Choices and Consequences in Computing

INFO 1260 / CS 1340

Lecture 9: Content Moderation

February 9, 2024

Platforms’ decision-making about content

Necessary and impactful choices about how to rank and display content, how and how much to personalize content to users…

… today, another set of platform decisions: how do platforms **moderate** undesirable content?

Platforms’ decision-making about content

- As we have seen, the law does very little to constrain speech online…
- … and under Section 230, platforms can legally moderate users’ content as much as they want, in any way they’d like to…
  - Including not at all
  - Including in a biased way
  - Including in an arbitrary and haphazard way
- But most platforms do moderate! Why? How?

Content note for today: this is difficult and controversial. We are less concerned with what the right policy is for any given piece of content, and more concerned with how platforms negotiate this decision-making.

Why moderate?

Given that they don’t legally have to – what are some reasons platforms might choose to moderate content?

Why moderate?

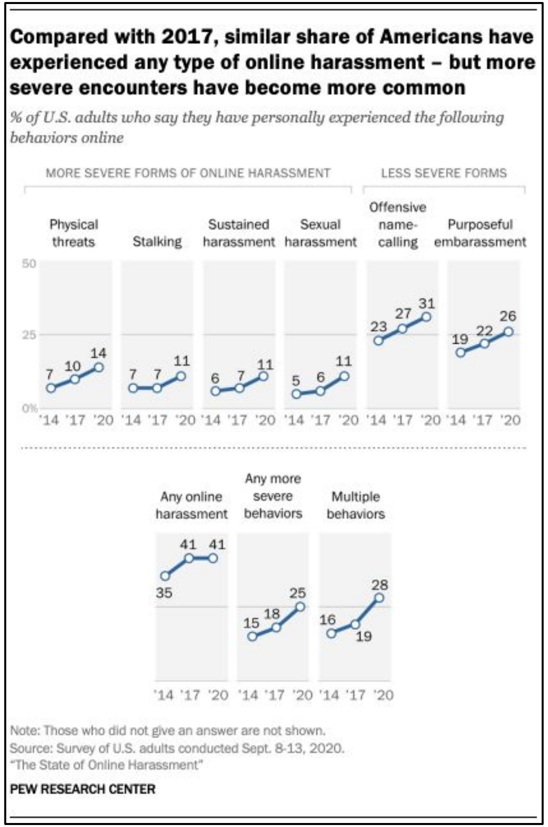
- Ideological commitments
- Happy users → more engagement
- Happy advertisers: the business model of the internet
- To prevent big-P policy (including to preserve 230 protections)
- Other external pressures (e.g. pressures from the infrastructural “stack”; public relations)

Context, purpose, audience, modality matter a lot for what kinds of choices platforms make about moderation!

Moderation includes not only decisions about whether to take down content, but also how to display it / whether to amplify it / etc.

There is plenty of content for platforms to moderate!

- Online harassment and abuse have been experienced by 41% of Americans
  - 68% of LGB Americans (51% have experienced severe behaviors like physical threats, stalking)
  - 33% of women under 35 report being sexually harassed online
  - Of those harassed, roughly half of Black and Hispanic respondents reported that they were targeted due to race (17% for white respondents)
  - Many more stats: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/
- In the worst cases, can have very disruptive effects on offline well-being and health (Citron)



Compared with 2017, similar share of Americans have experienced any type of online harassment – but more severe encounters have become more common

% of U.S. adults who say they have personally experienced the following behaviors online

Note: Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Sept. 8-13, 2020.
"The State of Online Harassment"
PEW RESEARCH CENTER

A global problem with fuzzy boundaries

- Figures from Thomas et al., “Hate, Harassment, and the Changing Landscape of Online Abuse” (2021)
- Relation to misinformation and radicalization, intimate partner surveillance / abuse (which we’ll discuss when we talk about privacy)

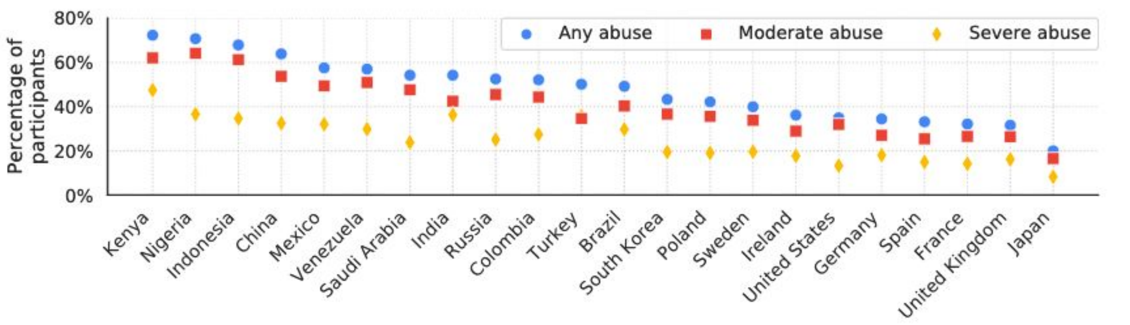


Fig. 2: Percentage of participants reporting any, moderate, or severe hate and harassment online per country, aggregated over 2016–2018.

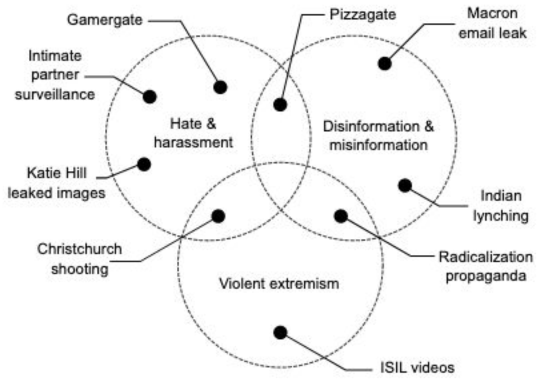


Fig. 1: Attackers engaged in hate and harassment may also engage in violent extremism and disinformation and misinformation. We argue there is not always a clear boundary between each class of abuse, but that the underlying intents differ.

Hard content moderation policy choices

General standards or specific rules?

- Somewhat more general (Reddit Rules) vs. more specific (Facebook community standards)
- Warning: even the rules contain references to lots of disturbing content
- Both have advantages and disadvantages
- Analog in the law: rules vs. standards
    - Rules: bright-line, clear—but inflexible
    - Standards: flexible, allow more discretion; but less consistently interpreted and administrable
    - E.g.: “the president must be 35 years old” vs “the president must be sufficiently mature” (Lawrence Solum)

The same for all groups, or different for some groups?

- A good Radiolab episode about the “Kill All Men” debate
  - Also removed: ”Men are the worst,” “Men are so useless”
  - Experts around the world have extremely different views about the appropriateness of pieces of content like this
  - An enormous number of social, cultural, and political contexts, constantly changing, and decisions must be made extremely quickly

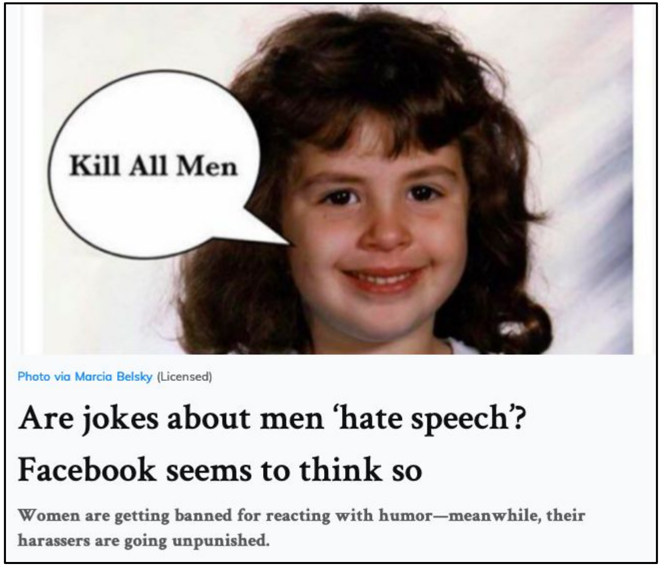


Photo via Marcia Belsky (Licensed)

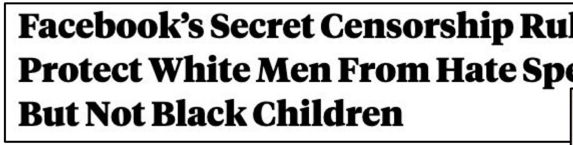
Are jokes about men ‘hate speech’? Facebook seems to think so

Women are getting banned for reacting with humor—meanwhile, their harassers are going unpunished.

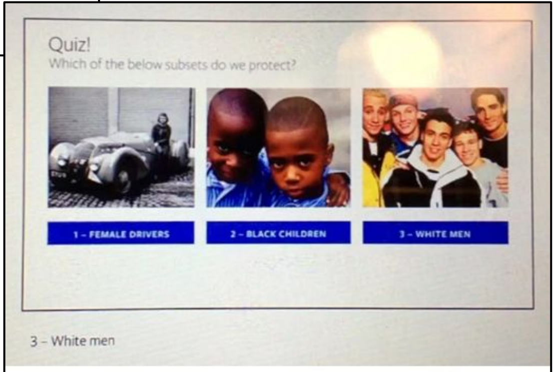
One example of policy change: are groups treated the same with respect to hate speech?

Facebook policy pre-December 2020: all “protected categories” treated the same (race, sex, gender, religion, national origin, ethnicity, sexual orientation, disability/disease)

- But: “subsets” are not protected
- Rules do not take historic discrimination/marginalization into account

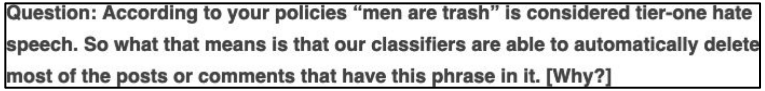


Facebook’s Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children

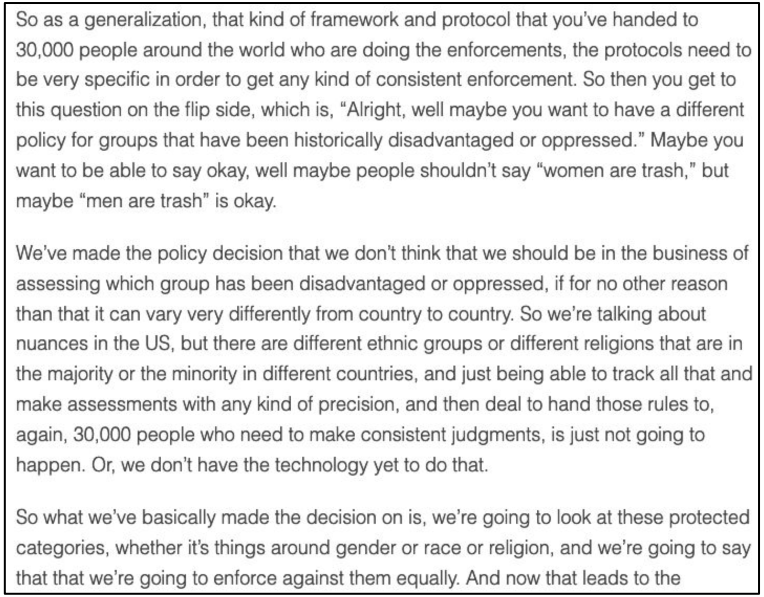


Quiz!
Which of the below subsets do we protect?

3 – White men



Question: According to your policies “men are trash” is considered tier-one hate speech. So what that means is that our classifiers are able to automatically delete most of the posts or comments that have this phrase in it. [Why?]



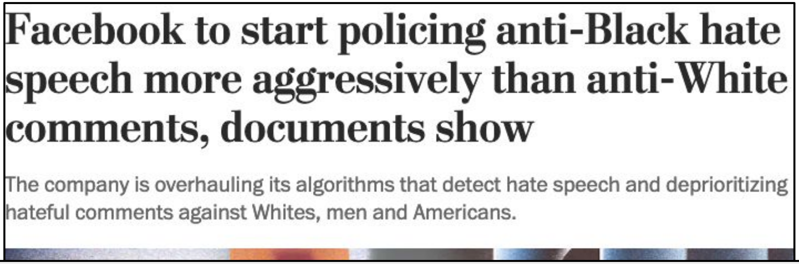
So as a generalization, that kind of framework and protocol that you’ve handed to 30,000 people around the world who are doing the enforcements, the protocols need to be very specific in order to get any kind of consistent enforcement. So then you get to this question on the flip side, which is, “Alright, well maybe you want to have a different policy for groups that have been historically disadvantaged or oppressed.” Maybe you want to be able to say okay, well maybe people shouldn’t say “women are trash,” but maybe “men are trash” is okay.

We’ve made the policy decision that we don’t think that we should be in the business of assessing which group has been disadvantaged or oppressed, if for no other reason than that it can vary very differently from country to country. So we’re talking about nuances in the US, but there are different ethnic groups or different religions that are in the majority or the minority in different countries, and just being able to track all that and make assessments with any kind of precision, and then deal to hand those rules to, again, 30,000 people who need to make consistent judgments, is just not going to happen. Or, we don’t have the technology yet to do that.

So what we’ve basically made the decision on is, we’re going to look at these protected categories, whether it’s things around gender or race or religion, and we’re going to say that that we’re going to enforce against them equally. And now that leads to the

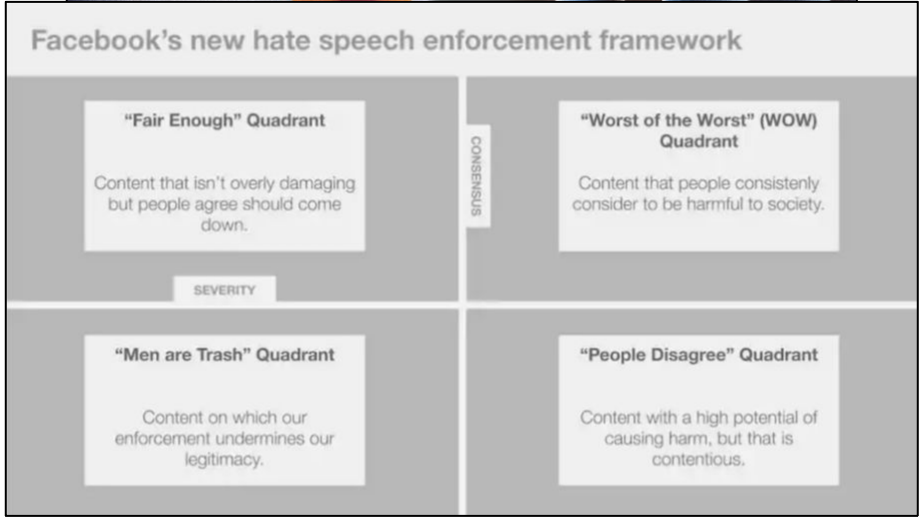
One example of policy change: are groups treated the same with respect to hate speech?

- Facebook policy overhaul, December 2020: “Project WoW”
- Spurred in part by civil rights audit and recognition that the most commonly taken down speech consisted of offensive characterizations of white people
- Comments against white people, men, and Americans could still be treated as hate speech, but not automatically deleted



Facebook to start policing anti-Black hate speech more aggressively than anti-White comments, documents show

The company is overhauling its algorithms that detect hate speech and deprioritizing hateful comments against Whites, men and Americans.



Facebook’s new hate speech enforcement framework

“Fair Enough” Quadrant

Content that isn’t overly damaging but people agree should come down.

“Worst of the Worst” (WOW) Quadrant

Content that people consistenly consider to be harmful to society.

“Men are Trash” Quadrant

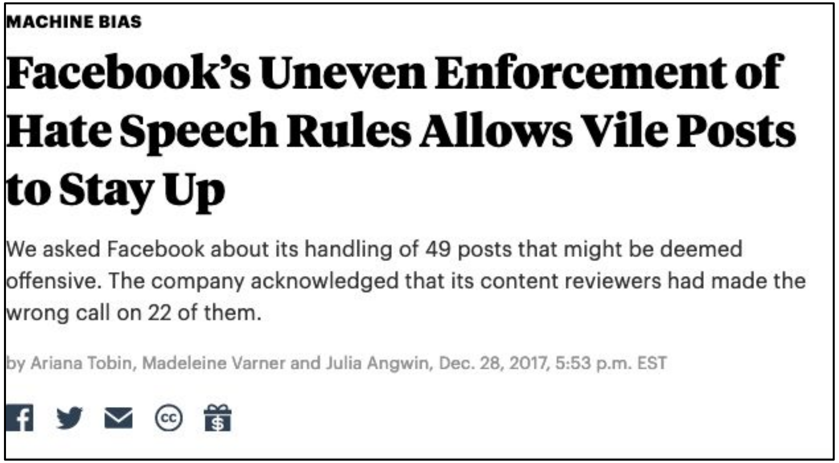
Content on which our enforcement undermines our legitimacy.

“People Disagree” Quadrant

Content with a high potential of causing harm, but that is contentious.

Inconsistent implementation

- Even when policies are clear, they are not always implemented consistently
- Speed and scale makes this inevitable
- Human content mods get about 8 seconds per piece of content
- You can see several examples here: https://projects.propublica.org/graphics/facebook-hate [many disturbing words and images, but each has a specific content warning]



MACHINE BIAS

Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up

We asked Facebook about its handling of 49 posts that might be deemed offensive. The company acknowledged that its content reviewers had made the wrong call on 22 of them.

by Ariana Tobin, Madeleine Varner and Julia Angwin, Dec. 28, 2017, 5:53 p.m. EST

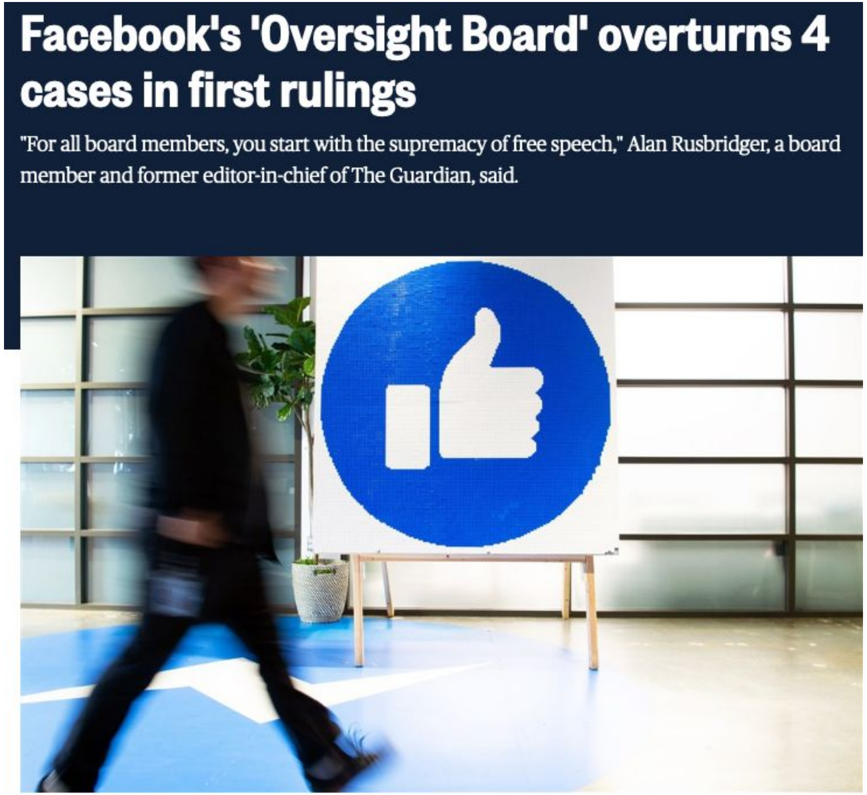
Hard content moderation implementation choices

Hard content moderation implementation choices

- Who does the work of content moderation?
- Automated vs. human
- Ex ante vs. ex post
- False positives vs. false negatives

Who does the work of content moderation?

- Users
  - As reporters
  - As volunteers
- Hired moderators
  - Often outsourced contract labor
- Companies/execs themselves
  - For high-salience decisions
- AI/automation
- Quasi-governmental appeals bodies

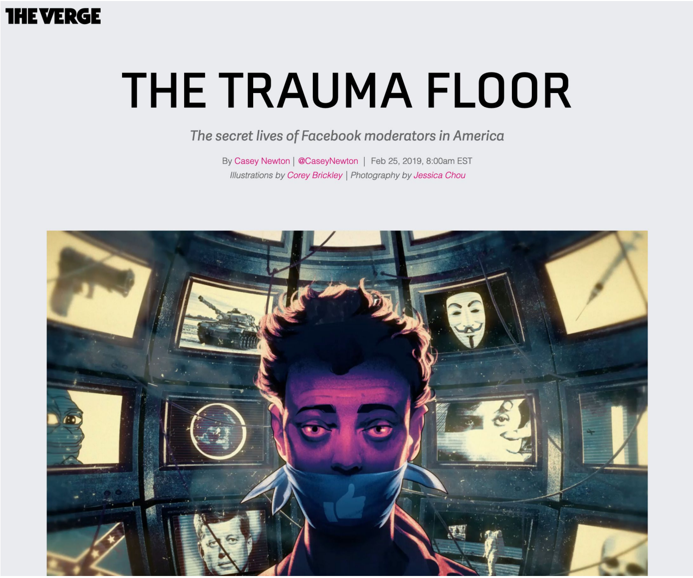


Facebook's 'Oversight Board' overturns 4 cases in first rulings

"For all board members, you start with the supremacy of free speech," Alan Rusbridger, a board member and former editor-in-chief of The Guardian, said.

Hired content moderators

- Extremely terrible, poorly paid, traumatizing jobs
- Often outsourced to contractors around the world in places with poor labor protections
- Be aware if you take a look at these pieces that they are extremely disturbing to read →



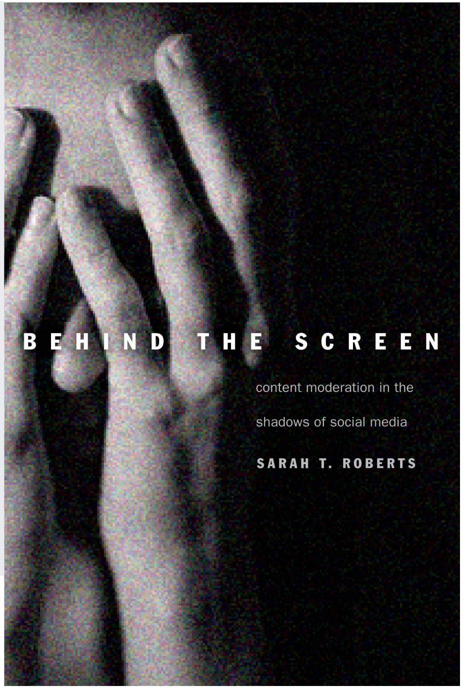
THE VERGE

THE TRAUMA FLOOR

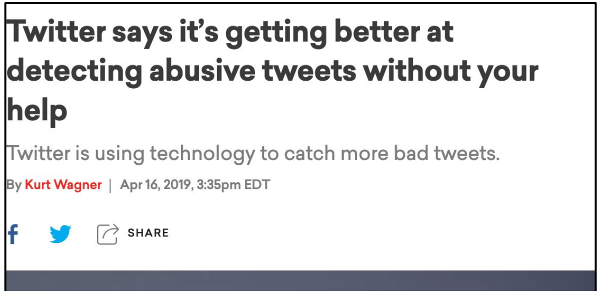
The secret lives of Facebook moderators in America

By Casey Newton | @CaseyNewton | Feb 25, 2019, 8:00am EST

Illustrations by Corey Brickley | Photography by Jessica Chou



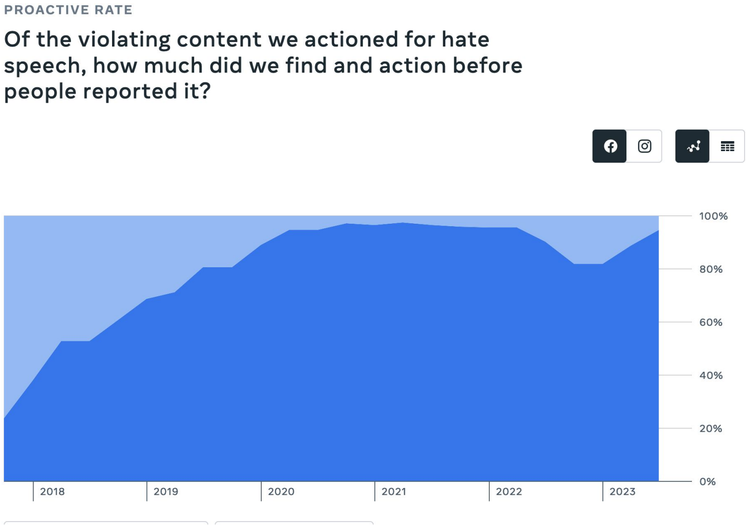
Movement toward more proactive moderation using AI



Twitter says it's getting better at detecting abusive tweets without your help

Twitter is using technology to catch more bad tweets.

By Kurt Wagner | Apr 16, 2019, 3:35pm EDT



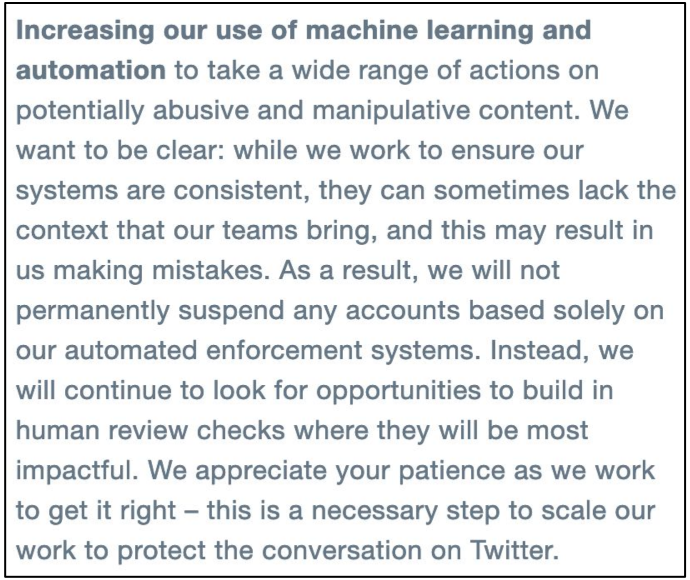
PROACTIVE RATE

Of the violating content we actioned for hate speech, how much did we find and action before people reported it?

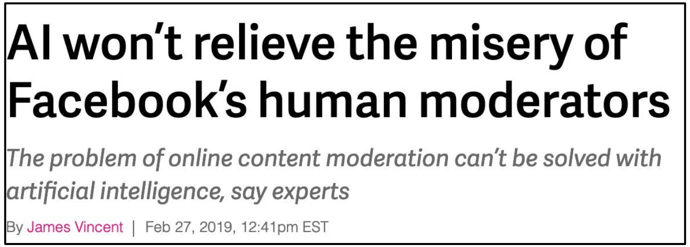
- Companies moving toward automating more content moderation
- In 2018, 0% of tweets involving "abusive behavior, hateful conduct, encouraging self-harm, threats" were identified proactively by the company; in 2019, 38% were
- Facebook community standards transparency reports

How much can AI help with content moderation?

- Making moderation easier for human moderators is desirable…
- But AI misses a huge amount of cultural and political context, subtlety, sarcasm, parody, meaning… and certainly can't replace human judgment
- Perhaps the best it can do is help with triage/flagging for human review
- James Grimmelmann: "AI can't solve political problems. What's fake news depends on who you ask. Kicking the question over to AI just means hiding value judgments behind the AI."



Increasing our use of machine learning and automation to take a wide range of actions on potentially abusive and manipulative content. We want to be clear: while we work to ensure our systems are consistent, they can sometimes lack the context that our teams bring, and this may result in us making mistakes. As a result, we will not permanently suspend any accounts based solely on our automated enforcement systems. Instead, we will continue to look for opportunities to build in human review checks where they will be most impactful. We appreciate your patience as we work to get it right – this is a necessary step to scale our work to protect the conversation on Twitter.

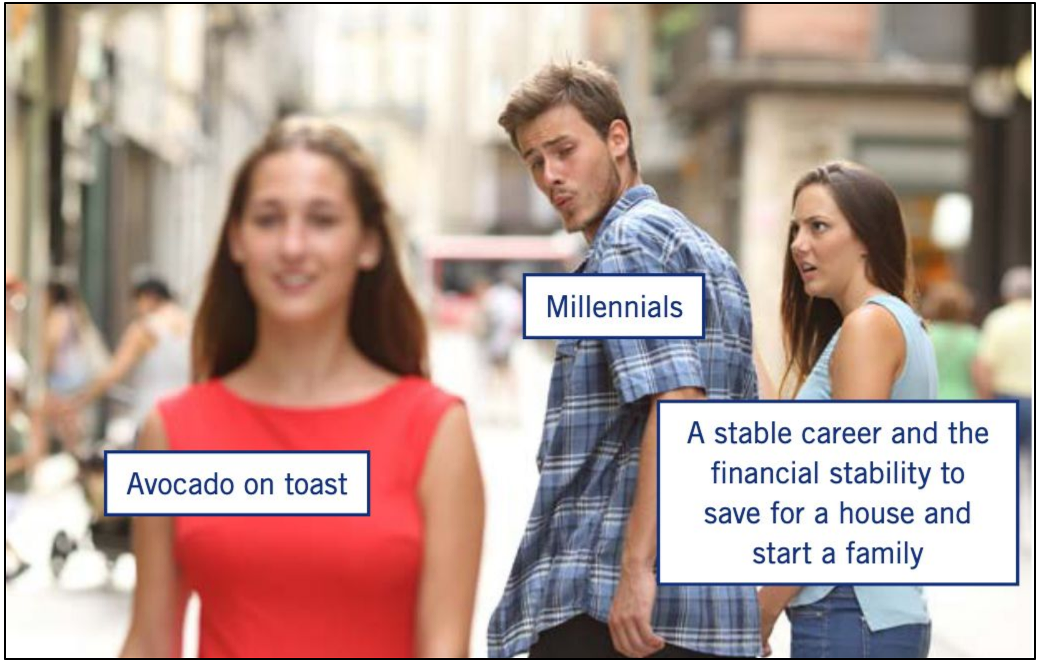


AI won't relieve the misery of Facebook's human moderators

The problem of online content moderation can't be solved with artificial intelligence, say experts

By James Vincent | Feb 27, 2019, 12:41pm EST

An intuitive example of why this is hard for AI

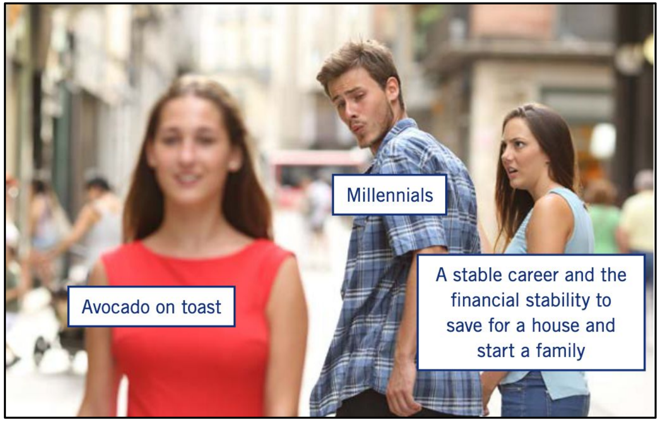

Millennials
A stable career and the financial stability to save for a house and start a family
Avocado on toast

Source: Ofcom, “Use of AI in Content Moderation” (2019)

Things you need to understand to interpret this:

- The structure of this meme
- What avocado toast signifies
- What “a stable career…” indicates
- Stereotypes about Millennials [not to mention judgment about whether this is an appropriate or offensive joke]
- What the facial expressions of the people indicate
- The presumed motives of the people in the picture
- Is this a joke or is it serious?

This is obviously very far from the most complex or controversial judgment to be made in content moderation – but the point is how nuanced this judgment is.

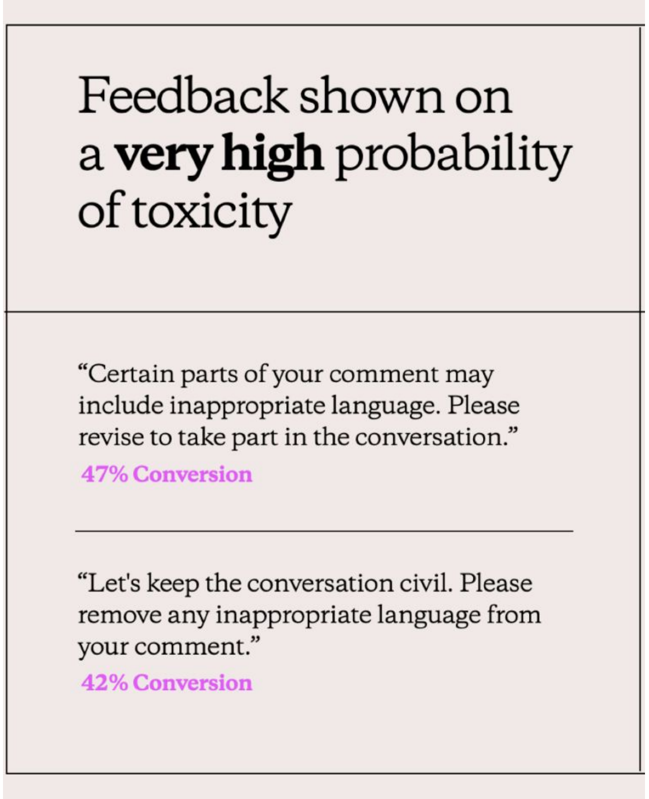

Avocado on toast
Millennials
A stable career and the financial stability to save for a house and start a family

Source: Ofcom, “Use of AI in Content Moderation” (2019)

Ex ante vs. ex post

(before the event or after the event)

- Ex ante:
  - Reviewing all content (e.g. NYT comments section)
  - Preventing people from posting things to begin with (e.g. blocked words)
  - Interstitial messages warning people not to post problematic content
- Trade-offs of ex ante moderation:
  - Prevents harm from viewing content (e.g. very violent or problematic videos)
  - Prevents virality
  - Possibly bad user experience if people can't post immediately
  - High overhead
  - Akin to "prior restraint" (censorship) – serious infringement on speech
  - Limited transparency



Feedback shown on a **very high** probability of toxicity

"Certain parts of your comment may include inappropriate language. Please revise to take part in the conversation."

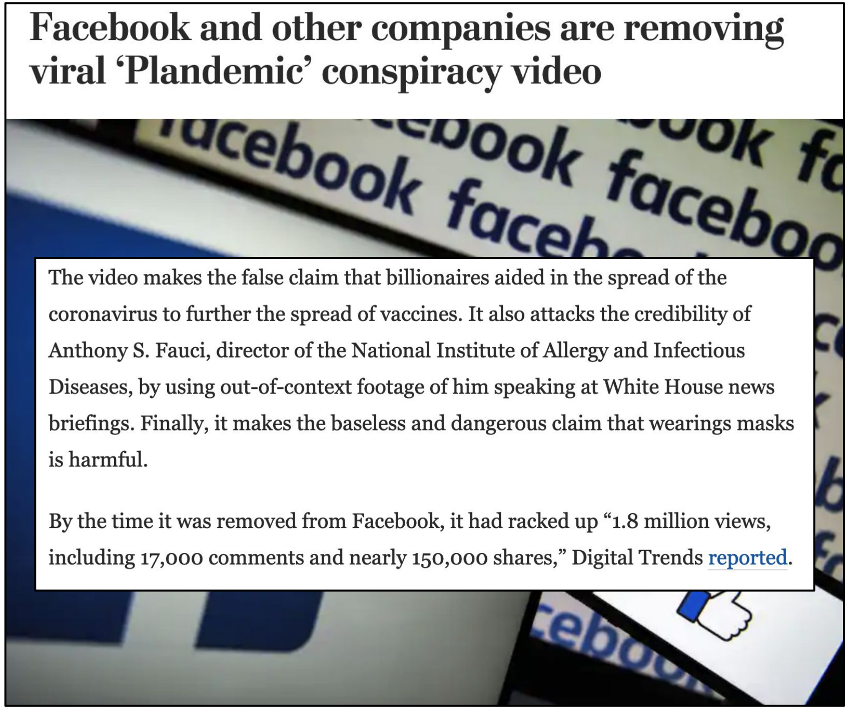
47% Conversion

"Let's keep the conversation civil. Please remove any inappropriate language from your comment."

42% Conversion

Ex ante vs. ex post

- Ex post:
  - Most common on the web
  - Reactive; waits until somebody finds out about the content (often users)
  - Imposes more of the burden on listeners (rather than speakers)
  - Processes can be very slow and impose harm
  - Whack-a-mole to catch new versions of viral content



Facebook and other companies are removing viral 'Plandemic' conspiracy video

The video makes the false claim that billionaires aided in the spread of the coronavirus to further the spread of vaccines. It also attacks the credibility of Anthony S. Fauci, director of the National Institute of Allergy and Infectious Diseases, by using out-of-context footage of him speaking at White House news briefings. Finally, it makes the baseless and dangerous claim that wearings masks is harmful.

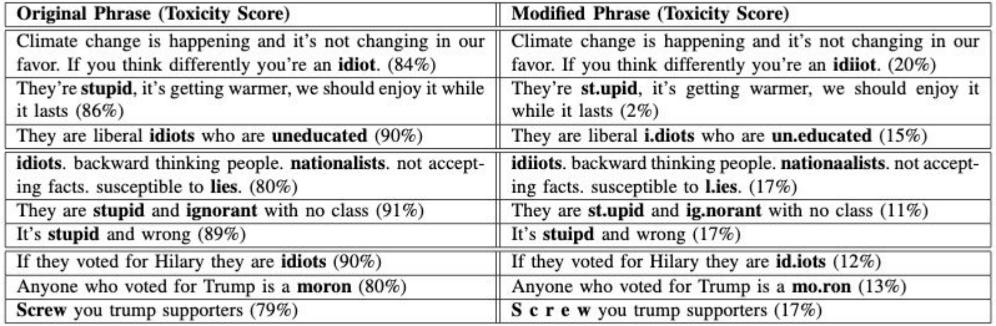
By the time it was removed from Facebook, it had racked up "1.8 million views, including 17,000 comments and nearly 150,000 shares," Digital Trends reported.

False positives vs. false negatives

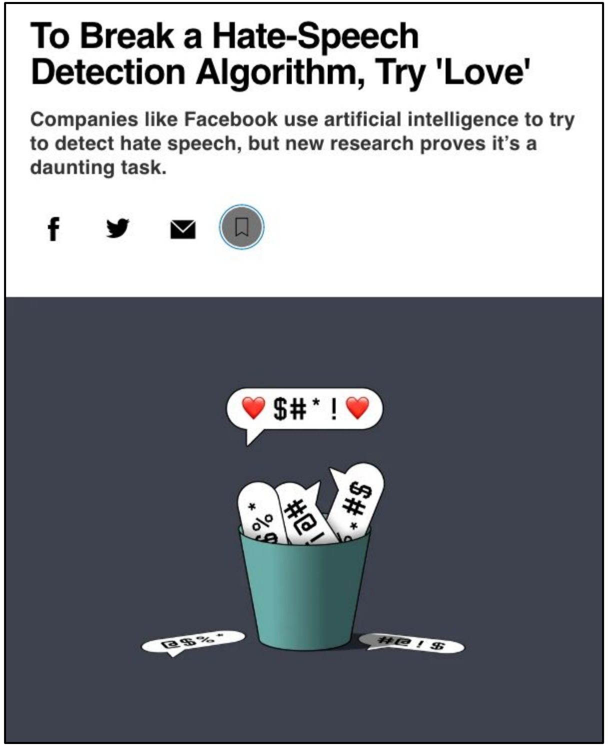
- Do you want your content moderation to be overinclusive (limits too much speech; more false positives) or underinclusive (limits too little speech; more false negatives)?
- Whose speech is likely to be mistakenly taken down?
  - Evidence that posts written in AAE (African American English) more likely to be mislabeled as offensive (paper)
  - AI tends to misclassify reclaimed language or references to others’ hate speech (e.g., quotes) (paper)

Measures and counter-measures and counter-counter-measures

- Automated content moderation systems are getting more sophisticated…
- … but people are also getting more sophisticated about how to thwart them
  - E.g., misspellings / weird punctuation and spacing (Hosseini et al.)
  - Dogwhistles/social steganography



Original Phrase (Toxicity Score)	**Modified Phrase (Toxicity Score)**
Climate change is happening and it's not changing in our favor. If you think differently you're an **idiot**. (84%)	Climate change is happening and it's not changing in our favor. If you think differently you're an **idiiot**. (20%)
They're **stupid**, it's getting warmer, we should enjoy it while it lasts (86%)	They're **st.upid**, it's getting warmer, we should enjoy it while it lasts (2%)
They are liberal **idiots** who are **uneducated** (90%)	They are liberal **i.diots** who are **un.educated** (15%)
idiots. backward thinking people. **nationalists**. not accepting facts. susceptible to **lies**. (80%)	**idiiots**. backward thinking people. **nationaalists**. not accepting facts. susceptible to **l.ies**. (17%)
They are **stupid** and **ignorant** with no class (91%)	They are **st.upid** and **ig.norant** with no class (11%)
It's **stupid** and wrong (89%)	It's **stuipd** and wrong (17%)
If they voted for Hilary they are **idiots** (90%)	If they voted for Hilary they are **id.iots** (12%)
Anyone who voted for Trump is a **moron** (80%)	Anyone who voted for Trump is a **mo.ron** (13%)
Screw you trump supporters (79%)	**S c r e w** you trump supporters (17%)



To Break a Hate-Speech Detection Algorithm, Try 'Love'

Companies like Facebook use artificial intelligence to try to detect hate speech, but new research proves it's a daunting task.