DRAFT

Ranking and Design Transparency

Data, Datasets, and Reports to track Responsible Algorithmic and Platform Design

2021-09-28 (v20211026)

Integrity Institute

Goal of this Presentation

DRAFT

- Clearly explain a set of transparency requirements that will
  - Enable the public to understand how the platforms are designed and how their ranking algorithms work
  - Enable the public to validate that the companies are using responsible, best design practices
- This will require a baseline understanding of how the ranking algorithms and platforms are designed

OUTLINE

- Goal of Transparency
- Ranking Algorithm Basics
- Examples of Harmful Design
- Best Practices
- Algorithmic Transparency
  - Data used
  - Machine Learning Models
  - Goals and Objectives

High Level Takeaways

DRAFT

- Ranking systems have multiple components
  - Platforms should release detailed descriptions of key components
- There are best practices is designing ranking systems to minimize risk of amplifying harm
  - Platforms should follow these best practices
  - Platforms should make public how they do so

What is the Integrity Institute?

DRAFT

- 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.

Goal of Transparency

DRAFT

- Ranking and design are where the mission and values of a social media company become encoded into the platform
- Ranking and design choices play a significant role in exposing users to harmful content
- The public needs transparency to see that the platforms are behaving responsibly here

Ranking Basics

DRAFT

- 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

DRAFT

Inventory

Ranking Basics

DRAFT

- 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

DRAFT

Inventory

Features

X, Y, Z

Ranking Basics

DRAFT

- 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

DRAFT

Inventory

Features

X, Y, Z

ML Models

Like?
Comment?
Retweet?

Ranking Basics

DRAFT

Inventory | **Features** | **ML Models**

- 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

DRAFT

Inventory

Features

X, Y, Z

ML Models

Like?
Comment?
Retweet?

Ranking

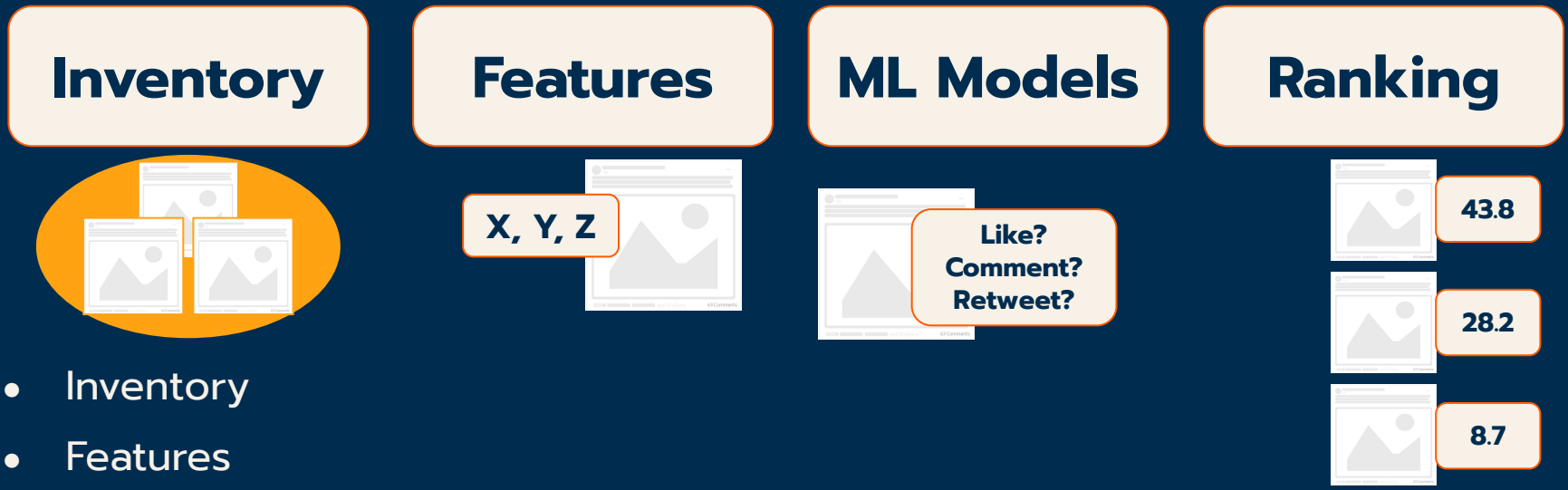
43.8

28.2

8.7

Ranking Basics

DRAFT



- Inventory
- Features
- ML Model Scoring
- Final Ranking Score
  - All the classifier scores are combined, business logic applied
  - Final sorting and list generated

Standard Design

DRAFT

- 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

DRAFT

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

Standard Design

DRAFT

- There are two basic frameworks or paradigms used in ranking systems
  - User Engagement focused ranking
  - Quality focused ranking

Standard Design: Engagement Ranking

DRAFT

- 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

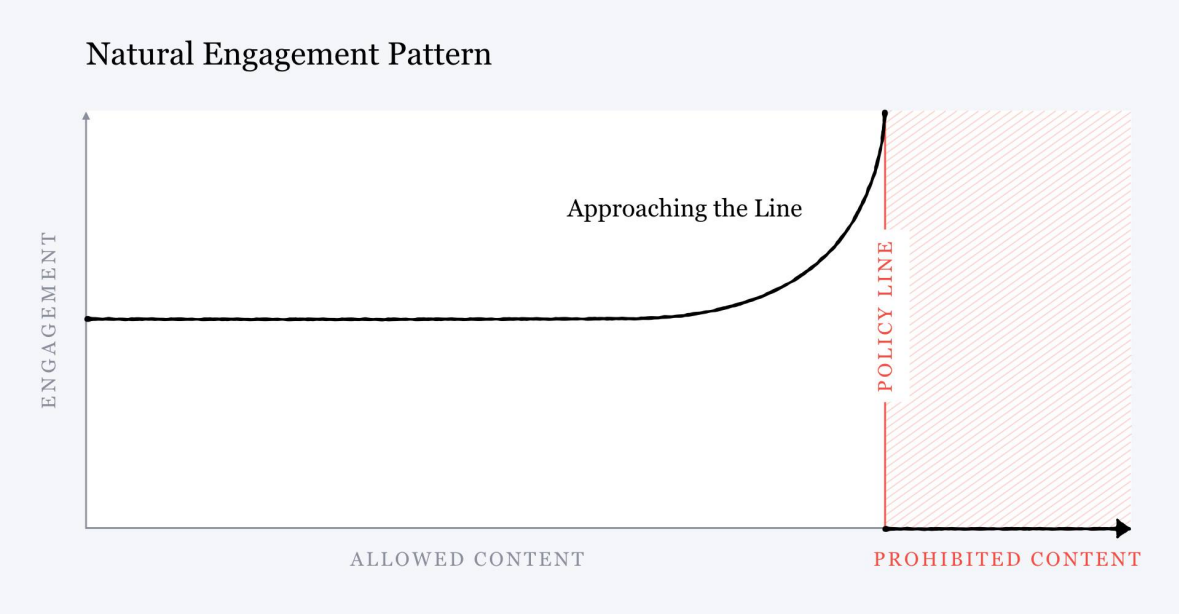
DRAFT

- 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

Harmful Algorithmic Design

DRAFT

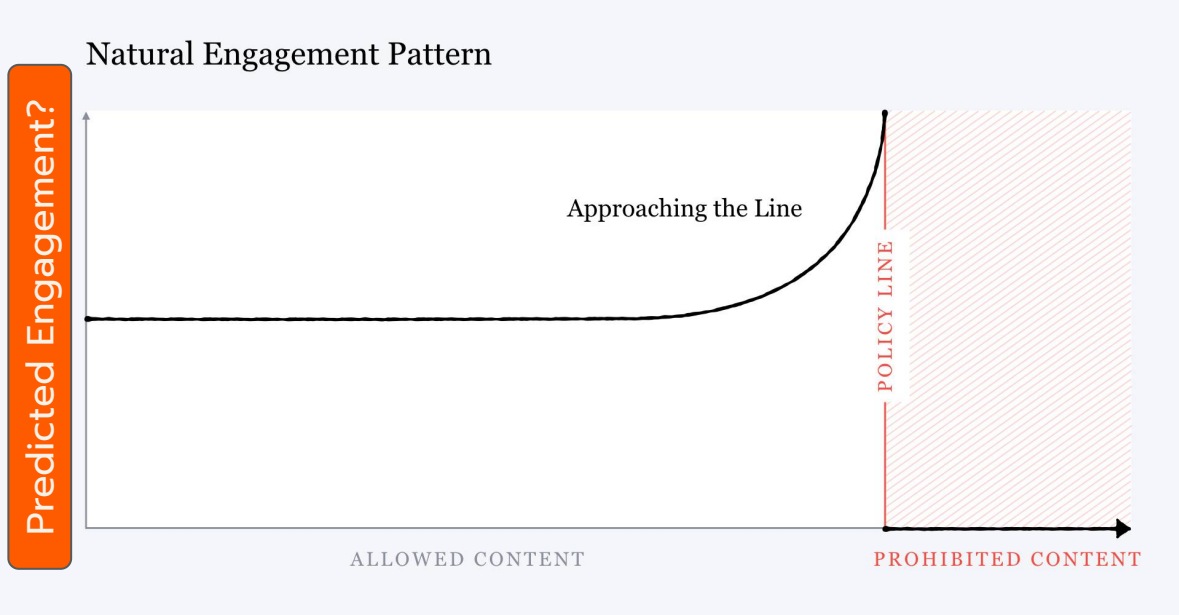
- Algorithms can be designed in a way that promotes harmful content
- Content that is more likely to violate policies (be harmful) will get more engagement [Source: https://www.facebook.com/notes/751449002072082/]
- This chart is from a public note by Mark Zuckerberg



Harmful Algorithmic Design

DRAFT

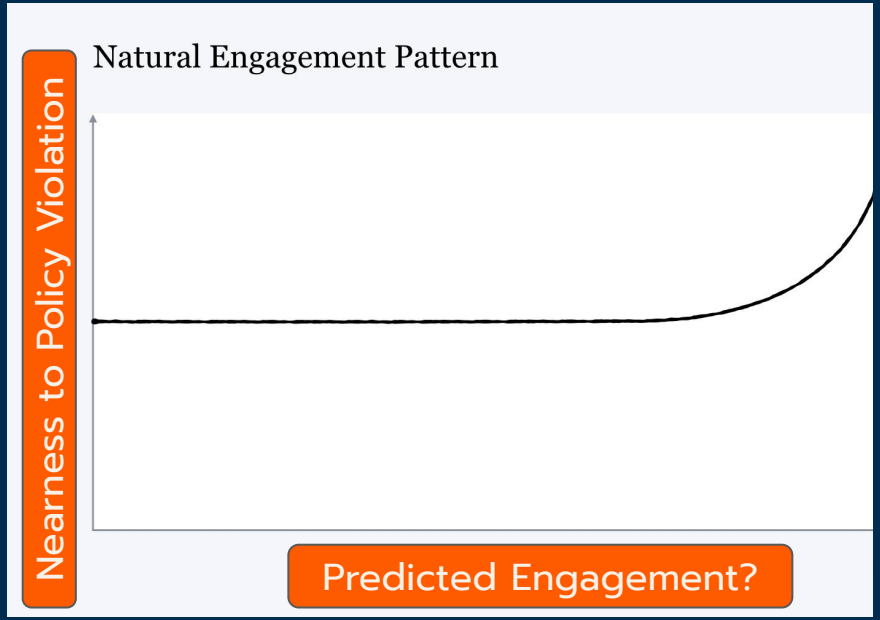
- Engagement focused platforms use ML models to predict if users will engage with content
- Would expect predicted engagement behaves the same as true engagement



Harmful Algorithmic Design

DRAFT

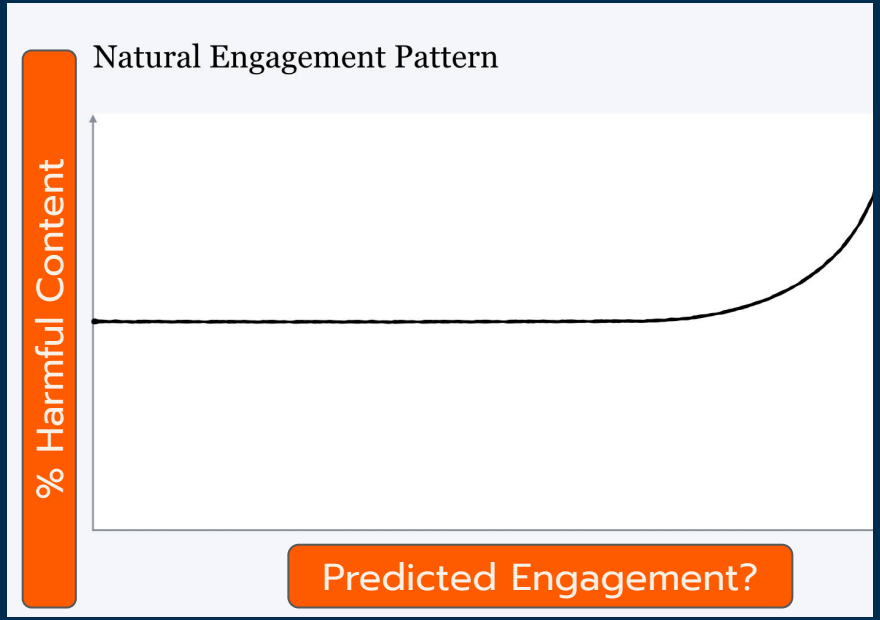
- Predicted engagement is the easier thing to track. So we can flip the axes so it is the X axis.
- Still expect an “Up and to the right” shape



Harmful Algorithmic Design

DRAFT

- The “Nearness to a policy violation” is not a “real” thing that can be measured.
- But, could ask “What % of the content is harmful?” as a function of the predicted engagement.

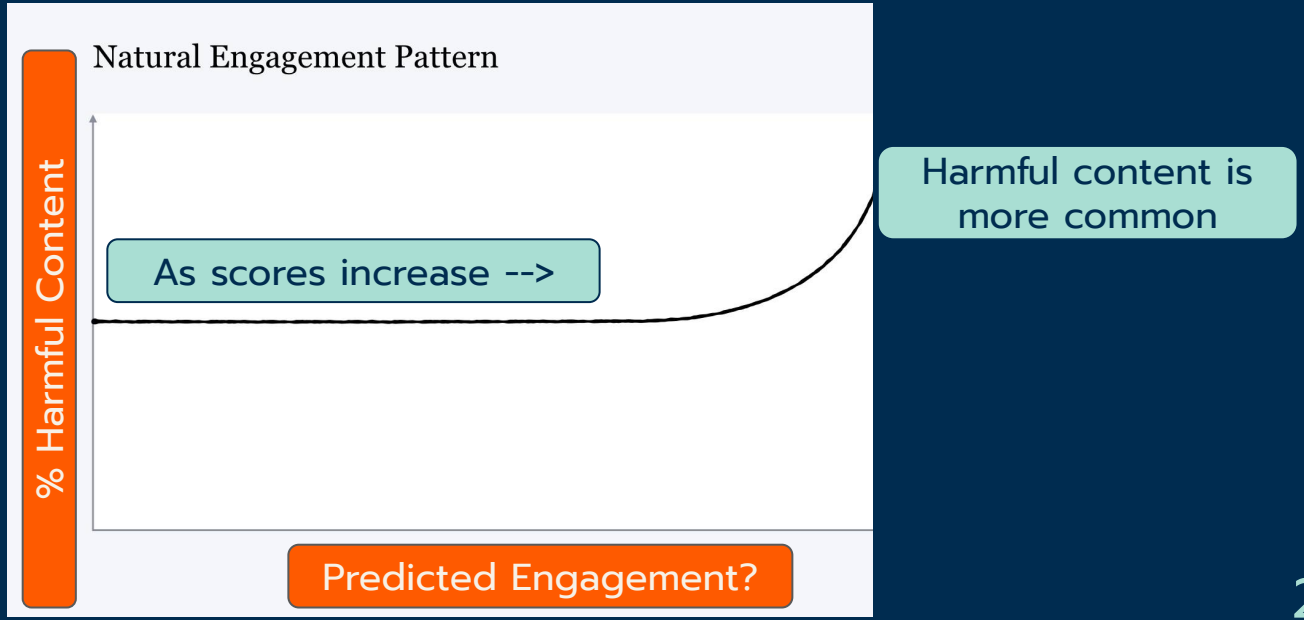


Harmful Algorithmic Design

DRAFT

So, from Zuckerberg's original chart, the natural expectation should be that harmful content will "float" to the top of lists that are ranked by predictions of engagement.

This is a chart that every platform could measure and release today.



Harmful Algorithmic Design

DRAFT

- We should have an expectation that engagement focused ranking systems will amplify harmful content
- This is an example of why the public needs transparency around ranking algorithms

This concludes the necessary background.

We will now share what reflects the consensus view of our community of Integrity Professionals.

Best Practices

DRAFT

- Platforms should establish a set of top line and A/B testing quality metrics for ranking that reflect company values and/or the company mission
  - These metrics should be independent of user engagements, growth of the platform, and user surveys about content
  - These could be used to help promote or nurture content that is good, which reduces the need to worry about demoting “bad” content
- Platforms should release the definition or description of the quality metrics
- These metrics should be used in addition to any other the companies want to use to evaluate ranking systems

Best Practices

DRAFT

- Platforms should expect various actors to try and exploit or game their ranking systems as well as novel forms of harmful content
  - Make known attack methods more costly
  - Consider rate-limiting or otherwise imposing costs on user actions in proportion to their role in abuse
    - Limiting new users ability to reach large audiences
    - Limit posting same or similar content across many spaces on the platform
    - Limit features from using multiple accounts or using anonymous accounts
  - Run every new feature through integrity modeling from the beginning and build with those concerns
    - An example of a feature to treat carefully are re-share (Re-tweet, re-blog) features
    - These are frequently used in platform abuse
- Platforms should release a summary of how they prevent bad actors from exploiting ranking systems

Best Practices

DRAFT

- Platforms should make integrity a key component in how changes to ranking systems are made
  - Elevate integrity metrics to highest priority and evaluate every ranking change against them
  - Use long running holdouts and other processes to track aggregate changes over time
  - Disclose if they have different processes for ranking different content topics, or how content topic classifiers impact ranking
  - Release their protocols for ranking changes during special events (Elections)
- Platforms should release an outline of their processes for determining product or ranking changes
  - With special attention given to any differences between features intended to increase integrity and features intended to increase business goals

Algorithmic Transparency

DRAFT

- Algorithmic transparency can be tied to the basic components of ranking systems

Algorithmic Transparency

DRAFT

- Platforms are intrinsically transparent on inventory
- By using the platform, it is possible to identify sources the platform uses for content in feeds

Algorithmic Transparency

DRAFT

- Platforms should release a list of the most important features in ranking
  - Should include an estimation of how important the features are in the models
  - Special attention should be given to any features that use user data
- Releasing these lists would not harm competitive advantage or proprietary information
  - In engagement focused systems, most important features are trivial
    - "How engaging is this post overall?" "Has the user engaged with this content producer previously?"
  - More general descriptions can be given for non-trivial features

Algorithmic Transparency

DRAFT

- Platforms should release a list of models which contribute to positive scores in the ranking system
  - Special attention should be given to any models predicting user actions
  - Ideally all or most models, but if there are many, give most important
- Releasing these lists would not harm competitive advantage or proprietary information
  - In engagement focused systems, these are again trivial
  - More general descriptions can be given for non-trivial features

Algorithmic Transparency

DRAFT

- Platforms should release how the final scoring calculation is computed
- There are ways of releasing this that respect proprietary information

Algorithmic Transparency

DRAFT

Top Line Company Metrics, Goals, Expectations

- Platforms should release their top line objectives for the ranking system and its specific definition
- Platforms should release the general process for making changes

Conclusion

DRAFT

- Ranking and design are where the mission and values of a social media company become encoded into the platform
- Ranking and design choices play a significant role in exposing users to harmful content
- The public needs transparency to see that the platforms are behaving responsibly here
  - Ranking systems have multiple components
    - Platforms should release detailed descriptions of key components
  - There are best practices is designing ranking systems to minimize risk of amplifying harm
    - Platforms should follow these best practices
    - Platforms should make public how they do so