Feed Ranking and Social Harms: A Trustworthy AI Problem

Naomi Shiffman
Fellow, Integrity Institute
Head of Data & Implementation, Oversight Board

What is the Integrity Institute?

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 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.

What I’ll cover today

- The standard design of ranked feeds
- How AI/ML/Algorithms can amplify harmful content
- Alternatives to engagement-based ranking
- How do we know if a ranked feed is contributing to societal harm?

The standard design of ranked feeds

Ranking Basics

Ranking systems all have similar components. The purpose of these components are to:

- Gather content
- Score content
- Produce final ranked list

Ranking Basics

Inventory

Ranking Basics

Inventory

- All applicable content is gathered (Posts, Tweets, Videos)
- Can include content from non-followed accounts (Reshares, Retweets, Friend Likes, Public videos on YouTube etc)

Ranking Basics

Inventory

Features

Ranking Basics

Inventory

Features

X, Y, Z

“Features” are discrete data about content and/or user

- Has the user liked, retweeted, content from the creator before?
- Do users “like the user” like, retweet, favorite the content?
- Has the user liked, retweeted, favorited content “like this content”?
- Does the content have external validation from other sources on the internet?

Ranking Basics

Inventory

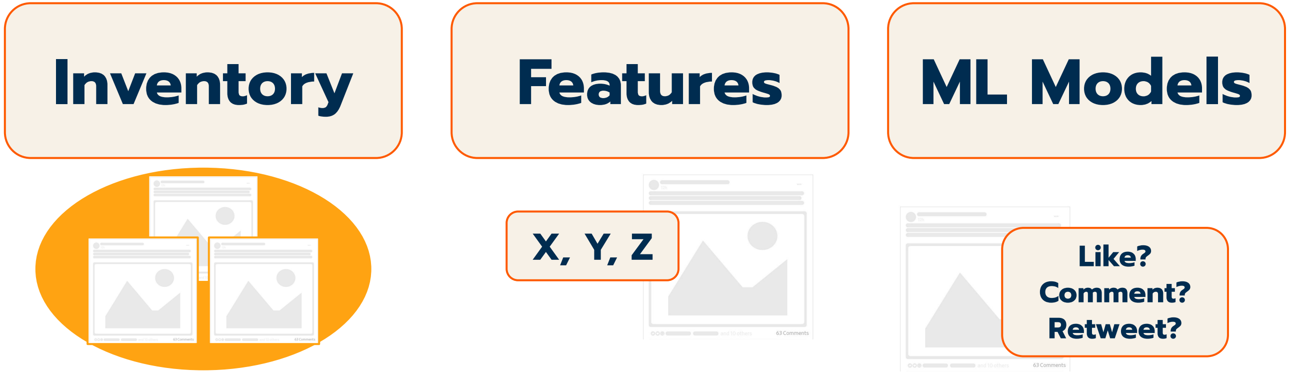
Features

ML Models

X, Y, Z

Like?
Comment?
Retweet?

Ranking Basics



Inventory

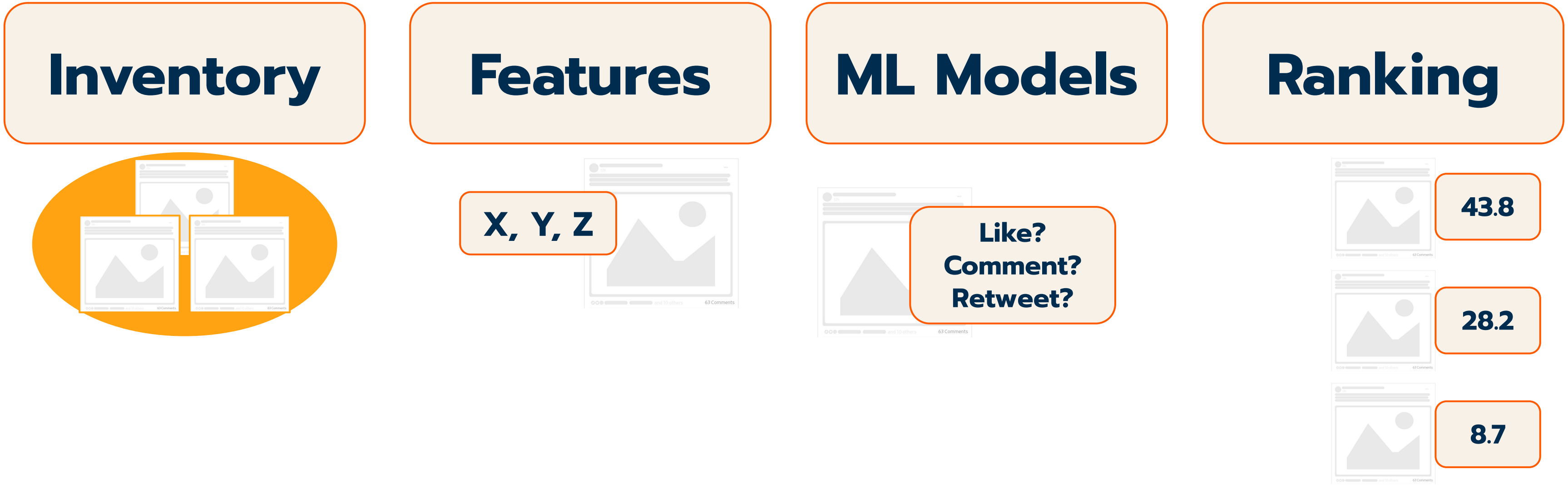
Features

ML Models

Machine learning models predict various outcomes

- “Will the user favorite this image?”
- “Will the user reshare this post?”
- “Is this content harmful?”
- “Is this content high quality?”

Ranking Basics



Inventory

Features

X, Y, Z

ML Models

Like?
Comment?
Retweet?

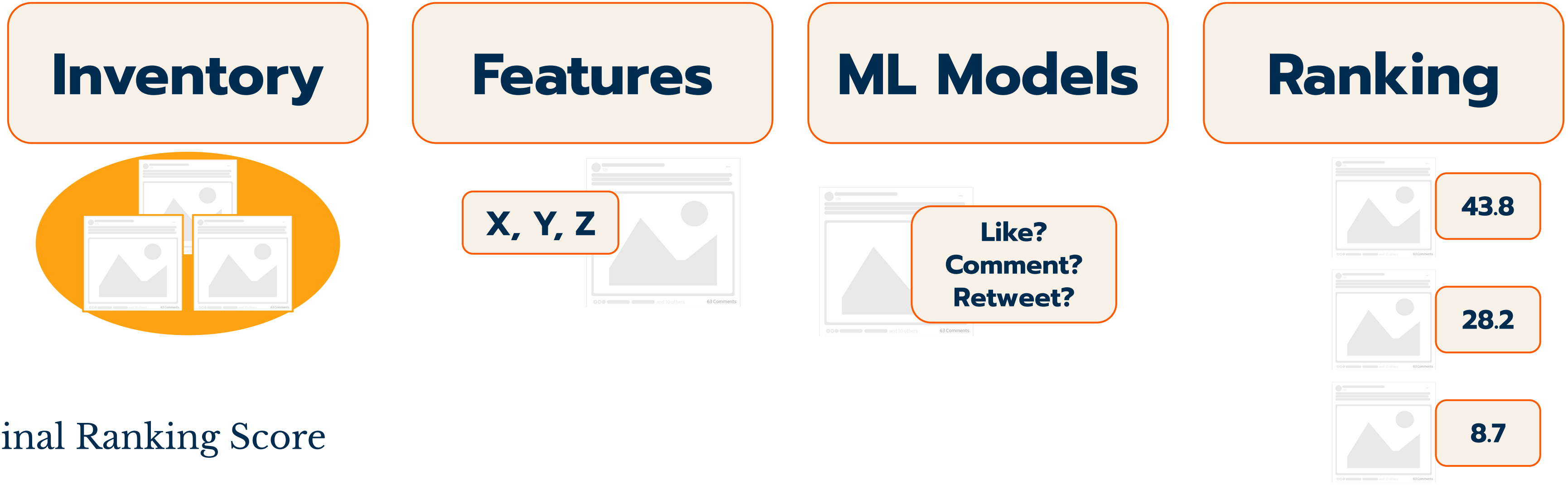
Ranking

43.8

28.2

8.7

Ranking Basics



Inventory

Features

X, Y, Z

ML Models

Like?
Comment?
Retweet?

Ranking

43.8

28.2

8.7

Final Ranking Score

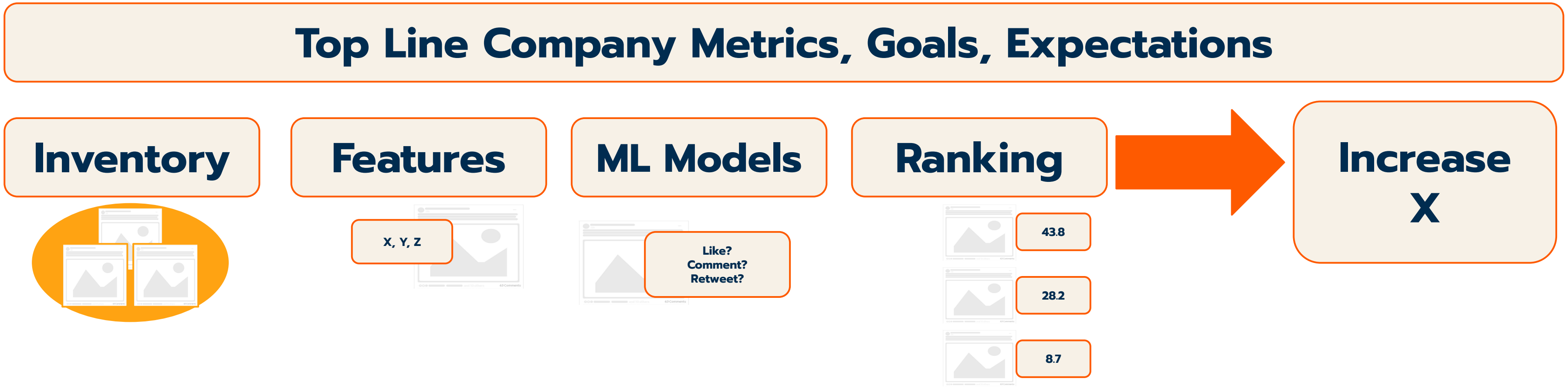
- All the classifier scores are combined, business logic applied
- Final sorting and list generated

Standard Design

- The ranking system is internal
- The company has objectives for the ranking system - “top-line” metrics
- The company and team have goals and metrics

Standard Design

The process is mediated by the companies' goals and experimentation process



Top Line Company Metrics, Goals, Expectations

Inventory

Features

ML Models

Ranking

Increase X

Standard Design: Engagement Ranking

Inventory

- Collect posts, including non-followed posts

Compute features

- Heavily influenced by individual user history

Run ML Models

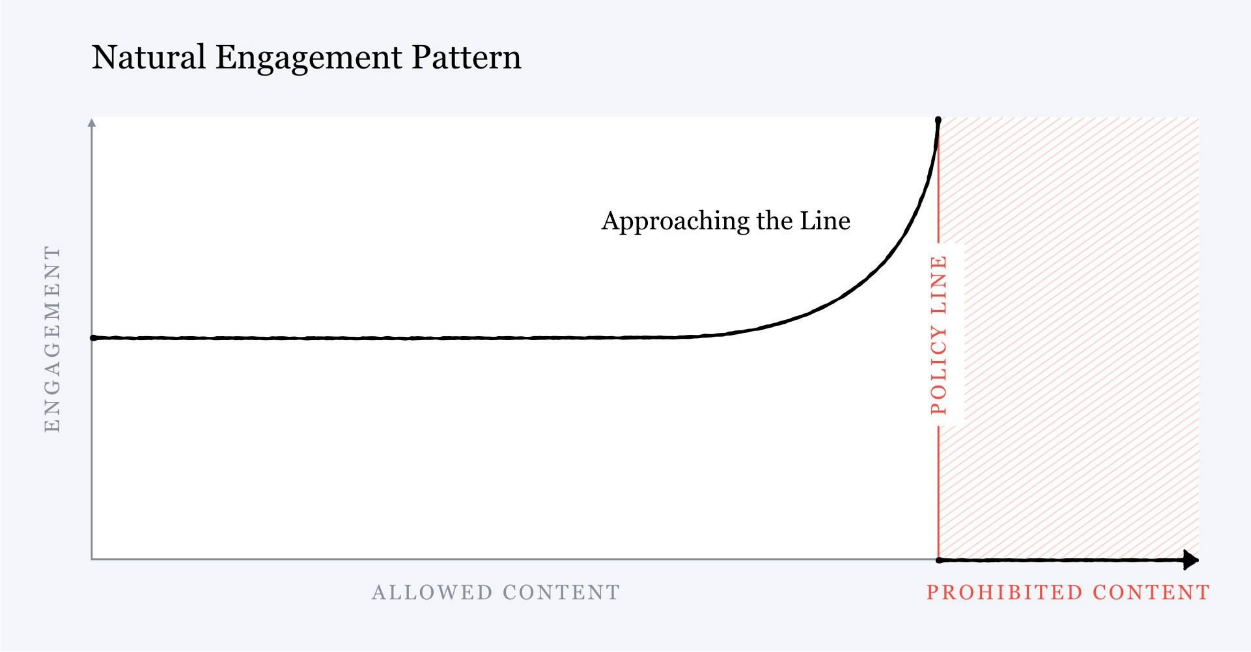
- Many predicted user engagement actions

Output final ranked list

- Scoring high on user engagement classifiers will push content up

How algorithms can amplify harmful content

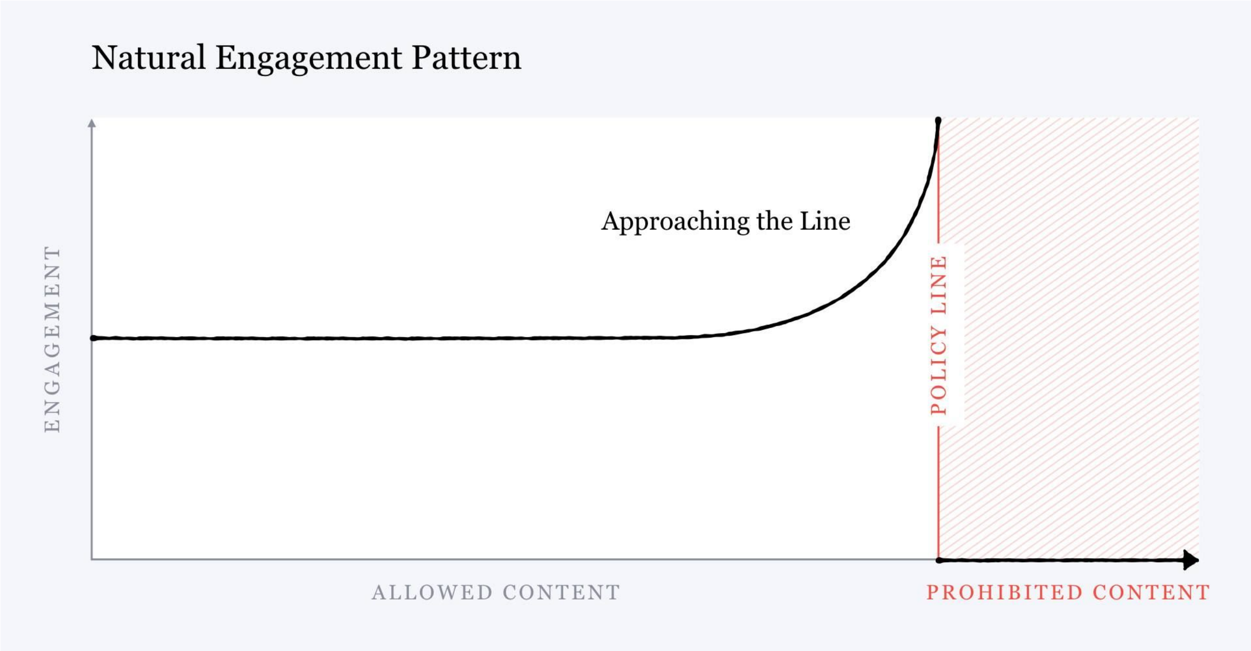
The Engagement Problem



Engagement: watching a video, clicking "like", re-sharing, commenting

Source: Mark Zuckerberg, 2018, facebook.com/notes/751449002072082/ aka: sahar.io/zucknote

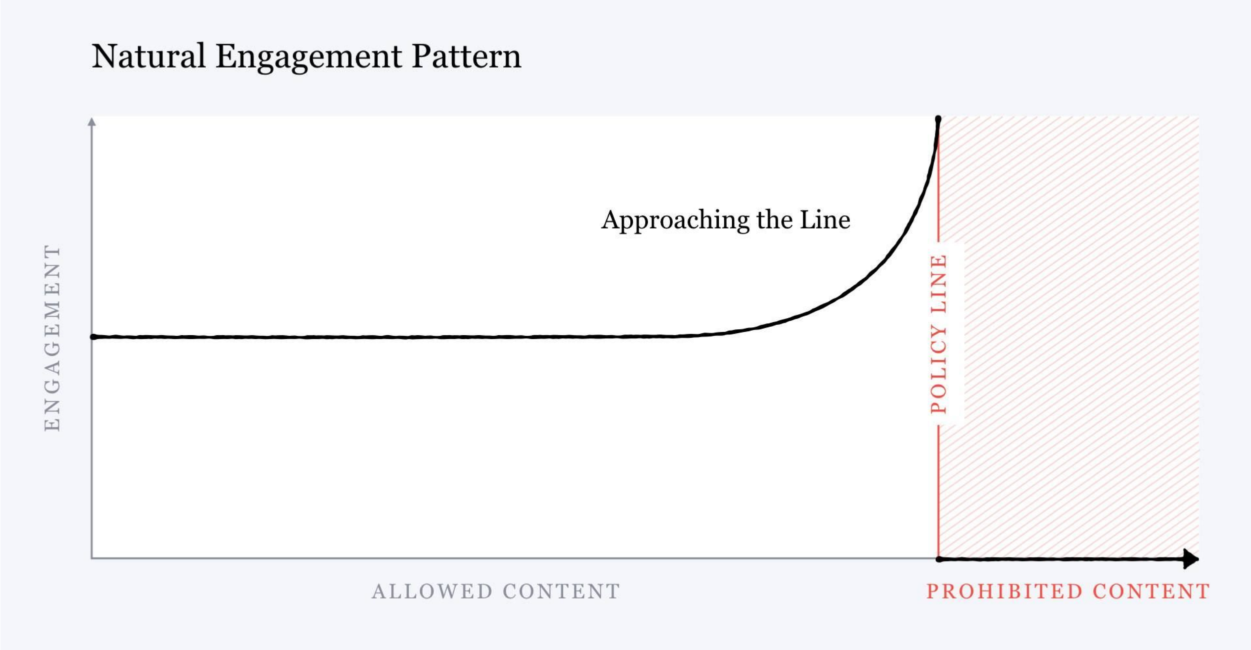
The Engagement Problem



What is allowed vs. prohibited?

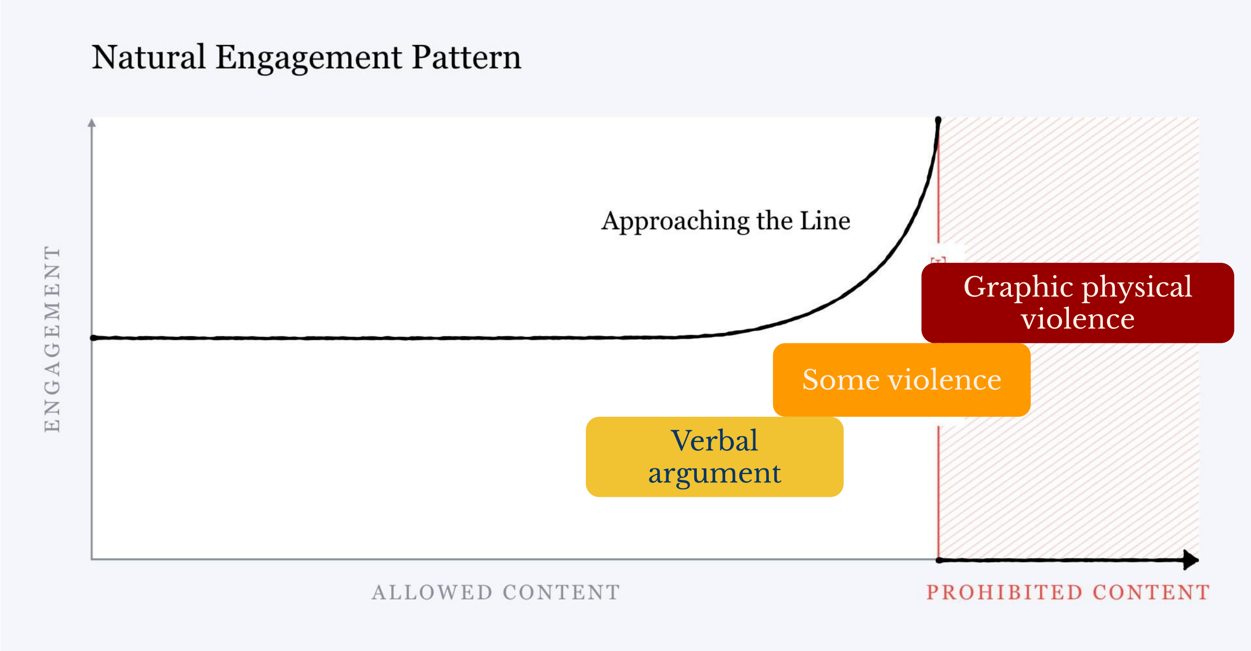
- Allowed content covers benign to borderline harmful
- Prohibited content is harmful

The Engagement Problem



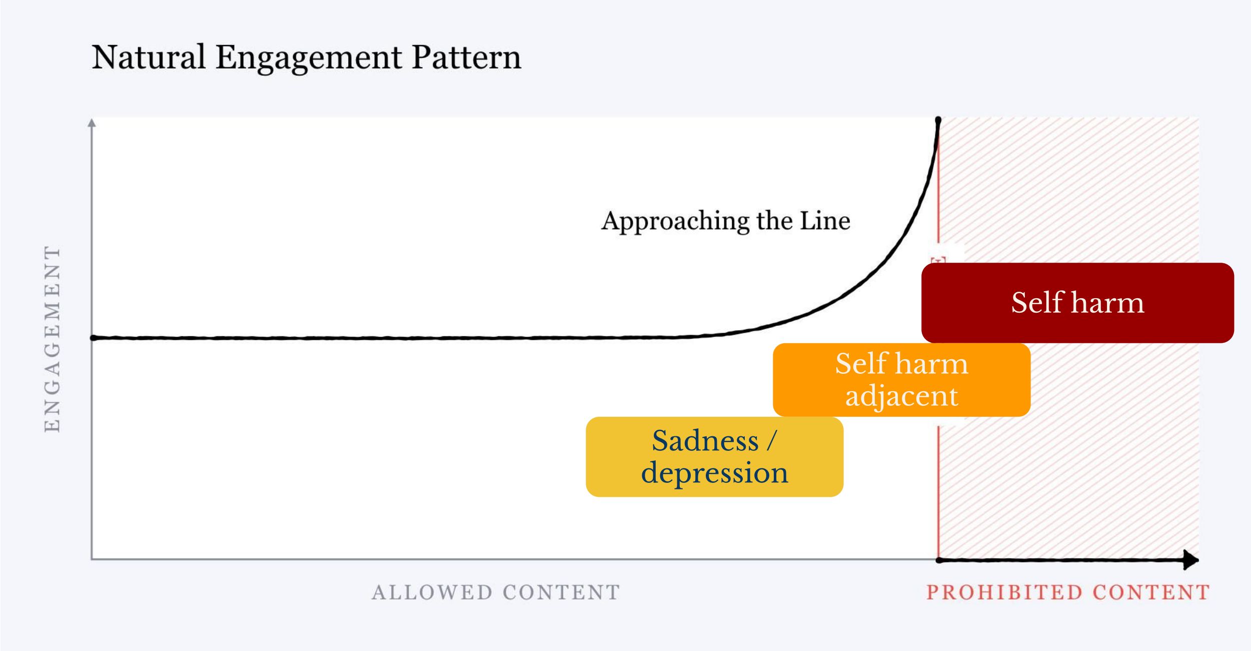
This is true across many types of potential harms

The Engagement Problem



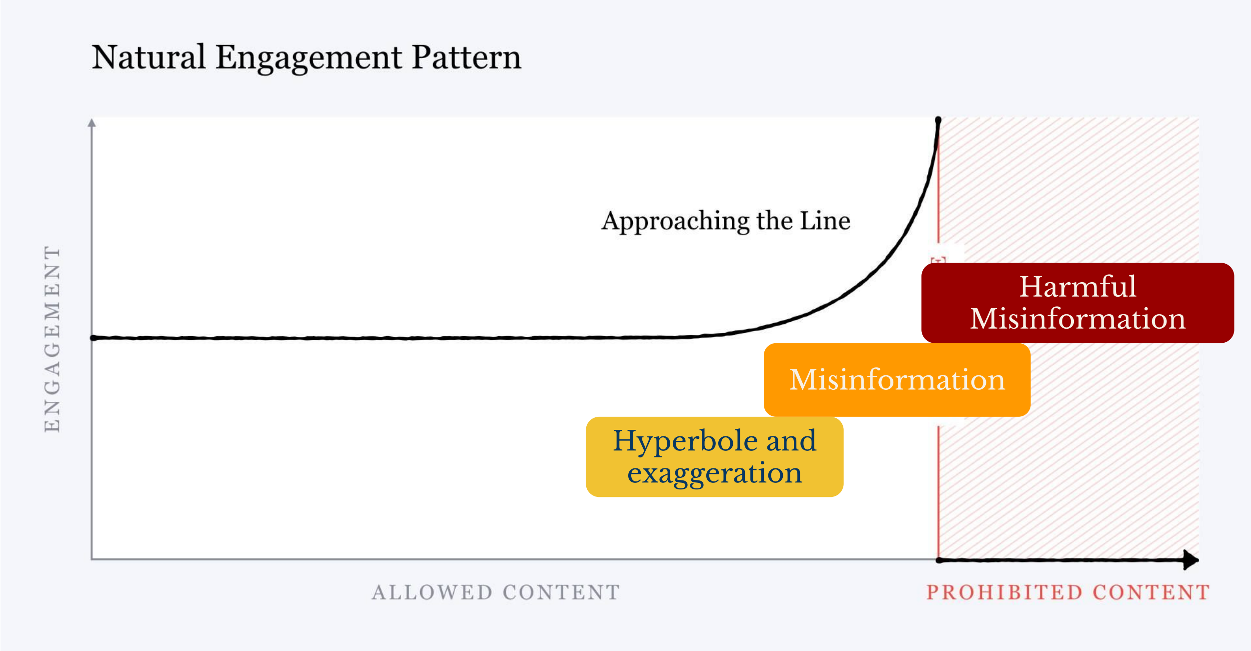
This is true across many types of potential harms

The Engagement Problem



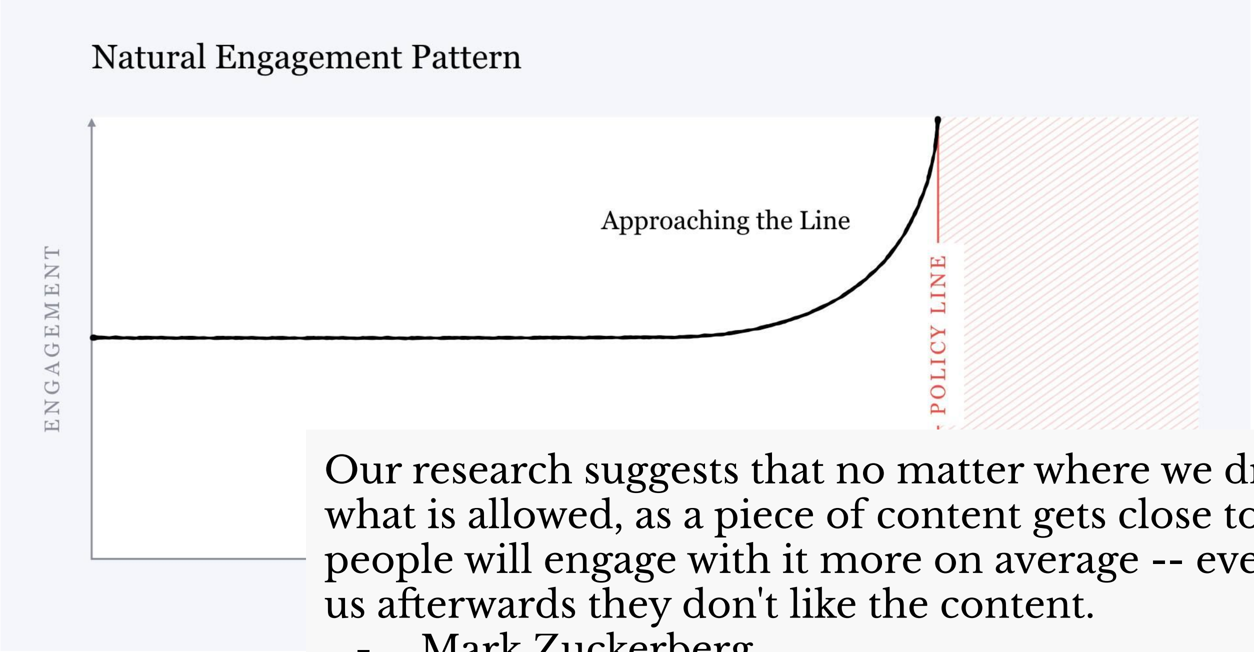
This is true across many types of potential harms

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

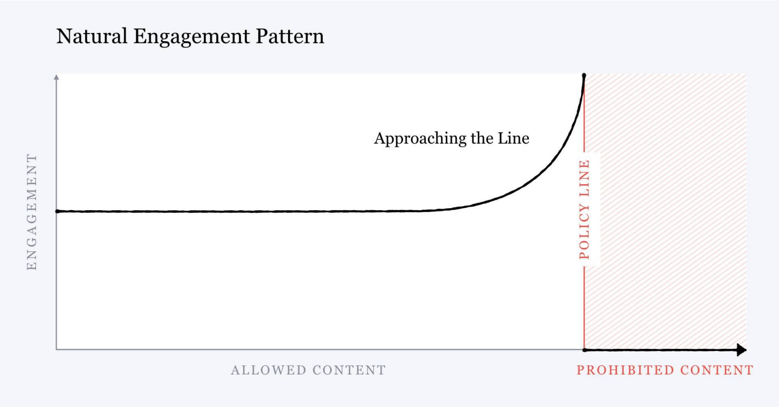
The Engagement Problem

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



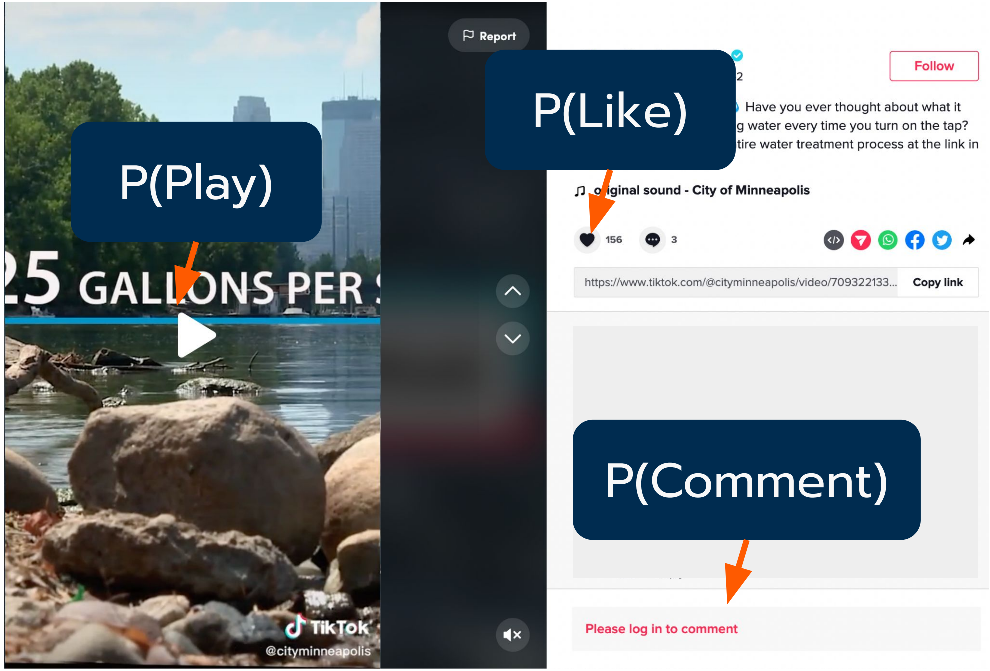
How Most Platforms Work

How do most platforms rank and order recommended content and accounts?

How Most Platforms Work

TikTok

- Probability user will…
  - Like a video
  - Comment on a video
  - Play a video
  - Watch a video for an extended time



Have you ever thought about what it
g water every time you turn on the tap?
tire water treatment process at the link in

♫ original sound - City of Minneapolis

https://www.tiktok.com/@cityminneapolis/video/709322133...

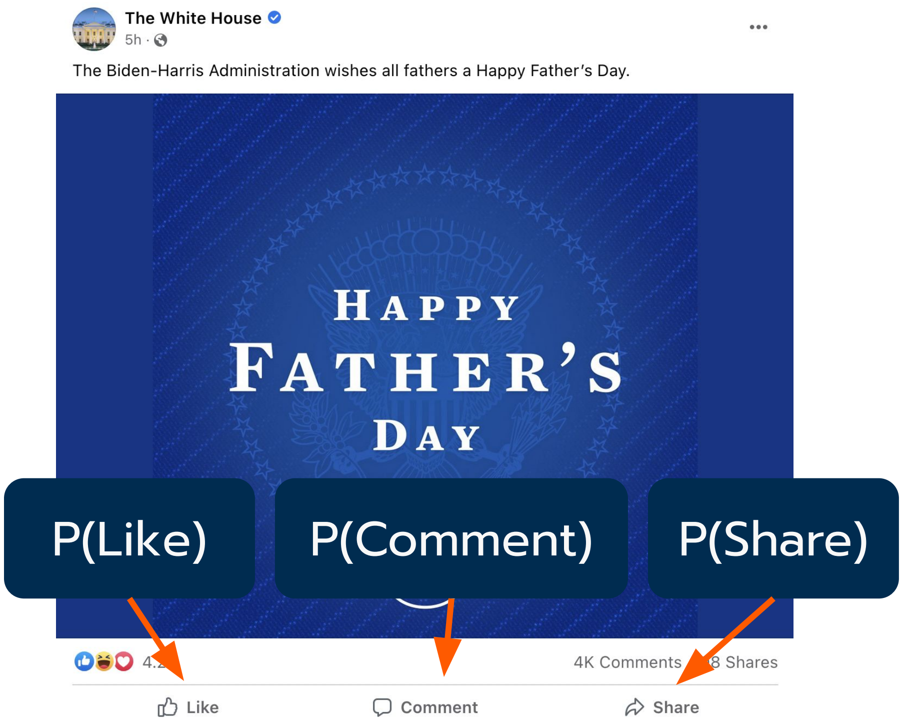
Please log in to comment

Source: NYTimes, 2021, https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html

How Most Platforms Work

Facebook

- Probability user will...
  - Like
  - Reaction
  - Comment
  - Reshare



The White House

5h ·

The Biden-Harris Administration wishes all fathers a Happy Father's Day.

P(Like)

P(Comment)

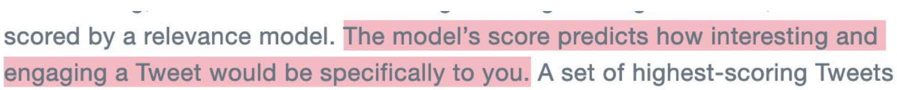
P(Share)

Source: Wall St. Journal, 2021, https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215

How Most Platforms Work

Twitter

- "Interesting and engaging"



scored by a relevance model. The model's score predicts how interesting and engaging a Tweet would be specifically to you. A set of highest-scoring Tweets

YouTube

- Clicks
- Watch Time
- Surveys

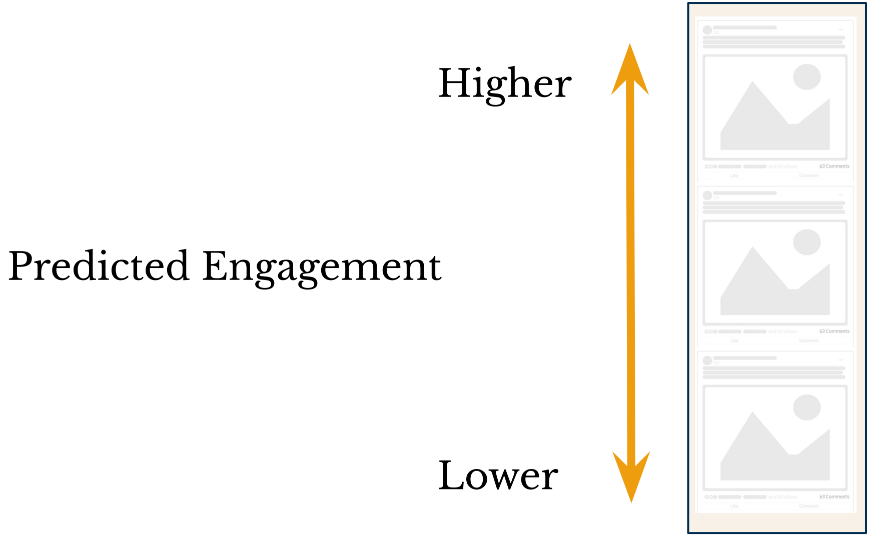
Source: Twitter, 2017, https://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-at-scale-in-twitters-timelines
YouTube, 2019, https://www.blog.google/around-the-globe/google-europe/fighting-disinformation-across-our-products/

How Most Platforms Work

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

How Most Platforms Work

Platforms recommend content and accounts most likely to be engaged with. Why does this matter?

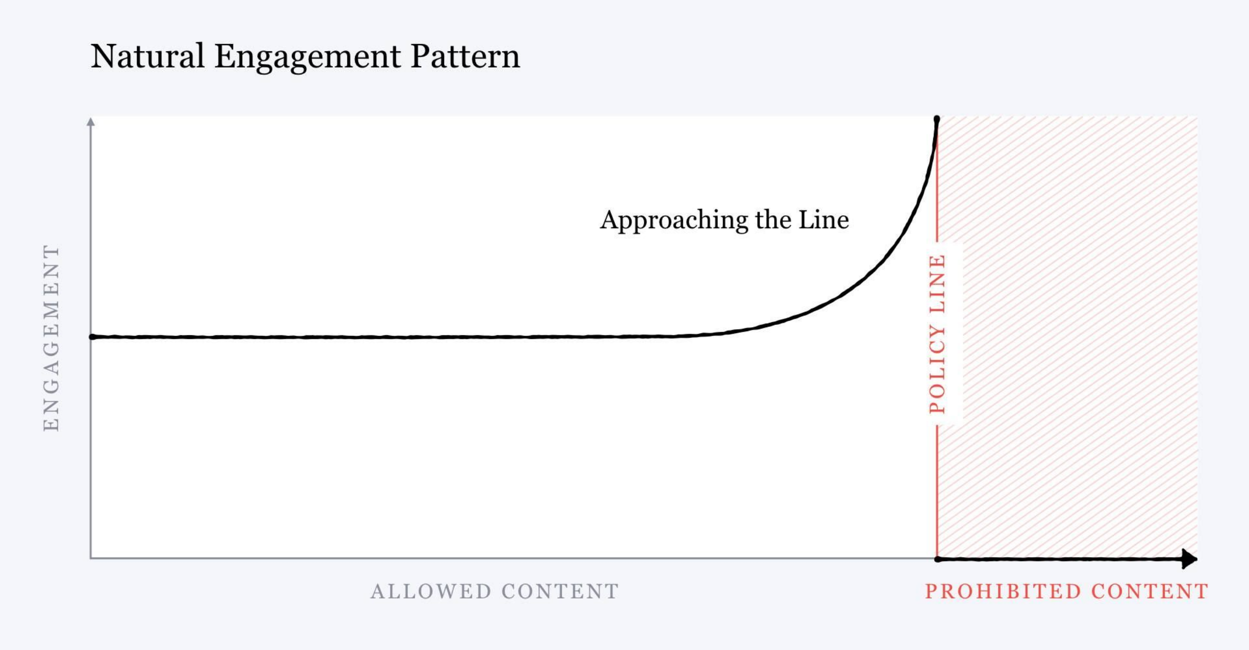


Higher

Predicted Engagement

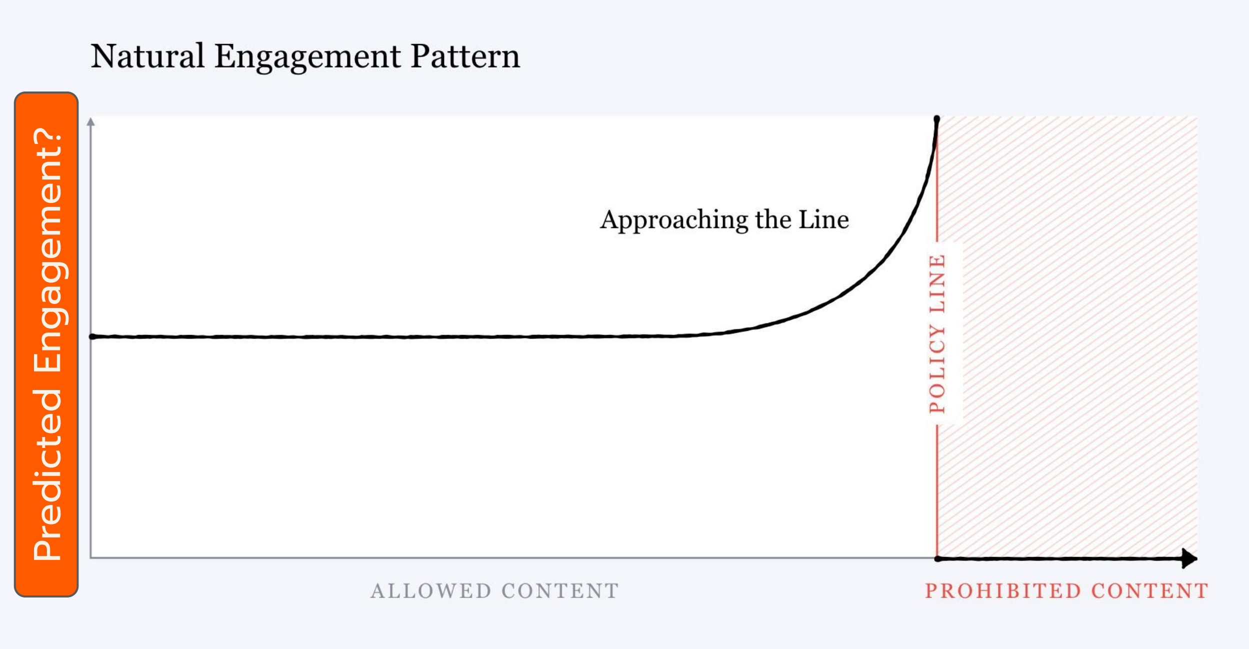
Lower

How Platform Design Can Amplify Harms



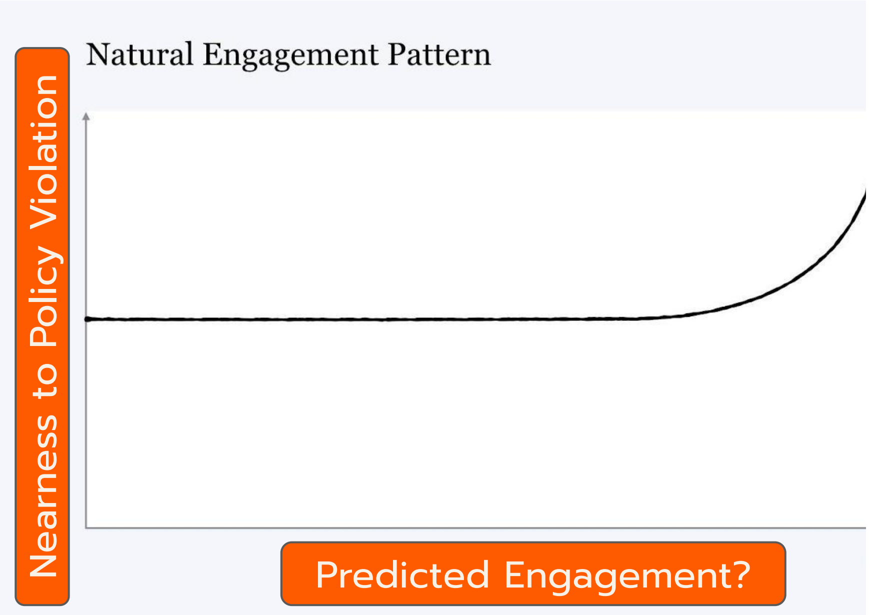
More engagement = more likely to be harmful

How Platform Design Can Amplify Harms



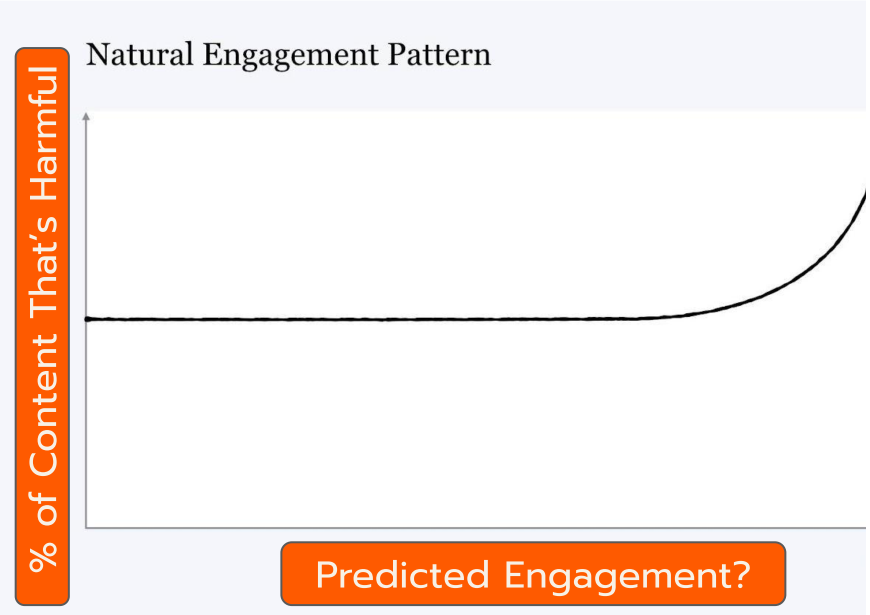
- Predicted engagement should follow actual engagement
- Content predicted to be engaging is more likely to be harmful

How Platform Design Can Amplify Harms



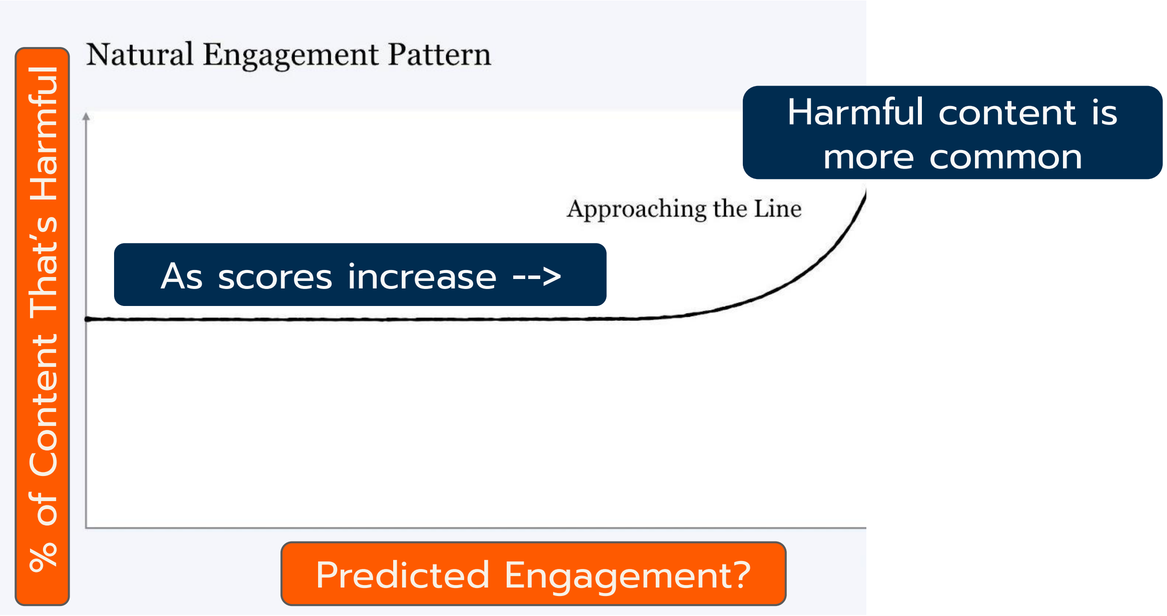
Let’s make it measurable: swap the X and Y Axes

How Platform Design Can Amplify Harms



“Nearness to policy” is not measurable; but % of content which is harmful is

How Platform Design Can Amplify Harms

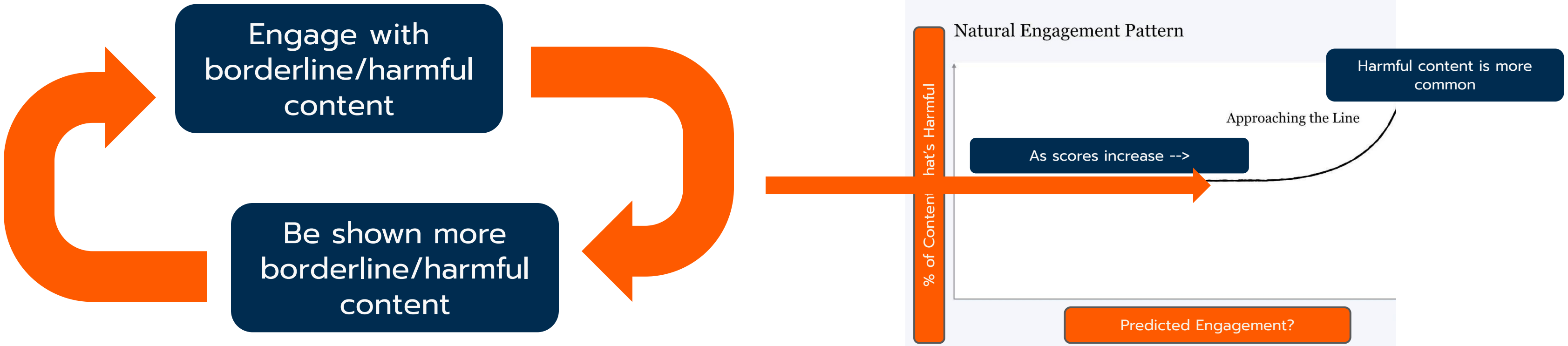


Harmful content will tend to "float to the top" of the ranking systems

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 a more extreme version of historical engagement
- Pushes people up and to the right on the ‘Natural Engagement Pattern’
- This is the “Rabbit Hole”



Engage with borderline/harmful content

Be shown more borderline/harmful content

Natural Engagement Pattern
% of Content That's Harmful
Harmful content is more common
Approaching the Line
As scores increase -->
Predicted Engagement?

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

The gravity of the system pulls towards bad behavior

This is the “gravity well” problem

Users will find a way to get around any specific restriction. The more you prevent it from happening, the greater the “potential energy”; the greater the rewards to those that do figure out how to bypass that barricade.

Instead, change the gravity – make it so that doing the *right* thing is where the gravity pulls.

This is the “gravity well” problem

The gravity problem isn’t just for user behavior – it applies to the *builders* of the platform too.

If you’re rewarded for tweaking the platform to maximize engagement, then not only is it a battle to make changes that don’t do that – you then have to prevent the rest of the company from reverting your changes accidentally.

Alternatives to engagement-based ranking

What Are Alternatives?

“Quality” focused ranking

- Google Search provides an example
- Define criteria for high and low quality content
- Release the criteria publicly for transparency and scrutiny
- Create ranking systems which estimate content quality

Source: Google Search, https://support.google.com/websearch/answer/9281931

Quality Ranking

Inventory

- Can be much broader, “All of internet”

Compute features

- Heavily influenced by “structural” features
- PageRank: How many links around the internet point to the content?

Run ML Models

- Used to predict objective quality and relevance assessments

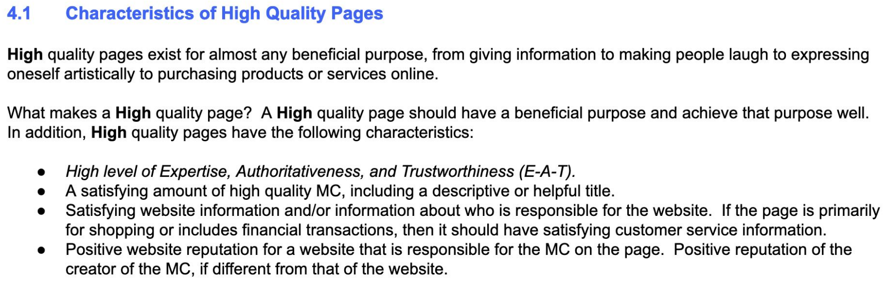
Output final ranked list

- Scoring high on quality ML models will push content up

Quality Ranking

High Quality

- Expertise, authoritativeness, and trustworthiness
- Information on who created and is responsible for content
- Positive reputation



4.1 Characteristics of High Quality Pages

High quality pages exist for almost any beneficial purpose, from giving information to making people laugh to expressing oneself artistically to purchasing products or services online.

What makes a **High** quality page? A **High** quality page should have a beneficial purpose and achieve that purpose well. In addition, **High** quality pages have the following characteristics:

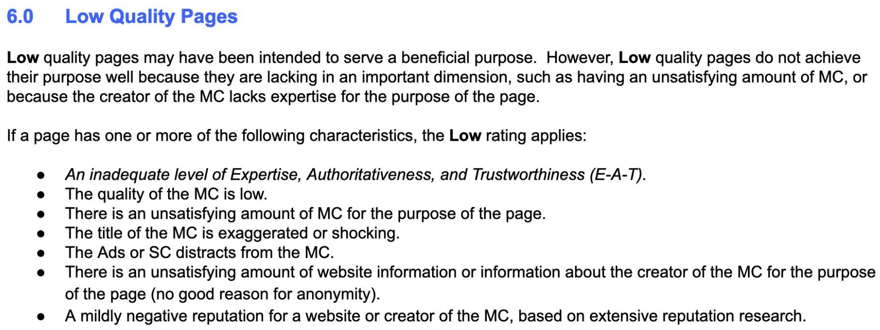
- *High level of Expertise, Authoritativeness, and Trustworthiness (E-A-T).*
- A satisfying amount of high quality MC, including a descriptive or helpful title.
- Satisfying website information and/or information about who is responsible for the website. If the page is primarily for shopping or includes financial transactions, then it should have satisfying customer service information.
- Positive website reputation for a website that is responsible for the MC on the page. Positive reputation of the creator of the MC, if different from that of the website.

Source: Google Search, https://support.google.com/websearch/answer/9281931

Quality Ranking

Low Quality

- Fails to serve a beneficial purpose or intended to be harmful
- Inadequate expertise
- Little information about who created content
- Negative reputation



6.0 Low Quality Pages

Low quality pages may have been intended to serve a beneficial purpose. However, **Low** quality pages do not achieve their purpose well because they are lacking in an important dimension, such as having an unsatisfying amount of MC, or because the creator of the MC lacks expertise for the purpose of the page.

If a page has one or more of the following characteristics, the **Low** rating applies:

- *An inadequate level of Expertise, Authoritativeness, and Trustworthiness (E-A-T).*
- The quality of the MC is low.
- There is an unsatisfying amount of MC for the purpose of the page.
- The title of the MC is exaggerated or shocking.
- The Ads or SC distracts from the MC.
- There is an unsatisfying amount of website information or information about the creator of the MC for the purpose of the page (no good reason for anonymity).
- A mildly negative reputation for a website or creator of the MC, based on extensive reputation research.

Source: Google Search, https://support.google.com/websearch/answer/9281931

Quality Ranking

It works!

- For conspiracy related searches, 2% of results on Bing are misinformation
- Vs. ~1% on Facebook overall (2016)

Source: Stanford Internet Observatory, 2019, https://cyber.fsi.stanford.edu/io/news/bing-search-disinformation
Source: Poynter, 2016,
https://www.poynter.org/fact-checking/2016/mark-zuckerberg-says-less-than-1-percent-of-facebook-content-is-fake-news-how-does-he-know/

How do we know if a ranked feed is contributing to societal harm?

Currently, the public doesn’t have the right data

Current regulatory environment:

- No requirement that platforms provide data demonstrating safety
- No requirement that platforms provide data on safety of design
- No requirement that platforms build responsibly

Data to Demonstrate Safety

- How many users are exposed to harmful content?
- Prevalence of harmful content
  - What % of all impressions on the platform are on violating content?
- Concentration of harmful content
  - Over a fixed time window, how many users are exposed to 1, 2, 3, 4 pieces of harmful content?
- Demographics of exposed users
  - Are certain ethnicities more likely to be exposed?
  - Are certain areas more likely to be exposed?
  - Are certain age groups?

Source: Integrity Institute, 2021, https://integrityinstitute.org/s/Metrics-and-Transparency-Summary-EXTERNAL.pdf

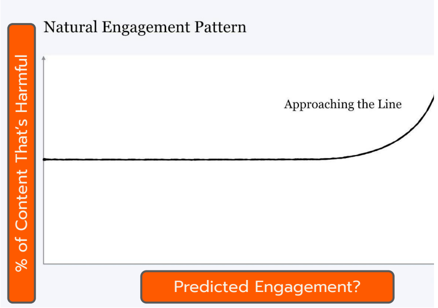
Data to Demonstrate Safety

- Random samples of impressions on public content
  - Released very regularly (daily, weekly)
  - Large number of samples (thousands, 10s of thousands)
- Random samples of impressions could be used by organizations monitoring social media
- Regularly report out on misinformation trends (medical, elections, etc.)

Safety of Design

Key check: Is platform in the “engagement problem”?

- Using all engagement actions a user has taken
- To predict all the future engagement actions a user might take
- For the purposes of maximizing engagement on the platform

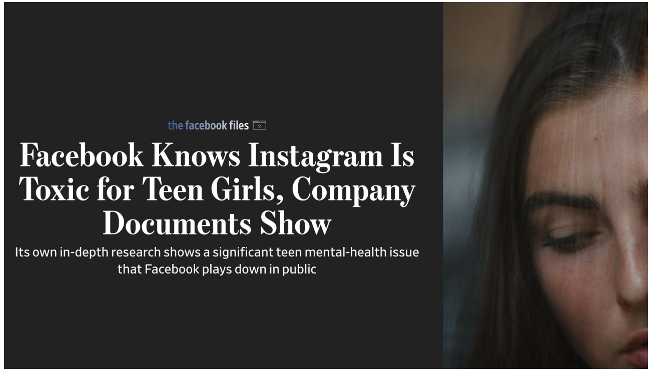


For models that influence ranking, how do they perform against harmful content?

Source: Integrity Institute, 2021, https://integrityinstitute.org/s/Ranking-and-Design-Transparency-EXTERNAL.pdf

Access to Users for Research

- Connect specific users to researchers
- How did platforms (IG) do this research?
    - Identify problematic usage
    - Get list of users that meet criteria
    - Reach out (email, in app notification)
    - Invite to participate in a study
- This process can be opened to valid external researchers in a privacy respecting manner



the facebook files

Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show

Its own in-depth research shows a significant teen mental-health issue that Facebook plays down in public

Source: Wall St. Journal, 2021, https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739

Conclusion

- Ranking by engagement is harmful.
- Topline metrics drive *user* behavior on platforms and *employee* behavior in companies.
- There are other ways to rank, such as quality-based ranking.
- Platforms need to provide:
  - Data on the scale and nature of harms on the platform
  - Public content datasets to raise awareness of harms
  - Reports on how ranking systems work
  - Access to users for valid research purposes

Thank you!