

Plaintiff's Exhibit Markers	
π PLAINTIFF π	United States District Court Northern District of California
	Case No. <u>MDL No. 3047 (N.D. Cal.)</u>
	Case Title <u>In re: Social Media Adolescent Addiction</u>
	Exhibit No. 2737
	Date Entered _____
By: _____ Mark Busby, Clerk _____, Deputy Clerk	

“Bad Experiences” Measurement

Plan for a 2021 plan



Goals

1. Share ambition and key examples of problems to solve
2. Share high-level approach
3. Feedback on next steps

Meeting Objectives

Non-Goals

1. Follow up on [Oct 15 review of TRIPS proxy success criteria](#)
2. Roadmapping-level detail
3. Product strategy for bad experiences
4. “Positive Experiences” measurement (ex. promoting positive well-being)

Our mission is to create the safest and most supportive global community.

We envision an Instagram where, everyone, especially teens, creators and underserved communities, feel safe and supported to express themselves and to push culture forward.

Supportive - Helping people with bad experiences, by reducing and resolving them. Not generally focused on enforcement.

Policy Violating - Problems defined in the company's Community Standards content policies

Borderline - Non-violating content policy intended to address part of policy narrowness

Bad Experiences - Problems defined by people's perception, as implemented in survey taxonomies like TRIPS and Hard Life Moments

Legitimacy - People and external stakeholders believe that our integrity work is effective at reducing harm and that our enforcement is defensible and fair. Both 1st and 3rd party perceptions

Reach - Measures the size of bad experiences (number of people affected by them or proxy)

Support Effectiveness - Measures how effectively we're resolving a bad experience. For people experiencing the problem (1st party), intersecting with 1st party part of legitimacy

Our company integrity mission is 'protect the community and its voice'. Our policies have to balance both protecting the community and its voice.

People feel they're having a bad experience or they don't. Their perception isn't constrained by policy balance.

Understanding “bad experiences” enables us to see the fuller picture -first picture of how people experience safety and support concerns, agnostic of policy.

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

Types of gaps between policy and bad experiences

Policy can be effective, but ...

1. incident doesn't meet the policy bar though clearly bad
2. incident doesn't fit current policy definitions
3. incident is hard to detect (by human or automated review)

Policy isn't the right tool when ...

4. "bad" is impossible to determine with any consistency
5. enforcement actions (hard or soft) cannot address the problem

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.

Billie Eilish Admits That Reading Instagram Comments Was “Ruining” Her Life

"The cooler the things you get to do are, the more people hate you."



Chrissy Teigen • Rita Wilson
• Wilfred Zaha • Dua Lipa
• Milana Vayntrub

HOME > CULTURE

The actress who plays AT&T's Lily is facing waves of online sexual harassment, including manipulated images and objectifying memes



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

 **Comment**



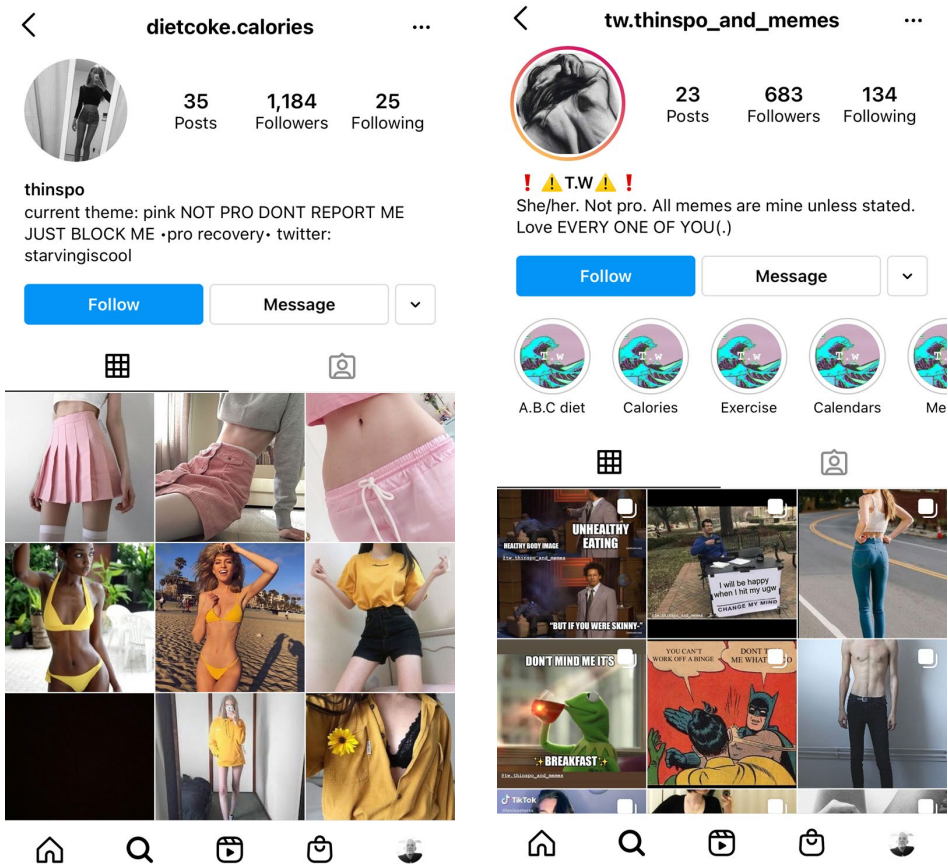
FIRST PERSON

Nicola Thorp 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.

CONTEXT - EXAMPLES: THINSPIRATION



"We're not teaching [others] to say, hey are you okay? We're teaching them to click a report button or to call the cops and try and trace the phone or trace the location based on a Tweet. I'm kind of afraid of the resources because of that."

2021 Proposal

What does success look like by end of 2021?

**Operationalize “bad experiences” measurement for
experiment launch decisions and goaling to
empower teams to work on problems in policy gaps.**

Measurement Tracks

Reach

Measure the number of people affected by bad experiences

Support Effectiveness

Measure how effectively we're resolving a bad experience

*Not covered here: 3rd party
("reputational") reach or
intensity separate from reach*

Reach measurement in 2020

*Not covered here: Hard Life
Moments for mental well-
being concerns*

Violating prevalence impressions

Labeled and classifier estimated metric for narrow policy violating definition. Available for a subset of problems based on CI + IGWB prioritization and operational cost.

FRX

Only negative user feedback with problem detail, with problems as defined by Community Standards policy. Available as aggregate for all reporting problems (dominated by spam) and by violation types and problem tags.

TRIPS

Survey for 13 integrity problems launched on IG in Dec'19. Also has Civic variant.

H2 learnings about Reach measurement

Violating prevalence impressions

While responsive to some changes, doesn't capture most personalized demotion impact in comments (similar on FB XI).

FRX

Per-product / per -surface cuts matter for seeing per-experiment impact. "FRX feedback" more inclusive than "FRX submit reports".

TRIPS

Responsive to IG Bullying H1 changes in comments slice (-5%).

Responsive to world events, including at sub -problem detail:
Since George Floyd protests in June, in the US "threats" portion of B&H and "political beliefs" for HS are elevated (ss)

H2 learnings about Reach measurement (continued)

USI

Most weight goes to FRX when predicting policy violations.
Operationally challenging to add new signals in current
central governance model.

pTRIPS

Super early results show some stat sig effects for 2 IG
Bullying experiments (Remove Preview Filter Toggle,
Personalized Demotions). TBD if surveys validate prediction.

Product cuts for both pTRIPS proxy and TRIPS surveys
probably quite important.

FB XI pursuing similar approach in close collaboration with
IGWB and CDS working group. Different in details.

Support Effectiveness measurement in 2020

Adoption

See how many people use our new support experiences (ex. Support Requests, Restrict, Bulk Comment Management)

Supportiveness

IX-developed survey: “How supportive was Instagram during this experience?” triggered after completing a specific flow

LEGIT

Covers much more scope than effectiveness for people experiencing problems (1st party). Responsive to actor transparency changes on IG.

How to Get There

Major Next Steps

1: Finish H2 execution

TRIPS holdout read for IG bullying, pTRIPS validation, USI development. We'll have new data before Dec 17 roadmaps.

2: Coordinate with CI IMI, OC, IX

At least CI IMI is actively proposing roadmap priorities that would help advance work in this area.

3: Prioritize H1 steps to unblock IG teams

Which metrics would unlock the most impact? What needs Foundation and IMI help vs. being driven by teams?

4: Decide IG Bullying goal / launch metrics

H1'21 is a key inflection point for IG Bullying in its 5th half of working in this direction. WB is taking Bullying P0 goal.

Discussion

Feedback on ambition and examples?

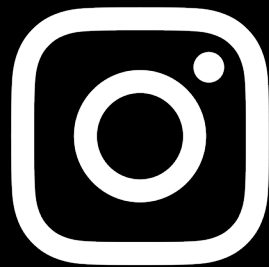
Is this ambitious enough? Too ambitious? Resonate as important and people-first?

Feedback on approach?

Is the framing intuitive? Leverage existing concepts and terminology for “extreme clarity”?

Feedback on next steps?

At a high-level, are we missing major steps for getting ready for H1 roadmaps?



CONTEXT · EXISTING METRICS

Track	Metric	Question metric answers
Reach	Violating prevalence	How many impressions are seen for content violating CS policy?
	FRX	How many people report content and accounts?
	TRIPS	How many people have seen or experienced the problem in last 7d? 5-point intensity?
	USI	What's the aggregate of reporting, blocking, unfollowing, and other actions? (with aggregate weighted for correlation with CS violations)
	pTRIPS	Of the aggregate of user actions and other signals, what's our prediction for TRIPS?
Support Effectiveness	Adoption	How many people adopt the supportive product solution? Does it cannibalize?
	Adoption within problem segment	How many people with a specific problem adopt the supportive product solution?
	Supportiveness	How supportive was Instagram during this experience?
	LEGIT	Do people believe that our integrity work is effective at reducing harm and that our enforcement is defensible and fair? 2737_0023

CONTEXT · ASSESSING EXISTING METRICS

Track	Metric	Shows progress in policy gaps?	Already proven in Nov'20?	More progress landing soon?	Works for low - reach?	Team(s) with most progress
Reach	Violating prevalence	No	Yes	Surface improvements landing	No (except UB Prevalence)	CI OC + IG Foundation
	FRX	Yes	Yes, for some use cases	Improving logging quality	?	CI IX, CI OC, IGWB Foundation, IG Bullying, US2020
	TRIPS	Yes	Yes, for some use cases	Major new results landing in Dec	No	FB XI, IG Bullying
	USI	No	No, not operationalized	H2 IG project in flight	?	CI IX, CI OC
	pTRIPS	Yes	No, not operationalized 2737_0024	Major new results landing in Dec	?	IG Bullying, FB XI

CONTEXT · ASSESSING EXISTING METRICS

Track	Metric	Shows progress in policy gaps?	Already proven in Nov'20?	More progress landing soon?	Works for low - reach?	Team(s) with most progress
Support Effectiveness	Adoption	?	Yes		Yes	IG Bullying
	Adoption within problem segment	Yes	No, initially proposed for Limited Profile		Yes	IG Bullying
	Supportiveness	?	Yes, but for different use case		?	CI IX
	LEGIT	?	1 early proof point for different use case		?	IGWB Foundation, CI Legitimacy

META3047MDL-019-00094148

Metadata

All Custodians	Andrews_Kyle;Bejar_Arturo;Castaneda_Diego;Diuk_Carlos;Hayashi-Branson_Hitomi;Malahy_Lori;Mummareddy_Satish;Otaru_Sayed;Parker_Sam;Rothschild_Miki;d'Alessandro_Brian	SEMANTIC
All Paths	/Andrews_Kyle;/Bejar_Arturo;/Castaneda_Diego;/Diuk_Carlos;/Hayashi-Branson_Hitomi;/Malahy_Lori;/Mummareddy_Satish;/Otaru_Sayed;/Parker_Sam;/Rothschild_Miki;/d'Alessandro_Brian	SEMANTIC
Primary Date	02/01/2021	DOC_TYP E_ALIAS