SMU Singapore Management University
School of Computing and Information Systems

2021-22 Term 1

IS457: Fairness in Socio-technical Systems

Week 4 - Auditing algorithms

KWAK Haewoon

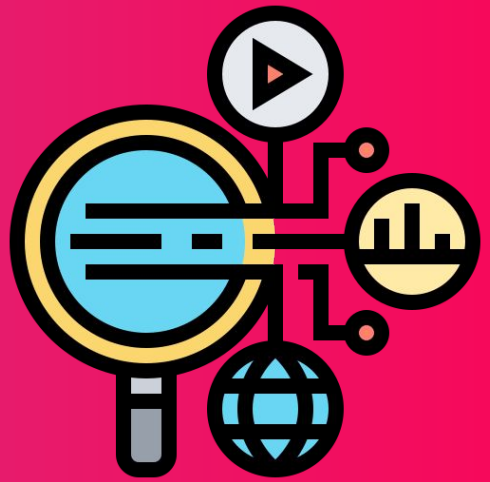


Image from Flaticon.com

Study questions

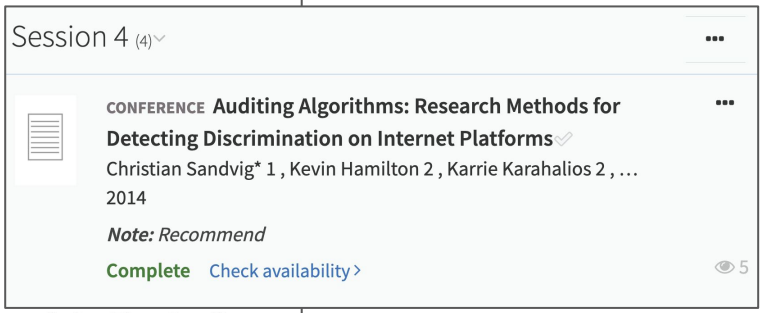
What is an “algorithm audit”?

What types of auditing techniques are there?

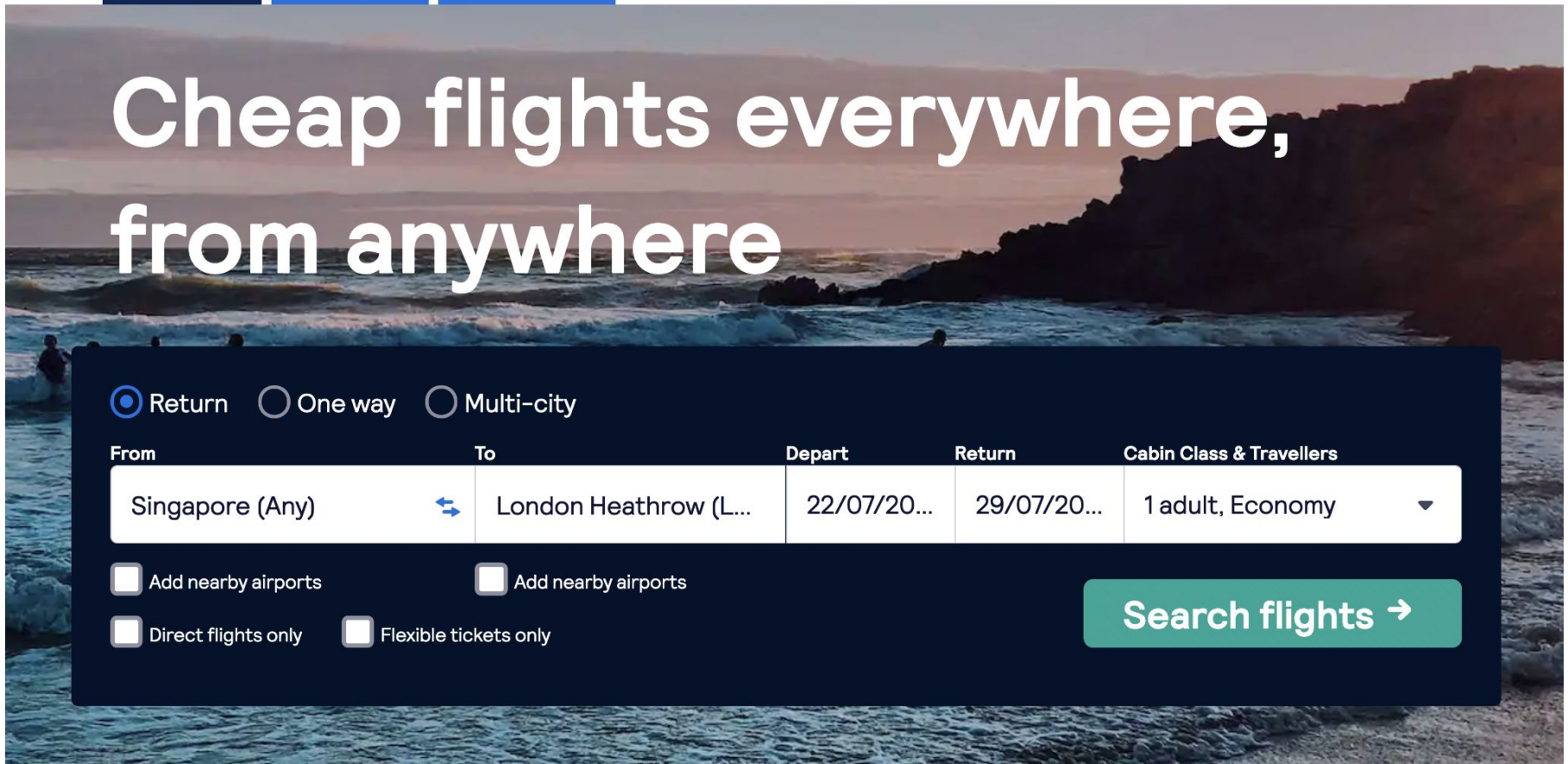
What are the characteristics of each auditing technique?

Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms

Christian Sandvig*[1], Kevin Hamilton[2], Karrie Karahalios[2], & Cedric Langbort[2]



Paper presented to "*Data and Discrimination: Converting Critical Concerns into Productive Inquiry*," a preconference at the 64th Annual Meeting of the International Communication Association. May 22, 2014; Seattle, WA, USA.


Skyscanner
Help
English (UK)
Singapore
$ SGD
Log in
Flights
Hotels
Car Hire
Cheap flights everywhere, from anywhere
Return
One way
Multi-city
From
To
Depart
Return
Cabin Class & Travellers
Singapore (Any)
London Heathrow (L...
22/07/20...
29/07/20...
1 adult, Economy
Add nearby airports
Add nearby airports
Direct flights only
Flexible tickets only
Search flights

Early days in the travel industry

Airline agents took reservations in person at airports or ticket offices, or on the phone, where they hand-wrote cards.

https://www.sabre.com/files/Sabre-History-rev2017.pdf
https://www.wired.com/2012/07/sabre/

SABRE (early 60s) - Computer Reservations Systems

Semi-**A**utomated **B**usiness **R**esearch **E**nvironment by American Airlines-IBM

SABRE connected 1,500 terminals across the U.S. and Canada. The hardware alone cost about $30 million -- or $230 million in 2012 dollars

Impact of SABRE on the industry

The average time to process a reservation: 90 minutes → a few seconds.

Instant and accurate update to seat inventory and passenger information becomes possible.

Other airlines adopted SABRE, too.

First page, first flight

Over 90% of tickets are sold for flights listed on the first screen of the CRS display, and over 50% for the first flight listed.

Beliefs about “unbiased” systems existed at this time.

"The Legal and Regulatory Implications of Airline Computer Reservation Systems." Harvard Law Review 103, no. 8 (1990): 1930-950. Accessed August 8, 2021. doi:10.2307/1341325.

Bias in CRS and antitrust investigations

Travel agents and competitors noticed that the first flight returned was often an AA flight that was much longer and more expensive than other alternatives.

The US Civil Aeronautics Board and the Department of Justice launched antitrust investigations.

Various ways injecting biases in CRS

Choose criteria for the display algorithm that match distinctive characteristics of its own flights, such as connecting points and nonstop service.

Fare information is omitted or delayed.

- Special deals from smaller carriers may not timely appear on the CRS display

Misinformation is displayed.

- Full flights would show up as available and flights with unsold seats would show up as full.

AA didn’t deny the manipulation

Speaking before the US Congress, the president of AA, Robert L. Crandall, boldly declared that biasing SABRE’s search results was in fact his primary aim.

He testified that “the preferential display of our flights, and the corresponding increase in our market share, is the competitive raison d’etre for having created the [SABRE] system in the first place.”

Crandall’s complaint: Why would you build and operate an expensive algorithm if you can’t bias it in your favor?

A new regulation about CRS was made (1984)

Regulation 15 CFR 255.4:

“Each [airline reservation] system shall provide to any person upon request the current criteria used in editing and ordering flights for the integrated displays and the weight given to each criterion and the specifications used by the system’s programmers in constructing the algorithm.”

Any algorithms may deserve scrutiny

Algorithms that appear to work well may still be dangerous.

Algorithms that do not satisfy their users would be unlikely to continue operation for very long, but algorithms can both satisfy their users and achieve other goals simultaneously.

Algorithms can be manipulated in ways that do not disadvantage their users directly or obviously.

Scrutiny of algorithms

An individual try of a target algorithm may detect some forms of harm.

But, a complete picture of the biases can be uncovered only via a systematic investigation.

"and so this men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right
And all were in the wrong!"

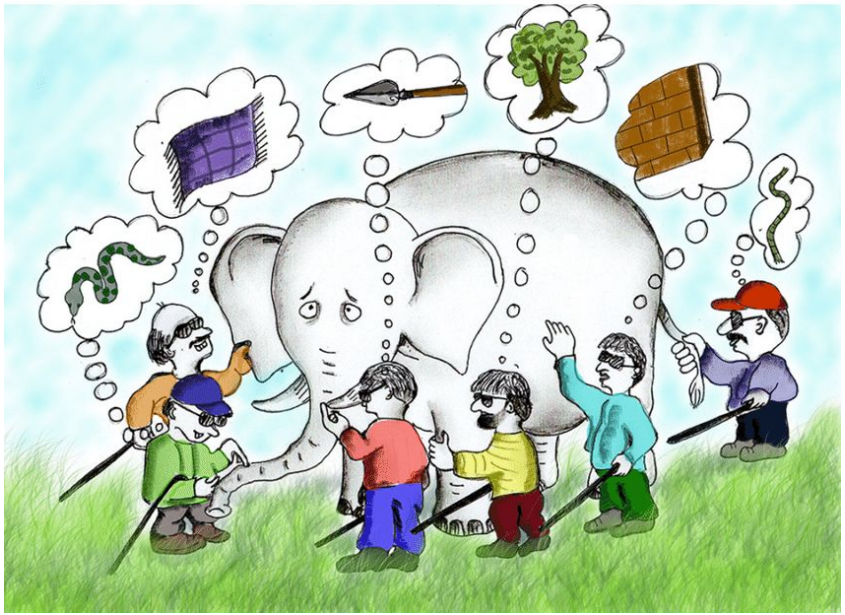
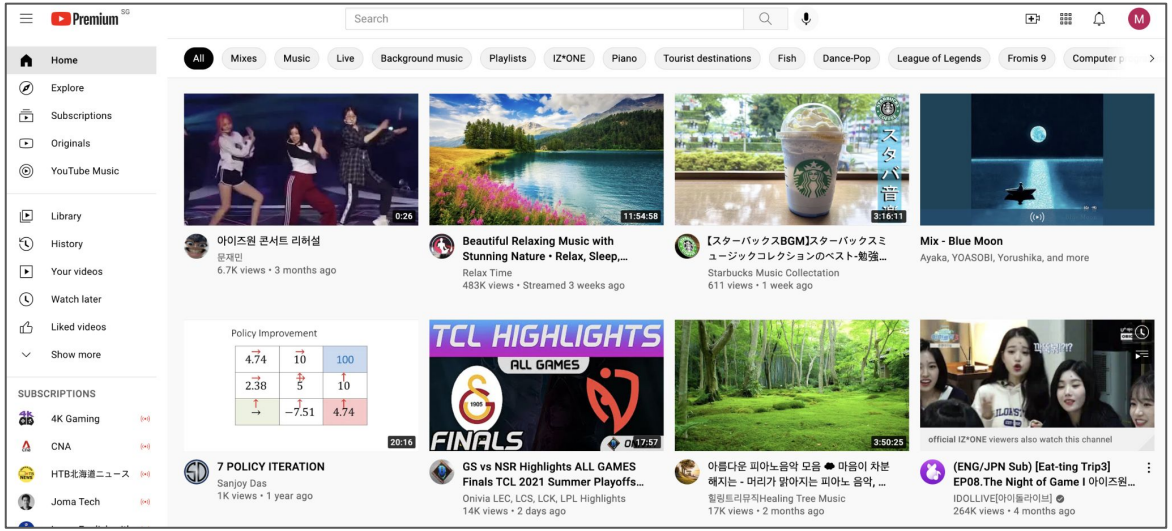


Image from https://www.researchgate.net/publication/320236661_Protein_Hydration_Dynamics_Much_Ado_about_Nothing

“Personalized” algorithms

Personalized algorithms make the problem complex.

- Individual investigations are unlikely to produce a broad picture of the system’s operation across the diversity of its users.



“Audit” study in social science

Field experiments to diagnose harmful discrimination in hiring.

Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

By Marianne Bertrand and Sendhil Mullainathan*

We study race in the labor market by sending fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perceived race, resumes are randomly assigned African-American- or White-sounding names. White names receive 50 percent more callbacks for interviews. Callbacks are also more responsive to resume quality for White names than for African-American ones. The racial gap is uniform across occupation, industry, and employer size. We also find little evidence that employers are inferring social class from the names. Differential treatment by race still appears to still be prominent in the U.S. labor market. (*JEL* J71, J64).

Bertrand, Marianne, and Sendhil Mullainathan. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." American economic review 94.4 (2004): 991-1013.

Experiment design

Collect employment ads from Chicago and Boston newspapers for > 6 months.

Create a bank of realistic and representative resumes.

Send 4 randomly selected resumes for each ad:

- Two higher-quality resumes (e.g., more experience, certificate degrees, etc.)
- Two lower-quality resumes

Randomly assign an African-American-sounding name to one HQR and one LQR.

Send 5k resumes for 1,300 employment ads in the sales, administrative support, clerical, and customer services job categories.

Measure callback rates.

Lower callback rates for African-American names

	Percent callback for White names	Percent callback for African-American names	Ratio	Percent difference (p-value)
Sample:				
All sent resumes	9.65	6.45	1.50	3.20
	[2,435]	[2,435]		(0.0000)
Chicago	8.06	5.40	1.49	2.66
	[1,352]	[1,352]		(0.0057)
Boston	11.63	7.76	1.50	4.05
	[1,083]	[1,083]		(0.0023)
Females	9.89	6.63	1.49	3.26
	[1,860]	[1,886]		(0.0003)
Females in administrative jobs	10.46	6.55	1.60	3.91
	[1,358]	[1,359]		(0.0003)
Females in sales jobs	8.37	6.83	1.22	1.54
	[502]	[527]		(0.3523)
Males	8.87	5.83	1.52	3.04
	[575]	[549]		(0.0513)

	Low	High	Ratio	Difference (p- value)
White names	7.18	13.60	1.89	6.42
	[822]	[816]		(0.0000)
African-American names	5.37	8.60	1.60	3.23
	[819]	[814]		(0.0104)

Audit study as a field experiment

Realistic settings and situations are used in order to ensure that the results are **generalizable** to real-world experience.

- Research design involves the random assignment to groups in a controlled setting.

Ethical challenges in the audit study

Audit studies waste the time of people who participate in them.

Audit study typically intends to show that participants in it are immoral, or even that they are criminals.

Participants **must not know** they are being audited, violating the basic principles for the ethical conduct of science.

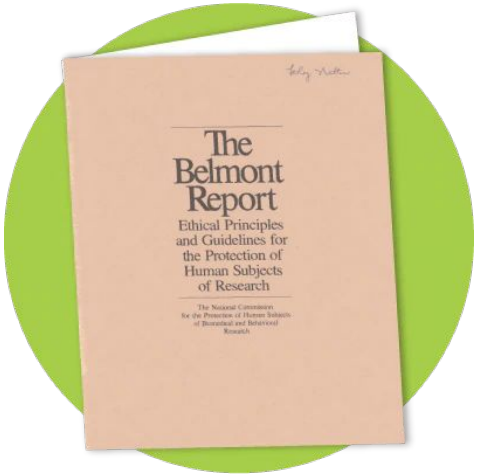
The Belmont Report (1976) emphasizes

Normal ethical practices (such as consent) may need to be set aside for the greater good in specific research studies.

When the evidence produced by audit studies is so important to society, researchers must be allowed:

- to waste a certain amount of other people's time
- to operate without informed consent

in order to secure meaningful evidence about these important social problems and crimes.



https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html
https://smu.sharepoint.com/sites/irb/ (SMU's Institutional Review Board website)

Consumer complaint study (2001) - Failed

Frank Flynn, a Columbia business school professor, mailed 240 restaurant owners a fabricated letter claiming he had got food poisoning at the restaurant.

...

A formal letter of apology from the dean of the business school, Meyer Feldberg, and two letters from Flynn were sent to all the owners later.


The New York Times

Scholar Sets Off Gastronomic False Alarm

https://www.nytimes.com/2001/09/08/nyregion/scholar-sets-off-gastronomic-false-alarm.html

Algorithm audit

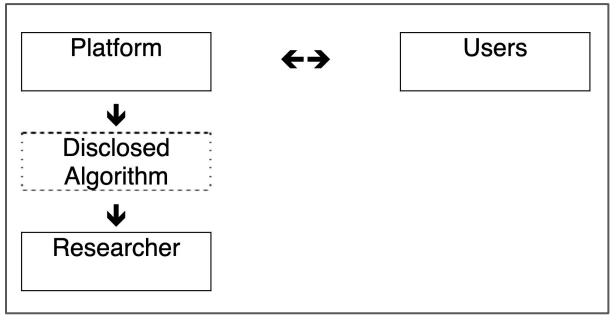
Investigates biases/discrimination led by algorithms in Internet platforms.

Primary differences from traditional audit (in social science):

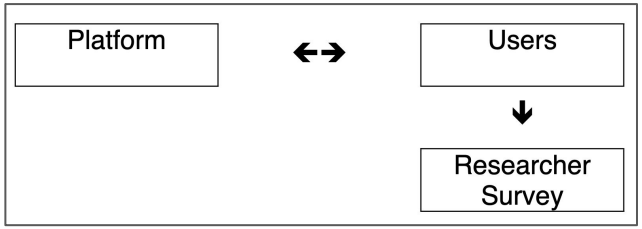
- Field experiment audit is typically designed to target a societal phenomenon.
- An algorithm audit usually seeks to target a particular platform.

Five forms of algorithm audit

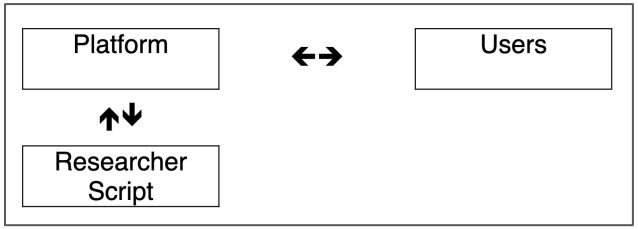
SMU Singapore Management University
School of Computing and Information Systems



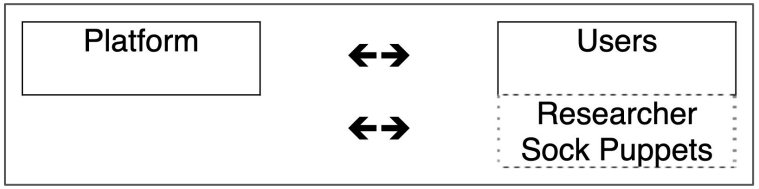
Code audit



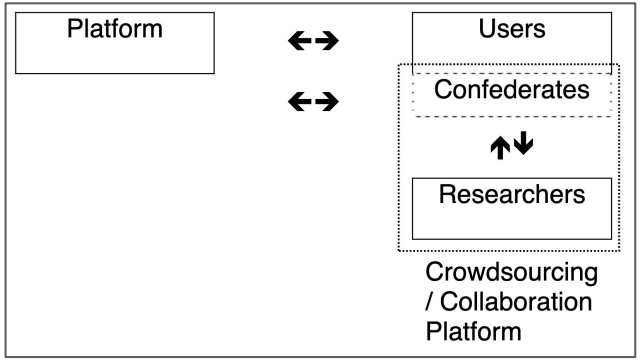
Noninvasive user audit



Scraping audit



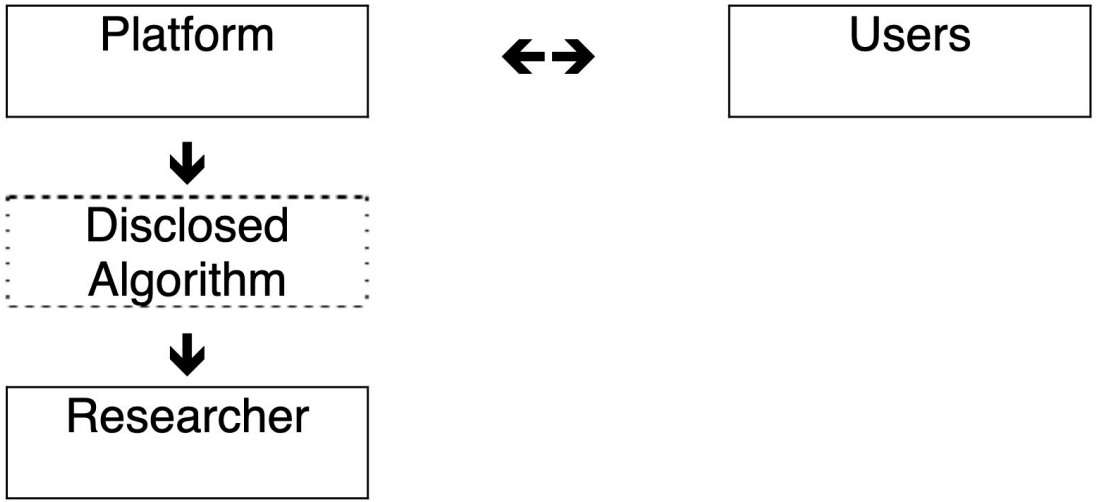
Sock puppet audit



Crowdsourced audit / collaborative audit

Code audit

If researchers could simply obtain a copy of the target algorithm, this could be a kind of algorithm audit.



Code audit - Considerations

Algorithms are typically **not available**:

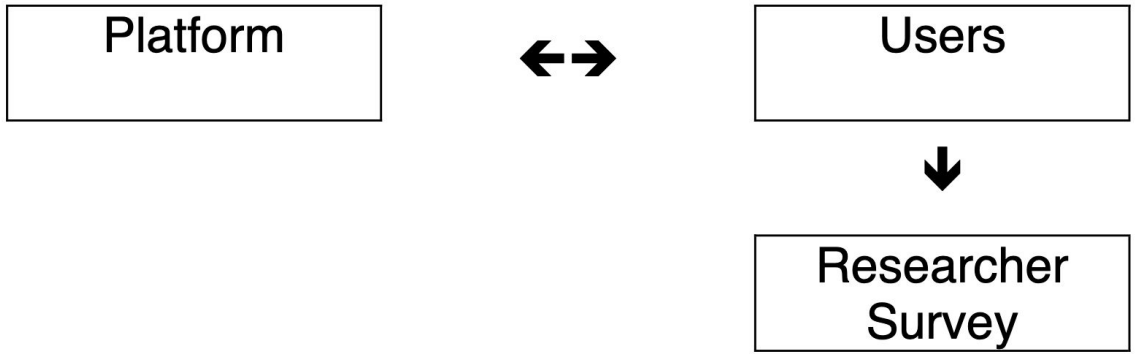
- Protected as *trade secret*.
- Public release of algorithms might produce negative consequences (abusers)

Some badly-behaving algorithms may produce their bad behavior only in the context of a particular dataset or application.

- Harmful discrimination could be modeled as a combination of an algorithm and its data, not as just the algorithm alone.

Noninvasive user audit

If users agreed to answer questions about what they did online or agreed to share all of their actions (e.g., search queries) and results, it might be possible to audit the operation of a platform's algorithm.



Noninvasive user audit - Considerations

Pros:

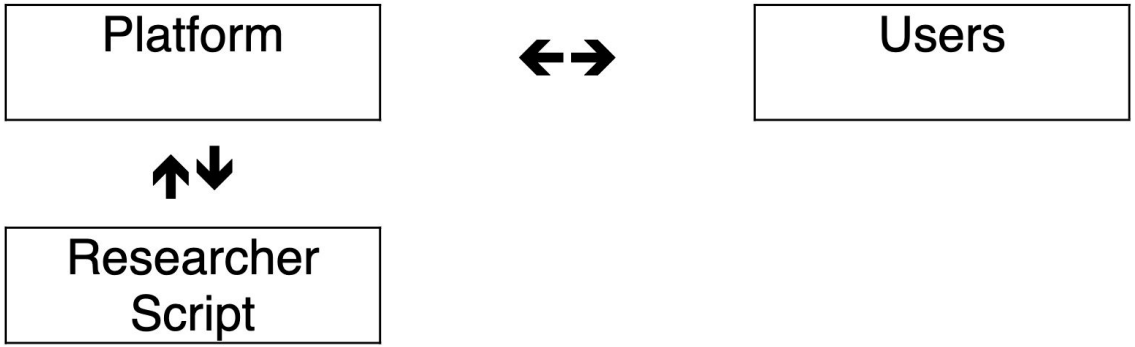
- Not perturbing the platform itself.

Cons:

- Difficult to infer causality from the observation results.
- Sampling problem.
- A survey-based audit is likely to have self-report biases.

Scraping audit

Researchers send repeated queries to a platform and observe the results.



Scraping audit - Considerations

Using APIs: Rate limiting (maximum # of requests / hour or day).

When there are no APIs: Check terms of service (TOS) of online platforms

- Sending repeated queries is likely to afoul of a platform's TOS.
- TOS often explicitly forbid the automatic downloading of any information from a web site, even if that information is public.

TOS banning automated scripts is common

"You may not use automated scripts to collect information from or otherwise interact with the Services"

"You can't attempt to [...] collect information in an automated way without our express permission."

Crawling against TOS?

HiQ Labs v. LinkedIn

From Wikipedia, the free encyclopedia

hiQ Labs, Inc. v. LinkedIn Corp, 938 F.3d 985 (9th Cir. 2019), was a United States Ninth Circuit case about web scraping. The 9th Circuit affirmed the district court's preliminary injunction, preventing LinkedIn from denying the plaintiff, hiQ Labs, from accessing LinkedIn's publicly available LinkedIn member profiles. hiQ is a small data analytics company that used automated bots to scrape information from public LinkedIn profiles.

2019

Data Scraping Survives! (At Least for Now) Key Takeaways from 9th Circuit Ruling on the HIQ vs. Linkedin Case

Monday, September 30, 2019

2021

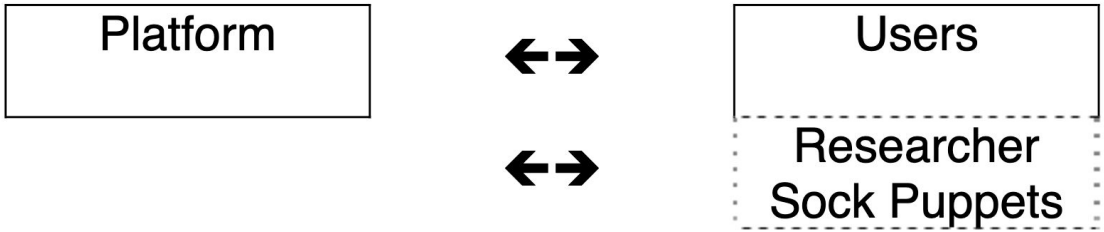
Supreme Court Vacates LinkedIn-HiQ Scraping Decision, Remands to Ninth Circuit for Another Look

Wednesday, June 16, 2021

https://www.natlawreview.com/article/data-scraping-survives-least-now-key-takeaways-9th-circuit-ruling-hiq-vs-linkedin
https://www.natlawreview.com/article/supreme-court-vacates-linkedin-hiq-scraping-decision-remands-to-ninth-circuit

Sock puppet audit

Researchers would use computer programs to impersonate users, likely by creating false user accounts or programmatically-constructed traffic.



Sock puppet audit - Considerations

Artificial user accounts can be

- used to investigate features of systems that are not public.
- assigned to important categories of discrimination that may be difficult to talk about (e.g., HIV seropositive status, sexual orientiation, poverty).

SPA involves **deception**: researchers are inventing false data and injecting it into the platform.

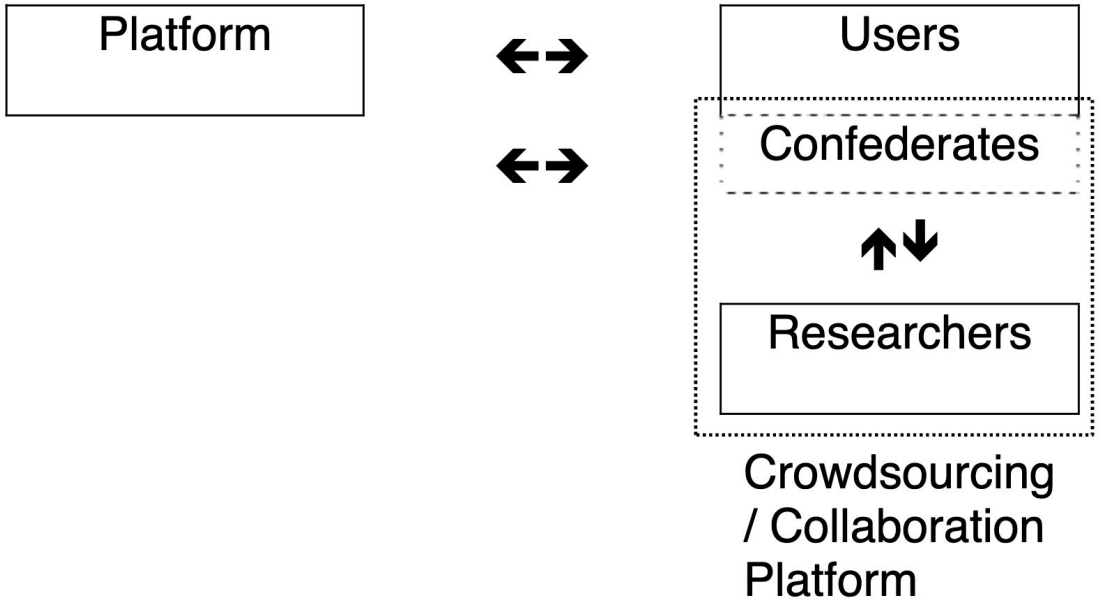
A large number of tests is expected for detecting misbehaving algorithms.

SPA: How TikTok's Algorithm Figures You Out

~4:16

Crowdsourced audit / Collaborative audit

Researchers recruit a large enough group of “testers” (users).



Crowdsourced audit - Considerations

Automated use of internet platforms by bots and computer programs (i.e., scraping audit or sock puppet audit) are usually prohibited in the Terms of Service given by platform providers.

But, **an actual human being** who is recruited to do something with an Internet platform is unlikely to trigger any of these prohibitions.

Sock puppet audit on price discrimination (2012)

Detecting price and search discrimination on the Internet

Jakub Mikians†, László Gyarmati⋆, Vijay Erramilli⋆, Nikolaos Laoutaris⋆
Universitat Politecnica de Catalunya†, ⋆Telefonica Research
jmikians@ac.upc.edu,{laszlo,vijay,nikos}@tid.es

Image from https://commons.wikimedia.org/wiki/File:Torre_de_hielo_en_Barcelona.JPG

Mikians, Jakub, et al. "Detecting price and search discrimination on the internet." Proceedings of the 11th ACM workshop on hot topics in networks. 2012.

What is price discrimination?

Price discrimination: customizing prices for some users

Price steering: changing the order of search results to highlight specific products

THE WALL STREET JOURNAL.

English Edition ▾ | Print Edition | Video | Podcasts | Latest Headlines

Home World U.S. Politics Economy Business **Tech** Markets Opinion Life & Arts Real Estate WS.

On Orbitz, Mac Users Steered to Pricier Hotels

Orbitz has found that Apple users spend as much as 30% more a night on hotels, so the online travel site is starting to show them different, and sometimes costlier, options than Windows visitors see. Dana Mattioli has details on The News Hub. Photo: Bloomberg.

https://www.wsj.com/articles/SB10001424052702304458604577488822667325882

Detecting price discrimination is not trivial

Which personal information are relevant and can cause or trigger discrimination?

How can information that is exposed while searching for price or search discrimination be controlled?

Information (types) to control

System-based information

- Different browsers (Chrome, Safari, Firefox, Internet Explorer)
- Different OSes (Windows, Mac OS, Linux)

Location-based information

- Deploy several proxy servers at 6 distinct sites (US East, US West, Germany, Spain, Korea, and Brazil)

Personal information

- Train personas: affluent customers vs. budget conscious customers
- Visit retail-jewelry/luxury goods/accessories sites vs. price aggregation sites

What does “visit websites” mean?

https://www.youtube.com/watch?v=8w4qPUSG17Y

Three types of information are enough?

Maybe not. What are other factors that can capture user's characteristics?

(+ Try https://browserleaks.com/)

Target products

35 product categories (e.g., clothing)

200 distinct vendors x 3 products per vendor

- low/mid/high price products per vendor.
- In case of hotels, three different dates (low/mid/high season) at multiple locations were selected.

Were there system-based differences in price?

Try different system-browser setups to get the price of 600 different products for four days from different vendors.

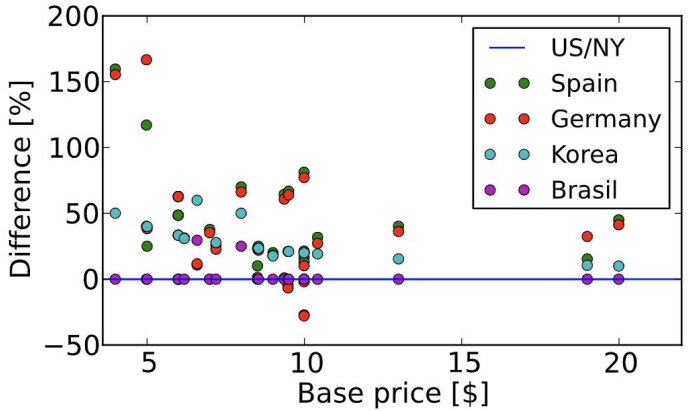
✓ No price differences between the system-browser setups.

Are there location-based differences in price?

Majority of the products do not show significant differences.

But, some show show a strong dependence on users' location

- Digital products: Amazon.com (e-book), Steampowered.com (games)

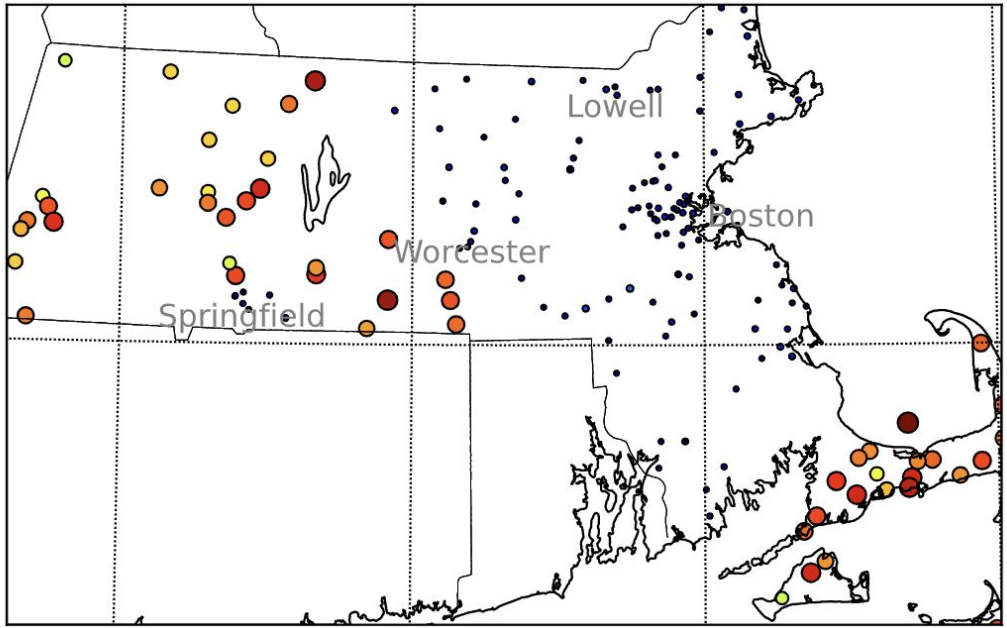


Results of 27 out of top 100 e-books

(max: 166%, med: 21%)

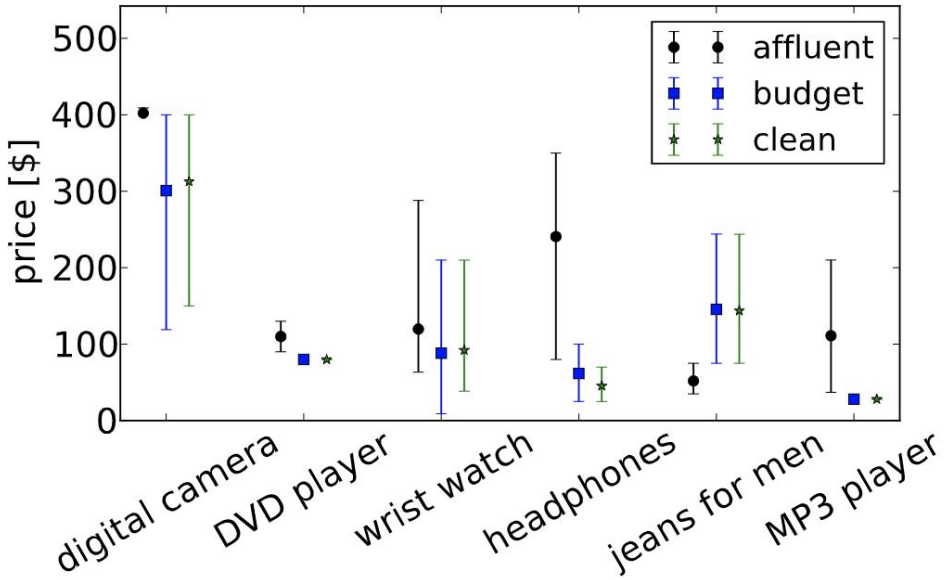
Price differences exist even within one state

Office products: staples.com (Price in city is higher than that in suburb)



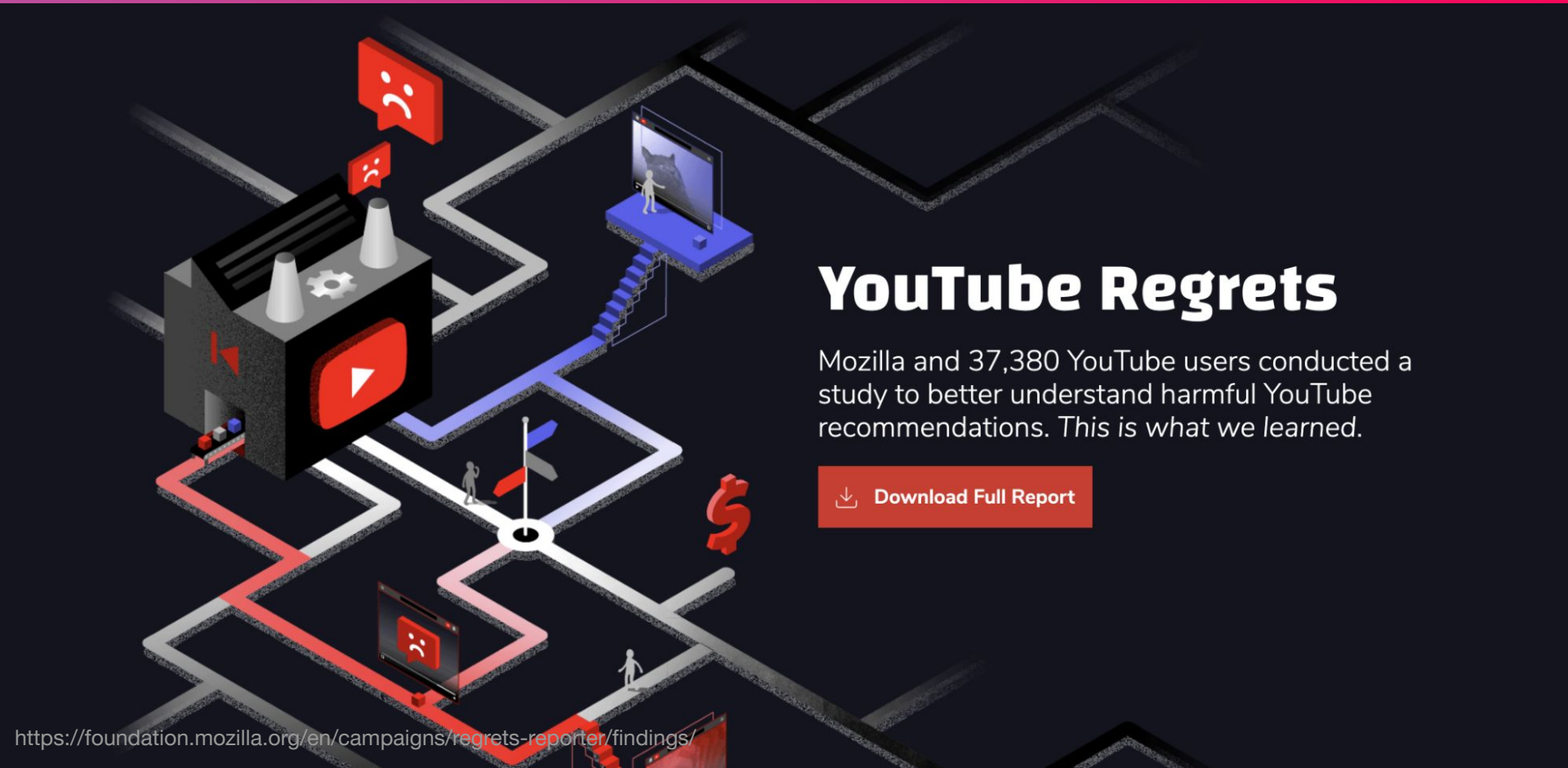
Are there persona-based differences in price?

No significant differences. But, Google is likely to recommend more expensive products for affluent customers (“price steering”).



Crowdsourced audit on Youtube recommendations

SMU
School of Computing and Information Systems



https://foundation.mozilla.org/en/campaigns/regrets-reporter/findings/

Opaque recommendation algorithms

YouTube's algorithm drives ≈700 million hours of watch time every single day.

But, the public knows very little about how it works; there are no official tools for studying it.

When a recommended video is harmful, there is no way to understand what happened and what could be changed to prevent the same thing from happening again.

#YouTubeRegrets campaign (2019)

Mozilla collected YouTube users' stories about the YouTube's recommendation engine leading to bizarre and sometimes dangerous contents.

01: A Deadly Fail

I started searching for "fail videos" where people fall or get a little hurt. I was then presented with a channel that showed dash cam videos from cars. At first it was minor accidents, but later it transitioned into cars blowing up and falling off bridges — videos where people clearly didn't survive the accident. I felt a little bit sick at that point, and haven't really sought out that type of content after that.

08: One Small Step for Conspiracies

I'm a teacher and I watched serious documentaries about Apollo 11. But YouTube's recommendations are now full of videos about conspiracy theories: about 9/11, Hitler's escape, alien seekers and anti-American propaganda.

https://foundation.mozilla.org/en/campaigns/youtube-regrets-about/

YouTube's responses to external research

SMU Singapore Management University | School of Computing and Information Systems

2 February 2018: You're d

The coverage: "Fiction is outr
algorithm disto
Chaslot & Grap

11 Augu
see that
The cove

25 July 2

27 September 2019: You're do

The coverage: "You
algorithm even

2 March, 2020: We're working on it 😉

The coverage: "A longitudinal analysis of YouTube's promotion of conspiracy videos" by researchers at University of California,
onspiracy Theorists?" by

12 May, 2021: Nothing... until someone noticed

The coverage: "A French coronavirus conspiracy video stayed on YouTube and Facebook for months" -- Politico

12 Fe
, trust us
age: "Youtube continues to push dangerous videos to
ceptible to extremism, white supremacy, report finds." --
uTube" --
) and

12 May, 2021: We're working on it

The coverage: "Youtube Kids has a rabbit hole problem" --Vox

6 April, 2021: We're worki

The coverage: "House panel c
as it opens investigation into a

13 April, 2021: We already 'fixed' this

The coverage: "Exploring YouTube And The Spread of
Disinformation" -- NPR

31
Th
propaga
Wash

2 June, 2021: No Comment

The coverage: "Senate Democrats urge Google to Investigate Racial Bias In Its Tools and The Company" -- NPR

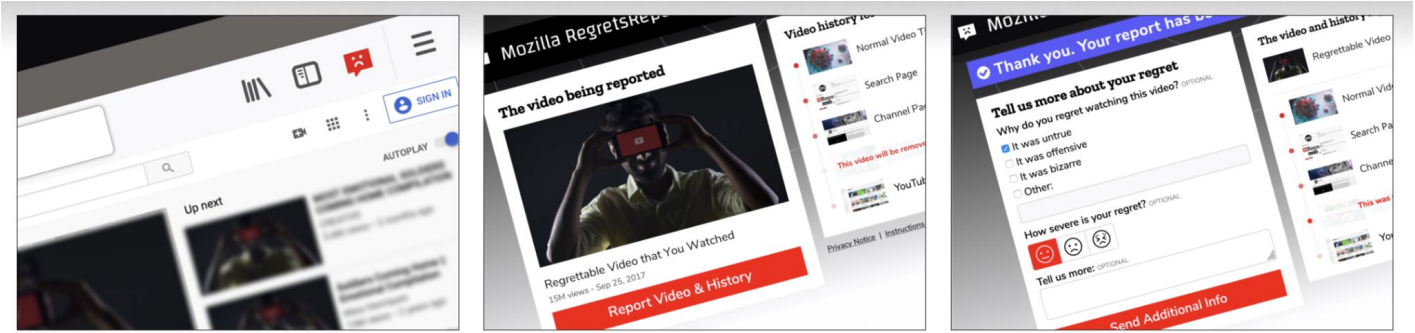
'YouTube Is A Pedophile's Paradise"--The
t

Compiled by Mozilla. https://foundation.mozilla.org/en/blog/congratulations-youtube-now-show-your-work/

RegretsReporter - Browser extension

SMU Singapore Management University
School of Computing and Information Systems

A crowdsourcing tool for scaling up the #YouTubeRegrets campaign by Mozilla.

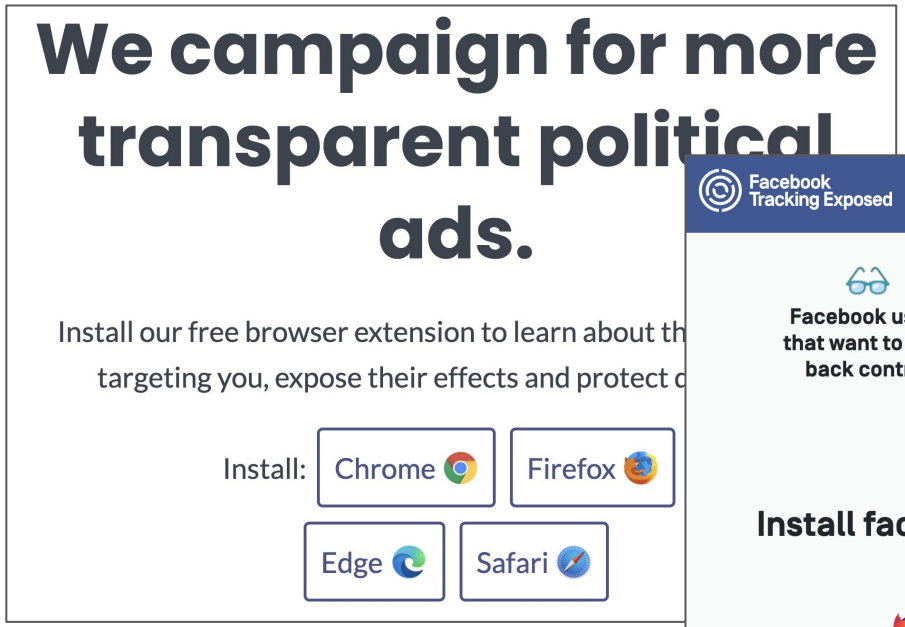


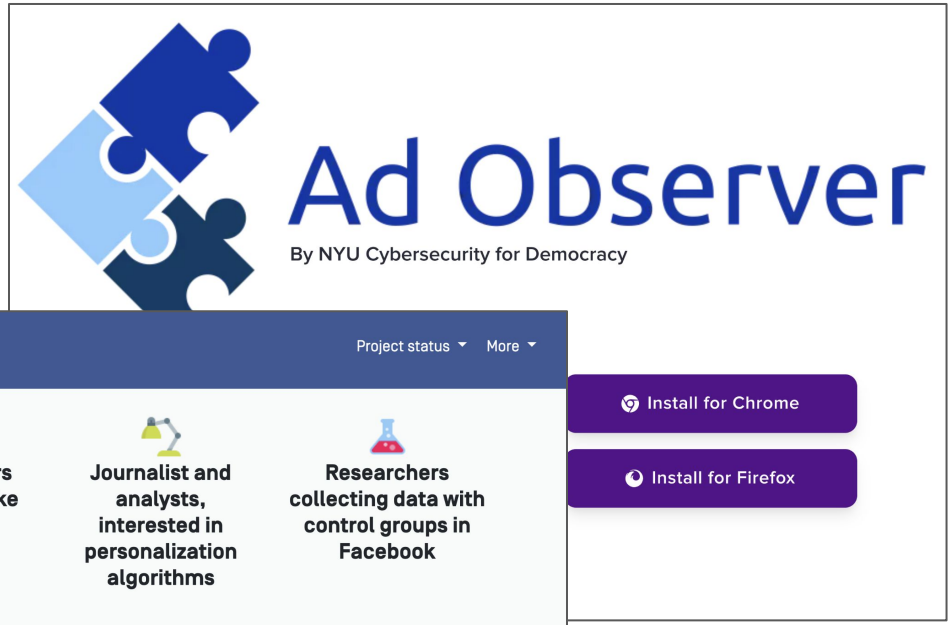
1. Click the extension icon in the browser bar

2. Report the video and recommendations that led you to it

3. Send any extra details you would like Mozilla to know

Browser extension as a widely used approach





https://whotargets.me/en/
https://adobserver.org/
https://facebook.tracking.exposed/

Crowdsourced audit based on extensions

SMU Singapore Management University
School of Computing and Information Systems

Our research is powered by real YouTube users.

Specifically: **37,380 volunteers** across **190 countries** installed Mozilla's RegretsReporter browser extensions for Firefox and Chrome.

FOR THIS REPORT, WE GATHERED FROM

3 362	1 662	91
REPORTS	VOLUNTEERS	COUNTRIES

Reports were submitted between **July 2020 - May 2021**. Volunteers who downloaded the extension but did **not** file a report were an important part of our study. Their data — for example, how often they use YouTube — was essential to our understanding of how frequent regrettable experiences are on YouTube and how this varies between countries.

“YouTube Regrets are disparate and disturbing

SMU Singapore Management University
School of Computing and Information Systems

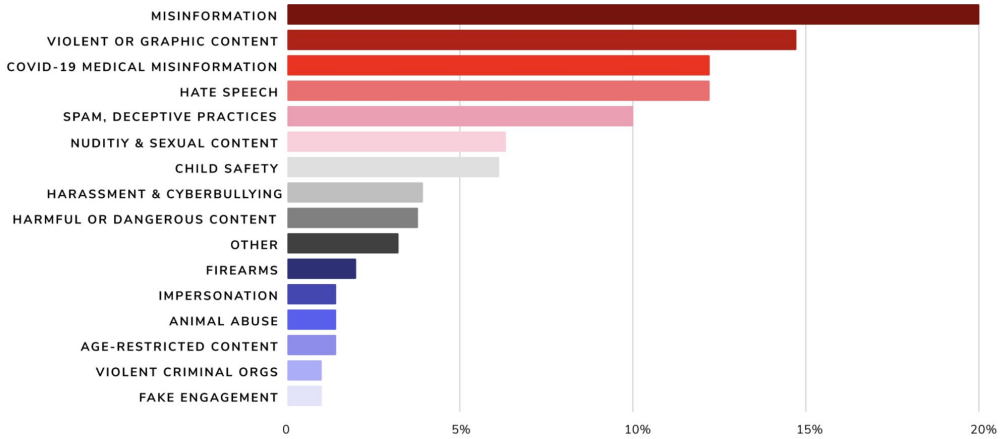
'Biggest fraud' in US history—up to 300,000 fake people voted in Arizona election: expert | NTD

965541 views - 1 Dec 2020

El Arca - It's Literally Furry Noah's Arc

119145 views - 22 Apr 2021

“12.2% of reported videos (95% confidence interval of 10.4 to 14.2%) either “should not be on YouTube” or “should not be proactively recommended,” based on YouTube’s Community Guidelines.”

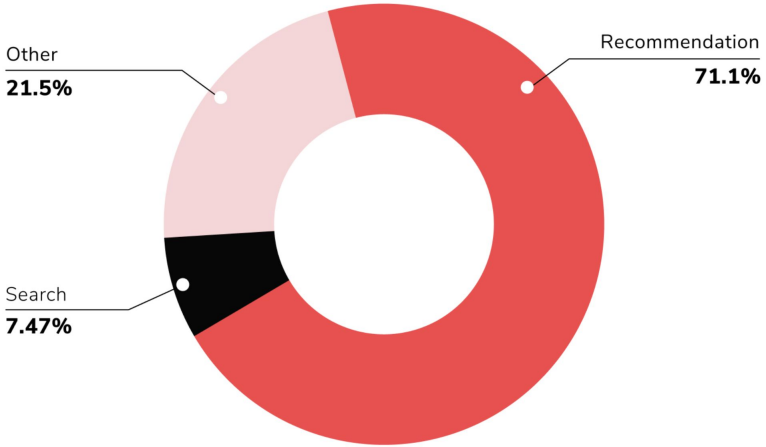


“The algorithm is the problem”

Around 9% of recommended Regrets were later removed from YouTube, but they had a collective 160 million views at the time that they were reported.

71% of all Regret reports came from videos recommended to the volunteers.

43.3% of recommendation was completely unrelated to the watch history.



Dual responsibilities of recommendation algorithm

The algorithm must weigh predictions about:

- How likely a video is to ‘engage’ someone
- How likely a video is to violate YouTube’s Community Guidelines.

But, several cases confirmed that:

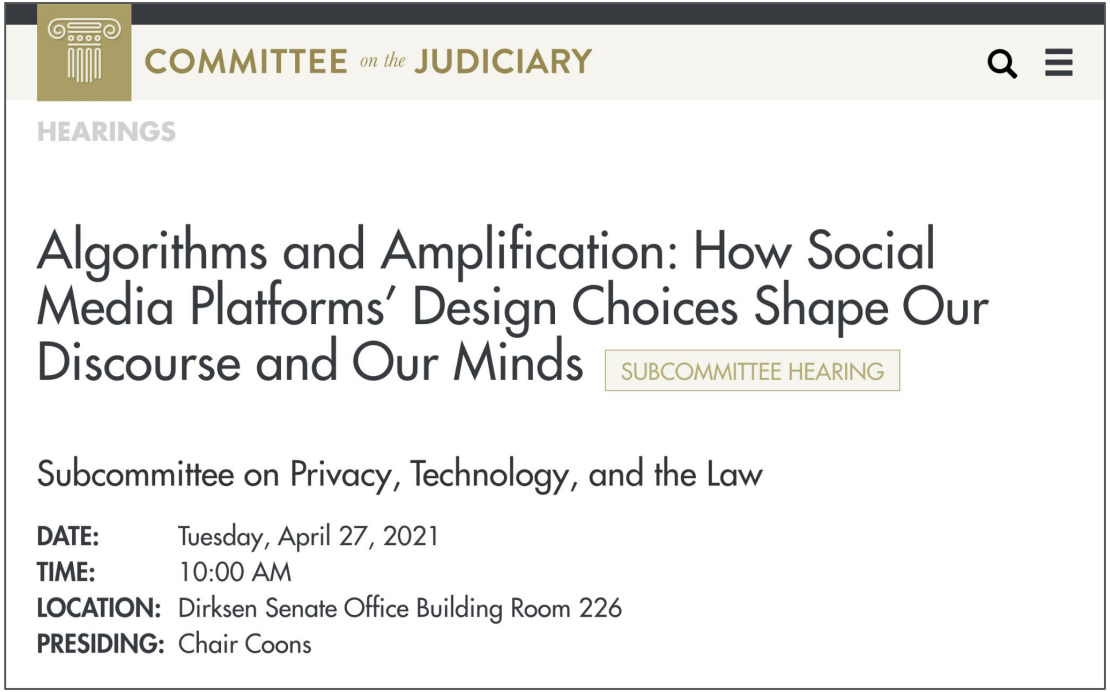
- Recommended videos that violated YouTube’s Community Guidelines were later removed, only after racking up millions of views.
- Reported videos had quickly gained significant viewership compared to other videos on the platform.

How does YouTube prioritize various decisions that their algorithm must make?

US Senate Judiciary Committee hearing (2021)

Senator Chris Coons
asking YouTube to commit to releasing information about how many times they recommend a given video

2:15:15~2:17:21



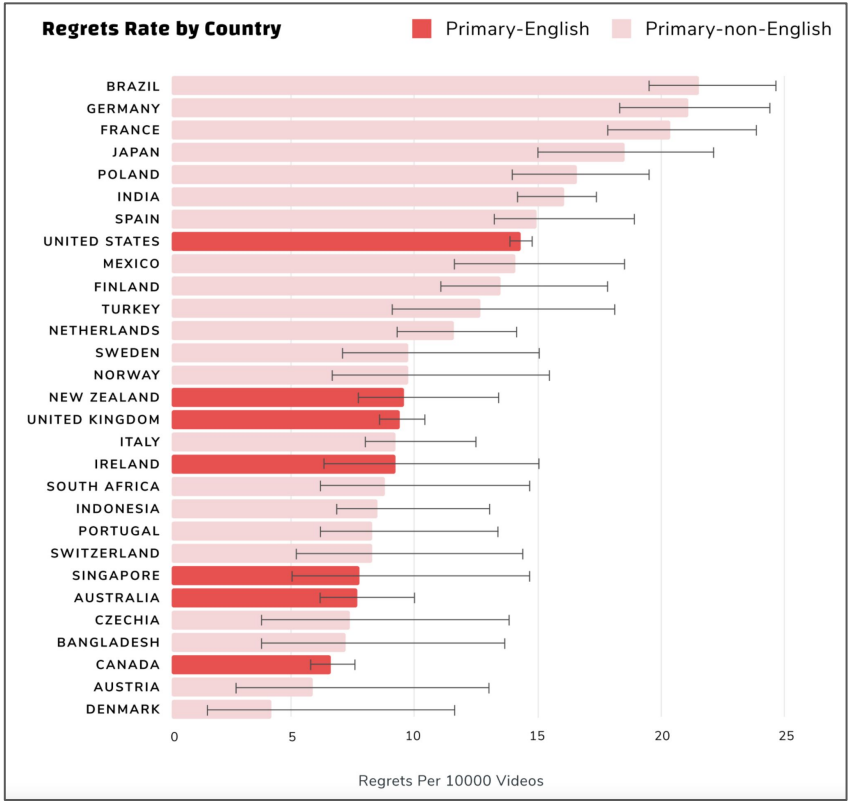
https://www.judiciary.senate.gov/meetings/algorithms-and-amplification-how-social-media-platforms-design-choices-shape-our-discourse-and-our-minds

“Bigger problem in non-English markets”

SMU Singapore Management University
School of Computing and Information Systems

The rate of YouTube Regrets is 60% higher in countries that do not have English as a primary language.

Covid-19-related reports are particularly prolific in non-English languages.



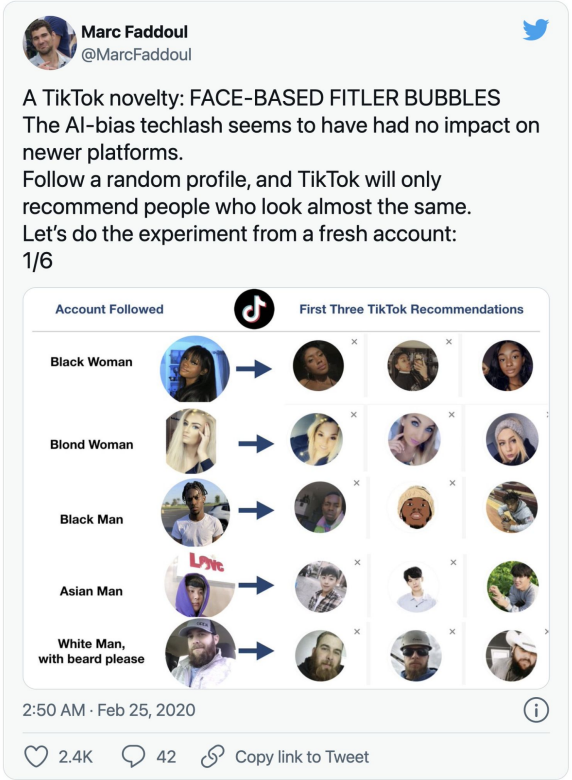
Group activity: design your audit studies

SMU Singapore Management University
School of Computing and Information Systems

Q. Does Google search favor big techs?: Someone argues that Google search shows positive news about big techs more frequently than negative news. Design your audit study to validate this claim.

Q. Do TikTok's recommended videos have a racial bias?: A racial bias in following recommendations are widely has been reported (see the left figure). Design your audit study to validate it.

Q. Apps you frequently used: Make a hypothesis about its bias and design your audit study to validate it.



https://twitter.com/MarcFaddoul/status/1232014908536938498

Example: Auditing social media posts by SG govt.

Q: Do Singapore government-related social media accounts represent each demographic group in a fair manner? (in terms of diversity)

Scraping audit: Data collection

- Find > 20 active government-related social media account on Twitter.
- Test GET /2/users/:id/tweets API endpoint (rate-limit: 900 reqs / 15-minute window).
- Collect 3,200 tweets from each of the accounts (It will take XX hours/days)
- Collect images embedded in tweets - There are no APIs for this.

Face labeling

- Examine gender / ethnicity of people in each image by using https://github.com/dchen236/FairFace

Evaluation

- Compare the proportion of each demographic group per social media account.

Reflection

https://smu.sg/IS457r4