Tennessee v. Meta
Case No. 23-1364-IV
PX-10232

Message

From: Andrea Saul [/O=THEFACEBOOK/OU=EXTERNAL (FYDIBOHF25SPDLT)/CN=RECIPIENTS/CN=[redacted]]
Sent: 12/1/2017 12:36:41 AM
To: Adam Mosseri [[redacted]@fb.com]; Tucker Bounds [[redacted]@fb.com]; Tessa Lyons-Laing [[redacted]@fb.com]; Lindsey [redacted] [[redacted]@fb.com]
CC: Elliot Schrage [[redacted]@fb.com]
Subject: Data transparency report re: fake news

Hey Adam/Tessa/Tucker/Shep-

TLDR: Community Integrity is going to release data as part of a re-branded transparency report in H1 of next year, and I think that will put added pressure on us to also release data in relation to the prevalence of fake news on the platform. This could be a good opportunity to release data on a variety of topics in the effort to build up trust through our transparency efforts.

Longer version: I have been working with policy and Community Integrity (formerly known as PAC) towards releasing a transparency report highlighting the prevalence of violating content on Facebook and how we handle it (longer plan below, if helpful). For example, how much [insert community standard issue here] we remove from Facebook and, where relevant, what percentage we find proactively and how quickly we remove it. We are going to take a look at the numbers before we are fully committed to releasing publicly, but the hope is that some short-term pain will be helpful for long-term gain.

Right now, we are aiming to be ready to put out numbers in March with a likely April release; we will tie this to the existing government request report, which is being re-branded as the transparency report this next half complete with a new website. We're tentatively planning on a Hard Questions post to (re-)educate the community on how these issues should be measured and then follow with the actual numbers.

I'm reaching out to you as fake news was raised as another issue people (i.e. Elliot, Rachel) think would be relevant to include. Curious if this is something you all are working on--or would like to be and just don't have the resources to dedicate to it. Let us know your thoughts and, if you think you'll head this direction, what may be needed so that we could do this in coordination for a broader transparency announcement.

Please let me know any questions you may have.

Thanks!
Andrea

Community Integrity Data Transparency Plan

Goal of this document
This document is a proposal for the proactive comms strategy for Community Integrity for 2018, specifically our intent to increase transparency and share metrics around Community Standards. Our objective today is to collect directional feedback on:

1. Areas we are missing
2. Approach to metrics

Non-goals

Overall company Integrity narrative. At a later stage, the result of this process will be combined into the overall FB Integrity narrative, as that comes together.

Mission

Support the company's Integrity / defense narrative (*"We get it, we act responsibly"*), by driving these themes:

1. The issues Facebook faces are difficult and complex
2. Facebook is acting responsibly, is capable and effective at managing content on its platform
3. Facebook is making a good faith effort to be transparent

Strategy

1. Identify key problem areas and stakeholders
2. For each area, define ideal messages to land
3. Where the narrative can be supported by metrics transparency:
    a) Select the best metric to measure our progress (ideally this aligns with our internal metrics)
    b) Compile metrics into regular transparency reports
    c) Proactively educate stakeholders on the right way to fairly measure the space
4. Provide input for and support product launches

Problem and Stakeholder Map

Problem area	Stakeholders	Concerns	Ideal messages
Policy and enforcement (Facebook accountability)	• Global policymakers • Global civil society & academics • Global media	Facebook makes global decisions about free speech with no oversight, accountability, or transparency	Facebook has robust systems to ensure that our enforcement actions are fair and effective
False positives	• Global policymakers (mostly US) • US minority leaders	• Facebook censors free speech, especially for minorities • FB's reporting system is leveraged by hate groups / trolls to coordinate reporting and silence voices ("brigading")	Facebook has robust systems to ensure that our enforcement actions are fair and effective
Hate speech	• EU policymakers • US policymakers • Global activists • US minority leaders	FB is not doing enough to combat online hate	FB effectively enforces against hate speech
Terrorism	Policymakers (global, mostly EU+US)	FB is a platform for terrorist propaganda	FB takes a strong stance against terrorism and can help governments counter terrorism
Graphic violence	• Global policymakers • Global media • Global civic society • Global public	FB is a cesspool for violent content and nothing is done until too late	FB effectively enforces against graphic violence
Brand safety (hate, terrorism, violence, porn)	Clients	Ads are shown next to problematic content	• Consumers do not associate ads with content above/below • There is a low probability of ads appearing next to problematic content
Fake accounts / authenticity	• Global policymakers • Countries with upcoming elections • Clients	• FB has a high number of fake accounts • FB is enabling election influence through fake accounts	FB effectively enforces against fake accounts
Suicide / self-injury	• Policymakers (global, mostly US+APAC) • NGOs • US public	FB increases exposure to content that increases contagion	FB saves lives through supporting people in time of crisis
Harassment	• Global media • Bloggers • Public figures	High-profile individuals are harassed on FB	FB is a safe place to express views
Child safety	Global policymakers	• FB puts minors at risk from sexual offenders • Minors can be exposed to alarming inappropriate content on FB	FB has mechanisms in place to protect minors on the platform

Women safety	Global policymakers (mostly APAC, EMEA, US)	Facebook is used to harass women	FB takes women's safety and has a number of policies, tools and programs in place to protect women

Metrics Approach

For each area, we will approach metrics using the below framework:

Question	Potential Metrics	What they say	Pros	Cons / Limitations	Recomm-endation
1. How much badness is experienced by our users?	VPV-prevalence, account prevalence, interactions prevalence	<1% of VPVs are bad 2-3% of accounts are fake	Accurately reflects user experience on the platform	• Harder to interpret because of confidence intervals and changes in coverage • Does not highlight proactive efforts • Sensitive to labeling quality	Need to use/ understand this to put #'s 2-4 In context, educate ppl on how to measure
2. How much badness do we catch?	Content/accounts actioned, Content/accounts reported	We action a lot (millions each week). User reports are not necessarily high quality	Available, easy to interpret, highlights FB efforts, highlights user report quality	• Attacker-driven • May not reflect how much badness people experience • Not VPV-weighted • Does not capture preventative efforts (blocking before reg / before post)	Use
3. How proactively do we detect it?	% proactive (of actioned)	We are very proactive for some violations (e.g.: Spam, CT), but not others (e.g.: Hate)	Highlights FB's proactiveness	• Can be gamed by reducing user reporting • Not VPV-weighted	Use
4. How quickly do we take it down?	Created-to-deleted lag, Reported-to-deleted lag	A large portion of what we take down is years old (e.g. CT) if we include retro-actioning	• Could establish trust with the regulators • Does give a sense on how long violating content stays up	• Paints a negative picture for some violations • Looks worse for violations that have retroaction and user-reporting.	Use

				Therefore, can create the wrong incentives. • Not VPV weighted	
5. How many mistakes do we make?	FP Reach, FP and FN rate for reps	• Very few (~0.05% FP Reach for Spam) are affected by FP • Many mistakes in user-report review	Helps establish trust	Not ready for most because it depends on GT	Not now
6. How many mistakes do we correct?	% appealable, # restored	• We don't review many appeals • A lot of actions are not appealable yet.	Helps establish trust	Not ready; appeals product is a WIP	Not now

External Metrics Timeline & Priorities

Based on a combination of comms priority + our ability to provide reliable metrics, we intend to prioritize external metrics reporting in the following sequence:

- 2017-Q4 (mostly ad-hoc / reactive needs)
  - Terrorism
  - Fake accounts
  - SSI
- 2018-H1 (planned comms, introduce regular transparency report tied to existing government requests report)
  - Hate speech
  - Violence
  - N&P
  - SSI