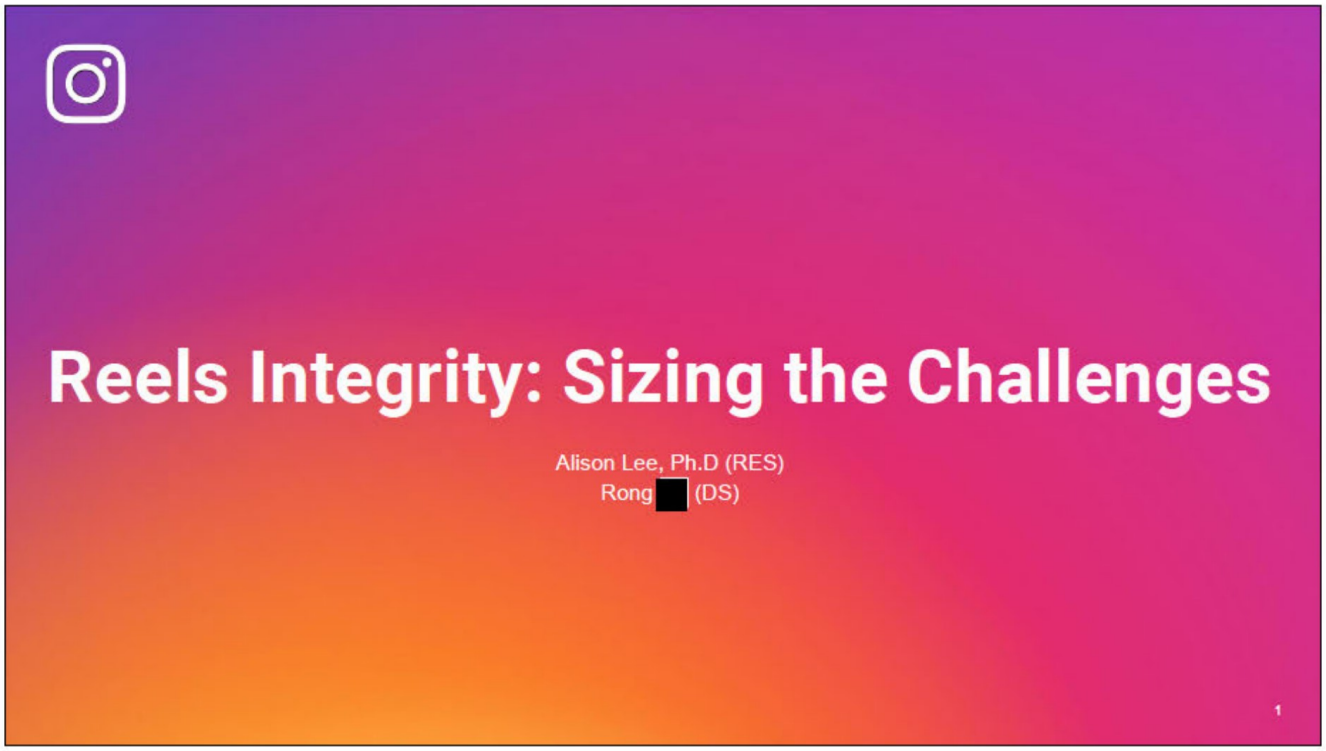
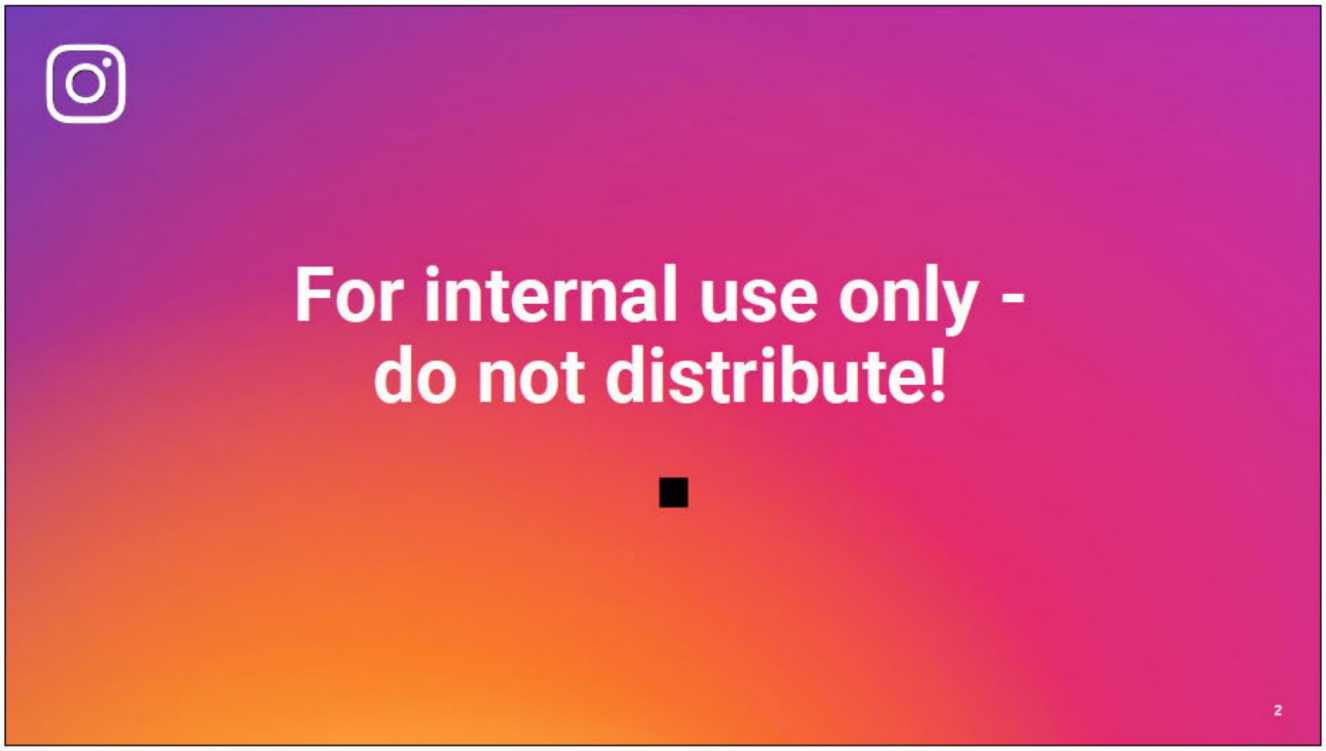
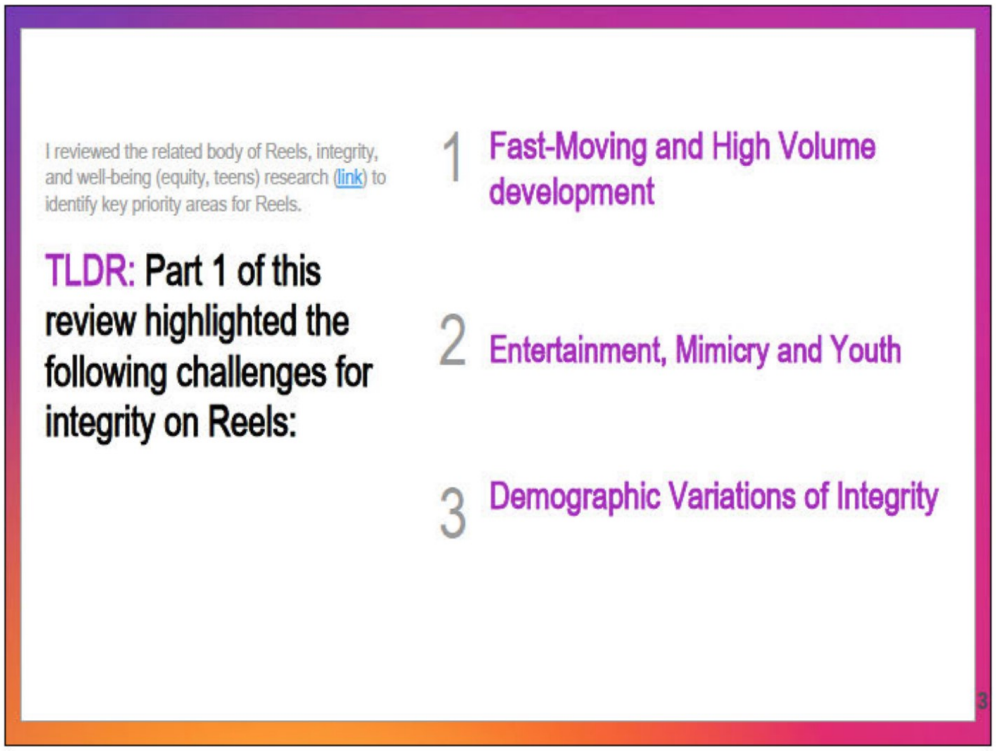
Tennessee v. Meta
Case No. 23-1364-IV
PX-00093

Document Provided in Native Format



This deck serves as a summary; full **Workplace Note with citations/literature here**

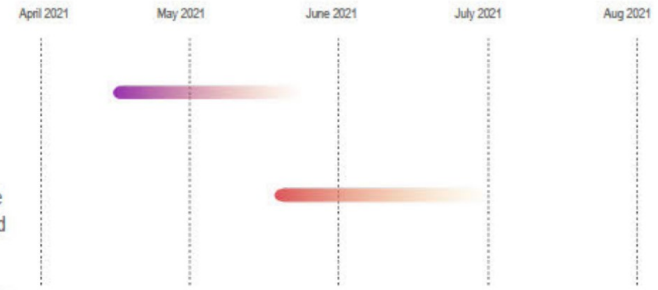

For internal use only -
do not distribute!
2



Workplace Note with citations/literature here

Timeline - Reels Integrity Review

Scoping and Prioritization



Part 1
Review Existing Literature

Part 2

Extant Data Investigations: use existing, internal data (FRX and prevalence data) coupled with account demographic data (gender, zip code, age, country) to size and scope integrity challenges

Part 1 & 2 to inform key product roadmap priorities for RI in H2 2021

4

There are multiple (15+) teams working on Reels launching a high volume of product improvements (200+ on 2021 H1).

This product pace and complexity is leading to **greater volatility in integrity.**

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

We released a high volume (200+ in H1 2021) of product launches, 30+ 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.

Gaps on recommended surface experimentation cause unexpected integrity impacts.

Example: biases between CEPd and PD can cause disparities between QE pretests and post-launch topline impact.

6

RONG - item 1 & 3

A full list of Reels launches: [redacted]

Reels presents new and unique integrity system challenges.

Subsurfaces on Reels have different integrity challenges.

Example: Reels Tab has higher prevalence of regulated goods, Cover of disturbing content, and RIFU of sexually suggestive content.

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.

Features of Reels (see next section) create unexpected challenges for integrity models.

For example, the fast paced nature of viral trends require faster recurring training for models to keep up with non-recommendable spikes.

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?

Trends, Entertainment, and Youth: great for engagement, not so great for integrity.

Who's at greatest risk?

Part 1: 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.

Part 2 Data Insights: Objectionable and Sexually suggestive content are higher on Reels.

Sexually suggestive (non-rec), hate (violating), and violent (violating) content are both reported and detected at higher rates on Reels. (cit 1; cit 2)

There are also significant discrepancies between what we detect, what users report, and what reports are actioned on, suggesting over/under-enforcement challenges.

Meaningful Interactions are highly related to harmful content.

Interactions such as clicks, likes, private shares, and comments are associated with sexually suggestive and hate speech content across IG. (link)

Trends contribute significantly to integrity challenges.

12-23 of Top 50 sexually suggestive media are part of trends (via weekly review). However, boosting trends overall seems to result in integrity wins (i.e., reduction of prevalence).

Sexually Suggestive Content featuring Minors (SSCM) is 2-3x higher on Reels than Explore*.

Female and unknown gender teens, as well as unknown gender non-teens, create the most SSCM. However, we're also over-enforcing by labeling birth announcements and parenting videos under this category. (DM me for quip on Protecting Youth!)

*for data pulled from 04/01-04/30

11

Alison - elaborate on #1

Part 2 Data Insights (cont'd): Users perceive and respond to different kinds of harm differently on Reels.

Profanity, Misinfo, and Civic Misinfo continued to be the most prevalent issues, but they also made people feel least bad compared to other integrity issues...

...While witnessing Hate, Bullying, and Violence on Reels were among the most upsetting.

Integrity issues may cost us engagement— people most often report scrolling past, exiting Reels, and exiting Instagram when having a negative experience on Reels rather than using other remediation tools. (1, 2)

The fourth most used remediation behavior, selecting "not interested", provides both users agency over their experience and gives us signals about how to improve. We should invest in these personalization tools to improve engagement.

The highest perceived reach integrity types are the least detected, and vice versa.

People didn't often report the highest labeled prevalence integrity types (sexually suggestive, violence, graphic content). Similarly, hate speech and bullying are some of the lowest detected prevalence on Reels, but participants reported witnessing those more often and report them as being more upsetting.

12

For more on users' perceptions of harm on Reels, check out this first of a series of notes: https://fb.workplace.com/notes/363135165322489

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 - part cularly 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.*)

13

Teens, women, LGBT people, and communities of color are more vulnerable to harm and marginalization.

How do we solve this in Reels?

14

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.

Part 2 Data Insights:
Teens on Reels...

Teens report (23-60%) more objectionable content across all violation types... compared to their DAU (Except for misinfo)

...but don't *consume* more objectionable and sexually suggestive content.
(Except Regulated Goods content on Reels cover, which they view more than non-teens).

Female and "Unknown Gender" Teens create 73% of "SSCM" content created by teens...

...While male teens *report* the most. "Custom Gender" Teens both create (1.2%) and report (1.9%) "SSCM" content at higher rates than their DAU (.4%).

But the group who creates the most SSCM, creating 32% of ALL SSCM content, are "unknown gender non-teens".

16

Rong


How do we tackle the multiple Youth Safety issues at play here?
1
"True" SSCM: Improving our detection of minors, as well as getting ahead of problematic trends (i.e., upcoming dance trends) that may put children at risk of producing suggestive content can help us keep them safe.
2
1
1
Benign but still risky content featuring minors: Children dancing/posing unto itself is not problematic, but it becomes an issue if bad actors see it. Tolerance for this content varies by country and is not preferred for many non-teen users. Preventing IIC actors from seeing/interacting with minors, more effective / child content personalization (SFPLT) or creating "teen-centered" ranking approaches for teens may help solve this problem.
3
1
1
Baby nudity. Much of what we're proactively detecting are baby and parenting videos that users do not find problematic. We should examine whether we need to revisit policies around baby nudity.
17

Alison

Inequities in Integrity:
We must start by naming that integrity work is inherently equity work.

Anytime we talk about harm, we cannot avoid its explicit connection to broader societal patterns of marginalization, representation, and power - and therefore, our role in addressing inequity of safety, agency, and belonging on our platform.

Part 1: 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 harmful content targeting marginalized groups are **under-enforced** and voices of marginalized groups are **over-enforced** across IG.

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." A culture of "non-relatability" and social comparison on Reels is discouraging for everyone, but 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.

Part 2 Data Insights: Demographic variations of harm are replicated in Reels.

(DM me for quip of these findings)

Globally, male users create and consume significantly more objectionable content than female users, while female users create more sexually suggestive content.

Of the objectionable content male users view, 2/3rds are regulated goods (drugs, guns, etc.) and ¼ are violent content.

India views more objectionable content, while Brazil and Argentina create and view sexually suggestive content (especially with minors) at greater rates.

By contrast, Turkey and Japan report disproportionately more sexually suggestive content, despite creating nearly none of it. These data point to cultural variations of tolerance for this content.

US zip codes with lower predicted education levels create higher rates of sexually suggestive content.

While education levels covary with other demographic factors like income and urbanicity, this raises questions of both inequitable protection (particularly of socio-economically marginalized women and girls) as well as inequitable over-enforcement.

Early "Fairness" assessments of Reels non-rec models surface biases.

False positive rates and other precision measures are weaker for communities of color on objectionable prevalence detection, and for women and custom gender users on sexually suggestive prevalence detection.

20

- RONG

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 Examine patterns of over/under-enforcement, and whose most marginalized/unprotected.

What segments of people experience harm and exclusion at greater rates on Reels, 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 youth-centered 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?

21

Final Thoughts

Investing in Equity-Centered ML

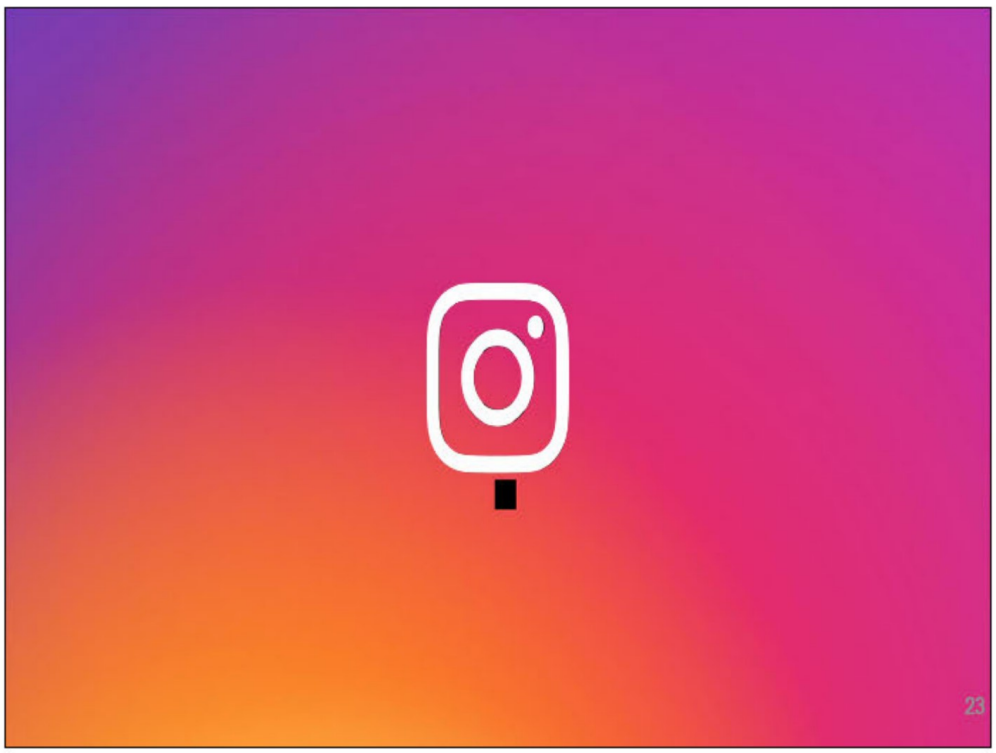
...could be a great opportunity to win in reducing harm, increasing engagement, and improving IG's public reputation.

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.


23

Citations and Literature

...can be found in the Workplace Note here, while data insights can be found in Rong [redacted] DS note, here.

Huge thanks to the following folks whose work and labor informed this review:
Lauren [redacted] Brian [redacted] Caro [redacted] Millie [redacted] Hitomi [redacted]
[redacted] Stacey [redacted] Randy David [redacted] Tal [redacted] Karen [redacted]
Daniel [redacted] Rong [redacted] Michael [redacted] Srilpa Modv. Femi [redacted] Rottem [redacted]
Alana [redacted] Lara [redacted] Jerry [redacted] Qiaojun [redacted] Jennifer [redacted] Zhanna
[redacted]

24

DOCUMENT SLIPSHEET

Bates Number: META3047MDL-072-00717468

BEGATTACH: META3047MDL-072-00717468

ENDATTACH: META3047MDL-072-00717468

Attachment Count: 0

PRODVOL: META3047MDL_VOL072

Custodian: [redacted]_UMER; LEE_ALISON; ROSEN_GUY

File Path: /LEE_ALISON; [redacted]_UMER; /ROSEN_GUY

Confidentiality Designation: HIGHLY CONFIDENTIAL (COMPETITOR)

HASHVALUE: D5C65E27951026AC7685BDED462E2C54

Author: [redacted]@META.COM

Family Date: 1/29/2022 12:00 AM

Create Date: 6/23/2021 12:00 AM

Last Modified Date: 1/29/2022 12:00 AM

LAST MODIFIED BY:

TRACK CHANGES: N

COMMENTS: N

HASHIDDENDATA: N

Filename: REELS INTEGRITY SIZING (PART 2)_1YJVK8_HQWB7E4WOZKK_89KYT9S3DRQ7QOXON7F9NRGK.PPTX

Title: REELS INTEGRITY SIZING (PART 2)

DOCEXT: PPTX

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: N