Ranked Feeds and the Harms They Pose

Sahar Massachi, 2022-07-06
ISD/DPL working group meeting

1

What is the Integrity Institute?

- Sahar Massachi, software/data engineer, data scientist, 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.
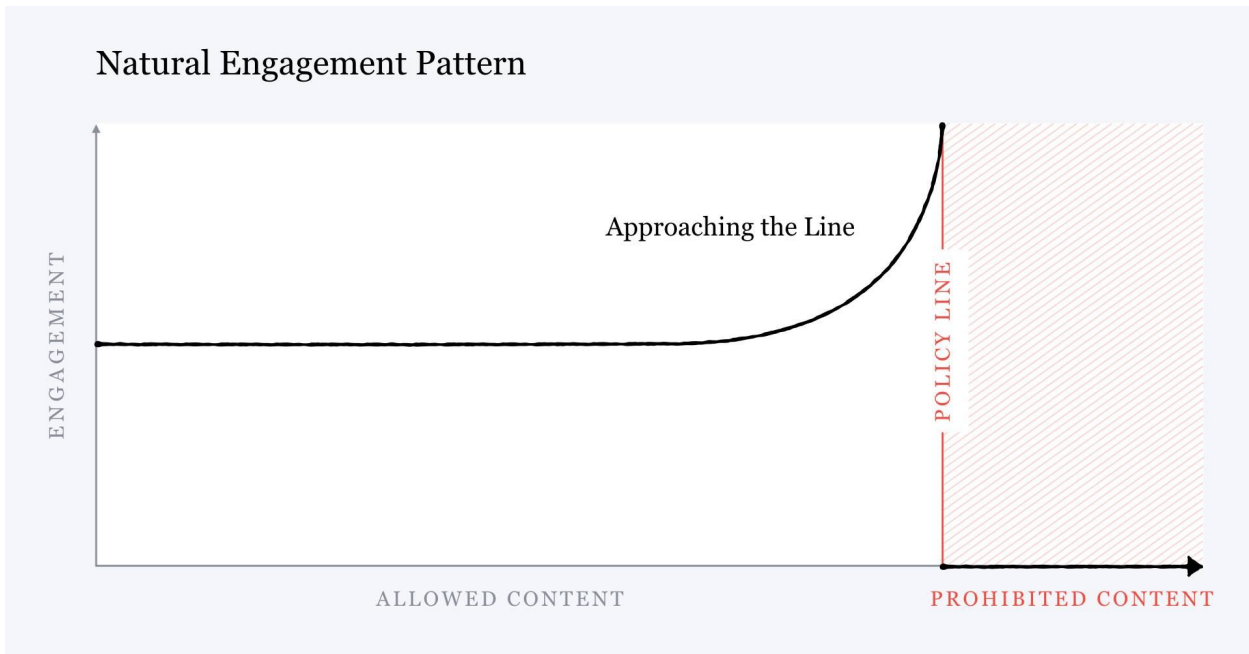- We are not comms professionals. Reach out if you have questions.

Outline

- How "algorithms" can amplify harmful and illegal content
- Understanding spread of content online (the lifecycle)
- The standard design of ranked feeds
- Conclusion
- Q/A
- *Appendix: Building with quality*
- *Appendix: What is Integrity Institute*

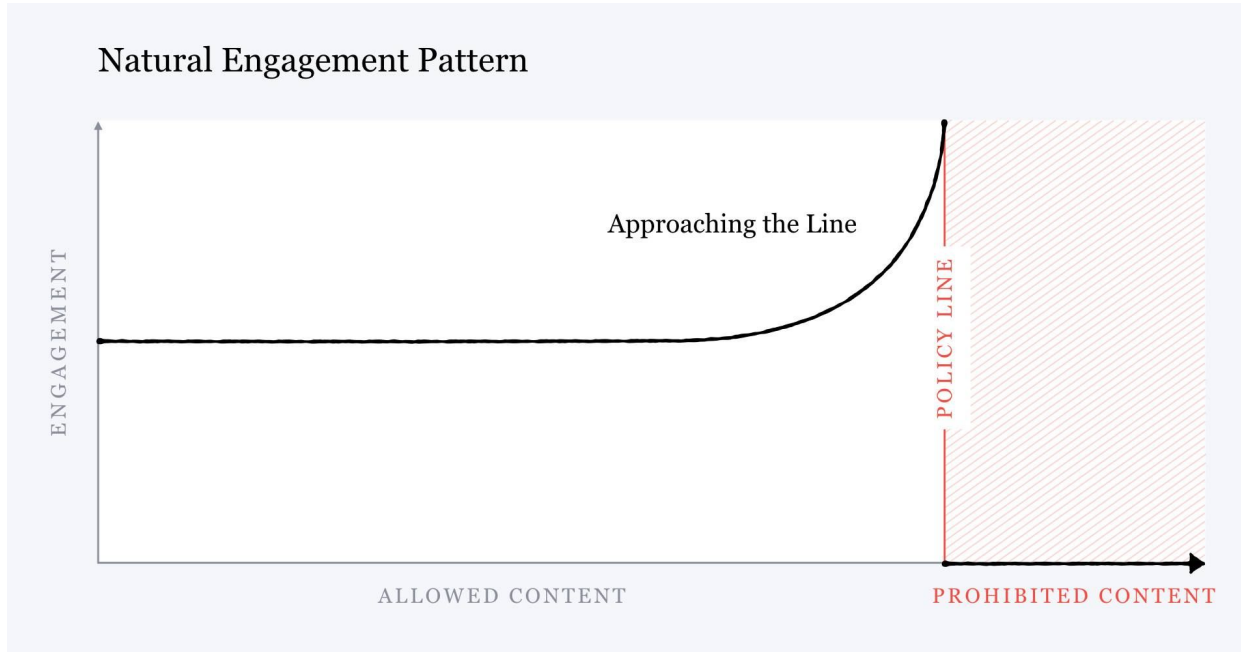# How algorithms can amplify harmful and illegal content

The Engagement Problem

Natural Engagement Pattern

Approaching the Line

ENGAGEMENT

POLICY LINE

ALLOWED CONTENT

PROHIBITED CONTENT

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

The Engagement Problem

Natural Engagement Pattern — chart showing ENGAGEMENT on the Y-Axis and content ranging from ALLOWED CONTENT to PROHIBITED CONTENT on the X-Axis, with a curve labeled "Approaching the Line" rising steeply toward the POLICY LINE.
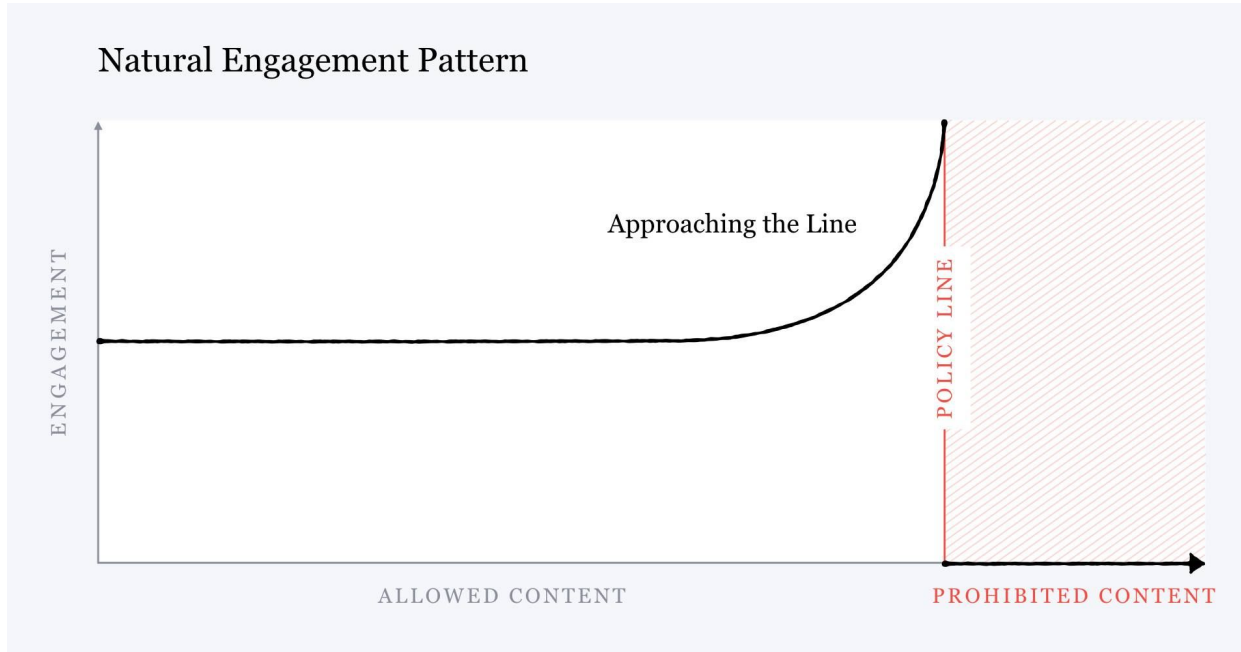
- X-Axis: What is allowed vs. prohibited?
  - Allowed content covers benign to borderline harmful
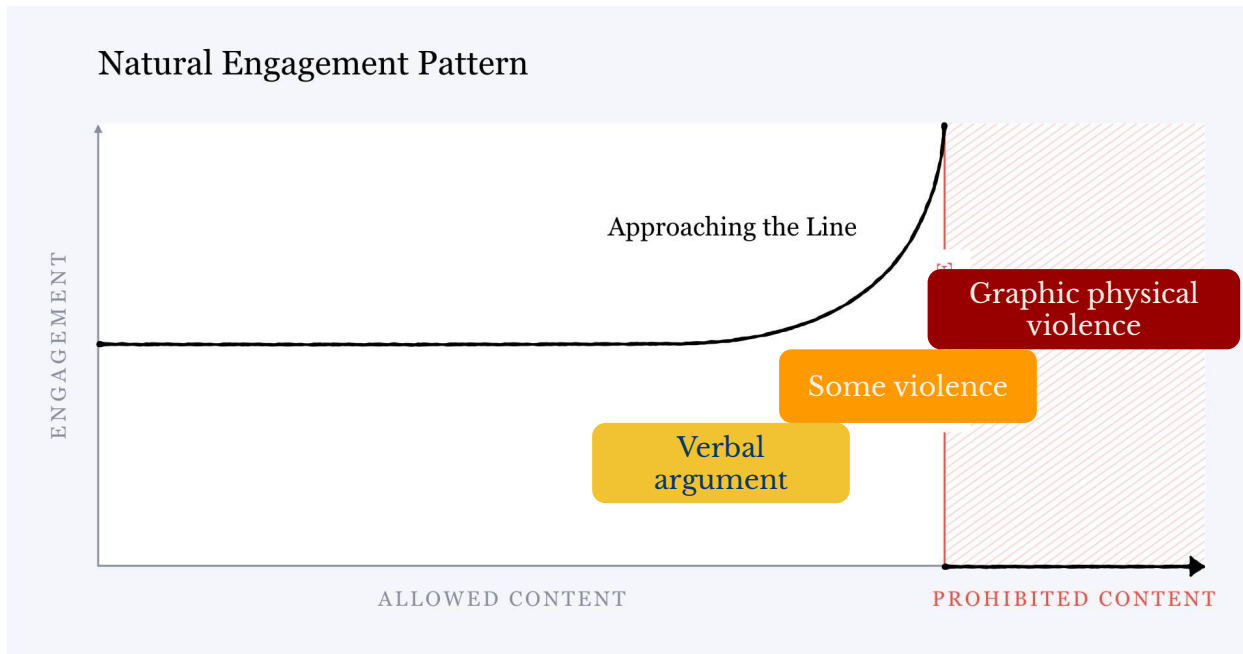 - Prohibited content is harmful

The Engagement Problem

Natural Engagement Pattern

ENGAGEMENT

Approaching the Line

POLICY LINE

ALLOWED CONTENT

PROHIBITED CONTENT

- This is true across many types of potential harms

The Engagement Problem

Natural Engagement Pattern

Approaching the Line

Graphic physical violence

Some violence

Verbal argument

ENGAGEMENT

ALLOWED CONTENT

PROHIBITED CONTENT

- This is true across many types of potential harms

The Engagement Problem

Natural Engagement Pattern

Approaching the Line

ENGAGEMENT

Self harm

Self harm adjacent

Sadness / depression

ALLOWED CONTENT

PROHIBITED CONTENT

- This is true across many types of potential harms

The Engagement Problem

Natural Engagement Pattern

Approaching the Line

Harmful Misinformation

Misinformation

Hyperbole and exaggeration

ENGAGEMENT

ALLOWED CONTENT

PROHIBITED CONTENT

- This is true across many types of potential harms

The Engagement Problem

Natural Engagement Pattern

Approaching the Line

ENGAGEMENT

POLICY LINE

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
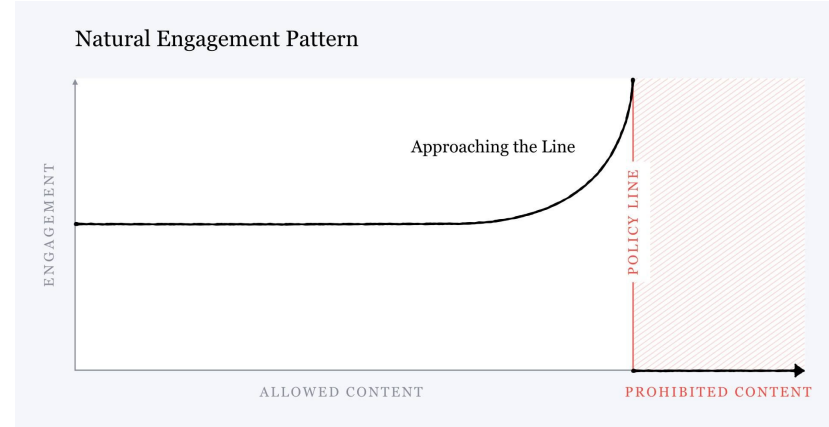
The Engagement Problem

- 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

Natural Engagement Pattern

Approaching the Line

ENGAGEMENT
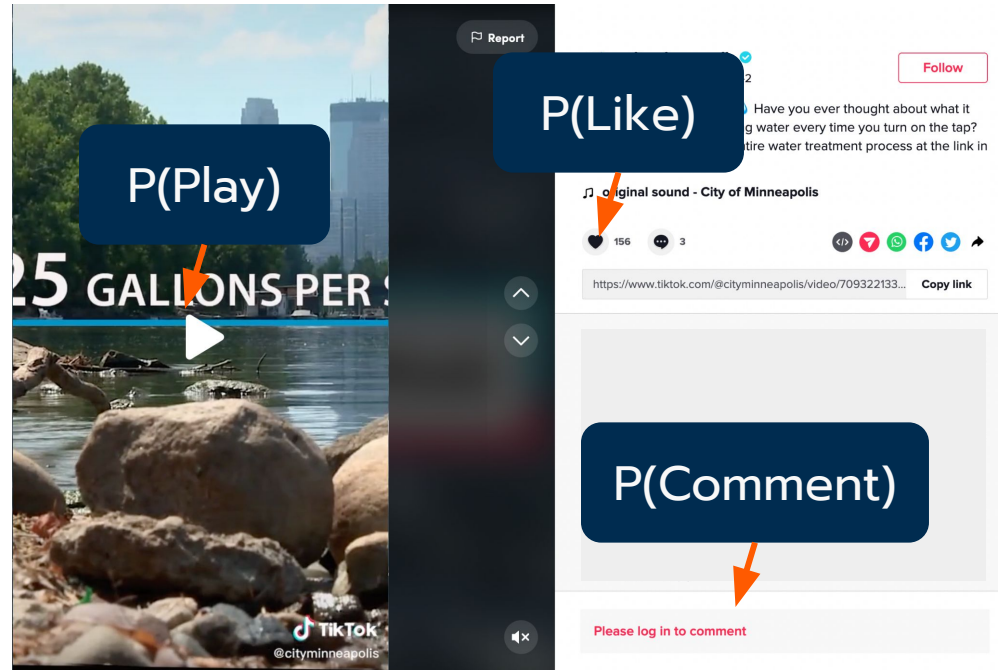
POLICY LINE

ALLOWED CONTENT

PROHIBITED CONTENT

How Most Platforms Work

- How do most platforms rank and order recommended content and accounts?
- We actually know for a number of them

How Most Platforms Work

- TikTok
- Predicting engagement
- Probability user will...
  - Like a video
  - Comment on a video
  - Play a video
  - Watch a video for an extended time

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.

HAPPY FATHER'S DAY

P(Like)   P(Comment)   P(Share)

👍😆❤️ 4.2   4K Comments   8 Shares

👍 Like   💬 Comment   ↗ Share

How Most Platforms Work

- Twitter
  - "Interesting and engaging"
- YouTube
  - Clicks
  - Watch Time
  - Surveys

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

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**

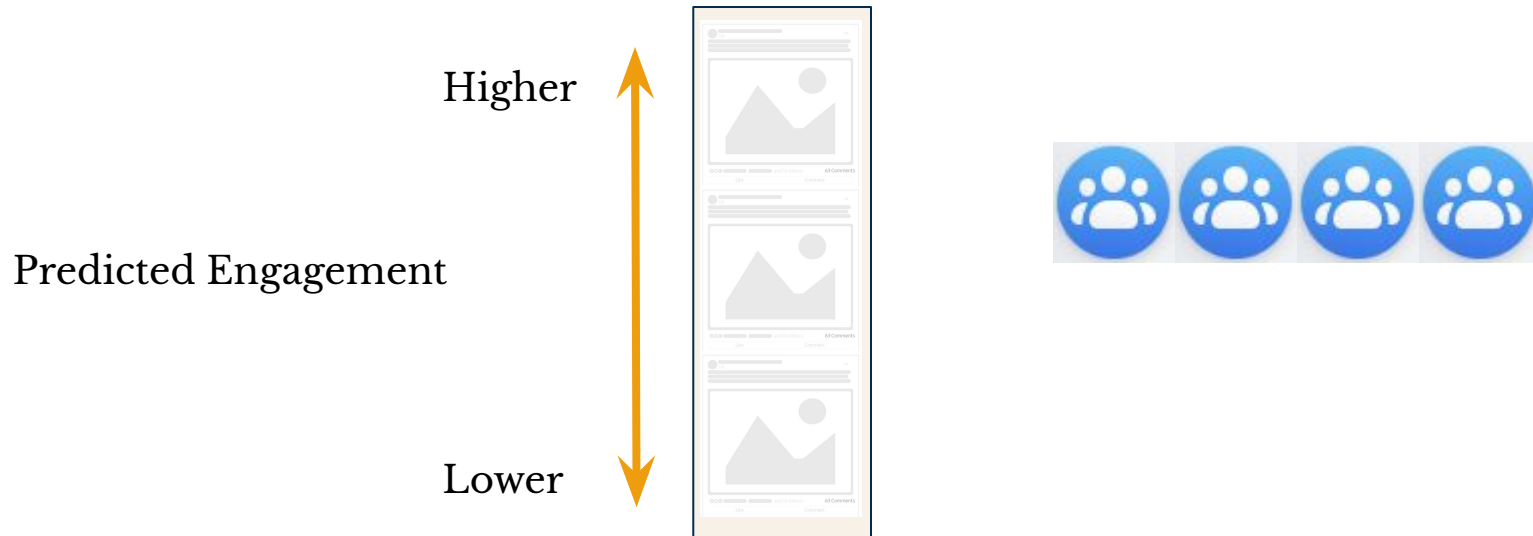**TikTok**   **Predicted Engagement**: Like, Comment, Watch

YouTube   **Predicted Engagement**: Clicks, Watch Time, Surveys
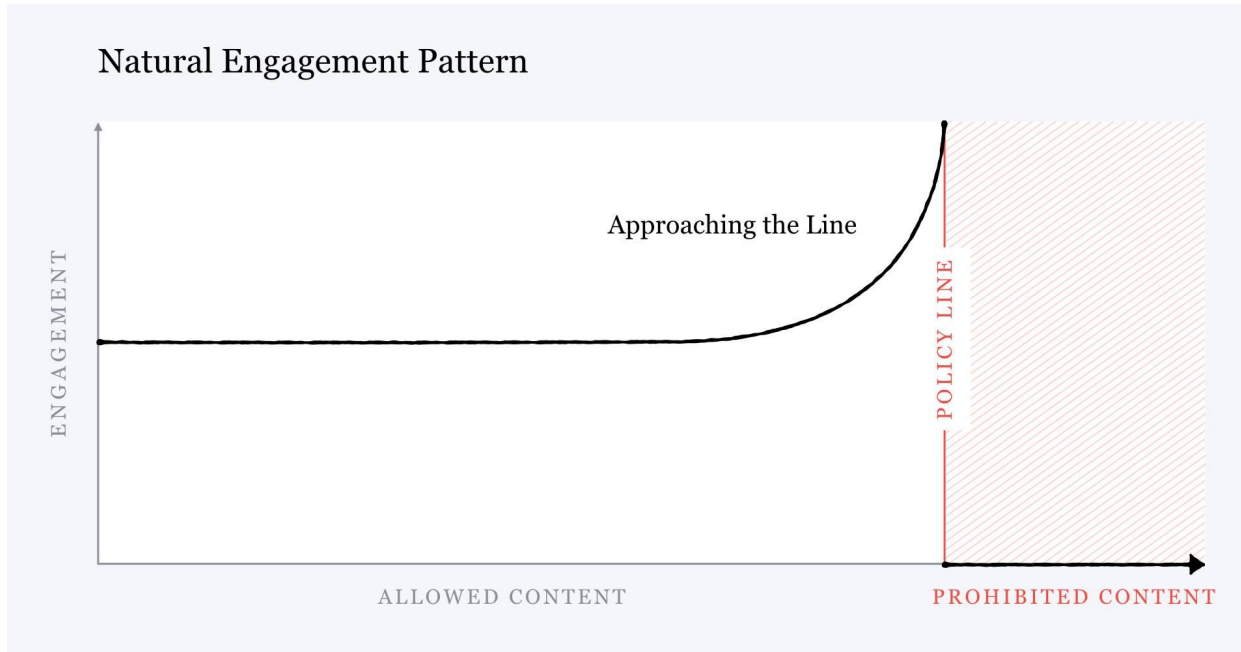
How Most Platforms Work

- Platforms recommend content and accounts most likely to be engaged with.

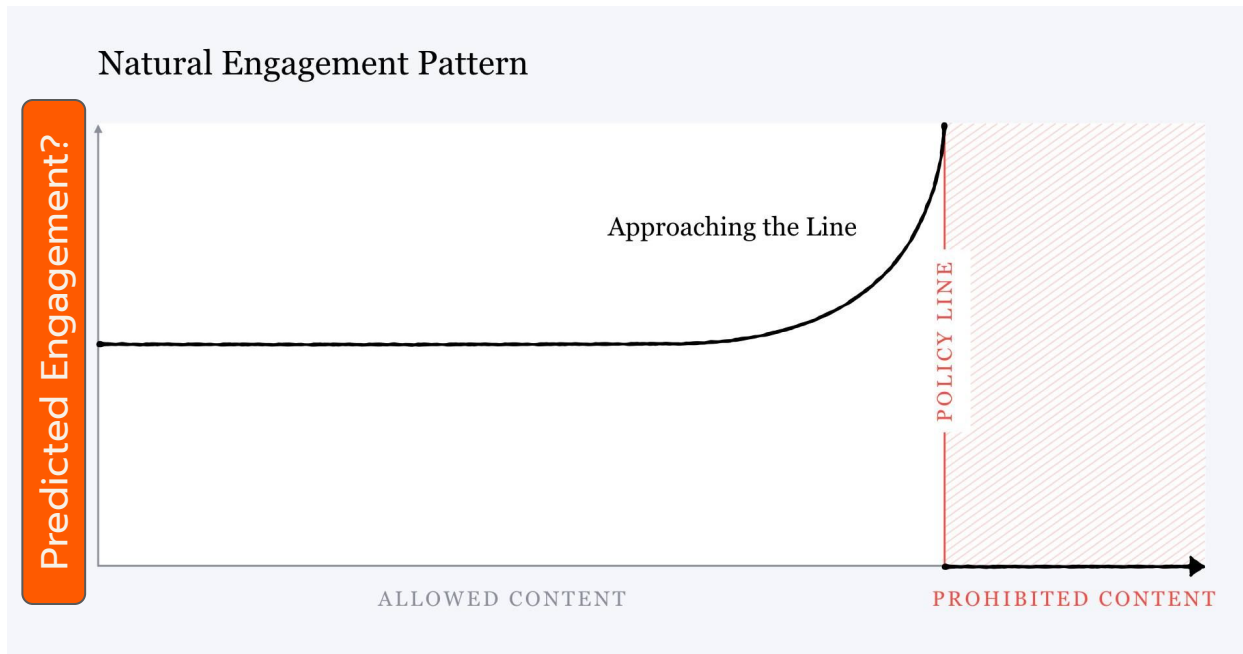- Why does this matter? Back to Zuckerberg's chart

Higher

Predicted Engagement

Lower

How Platform Design Can Amplify Harms

Natural Engagement Pattern

Approaching the Line

ENGAGEMENT

POLICY LINE

ALLOWED CONTENT

PROHIBITED CONTENT

- More engagement, more likely to be harmful

How Platform Design Can Amplify Harms

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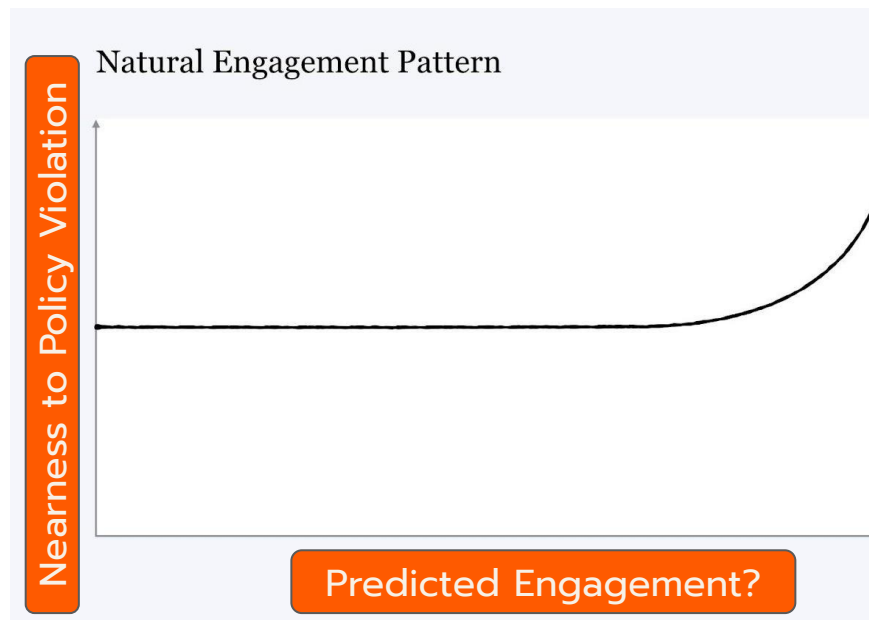
- Predicted engagement should follow actual engagement
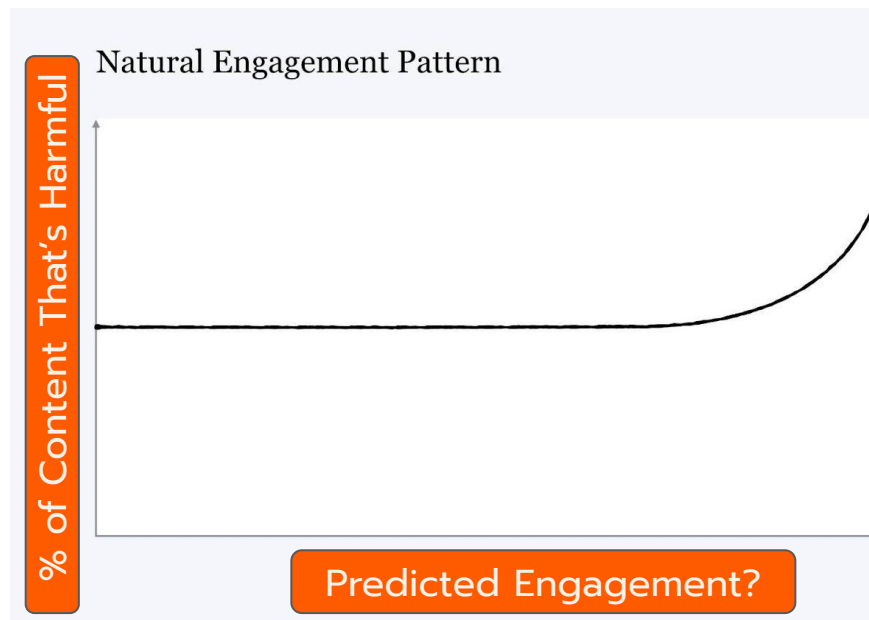- Content predicted to be engaging is more likely to be harmful

How Platform Design Can Amplify Harms

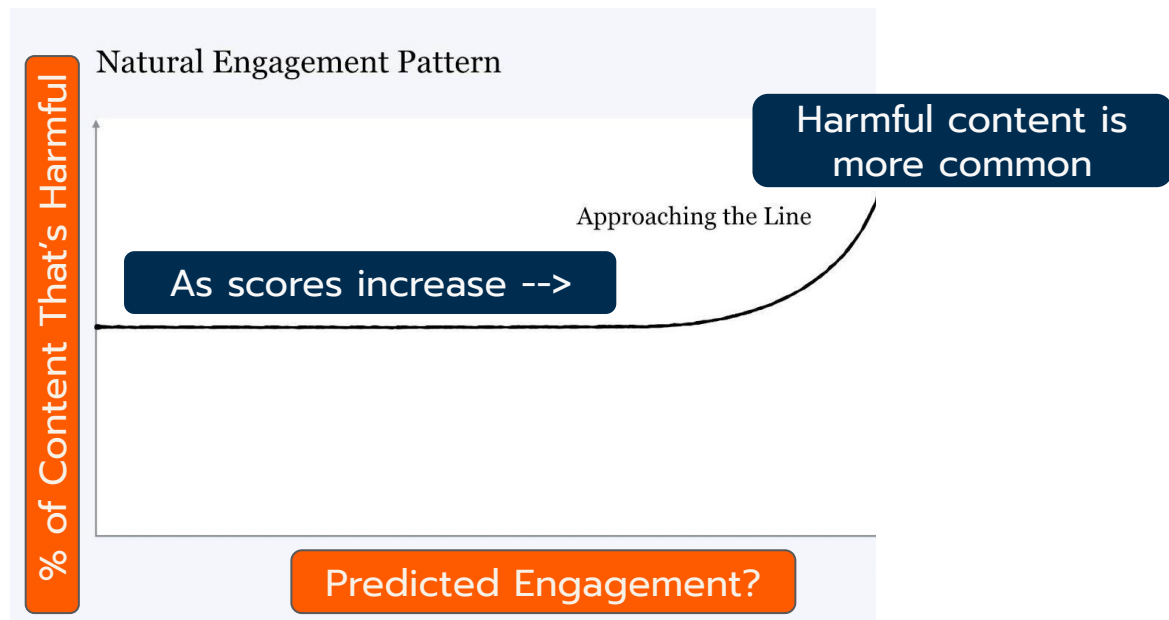- Let's make it measurable
- Swap the X and Y Axes

How Platform Design Can Amplify Harms

- "Nearness to policy" is not measurable
- % of content which is harmful is
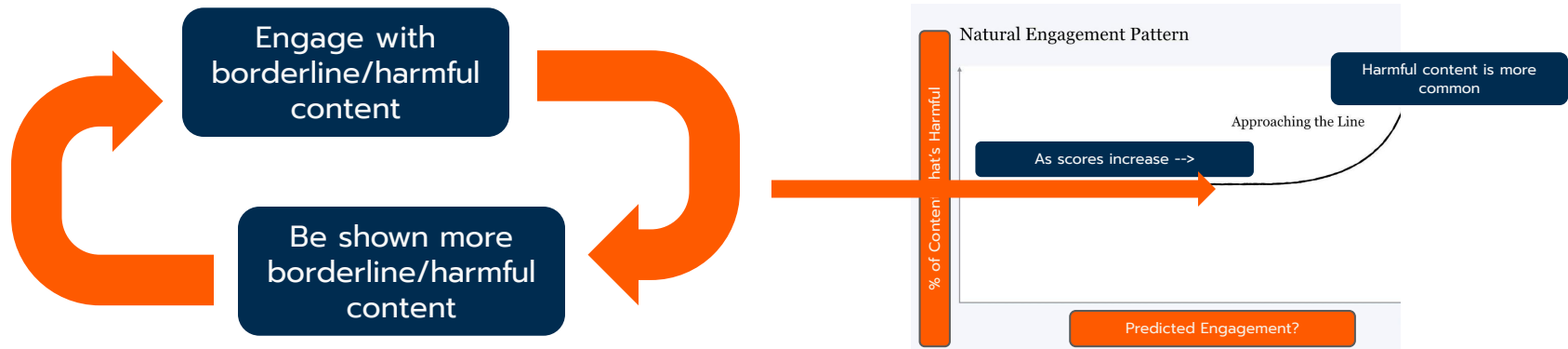
How Platform Design Can Amplify Harms

- 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
- 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

How Does This Problem Manifest?

- Platforms track everything users engage with
- They u~~~~dn~~~~ ~~~~l~~~~ ~~~~ll~~~~ ~~~~id~~~~ ~~~~f~~~~
- 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.

Be shown more borderline/harmful content

~~~~ul~~~~ content is more common

% of Con

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

- User behavior that maximizes engagement gets promoted by the platform
- But *some* user behavior that maximizes engagement is a terms of service violation
- The "normal" way to deal with this is via banning that specific type of behavior
- The gravity of the platform still pulls towards maximizing engagement
- 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.
- Like trying to put a little 10 foot damn in front of a river's path into the valley – the river will find a way.
- **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.
- You want to change the gravity so that doing the right thing is easy, doing the harmful thing is hard.

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
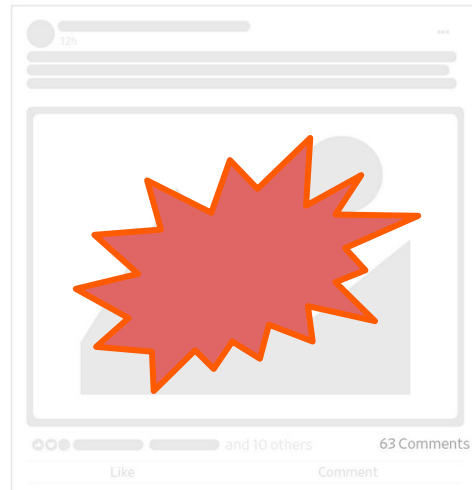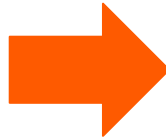
Lifecycle of Harmful Content

- 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
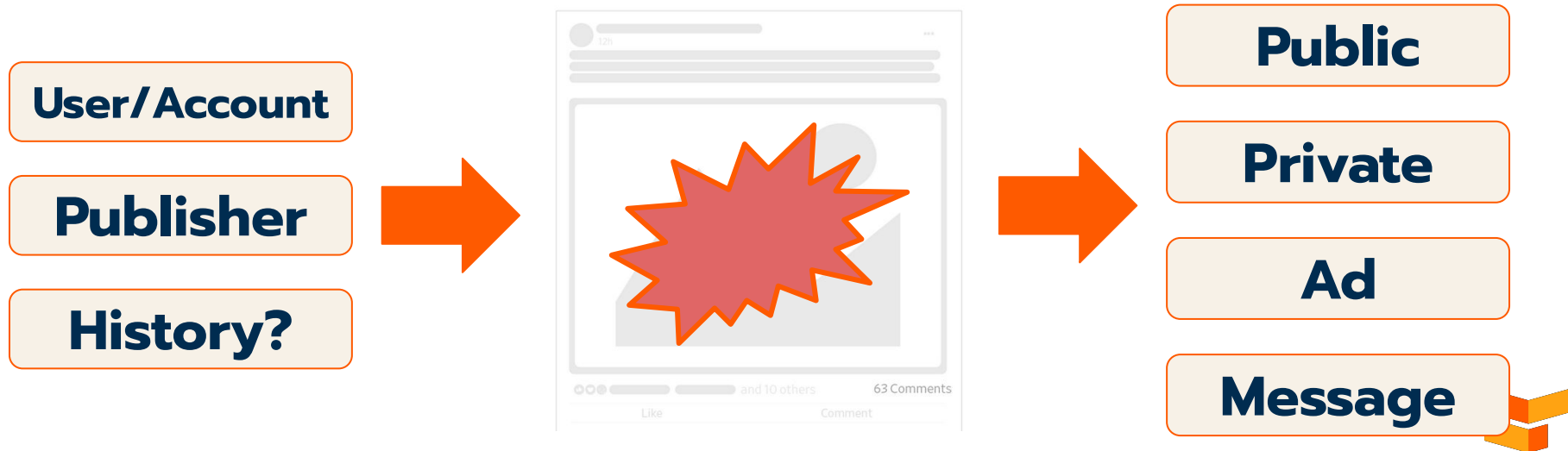
User/Account

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)

User/Account

Publisher

History?

Public

Private

Ad

Message

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?
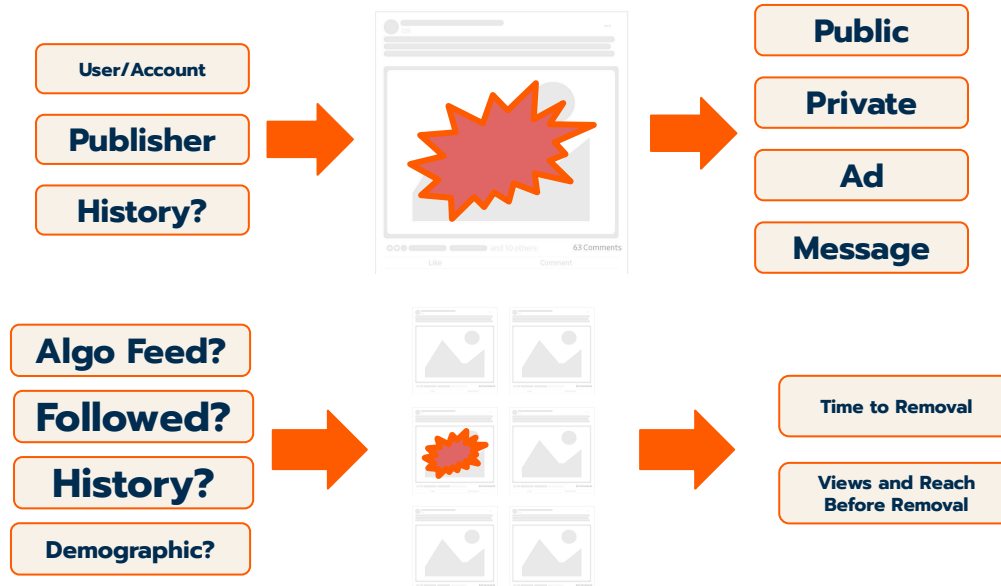
Lifecycle of Harmful Content

- 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

User/Account

Publisher

History?

Public

Private

Ad

Message

Algo Feed?

Followed?

History?

Demographic?

Time to Removal

Views and Reach Before Removal

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

The Standard Design of Ranked Feeds

Ranking Basics

- ▷ Ranking systems all have similar components

    - ➢ Specifically for all the social platforms

- ▷ The purpose of these components are

    - ➢ Gather content

    - ➢ Score content

    - ➢ Produce final ranked list

Ranking Basics

Inventory

Ranking Basics

Inventory

- ▷ 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

X, Y, Z

Ranking Basics

Inventory

Features

X, Y, Z

▷ Inventory

▷ Features

> ➢ "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**

X, Y, Z

Like?
Comment?
Retweet?

▷ Inventory

▷ Features

▷ ML Model Scoring

    ▷ 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?"

    ▷ Basically, each model predicts the probability of a specific user action or property of the content

Ranking Basics

Inventory

Features

X, Y, Z

ML Models

Like?
Comment?
Retweet?

Ranking

43.8

28.2

8.7

Ranking Basics

Inventory	Features	ML Models	Ranking

X, Y, Z

Like? Comment? Retweet?

43.8

28.2

8.7

▷ Inventory

▷ Features

▷ ML Model Scoring

▷ Final Ranking Score

    ▷ All the classifier scores are combined, business logic applied

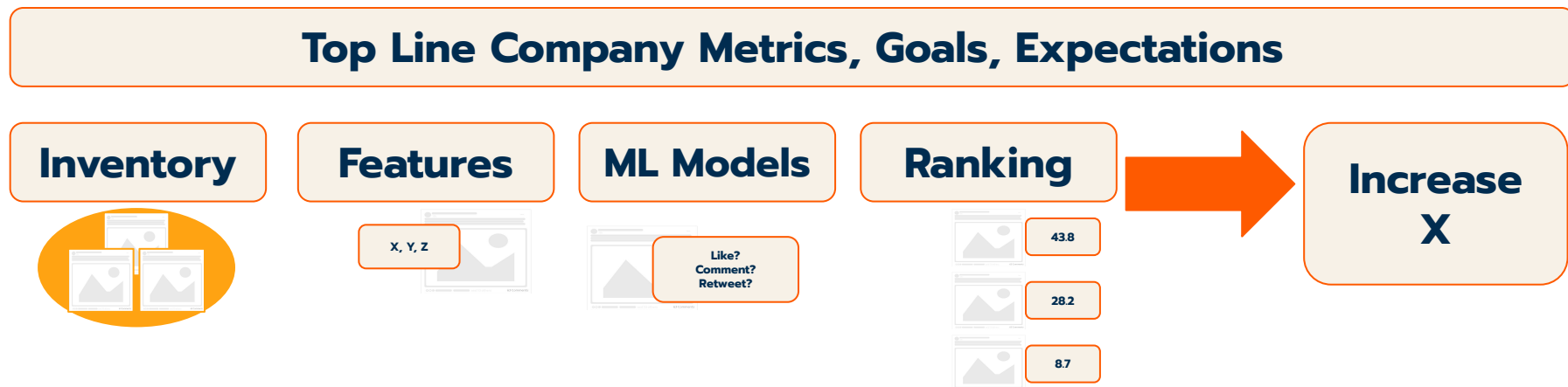   ▷ Final sorting and list generated

Standard Design

- ▷ The ranking system is internal
  - ➢ Managed by the company, by the team responsible for it
- ▷ The company has objectives for the ranking system
  - ➢ This could be "top line" metrics they use to evaluate how well the ranking system is doing
- ▷ The company and team have goals and metrics
  - ➢ These are how they decide whether to launch changes

Standard Design

▷ This process is mediated by the companies' goals and experimentation process

Standard Design

▷ This process is mediated by the companies' goals and experimentation process

Inventor

ncrease X

8.7

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

Standard Design

▷ This process is mediated by the companies' goals and experimentation process

Inventor

ncrease X

8.7

Key Takeaway: While all parts are important, the topline company metrics and goals are the foundational "gravity" that drives everything else.

Standard Design

▷ There are two basic frameworks or paradigms used in ranking systems

> User Engagement focused ranking

> Quality focused ranking

Standard Design: Engagement Ranking

▷ Inventory
  - ➢ Collect posts, including non-followed
  - ➢ Almost all platforms have mechanism for unfollowed accounts
    - ○ Retweets, reshares, feed of all public content
    - ○ This enables huge reach of content, beyond initial audience

▷ 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

▷ All in service of company level goals, often quantifiable engagement metrics

Standard Design: 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

▷ All in service of company level goals, often quantifiable quality estimates

Conclusion: understanding ranked feeds

Conclusion

▷ Ranking by engagement is harmful
  ▷ Bad content rises to the top
  ▷ Patchwork fixes can't stop that general trend
  ▷ True for any type of bad content: harassment, hoaxes, hate speech, self-harm, spam, propaganda, etc
  ▷ This is a gravity that makes bad *behavior* easier and good behavior harder
▷ We can understand ranking systems through several lenses of transparency
  ▷ Lifecycle of Harmful Content – transparency around bad content and who sees it and *why*
  ▷ Standard Design – transparency around the creation of inventory, and each step of the way
  ▷ The Engagement Trap – are platforms turbocharging engagement rabbitholes?
  ▷ Four Pillars (not discussed here, but on our website): Aggregate Metrics, Content Samples, Ranking Transparency, Process Transparency
▷ Topline metrics drive *user* behavior on platforms and *employee* behavior in companies
  ▷ They're the skeleton key to understanding the entire black box
▷ You can measure societal harms in many ways
  ▷ On-platform data, off-platform data, surveys of specific users
  ▷ Two on-platform lenses: overall prevalence, and concentrated problems

Appendix: Quality and Transparency

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

What Are Alternatives?

- 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

What Are Alternatives?

- 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

What Are Alternatives?

- And it helps!
- For conspiracy related searches, 2% of results 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/

What Data Do We Need From Platforms?

- 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

What Data Do We Need From Platforms?

- Current regulatory environment
- **No requirement that platforms provide data demonstrating safety**
- **No requirement that platforms provide reports on safety of design**
- No requirement that platforms build responsibly

Data to Demonstrate Safety

- This is a huge topic, but highlights
- We have briefing on "Lifecycle of Harmful Content"
- 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
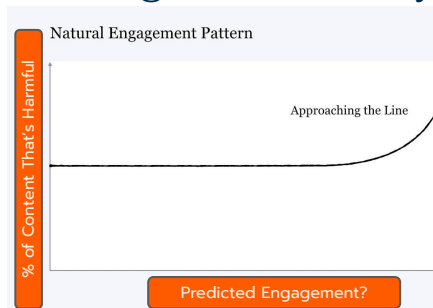
Data to Demonstrate Safety

- Random samples of impressions on public content
  - Released very regularly (daily, weekly)
  - Large number of samples (thousands, 10's of thousands)
- If the platforms are going to show medical conspiracy theories to 22M people, they need to report that fact sooner than 3 months after the fact
- Random samples of impressions could be used by organizations monitoring social media
- They could regularly report out on medical misinformation trends
- So you can be aware of misinformation trends before they show up in your office
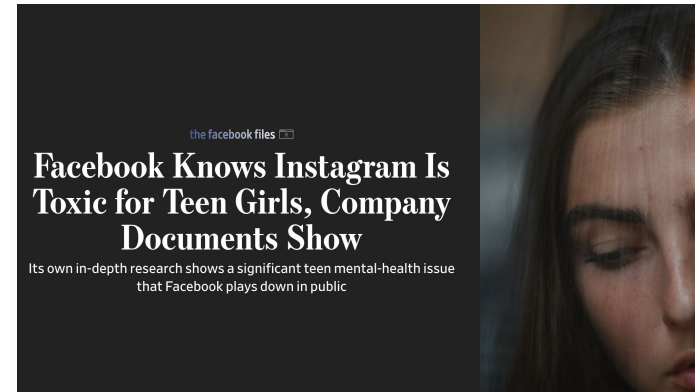
Safety of Design

- Again, this is a huge topic
- We have briefing on "Ranking and Design Transparency"
- 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

Conclusion

- The "Engagement problem"
  - Most platforms use it
  - Can exploit cognitive biases
  - Amplify harmful content
    - Harassment, Hoaxes, Hate Speech.
    - Addictive behavior by users. Bad behavior by attackers.
- There are alternatives some platforms use
  - Google Search and Quality Focused 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

Appendix: Integrity Institute

Integrity Institute

Pitch Deck

The social internet should help individuals, society, and democracy thrive

Instead, we're seeing…

This Tweet violated the Twitter Rules about glorifying violence. However, Twitter has determined that it may be in the public's interest for the Tweet to remain accessible. Learn more

View

The Washington Post
Democracy Dies in Darkness

Sections

THE TECHNOLOGY 202 • Analysis

Facebook's most popular posts show how it's vulnerable to exploitation, report finds

By Cristiano Lima

March 30, 2022 at 9:06 a.m. EDT

BLEEPINGCOMPUTER

NEWS · DOWNLOADS · VIRUS REMOVAL GUIDES · TUTORIALS

Twitter bots pose as support staff to steal your cryptocurrency

By Lawrence Abrams · December 7, 2021

▷ Spam, hoaxes, bots, harassment, hate speech, misinformation, disinformation, and other harms

▷ Bad behavior being **structurally rewarded** by platforms - untrusted companies acting in untrustworthy ways

▷ **Global spread**: viral lynchings in India; incitement of genocide in Myanmar; civil violence in Ethiopia

▷ **Community-wide impact**: even if you abstain from using a social media platform, your neighbors, family members, and fellow voters still do

Many groups try to address this problem, but...

THEY ALL LACK EXPERIENCE ACTUALLY WORKING ON PLATFORMS

This is important because people with experience can:

- **Credibly explain** what approaches actually work
  - Educate stakeholders on issues and how to fix them
    - Policymakers, advocacy groups, NGOs, journalists, and academics
  - Help companies decide what to build and how to do it
  - Share best practices around technical solutions
- **Credibly rebut** bad-faith arguments
  - "Company X is bad because they did Y"
  - "Proposed regulation Z is impossible to implement, and thus shouldn't be adopted"
- Frame the problem correctly and **avoid rabbitholed arguments**
  - "Censorship vs. Free Speech" is a distraction, and we can tell you why
  - "This isn't actually hate speech" vs. "Yes, it is" is not relevant

Integrity Professionals Are The Key

Integrity Professionals are tech workers with experience in **Integrity roles** - roles dedicated to addressing harms to people and societies within social Internet platforms

Integrity Professionals have experience in tackling these and other issues on behalf of social Internet companies:

- ▷ Ethical Design
- ▷ Harassment
- ▷ Hate Speech
- ▷ Hoaxes
- ▷ Spam
- ▷ Inauthentic Behavior
- ▷ Data Transparency and Reporting
- ▷ Information Operations
- ▷ Content Quality
- ▷ Espionage

- ▷ Disinformation
- ▷ Misinformation
- ▷ Digital Crime
- ▷ Information Ecosystems
- ▷ Child Safety
- ▷ Counter-terrorism
- ▷ Human Trafficking
- ▷ Toxicity
- ▷ Impersonation
- ▷ … and more!

This expertise was locked inside a few big companies until the Integrity Institute

Our members have experience working at platforms across the industry
(including all the companies seen here)

The Integrity Institute brings on-platform experience to the people theorizing, building, and governing the social Internet

▷ Diversity in experience: **over 70 active members** spanning **more than 18 platforms**

▷ Ethics-first: **non-profit 501(c)(3)**, foundation-funded

▷ Breadth of knowledge: active global contributions **across the Integrity spectrum** - developing the theory of Integrity, writing effective legislation, advising on technical implementation, monitoring for compliance, and more

OMIDYAR NETWORK

*"We are proud to support the Integrity Institute's efforts to shape a better social internet. The organization and its network has already **achieved so much in a short period of time, and is key to our collective ability to hold platforms accountable to the public interest**."*
– Anamitra Deb, Omidyar Network

THE WALL STREET JOURNAL

*"The institute is now advising **lawmakers and think tanks around the world**"*

The New York Times

2021 Good Tech Award

We Bridge Integrity Professionals To The World

Tech Companies
▷ How do we build a platform that will work over a long period of time?
▷ How do we stop harms before PR crises?
▷ How do we get credit when we do well?
▷ What are easy solutions that work?

Policymakers and NGOs
▷ How do platforms work?
▷ What regulation would have a positive impact on business incentives?
▷ What are the real problems to solve?

Other Integrity Workers
▷ Has anyone worked on this problem before?
▷ How do I convince my company to prioritize my team?
▷ How do we get regulation that helps us do our job?

Academia and Researchers
▷ Are we studying the right issues?
▷ Are we using the right data?
▷ How do you study platforms from the inside?

Our Team

Jeff Allen
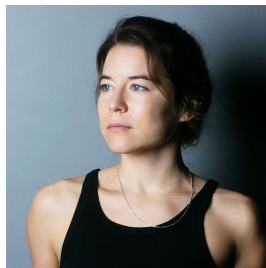Co-founder, Chief Research Officer

Sahar Massachi
Co-founder, Executive Director

Rachel Fagen
Director of Operations

Cassandra Marketos
Advisor, Community Engagement

Tim Gavin
Community Product Manager

Katie Harbath
Community Advisory Board Chair

Samidh Chakrabarti
Advisory Board Member

Our Fellows

We have validated key parts of this model

▷ There is huge need and demand for this expertise
  ➢ Governments, NGOs, advocacy groups, academics, and platforms
▷ Our members can meet this demand
  ➢ They have the necessary knowledge and experience
  ➢ However, their time is precious and limited
▷ We need more staff to unlock this impact
  ➢ **Add scale**: triage and prioritize the flood of interest
  ➢ **Add structure:** maintain and execute clear async workflows
  ➢ **Add speed:** help push out published artifacts across all active workflows
  ➢ **Add strategy:** understand what are the most important problems to tackle
  ➢ **Add support:** nurture the community and its members

How We Deliver Impact

▷ Social internet companies **do Integrity better**
  ➢ Through legislation, regulation, self-regulation, compliance, etc.
  ➢ We shine a clear light on the *what*; we provide guidance on the *how*

▷ People **understand Integrity better**
  ➢ We give shape to the idea of an **Integrity worker**
  ➢ We bring together various Integrity roles (e.g. trust and safety, anti-spam, anti-ad fraud, threat intelligence, feed ranking) **under a shared identity, working on a shared project**

▷ Integrity Professionals have **more power**
  ➢ Integrity workers, by and large, know *how* to fix the social Internet, but they often lack the power and position inside companies to do so
  ➢ We advocate for companies to change their organizational design to **give these professionals the means and the power to solve these problems** in an ongoing fashion
  ➢ The role of "doing the right thing" is **honored, not silenced**

Our Current Funding

We have raised **over $1M** to date from foundations and small donors

We are committed to remaining **independent of grants from platforms**

To implement the work already underway, we need an additional **$1M in 2022, $2.8M in 2023, and $3.6 M in 2024**
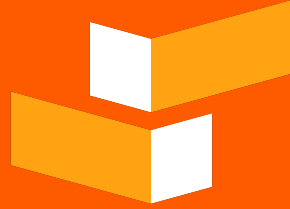
We have other time-sensitive, high-impact opportunities that can be unlocked if we receive additional funding beyond our current commitments (est. $2M or more)