

DS 593: Privacy in Practice

Privacy in Tension: Content Moderation

News?



Sweden's Tax Authority Accused of Selling People's Data to Advertisers

April 4, 2025 By [Amar Ćemanović](#) — [Leave a Comment](#)



<https://cyberinsider.com/swedens-tax-authority-accused-of-selling-peoples-data-to-advertisers/>

Last time

- Exploring some of the tensions and trade-offs with regards to privacy
- The Exceptional Access Debate

Today

- Content moderation and Privacy

What is content moderation?

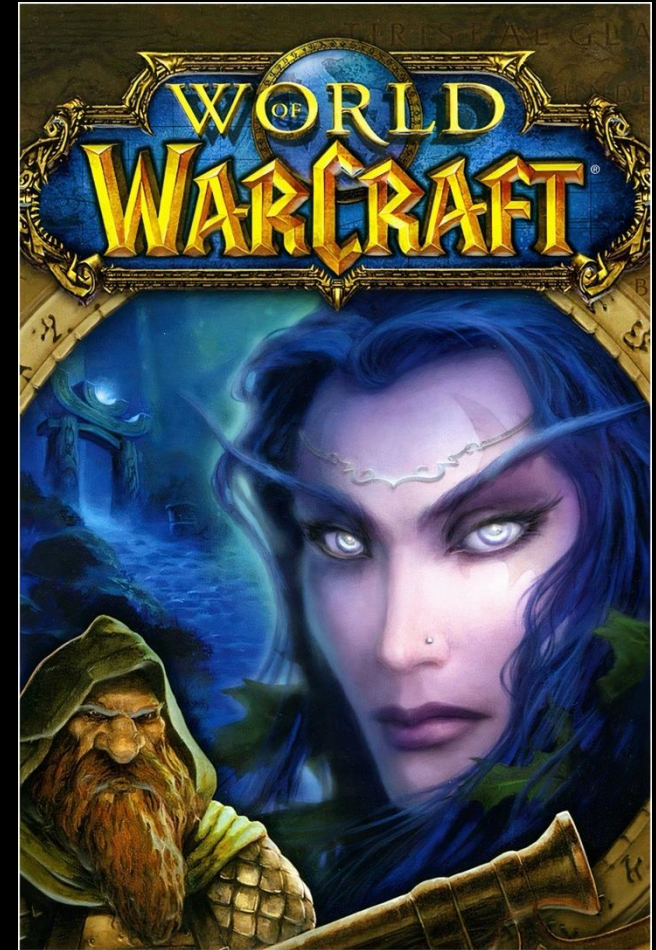
What is content moderation?

- Content moderation manages the experience of users in a digital space
- Essentially, how to make the space somewhere people want to be?
- Not quite a 1A issue

What is the content we don't want?

- Hateful content
- Harassment of users
- Illegal content
- Violent/disturbing content
- Other toxic content

How it started



<https://slate.com/culture/2024/12/video-games-world-warcraft-multiplayer-call-duty-halo.html>

How it's going





Type	Abuse mechanism	2016–2018 Global	2016–2018 US-only	Pew 2017 (US-only)	DS 2016 (US-only)	ADL 2018 (US-only)	DCI 2018 (Global)
Moderate	Been exposed to unwanted explicit content	19%	16%	–	–	–	23%
	Been insulted or treated unkindly	16%	14%	–	–	–	–
	Had someone make hateful comments	16%	14%	–	–	–	–
	Been called offensive names [†]	14%	13%	27%	25%	41%	20%
	Been concerned because specific information about me appeared on the Internet	11%	8%	–	–	–	–
Severe	Been stalked [†]	7%	5%	7%	8%	18%	5%
	Had an account hacked by someone I know	6%	3%	–	–	–	–
	Been sexually harassed [†]	6%	3%	6%	8%	18%	–
	Been harassed or bullied for a sustained period [†]	5%	4%	7%	5%	17%	4%
	Had someone post private photos of me to embarrass me	5%	3%	–	5%	–	3%
	Been impersonated by someone I know	5%	2%	–	6%	–	–
	Been physically threatened [†]	4%	2%	10%	11%	22%	5%
	Had someone I know use spyware to monitor my activities	4%	1%	–	–	–	4%
Aggregate	Been target of any online abuse	48%	35%	41%	36%	53%	40%
	Been target of any moderate online abuse	40%	32%	22%	–	–	–
	Been target of any severe online abuse	25%	13%	18%	–	37%	–

TABLE II: Frequency that participants reported experiencing hate and harassment online. We compare our results against previous surveys. We denote questions where the framing exactly matches a previous PEW survey with a dagger †. Our question framing differs from the other listed surveys, though the abuse mechanisms studied overlap.

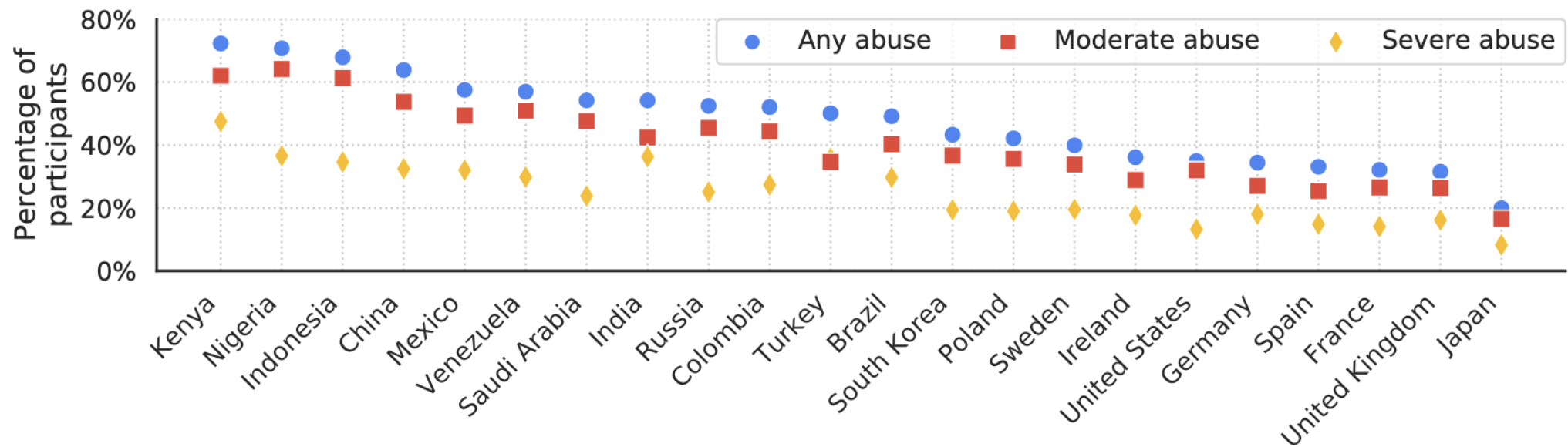
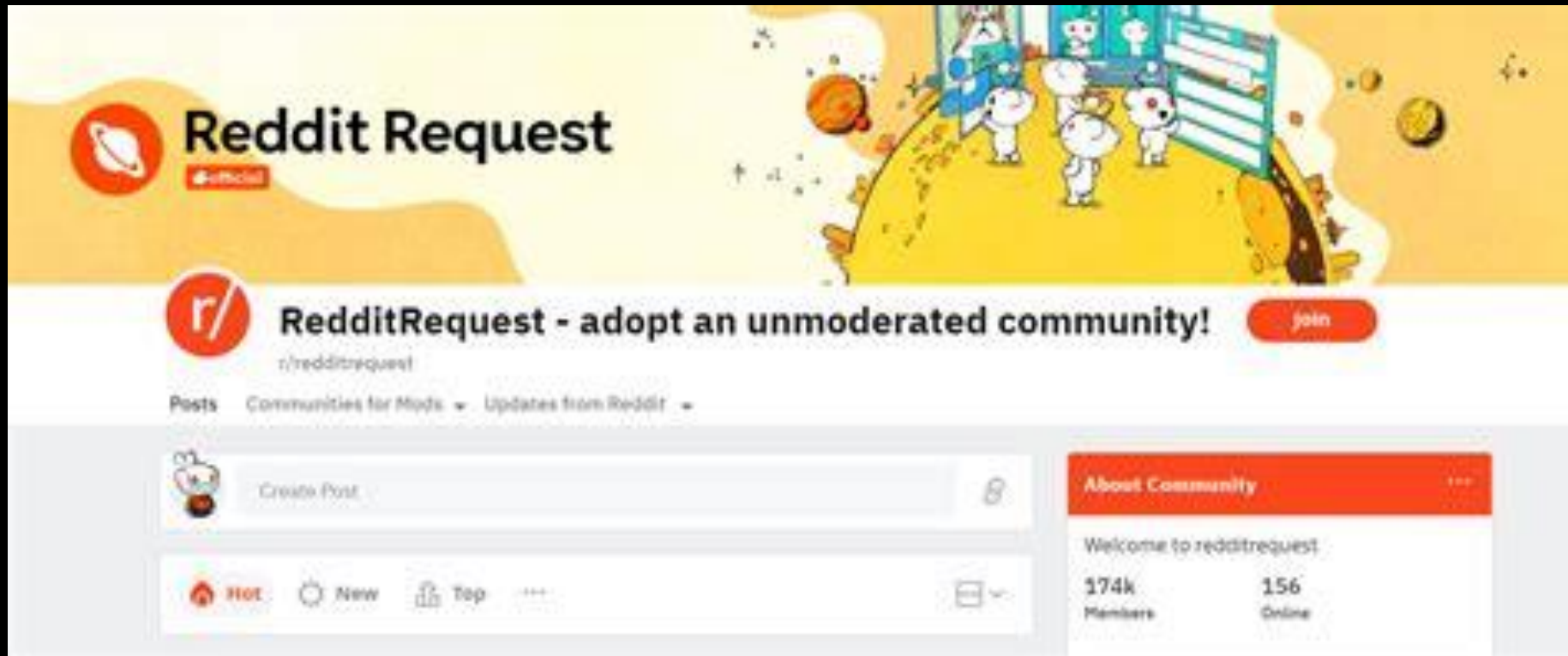


Fig. 2: Percentage of participants reporting any, moderate, or severe hate and harassment online per country, aggregated over 2016–2018.

How is content moderation performed?

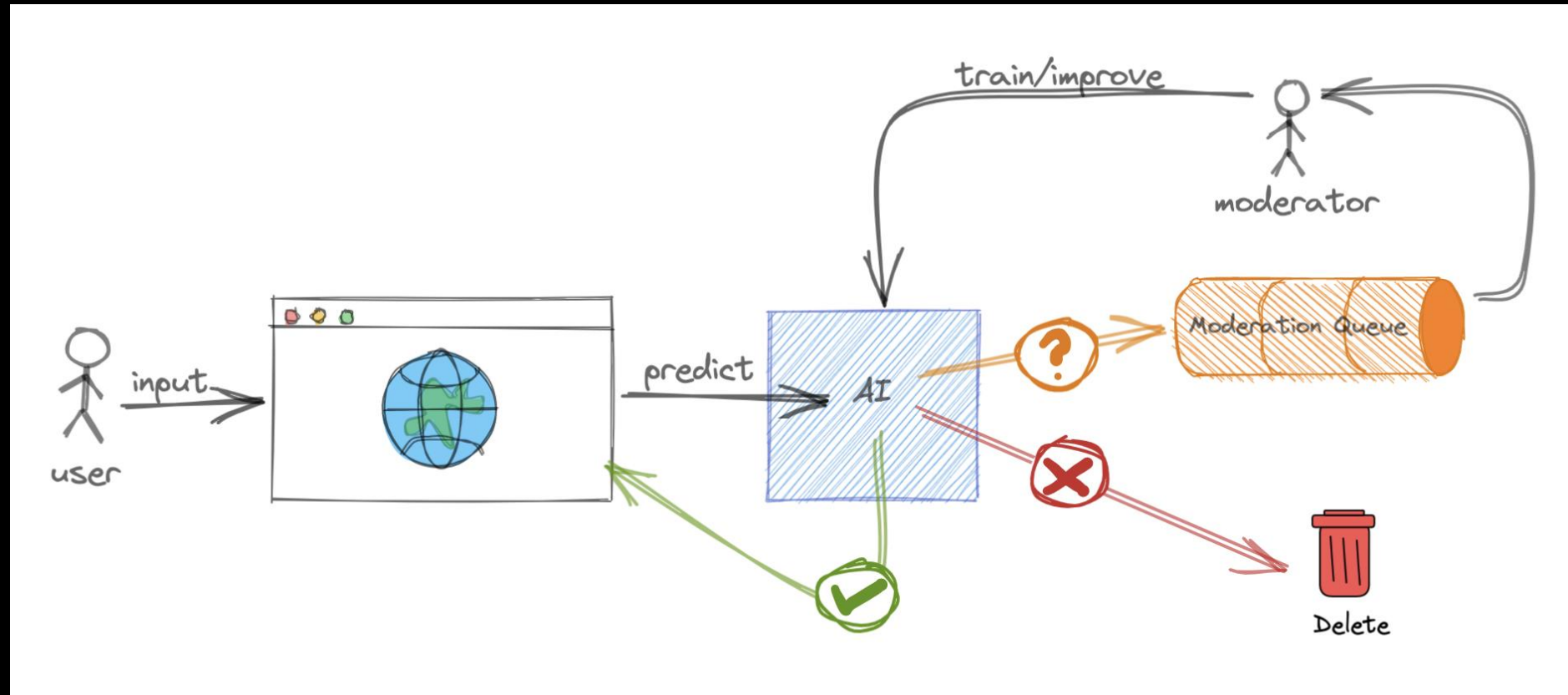
Community Moderation



Moderation Teams



Automated Moderation



Moderation Actions

- Reporting
- Flagging
- Removing
- Banning
 - Shadow banning
- Deplatforming

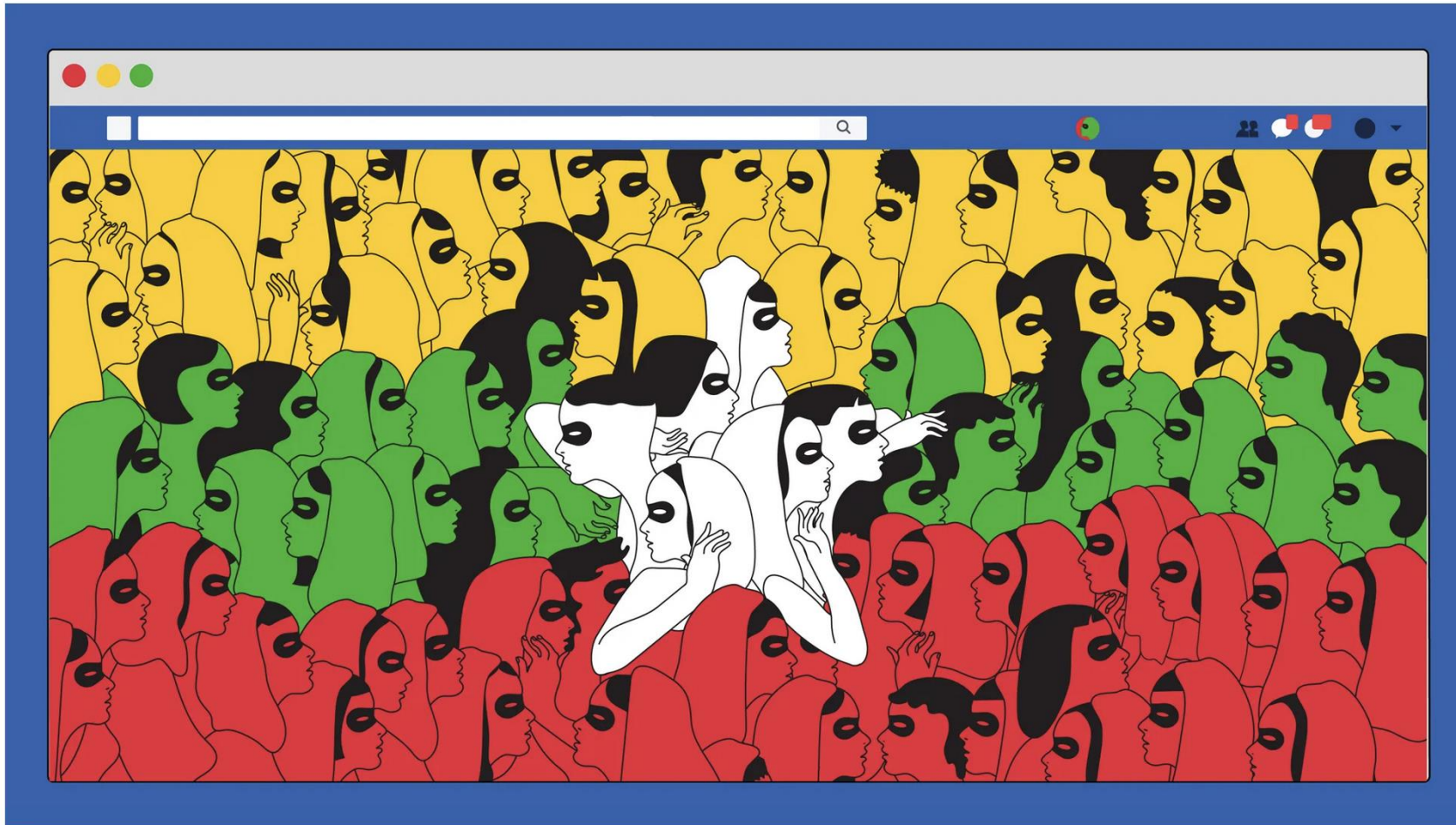
Core challenge: how to decide what is unacceptable

Core challenge: how to decide what is unacceptable

- Moderator Discretion
- Platform Discretion
- Community Standards
- Training a classifier

How Facebook's Rise Fueled Chaos and Confusion in Myanmar

The social network exploded in Myanmar, allowing fake news and violence to consume a country emerging from military rule.



LYDIA ORTIZ/PATRICK RAFANAN

<https://www.wired.com/story/how-facebooks-rise-fueled-chaos-and-confusion-in-myanmar/>

DOCUMENTING HATE


How One Major Internet Company Helps Serve Up Hate on the Web

Cloudflare, a prominent San Francisco outfit, provides services to neo-Nazi sites like The Daily Stormer, including giving them personal information on people who complain about their content.

by Ken Schwencke, May 4, 2017, 8 a.m. EDT

Cloudflare's CEO has a plan to never censor hate speech again

"We needed to change the conversation," CEO Matthew Prince told Ars.

TIMOTHY B. LEE - DEC 4, 2017 11:37 AM |  675

The Terror Queue

These moderators help keep Google and YouTube free of violent extremism – and now some of them have PTSD

OPINION

YouTube Is Erasing History

Under pressure to remove “extremist content,” platforms are purging vital human rights evidence.

Oct. 23, 2019

How LGBTQ+ Content is Censored Under the Guise of "Sexually Explicit"

BY JILLIAN C. YORK | AUGUST 18, 2021



Now mix in privacy

Now mix in privacy

- Fundamentally, content must be “seen” to be “moderated”
- Anonymous and E2EE systems limit the ability to perform traditional moderation
- Where does this leave us?

Apple's idea for combating CSAM

- E2EE deployment for iMessage lead to concerns of its potential for being abused and used to distribute CSAM
- Traditionally, this would be detected on the server as content passes through the service
- Server no longer sees it, could enforcement be done on device?

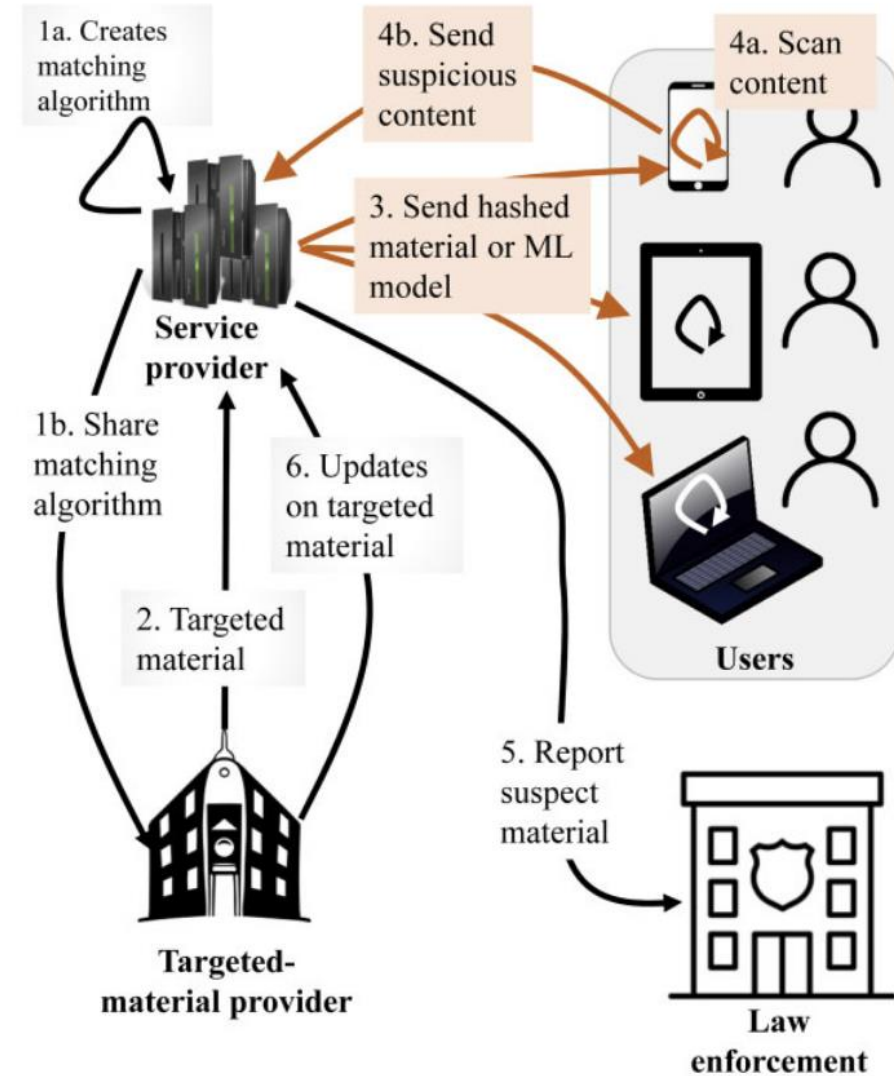
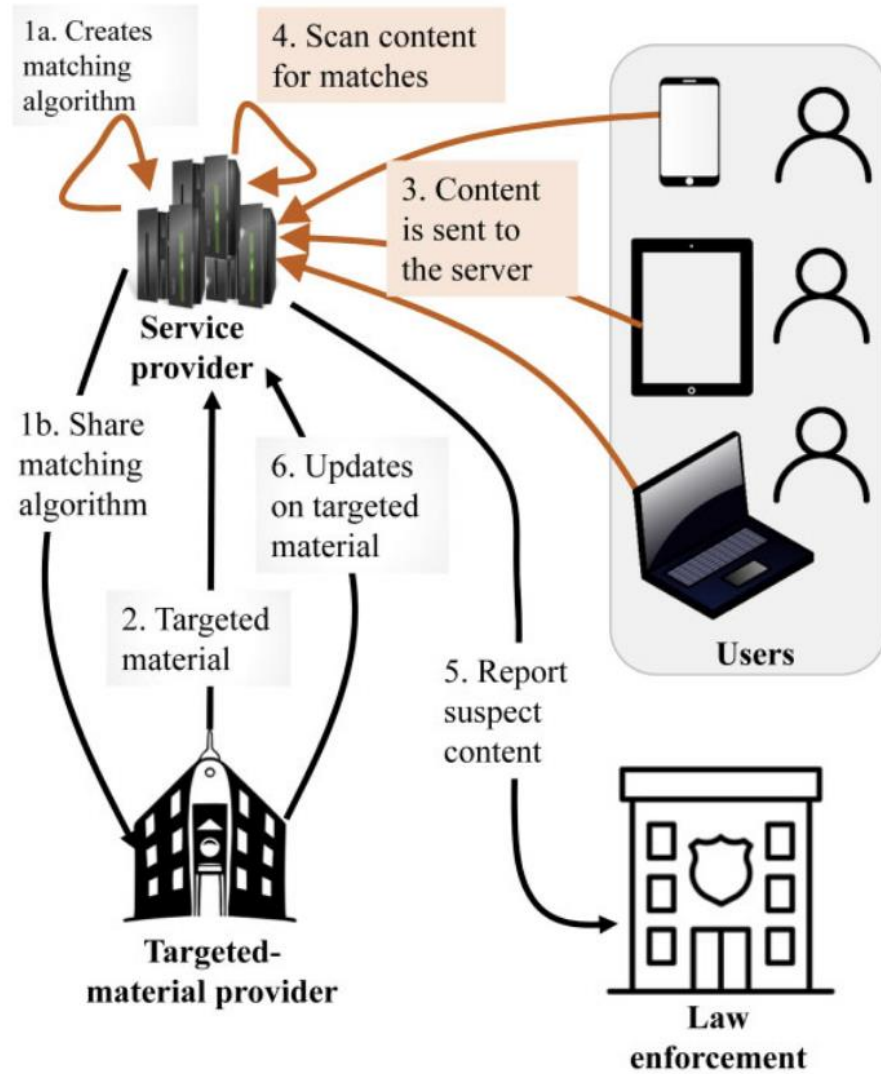


Figure 1. Scanning operation flows. *Left:* Server-side scanning. *Right:* Client-side scanning (the main changes are in orange).



Figure 3. Collisions of the NeuralHash function extracted from iOS 14. *Top:* A pair of accidentally colliding images in the ImageNet database of 14 million sample images; *Bottom:* An artificially constructed pair of colliding images.

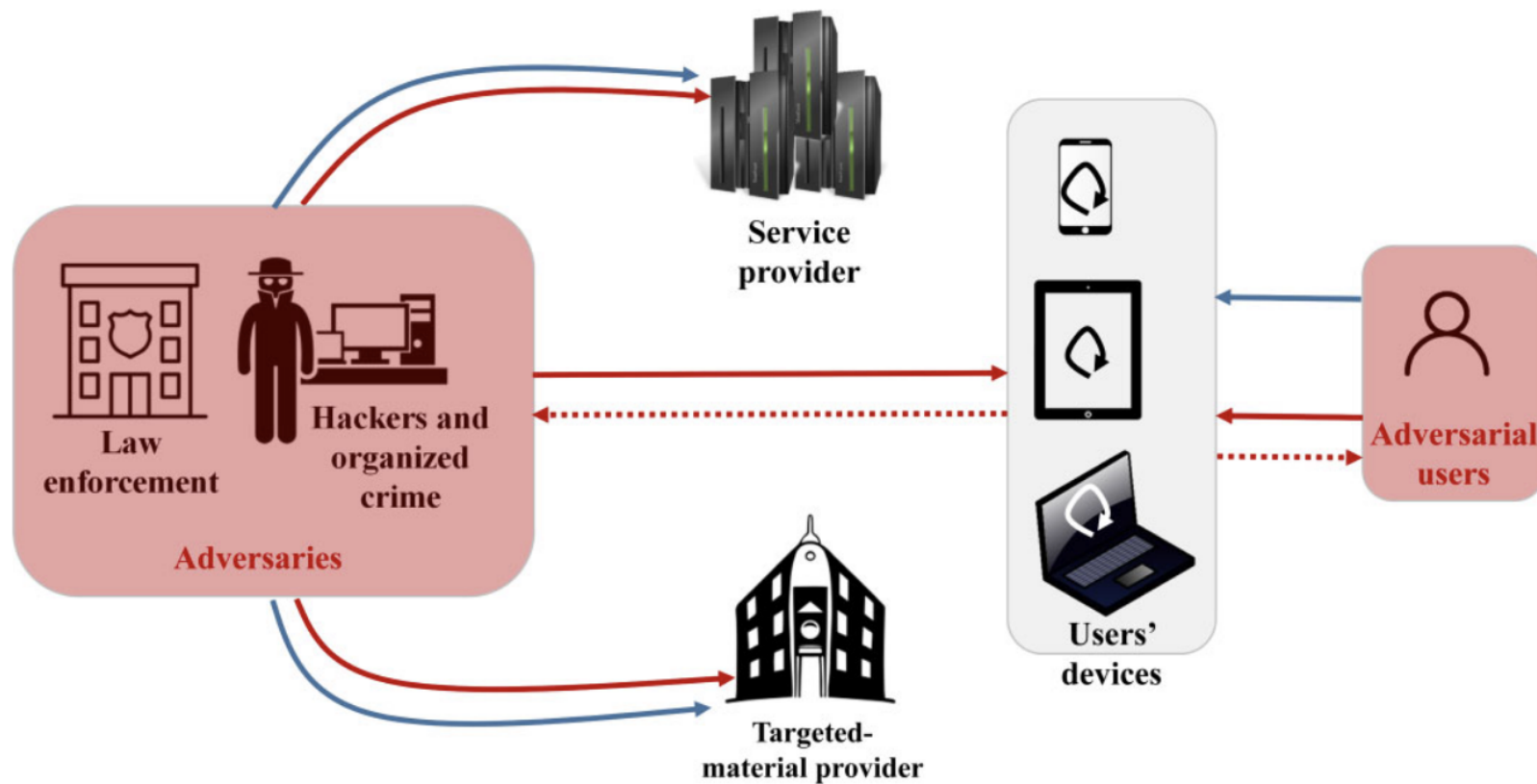


Figure 2. From server-side to client-side: New compromise paths and advantage points for adversaries (—→: compromise paths in server-side scanning; —→: compromise paths in CSS; - - -→: knowledge gained by adversary in CSS).

What is the message, only bad news?

What is the message, only bad news?

- The primary challenge is avoiding a system that is
 - Overly centralized
 - Too broad
 - Not transparent
 - Has large scale consequences if broken
 - Data-hungry

Promising Research Directions

- On device nudges
- Offline Moderation
- Community-local Moderation
- Anonymous Reporting Systems
- Thresholded Source Tracing for E2EE
- Zero-knowledge Credentials

Next Time

Privacy and Decentralization