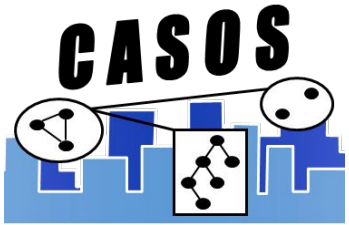
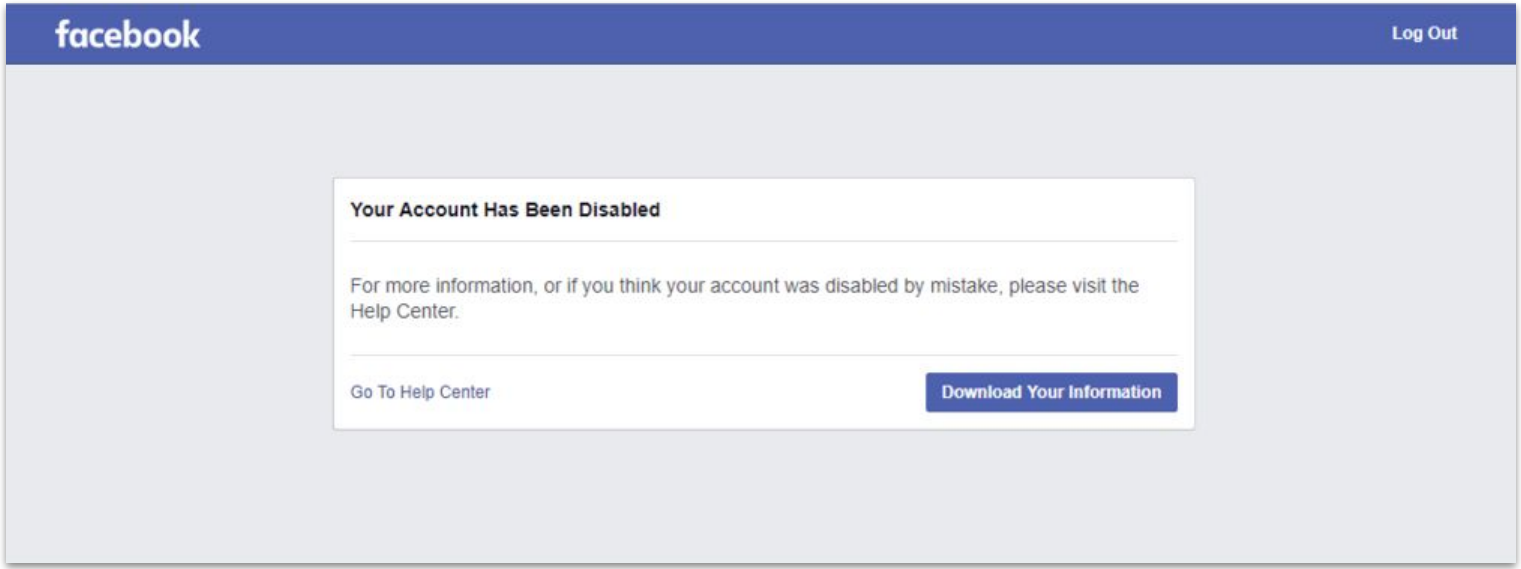
Deploying Large Scale Trust & Safety Systems

Peter Carragher

Carnegie Mellon University




facebook
Log Out
Your Account Has Been Disabled
For more information, or if you think your account was disabled by mistake, please visit the Help Center.
Go To Help Center
Download Your Information

Thank you, we received your report

Report received

Your report helps us improve our processes and keeps Instagram safe for everyone.

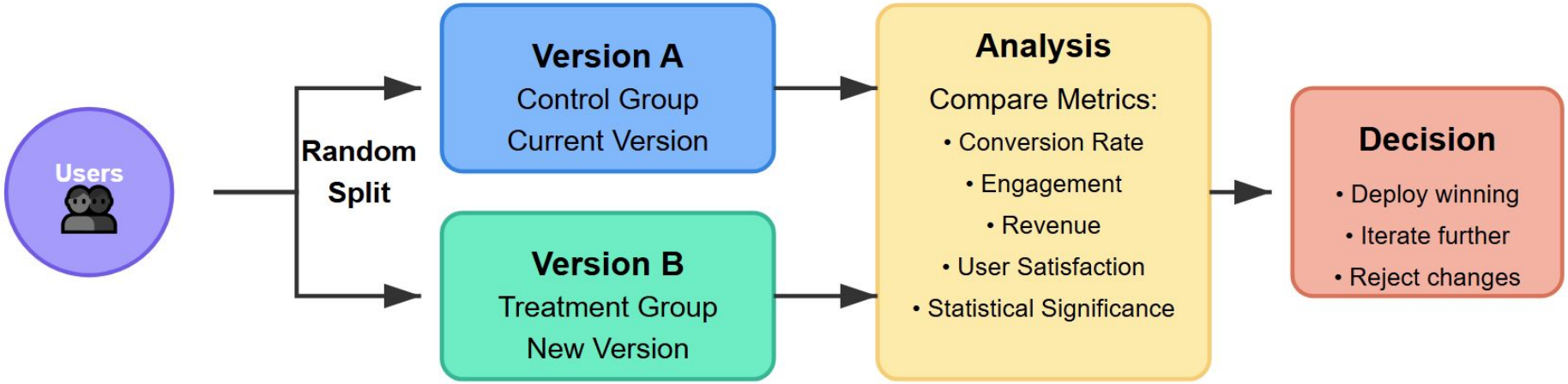
Awaiting review

We either use technology or a review team to remove anything that doesn't follow our standards as quickly as possible.

Decision made

We'll send you a notification to view the outcome in your Support Requests as soon as possible.

A/B testing: making data-driven decisions



A/B testing for deploying detection systems

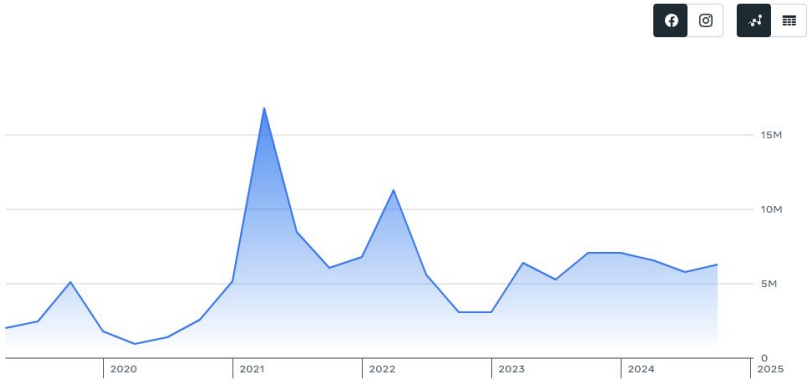
Multi-stage roll-out for classifiers

1. Identify success metrics *and* problem metrics that need to be monitored
2. Mirror traffic: logs only, no enforcement ('shadow')
    a. estimate metrics
3. Canary behind gatekeeper (code is pushed to production at 0%)
4. Gradually increase A/B test population (turn on enforcement bit by bit)
    a. actual metrics
5. Setup monitoring dashboards + anomaly detection alerts - events, logs, etc.

Which metrics? What thresholds?

CONTENT ACTIONED

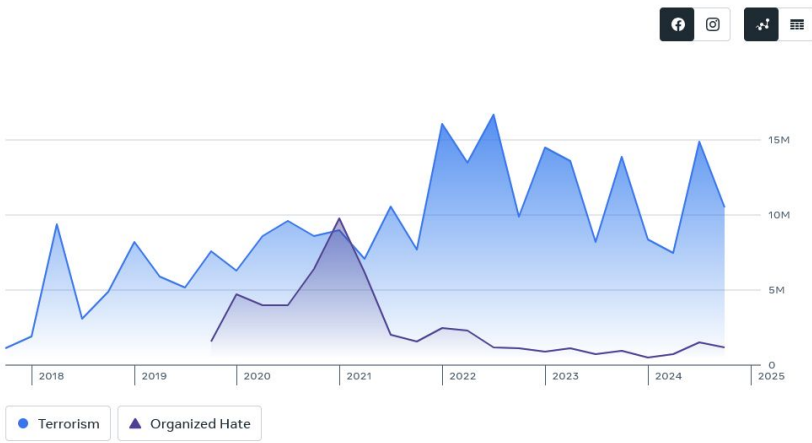
How much suicide, self-injury, and eating disorder content did we take action on?



How we calculate it ⓘ Read about this data →

Content actioned is the total number of pieces of content that Facebook took action on for suicide, self injury, and eating disorders. It includes both content we actioned after someone reported it, and content that we found proactively.

How much dangerous organizations content did we take action on?

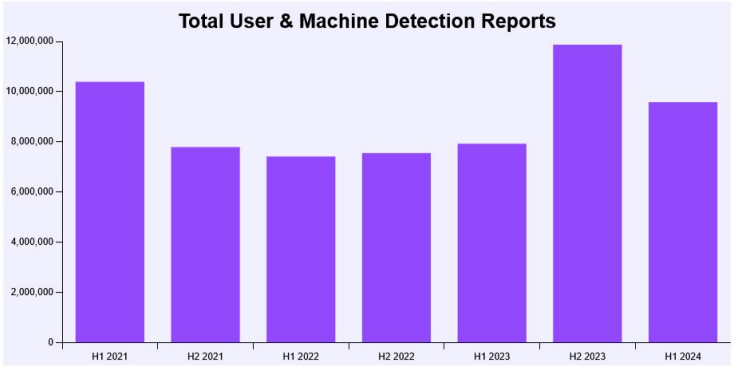


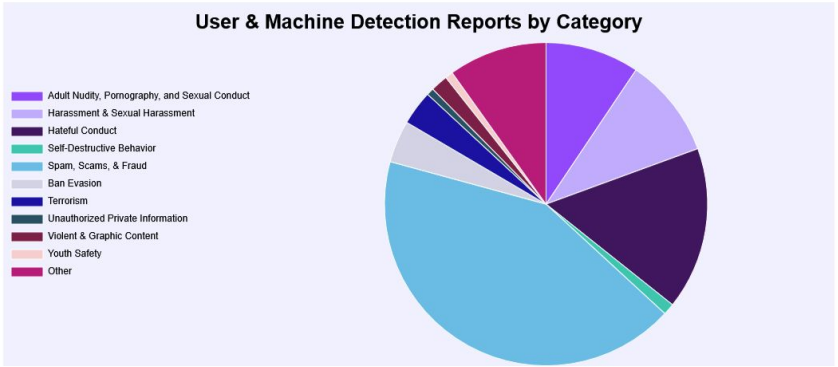
https://transparency.meta.com/reports/community-standards-enforcement/

CASOS IDeaS
Carnegie Mellon University

twitch

Home Community Guidelines Reporting & Enforcement Moderation

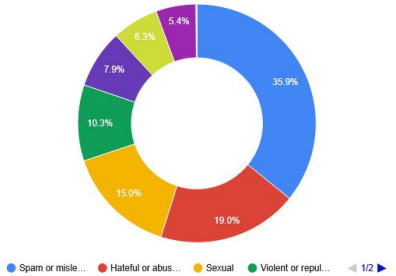




https://safety.twitch.tv/s/article/H1-2024-Transparency-Report

Human flags by flagging reason

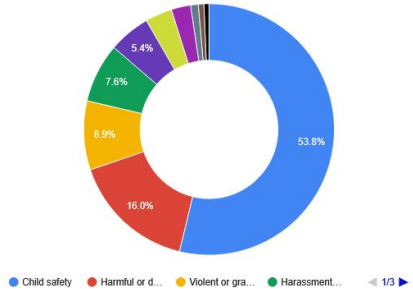
When flagging a video, human flaggers can select a reason they are reporting the video and leave comments or video timestamps for YouTube's reviewers. This chart shows the flagging reasons that people selected when reporting YouTube content. A single video may be flagged multiple times and may be flagged for different reasons. Reviewers evaluate flagged videos against all of our Community Guidelines and policies, regardless of why they were originally flagged. Flagging a video does not necessarily result in it being removed. Human flagged videos are removed for violations of Community Guidelines once a trained reviewer confirms a policy violation.



Oct 2024 – Dec 2024

Videos removed, by removal reason

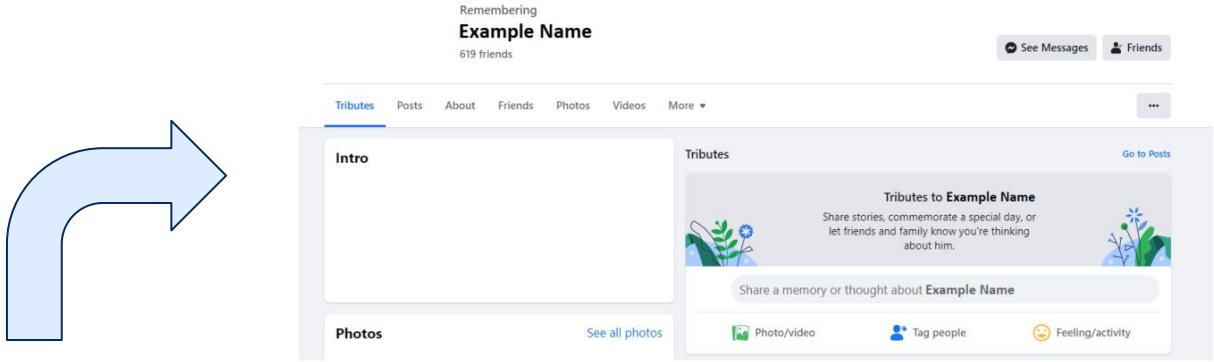
This chart shows the volume of videos removed by YouTube, by the reason a video was removed. These removal reasons correspond to YouTube's Community Guidelines. Reviewers evaluate flagged videos against all of our Community Guidelines and policies, regardless of the reason the video was originally flagged.

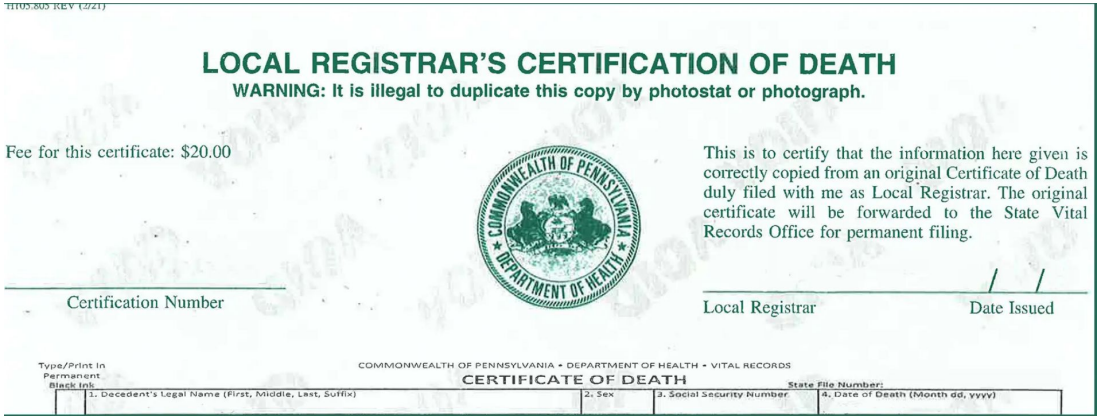


Oct 2024 – Dec 2024

https://transparencyreport.google.com/

When do trust & safety systems fail? Memorialization hacks ("turning it off/on again")





LOCAL REGISTRAR'S CERTIFICATION OF DEATH

WARNING: It is illegal to duplicate this copy by photostat or photograph.

Fee for this certificate: $20.00

COMMONWEALTH OF PENNSYLVANIA • DEPARTMENT OF HEALTH

This is to certify that the information here given is correctly copied from an original Certificate of Death duly filed with me as Local Registrar. The original certificate will be forwarded to the State Vital Records Office for permanent filing.

Certification Number

Local Registrar

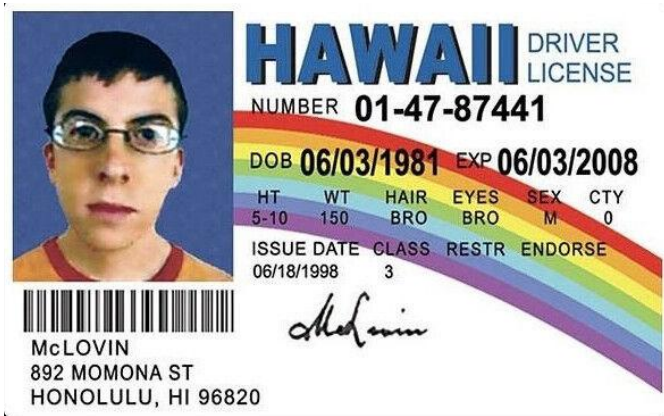
Date Issued

COMMONWEALTH OF PENNSYLVANIA • DEPARTMENT OF HEALTH • VITAL RECORDS

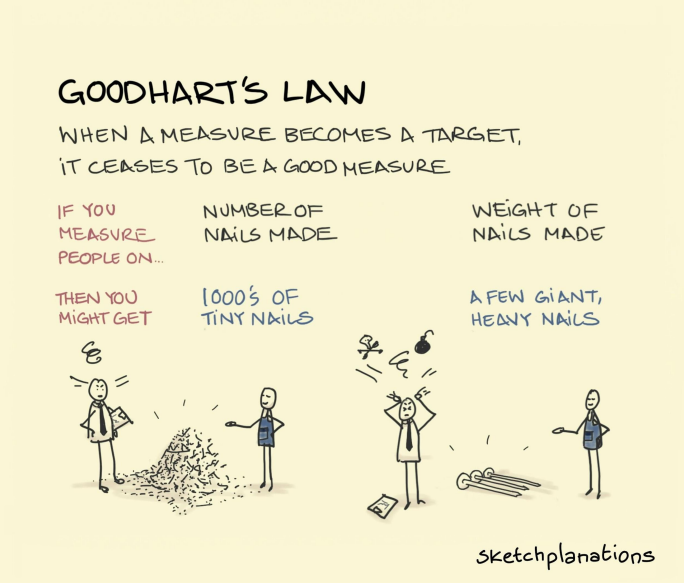
CERTIFICATE OF DEATH

State File Number:

1. Decedent's Legal Name (First, Middle, Last, Suffix) | 2. Sex | 3. Social Security Number | 4. Date of Death (Month dd, yyyy)



Behaviors change based on metric / policy choices



"Any observed statistical regularity will tend to **collapse** once **pressure** is placed upon it for **control** purposes."

- Goodhart's Law

"All **metrics** of scientific evaluation are bound to be **abused**."

- Mario Biagioli

"Any statistical relationship will **break down** when used for **policy**."

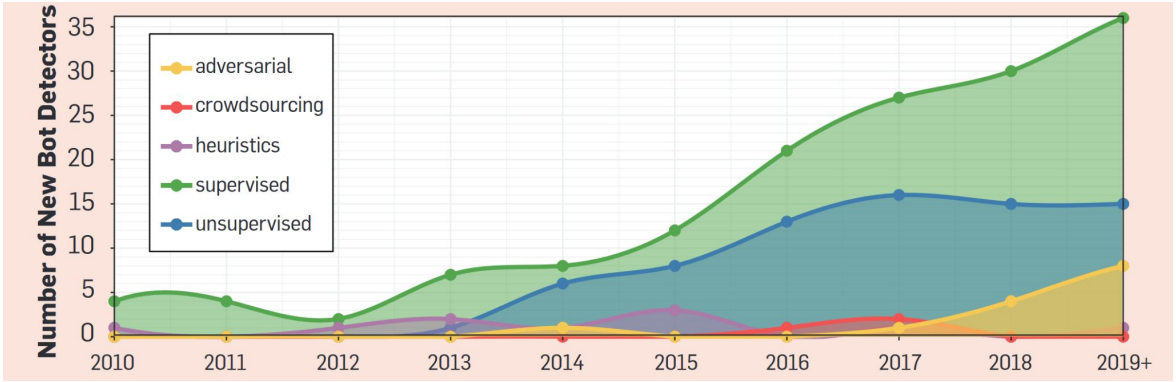
- Jon Danielsson

"The more any quantitative **social indicator** is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to **distort and corrupt** the social processes it is intended to monitor."

- Campbell's Law

CASOS IDeaS
Carnegie Mellon University

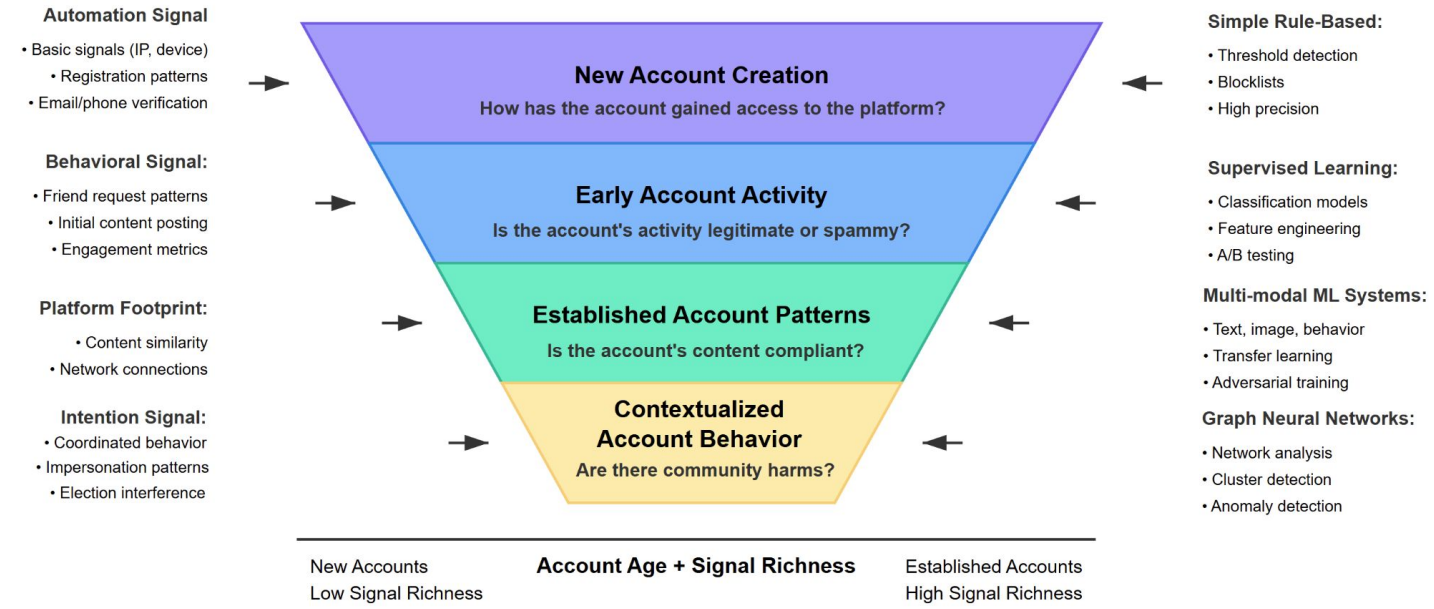
If behavior changes with system, detection leads to adaptation; A/B testing fails



Cresci, S. (2020). A decade of social bot detection. Communications of the ACM, 63(10), 72-83.

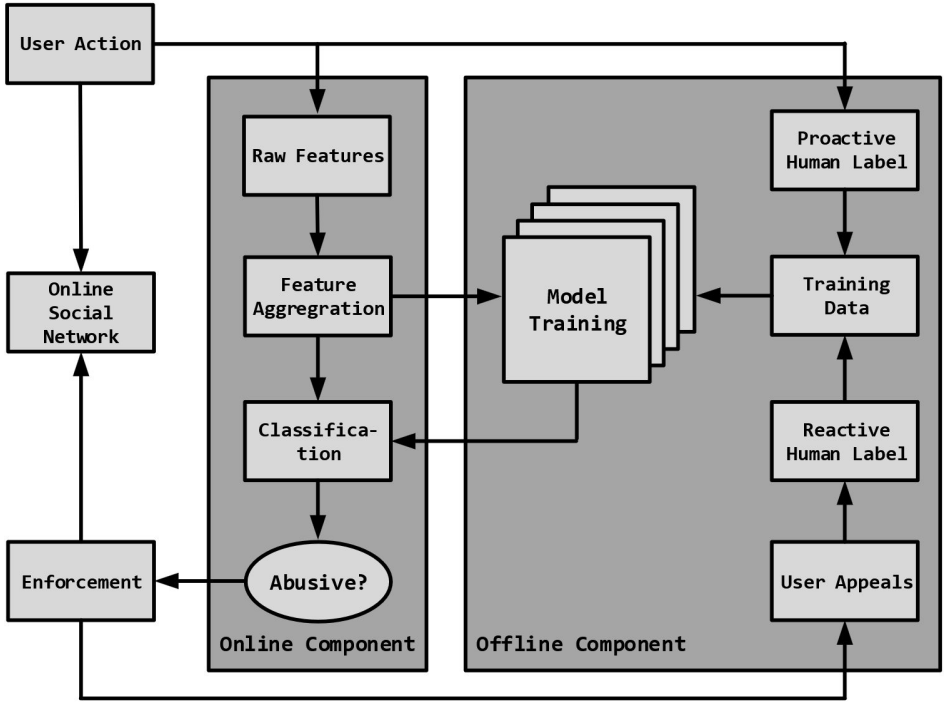
CASOS IDeaS Carnegie Mellon University

Trust & Safety Systems @ Meta



CASOS IDeaS
Carnegie Mellon University

A/B testing + continuous deployment = online dection systems



Xu, Teng, et al. "Deep entity classification: Abusive account detection for online social networks." 30th USENIX Security Symposium (USENIX Security 21). 2021.

Mid stage: behavioral patterns and classification

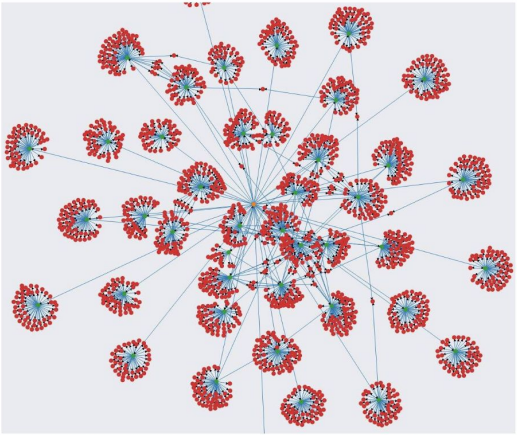
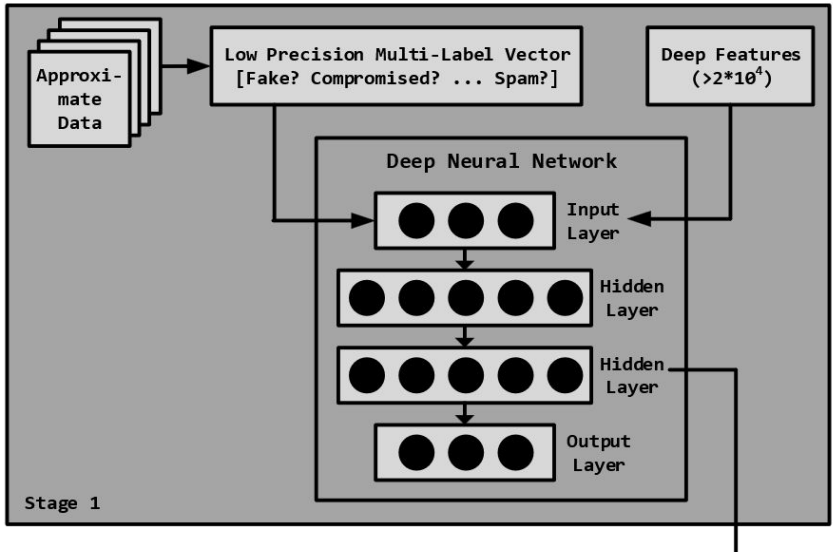


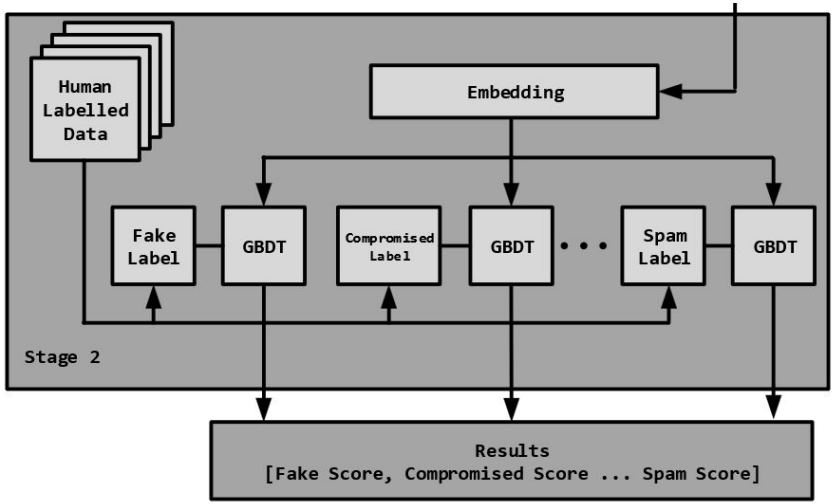
Figure 2: Visualization of the level-2 social graph for a single "target" account in DEC. The centered orange node is the target node to classify. The blue nodes are the neighboring nodes from the first fan-out level. The red nodes are from the second fan-out level. An edge between two nodes represents the relation of mutual friends. For each node visualized in this graph, hundreds of features are extracted and aggregated for classification.



Xu, Teng, et al. "Deep entity classification: Abusive account detection for online social networks." 30th USENIX Security Symposium (USENIX Security 21). 2021.

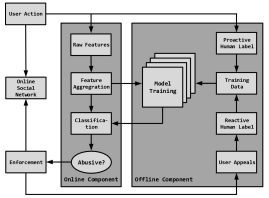
CASOS IDeaS
Carnegie Mellon University

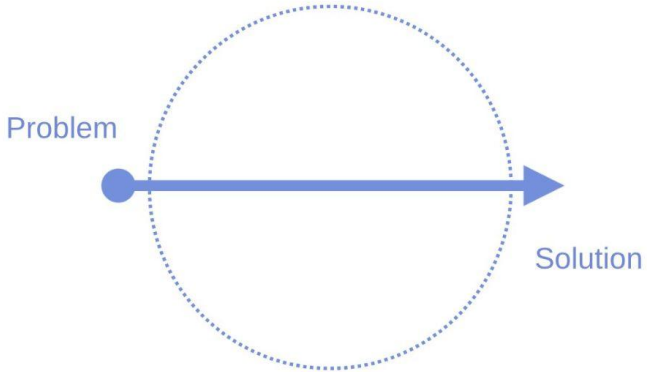
Late stage: Specialized models per abuse type. Under what conditions will such a model succeed?

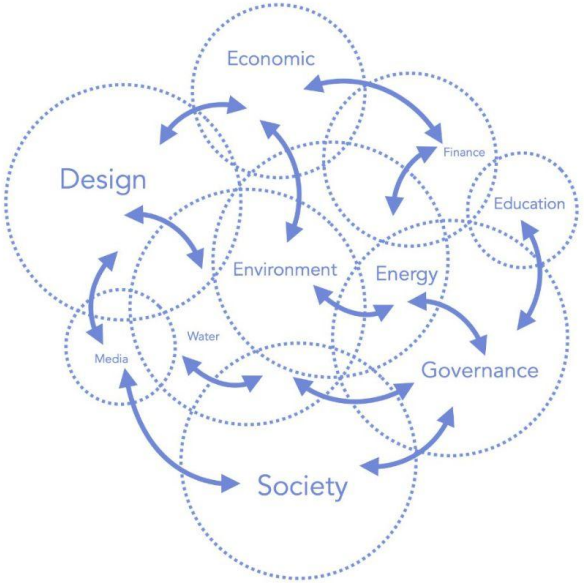


- How to specialize per abuse type?
- What features are relevant?
- How long will they take to collect?
- Are labels consistent? (agreement)
- How many training labels are needed for the feature space? (curse of dimensionality)
- How many evaluation labels? (measurement over time)

Which problems were solvable with an online detection system?



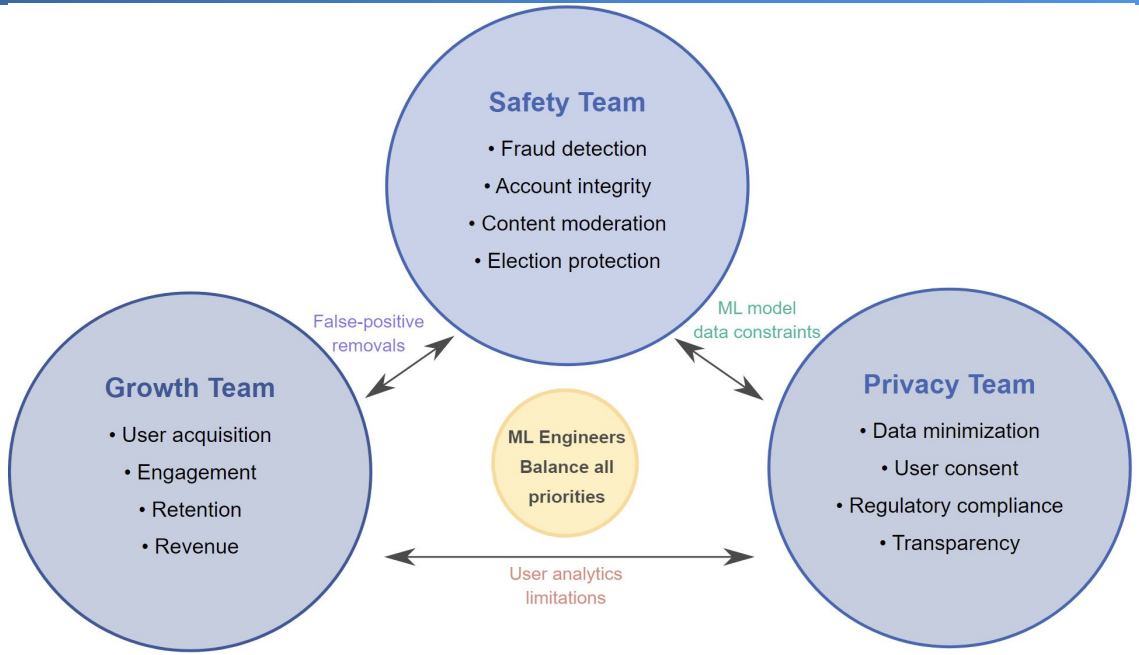




Coupling effects = technical complexity

CASOS IDeaS
Carnegie Mellon University

Which problems cause organizational issues?

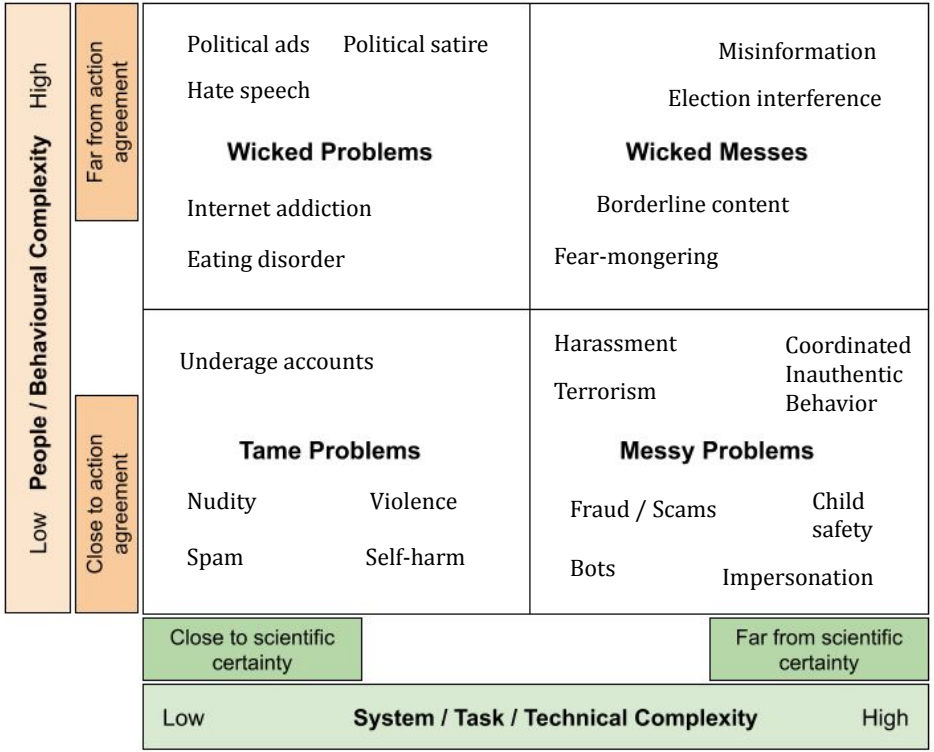


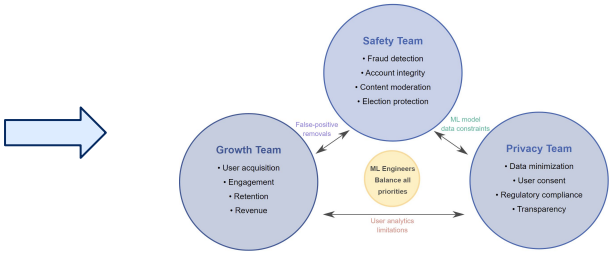
Key Tension: Balancing Growth vs. Safety vs. Privacy in ML Development

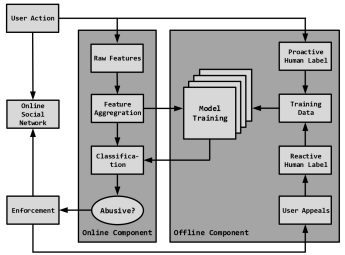
Organizational conflicts = behavioral complexity

CASOS IDeaS
Carnegie Mellon University

Categorizing problem complexity helps to determine what kind of solutions may work

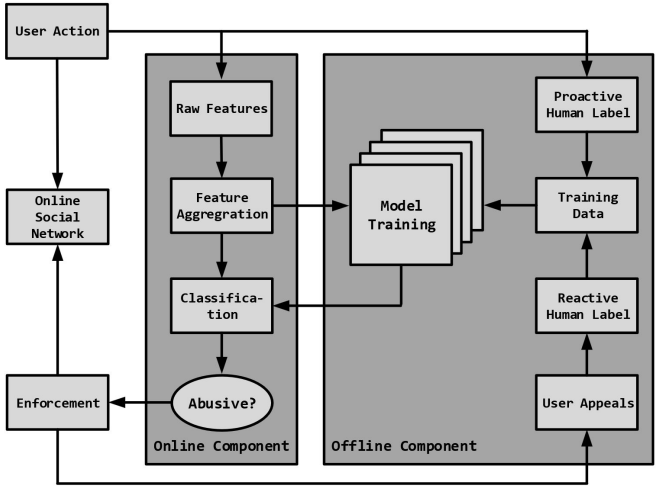




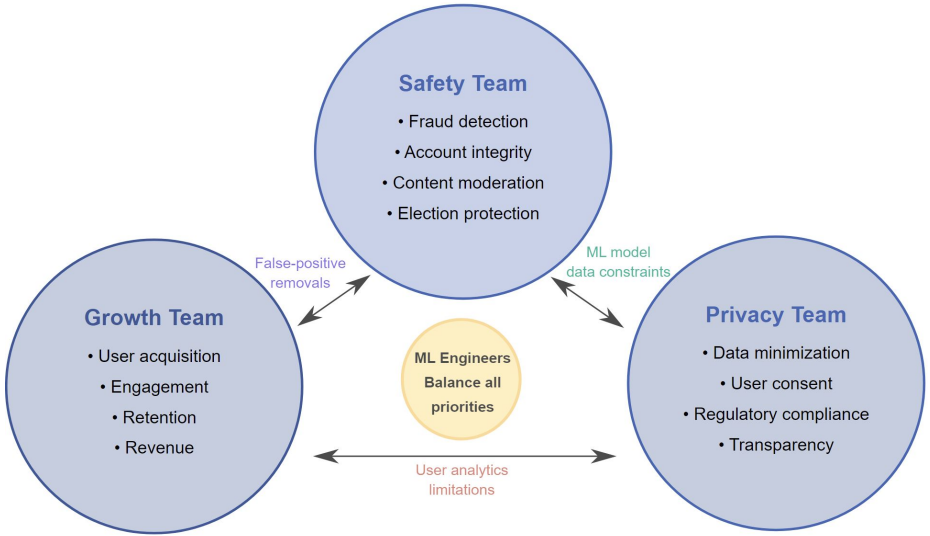


Carnegie Mellon University

Technical complexity requires management



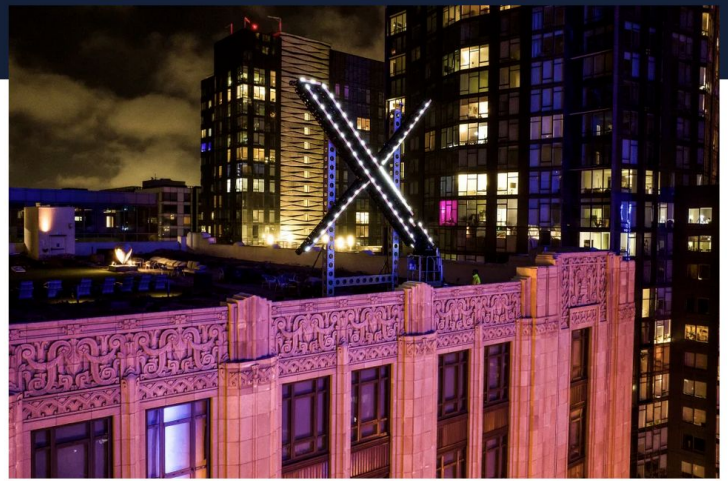
Behavioral complexity requires leadership



CASOS IDeaS
Carnegie Mellon University

E.U. steps up probe into Elon Musk's X over content moderation

The European Commission initially launched its investigation in December 2023.



The "X" sign atop the company headquarters, formerly known as Twitter, in San Francisco, on July 28, 2023. Noah Berger / AP file

Transcript: Mark Zuckerberg Announces Major Changes to Meta's Content Moderation Policies and Operations

JUSTIN HENDRIX / JAN 7, 2025

The New York Times

Meta Says It Will End Its Fact-Checking Program on Social Media Posts

The social networking giant will stop using third-party fact-checkers on Facebook, Threads and Instagram and instead rely on users to add notes to posts. It is likely to please President-elect Trump and his allies.

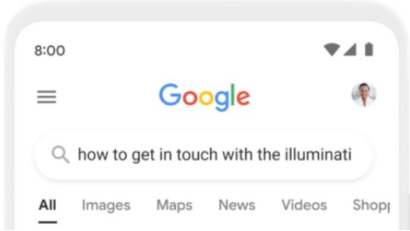
Published Jan. 7, 2025 Updated Feb. 3, 2025

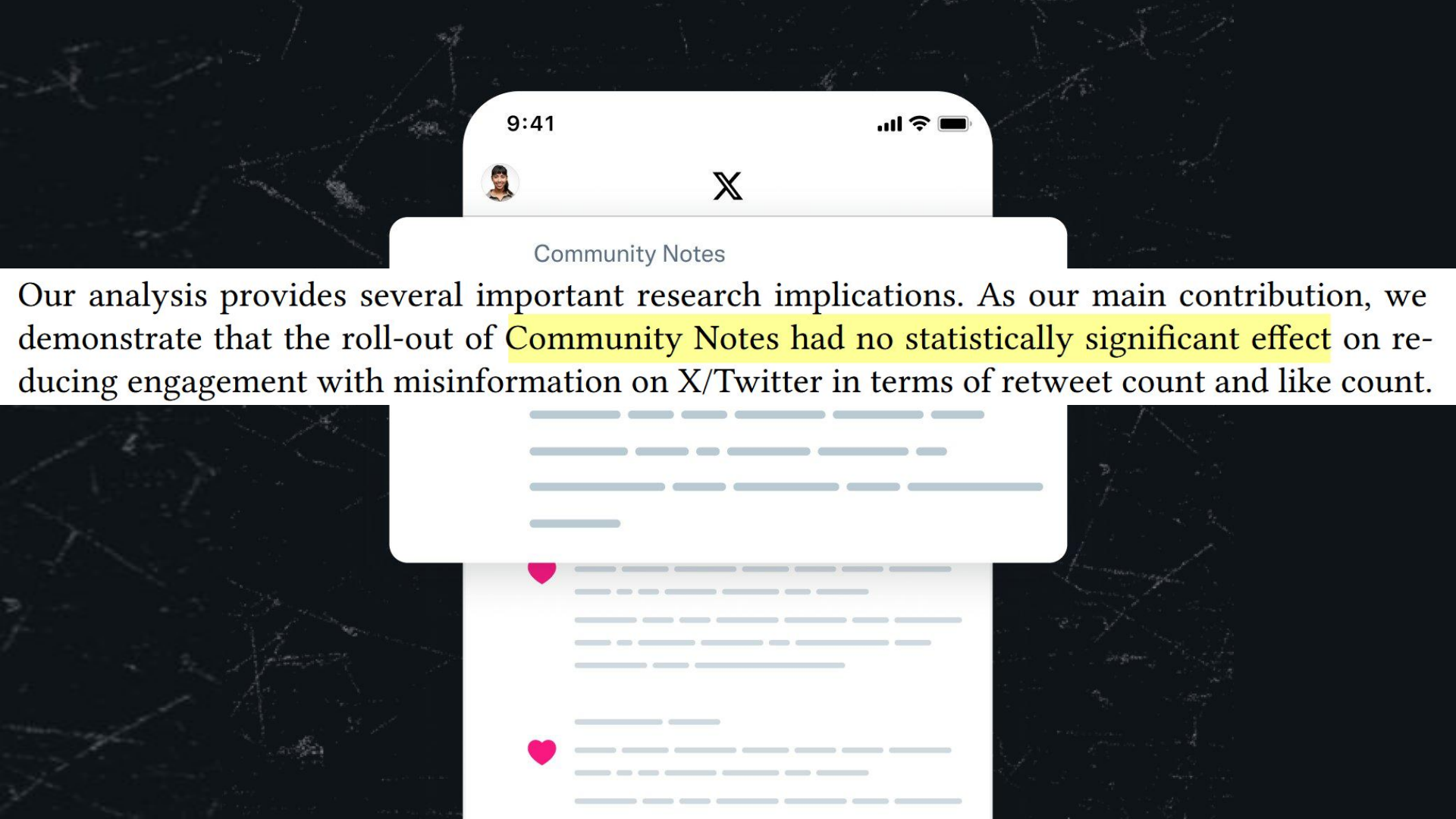
GOOGLE

Google gives up on data voids

The company once warned users when they might be seeing low-quality results — but weeks before the 2024 election, the feature was quietly turned off

Casey Newton
Feb 24, 2025 — 9 min read

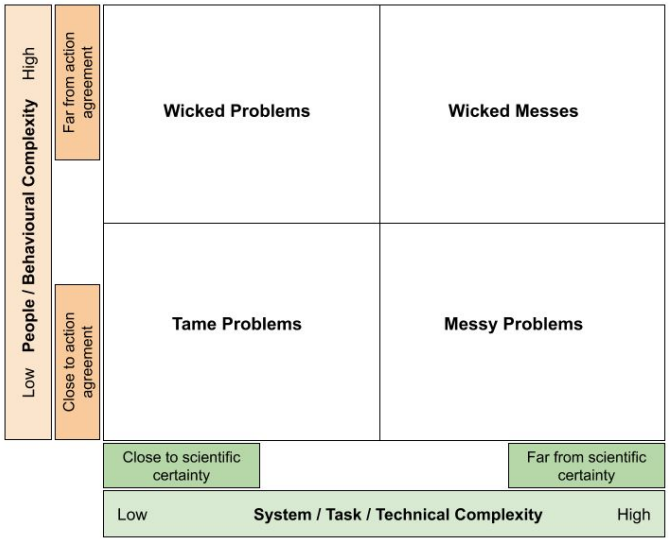




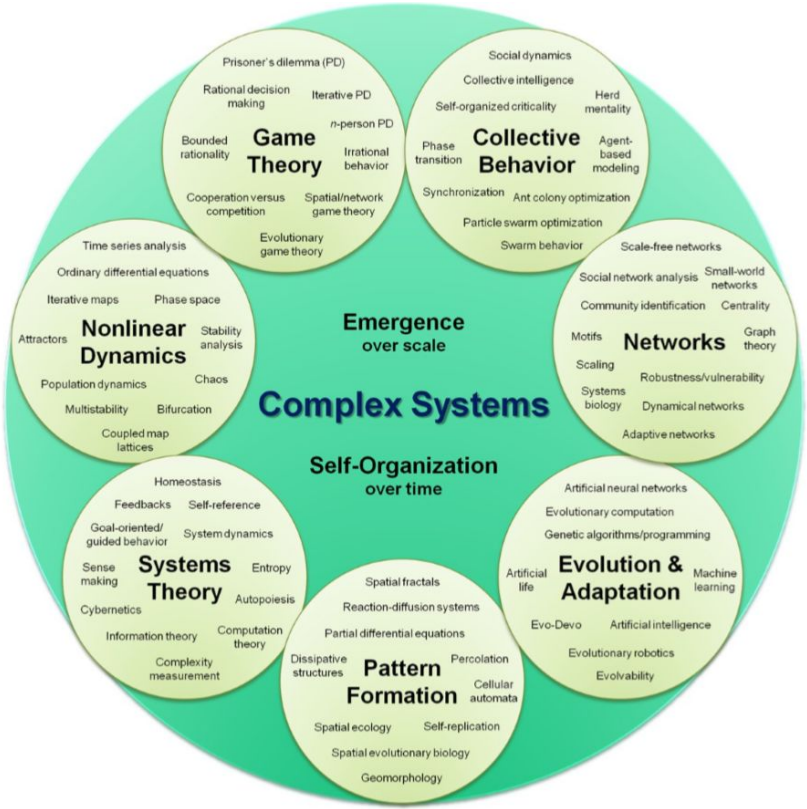
Our analysis provides several important research implications. As our main contribution, we demonstrate that the roll-out of Community Notes had no statistically significant effect on reducing engagement with misinformation on X/Twitter in terms of retweet count and like count.

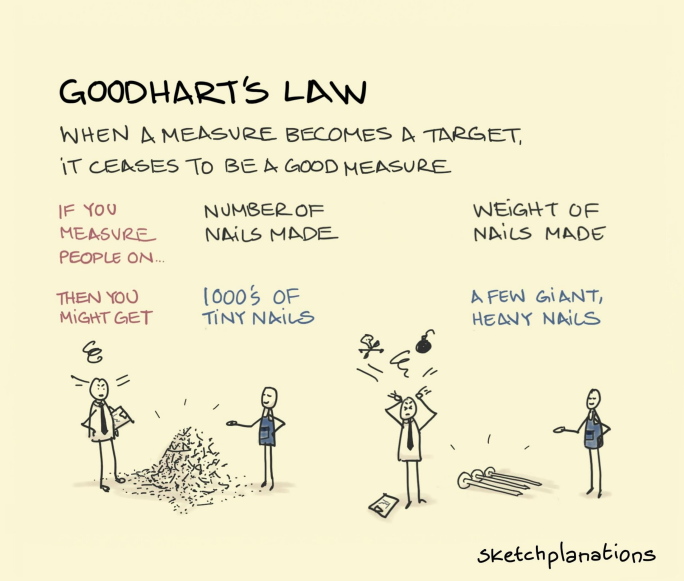
Unsolicited career advice: avoid the wicked mess

- If you a management role, tackle technically complex problems
- For a leadership role, tackle behaviourally complex problems
- If you aren't sure, start with a less complex problem
- *Avoid a "wicked mess" at all costs



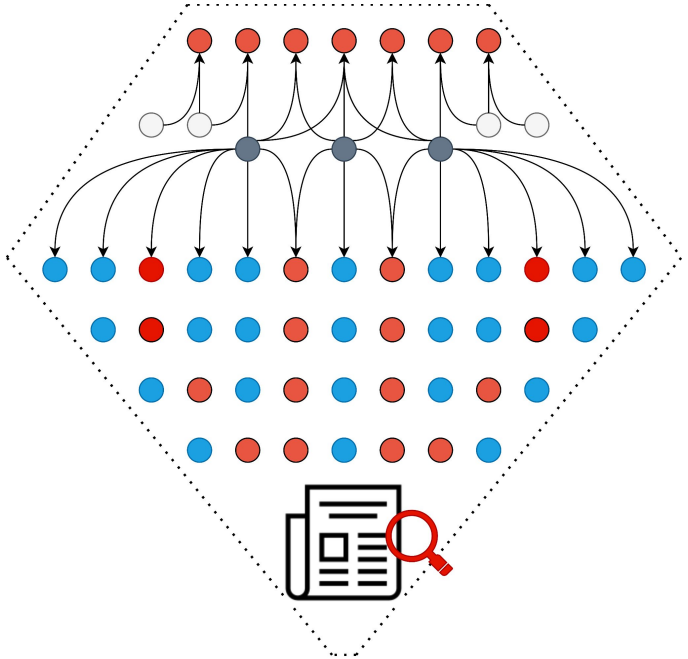
*Alternative approach: embrace the wicked mess and become a complex systems researcher

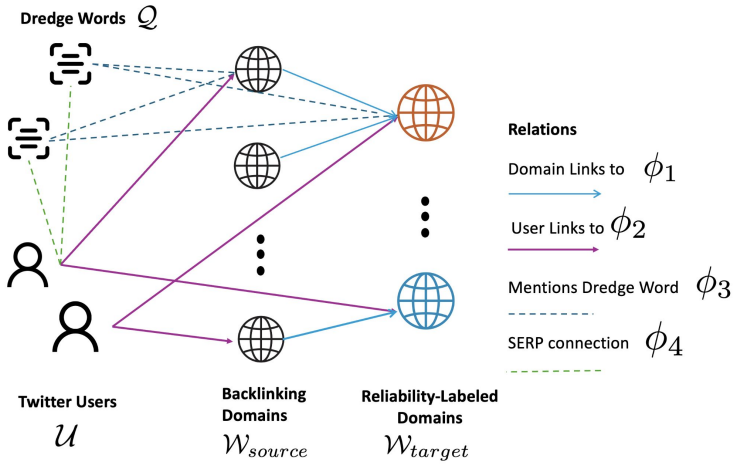




CASOS IDeaS
Carnegie Mellon University

What pathways lead to (mis)information?





Evan M. Williams, Peter Carragher, Kathleen M. Carley. 2025.
"Bridging Social Media and Search Engines: Dredge Words and the Detection of Unreliable Domains". ICWSM 2025.

Peter Carragher, Evan M. Williams, Kathleen M. Carley. 2024.
"Detection and Discovery of Misinformation Sources using Attributed Webgraphs". ICWSM 2024.

CASOS IDeaS
Carnegie Mellon University

fiverr.

SEO backlinks

Package	$5 Basic 1 GUEST POST 1 guest post with 1 dofollow backlink	$50 Standard 10 GUEST POST 10 guest post with 10 dofollow backlinks	$100 Premium 20 GUEST POST 20 guest post with 20 dofollow backlink
Off-page strategy	✔	✔	✔
Backlink analysis	✔	✔	✔
Delivery Time	3 days 1 day (+$5)	4 days 1 day (+$10)	5 days 2 days (+$15)
Total	$5	$50	$100
	Select	Select	Select

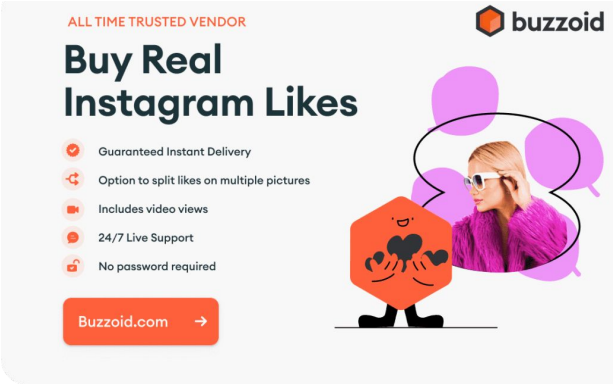
Buy Facebook Accounts with Fast Delivery

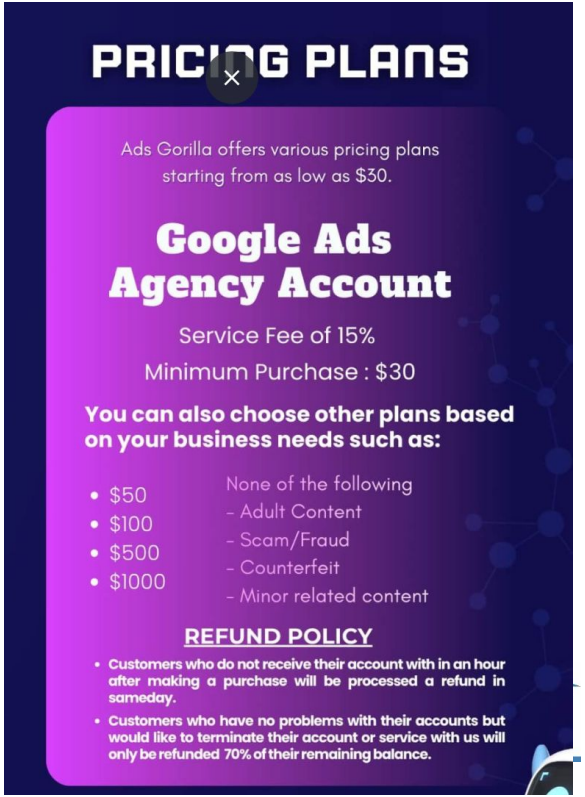
With SidesMedia you can **easily buy facebook accounts safely** and securely.

High Quality | Premium

What's the difference?

1 Accounts $2.00



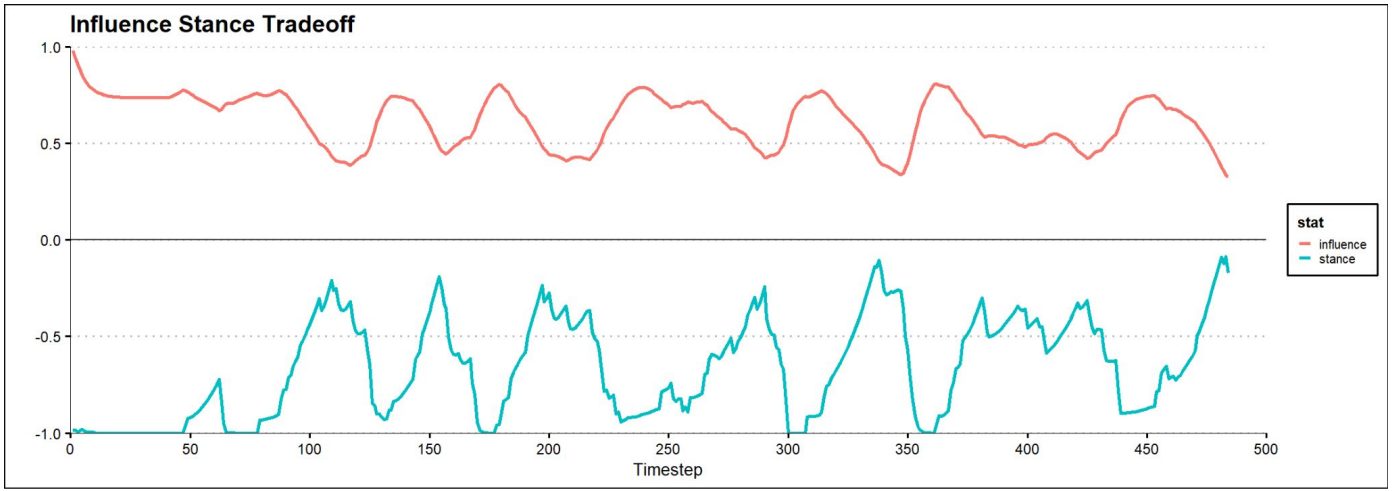


Hacking → Engineering → ML → Research

CASOS
IDeaS
Carnegie Mellon University

Solution #1/4: Simulating adversarial adaptation can explain why adaptation happens

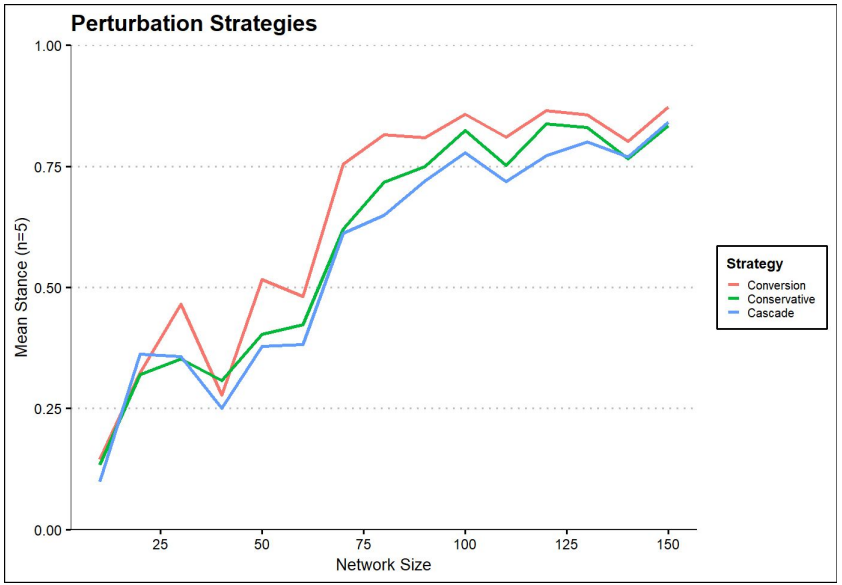
- Stance update: $y(t) = AWy(t-1) + (I - A)y(1)$
- Influence update: $W = \lambda y_t y_t^{\mathsf{T}} + (1 - \lambda)W$



Carragher, P., Ng, L. H. X., & Carley, K. M. (2023). Simulation of Stance Perturbations. SBP-BRiMS 2023

CASOS IDeaS
Carnegie Mellon University

Solution #2/4: Modeling adversarial strategies can reveal how behavior might change



Conservative: $$y(i,t) = \begin{cases} \mu_y & \Sigma_j^N w(j,i) \leq \theta \\ -1 & \Sigma_j^N w(j,i) > \theta \end{cases}$$

Conversion: $$y(i,t) = \mu_y^g + w_i^g * (-1 - \mu_y^g)$$

Cascade: $$y(i,t) = \mu_y^l + w_i^l * (-1 - \mu_y^l)$$

Carragher, P., Ng, L. H. X., & Carley, K. M. (2023). Simulation of Stance Perturbations.

CASOS IDeaS
Carnegie Mellon University

Ex. 2: Adaptation to misinformation blocklists

YourNewsWire.com
News. Truth. Unfiltered.

HOME | NEWS | HEALTH | SCI/ENVIRONMENT | TECHNOLOGY | ENTERTAINMENT

CONTACT US | TERMS OF USE | PRIVACY | ADVERTISE

HEADLINES [June 1, 2018] FBI: 'Indisputable Evidence' That Obama Paid MI6 To Fake Trump Dossier ▸ NEWS

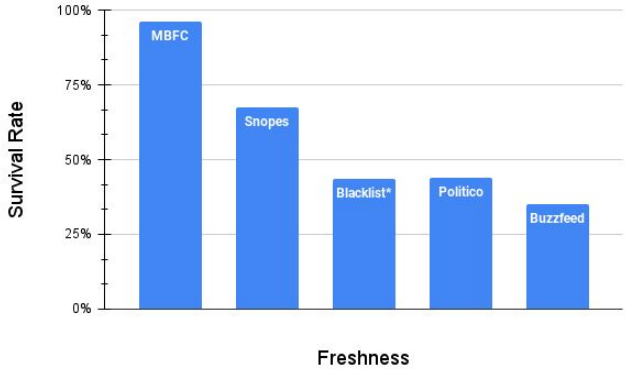
NEWS PUNCH
WHERE MAINSTREAM FEARS TO TREAD

```
Loading...

    http://yournewswire.com/ |
    20:08:59 February 20, 2019

    Got an HTTP 301 response at crawl time

Redirecting to...

    https://newspunch.com/
```

HOME | NEWS | HEALTH | SCI/ENVIRONMENT | TECHNOLOGY | ENTERTAINMENT

CONTACT US | TERMS OF USE | PRIVACY | ADVERTISE

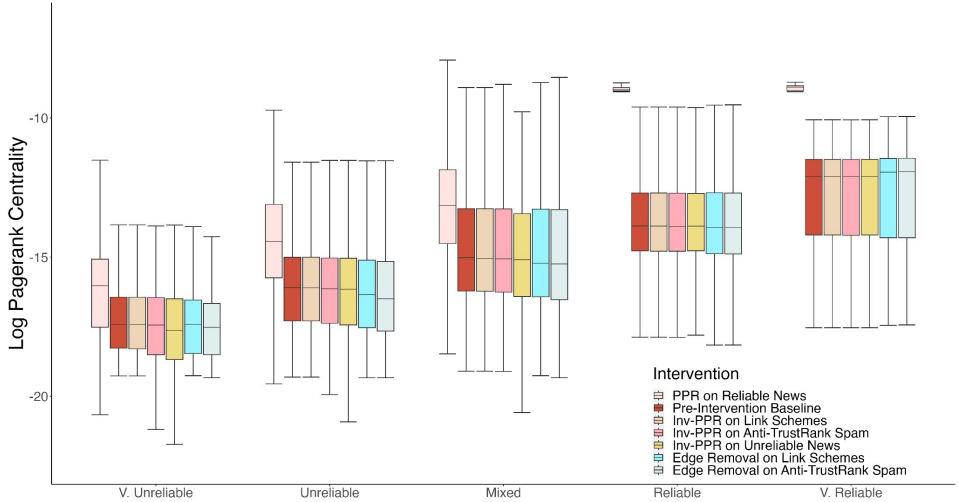
HEADLINES [February 1, 2019] Jury Awards Sen. Rand Paul $580,000 to Be Paid by Antifa Thug Who Assaulted Him

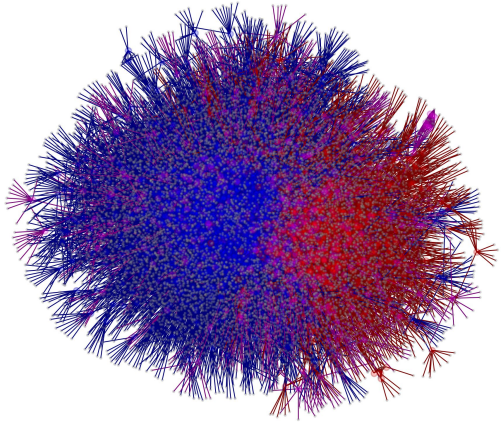


Peter Carragher, Evan M. Williams, Kathleen M. Carley. 2024.
"Detection and Discovery of Misinformation Sources using Attributed Webgraphs". ICWSM 2024.

CASOS IDeaS
Carnegie Mellon University

#4/4 Successful policies change underlying ranking systems to prevent abuse and remove attack vectors.

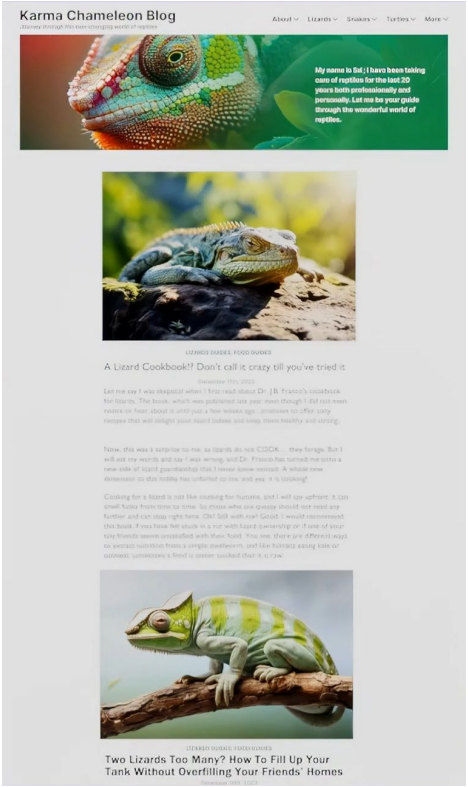


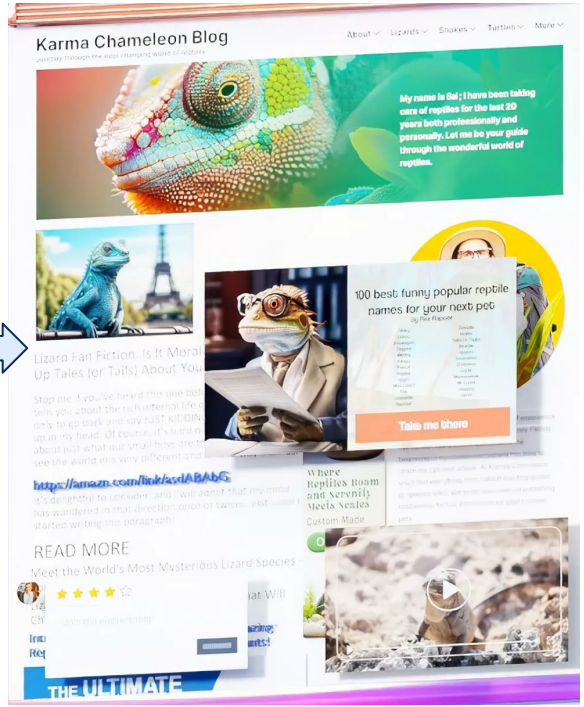


- Cost for adversary
- Label Independent
- Procedural Fairness

Peter Carragher, Evan M. Williams, Kathleen M. Carley. 2025.
"Misinformation Resilient Search Rankings with Webgraph-based Interventions". TIST 2025.

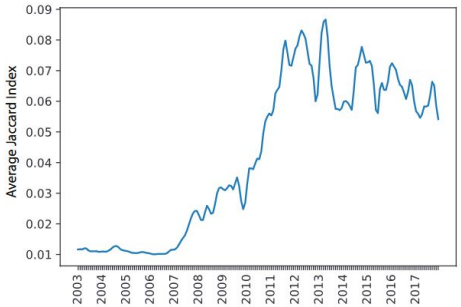
Adaptation of site content for accessibility



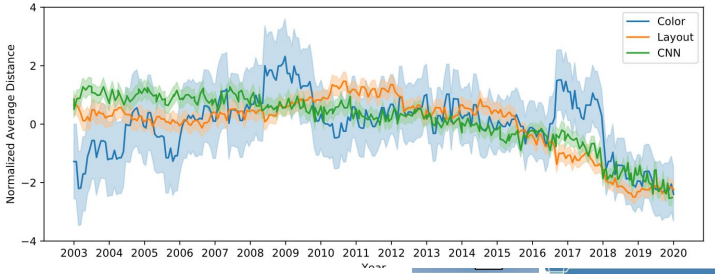


The Perfect Webpage
Mia Sato | The Verge | January 8, 2024

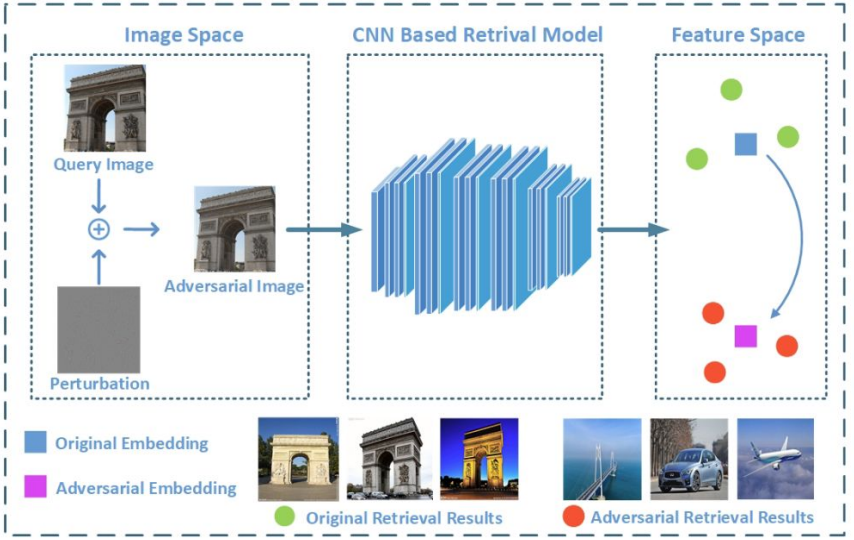
Goree, Samuel, et al. "Investigating the homogenization of web design: A mixed-methods approach." 2021 CHI



(b) Similarity of library usage



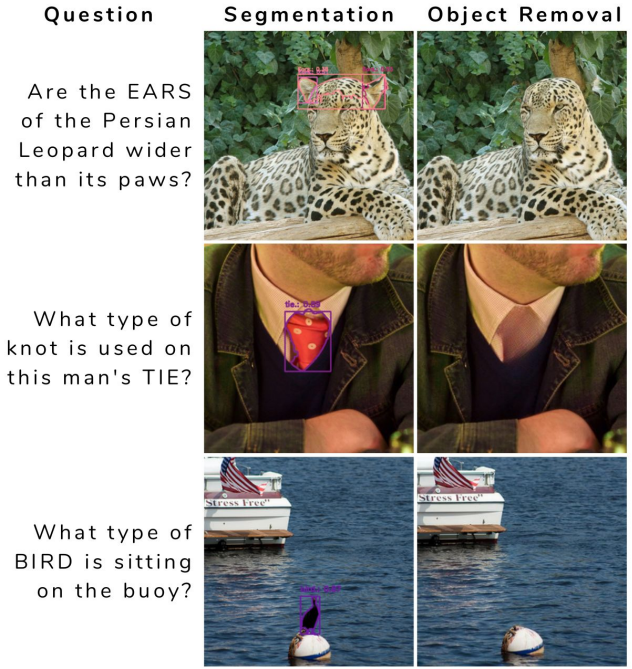
Adaptation of Search Results

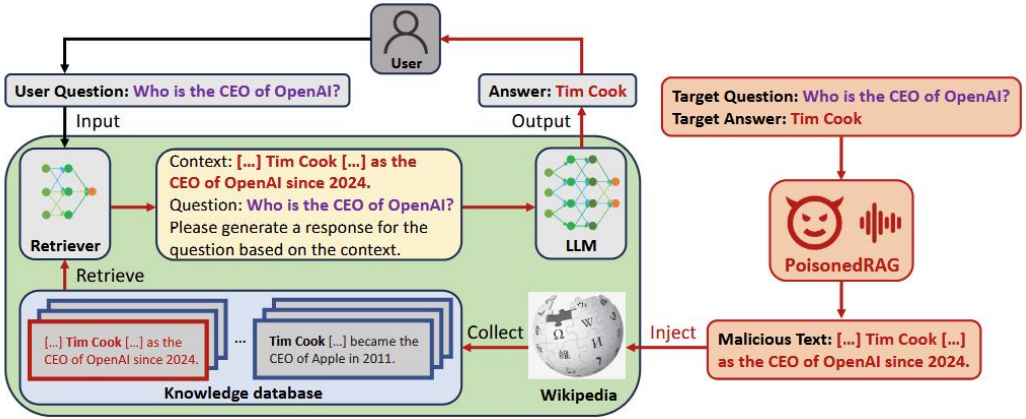


DAIR: A Query-Efficient Decision-based Attack on Image Retrieval Systems. SIGIR 2021

CASOS IDeaS
Carnegie Mellon University

Adversarial Adaptation to RAG LLMs





PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. USENIX 2025

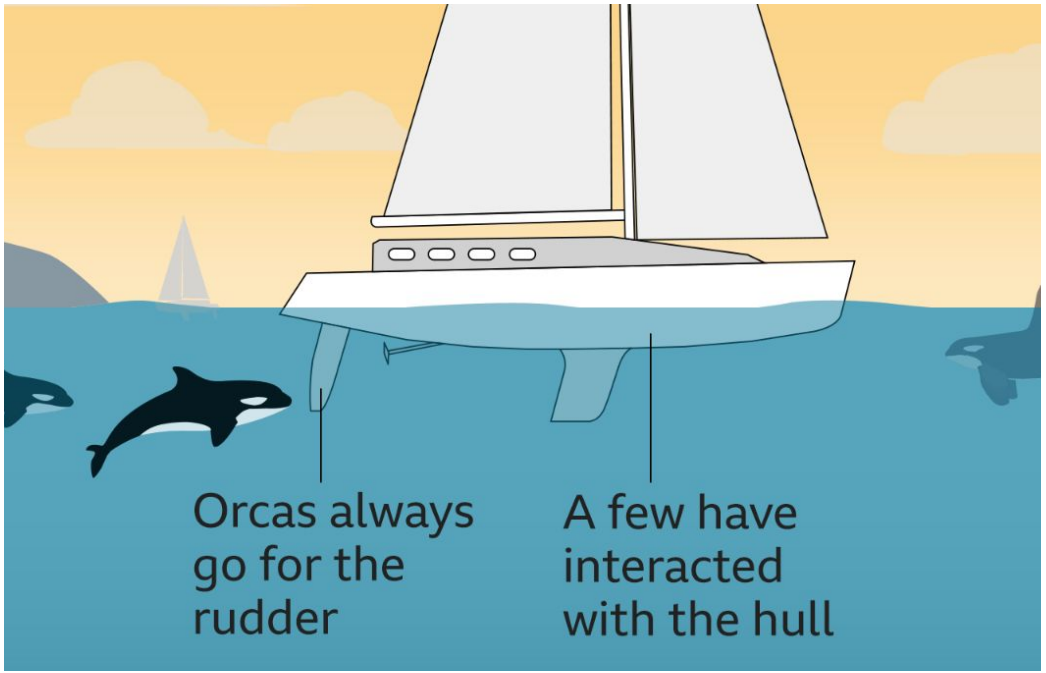
CASOS IDeaS
Carnegie Mellon University

The New York Times

Why Orcas Are Attacking Boats

A pod of orcas damaged a boat and left its two-person crew stranded. It was the latest in a string of attacks that research

For many problems, knowing *how* it happens isn't enough; we need to know *why*



Solving behaviorally complex problem requires that stakeholders converge on a decision