Tennessee v. Meta
Case No. 23-1364-IV
PX-04628

Emerging Reels Integrity Challenges

Alison Lee

1

EXHIBIT
4
depobook.com

I reviewed the related body of Reels, integrity, and well-being (equity, teens) research (link) to identify key priority areas for Reels.

TLDR: The Reels ecosystem makes it more vulnerable to integrity challenges.

1 There are **multiple** Reels teams launching a **high volume of product improvements**, leading to increases in sexually suggestive and objectionable content.

Reels product teams should **coordinate** to release launches strategically, and **consider integrity impacts** proactively for new product launches

2 The focus on **entertainment / mimicry,** and the **teen audience** leaves Reels more open to 1) sexually suggestive and offensive content and 2) vulnerability of younger people.

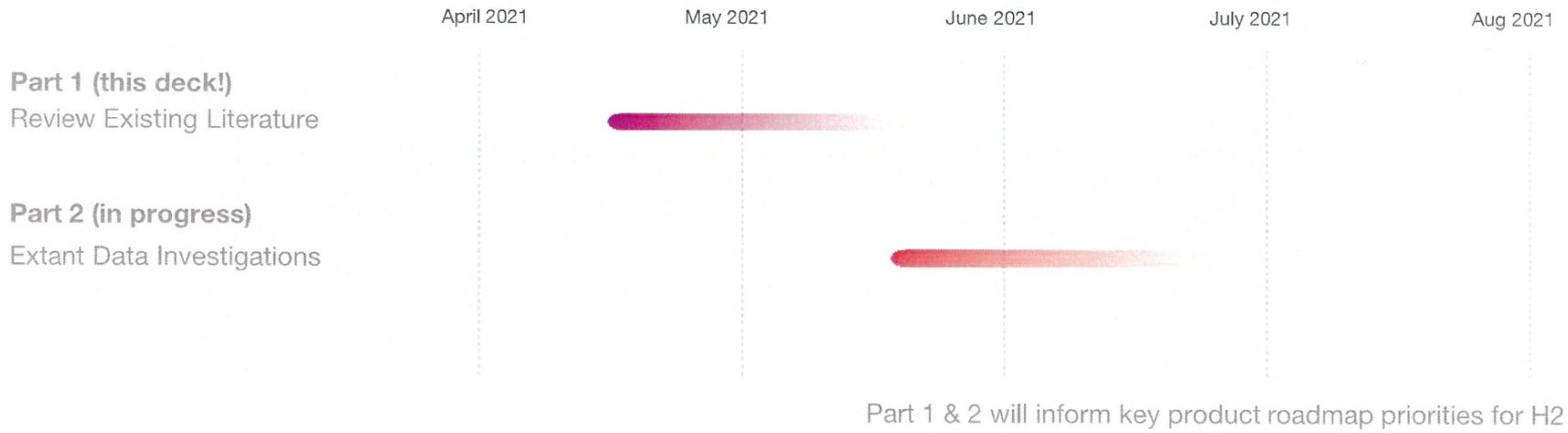
Reels teams should build user-centered tools (audience controls, recommendation feedback) **and preventive mechanisms** (i.e., tuning ranking models, protections for minors) to keep users safe.

3 **Teens, women, LGBT people, and communities of color** are more vulnerable to harm and marginalization on IG. Early research suggests these inequitable patterns are reproduced on Reels.

Reels teams should use a **targeted universalism approach** to reduce inequities of harm and improve socio-culturally diverse content, resulting in **wins in integrity, engagement, and public perception.**

Timeline - Reels Integrity Review

Scoping and Prioritization



Part 1 & 2 will inform key product roadmap priorities for H2

There are multiple (15+) teams working on Reels launching a high volume of product improvements (100+ on 2021 H1).

This product pace and complexity is leading to **greater volatility in integrity, especially increases in sexually suggestive content.**

Reels is a quickly evolving surface, with lots of moving parts.

We released a high volume (100+ in H1 2021) of product launches, 17 leading to integrity regressions.

The pace at which these launches occur make it difficult for us to identify the additive, interaction or long-term integrity effects of each launch.

Multiple Teams simultaneously work on Reels.

Including the Reels Recommendation Relevance team (3 sub-teams), Reels & Camera team (4 sub-teams), Reels Sharing (3 sub-teams), Reels Engagement, Family Reels, as well as other non-surface specific teams such as Integrity, Well-being, and Equity.

We have less sophisticated knowledge + tools for empowering users to address harm.

Example: tuning the 'Not Interested' button to better customize Reels ranking can help recommendations teams better personalize recommendations and integrity teams document "grey area" content.

Reels presents new and unique integrity system challenges.

Subsurfaces on Reels have different integrity challenges.

Example: increases in objectionable content promoted by our ranking models result in bigger spikes especially on Reels Cover and In-Feed Unit (RIFU).

System for identifying harmful content in Reels is harder than other content review systems.

Systems for identifying harmful content in short form videos works less well compared to images and text, takes longer, and is harder to judge/less accurate.

Less is known about what, how, and why certain kinds of integrity challenges come with the surface.

For example, while the child safety team has done extensive work to improve child safety generally across IG, Reels is not yet a focus. This is concerning, given much of Reels (and TikTok) came from viral dance trends consumed and created by teens and kids.

So what?
Now what?

1. How can teams more proactively consider **integrity impacts in their product decisions and strategy**, both in early design phases as well as in process (i.e., using guardrails)?

2. How can teams communicate and launch new product features strategically so that **we can move fast and in sync without compromising safety?**

3. How might Reels product teams understand **where (and for whom) disproportionate harm is limiting engagement?** What improvements would lead to reduction of harm and more compelling Reels experiences?

Mimicry, trends, and entertainment: great for engagement, not so great for integrity.

Who's at greatest risk?

Trends, mimicry, and entertainment are leading to increased risk of harm.

Ranking models cause integrity regressions by increasing the prevalence of non-recommendable content.

Increasing **topic diversity** increased all objectionable content, and recommending based on **P(click), Mimicry**, and **downranking watermarked media** increased the prevalence of sexually suggestive content.

Trends, Entertainment, and Mimicry can lead to harm.

Risky trends (i.e., the TikTok WAP trend for teens or trends mocking disabilities) may leave people open to harm, ranging from seeing harmful content to bullying or predatory behaviors. **Offensive content** (hate speech, violence, bullying) may also be delivered under the guise of "entertainment" at higher rates on Reels.

Unconnected dancing videos featuring young girls raise safety concerns.

Studies and internal feedback suggests that people are uncomfortable with seeing and concerned about the safety of unconnected Reels content of tween and young girls dancing and posing.

So what?
Now what?

1. How might we leverage **multiple ML approaches** (relevance, non-recommendable content, age modeling) to identify when particular trends may leave people - especially young girls - particularly vulnerable to harm?

2. What expansions in **non-recommendable policies** might we need to implement to ensure ranking efforts are not pushing unpleasant or uncomfortable trends or content to users?

3. What **design** (self-remediation) and **engineering** approaches might better support safer experiences? (*e.g., giving people more control over recommended content, limiting who can see content featuring young people, etc.*)

Teens, women, LGBT people, and communities of color are more vulnerable to harm and marginalization on IG.
This holds true for Reels.

Teens are the target audience for Reels, but we have limited knowledge, tech, and tools to keep them safe.

But, here's what we do know.

Teens hear about and see integrity issues at higher rates on Reels, but they're less bothered by them than adults.

Teens especially see and hear about more profanity and nudity, as well as (civically-related and other) bullying and hate.

Teens don't think deeply about safety risks until something bad happens.

Consider this with 1) how frequently content featuring minors shows up on Reels recommendations (see slide 9), 2) teens are fueled to produce by likes and comments, and 3) teens are more willing to talk to people who are "friends of friends".

Teens want more control over who sees their content on Reels

While this finding focused on audience control out of social risk (*i.e., being embarrassed by content)*, it aligns with research on teens' preference for privacy, control, and awareness to address harm.

We must start by naming that integrity work is inherently equity work.

Anytime we talk about harm (who is most vulnerable to it, whose needs and experiences are represented and designed for), we cannot avoid its explicit connection to broader societal patterns of marginalization and power - and therefore, our role in addressing inequity of safety, agency, and belonging on our platform.

Inequities in safety, belonging, and engagement are replicated in Reels.

Harm not effectively covered by our policies has led to users perceiving IG as unsupportive, exclusionary, and silencing.

Users perceive that harm towards marginalized groups are **under-enforced** and voices of marginalized groups are **over-enforced** across all IG products.

Lack of Representation and Social Comparison are barriers to engagement on Reels.

Low distribution, lack of representation, and no credit to original creators lead to under-engagement and people believing that "Reels is not for people like me." Social Comparison on Reels may especially impact teen girls and non-binary teens.

Global differences in integrity needs require inclusive solutions.

Different countries vary on how they (esp. young women) experience and respond to harm. Empowering global users would require that integrity tools are designed for varying levels of language fluency and global needs.

So what? Now what?

1

Promote Socio-Culturally Diverse Creators.

TikTok is perceived to be more effective at distributing socioculturally diverse and inclusive content. **Promoting socio-cultural diversity of creators** may both increase engagement and mitigate feelings of inequity, and help us win with **Gen Z,** who are more racially and ethnically diverse, and more socially progressive than past generations.

2

Design for people closest to harm and exclusion through a targeted universalism approach.

What segments of people experience harm and exclusion at greater rates, and what can we do to improve their experience? What kinds of inequitable harmful experiences (under/over-enforcement) are not effectively covered by our community guidelines and integrity systems?

3

Invest in age-specific engineering, research, and design.

We need to better understand how to protect and support teens in Reels. How might this work sync up with broader priorities to support and protect younger audiences across IG and the family of apps?

Final Thoughts

Investing in Equity-Centered ML

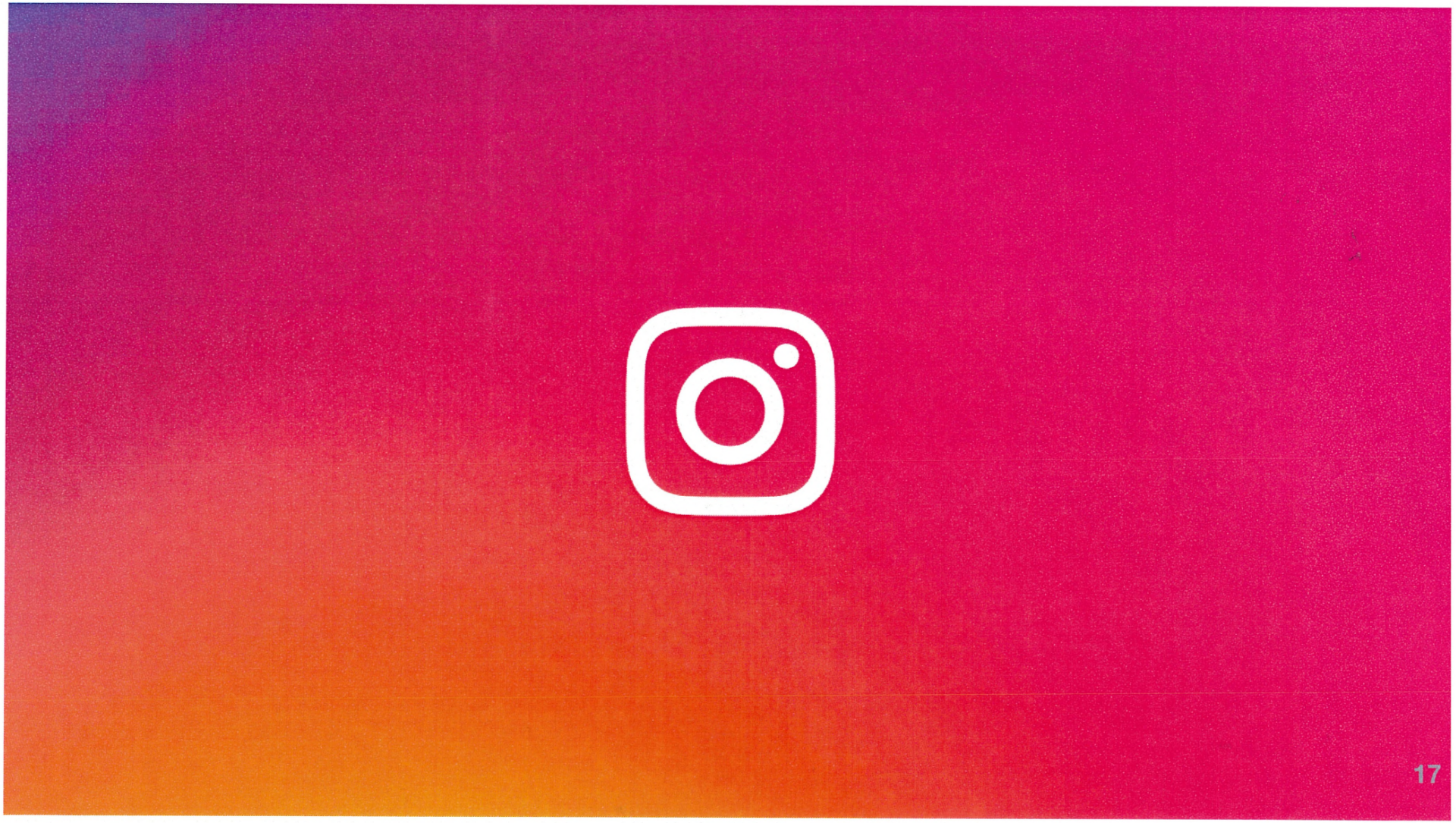
...could be a great opportunity to win in reducing harm, increasing engagement, and improving IG's public reputation.

Designing for Teens and Youth

...requires us to know who they are, and what they need to be safe. We need to invest in both user controls and engineering approaches that keep them safe, especially in a product designed for young users.

Integrity is central, not in opposition, to engagement

Although some approaches to mimicry, trends, and entertainment may lead to small engagement wins, they ultimately make the space less safe, less inclusive, and less friendly for diverse and global audiences. We must center designs that enable us to win at both engagement and integrity.


17

Citations and Literature

...can be found in the Workplace Note here.

Huge thanks to the following folks whose work and labor informed this review: Lauren Cheatham, Brian d'Alessandro, Caro Diaz, Millie Harrison, Hitomi Hayashi-Branson, Stacey Houston, Randy Illum, David Kille, Tal Kimerling, Karen Krieb, Daniel Lin, Rong Liu, Michael Miner, Shilpa Mody, Femi Olanubi, Rottem Sagi, Alana Shine, Lara Vujovic, Jerry Yu, Qiaojun Wang, Jennifer Yih, Zhanna Zhanabekova.