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

DX3790

# Document Provided in Native Format

# Company Well-being Strategy: Bullying and Harassment

Summit September 16, 2021

FACEBOOK

# Agenda

1. Discuss company unified well-being strategy (more info).
2. Outline external views and research (more info).
3. Upcoming research and research-based approaches to product development.


Company Well-Being Strategy
FACEBOOK

Company Well-being topics

The company is focusing on four topics:

1. Bullying and Harassment **[focus today]**
2. Problematic Use
3. Connections
4. Emotional impact

## How did we get to those topics?

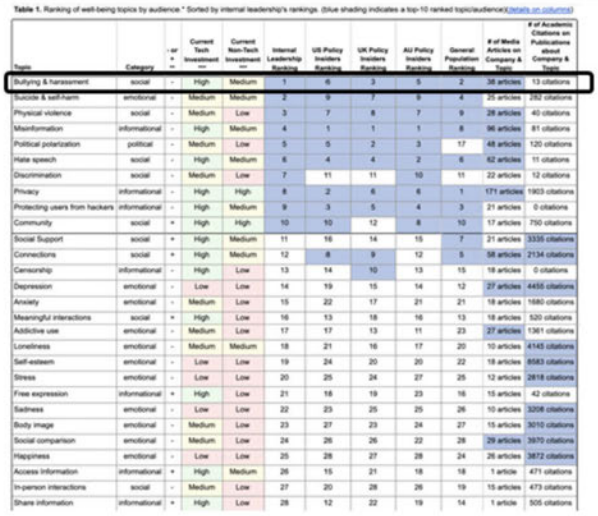
- Identified 1000+ topics during a series of literature reviews (400+ academic articles, 900+ media articles, 150+ internal research articles).
- We narrowed the list down to the top 28 most commonly referenced topics.
- Conducted a MaxDiff exercise with 91 leaders at Facebook, 402 policy insiders, and 12,002 people from the general population and they ranked the importance of Facebook focusing on each topic.
- Finally, conducted an internal investment audit, to find the highest impact opportunities.

## B&H is a Priority

Users chose bullying second out of 28 other well-being topics as a an important focus.

It also falls in the top 10 important focus areas by internal leadership and policy insiders. The media regularly reports on it.

Internal analysis shows those who feel bullied are more likely to deactivate and disengage.



| Topic | Category | - or + or ~ | Current Tech Investment *** | Current Non-Tech Investment *** | Internal Leadership Ranking | US Policy Insiders Ranking | UK Policy Insiders Ranking | AU Policy Insiders Ranking | General Population Ranking | # of Media Articles on Company & Topic | # of Academic Citations on Publications about Company & Topic |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bullying & harassment | social | - | High | Medium | 1 | 6 | 3 | 5 | 2 | 38 articles | 13 citations |
| Suicide & self-harm | emotional | - | Medium | Medium | 2 | 9 | 7 | 9 | 4 | 25 articles | 282 citations |
| Physical violence | social | - | Medium | Low | 3 | 7 | 8 | 7 | 9 | 28 articles | 40 citations |
| Misinformation | informational | - | High | Medium | 4 | 1 | 1 | 1 | 8 | 96 articles | 81 citations |
| Political polarization | political | - | Medium | Low | 5 | 5 | 2 | 3 | 17 | 48 articles | 120 citations |
| Hate speech | social | - | High | Medium | 6 | 4 | 4 | 2 | 6 | 62 articles | 11 citations |
| Discrimination | social | - | Medium | Low | 7 | 11 | 11 | 10 | 11 | 22 articles | 12 citations |
| Privacy | informational | - | High | High | 8 | 2 | 6 | 6 | 1 | 171 articles | 1903 citations |
| Protecting users from hackers | informational | - | High | Medium | 9 | 3 | 5 | 4 | 3 | 21 articles | 0 citations |
| Community | social | + | High | High | 10 | 10 | 12 | 8 | 10 | 17 articles | 750 citations |
| Social Support | social | + | High | Medium | 11 | 16 | 14 | 15 | 7 | 21 articles | 3335 citations |
| Connections | social | + | High | Medium | 12 | 8 | 9 | 12 | 5 | 58 articles | 2134 citations |
| Censorship | informational | - | High | Low | 13 | 14 | 10 | 13 | 15 | 18 articles | 0 citations |
| Depression | emotional | - | Low | Low | 14 | 19 | 15 | 14 | 12 | 27 articles | 4455 citations |
| Anxiety | emotional | - | Medium | Low | 15 | 22 | 17 | 21 | 21 | 18 articles | 1680 citations |
| Meaningful interactions | social | + | High | Low | 16 | 13 | 18 | 16 | 13 | 18 articles | 520 citations |
| Addictive use | emotional | - | Medium | Low | 17 | 17 | 13 | 11 | 23 | 27 articles | 1361 citations |
| Loneliness | emotional | - | Medium | Medium | 18 | 21 | 16 | 17 | 20 | 10 articles | 4145 citations |
| Self-esteem | emotional | - | Low | Low | 19 | 24 | 20 | 20 | 22 | 18 articles | 8583 citations |
| Stress | emotional | - | Low | Low | 20 | 25 | 24 | 27 | 25 | 12 articles | 2818 citations |
| Free expression | informational | + | High | Low | 21 | 18 | 19 | 23 | 16 | 15 articles | 42 citations |
| Sadness | emotional | - | Low | Low | 22 | 23 | 25 | 25 | 26 | 10 articles | 3208 citations |
| Body image | emotional | - | Medium | Low | 23 | 27 | 23 | 24 | 27 | 15 articles | 3010 citations |
| Social comparison | emotional | - | Medium | Low | 24 | 26 | 26 | 22 | 28 | 29 articles | 3970 citations |
| Happiness | emotional | - | Low | Low | 25 | 28 | 27 | 28 | 24 | 26 articles | 3872 citations |
| Access information | informational | + | High | Medium | 26 | 15 | 21 | 18 | 18 | 1 article | 471 citations |
| In-person interactions | social | - | Medium | Low | 27 | 20 | 28 | 26 | 19 | 15 articles | 473 citations |
| Share information | informational | + | High | Low | 28 | 12 | 22 | 19 | 14 | 1 article | 505 citations |

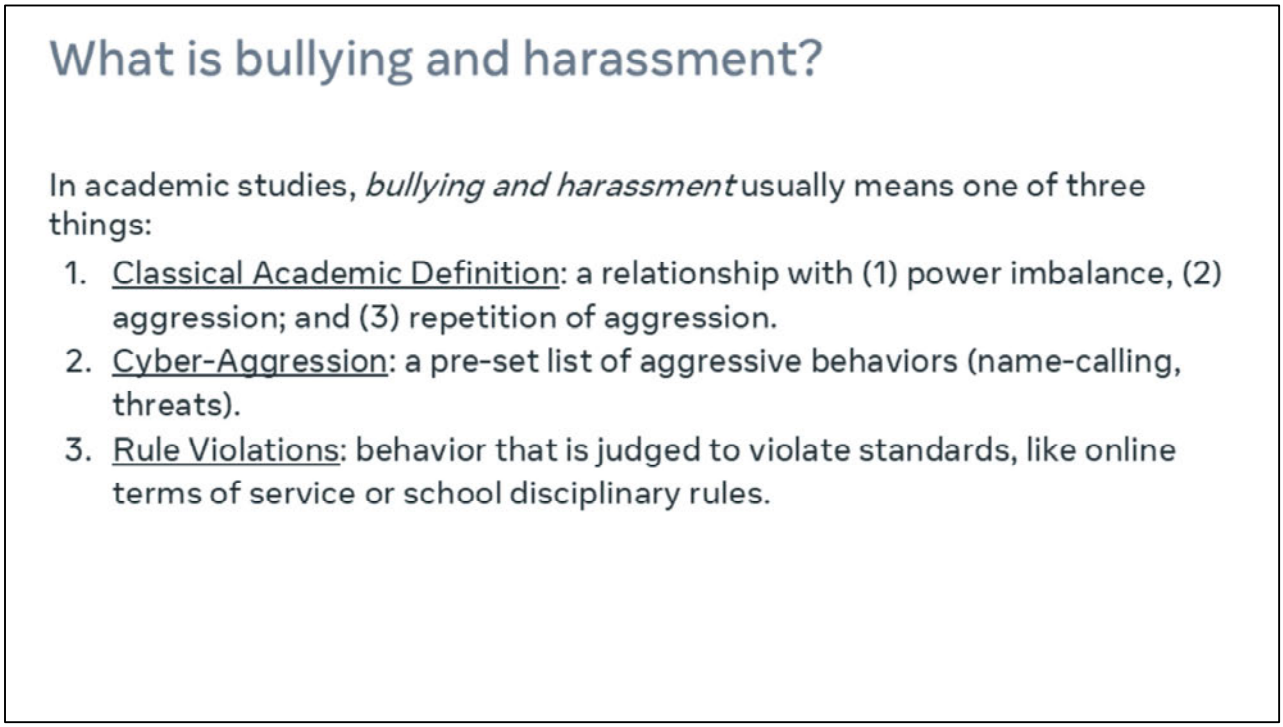
[source]

*Disclaimer: This data best represents adults in the US, UK, and Australia as well as adults; we replicated the study among teens and found highly consistent results.*

This is a summary of many of the results from the exercises I just outlined discussed.

Highlighting the bullying and harassment row (near the top),you can see that bullying and harassment is a high priority across a number of axes. In fact, it's generally one of the higher ranked issues among the various groups (apart from academics).


External Narratives and Research
About Bullying and Harassment
FACEBOOK



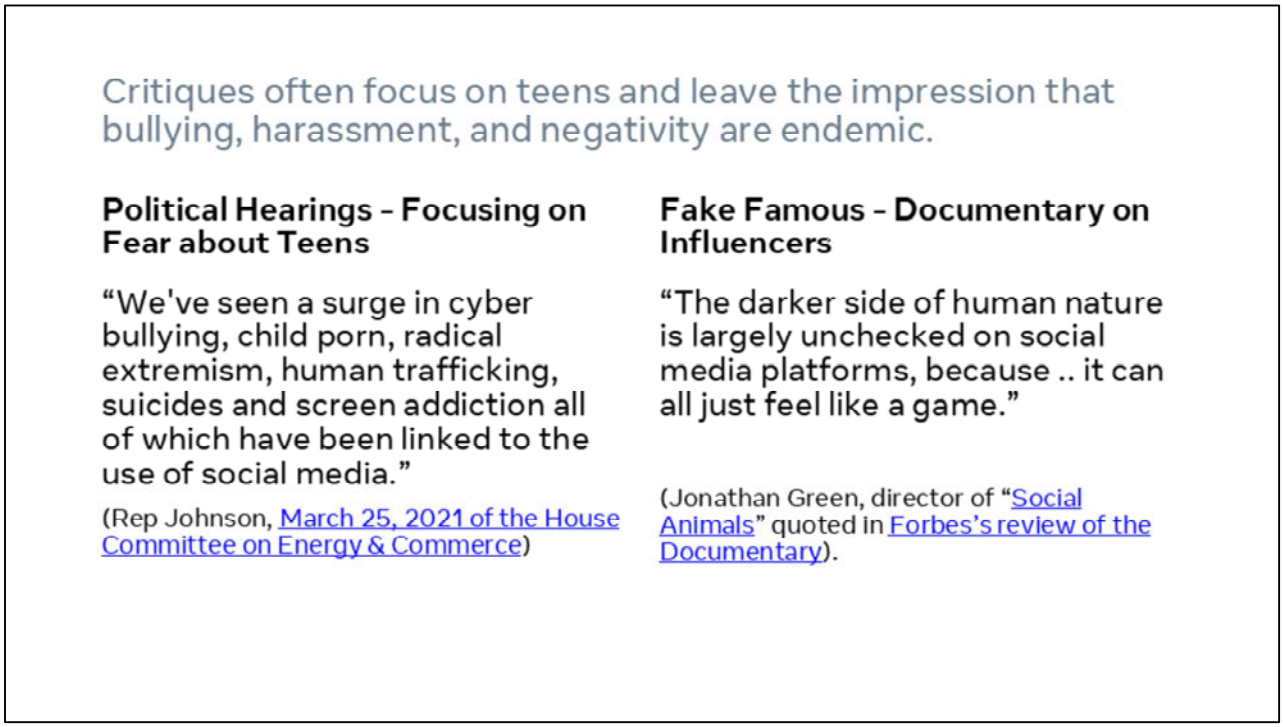

[Detail]

- Based on our literature review, academics are still evaluating how to appropriately measure bullying and harassment, especially in online contexts.
- "Bullying and Harassment" or "cyberbullying" are used very loosely and usually fit into one of three major categories:
    - **First, there is the technical definition.** There is the classic definition (which is commonly used by official government tracking organizations, like the US CDC).
    - **Second, there is the looser definition of "aggression".** There is a looser and more broad category of "Aggression" which is a list of behaviors.
        - This is the approach we use to measure bullying and harassment internally, using the Tracking Reach of Integrity Problems Survey (TRIPS) [source]
            - Insults and disrespects
            - Inappropriate contact
            - Damaging reputation
            - Threatening
            - Excluding or leaving out
        - This is also the approach used by PEW research when they publish a survey of online harassment roughly every two years [source]
    - **Third, there is "violating behavior."** This is commonly used by external researchers testing school-based programs -- they use disciplinary reports for violation of school standards and policies against bullying as a measure

[example]. For the company, we use this definition when we use our "community standards report"

[Support and Additional Info]
Literature review section, here.

- Differences in definitions are not just academic; they can show up in empirical studies and spark confusion.
  - One article we uncovered looked at 76 articles of longitudinal studies covering adolescent populations which were published between 2007 to 2017 (Camerini et al. 2020). They show an extremely wide range of prevalence between 1.9% to 84%.
  - We also see very large differences in the magnitude of our internal measures where our "classifier estimated prevalence" estimates the proportion of VPVs that violate our community standards and run fractions of a percent [dashboards]; around 0.2%. In contrast, TRIPS estimates perceptions of bullying at 10% (for targets) to 20% (for witnesses) [dashboards].




[Detail]

- There are several external critiques regarding online bullying and harassment. Our evaluation so far has been based on reviews of news media, documentaries tv/movies, government websites, and nonprofit websites (Media Searches (Including Movies, TV, and Documentaries; External Critique List; Press/Media Eval from Comms; Press/Media Eval from Policy; External Critique Overview).
- Examples of two common critiques are outlined here:
  - Critical voices often make pleas about children and teens, making social media sound like a very dangerous place, the left pane summarizes a sentiment that some researchers call a "moral panic".
  - We also here critiques about public figures or creators being subject to more bullying and harassment than others; the right pane comes from a documentary about young adults in this context.
- There is not a lot of evidence for a bullying epidemic online or on our platforms; but some external areas to highlight areas of concern for certain public figures.

[Additional Info]

Literature review source

Andrew Przybylski PhD. 2020 Why scientists don't actually know if social media is bad for you covers the impulse toward moral panic quite well.

## Estimates of Prevalence from Academic Articles

Recent academic review (2020) summarizes (peer reviewed) estimates:

1. Perpetration ranges from 5.3% to 66.2%
2. Victimization ranges from 1.9% to 84%

"Failure to have a consensus about a definition is not a trivial problem."

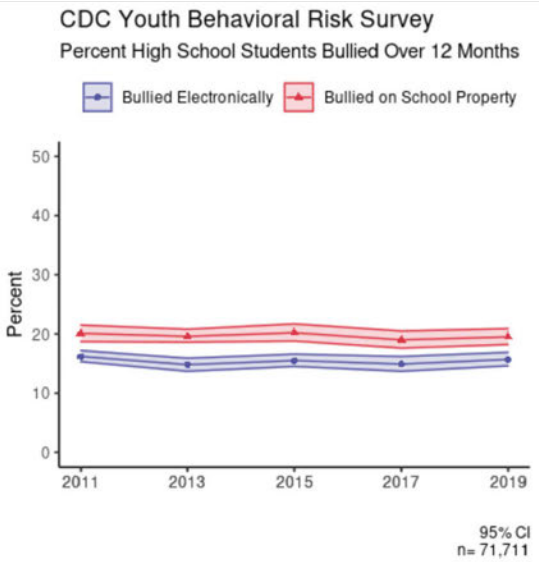
(Ansari 2020)

## Bullying among US Teens did not increase in the last decade.

The United States Centers for Disease Control & Prevention tracks the incidence of bullying in US high schools.

Both at-school and online bullying have remained consistent through the tracking period (between 2011 and 2019).

Studies among teens in the UK are similar. Pew Research Center finds no change in the frequency of online harassment for adults between 2017 and 2020, but did see an increase in self-reported severity.




[Detail]

- Since the CDC began tracking this data, among schools, rates of at-school and online bullying have remained consistent at around 20% for in-person and around 16% or 17% for “electronic”
- While this slide highlights the US survey, there are similar results in other countries. The UK department of education reports on a longitudinal survey of adolescents called the Crime Survey of England and Wales (accessed 3/14/2021).
- For adults, Pew research Pew Research Center conducts a survey every 3 years among internet connected adults in the United States (2014, 2017, and 2020).
- Pew found no change in the frequency of online harassment between 2017 and 2020.

[Support and Additional Info]

Literature review section, here.

- Since 2009, the CDC has tracked bullying on school property in a survey that is representative of high school students in the United States, with the exception of Minnesota, Oregon, Washington, and Wyoming (which do not participate).
- In 2011, they added a question about electronic bullying, and revised that question in 2017 to specifically references Instagram and Facebook.
- While there are some differences in their measures (the UK data shows less bullying overall), the patterns between 2013 and 2017 are the same: fairly steady rates, with “bullying: exceeding “cyberbullying” a few points. In this study of England

and Wales, they found that 63% of all bullying happened exclusively on school grounds.

- Pew does not rely on the technical definition of bullying, but use a broader list of "aggression" like TRIPS
    - Their estimates are higher (near 43%-45%), but did note an uptick in feelings of severity. Their report also highlighted increasing civic harassment.
        - Notably, the 2020 wave was conducted Sept. 8-13, 2020 during the pandemic, and before the presidential election.
        - So the stability of estimates occurred after the summer and as the presidential election was heating up.

[Survey Language]

**[YRBS data]**
Notably, the language of the electronic bullying question specifically references Instagram and Facebook since 2017:

> **[Language of CDC Survey]**
> Bullying is when 1 or more students tease, threaten, spread rumors about, hit, shove, or hurt another student over and over again. It is not bullying when 2 students of about the same strength or power argue or fight or tease each other in a friendly way. During the past 12 months, have you ever been electronically bullied? (Count being bullied through texting, Instagram, Facebook**,** or other social media.)
>
> **Parenthetical used 2011-2015:** (Count being bullied through e-mail, chat rooms, instant messaging, Web sites, or texting.)
>
> More information about the YRBSS here; 2021 questionnaire, here [bullying questions are unchanged].

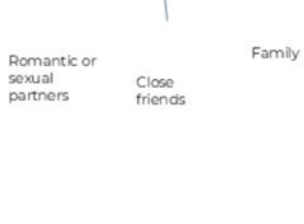
- They have conducted the survey every two years, usually in the spring, but the 2021 data collection has been delayed (and probably won't refresh until 2021)

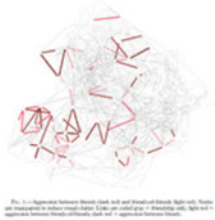
**[PEW data]**

> **[Language of Pew Survey]**
> Which, if any, of the following have happened to you, personally, ONLINE?
>
> 1. Been called offensive names
> 2. Been physically threatened
> 3. Been harassed for a sustained period
> 4. Been stalked
> 5. Had someone try to purposely embarrass you
> 6. Been Sexually Harassed

**Bullying and harassment can occur across relationships of varying strengths -- context can change the meaning of particular interactions.**



Bullying and harassment can be highly context-dependent.

Some research suggests students are three times more likely to experience aggression from friends (or frenemies) and friends of friends than strangers [source]; studies of online aggression in New York schools show similar patterns [source].

Distinguishing between playful banter among friends and bullying can be difficult and complex [*e.g.* source, source].

## Some public figures report higher experiences of bullying online.

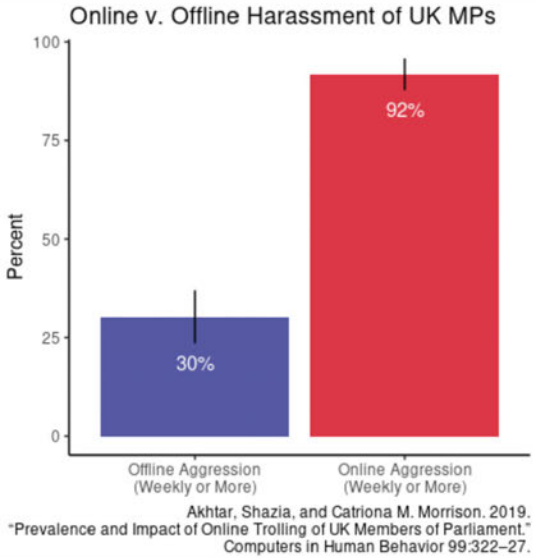
There is a lot of news, but very few quantitative studies on public individuals.

In one external study:

30% of politicians reported weekly in-person aggression (abusive language, threats)

92% reported online aggression through Facebook and Twitter.

Internal research on Instagram estimates creators experience bullying 1.3 to 1.9 times more often than other IG users [source]




[Detail]

- Some bullying behaviors, such as trolling, are more rare overall but seem to occur more frequently for users with large followings, such as creators, athletes, celebrities, politicians, and journalists.
- There is actually little empirical work both within the company and outside of the company on this phenomena.
- One study invited MPs in the UK to take part in a survey on their experience with online harassment. They found that 30% of the politicians reported experiencing offline verbal aggression (threats, racist or sexist comments, and other abusive language) each week, while 92% reported experiencing online aggression each week through Facebook or Twitter -- unfortunately the authors did not disaggregate these platforms, aso we can't distinguish *where* these threats occured.
- Internally, Instagram has begun to investigate experiences of bullying and harassment among creators (this includes both aspiring and experimental creators) and finds they are also more likely to experience bullying and harassment than other Instagram users. Thankfully, IG creators do not report nearly as a high a volume of physical threats (4%-6%, much lower than the UK MPs).

[Support and Additional Info]

Section in the literature review [link]

- Bauman 2019, Chapter 4 discusses the lack of evidence and studies about trolling in general.

There is a small but important academic literature on how social media can provide support--this is nearly always qualitative.

**Support for Transgender Youth**

"In online environments, transgender adolescents may feel they are able to express ideas that they would not be able to express offline because of stigma and bias."

Article in Journal of Adolescent Health

**Analysis of the #metoo Movement**

Concluding that "the internet and social media" were critical. "That is what propelled this movement and what has given a voice to a group that has historically felt ignored and demeaned."

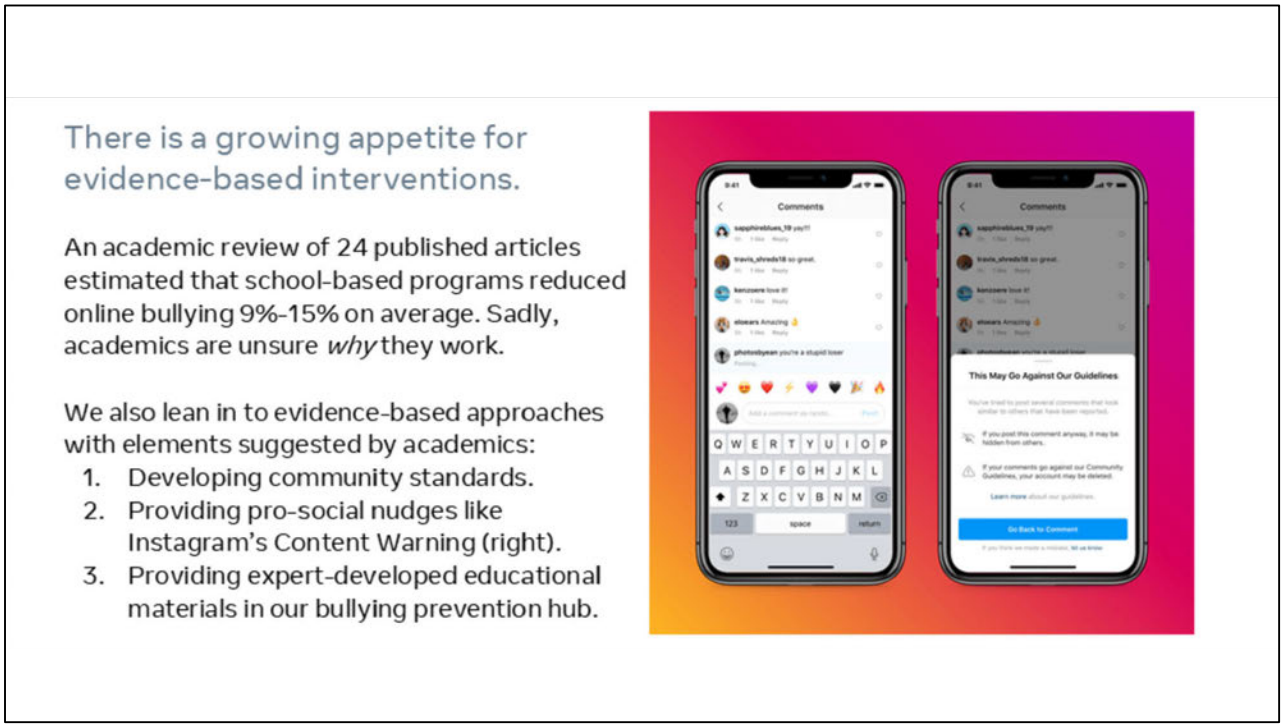
*Political Cyberbullying* (2019), p. 114

[Detail]

- There is a small, but growing literature covering how social media can provide support for bullying and harassment in real life.
- We did not find significant quantitative studies about our platforms in academic literature (though there has been some analysis of the metoo movement on twitter). But here are two examples that will hopefully give you an overview of the influence online platforms can have even on in-person contexts.
- In both cases, the researchers provide a balanced picture of social media; noting difficulties.
    - The left panel is a research article by Elie Selkie, a young career public health professional with a growing interest and research in cyberbullying and support in online spaces. She interviews several transgender adolescents, outlining the importance of having a place to connect with others in order to get social support.
    - The right panel is an excerpt from a book by Sheri Bauman, an international expert in online bullying and the research lead for *Ditch the Label*. She devotes an entire chapter to online harassment, including a discussion of online slut shaming and revenge porn. She ends this chapter on a very positive note, however, noting that social media (specifically citing Facebook and Twitter) were critical tools in shifting important norms in both online and offline spaces.

- Both are light on quantitative data, and both outline difficulties and problems, but they each end on a hopeful note. We hope these studies may represent a shift toward evaluating the positive and supportive aspects of social media and our platforms.

[Support]




[Talking points]

- Governments across the world are investing in school-based programs to reduce bullying and harassment; the US has two ongoing grants through NIH and there is a large school-based study related to WhatsApp funded by the Israeli government that has begun producing empirical research.
- There is a growing and robust academic literature testing school-based programs that include education for students, parents, and teachers, in order to reduce bullying and aggression. Research estimates that these programs may reduce online bullying by as much as 15%.
- Notably, however, many of the recommendations from the literature are already in force on our platforms, such as enforcing publicly available community standards, developing automatic detection for moderation assistance and pro-social nudges.
- For example, we've worked with the Yale Center for Emotional Intelligence to create accessible and free evidence-based guidance on how to start some important conversations for people being bullied, parents who have had a child being bullied or accused of bullying, and educators who have had students involved with bullying

[Support and Additional Info]

Links to the israeli studies on WhatsApp: (Kashy-Rosenbaum and Aizenkot 2020; Aizenkot 2020).

**Narrative Context and Background on Product proposals from literature review**

For many years researchers have developed and studied intervention programs among school-age youth and have even published technical packages outlining programs that work. These approaches are primarily school-based programs and community partnerships (like community-hospital partnerships). Some interventions of these have been tested in the online context, with a recent meta-analysis using randomized controlled trials estimating school-based programs reduce online bullying by an average of 10-15%. Other research reviews outline approaches outside of schools (though these have not been tested in randomized controlled trials). They offer suggestions for parents, peers, health care providers, policy-makers, and social media companies. External Academics have hypothesized (Ansary 2020, at pp. 6-7) that several proposed actions by social media companies may reduce online bullying and harassment, many of which we are already adopting:

1. Enforce community standards and provide reporting mechanisms.
2. Detect bullying and harassing behavior with AI systems.
3. Display notifications designed to promote self-reflection.
4. Message about appropriate online behavior.
5. Prevent anonymous posting.
6. Have robust requirements for user authentication.
7. Promote positive intervention by bystanders.

Unlike school-based programs, there are no randomized controlled trials testing these interventions in published literature. Social science researchers have no estimates how much (if any) these may impact bullying and harassment frequency.

**One-pager Partnership context (with links)**
With a desire to impact this space both on and off our platforms, the company has invested significantly in partnerships with expert organizations. For example, we've partnered with organizations like Project Rockit, on outreach to schools and communities. Through the support of Facebook and Instagram, Project Rockit was able to scale their youth-driven approach to grow 10,000 Digital Ambassadors from 600 schools over 2019-2020 and connect them to a powerful network of peers creating social change. We've also worked with the Yale Center for Emotional Intelligence to create accessible and free evidence-based guidance on how to start some important conversations for people being bullied, parents who have had a child being bullied or accused of bullying, and educators who have had students involved with bullying. Through a partnership with the International Bullying Prevention Association, we provided in-person and virtual bullying prevention training to over 1,000 participants around the globe and developed new resources for families, youth and educators that are now available on the Facebook safety center.


Next Steps
1. Additional Research Investments
2. Research Supported Product Strategy
FACEBOOK

## We are Building off of Significant Investment

There are a number of research teams and product teams who have been working in this important area, and making significant progress -- many of the presentations you see later today will incorporate their work. Below, I highlight a few current research POCs.

- Instagram (Diego Castenda; Matt [redacted]
- Facebook App (Rottem [redacted]
- Messenger (Mary [redacted] Hannah [redacted]
- Central Integrity (Younghee [redacted] Mark [redacted]

How can we best leverage our collective wisdom to tackle the problem and have closer parity?

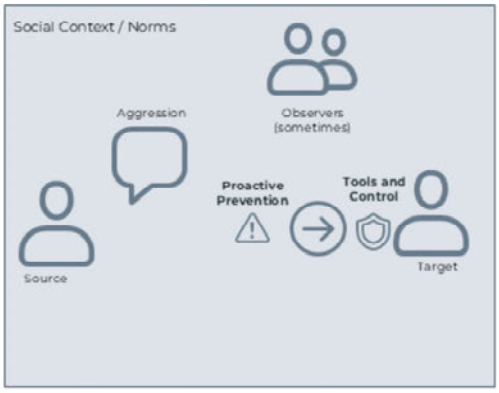
I wonder if this should go right after slide 16. So the flow would be something like, "first, a LOT of good work has been done in this space. you'll hear more about it later today. Here's what we're doing specifically in terms of foundational research."

## Snapshot of B&H research from company well-being working team

| Projects | Short Description | Research Lead(s) | Status | ETA |
|---|---|---|---|---|
| Strengthen Internal Foundation | Understand recent social science on each topic to ensure that the company's communication, research, and product work are guided by the strongest evidence. (*E.g.* External Literature Review; Social Science Note) | Team | Delivered | 6/2021 |
| Bullying interventions | Collaboration with Instagram to evaluate and publicize the impact of products intended to reduce bullying and harassment. [link] | Bryce | Execution | 11/2021 |
| Bullying policies | A survey experiment evaluating the impact, effectiveness, and interpretation of publicly accessible resources to prevent bullying and harassment, including things like our Community Standards. [link] | Iffat | Execution | 12/2021 |

## Research-Supported Product Strategy

1. **Proactively prevent and reduce** the amplification of bullying and harassment on platform (e.g., ranking, goaling).
2. **Provide people with the right tools at the right time** to prevent and/or navigate bullying and harassment on platform (e.g., controls).




"we also did an audit to see how well we're doing in our product strategy and learned some things that hopefully will inspire this group:
Overall, the company has made a large investment in bullying and harassment, especially on prevention focused tools on Instagram and on reducing amplification via proactive detection of policy violating content on FB and IG. "

Then as you move into examples:
"However, the different apps are not holistically aligned in their approach, and there are still significant gaps in our mechanics for high responsibility issues like brigading and mass harassment, and in the support and resources we offer to people who encounter bullying and harassment, particularly youth."

## Examples of Potential Parity Across Apps

*Proactive Prevention*

1. Interaction "cooldowns" for people who have repeatedly been reported or flagged by our classifiers
2. Improve ML detection of mass harassment / brigading by investing in proactive detection
3. Kindness reminders and nudges

*Support & Resources*

1. Additional moderation tools (bulk hide, delete, report option) across surfaces and apps
2. Ability to turn off comments/messages across surfaces and apps
3. Better timely upsells of safety features (ex: turning off message sends) to public figures/creators


Appendix
FACEBOOK

## Project at-a-glance: Bullying interventions

| | |
|---|---|
| **Description** | On-platform experiments on Facebook, Instagram, and/or Messenger to evaluate and publicize the impact of products intended to reduce bullying and harassment. The goal is to demonstrate concrete proof points that we develop products to *prevent* bullying and harassment and *empower* users to act. |
| **Research questions** | 1. How much do our product interventions impact bullying and harassment?<br>2. What are the most effective interventions in reducing bullying and harassment? |
| **Method** | On-platform experimental approach (a/b tests). Outcomes will include log data and behavioral outcomes as well as survey measures (knowledge, awareness, etc.). |
| **Impact** | - Sample talking points: "product xyz reduced estimated incidence of bullying and harassment by xx%."<br>- Deepen existing anti-bullying partnerships, and develop relationships with academics.<br>- Blog publication co-authored with current partners (e.g., International Bullying Prevention Association, Cyberbullying Research Center, or the ABC Bullying Centre) and a publicly recognizable internal leader. A technical paper and data would be available on request for replication. |

## Project at-a-glance: Bullying & Harassment policies

| | |
|---|---|
| **Description** | A survey experiment that evaluates the impact, effectiveness, and interpretation of publicly accessible Community Standards that detail our guidelines on bullying and harassment. |
| **Research questions** | 1. Do people understand Facebook's bullying and harassment Community Standard?<br>2. How closely do such resources match with lay interpretations of bullying and harassment?<br>3. How well do target audiences of our bullying and harassment prevention hub (teens, parents, etc) understand the resources? |
| **Method** | Survey experiment. People review and evaluate a vignette via survey: half evaluate the vignette before reading the resource and half after reading the resource. Both sets answer questions about bullying and harassment and their interpretation of the resource. |
| **Impact** | - Identify areas of strength and areas for improvement in our public-facing resources<br>- Identify what people think will *improve* their understanding<br>- Get measurable insights on demographic variation (locale, age, etc.) in interpretation of activity believed to be bullying and harassing. |

## Prior studies in the same vein …

| | |
|---|---|
| **Hate Speech** | 1. Reviewing Facebook's policy does not improve accuracy and may even lead to a reduction in accuracy in assessing what is and is not allowed on the platform.<br>2. Most adults find some part of the policy hate speech unclear. Phrases that detail "grey" areas or exceptions are most often selected as unclear.<br>3. Peoples (mis)interpretations of the policy is associated with their sentiment towards the policy. |
| **Regulated Goods** | 1. Reviewing the complete regulated goods policy language leads people to become both less accurate and more certain in assessing whether content would or would not be allowed on our platform.<br>2. Again, a majority of U.S. adults find some part of the policy language unclear, especially "grey areas".<br>3. There is a counter-intuitive relationship between approval and reviewing Facebook's regulated goods policy. Those who were less accurate were more likely to approve. |
| **In-progress** | 1. Evaluating Facebook's V&I policy and updated hate speech policy in 5 markets including the US, UK, FR, MX, IN and ZA(hate speech only)<br>2. Deeper analysis on why language is selected as unclear |

## Instagram

| Theme | Problem Statement | Outcomes |
|---|---|---|
| Awareness of interventions | 1. Users experiencing harmful interactions can be unaware and therefore potentially underutilizing systematic self-remediation tools that could protect them from bullying and harassment. | 1. Users are upsold SR tools when they need them most.<br>2. Users access SR tools to effectively reduce bullying and harassment experiences. |
| Impact of interventions | 1. Users are faced with a growing number of self-remediation tools, some of which may be sub-optimal at improving user perceptions of supportiveness, effectiveness, and prevalence of bullying and harassment.<br>2. We don't understand how to optimally unify Reduce and Support for improving user experience. | 1. Users are upsold the interventions that have the most positive impact.<br>2. We identify which types of tools are most impactful to guide further development and investment, as well as which are least impactful and may need revisions<br>3. We develop a framework to decide if reduce or support strategy should be used for specific user problems. |
| Reduce and Prevent Harmful Interactions | 1. Users (particularly high value users / athletes) continue to experience harmful interactions that are preventable on our platform.<br>2. We don't know how the comments we show incentivize or deter the initiation of HI, or report of violations.<br>3. We don't know how best to leverage account value signals in comments.<br>4. We don't know how to effectively identify new ways of crowd-source community moderation of bullying, harassing, and racist comments, including those that do not necessarily violate the community guidelines. | 1. We are able to identify and prevent more harmful interactions from reaching targets.<br>2. Insight into comment ranking/showing signals to reduce HI.<br>3. Insight informing possible new products or design features enabling crowd-source community moderation of racist comments. including those that do not necessarily violate the community guidelines. |

**Messenger TRIPS data shows that B&H has higher reach relative to other high severity integrity problems.**

