported.

This is a **material gap** in how Meta manages and resources preventing the harm their products enable. This gap disproportionately affects children and young teens who are developmentally at risk.

Meta's transparency reports do not present an accurate picture of risks and harms to parents or lawmakers. Prevalence is misleading.

Major risks for minors on Meta's products

Inappropriate content recommendations to teens, and the mental health impact of extended exposure.

Inappropriate contact through direct messaging and comments.

Under 13 users on Instagram.

Rewarding content that sexualizes minors.

Unwanted/inappropriate content
in the last 7 days

Have you seen violent, bloody, disturbing images that bothered you?

12.8% of 13-15 year olds said yes.

Have you seen nudity or sexual content that you did not want to see?

19.2% of 13-15 year olds said yes.

Have you seen someone harm themselves, or threaten to do so?

8.4% of 13-15 year olds said yes.

Have you seen discrimination against people because of their gender, religion, race, sexual orientation, or another part of their identity?

26% of 13-15 year olds said yes.

	Overall %	13-15	16-17
		Column N %	Column N %
commerciality	48.2%	37.7%a	42.5%b,f,g
misinfo	30.3%	24.4%a	29.2%b
bully witness	28.3%	27.2%a	29.4%b
audience limitation	26.8%	28.2%a,b	29.3%a
usability passive	25.5%	30.1%a	30.0%a
hate witness	25.3%	26.0%a	28.5%b
data privacy	24.4%	17.9%a	21.4%b
perceived control	23.9%	18.4%a	20.0%a
fake acct 1st	21.8%	15.8%a	21.2%b
negative comparison	19.2%	21.4%a	19.5%b
transparency	17.5%	14.9%a	19.1%b,f
political posts	17.0%	14.3%a,d	16.3%a,b
nudity	16.3%	19.2%a	18.4%a
over enforcement	14.8%	14.1%a,c	16.8%b
violence	12.8%	12.8%a,b	14.3%a
unwanted advances	11.9%	13.0%a	14.1%a
usability action	10.7%	11.2%a,e	11.1%a,e
bully target	8.1%	10.8%a	9.7%a
self harm	6.7%	8.4%a	7.2%a,b
acct security	3.9%	5.2%a	3.2%b,c
drugs	3.9%	3.7%a,b	3.5%a,b
impersonation 1st	3.7%	5.5%a	3.5%a

Negative Comparison

Have you felt worse about yourself because of of other people's posts on Instagram?

21.4% 13-15 year olds said yes.
80% more than once, 16% more than 10 times.

'She opens the app and gets bombarded'

'There is nothing we can do about it'

[redacted]



Self-harm and product design

Testing Instagram behavior for the account of a 13 year old

In any search engine, searching 'I want to hurt myself' the top links are resources offering help .

On Instagram, for the last 2 years including today, the top recommendation is an account whose profile image is of cutting which cannot be reported.

You then get a feed of depressive content recommended, similar to most of the content Molly Russell got recommended prior to committing suicide.

In the tags tab, you get images of a teddy bear hung with the word 'suicide', and a black and white image of murder-suicide between Mickey and Minnie Mouse.

I can provide screenshots, or it is easy to reproduce from your own phone.

https://integrityinstitute.org/blog/why-is-instagram-search-more-harmful-than-google-search

Content policy is not the right tool

CONTEXT · **EXAMPLES OF BAD EXPERIENCES**

Examples of bad experiences often in policy gaps

Mass harassment

Very intense, low reach. Disproportionately affects creators. Portion isn't CS policy violating but still felt as intense.

Non-credible or non-violent threats

Doesn't meet CS violating policy bar, but felt intensely with moderate reach.

Thinspiration

Easy to find content adjacent to eating disorder part of SSI and to broader social comparison and body image issue-triggers.

Reducing unwanted content for teenagers does not affect freedom of speech

The right of someone to post non violating self-harm, sexual, or violent content, does not mean that it should be recommended to a 13 year old.

Recommendation surfaces like reels or feed should be reasonably safe, and reasonably free of unwanted content for teens.

Amplification from engagement can be profoundly harmful. Watching a video out of disgust in results in more of that content being recommended.

When a teenager gets recommended unwanted content, it should be rewarding and effective to flag it and why. The 'not interested' button does not work.

https://panoptykon.org/sites/default/files/2023-12/panoptykon_algorithms-of-trauma-2_case-study-report_dec-2023.pdf

Inappropriate/unwanted contact

Have you received unwanted sexual advances on Instagram?

13%, 1 in 8 kids, yes in the last 7 days.
68.6% on direct messaging.
18.5% comments on their posts.

There is <u>still</u> no way for a teenager to indicate when this happens in product.

Without inappropriate contact data, how can the issue be understood and addressed?


CONTEXT · EXAMPLES: THREATS

When I received rape threats in my DMs, Instagram offered me no help at all

Comment

FIRST PERSON [redacted] Friday 24 Jul 2020 3:52 pm

I've been on the receiving end of an unsolicited dick pic or two in my time. I laughed them off at first.

Then the same man started sending explicit descriptions of violent sexual fantasies he had about me. This included rape, choking, putting me in crutches and forcing my mother and father to watch.

Understanding unwanted advances, sexual harassment, and threats over DMs

"I look forward to seeing you tonight."

- What if you get that message from a stranger? Or someone who has made you feel uncomfortable at school?
- How many ways are there to express unwanted sexual advances and threats? Across cultures, age groups, and different languages?
- Your messages should be a safe space for you, message content doesn't matter. What matters is someone contacted you and then made you feel harassed.

There needs to be a way in the product to learn of inappropriate contact

- Add a button in messages, or after blocking, where teens can easily flag inappropriate contact & the reason why.
- This is not 'reporting'.
- Language in the button needs to match experience:
  - 'this is not for me'
  - The reason why e.g. 'creepy', 'they are pretending to be someone they're not', etc.
- When someone initiates a few times, give them feedback.
- When someone initiates many times, investigate.

Under 13 users and rewarding exploitative content

It is trivial to find large cohorts of teenagers under 13 using Instagram. Basic testing with Reels can create a feed made up almost exclusively of under 13 accounts with large reach, primarily posting content that sexualizes minors but is not CSAM.

Posts of minors that are sexually exploitative posts have tens or hundreds of thousands of views vs. more innocent content that has a few hundred views.

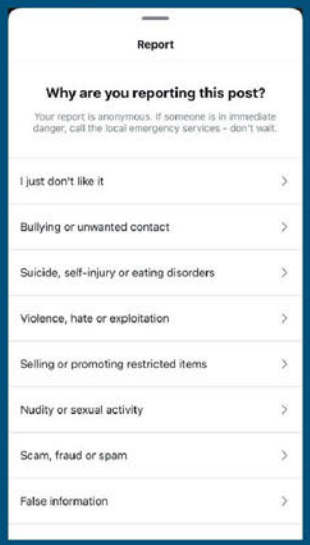
Anonymized public comments from posts by minor presenting accounts:

- GIFs that include "Boner Alert" and images implying masturbation.
- "Sexy."
- "You should go get a chair and a rope."

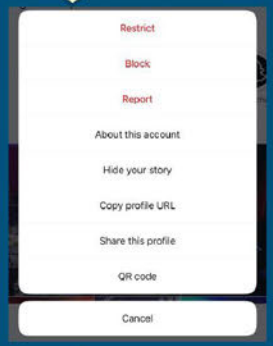
Bad faith design

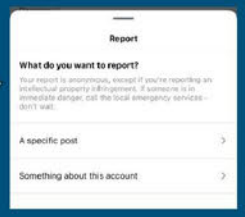
Under 13 reporting

You cannot report directly from content

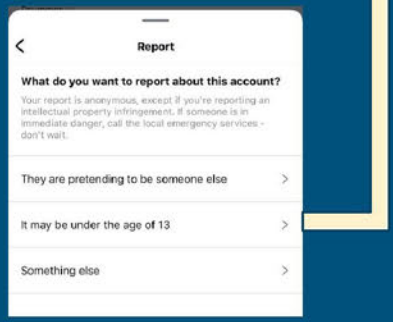


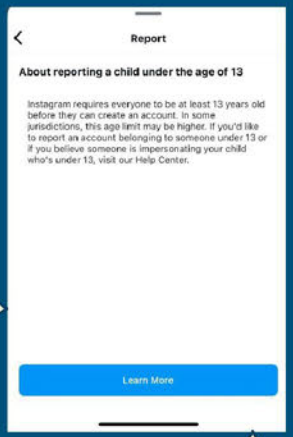
3 taps to get to the profile and tap 'Report'.

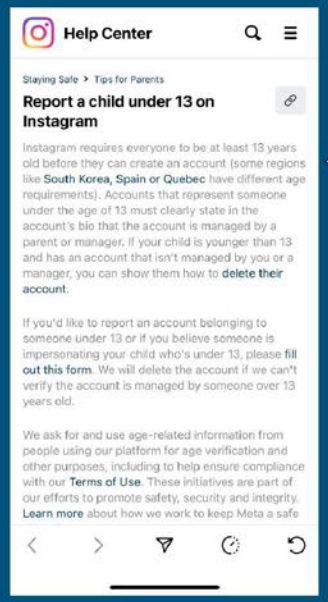




No way to submit report in the product..

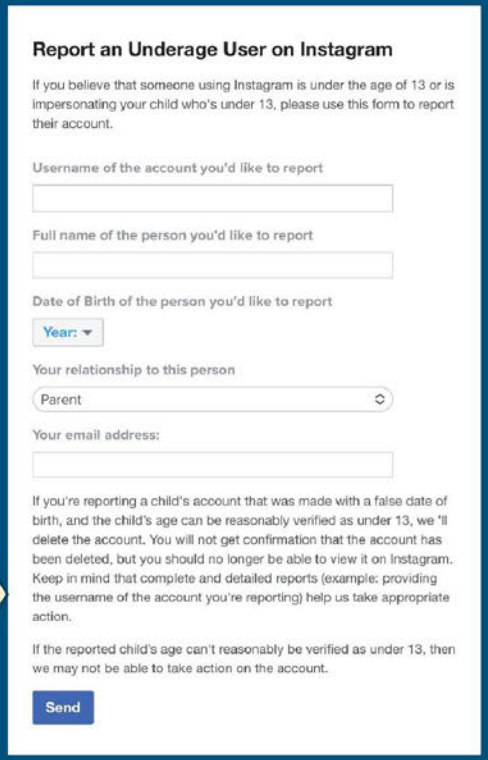






Help center, link to submit report inline in second paragraph.

There is no way to cut and paste the account name from the app (context which they had), no way to know the name of the person, or the birth year, asks for your e-mail address.



The impact of Meta's negligence

One of the most critical problems with all of these issues, is that Meta's unwillingness to understand and address these issues creates an environment where harmful content and behavior is **normalized**.

The problems we are seeing today with gender or identity based online violence, the attacks towards women and minorities, minor sexualization, etc. All stem from an environment where harmful content is rewarded by distribution. When a teen experiences harm, Meta will not help.

My daughter, and all of her friends, have experienced most of the harms covered in this presentation directly. She tried reporting them every time. When asked when she got help, she said: 'Not once'.

Every risk discussed here could be meaningfully reduced in a short amount of time if it was a priority for Meta.

What could improve child safety?

Transparency that incentivizes the reduction of harms that teens experience in Meta's products.

Ensuring that for unwanted content and unwanted contact, there is an easy way for teens to flag in the product when it happens.

Require that independent academic experts are embedded in product teams that are working on issues of teen health and safety.

An investigation on under 13 accounts, and content that sexualizes and exploits minors.

A process to ensure there are lessons learned and reasonable product changes after a teen dies or ends up in hospital.

Suggestions for best practices

Product principles, basic safety rights of teens:

- **Right to safety in your messages.**
  - Measured by unwanted advances or inappropriate contact.
- **Right to safe recommendations**.
  - Measured by unwanted content on recommendation surfaces.
- **Right to a safe space where you can express yourself**.
  - Measured by inappropriate or harassing comments.

Agency measured by asking: When I have a bad experience, I have a tool that helps me, which also helps make the community safer.

Measuring teen support tools, by whether the teen felt supported.

Full access for domain experts is possible and necessary

Reputable academic experts should be embedded as independent consultants in product and engineering teams, with agreements to do privacy preserving publishing and research.

The way Meta currently engages with experts is safety theater. The meetings with the policy team and discussions rarely result in product changes.

A note on corporate confidentiality

All of the issues surrounding harmful experiences for minors should be discussed open and transparently.

Any aggregate, non re-identifiable information around teen harm should be public.

This is not a competitive issue, the more there is transparency the safer these products will be for teens.

The only reason for confidentiality in this domain is to avoid embarrassment for executives and reputational damage when a company is negligent.

Questions that Meta can answer

Unwanted advances/inappropriate contact

How do you know when a teen experiences an unwanted advance?

Estimate the number of teens that experienced unwanted sexual advances in the last quarter.

When a teen experiences an unwanted sexual advance:

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

One of the actions a teen can take is blocking, do you ask them why? If you don't ask, how can you know if the teen blocked because they got an unwanted advance?

How many submit a report? What categories do they use?

What percentage of reports is acted on?

What is the distribution of response times?

Questions that Meta can answer

Bullying and harassment

How do you know when a teen experiences bullying and harassment?

Estimate the number of teens that experienced bullying and harassment in the last quarter.

When a teen experiences bullying and harassment:

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

How many submitted a report?

Did teens feel supported after submitting a report?

What percentage of reports is acted on? What is the distribution of response times?

If the teen blocked, did you ask them why? If you don't ask, how can you know if the teen got bullied and provide them appropriate support?

Questions that Meta can answer

Unwanted content: self-harm, violent, sexual, etc.

How do you know when a teen experiences unwanted [harm type] content?

What percentage of teens experienced unwanted [harm type] content in the last quarter?

When a teen experiences unwanted [harm type] content:

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

How many submit a report? What categories do they use?

What percentage of reports is acted on?

What is the distribution of response times?

What is your analysis of the content of reports that is not acted on?

Questions that Meta can answer, for the last quarter
Under 13 accounts and content that exploits minors

When someone tries to report an account that looks to be under 13 years old:

- How many people take the first step?
- How many people submit a report?
- How many reports did you act on?

When someone comes across content where a kid under 13 is posting sexually suggestive or exploitative - but not explicit content.

- What actions can they take?
- What actions should they take?
- What actions do they take and what is the result?

When inappropriate or suggestive comments of minors happen (gifs, emojis, etc).

- What actions should the creator take? What actions do they take?
- What actions should an observer take? What actions do they take?

For areas where there are reports, what percentage of reports is acted on?

What is your analysis of the reports that are not acted on?

DOCUMENT SLIPSHEET

Bates Number: BEJAR0000444-BEJAR0000469

BEGATTACH:

ENDATTACH:

Attachment Count: 0

PRODVOL:

Custodian: UNSPECIFIED CUSTODIAN

File Path: /BEJAR0000001-BEJAR0002431/PART 1/BEJAR0000444.PDF

Confidentiality Designation:

HASHVALUE:

Author:

Family Date: 2/28/2025 6:48 PM

Create Date: 2/28/2025 6:48 PM

Last Modified Date: 2/28/2025 6:48 PM

LAST MODIFIED BY:

TRACK CHANGES:

COMMENTS:

HASHIDDENDATA:

Filename: BEJAR0000444.PDF

Title:

DOCEXT:

Email From:

Email To:

Email CC:

Email BCC:

Received Date: 12/31/9999 7:00 PM

Sent Date: 12/31/9999 7:00 PM

Subject:

THREADID:

REDACTION TYPE:

REDACTIONS: