

Moderation is mainstream

The Evolution of Trust & Safety pre- and post 2017

Today's session

- 01 Why I started a content moderation newsletter
- 02 A very brief history of Trust & Safety
- 03 What I'm watching in 2025

*“There’s something **strange and also amazing** about the fact that content moderation (a job that was not too long ago described as ‘the worst job going’) is now being discussed, seriously, on a US conservative radio show with 13.5 million weekly listeners. How far we've come.”*

***‘Moderation is mainstream’**, Everything in Moderation*, August 9, 2018*



EVERYTHING IN MODERATION*



EVERYTHING IN MODERATION*



EVERYTHING IN MODERATION*

In numbers

309

weekly editions

4200

newsletter subscribers

75

episodes of *Ctrl-Alt-Speech*

What readers/listeners say...

"The best compilation of information and sources out of many newsletters I am subscribed to. All in one place, the best snippets and also the author's thoughts along the way makes reading complicated topics more entertaining"

"Ben has a lot of context and understanding about the industry and pulls relevant resources together (with some informed commentary). I always recommend to people who work in the industry or who want to learn more about the industry"

"...like two internet sheriffs have deputized themselves to call out the sometimes asinine and semi-ridiculous attempts by well-meaning but ill-informed humans to "make the internet safer."

A brief history of Trust & Safety

The internet's first commercial spam?

Path: gmd.de:link.net:rz.uni-karlsruhe.de:news.uni-stuttgart.de:news.be
From: nikel@indirect.com (Laurence Canter)
Newsgroups: rec.juggling.us.legal
Subject: Green Card Lottery- Final One?
Date: 12 Apr 1994 08:12:17 GMT
Organization: Canter & Siegel
Lines: 34
Message-ID: <2od151\$45t@herald.indirect.com>
NNTP-Posting-Host: id1.indirect.com

Green Card Lottery 1994 May Be The Last One!
THE DEADLINE HAS BEEN ANNOUNCED.

The Green Card Lottery is a completely legal program giving away a certain annual allotment of Green Cards to persons born in certain countries. The lottery program was scheduled to continue on a permanent basis. However, recently, Senator Alan J Simpson introduced a bill into the U. S. Congress which could end any future lotteries. THE 1994 LOTTERY IS SCHEDULED TO TAKE PLACE SOON, BUT IT MAY BE THE VERY LAST ONE.

PERSONS BORN IN MOST COUNTRIES QUALIFY, MANY FOR FIRST TIME.

The only countries NOT qualifying are: Mexico; India; P.R. China; Taiwan, Philippines, North Korea, Canada, United Kingdom (except Northern Ireland), Jamaica, Dominican Republic, El Salvador and Vietnam.

Lottery registration will take place soon. 55,000 Green Cards will be given to those who register correctly. NO JOB IS REQUIRED.

THERE IS A STRICT JUNE DEADLINE. THE TIME TO START IS NOW!!

For FREE information via Email, send request to cslaw@indirect.com

Canter & Siegel, Immigration Attorneys
3333 E Camelback Road, Ste 250, Phoenix AZ 85018 USA
cslaw@indirect.com telephone (602)661-3911 Fax (602) 451-7617

“THERE IS A STRICT JUNE DEADLINE. THE TIME TO START IS NOW!!”

For FREE information via Email, send request to cslaw@indirect.com”

Usenet, April 12, 1994



[home](#) | [my eBay](#) | [site map](#) | [sign in](#)

[Browse](#) | [Sell](#) | [Services](#) | [Search](#) | [Help](#) | [Community](#)

[overview](#) | [news](#) | [chat](#) | [newsletter](#) | [library](#) | [charity](#) | [eBay store](#) | [about eBay](#)

 Buyers! **\$1 Off** every time you use eBay Online Payments!

[Smart Search](#)

☐ Search titles **and** descriptions

About eBay

▼ Press Releases

[December 1999](#)
[November 1999](#)
[October 1999](#)
[September 1999](#)
[August 1999](#)
[July 1999](#)
[June 1999](#)
[May 1999](#)
[April 1999](#)
[March 1999](#)
[February 1999](#)
[January 1999](#)
[December 1998](#)
[November 1998](#)
[October 1998](#)
[September 1998](#)
[August 1998](#)
[July 1998](#)
[June 1998](#)
[May 1998](#)



Press Releases

You've come to the right source for a complete archive of company press releases. Feel free to peruse eBay's many announcements about new product features, business initiatives, on-site promotions, business partnerships, user services, and much more.

- [01/29/99](#) - eBay Soars To New Heights
- [01/29/99](#) - eBay's response to the investigation by the New York City Department of Consumer Affairs
- [01/25/99](#) - Compaq and eBay Offer Easier Access To Leading Online Trading Community
- [01/25/99](#) - A Picture Is Worth A Thousand Words
- [01/15/99](#) - eBayTM Launches The Most Comprehensive Trust And Safety Upgrades To The World's Largest Person-To-Person Trading Site
- [01/05/99](#) - Mark McGwire and Sammy Sosa 1998 Home Run Balls To Be Auctioned Online and On-Live

*“While less than 1/100 of 1 percent of the millions of transactions have a reported fraud complaint, eBay takes every one seriously. Since it was founded in 1995, eBay has instituted a number of safety tools and resources including the Feedback Forum in Feb. '96 and SafeHarbor in Feb. '98. to actively work with the community and law enforcement agencies. eBay is now expanding this industry-leading program and providing even more with **new tools and safeguards to make the site a safe place to trade online.**”*

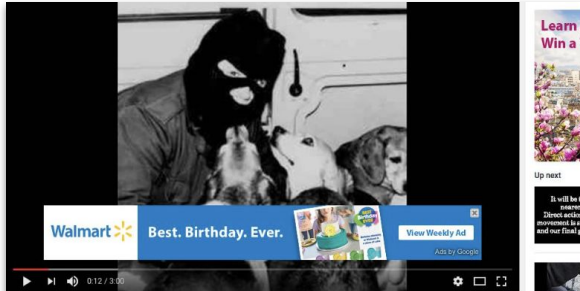
eBay, January 15, 1999

“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.”

Section 230 of the Communications Decency Act, 1996

Media flash points?

THE TIMES



Rise Against - Black Masks & Gasoline [Animal Liberation Front]

The extremist Animal Liberation Front will have made money from adverts for Walmart which have appeared on its videos. The company declined to comment

VIDEO

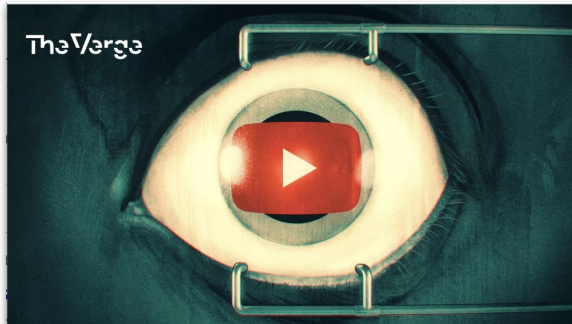
Top brands pull Google adverts in protest at hate video links

Alexi Mostrous, Head of Investigations | [James Dean](#), US Business Editor

Thursday March 23 2017, 12.00am, The Times

2017

The Verge



GOOGLE

THE TERROR QUEUE

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

By Casey Newton | [@CaseyNewton](#) | Dec 16, 2019, 11:00am EST

2019

THE WALL STREET JOURNAL.

the facebook files

A Wall Street Journal investigation

[Facebook](#) Inc. knows, in acute detail, that its platforms are riddled with flaws that cause harm, often in ways only the company fully understands. That is the central finding of a Wall Street Journal series, based on a review of internal Facebook documents, including research reports, online employee discussions and drafts of presentations to senior management.

Time and again, the documents show, Facebook's

2021

What I'm watching in 2025

The overdue promise of transparency

DSA Transparency Database overview



9 503 459 849

Total number of statements of reasons submitted



Most Reported Violations

1. Scope of platform service
2. Unsafe and/or illegal products
3. Illegal or harmful speech



Top Restriction Types

1. Disabling access to content
2. Removal of content
3. Other restriction (please specify)



118

Number of active platforms



48%

of fully automated decisions

A user appeals revolution?

*“TikTok’s recent adoption of decisions by User Rights represents a meaningful shift. The platform has **acknowledged restrictions and reinstated content and accounts previously removed....** In another instance, User Rights found an account suspension unjustified due to the absence of a serious policy violation or repeated misconduct and recommend the account’s reinstatement’*

User Rights, December 18, 2025



Press Release

Out-of-Court Dispute Settlement Shows Impact: TikTok Implements User Rights’ Decisions

Berlin, December 18th, 2024 – TikTok has become the first platform to implement decisions issued by [User Rights](#) under the EU’s Digital Services Act (DSA), marking a major milestone in establishing out-of-court dispute settlement for social media platforms.

In August 2024, User Rights was certified as the first German dispute settlement body under Article 21 of the DSA. It is the first body Europe-wide focusing on social media platforms. Since then, the organization has reviewed and resolved hundreds of cases submitted by users. Before the creation of out-of-court dispute settlement for social media platforms, users had limited options for challenging moderation decisions, relying either on internal platform mechanisms or pursuing costly legal action.

TikTok’s recent adoption of decisions by User Rights represents a meaningful shift. The platform has acknowledged unjustified restrictions and reinstated content and accounts previously removed. “TikTok’s decisions don’t always align with its own guidelines,” explains Niklas Eder, co-founder of User Rights. “For example, TikTok mistakenly classified a satirical meme as hate speech. Our review determined that satire, as a protected form of expression, did not violate their policies.” In another instance, User Rights found an account suspension unjustified due to the absence of a serious policy violation or repeated misconduct and recommended the account’s reinstatement.

This development underscores the value of independent out-of-court dispute resolution. “Protecting fundamental rights, especially freedom of expression in the digital space, is crucial,” emphasizes Raphael Kneer, co-founder of User Rights. “We’re proud to be Europe’s first certified dispute settlement body and to play a key role in defending these rights.”

User Rights offers an independent and external mechanism for reviewing platform decisions. Users can submit complaints free of charge to challenge content restrictions or to address platforms’ refusals to remove potentially illegal content.

The light and shade of AI

"Despite the ubiquity of moderation, performing the task currently remains labor intensive, even with the help of machine learning. We believe that Large Language Models (LLMs) like GPT-4 represent the greatest change to the dynamics of content moderation in at least a decade. **LLMs can now directly automate the core activity of moderation** – the classification of content according to a set of written, human intelligible, policies."

Dave Willner, January 29, 2024



"Recently I was made aware that AI of 'me' falsely endorsing Donald Trump's presidential run was posted to his site. It really conjured up my **fears around AI, and the dangers of spreading misinformation**. It brought me to the conclusion that I need to be very transparent about my actual plans for this election as a voter. The simplest way to combat misinformation is with the truth."

Taylor Swift, September 11, 2024



Professionalisation of T&S

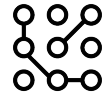
“As the industry formalises, we also recommend further professionalisation of the Trust and Safety sector through the **use of standardised trainings and potential qualification pathways**. We expect this to provide Trust and Safety professionals with a greater sense of professional accountability within the sector like lawyers and accountants who have duties to provide for their clients but also to uphold professional standards”

Technology and Trust and Safety Report, October 2024

Source: [Department of Science, Innovation and Technology](#)



Global Internet Forum to Counter Terrorism (2017)



Tech Against Terrorism (2017)



All Tech Is Human (2018)



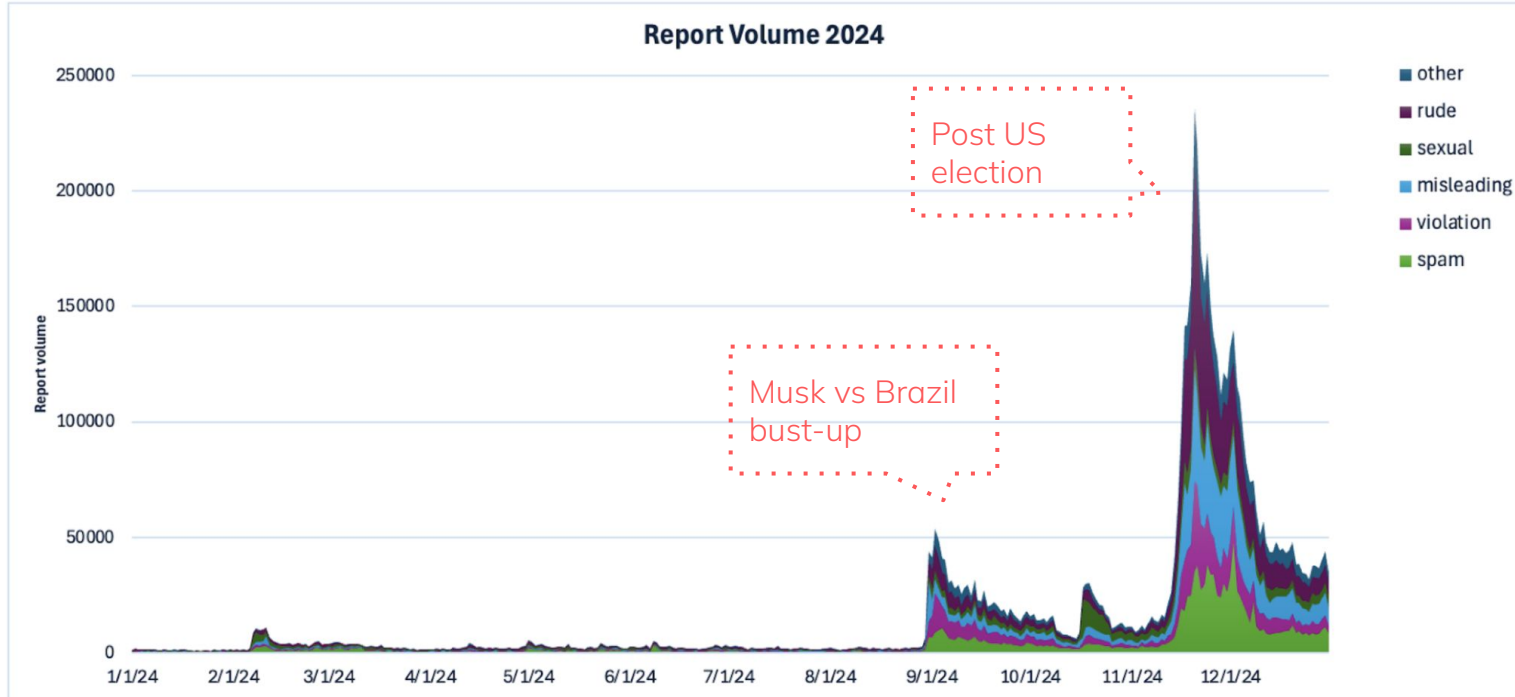
Trust and Safety Professional Association (2020)



Digital Trust and Safety Partnership (2021)

Decentralisation means everyone is a moderator

Bluesky report volume in 2024





EVERYTHING IN MODERATION*

Q&A

Thank you for inviting me today

Ben Whitelaw

Founder and editor, *Everything in Moderation**

Co-host, *Ctrl-Alt-Speech* podcast

ben@everythinginmoderation.co