Social Media and the Spread of Harmful Content

And what data could help mitigate harm within medical community

Jeff Allen, 2022-06-22

Disclosure of Relevant Financial Relationships/Content Validation/ Copyright/Evidence Citations

- Under the ACCME Standards for Integrity and Independence, everyone who is in a position to control the content of an education activity must disclose all relevant financial relationships with any ineligible companies. An "ineligible company" is any entity whose primary business is producing, marketing, selling, re-selling, or distributing healthcare products used by or on patients. Financial relationships of any dollar amount are defined as relevant if the educational content is related to the business lines or products of the ineligible company.
- Today's Presenters/Speakers, Planners and Committee Members have disclosed that they have no relevant financial relationships with ineligible companies. The CME Department has reviewed their disclosure information for the planners/presenters and/or committee/faculty for this program and they do not have relevant financial relationships with ineligible companies.

Ensure Content is Valid

Accredited providers are responsible for ensuring education is fair and balanced and that any clinical content presented supports safe, effective patient care.

1. All recommendations for patient care in accredited continuing education must be based on current science, evidence, and clinical reasoning, while giving a fair and balanced view of diagnostic and therapeutic options.
2. All scientific research referred to, reported, or used in accredited education in support or justification of a patient care recommendation must conform to the generally accepted standards of experimental design, data collection, analysis, and interpretation.
3. Although accredited continuing education is an appropriate place to discuss, debate, and explore new and evolving topics, these areas need to be clearly identified as such within the program and individual presentations. It is the responsibility of accredited providers to facilitate engagement with these topics without advocating for, or promoting, practices that are not, or not yet, adequately based on current science, evidence, and clinical reasoning.
4. Organizations cannot be accredited if they advocate for unscientific approaches to diagnosis or therapy, or if their education promotes recommendations, treatment, or manners of practicing healthcare that are determined to have risks or dangers that outweigh the benefits or are known to be ineffective in the treatment of patients.

Copyright

- Today's Presenters/Speakers, Planners and Committee Members have been advised to obtain required permissions for all copyrighted materials being used from textbooks, journals, or other print or electronic media.

Evidence Citations

- Evidence citations are required for clinical presentations. Presenters/Speakers, Planners and Committee Members are required to list or attach evidence-based references that support the clinical practice concerns and recommendations that you will be presenting.

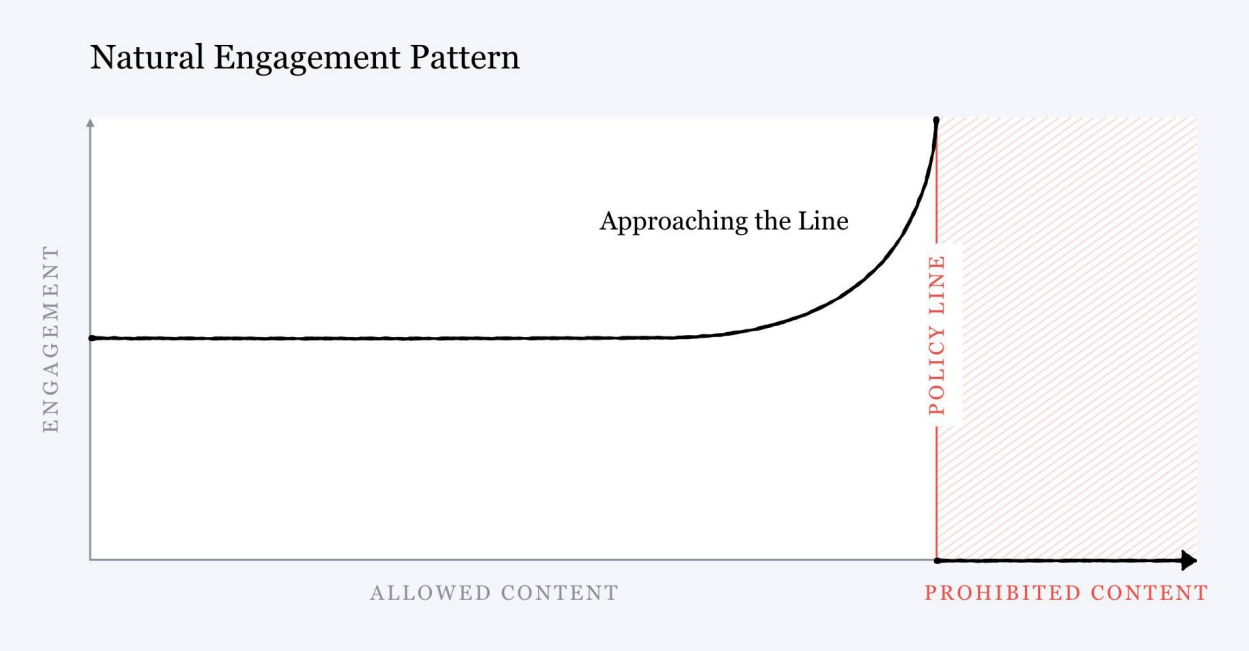
Background

- Jeff Allen, a doctor but not that kind of doctor (Physics), data scientist by trade (Facebook, Instagram)
- 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

- The engagement problem
- How most platforms work
- How platform design can amplify harms
- Alternatives to engagement centric design
- The data platforms need to provide

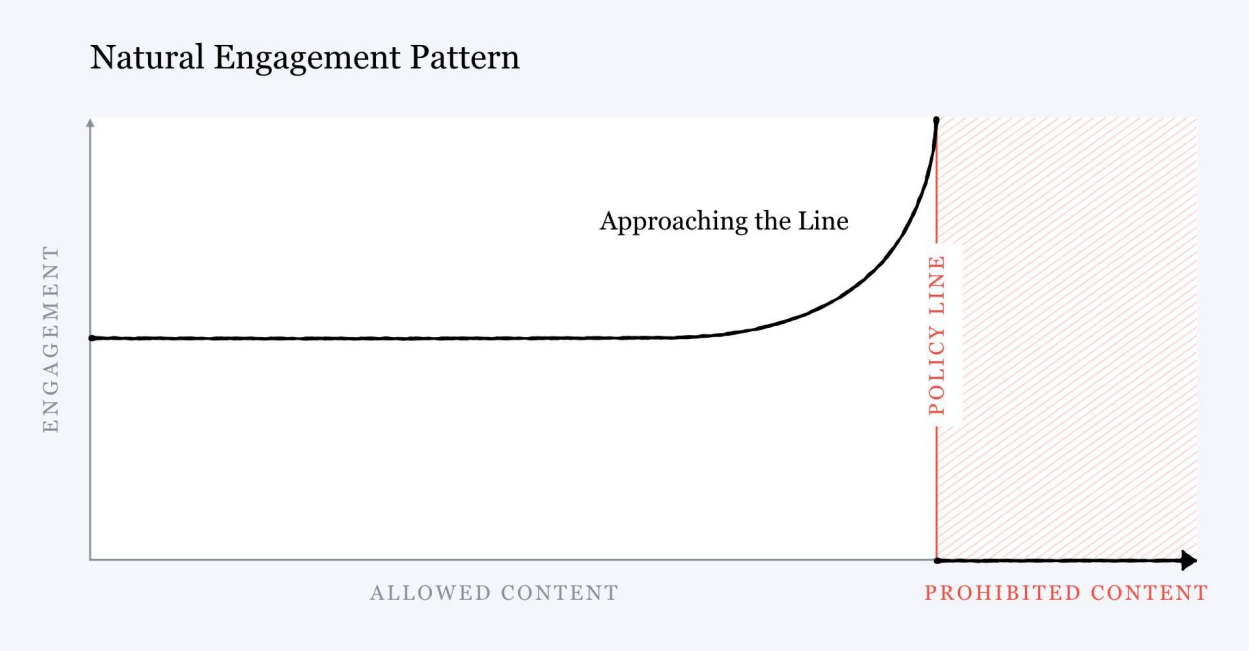
The Engagement Problem



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

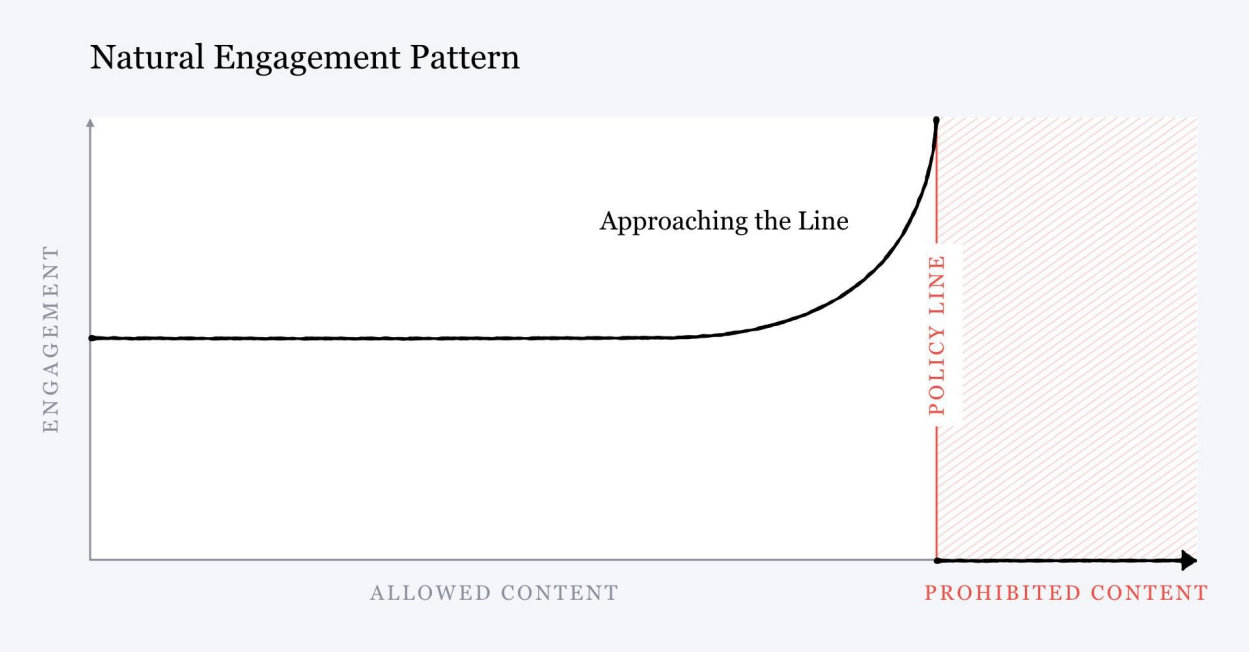
Source: Mark Zuckerberg, 2018, https://www.facebook.com/notes/751449002072082/

The Engagement Problem



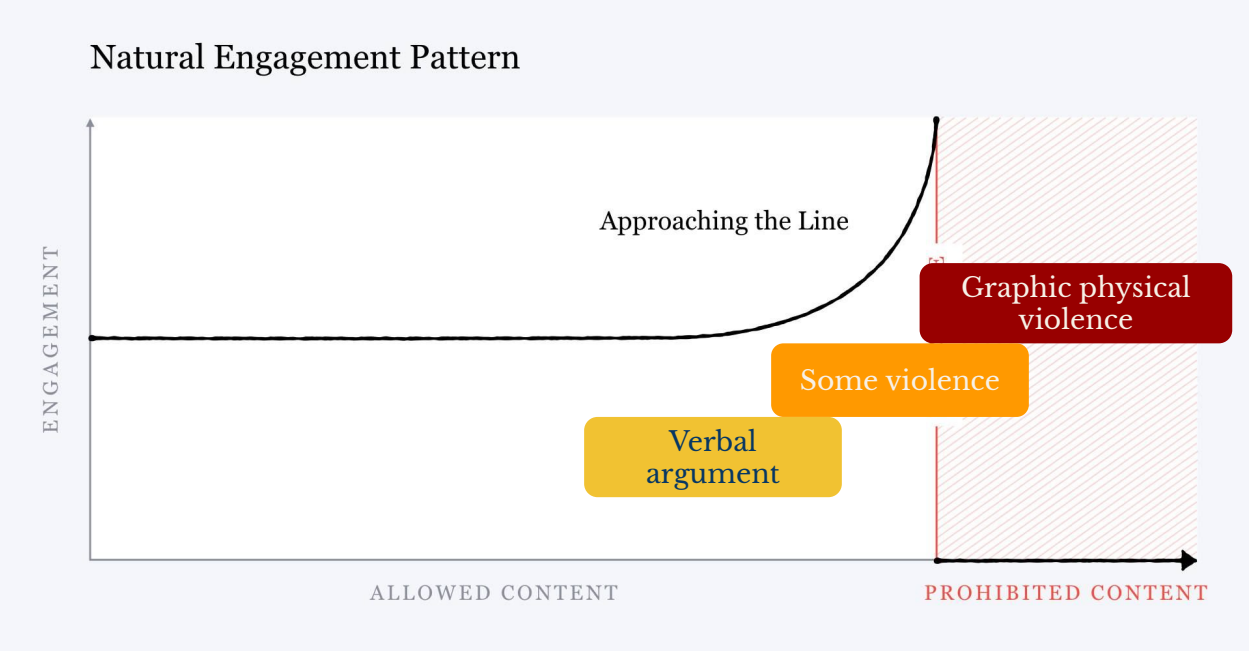
- X-Axis: What is allowed vs. prohibited?
  - Allowed content covers benign to borderline harmful
  - Prohibited content is harmful

The Engagement Problem



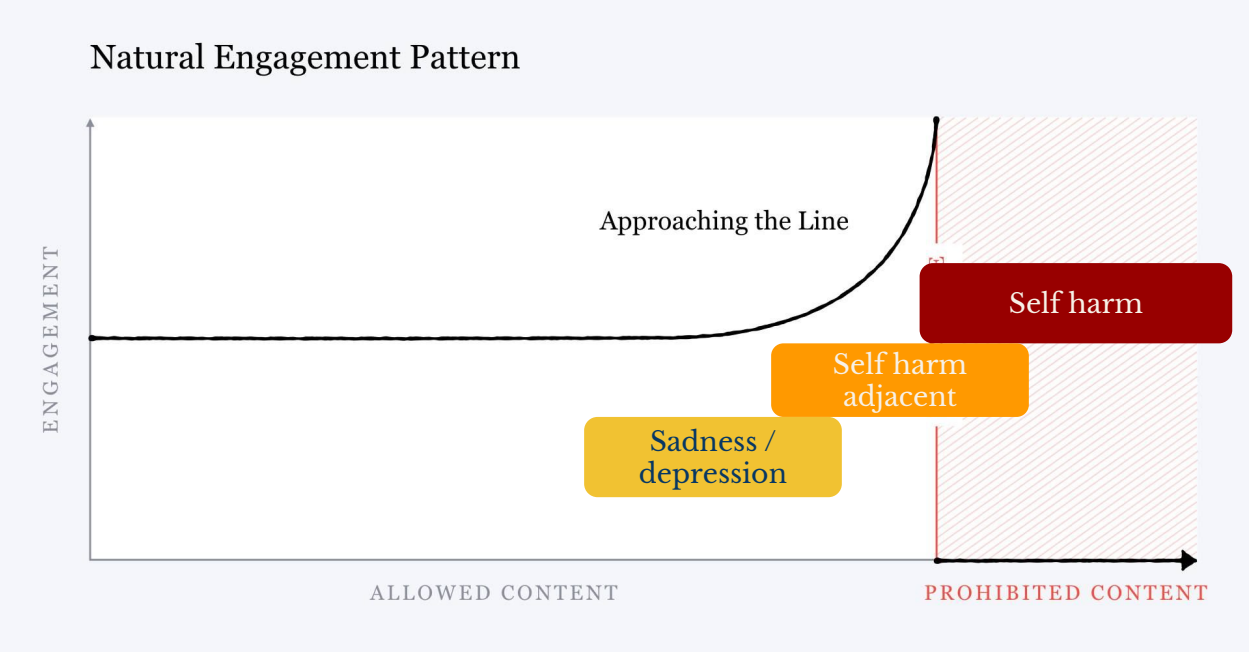
- This is true across many types of potential harms

The Engagement Problem



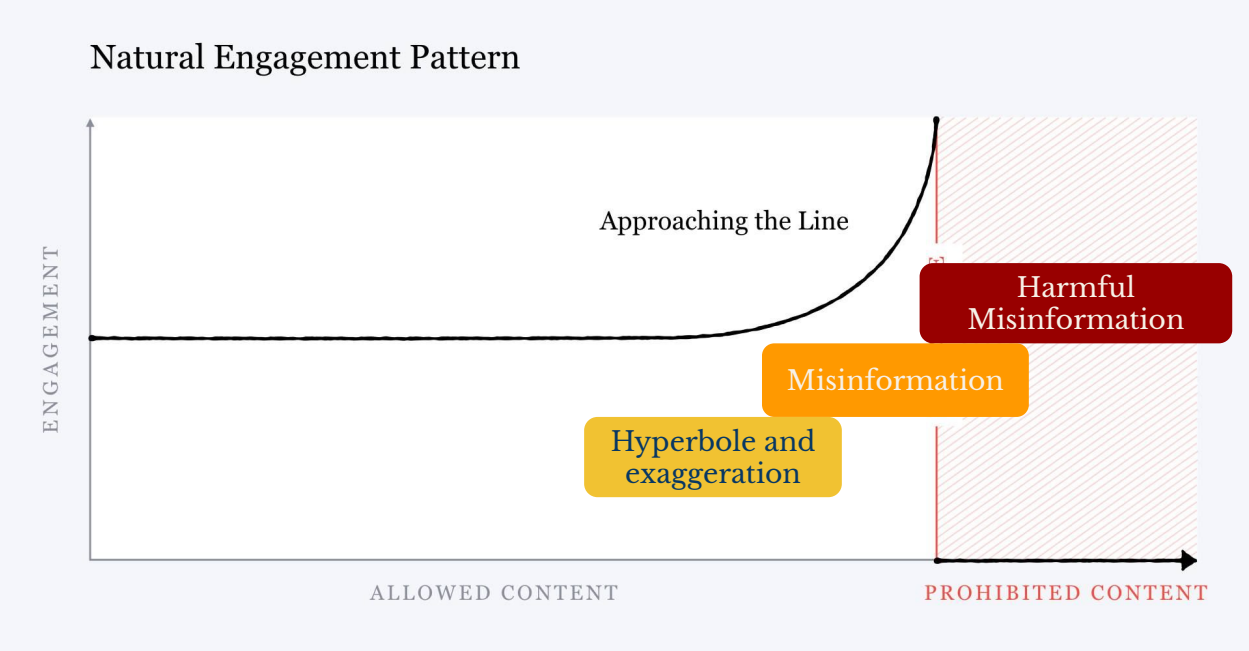
- This is true across many types of potential harms

The Engagement Problem



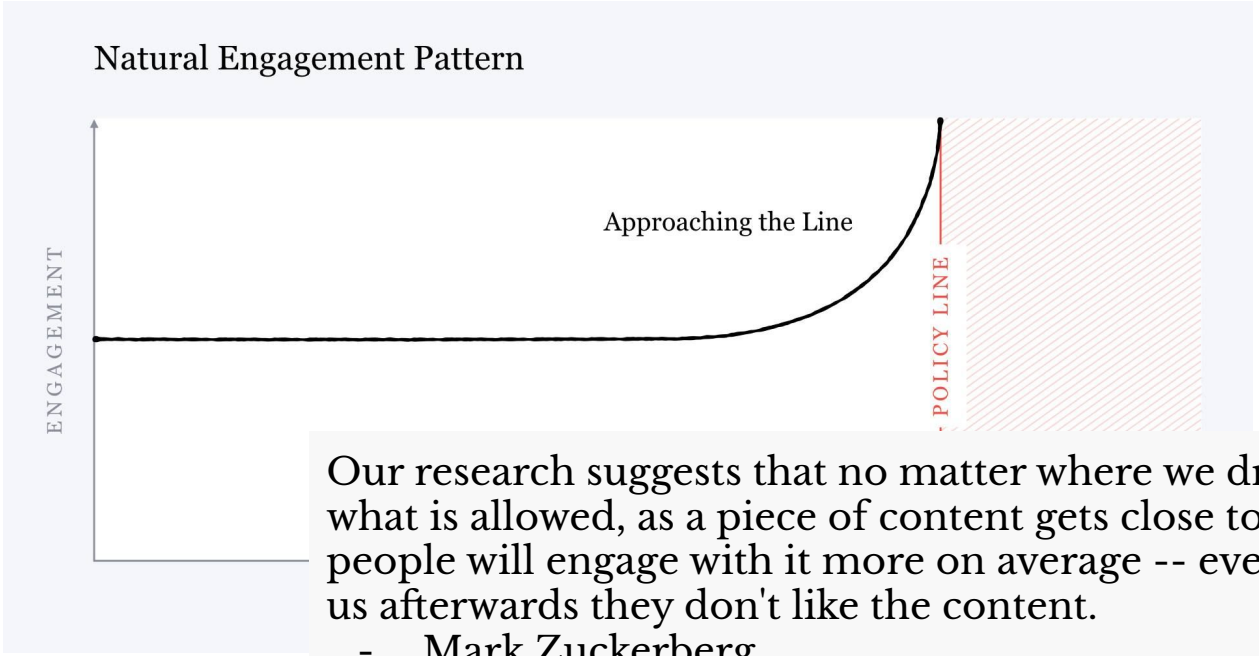
- This is true across many types of potential harms

The Engagement Problem



- This is true across many types of potential harms

The Engagement Problem

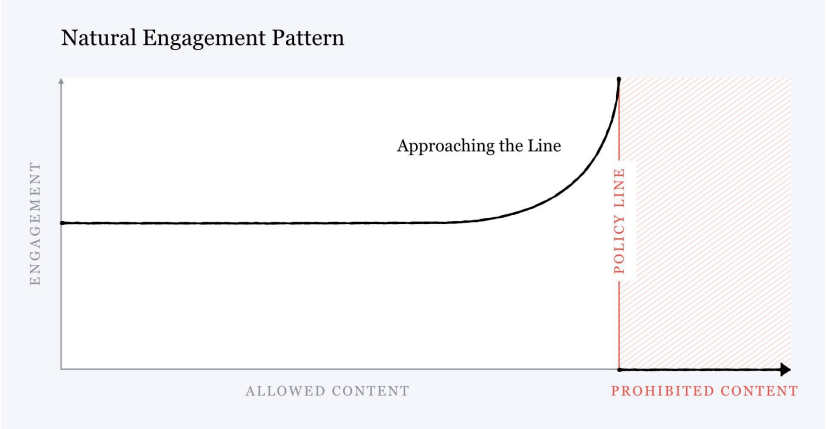


Our research suggests that no matter where we draw the lines for what is allowed, as a piece of content gets close to that line, people will engage with it more on average -- even when they tell us afterwards they don't like the content.

- Mark Zuckerberg

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

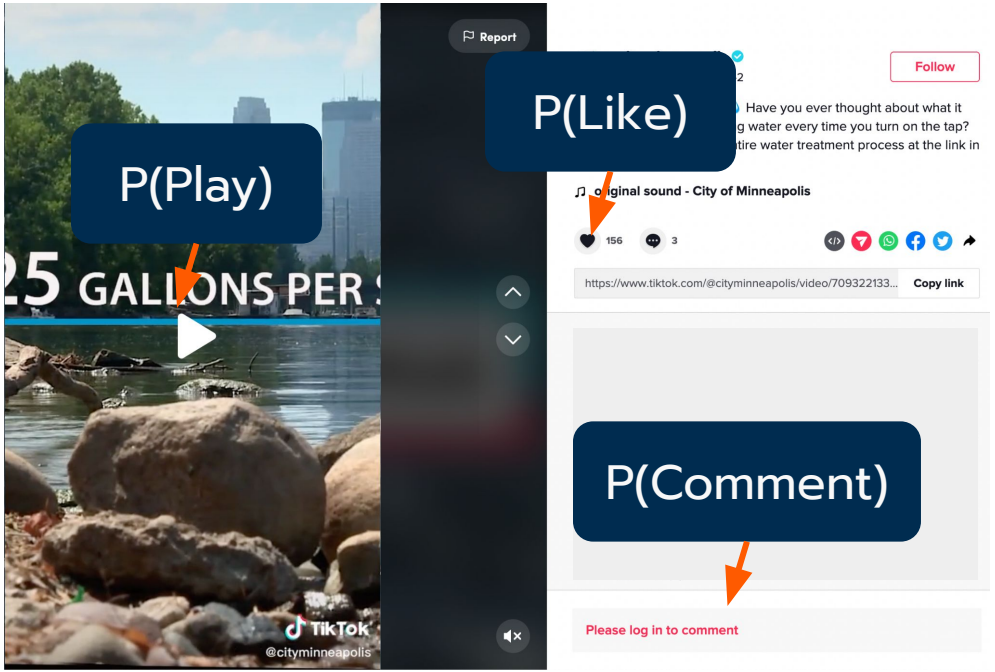


How Most Platforms Work

- 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



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

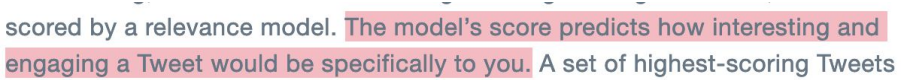
How Most Platforms Work

- Facebook
- Probability user will…
  - Like
  - Reaction
  - Comment
  - Reshare

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

How Most Platforms Work

- Twitter
  - “Interesting and engaging”
- YouTube
  - Clicks
  - Watch Time
  - Surveys



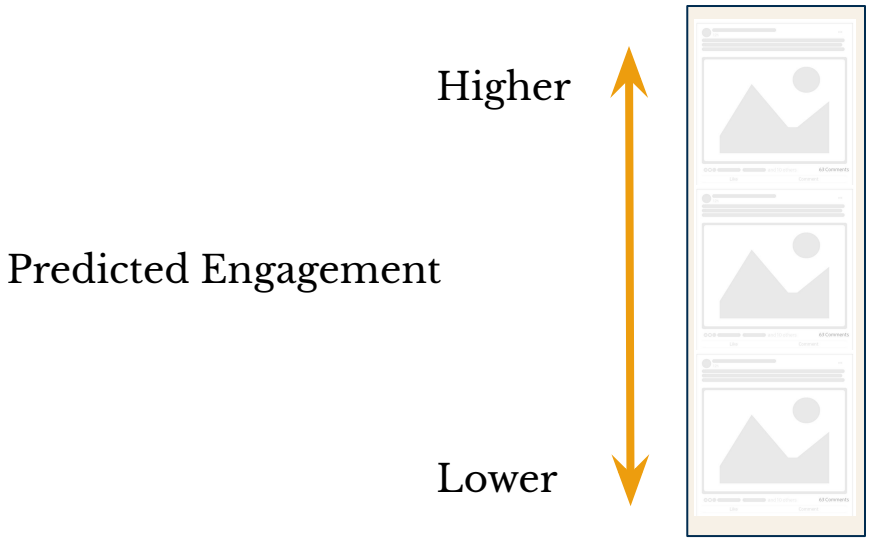
Source: Twitter, 2017, https://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-at-scale-in-twitters-timelines
YouTube, 2019, https://www.blog.google/around-the-globe/google-europe/fighting-disinformation-across-our-products/

How Most Platforms Work

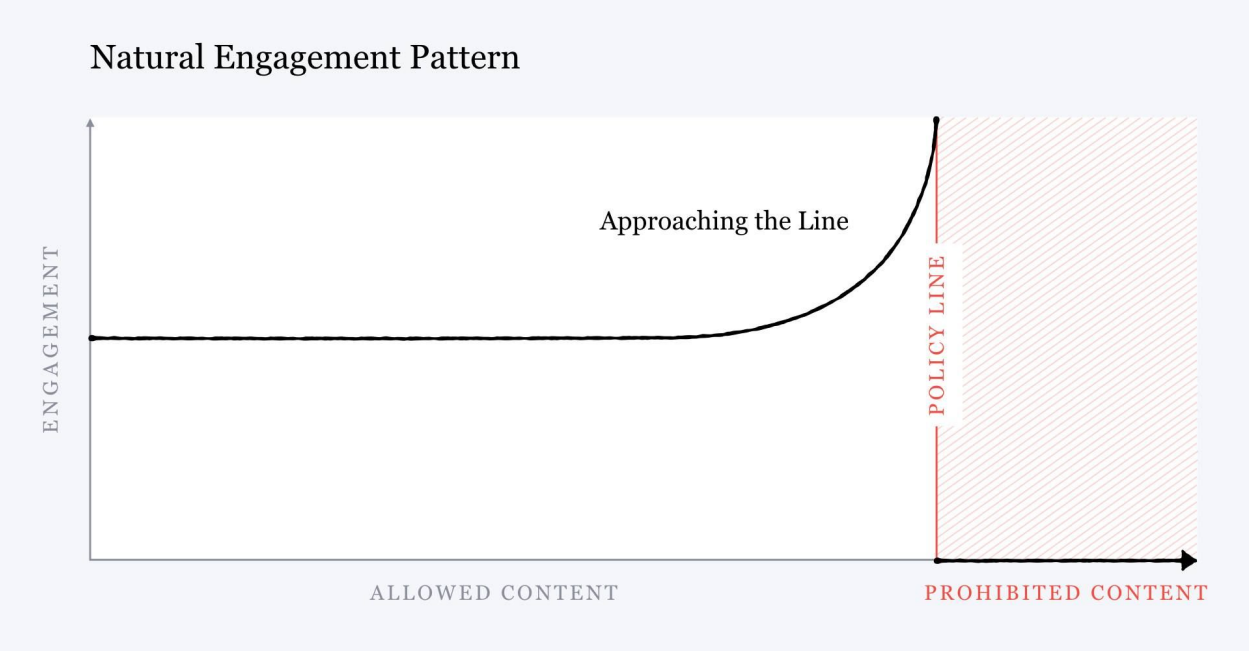
Facebook	**Predicted Engagement**: Like, Reaction, Comment, Share
Twitter	**Predicted Engagement**
TikTok	**Predicted Engagement**: Like, Comment, Watch
YouTube	**Predicted Engagement**: Clicks, Watch Time, Surveys

How Most Platforms Work

- Platforms recommend content and accounts most likely to be engaged with.
- Why does this matter? Back to Zuckerberg's chart

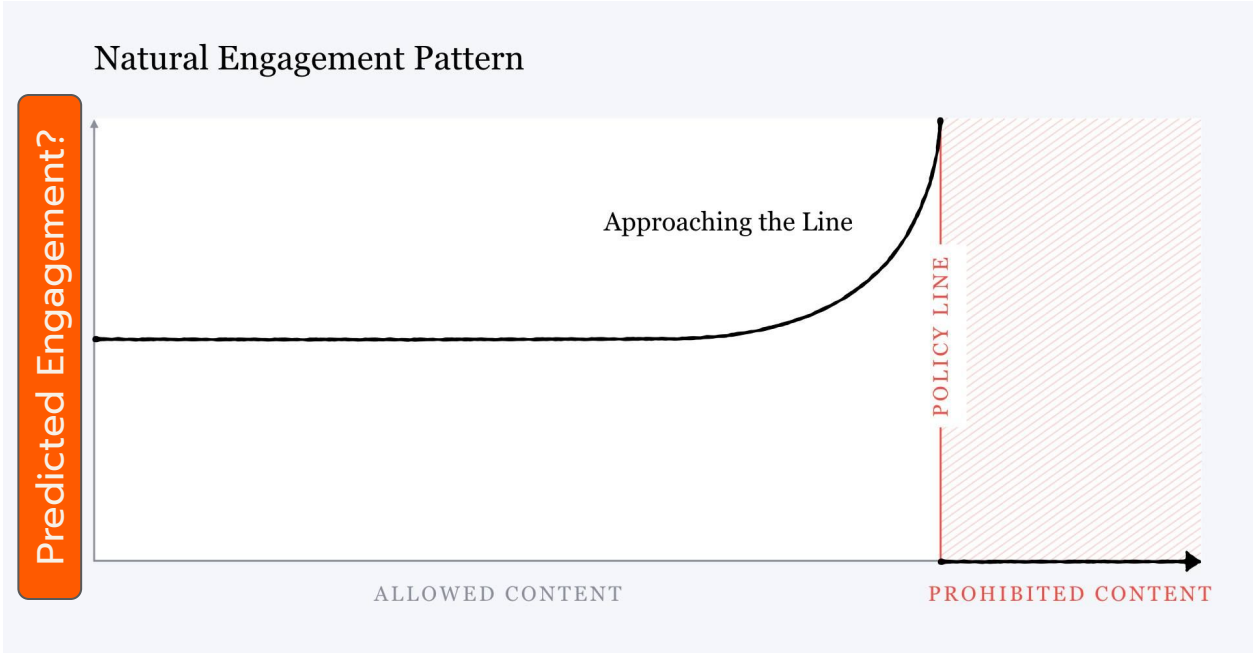


How Platform Design Can Amplify Harms



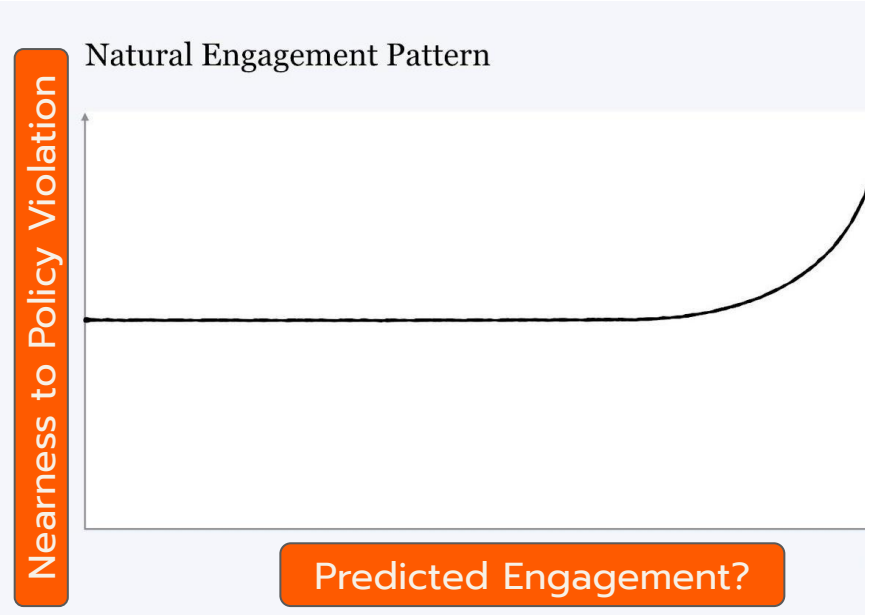
- More engagement, more likely to be harmful

How Platform Design Can Amplify Harms



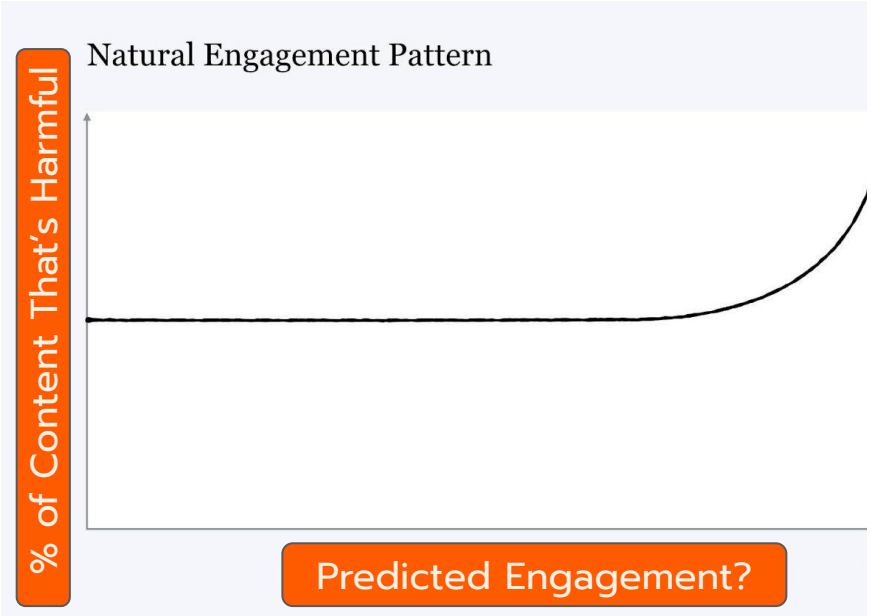
- Predicted engagement should follow actual engagement
- Content predicted to be engaging is more likely to be harmful

How Platform Design Can Amplify Harms



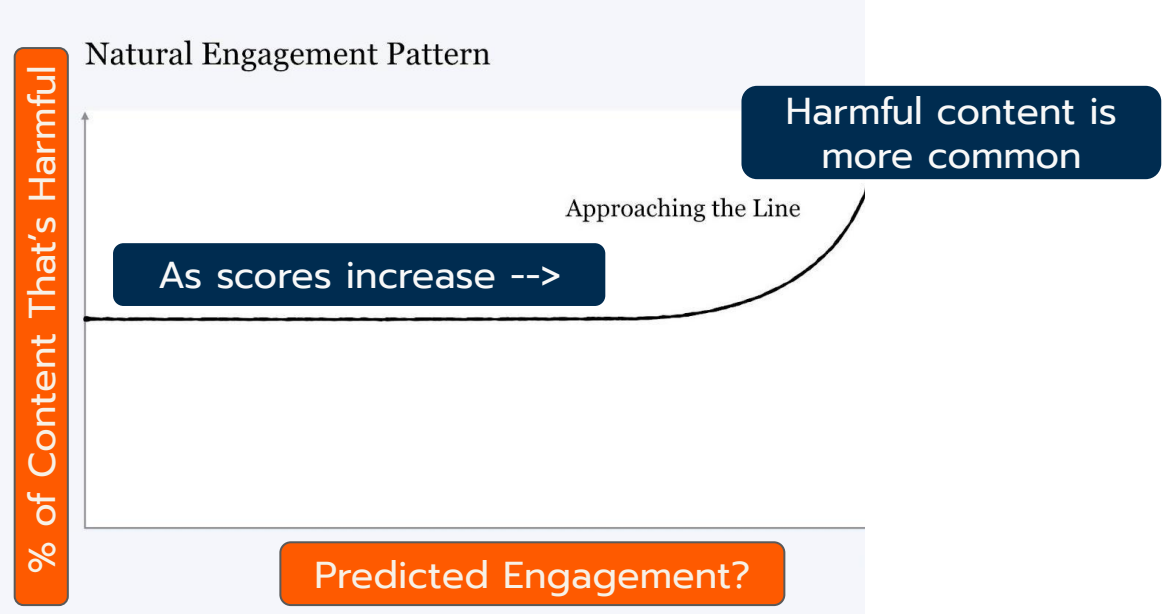
- Let's make it measurable
- Swap the X and Y Axes

How Platform Design Can Amplify Harms



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

How Platform Design Can Amplify Harms

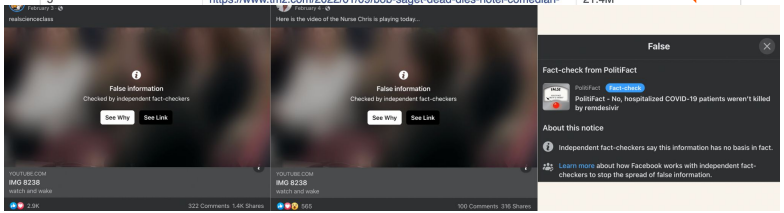


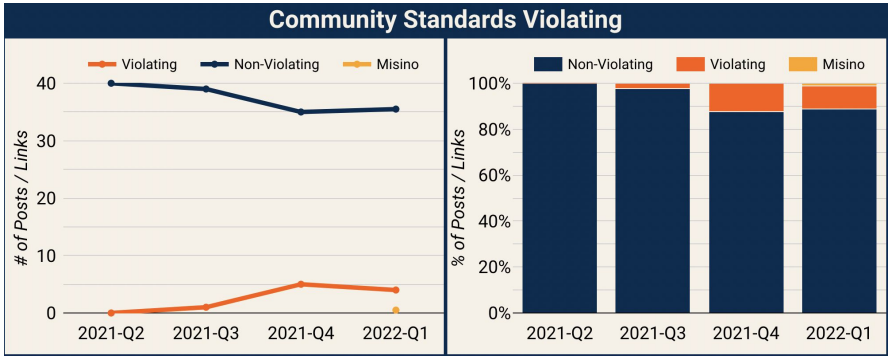
- Harmful content will tend to “float to the top” of the ranking systems
- This chart is measurable! Every platform could report it publicly

How Does This Problem Manifest?

- Viral, most seen content is more likely to be violating and harmful
- From Facebook's Widely Viewed Content Report
  - 12% of the most viewed content is violating or comes from violating accounts
  - #4 most viewed link (22M people) was a COVID conspiracy theory

Rank	Link	Content Viewers
1	**We blocked this link for violating our Inauthentic Behavior policy.** **Details:** The people behind this link's domain, nayenews24[.]info, used spam tactics to mislead people and drive them to their website. Content with links to this domain can no longer be created on Facebook.	31.6M
2	**We blocked this link for violating our Inauthentic Behavior policy.** **Details:** The people behind this link's domain, nayenews24[.]info, used spam tactics to mislead people and drive them to their website. Content with links to this domain can no longer be created on Facebook.	27.7M
3	https://www.youtube.com/watch?v=h3NhX6-5mO0	23.5M
4	https://www.youtube.com/watch?v=5SVO0lc_1_o	22.1M
5	https://www.tmz.com/2022/01/09/bob-saget-dead-dies-hotel-comedian-	21.4M

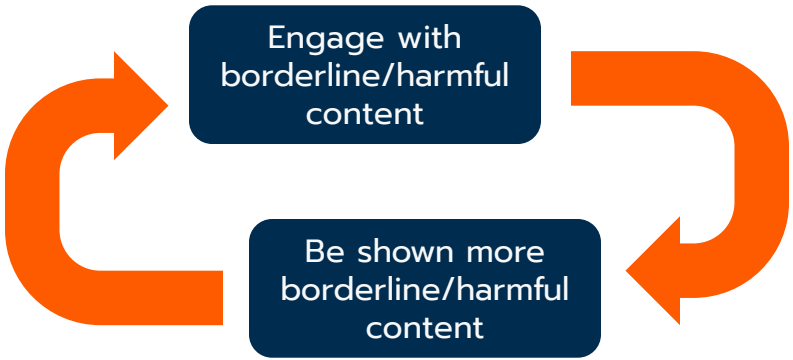


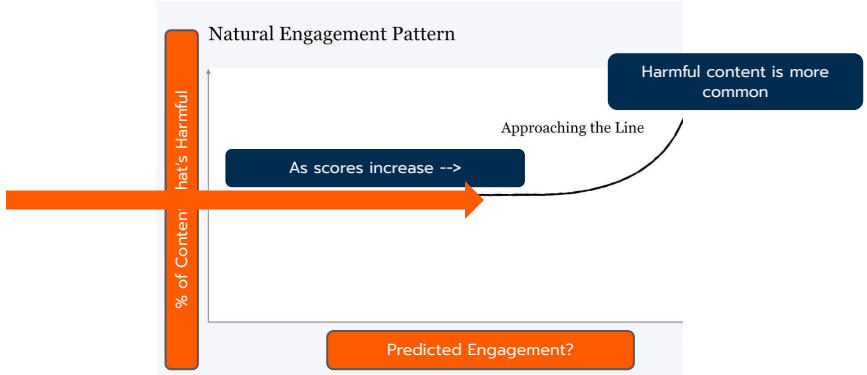


Source: Integrity Institute, 2022, https://integrityinstitute.org/widely-viewed-content-analysis-tracking-dashboard

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”





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

How Does This Problem Manifest?

- Example: Instagram hearing on child protections in Senate
- Adam Mosseri of IG: “Only 0.05% of impressions on IG on on self harm imagery”
- Senators: “When we create fake accounts that like self harm content, we see tons of recommendations around self harm”
- Who’s right? Both.
- If only a subset of users has a high prevalence, the overall average prevalence can still be low
    - 1% of users with 5% prevalence = 0.05% overall

Source: C-SPAN, 2021, https://www.c-span.org/video/?516470-1/senate-hearing-online-protections-children

How Does This Problem Manifest?

- This is inherently addictive design
- Does social media meet medical definition of addiction?
  - Maybe (Anecdotes)
- But this design maximizes any potential addictive nature

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

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.

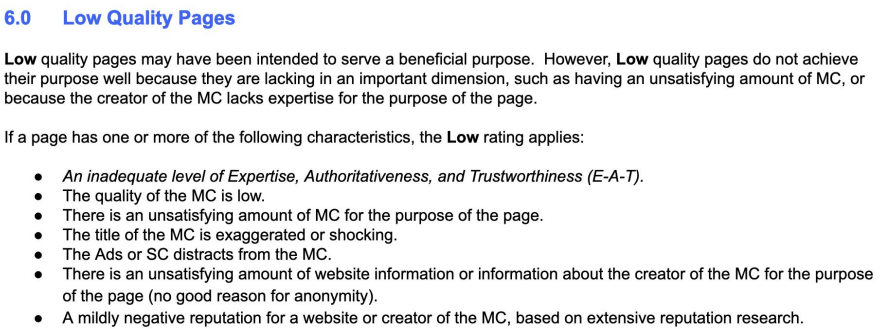
What makes a **High** quality page? A **High** quality page should have a beneficial purpose and achieve that purpose well. In addition, **High** quality pages have the following characteristics:

- *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?

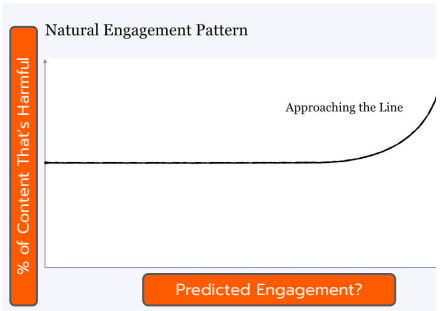
Source: Integrity Institute, 2021, https://integrityinstitute.org/s/Metrics-and-Transparency-Summary-EXTERNAL.pdf

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

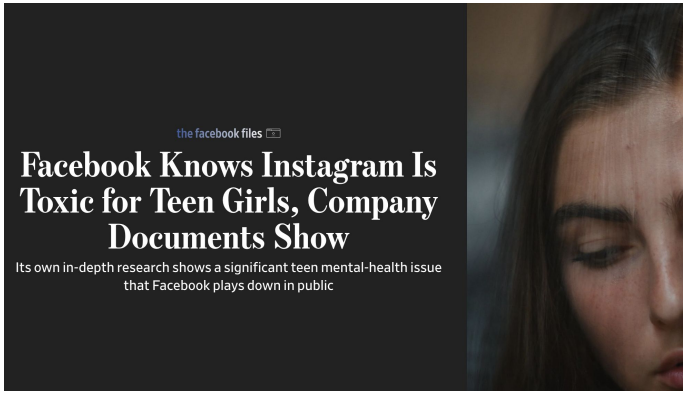
- Again, this is a huge topic
- We have briefing on “Ranking and Design Transparency”
- Key check: Is platform in the “engagement problem”
  - Using all engagement actions a user has taken
  - To predict all the future engagement actions a user might take
  - For the purposes of maximizing engagement 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



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

Conclusion

- The “Engagement problem”
  - Most platforms use it
  - Can exploit cognitive biases
  - Amplify harmful content
    - Medical misinfo, depression, self harm, eating disorder
  - Amplify any addictive nature
- 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