Sculpting the Future of Social Media through Incentives and Regulation

Jeff Allen, 2022-06-23
RAN Policy Meeting

What is the Integrity Institute?

- Jeff Allen, data scientist by trade (Facebook, Instagram), co-founder of II
- We are growing a community of tech workers with experience working at social media companies on problems that lie at the intersection of technology, policy, and society. We use our community as infrastructure to support the public, policy makers, academics, journalists, and social media companies themselves as they try to understand best practices and solutions to the problems posed by social media.
- We believe in a social internet that helps societies, democracies, and individuals thrive
- We build towards this vision through three pillars:
    - Building a community of integrity professionals
    - Disseminating and enriching the shared knowledge inside that community
    - Building the tools and research of an open-source integrity team.
- We are not comms professionals. Reach out if you have questions.

Outline

- Understanding spread of content online
- How algorithms can amplify harmful and illegal content
- Current and upcoming platform landscape
- Current policy landscape

Understanding the spread of harmful content

Lifecycle of Harmful Content

- This is a piece of harmful content
- It contains misinfo, hate speech, illegal content, content from extremist group

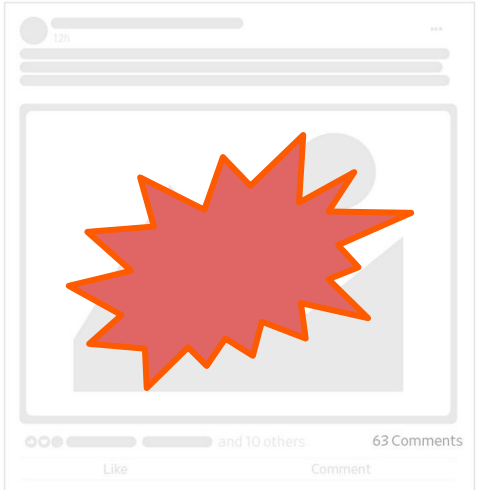
Lifecycle of Harmful Content

- It was uploaded to the platform by
  - A user, an account, a channel, a publisher/business
- The user/account/business may have a history of harmful content

User/Account

Publisher

History?



Lifecycle of Harmful Content

- They distributed it
  - Publicly, privately to followers, in a private group/channel, via an ad, in a direct message to another user(s)

Lifecycle of Harmful Content

- This is a harmful exposure
- A user saw the harmful content

Lifecycle of Harmful Content

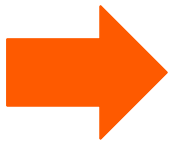
- The exposure happened on a “surface” (Feed, or part of app that shows content)
- The user may follow the creator, or followed the creator due to platform recommendation
- Or have history of exposure
- Or be in vulnerable group

Algo Feed?

Followed?

History?

Demographic?



Lifecycle of Harmful Content

- If harmful content is detected by platform (user reports, algorithmic flag) it can be moderated
- Moderation could be removal, labeling or screening, downranking, require user remove it, etc.

Algo Feed?

Followed?

History?

Demographic?

Time to Moderation

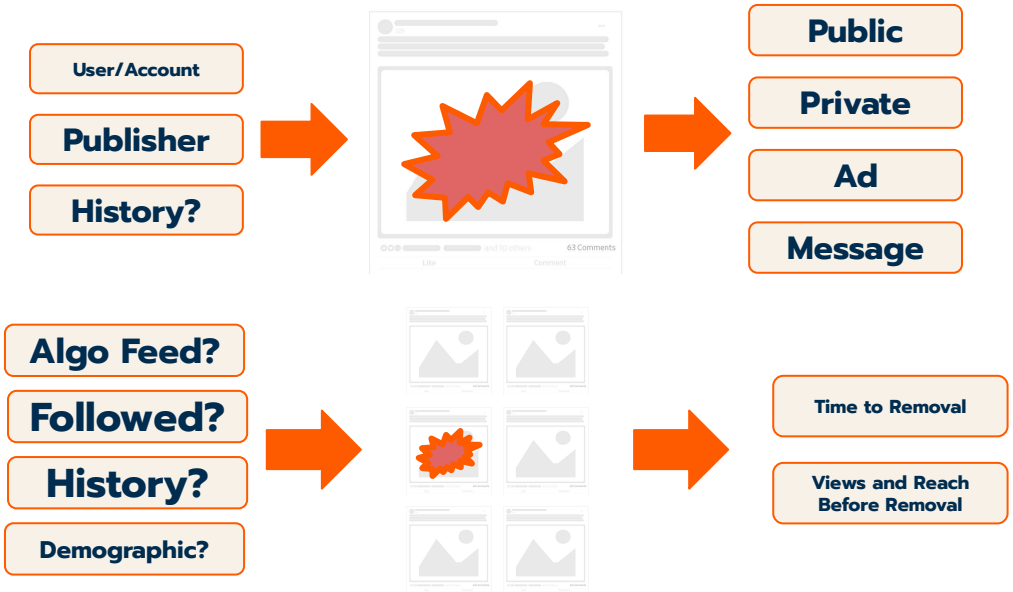
Views and Reach Before Moderation

*Reach: The number of users who view the content

Lifecycle of Harmful Content

- Changes and decisions for this happens under company goals and processes

Top Line Company Processes and Goals



Lifecycle of Harmful Content

- Changes and decisions for this happens under company goals and processes

Top Line Company Processes and Goals

Key Takeaway: It is the consensus view of Integrity Professionals that platforms make this *entire* lifecycle fully transparent and provide all metrics to quantify it.

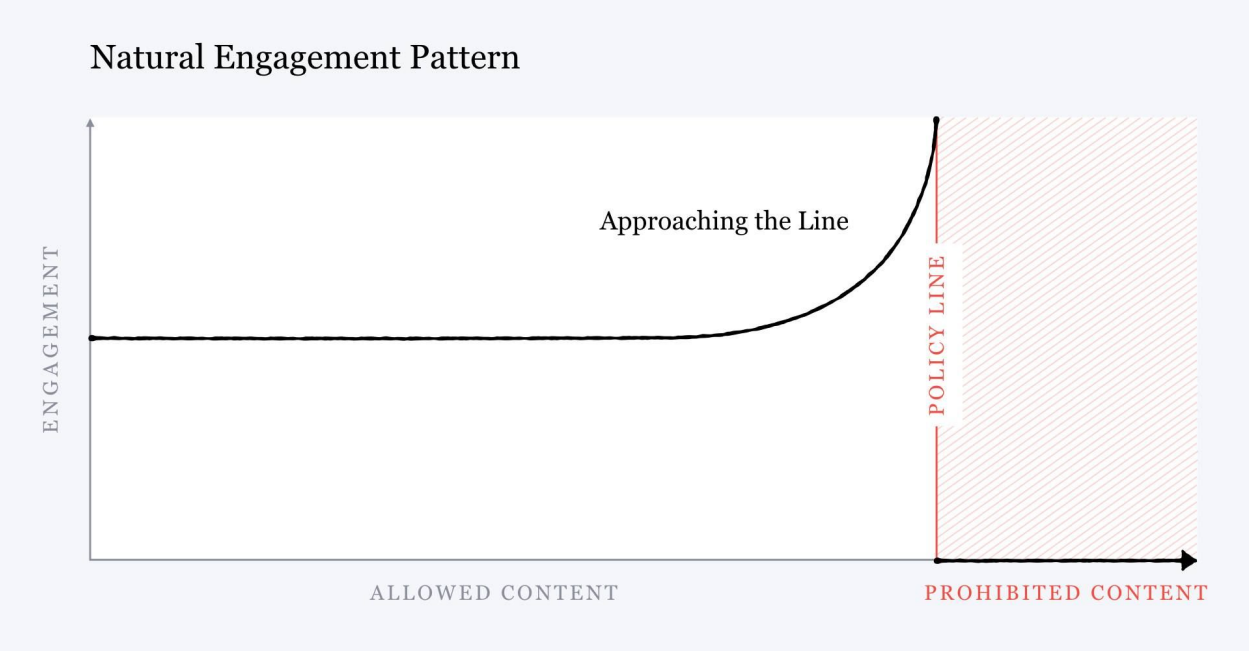
History?

Demographic?

Views and Reach Before Removal

How algorithms can amplify harmful and illegal content

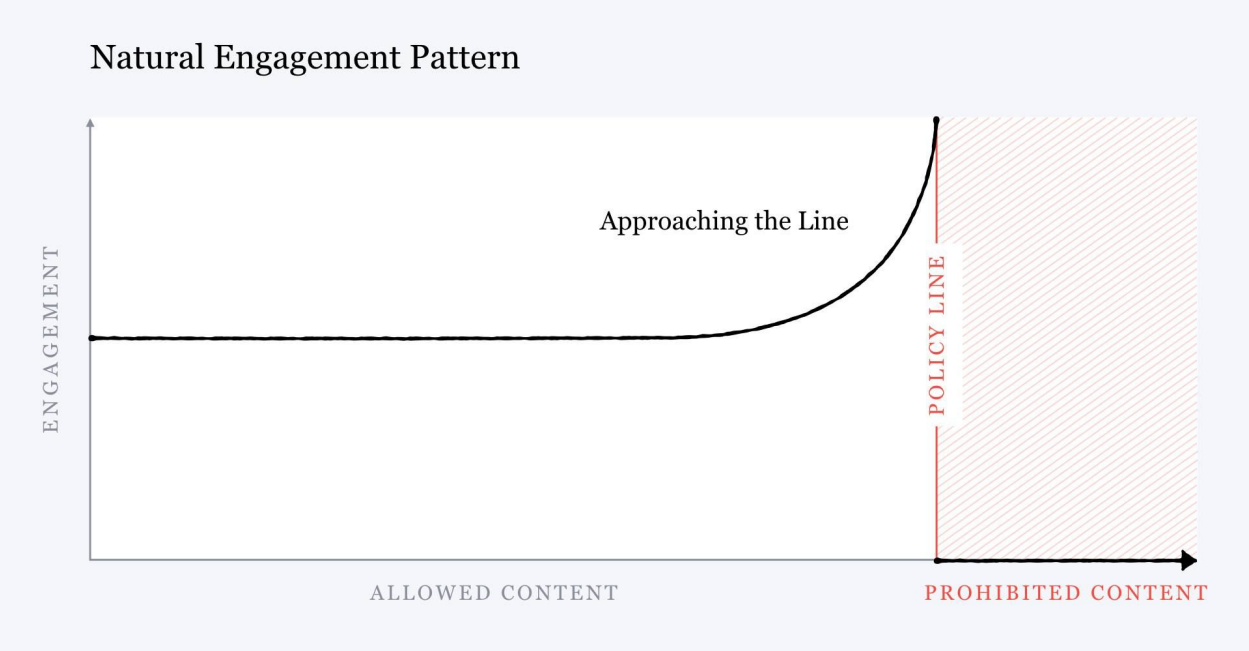
The Engagement Problem



- Y-Axis: What is engagement?
  - Watching a video, clicking “like”, re-sharing, commenting

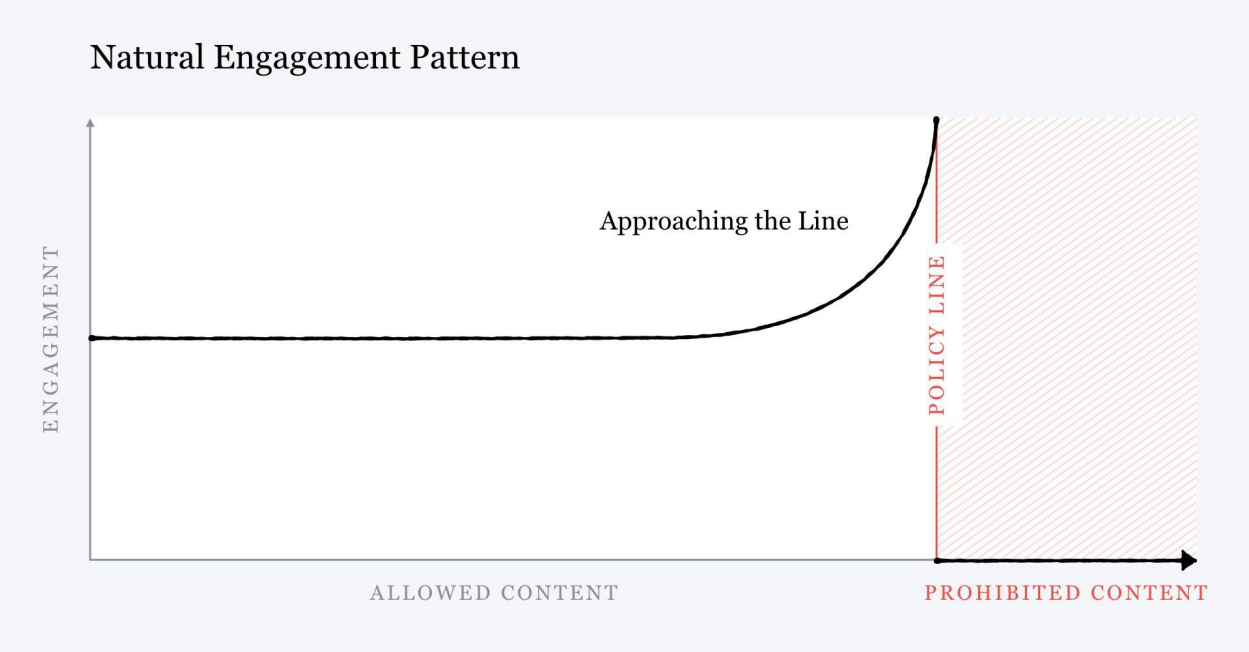
Source: Mark Zuckerberg, 2018, https://www.facebook.com/notes/751449002072082/

The Engagement Problem



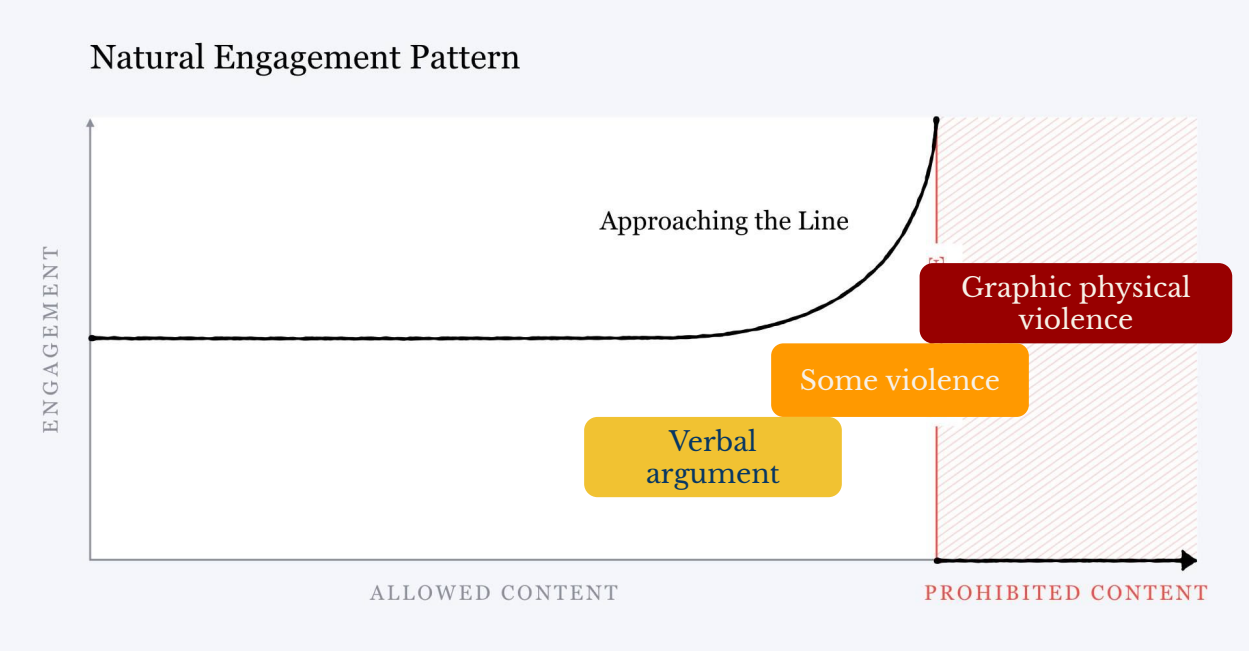
- X-Axis: What is allowed vs. prohibited?
  - Allowed content covers benign to borderline harmful
  - Prohibited content is harmful

The Engagement Problem



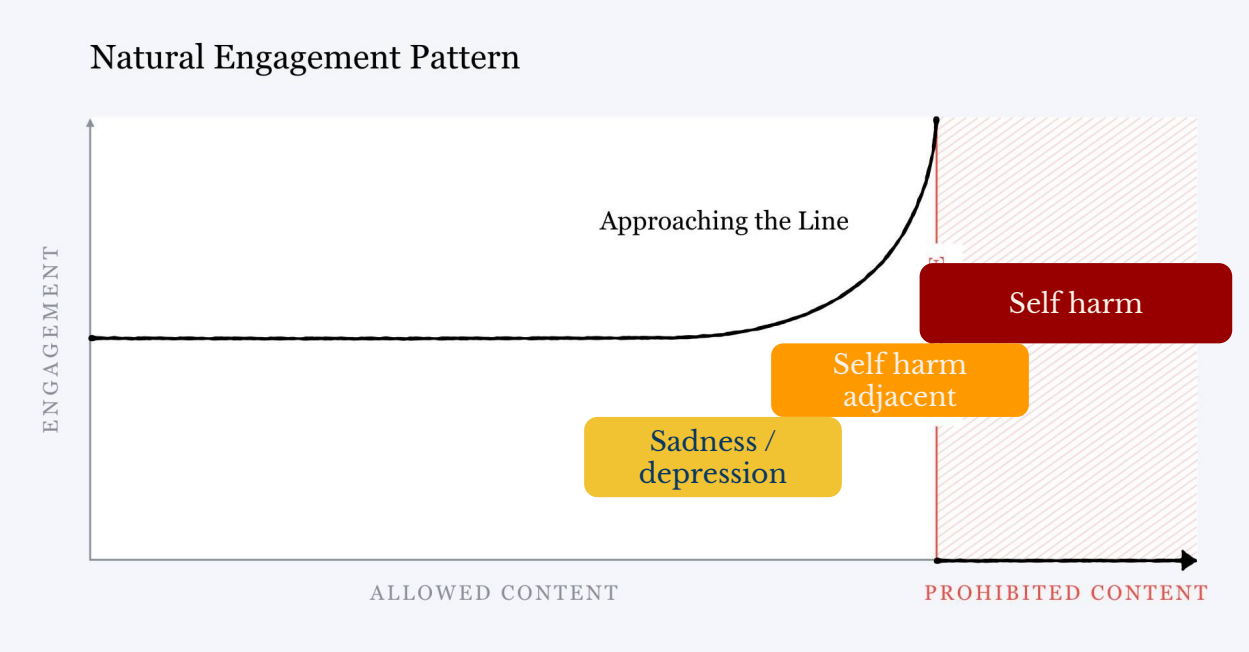
- This is true across many types of potential harms

The Engagement Problem



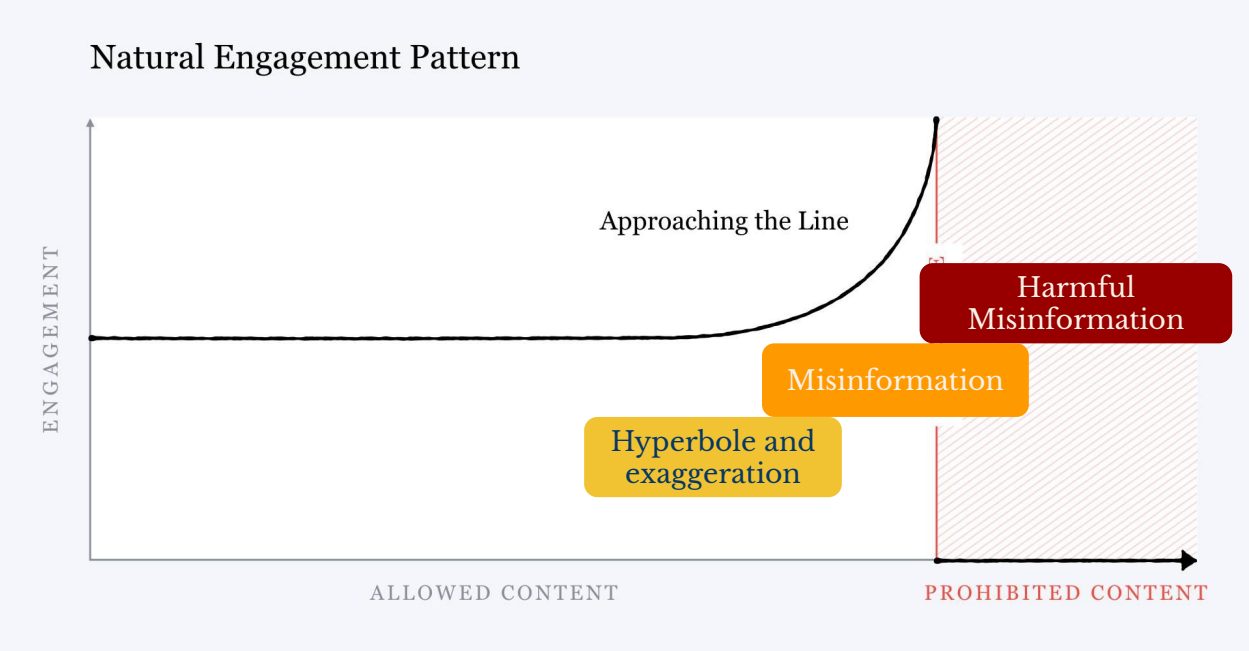
- This is true across many types of potential harms

The Engagement Problem



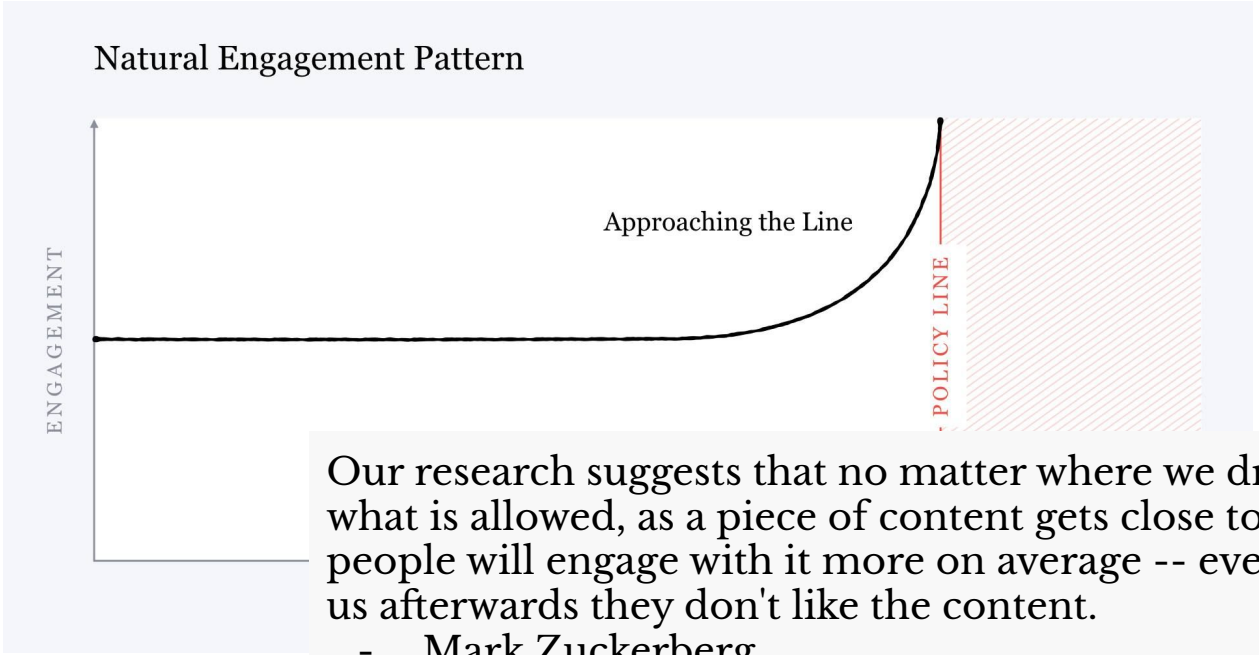
- This is true across many types of potential harms

The Engagement Problem



- This is true across many types of potential harms

The Engagement Problem



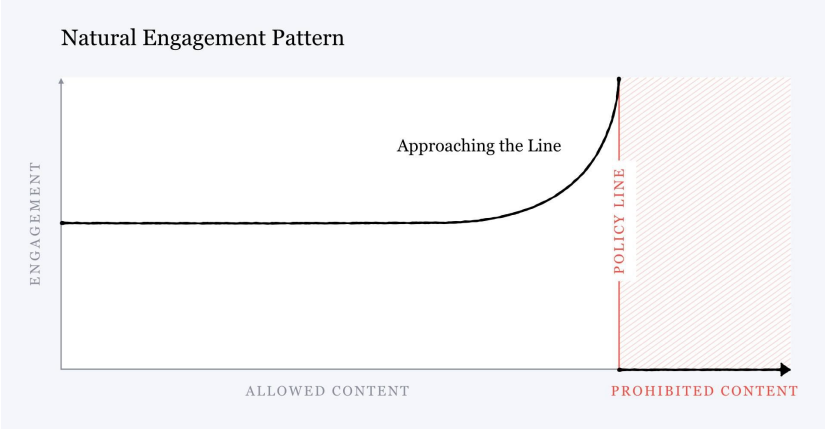
Our research suggests that no matter where we draw the lines for what is allowed, as a piece of content gets close to that line, people will engage with it more on average -- even when they tell us afterwards they don't like the content.

- Mark Zuckerberg

Source: Mark Zuckerberg, 2018, https://www.facebook.com/notes/751449002072082/

The Engagement Problem

- And this shouldn't be surprising
  - "If it bleeds it leads" nightly news
  - Tabloids near checkout in grocery stores
  - People "rubbernecking" at accidents
- But, social media brings new aspects
  - "Connected world" means connected to bad actors
  - Many more "content subjects"
  - Little/No human editorial oversight

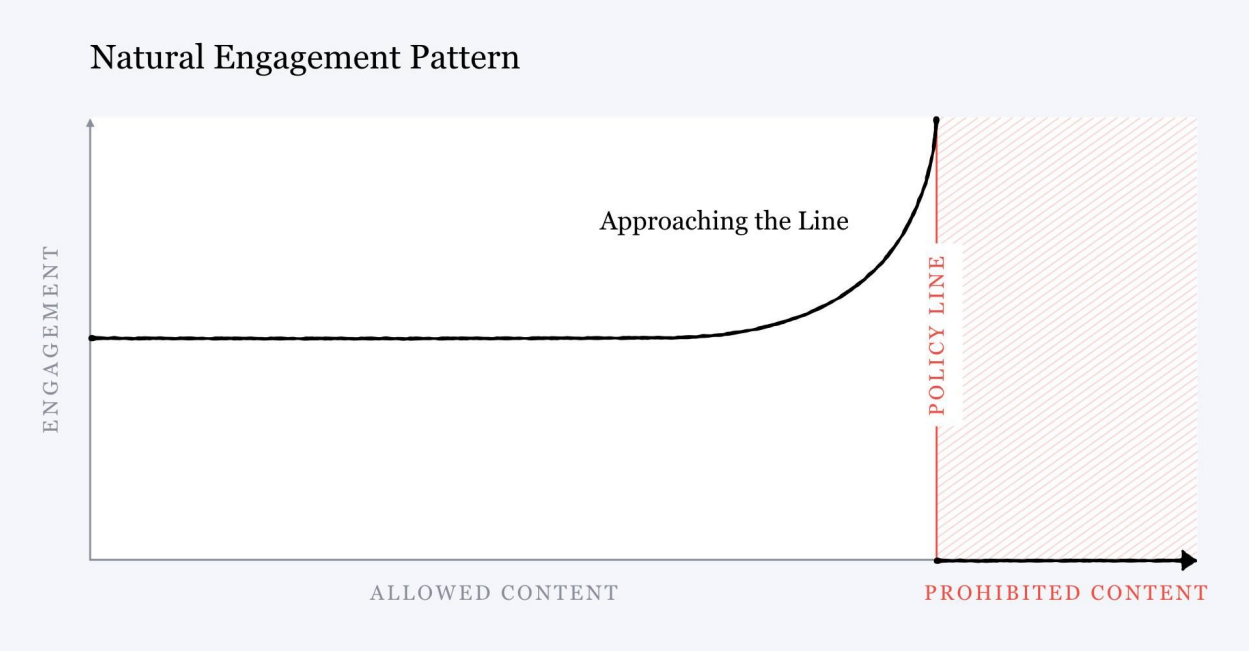


How Most Platforms Work

Facebook	**Predicted Engagement**: Like, Reaction, Comment, Share
Twitter	**Predicted Engagement**
TikTok	**Predicted Engagement**: Like, Comment, Watch
YouTube	**Predicted Engagement**: Watch Time, Surveys

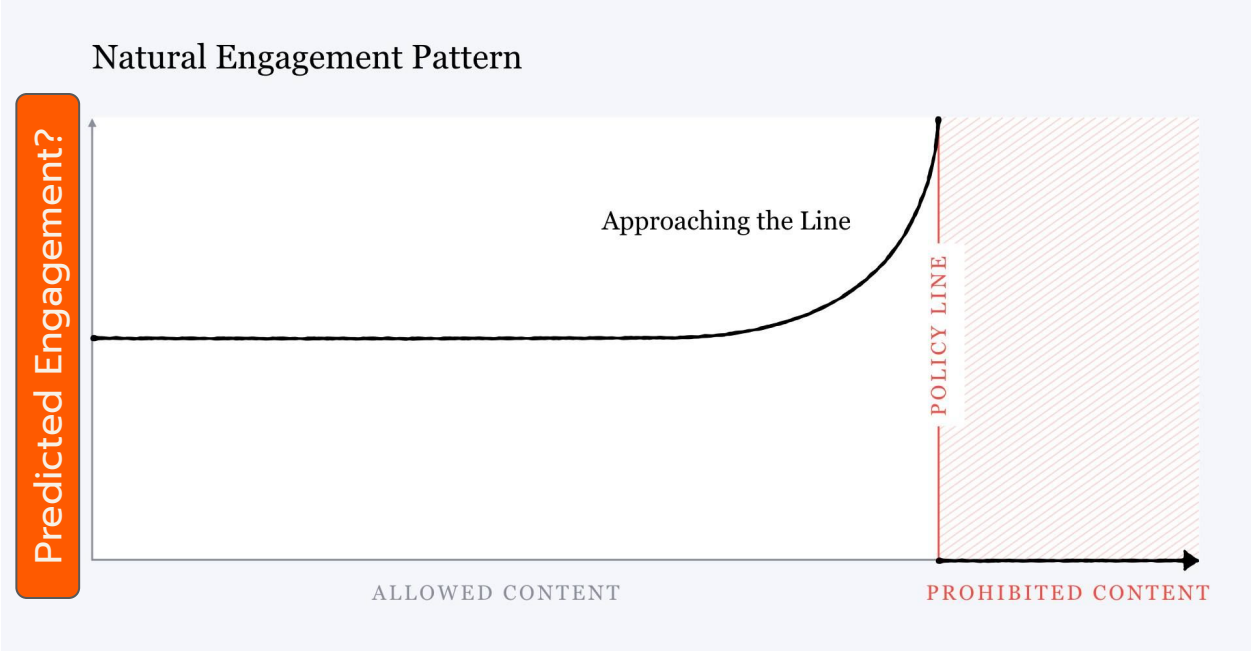
Source: NYTimes, 2021, https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html
Source: Wall St. Journal, 2021, https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215
Source: Twitter, 2017, https://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-at-scale-in-twitters-timelines
Source: YouTube, 2019, https://www.blog.google/around-the-globe/google-europe/fighting-disinformation-across-our-products/

How Platform Design Can Amplify Harms



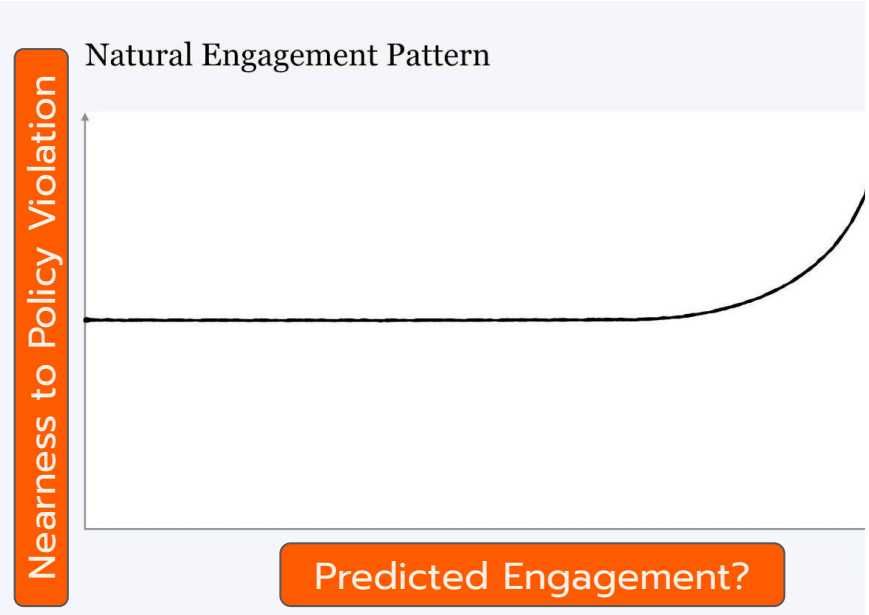
- More engagement, more likely to be harmful

How Platform Design Can Amplify Harms



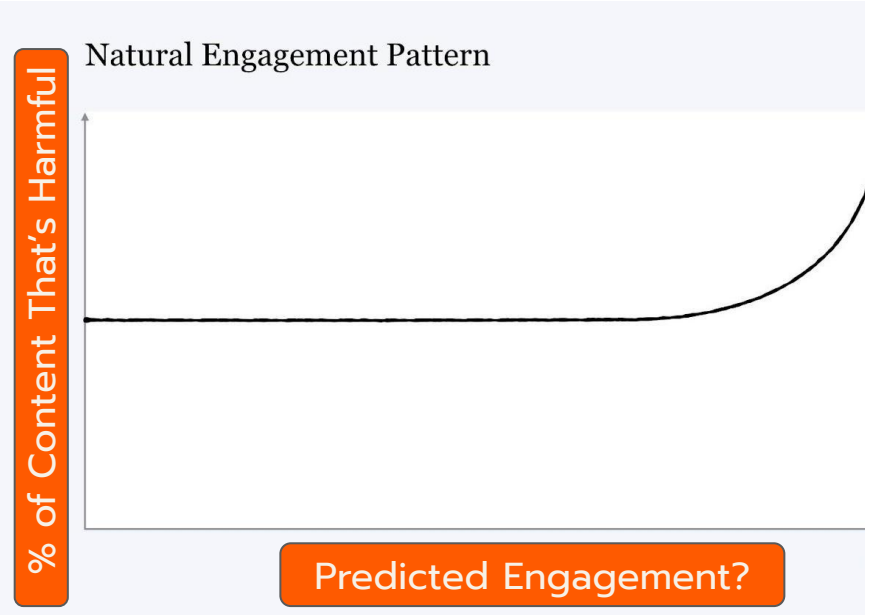
- Predicted engagement should follow actual engagement
- Content predicted to be engaging is more likely to be harmful

How Platform Design Can Amplify Harms



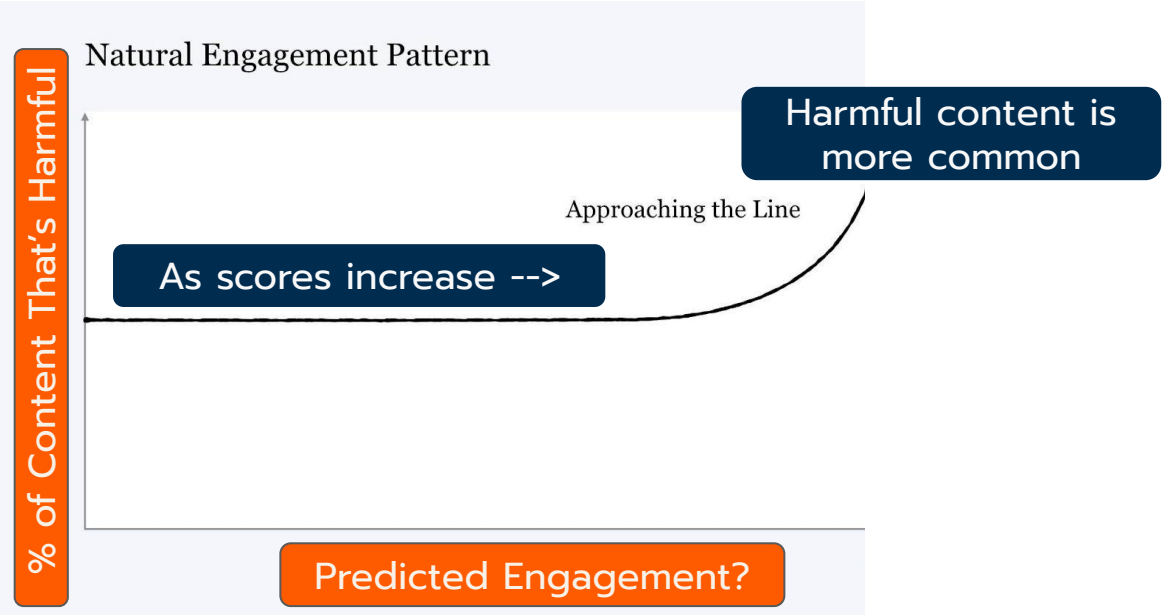
- Let's make it measurable
- Swap the X and Y Axes

How Platform Design Can Amplify Harms



- "Nearness to policy" is not measurable
- % of content which is harmful is

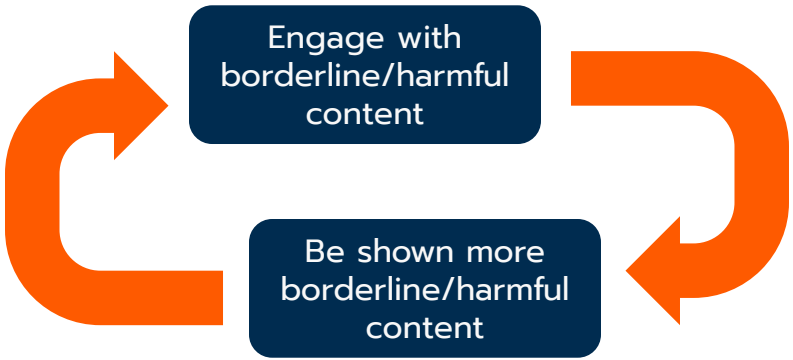
How Platform Design Can Amplify Harms

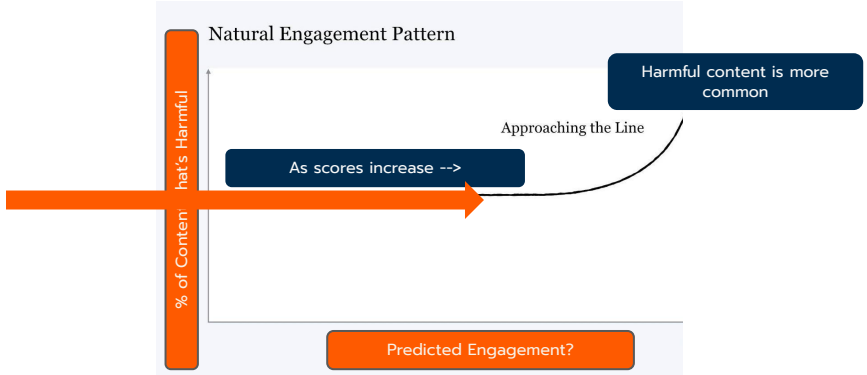


- Harmful content will tend to “float to the top” of the ranking systems
- This chart is measurable! Every platform could report it publicly

How Does This Problem Manifest?

- Platforms track everything users engage with
- They use that to predict what users will engage with in future
- The systems are biased to show more extreme version of historical engagement
- This is the “Rabbit Hole”





Source: New York Times, 2020, https://www.nytimes.com/column/rabbit-hole

How Does This Problem Manifest?

- Platforms track everything users engage with
- They u
- The ment
- This

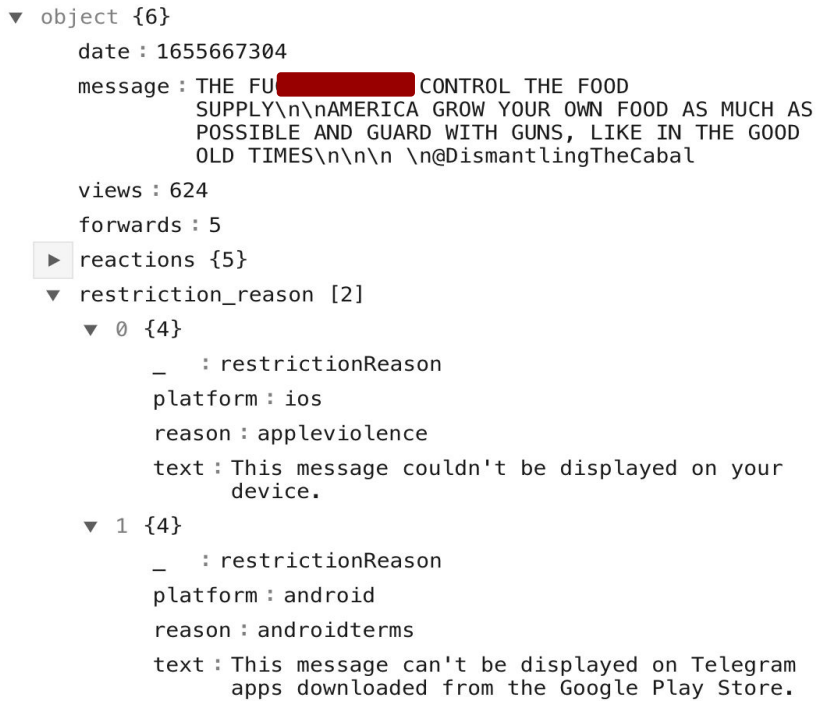
Key Takeaway: It is the consensus view of Integrity Professionals that platforms make it transparent how their ranking and recommendation systems work, with enough detail to audit if they are exploiting the engagement problem.

Source: New York Times, 2020, https://www.nytimes.com/column/rabbit-hole

Current and upcoming platform landscape

Current Trends: Hostility to Integrity

- We cannot expect platforms to be amenable to cooperation with external partners
- For all the bad press Facebook gets, they have invested in integrity and cooperated with external partners
- This will not always be the case
- Example: Telegram
  - Telegram knows dangerous groups are using it
  - Allow them even after detection



Current Trends: Ephemerality

- Idea of ephemerality is old
  - Snapchat Story
  - But continues to expand
- Clubhouse: Ephemeral chat rooms
- Twitter Spaces: Ephemeral chat rooms
- Facebook has explored
- Open question:
  - To what extent should platforms report and make available ephemeral content which has broad distribution?

Current Trends: Private Chat

- Private chat growing
  - Some E2EE, some not
  - Telegram, WhatsApp current biggest
  - Hidden Facebook Groups offer similar functionality
- Private chat means
  - Chat is not “discoverable” within the app
  - Can only be made aware through an invite from a current member
- Open question:
  - How big should private spaces be allowed to get? At what point is it no longer private?
  - WhatsApp: 512 members
  - Signal: 1,000 members
  - Telegram: 200,000 members
  - Facebook Groups: Millions

Potential Upcoming Trend: Virtual Reality

- Meta will be investing in “Metaverse”
- From integrity point of view, a lot we can learn from the past
    - Second Life, Gaming (World of Warcraft, Minecraft, Roblox)
- There are some differences from current, most popular platforms
- But also some similarities

Potential Upcoming Trend: Virtual Reality

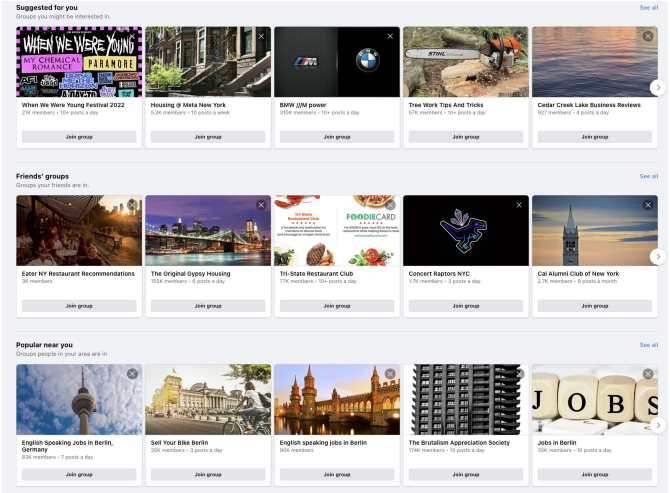
- Significant part of VR experience is a “chat room”
- Users meet in virtual space, and talk and interact
- Chat based apps give reasonable model of what to expect
    - Discord, Telegram, etc

Potential Upcoming Trend: Virtual Reality

- Virtual spaces can be created by users, and recommended by platform systems

Potential Upcoming Trend: Virtual Reality

- Virtual spaces can be created by users, and recommended by platform systems
- Existing account / group / channel recommendations are similar
    - Facebook Group recommendations give a reasonable model for what to expect here



Current regulatory landscape

What Could Be Productive Regulation Wise?

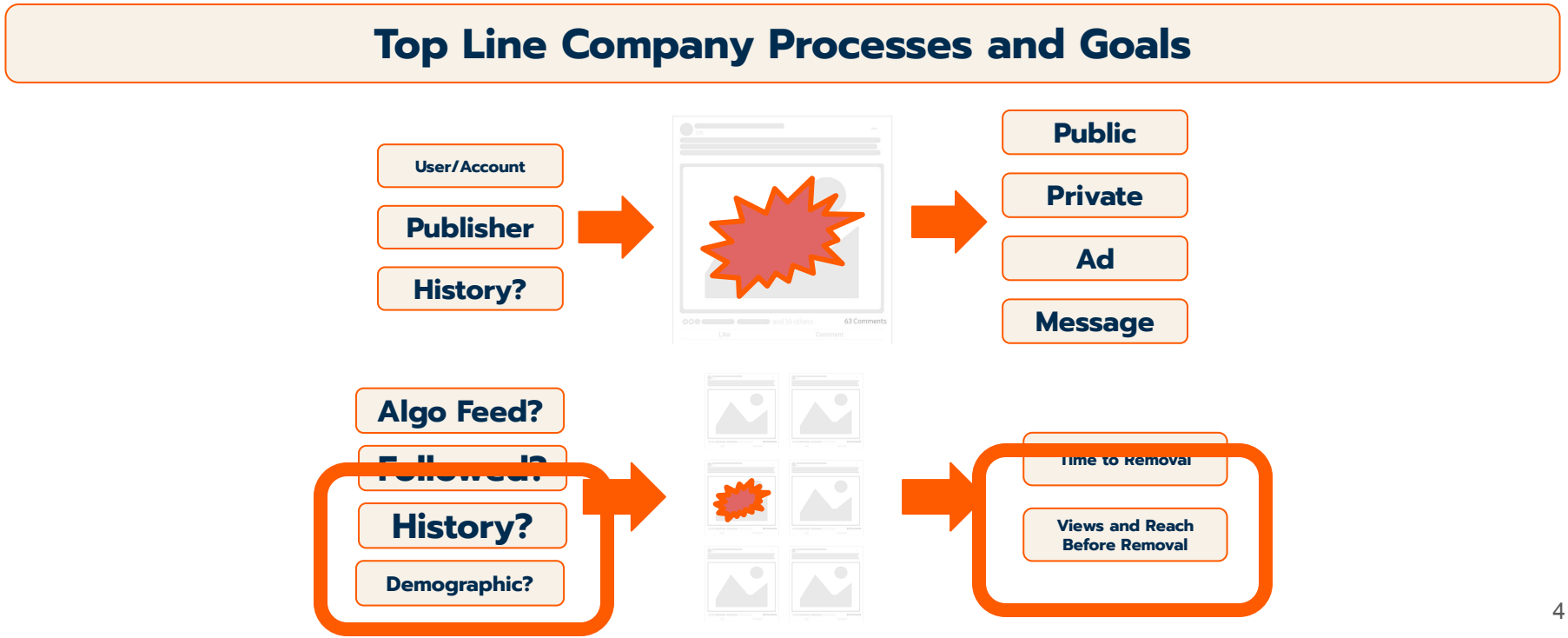
- There is definitely progress in the right direction (DSA)
- Society could use
  - Platforms provide data demonstrating safety
  - Platforms provide data on risk of design
  - Regulation that is flexible and updatable

Transparency Can Change Business Decisions

- Example: Instagram Kids
  - Spring 2021: “Facebook Is Building An Instagram For Kids”
  - September 14, 2021: “Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show”
  - September 27, 2021: “Instagram for kids paused after backlash”
- What were the pieces of data that, when made public, changed Instagrams decision? What would have changed their decision to announce it in Spring of 2021?

Transparency Can Change Business Decisions

- Lifecycle of Harmful Content covers it



Transparency Can Change Business Decisions

- Key Data Points
  - Self harm and eating disorder content is concentrated
  - Self harm and eating disorder content primarily impact young people
- Concentration
  - Prevalence of self harm content on Instagram is ~<0.05%
  - How do you get that prevalence?
    - A: Every user sees self harm in 5 out of every 10,000 posts
    - B: 1% of users see self harm in 1 out of every 20 posts, everyone else ~never
  - Self harm and eating disorder content, and most harms, follow B
- Young People
  - What is the prevalence by demographic breakdown?
  - Prevalence of self harm and eating disorder content is higher for young people

Transparency Can Change Business Decisions

- Example: Instagram Kids
  - The DSA is not quite there yet in mandating this kind of transparency
  - But DSA is also adaptable
    - Audits can grow, CoP can grow, risk assessments can grow
- Adaptable legislation is crucial
  - TikTok will not be the last new platform to reach 1B users
  - We have not seen all forms of social media
  - We can't keep waiting ~10 years between updates

Wrap up

Conclusion

- If you want to know what data integrity workers use on the inside to study harms on platforms, ask us!
  - Lifecycle of Harmful Content
  - Ranking and Design Transparency
  - Follow the topline company goal metrics
- There are known irresponsible design patterns in social media
  - There are also known more responsible design patterns
- Social media continues to evolve and develop
  - New companies, new features, new communication methods
- But there is genuine signs regulation could impact how companies make decisions