Metrics & Transparency

Data and Datasets to Track Harms, Design, and Process on Social Media Platforms

2021-09-22

Integrity Institute

Goal of this Presentation

- Clearly explain a comprehensive set of transparency requirements that will
  - Enable the public to understand the scale and cause of harms occurring on social media platforms
  - Enable to public to validate that social media companies are using best practices in responsibly designing and building their platforms
- This will require a baseline understanding of how content is distributed on platforms, which we will provide

OUTLINE

- Goal of Transparency
- Lifecycle of Harmful Content
- Summary Data
- Content Datasets
- Algorithmic Transparency
- Process Transparency

High Level Takeaways

- Transparency is a crucial tool towards improving platforms
- Making progress will require *significant* requirements of platforms to be comprehensive
- The public needs metrics, data, and reports to illuminate the following aspects of the platforms
    - The scope and mechanisms of harmful content distribution
    - The role that algorithms and design choices made by the platforms play in that distribution
    - The processes within the company for making product changes that may impact that distribution
    - In addition, the public needs enough information such that all claims can be confirmed and validated by the public and auditors
- These requirements will be met by
    - Data based reports
    - Public data sets
    - Comprehensive algorithmic transparency
    - Comprehensive transparency into their company processes

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 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 have not launched publicly. Please treat our work with discretion.

Goal of Transparency

- The public needs transparency to fully understand
  - What is the societal scale of harm?
  - How, why, and to whom do harms occur?
  - Are the platforms following best practices to minimize harms?
  - How can we track systemic issues and progress rather than anecdotes?

Goal of Transparency

- Increased transparency will
  - Help the public make informed decisions about the platforms
  - Incentivise to the platforms to follow best practices
  - Elevate the influence of teams trying to minimize harms within the company

Lifecycle of Harmful Content

This is a piece of harmful content.

It contains: Misinfo, Hate Speech, Self Harm, etc

Lifecycle of Harmful Content: Creation

- It was uploaded to the platform by a user, account, channel, publisher/business
- These actors have different motivations
  - Share messages, express beliefs, or financial profit
- This user/account/publisher may have a history of this harm

User/Account

Publisher

History?

Lifecycle of Harmful Content: Creation

- They distributed it
  - Publicly
  - Privately to Followers
  - Privately to a Private Group
  - Via an ad
  - In a direct message to another user

User/Account

Publisher

History?

Public

Private

Ad

Message

Lifecycle of Harmful Content: Exposure

This is a harmful exposure.

A user saw the harmful content

Lifecycle of Harmful Content: Exposure

- The exposure happened on a “surface”, a feed or other part of the app that shows content
  - Algorithmically ranked feed
  - A non-algorithmically ranked surface
    - User visited the accounts timeline, or directly visited the content
- The user may have had a connection to the account
- The user may have a history of being exposed
- The user may be part of a demographic that is particularly vulnerable

Algo Feed?

Followed?

History?

Demographic?

Lifecycle of Harmful Content: Exposure

- If the content is flagged (by users or algo), then after being on the platform for a period of time, and seen by some users, it can be moderated
- Moderation can include removal, labeling or screening, downranking, or requiring user to remove it

Algo Feed?

Followed?

History?

Demographic?

Time to Moderation

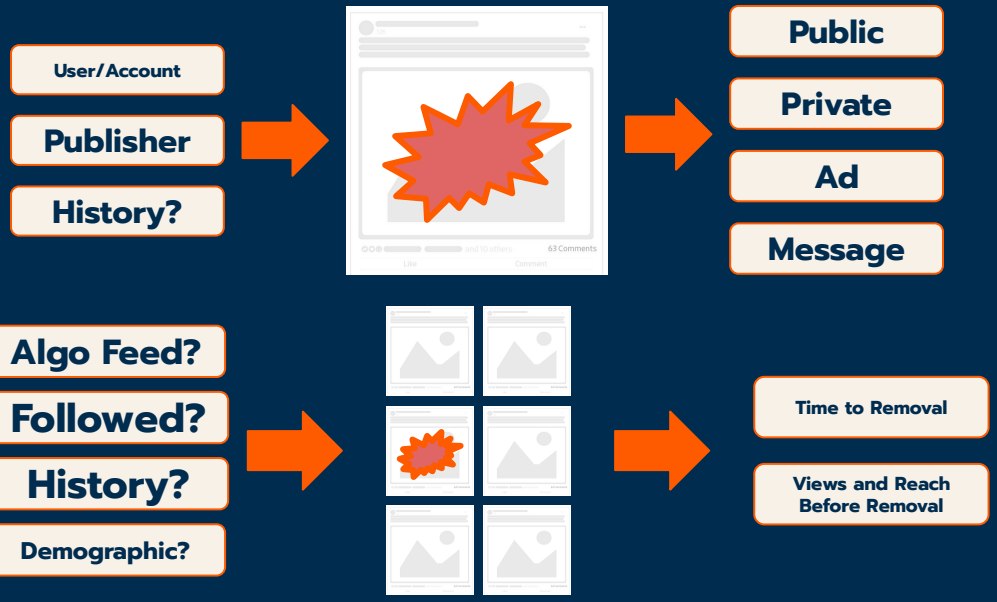
Views and Reach Before Moderation

*Reach: The number of users who view the content

Lifecycle of Harmful Content

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

Top Line Company Processes and Goals



Lifecycle of Harmful Content

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

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

Metrics

- Transparency can come in the form of
  - Aggregated and summary data
    - Example: Number of users exposed to harmful content
  - Content based datasets
    - Example: Top 10k most viewed pieces of content
  - Details and descriptions of algorithms and design
    - Example: Classifiers used in ranking feed content
  - Details and descriptions of process
    - Example: Metrics companies use in experimentation
- Feel free to skip to any section of particular interest

Aggregated and Summary Data

- Several platforms release “prevalence” measures
  - Prevalence: What % of content views are harmful exposures?
  - YouTube and Facebook reports
- These are inadequate
  - What is the societal scale?
  - Why are these harmful exposures happening?
- The goal of summary data is to make “lifecycle of harmful content” fully transparent

Aggregated and Summary Data

- **The consensus view of Integrity Professionals is that platforms release**
  - The prevalence of harmful exposures
  - The reach of harmful content over 7, 30, and 90 day windows
  - The distribution of frequency of exposures for users
    - How many users had 0, 1, 2, 3, 4, 5+ harmful exposures?
  - The frequency of underlying causes of harmful exposures
    - What % take place in algorithmic feeds?
    - What % are from creators the user follows?
    - What % of those follows were from a recommendation?
    - What % of harmful exposures are on public content?
    - What % of harmful exposures are from creators who have previous offenses?

Aggregated and Summary Data

- Additional data to release could include
  - The time delay between harmful content being posted and moderated
  - The average views and reach of harmful content before it is moderated
  - Basic demographic statistics on the viewers of harmful content
    - Age, Region
    - Platforms should work with experts and the user communities to determine when additional details on people exposed to harmful content should be included
      - When to include protected classes
  - Additional data could be provided on the targets of harmful content, rather than the viewers
    - Such as what details to provide on targets of harassment and bullying

Content Based Datasets

- The goal of content-based datasets is
  - Allow public to validate official data reports
  - Raise awareness of new harms and issues
- Platforms already make datasets like this available
  - Facebook's Widely Viewed Content report
  - Twitter's streaming data
  - Reddit API
- In addition, many companies make this data available at a cost
  - Tubular Insights for YouTube
  - NewsWhip for Facebook
  - BuzzSumo for Facebook, Twitter, YouTube

Content Based Datasets

- **The consensus view of Integrity Professionals is that platforms release**
  - The Top N pieces of public content
    - N should be at least 10,000
    - Released on a weekly basis
    - Data should include all public data with the content
  - A random sample of N impressions on public content
    - N should be at least 10,000
    - Released on a weekly bases
    - Data should include all public post content
    - Data should include key ML model scores for the content
      - Specifically any engagement predictions

Algorithmic & Design Transparency

- The goal of algorithmic and design transparency is to ensure platforms are following best practices to mitigate harms
- Transparency here is a complicated subject, and we have a special deck dedicated
- Please see the “Algorithmic and Design Transparency” deck for more details on algorithmic design
    - https://docs.google.com/presentation/d/1hPkgBR0b3iN1eqdISRRn80uv5yyuvGI-0KiYamkFR20/
- Summary provided here

Algorithmic & Design Transparency

- **The consensus view of Integrity Professionals is that platforms release**
  - The Top N most important features in the ranking system
    - N should be > 10
    - “Importance” of the features should be assessed using accepted practices for the model design
  - A list of ML models and what they try to predict
    - Special attention to any ML models that involve predicting user actions
  - The top-line objectives for the ranking systems and their specific definitions
  - Full disclosure if there are any different ranking processes for content topics and how content classifiers impact ranking
  - How they prevent bad actors from “gaming” the ranking systems and make the system adversarially robust

Process Transparency

- The platforms operations are largely opaque to the outside
  - How do they make decisions?
  - What metrics do they use to evaluate their platforms?
  - How do integrity concerns get weighed against top-line company goals?
- Transparency here applies pressure on the platforms to build responsibly

Process Transparency

- The consensus view of Integrity Professionals is that platforms release
  - Core metrics that are used in experimentation processes and their exact definitions
  - An outline of their processes for determining product or ranking changes
    - Specifically for features to reduce harms on the platform
    - And for “normal” features
  - Their process for platform changes around significant events (Elections)
  - Staffing levels on integrity and trust and safety teams
  - Platforms should release how they assess content quality
    - Specifically any quality assessments related to integrity
    - Should include positive definitions of content quality as well as negative

Conclusion

- Transparency is a crucial tool towards improving platforms
- The public needs transparency from platforms to illuminate
  - The scope and mechanisms of harmful content distribution
  - The role that algorithms and design choices made by the platforms play in that distribution
  - The processes within the company for making product changes that may impact that distribution
  - In addition, the public needs enough information such that all claims can be confirmed and validated by the public and auditors
- To do that, we need the platforms to
  - Release data based reports
  - Release public data sets
  - Provide comprehensive algorithmic transparency
  - Provide comprehensive transparency into their company processes
- This is a significant requirement of platforms, but is possible, and if done comprehensively will have a meaningful impact on the incentives of the platforms