Meaningful Platform Transparency

Transparency to Inform Society & Prepare for Upcoming Regulations

Jeff Allen, Integrity Institute, 2023-07-11

Outline

- What is meaningful transparency?
- What are the current and upcoming regulations?
- What is risky content?
- Transparency Frameworks

What is the Integrity Institute?

The Integrity Institute is a **growing community of 250+ tech workers** who believe in building integrity-first online platforms that help individuals, communities, and democracies thrive.

With years of experience mitigating harms to people and communities within **55+ online platform companies**, we bring seasoned, insider knowledge to leaders theorizing, building, and governing online platforms and help them put integrity front and center.

Here's how we do it:

- We build and empower a community of integrity professionals in tech, giving them the tools and research they need to make online platforms safer and healthier for people and societies
- We advise online platforms, policymakers, and academics to put integrity at the heart of company governance, compliance, and tech regulation.
- We educate the public about what an integrity-first future looks like for the social internet.

Talk to us!

This deck is significantly a reflection of the work of the Transparency Working Group at the Integrity Institute

- You can see our existing transparency briefings and our newest on when it releases (soon!) at https://integrityinstitute.org/resources
- We have many world leading experts in and all aspects of integrity and T&S work in our membership
- We love talking about how to make their transparency reports better and more comprehensive, for both civil society and companies
- We want to help everyone get this right.

Please reach out! hello@integrityinstitute.org

What is meaningful transparency?

What is meaningful transparency?

- Lots of new regulations mandating transparency
- This is great!
- Transparency is one of the most impactful policy options
  - Can increase the influence of trust & safety workers inside of companies
  - Can change business incentives
- But there are absolutely right ways and wrong ways to do transparency

Quick Example: Prevalence

- Platforms like Facebook, Instagram, and YouTube are releasing prevalence metrics in transparency reports
  - This is good! A nice proactive step from the platforms
  - And, when you're inside the companies, prevalence is the most useful metric (A/B testing, goal setting)
- But…. from the outside… who cares?
  - How should the public interpret them? Is 0.02% prevalence of hate speech good? Or bad? Responsible or irresponsible?
- Simply put, **the public doesn't really care about prevalence**
  - They care the true scale: How many exposures? How many people?
  - They care about why the exposures happen: From platform recommendations? From users DM'ing each other?
  - They care about the nature of the harms: How frequently are viewers exposed? What's borderline and violating content like?

https://transparencyreport.google.com/youtube-policy/views

What is meaningful transparency?

- Meaningful transparency means
  - Informing the public about the scale, cause, and nature of harms that occur on platforms
  - Giving the public enough information to validate the claims of the platforms
  - Giving the public enough information to ensure platforms are designed responsibly, using best practices
  - Broadly aiding the public in understanding platforms and their impact

- Meaningful transparency will provide accountability
  - Incentivize companies to take concerns and recommendations from integrity and trust & safety teams more seriously
  - Create a stronger business incentive to design the platform more responsibly

Transparency should empower people inside of companies to do the right thing and strengthen the health of platforms for the long term, from both business and user point of views.

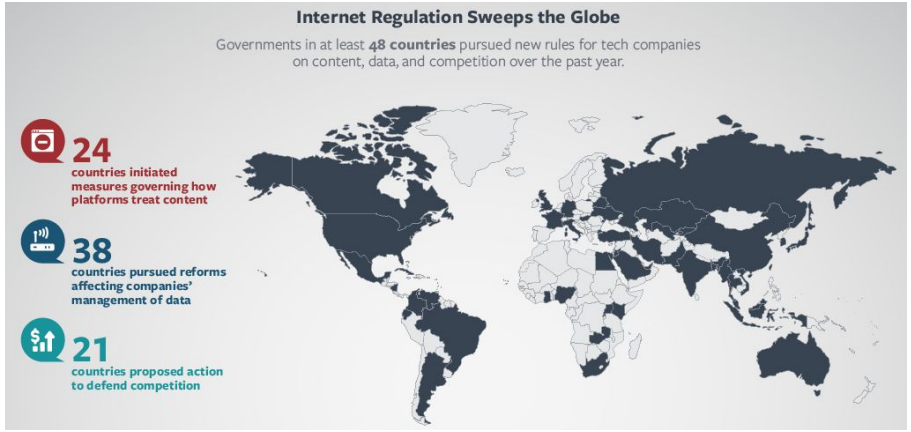
Transparency should empower society to make informed decisions around responding to harms on the social internet.

The Coming Regulation

Source: Ned Shark

Transparency Regulation

- Regulation that involves transparency passed in
  - California
  - Australia
  - The UK
  - The EU
- Regulation that involves transparency has proposed in
  - US (Federal)
  - Brazil
  - India
- Basically, it's coming, and when combined, it's going to be comprehensive



https://freedomhouse.org/report/freedom-net/2021/global-drive-control-big-tech

Transparency Regulation

- Regulation is coming in the form of
  - Audits
  - Documentation of product functions and policies
  - Reports on content moderation
  - Risk assessments
  - Data access
- In theory, this could be a legal mess!
  - Each regulation asking for things widely different things
- It doesn't look like this is the case, there's a lot of alignment between regulation

Transparency Regulation

DSA Article 26 [EU]

5. When conducting risk assessments, very large online platforms shall take into account, in particular, how their content moderation systems, recommender systems and systems for selecting and displaying advertisement influence any of the systemic risks referred to in paragraph 1, including the potentially rapid and wide dissemination of illegal content and of information that is incompatible with their terms and conditions.

PATA [US]

(b) TRANSPARENCY OF CERTAIN CONTENT AND USER ACCOUNTS.—

(1) IN GENERAL.—Not later than 1 year after the date of enactment of this Act, the Commission shall, in accordance with section 553 of title 5, United States Code, and subject to subsection (g), issue regulations to require platforms to make available to the public on an ongoing basis, in a specific section of their online interface, through a searchable and reliable tool that allows multicriteria queries and through application programming interfaces, a repository containing information regarding reasonably public content on the platform that—

(A) has been highly disseminated; or

(B) was originated or spread by major public accounts.

(2) DISCLOSURE OF PUBLIC CONTENT SAMPLINGS.—The regulations issued under paragraph (1) shall further require platforms to disclose on an ongoing basis statistically representative samplings of reasonably public content, including, at a minimum, a sampling that is weighted by the number of impressions the content receives.

AADC [California]

(i) Whether the design of the online product, service, or feature could harm children, including by exposing children to harmful, or potentially harmful, content on the online product, service, or feature.

Online Safety Bill [UK]

(f) the different ways in which the service is used, and the impact of such use on the level of risk of harm that might be suffered by children;

(g) the nature, and severity, of the harm that might be suffered by children from the matters identified in accordance with paragraphs (b) to (f), giving separate consideration to children in different age groups;

(h) how the design and operation of the service (including the business model, governance, use of proactive technology, measures to promote users' media literacy and safe use of the service, and other systems and processes) may reduce or increase the risks identified.

https://digitalservicesact.cc/dsa/art26.html
https://leginfo.legislature.ca.gov/faces/billCompareClient.xhtml?bill_id=202120220AB2273&showamends=false
https://bills.parliament.uk/publications/51870/documents/3679
https://www.congress.gov/bill/117th-congress/senate-bill/5339/text#toc-id058DC41D2F27402882F730E0DE93B809

Source: Gollum

What is “risky” content?

- Bills call out “risk assessments.” But what risks? What content?
  - Illegal content
  - Content that risks human rights
  - Content that harms children

- But most platforms have already done this work
  - Platforms do have good incentives to develop comprehensive content policies
  - Codified all this content in their content policies

- So, let’s go with the easiest answer: Content that violates platform policy
- **“Violating content” will be a shorthand for “risky content,” “illegal content,” and “harmful content”**

Transparency Frameworks

Transparency Frameworks

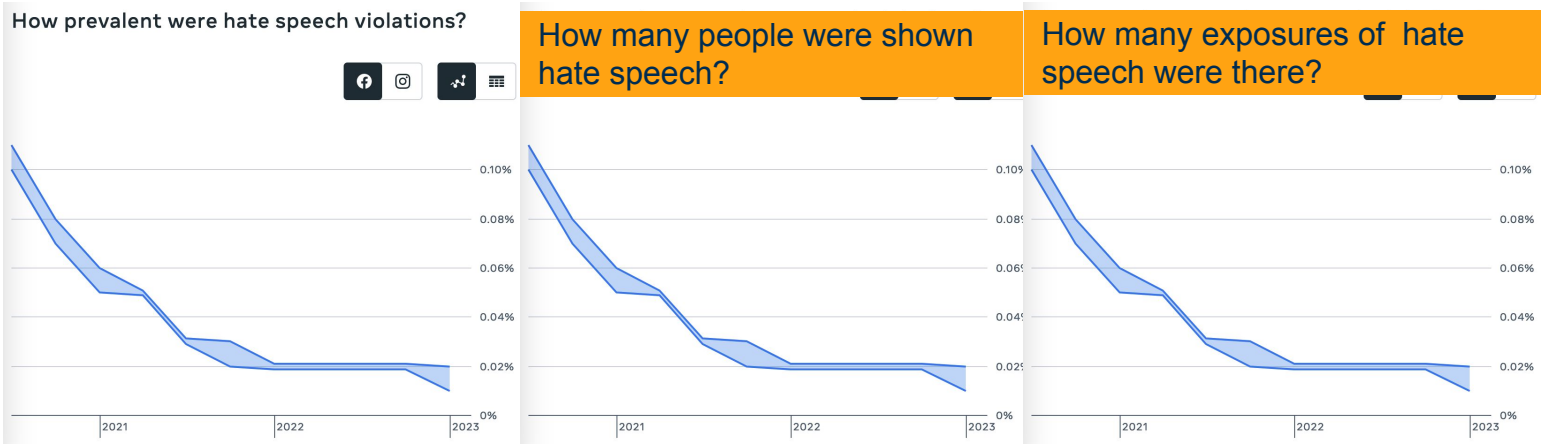
Goal of the frameworks

- Meet the spirit and the letter of the risk assessment laws
- Make clear the scale, cause, and nature of exposures to and risks from violating content
- Enable external parties to validate the report
- Explain how algorithms and platforms are designed and the role they play in exposures to and risks from violating content
- Explain how company processes play a role in exposures to and risks from violating content

Scale of harms and risks

Delivered in a (public) transparency report

- What is the prevalence of violating content?
- How many exposures to violating content are there in a given time window (30 days, etc.)?
- How many users are exposed to violating content in a given window?
- All broken down by
  - Violation type
  - Basic demographics (region, age bucket)
  - Other additional, aggregate details, (protected classes or vulnerable groups) when appropriate

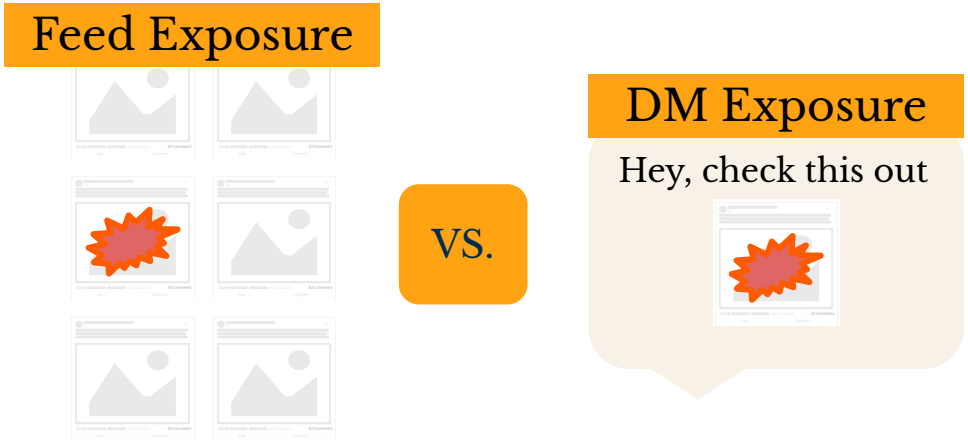


https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/

Cause of harms and risks

Delivered in a (public) transparency report

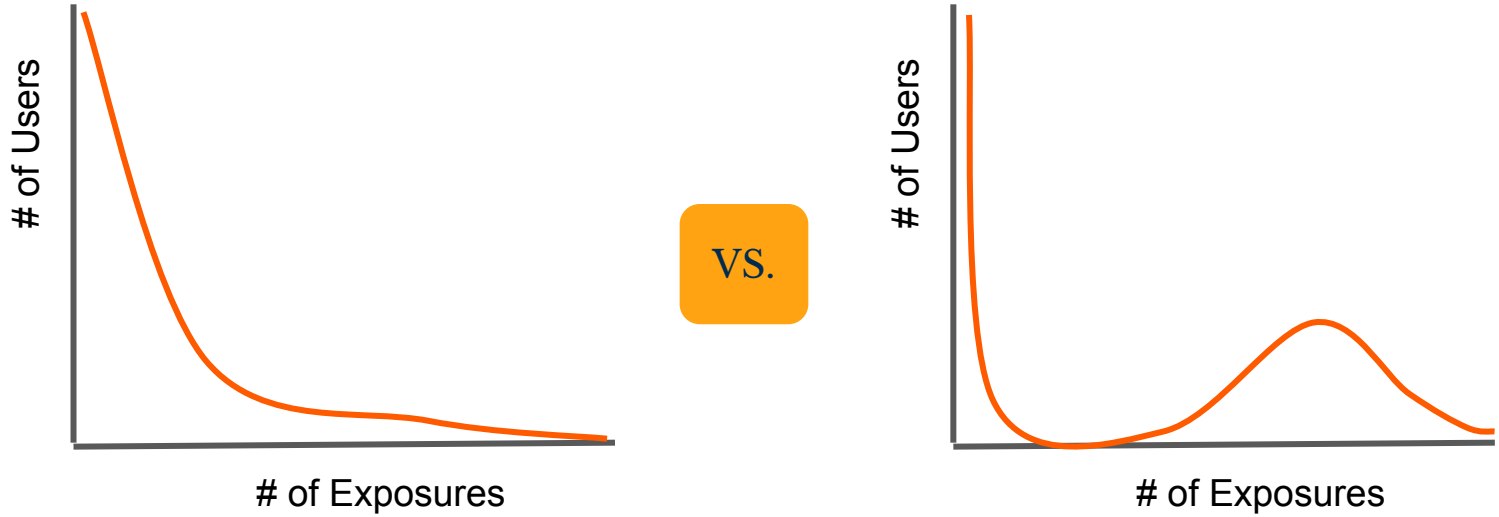
- The frequency of underlying causes of the exposures to violating content
  - What % take place in an algorithmic feed?
  - What % are from creators that the user follows?
    - What % of those follows came from an algorithmic recommendation?
  - What % of exposures are on public content?
  - What % of exposures are from creators that have previously posted violating content?
- Different means of exposure have different implications for risk
  - Users DM'ing each other violating content has a different risk profile than algorithms recommending it



Nature of harms and risks

Delivered in a (public) transparency report

- For users that are exposed to violating content, what is the distribution of # of exposures in a given time window (30 days)?
- All users seeing a low level of exposures to violating content has a different risk profile than a few users with high levels of exposure



Enabling external validation

Delivered in (public) Data Sets

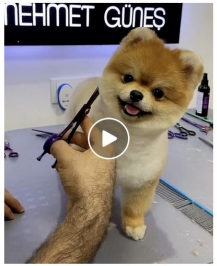
- The top N pieces of public content (Ideally by views)
  - N should be 10,000+
  - Released on a regular and fast cadence (Daily or weekly)
- A random sample of N impressions on public content
  - N should be at least 10,000+
  - Released on a regular and fast cadence (Daily or weekly)
- What should be in the dataset? **Could be very minimal and privacy respecting**
  - For Top N, could be literally two columns: URL to content, # of views in relevant time window
  - For random sample could be ~5 columns:
    - URL to content
    - Where impression took place (Which surface/product in the platform)
    - What algorithmic systems played a part in making the impression
    - If the user followed the creator of the content or not
    - If the user followed the creator due to an algorithmic recommendation
    - Maybe: basic demographic information (region, age range)
    -
- Ideally: **Real time to support civil society groups** trying to protect their communities
  - Especially around critical events
- **These are all reasonable datasets to ask for**. Platforms are already providing versions of them.

Enabling external validation

Meta's CrowdTangle

crowdtangle

FB WVCR

Rank	Post Link	Post Image	Content Viewers
1	facebook.com/526392 9853637131	This post is no longer publicly available. Its owner may have removed it, changed the post's privacy settings or changed their account settings.	53M
2	facebook.com/101594 99731211026		51.9M
3	facebook.com/646950 5544430315		46M
4	facebook.com/102265 50523423452		44.9M

Twitter Sample Strem

Volume streams

GET /2/tweets /sample/stream

Streams about 1% of all Tweets in real-time.

If you have Academic Research access, you can connect up to two redundant connections to maximize your streaming up-time.

Run in Postman › | Build request with API Explorer ›

Endpoint URL

`https://api.twitter.com/2/tweets/sample`

https://transparency.fb.com/data/widely-viewed-content-report/

https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams/api-reference/get-tweets-sample-stream

Algorithmic Transparency

Delivered in a (public) transparency report

- Summary of basic components:
  - Inventory, Features, Machine Learning Models, Value Model, and System Evaluation
- Features
  - Most important features in the machine learning models
  - Especially features that rely on the **individual users data and engagement history**
- Machine Learning Models
  - Most important machine learning models and what they are trying to predict
  - Especially models that are **predicting user actions**
  - How violating content performs in the models*
- Value Model
  - Most significant contributions to the value model
  - Ideally just the straight value model
- System evaluation
  - Topline metrics used to evaluate the ranking system
  - Most important metrics used in evaluating A/B tests
- **These are all reasonable to ask for**. Platforms are already providing versions.

Algorithmic Transparency

Meta Model Cards

How likely you are to reply with text to a story in an author's collection that you've started to watch

Signals influencing this prediction include:

- How many text replies to stories you've sent
- Where data privacy laws permit:
  - How many times you've communicated with the story's author such as through chat messages
  - How many times you've exchanged messages on a mobile device with the story's author
- The total number of stories that people have closed by navigating right in the author's collection
- The total amount of time people have spent viewing the author's stories

The predicted amount of time you will spend watching each new story added to an author's collection

Signals influencing this prediction include:

- The average number of stories from the author that people have watched
- The average amount of time that people have spent viewing the author's stories
- The total number of stories from the author that have advanced automatically to the next story as people have watched them
- How many photos are in the author's collection of stories
- The total number of stories that people have closed by navigating right in the author's collection

How likely you are to watch stories in an author's collection for longer than the average amount of time that other people, with the same number of unseen stories as you have, spend watching stories

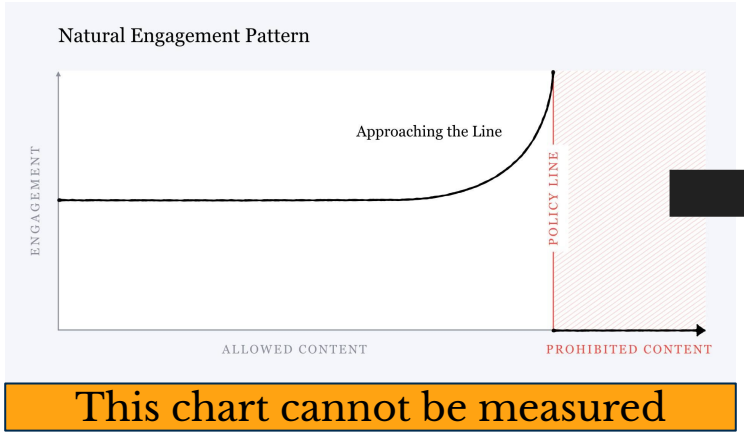
Twitter "The Algorithm" Source Code

```
Twitter Ranking Score =
    75  * is_replied_reply_engaged_by_author
  + 27  * is_replied
  + 12  * is_profile_clicked_and_profile_engaged
  + 11  * MAX(
            is_good_clicked_convo_desc_favorited_or_replied,
            is_good_clicked_convo_desc_v2
          )
  + 1.0 * is_retweeted
  + 0.5 * is_favorited
  + 0.005 * is_video_playback_50
  - 74  * is_negative_feedback_v2
  - 369 * is_report_tweet_clicked
```

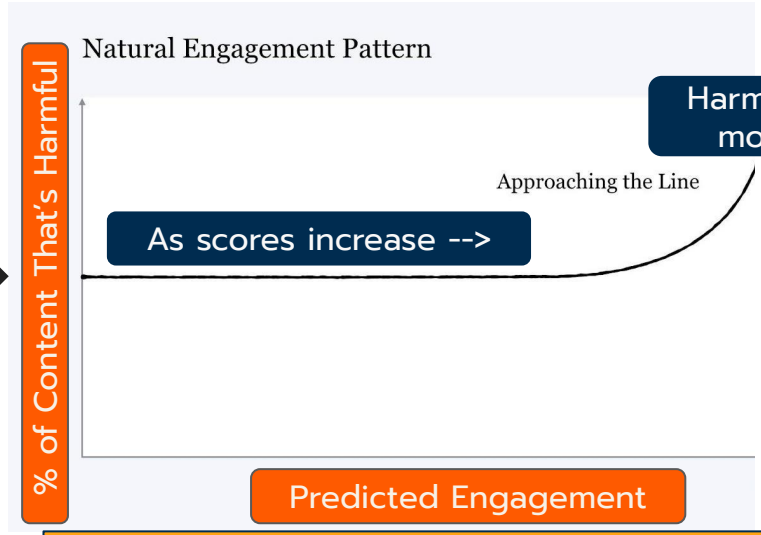
https://transparency.fb.com/features/explaining-ranking
https://github.com/twitter/the-algorithm

Algorithmic Transparency

- It is possible to evaluate how much algorithmic systems amplify violating content
- The "Natural Engagement Pattern" from Zuckerberg sets expectations
  - Engagement based ranking will amplify violating content
- This can (has been) measured at platforms, can be made public
- Platforms should publish how known violating content performs in their engagement classifiers



This chart cannot be measured



But this one can

https://www.facebook.com/notes/751449002072082/

Transparency of company processes

Delivered in a (public) transparency report

- Summary of company processes that can impact exposures to violating content
- Core metrics used in A/B testing
- An outline of their processes for product or ranking changes
  - Compare process for features to reduce harms on the platform and “normal” growth 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

Wrap Up

Key Takeaways

- Regulation mandating transparency is coming, from all over the world
- This is most likely going to be a good thing, and it could be a great thing
  - **Meaningful transparency can raise the influence of integrity and T&S teams internally**
- Goals of transparency:
  - Informing the public about the scale, cause, and nature of harms that occur on platforms
    - Giving the public enough information to validate the claims of the platforms
    - Giving the public enough information to validate the platforms are designed responsibly
    - Broadly aid the public in understanding platforms and their impact
  - Meaningful transparency will provide accountability
    - Incentivize companies to take concerns from integrity and trust & safety teams more seriously
    - Incentivize companies to take recommendations from integrity and trust & safety teams more seriously
    - Create a stronger business incentive to design the platform more responsibly
- These can actually all be achieved through comprehensive transparency
  - And **an** interpretation of all the bills is close to getting us here
  - TBD if that interpretation will win in the end…

Integrity Institute