Ethics, Public Policy, and Technological Change

Rob Reich
Mehran Sahami
Head TA: Roberta Fischli

Housekeeping

- WIM Students (CS182W) only: The deadline for the WIM revision of the philosophy paper is revised to **March 3rd at 11:59pm PST**
  - If you haven't met with a Technical Communication Program (TCP) tutor for the philosophy paper assignment yet, please make sure to sign up for a session before the revision deadline
  - This is a requirement for WIM credit

Today's Agenda

- Value trade-offs in free speech
- Methods and challenges in content moderation
- When regulation and technology collide
  - Revisiting CDA 230
  - Platform interoperability
  - Anti-trust: what's old is new again

Why Free Speech?

Recall, a commitment to free speech protects:

- Minority groups from the tyranny of majority opinion
- Eccentric/idiosyncratic ideas from conventional wisdom
- Individuals to pursue their own interests, exercise their own conscience, regardless of social norms or government truths.

It leaves space for people to counter any official orthodoxy.

Comm. Decency Act, Section 230

"No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."

"No provider or user of an interactive computer service shall be held liable on account of--
`(A) any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected"

- Electronic Frontier Foundation: "The Most Important Law Protecting Internet Free Speech"

Today's Agenda

- Value trade-offs in free speech
- **Methods and challenges in content moderation**
- When regulation and technology collide
  - Revisiting CDA 230
  - Platform interoperability
  - Anti-trust: what's old is new again

Examining Incentives

- Consider: why do spam, fake news, etc. exist?
  - People engage with it
- Content-provider incentives
  - More content engagement = more monetization and/or influence
  - Produce content that users are likely to agree with or find titillating
  - Repeatedly modify content until it passes filter
- Platform incentives
  - More engagement = more monetization
    - Maximize click-through rates (*you get what you measure)*
  - Promote more eye-catching and extreme content
  - How to balance short-term and long-term interests?
    - You're grappling with this on Assignment #3

The Reality of Content Moderation

- Human moderators
  - Requires large number (many thousands) of moderators
  - Try to abide by guidelines, but can have different individual standards for objectionable content
  - Psychological toll on moderators
    - Some reported PTSD-like symptoms
    - Many moderators paid under minimum wage

TECHNOLOGY

In Settlement, Facebook To Pay $52 Million To Content Moderators With PTSD

MAY 12, 2020 · 10:52 PM ET

Bobby Allyn

TECH

Tech layoffs ravage the teams that fight online misinformation and hate speech

PUBLISHED FRI, MAY 26 2023•9:00 AM EDT | UPDATED SAT, MAY 27 2023•7:02 AM EDT

Hayden Field
@HAYDENFIELD
Jonathan Vanian
@IN/JONATHAN-VANIAN-B704432/

KEY POINTS

- Meta, Amazon, Alphabet and Twitter have all drastically reduced the size of their teams focused on internet trust and safety as well as ethics as the companies focus on cost cuts.
- As part of Meta's mass layoffs, the company ended a fact-checking project that had taken half a year to build, according to people familiar with the matter.

The Reality of Content Moderation

- Human moderators
  - Requires large number (many thousands) of moderators
  - Try to abide by guidelines, but can have different individual standards for objectionable content
  - Psychological toll on moderators
    - Some reported PTSD-like symptoms
    - Many moderators paid under minimum wage
- Automated filtering
  - More easily scales to large quantities of content
  - More uniform filtering standard
  - Difficult to match human accuracy
    - Context is often critical for content moderation decisions

Automated Content Moderation

- Example of the challenges in automated content moderation
  - Consider a piece of content with the word “breast”
  - How would we rate this?
  - “Chicken breast” → Rating: G
  - “Breast cancer” → Rating: PG
  - “Teen breast” → *It’s complicated*
    - “Teen breast development” → Rating: R
    - “Sexy teen breast” → Rating: X
    - “Underage sexy teen breast” → Likely illegal
- Lots of context may be needed for accurate classification
- Story time: *There are certain combinations of words I never want to see together again*

It’s Hard for Machines

Multi-Modal Challenges

- Do we need external knowledge to understand multi-modal content?
- Do we need new types of representations?

Source: A. Halevy *et al*, *Preserving Integrity in Online Social Networks*, Communications of the ACM, Feb. 2022, 65(2).

It’s Also Hard for Humans

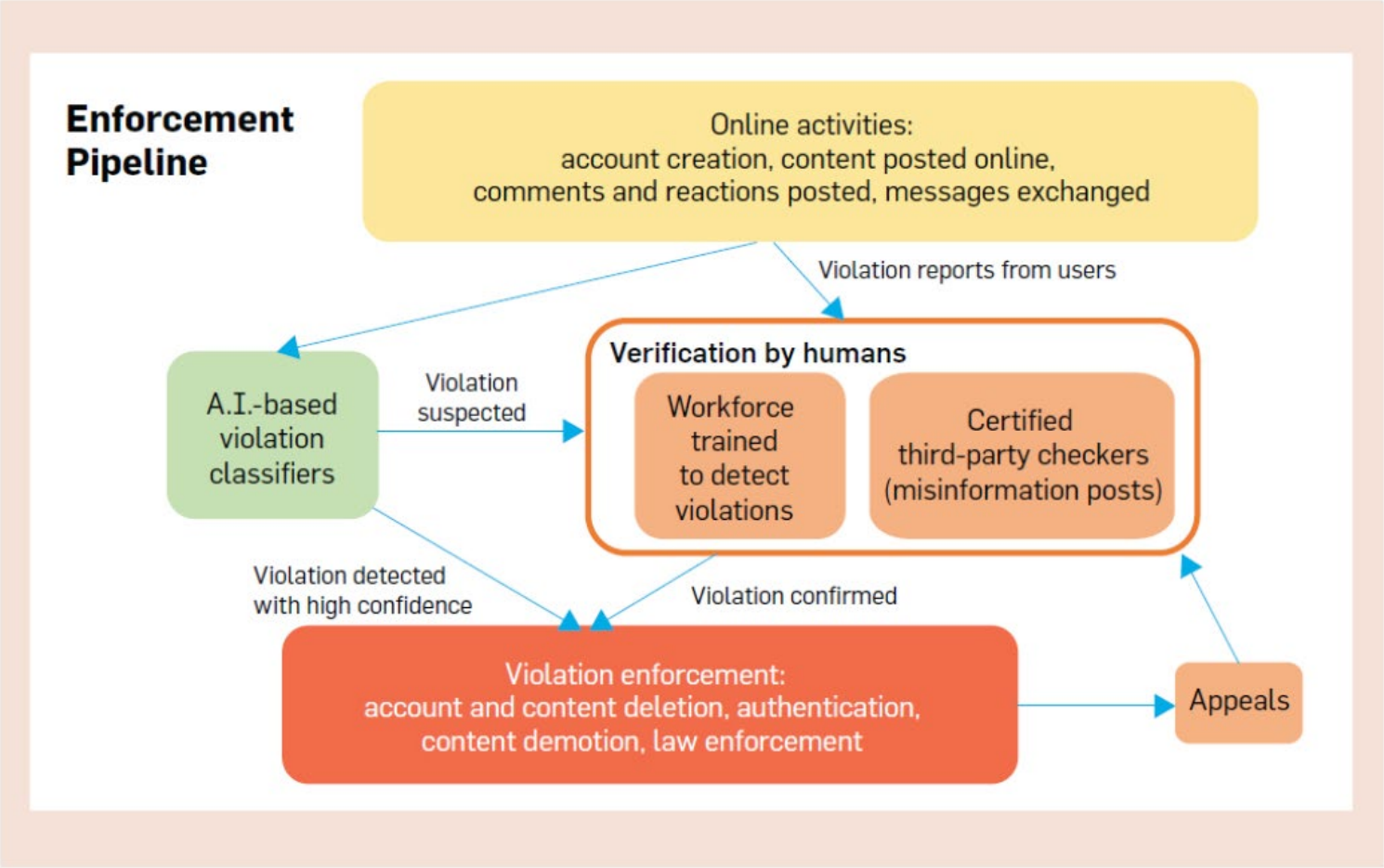
Computer vision: broccoli vs. marijuana
(sometimes better than humans!)

Broccoli (fried)

Marijuana

Source: A. Halevy *et al*, *Preserving Integrity in Online Social Networks*, Communications of the ACM, Feb. 2022, 65(2).

Human/Machine Hybrid Systems



Source: A. Halevy *et al*, *Preserving Integrity in Online Social Networks*, Communications of the ACM, Feb. 2022, 65(2).

Types of Content Moderation

- Flag questionable/misleading content
  - Tags can be ignored (mixed results due to confirmation bias)
- Down-rank or limit/slow distribution of objectionable content
  - Content is still available
  - Reactions/reposting by viewers can potentially up-rank/spread content
- Delete objectionable content
  - Modified versions (that get through filters) can posted
  - May take time before content is detected/deleted
- Sequester user/content (i.e., shadow banning)
  - Poster sees content and believes it is posted normally
  - Content not seen by other users in the system
  - Lack of response aims to discourage objectionable content
- Suspend or expel user (content poster) from platform
  - More effective for well-known personalities (e.g., Donald Trump)
  - Can come back under a different username or be reinstated

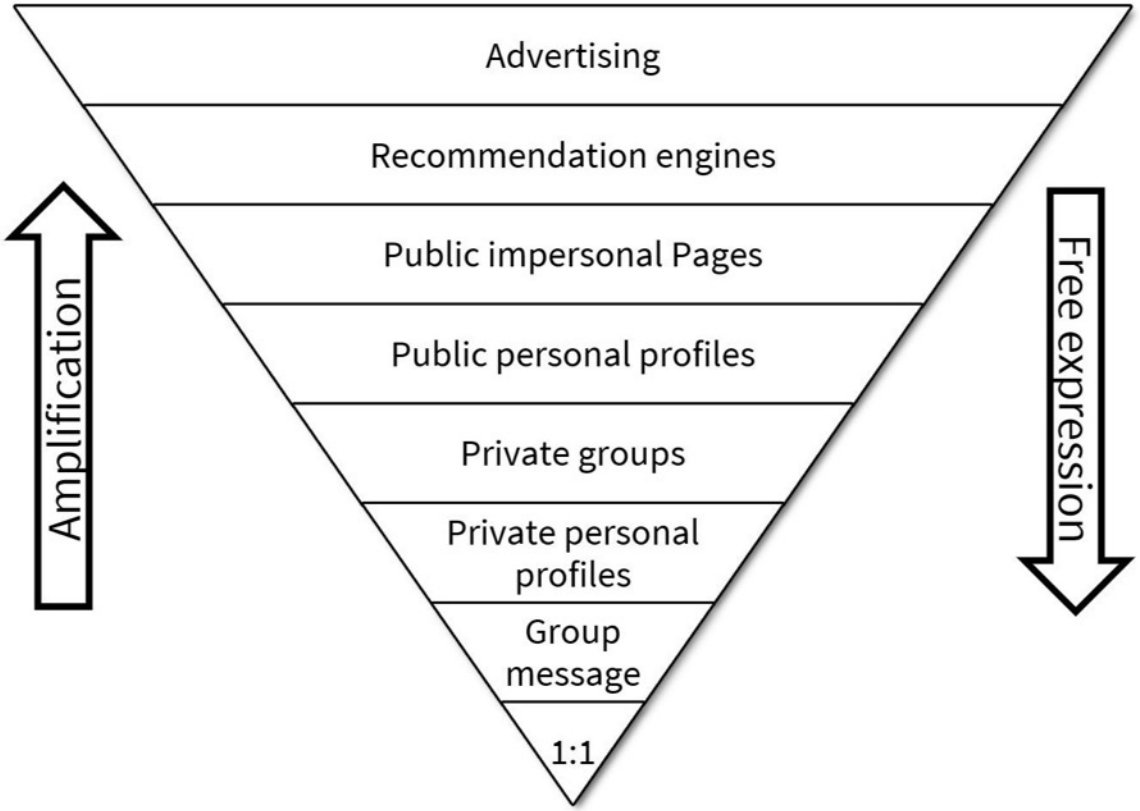
Freedom Speech Does Not Mean Freedom of Reach

Distinguish between the core principle of **freedom of expression** and the idea that this entitles any person to **algorithmic amplification**.

Question: Does downranking or de-platforming undermine freedom of speech?

Different Impact, Different Standards

- Could use different standards/methods for moderation based on level of amplification provided by platform/application

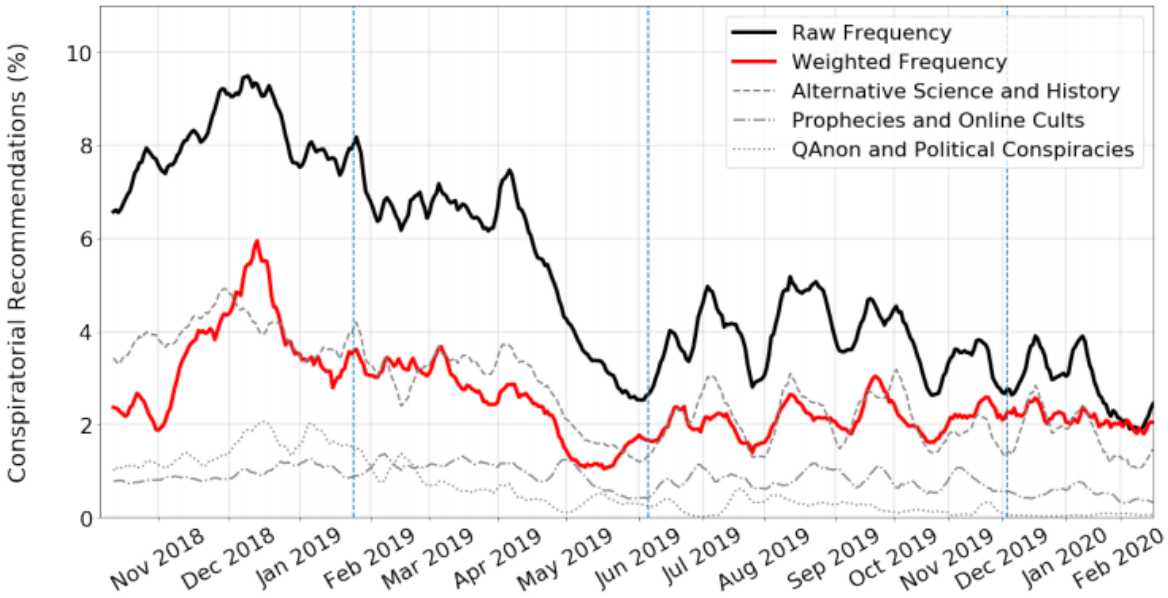


Source: Courtesy of Alex Stamos

Analysis of YouTube Video Promotion

- "A longitudinal analysis of YouTube's promotion of conspiracy videos" by Faddoul, Chaslot, and Farid (March 2020)
  - E.g., the moon landing was faked, the pyramids of Giza were built by aliens, end of the world prophecies, etc.
  - "Overall reduction of conspiratorial recommendations is an encouraging trend. Nonetheless, this reduction does not make the problem of radicalization on YouTube obsolete nor fictional, as some have claimed."

- **Raw Frequency** is number of times a video was recommended multiplied by the probability that each video is conspiratorial.
- **Weighted Frequency** is computed by weighting the Raw Frequency by the number of views of the source video.
- **Dotted vertical lines** represent 3 YouTube announcements about fighting conspiratorial content.



Source: "A longitudinal analysis of YouTube's promotion of conspiracy videos" by Faddoul, Chaslot, and Farid (March 2020)

Content Moderation as a Process

- Pitfall of “paralysis by analysis”
  - Outside of clearly illegal content, what should guidelines for content be?
- Creating culture of documentation
  - Explicitly documenting decisions that go into code/models
  - Often “code/model decides”, but important to explicitly understand why
- Iteration as a way to improve
  - There may not be one right decision
  - Get feedback from community as to what is acceptable
  - Careful to balance principles vs. “tyranny of the masses”
  - Content generators will respond
    - E.g., spammers modify content to get through filters
    - It’s an on-going “arms race”
    - Generative AI is latest tool/weapon in this battle
    - Open question as to whether platforms will have CDA 230 protection for content produced by their own generative AI

Today's Agenda

- Value trade-offs in free speech
- Methods and challenges in content moderation
- **When regulation and technology collide**
  - Revisiting CDA 230
  - Platform interoperability
  - Anti-trust: what's old is new again

Recall, CDA 230

"No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."

"No provider or user of an interactive computer service shall be held liable on account of--
`(A) any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected"

Amending CDA 230: H.R. 2154

- "Protecting Americans from Dangerous Algorithms Act"
  - Introduced in House on March 23, 2021
  - Co-sponsored by Anna Eshoo in CA, 18th District (you're sitting in it)
  - Creates exception in immunity provided by CDA 230
- Allows legal case to be brought against computer service when content is *algorithmically amplified*:

  "the claim involves a case in which the interactive computer service used an **algorithm, model, or other computational process to rank, order, promote, recommend, amplify, or similarly alter the delivery or display of information** (including any text, image, audio, or video post, page, group, account, channel, or affiliation) provided to a user of the service if the information is directly relevant to the claim."

Source: https://www.congress.gov/bill/117th-congress/house-bill/2154/text

An Exception to the Exception

- H.R. 2154 provides caveat to notion of algorithmic amplification:

  "the requirement is not met if—

  (I) the information delivery or display is ranked, ordered, promoted, recommended, amplified, or similarly altered in a way that is obvious, understandable, and transparent to a reasonable user based only on the delivery or display of the information (without the need to reference the terms of service or any other agreement), including sorting information—

  (aa) chronologically or reverse chronologically;

  (bb) by average user rating or number of user reviews;

  (cc) alphabetically;

  (dd) randomly; and

  (ee) by views, downloads, or a similar usage metric; or

  (II) the algorithm, model, or other computational process is used for information a user specifically searches for."

- Seems to have "died in committee"

CDA 230 and Generative AI

- Does CDA 230 protect AI-generated text and images?
  - CDA 230 inoculates platforms from liability for content hosted on them
  - But now platforms aren't just hosting content, they are <u>generating</u> it
    - Example: Bing and Google's Search AI; ChatGPT

- So, should companies be held liable for the content they produce?

Justice Neil Gorsuch, Gonzalez *v* Google (quoted in VentureBeat, Feb. 21, 2023):

"Artificial intelligence generates poetry… It generates polemics today that would be content that goes beyond picking, choosing, analyzing or digesting content. ***And that is not protected.*** *Let's assume that's right. Then the question becomes, what do we do about recommendations?"*

Image screenshot from: Sharon Goldman, "Could Big Tech be liable for generative AI output? Hypothetically 'yes,' says Supreme Court justice", VentureBeat, Feb. 21, 2023. https://venturebeat.com/ai/could-big-tech-be-liable-for-generative-ai-output-hypothetically-yes-says-supreme-court-justice/

Platform or Publisher?

How should we think of Meta/Facebook, Twitter/X, YouTube, Instagram, and TikTok?

- As infrastructure that hosts content (and therefore have limited liability for content hosted)? Like the telephone company?
- Or are they publishers, responsible for the content in the way that a newspaper or television show would be?
- Neither! The core function of a platform is algorithmic sorting of content.

Here's the Problem…

- Without the legal immunity provided by Section 230, platforms would have to exercise strong judgment over everything users produce as if they were newspaper editors
  - This is difficult to do technically
  - This is very expensive to do (and large companies would be better able to do it, entrenching their power)
  - It would impose serious constraints on on-line speech

- Moreover, it's not clear how far liability regimes get us when it comes to misinformation and disinformation

Broader Views on Regulation

1. Treat as “platform utility” with common carriage/equal access to address gatekeeper power (e.g. railroads, “net neutrality”, potentially ban targeted advertising)

2. Introduce a “separations regime” to prevent dominant platforms from leveraging their platform in other business lines (e.g. Amazon selling own version of products)

3. Tackle accumulation of power/concentration by preventing mergers/acquisitions or breaking up major companies

4. Require interoperability between platforms to prevent monopolistic lock-in and consumer choice

Broader Views on Regulation

1. Treat as “platform utility” with common carriage/equal access to address gatekeeper power (e.g. railroads, “net neutrality”, potentially ban targeted advertising)

2. Introduce a “separations regime” to prevent dominant platforms from leveraging their platform in other business lines (e.g. Amazon selling apparel)

3. Tackle accumulation of power/concentration by preventing mergers/acquisitions or breaking up major companies

4. **Require interoperability between platforms to prevent monopolistic lock-in and consumer choice**

Platform Interoperability

- Define APIs that allow access to data within a social network
  - Access to friend connections, profile information, etc.

- Allows for easier migration between platforms
  - Your data/friends available across platforms
  - You choose the platform with policies that best align with your values

- Middleware for content moderation/algorithmic ranking (Fukuyama *et al*, 2020)
  - Allow user to choose from multiple “middleware” options for ranking/moderating content their receive

Long Live OpenSocial!

Last updated Dec 8, 2017

OpenSocial

The main aim of OpenSocial ↗ is to define a common API for social applications across multiple websites. Using standard JavaScript and HTML, developers can create applications that access a social network's friends and feeds. Applications that use the OpenSocial APIs ↗ can be embedded within a social network itself or access a site's social data from anywhere on the web. OpenSocial applications are based on Google gadgets technology, and gadgets have been standardised as part of the OpenSocial effort. The Atlassian Gadgets framework is based on version 0.8 of the gadget specification ↗ from OpenSocial, but does not currently support other parts of OpenSocial, such as the social data or web service APIs.

Source: https://developer.atlassian.com/server/framework/gadgets/opensocial/

OpenSocial is Dead!

WIRED SECURITY POLITICS GEAR THE BIG STORY BUSINESS MORE SIGN IN SUBSCRIBE

ADAM DUVANDER NOV 13, 2008 2:00 PM

Where is the OpenSocial Revolution?

Yahoo points out that OpenSocial is a year old. The collection of APIs is a write-once approach to bringing the Facebook platform to any social website. Developers have not clamoured to develop OpenSocial apps. What's the deal? While Google was the instigator of OpenSocial, it found many supporters in fellow Facebook competitors: MySpace, Orkut, Friendster, [...]

SAVE

OpenSocialYahoo points out that OpenSocial is a year old. The collection of APIs is a write-once approach to bringing the Facebook platform to any social website. Developers have not clamoured to develop OpenSocial apps. What's the deal?

While Google was the instigator of OpenSocial, it found many supporters in fellow Facebook competitors: MySpace, Orkut, Friendster, Hi5, and more.

Source: https://www.wired.com/2008/11/where-is-the-opensocial-revolution-/

Is Anti-Trust Really a Remedy?

- U.S. v. IBM
  - DOJ charged IBM in Jan. 1969 with monopolizing digital computer market
  - Trial began in May 1975
  - Case is dropped in 1982, dismissed as “without merit”
- U.S. v. Microsoft
  - Microsoft charged with monopolizing OS market to stifle competition
    - Specifically, bundling Internet Explorer with Windows
  - Trial began in May 1998. Microsoft ordered to be broken up in 2000.
    - Overturned after appeal
  - Case settled in 2001 with agreement by Microsoft to share APIs with third parties (and appointment of oversight panel)
  - Real impact: Microsoft did take care to avoid actions that could be perceived as anti-competitive
- *Story time: what happened to my online greeting cards?*

Let's Try That Again

Congressional Research Service
Informing the legislative debate since 1914

Legal Sidebar

Federal Court Endorses Behavioral Remedies, Rejects Structural Relief, in Google Search Antitrust Litigation

"In August 2024, the district court held that ***Google had unlawfully monopolized the markets for general search services and general search text ads.*** *The court concluded that Google had monopoly power in both markets based on its dominant market share and the presence of significant entry barriers."*

Source: https://www.congress.gov/crs-product/LSB11362

Proposed Remedies

The proposed final judgment, or PFJ, filed with the court seeks to end Google's illegal monopoly and restore competition in several ways. The PFJ ends Google's search distribution contracts and revenue sharing agreements by ***prohibiting Google from paying to be the initial default search engine on any phone, device, or browser****.* ***Google is also required to share its data and information****—unlawfully obtained through its monopoly power—with rivals to improve the competitive choices available to consumers. This data will be shared in a manner that safeguards personal privacy and security.*

…

Additionally, the PFJ seeks ***the divestiture of Chrome****, the Google browser through which a significant percentage of all Google searches are made.*

Source: https://www.oag.state.va.us/media-center/news-releases/2807-november-21-2024-virginia-justice-department-seek-limits-to-google-business-practices-to-end-search-engine-monopoly

What Actually Happened

"On September 2, 2025, the U.S. District Court for the District of Columbia released its remedies decision in United States v. Google, an antitrust case involving Google's conduct in certain markets related to online search and search advertising.

In the decision, ***the court rejected the plaintiffs' proposals for an immediate divestiture of Google's Chrome web browser and a contingent divestiture of the Android operating system****, but ruled that* ***Google will be subject to several behavioral remedies****, including a prohibition of exclusive contracts relating to the distribution of Google Search, the Chrome browser, and certain artificial-intelligence (AI) products."*

Source: https://www.congress.gov/crs-product/LSB11362

But We're Not Done Yet

"Google has appealed a US district judge's landmark antitrust ruling that found the company illegally held a monopoly in online search.

…

In its announcement on Friday, Google said the ruling by Judge Amit Mehta didn't account for the pace of innovation and intense competition the company faces.

…

The company is requesting a pause on implementing a series of fixes - viewed by some observers as too lenient - aimed at limiting its monopoly power."

Source: https://www.bbc.com/news/articles/clyn0ek5rdpo

Two Can Play That Game


PYMNTS

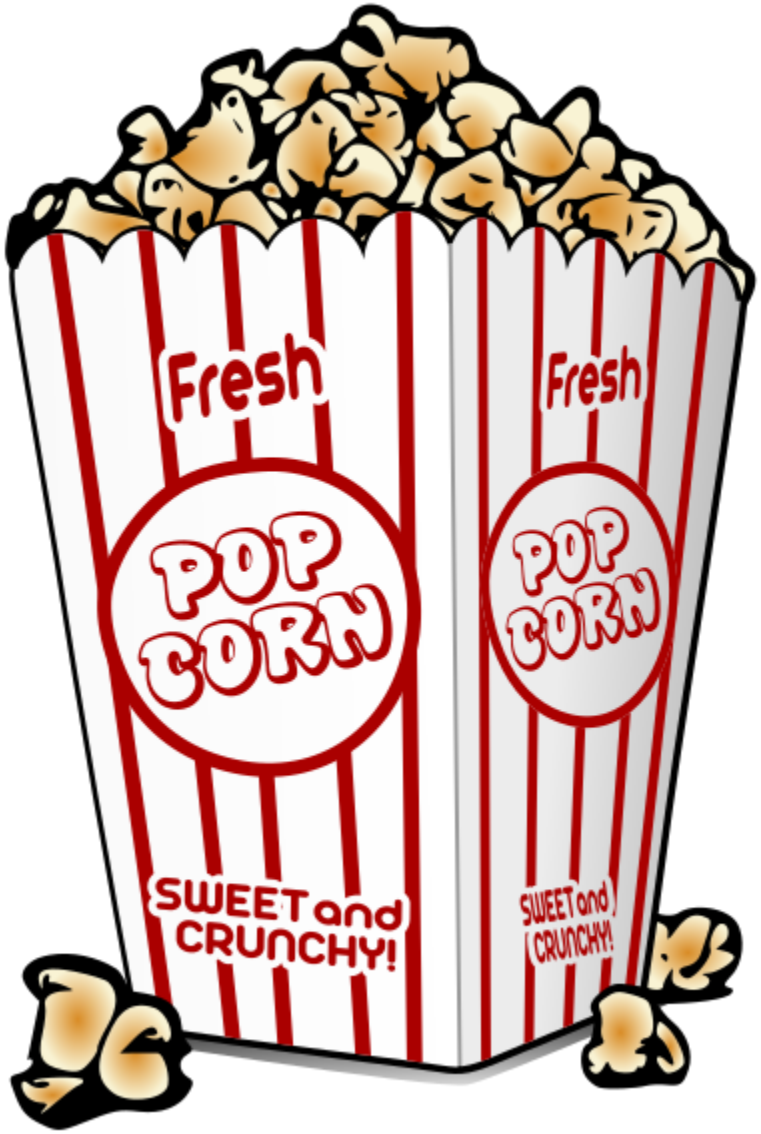
Justice Department to Appeal Ruling in Google Search Antitrust Case

BY **PYMNTS** | FEBRUARY 3, 2026

"The Justice Department and 35 states will appeal a September 2025 court ruling that allowed Google parent company Alphabet to keep its Chrome browser after losing an antitrust case.

The plaintiffs said Tuesday (Feb. 3) in court papers that they will appeal the ruling, Reuters reported Tuesday."

Source: https://www.pymnts.com/google/2026/justice-department-to-appeal-ruling-in-google-search-antitrust-case/


Fresh
POP CORN
SWEET and CRUNCHY!
Fresh
POP CORN
SWEET and CRUNCHY!