How do Tech Companies Work & the Basics of Trust and Safety Design

CS 152 — Trust and Safety

Alex Stamos

Stanford Cyber Policy Center

April 2, 2026

Practical Stuff

- Milestone 1 will be posted Friday. Please attend section next week to hear from your TA about the project and brainstorm ideas.
- Be creative.
- Waitlist drama…

What will you learn today?

- The critical roles in a tech company
- Basic concepts in online product design
- The lifecycle of a trust and safety issue
- The basics of measuring Trust and Safety
- The tradeoffs inherent in metrics

Surfaces & Responses

Basic definitions

Surface: a technical "entry point" to interact with a system

- One product can have multiple surfaces. Facebook has:
  - Public profiles
  - Private groups
  - Live video streaming
  - Private messaging

Features -> Affordances -> Outcomes

- Features: "design elements that offer specific types of capabilities offered by the system"
- Affordances: "possibilities for action available in a given environment"
- Outcomes: "actions or other behaviors connected with the goals of the user"

Surfaces are for users and abusers

When we think about entry points for abuse in a system…

- What design elements do people create content in vs. consume content from?
- What do these design elements let people do?
- What do people do with them, and what are the results?

Example surfaces

Product Type	Example Products	Example Surfaces
Social networks	X, Threads, Bluesky, Instagram	profiles, feeds
Social news / discussion	reddit, Quora	comments, voting
Messaging platforms	WhatsApp, Discord, Signal, Telegram	private messaging, audio, video
Video hosting	YouTube, TikTok, PornHub	video, comments
E-Commerce	Amazon, Shein, FB Marketplace	product search, purchasing, ratings
Gaming platforms	Minecraft, Roblox, Steam	3d worlds, chat, audio
Generative AI	ChatGPT, Midjourney, Flux	text, image, video generation

ABC framework

- **Actors**: abusive actors
  - Who are these users? What is their intent? What networks are they situated in?
- **Behaviors**: deceptive behaviors
  - What do the actors do? What are actions that repeat offenders take?
- **Content**: harmful content
  - What kinds of content is created or viewed that can be harmful?

ABC framework

- **Actors**: abusive actors
  - Who are these users? What is their intent? What networks are they situated in?
- **Behaviors**: deceptive behaviors
  - What do the actors do? What are actions that repeat offenders take?
- **Content**: harmful content
  - What kinds of content is created or viewed that can be harmful?

ABC framework

- **Actors**: abusive actors
  - Who are these users? What is their intent? What networks are they situated in?
- **Behaviors**: deceptive behaviors
  - What do the actors do? What are actions that repeat offenders take?
- **Content**: harmful content
  - What kinds of content is created or viewed that can be harmful?

ABC framework

- **Actors**: abusive actors
  - Who are these users? What is their intent? What networks are they situated in?
- **Behaviors**: deceptive behaviors
  - What do the actors do? What are actions that repeat offenders take?
- **Content**: harmful content
  - What kinds of content is created or viewed that can be harmful?

ABC framework

- **Actors**: abusive actors
  - Who are these users? What is their intent? What networks are they situated in?
- **Behaviors**: deceptive behaviors
  - What do the actors do? What are actions that repeat offenders take?
- **Content**: harmful content
  - What kinds of content is created or viewed that can be harmful?

ABC framework

- **Actors**: abusive actors
  - Who are these users? What is their intent? What networks are they situated in?
- **Behaviors**: deceptive behaviors
  - What do the actors do? What are actions that repeat offenders take?
- **Content**: harmful content
  - What kinds of content is created or viewed that can be harmful?

DEF responses

- What are the **design** visions that engineers, designers, and other technology workers imagine will impact behavior when they are building these systems (ie., that will "afford" certain technology-mediated activities within their platforms)?
- What **evaluations** do technology workers conduct and measure to understand if users will adopt those activities… or if actors will take other actions to create abusive behaviors/content?
- What **features** are ultimately produced, launched, and scaled within technology platforms? Which features are successful? Where do they fail because actors manipulate them to produce abusive behaviors/content?

DEF responses

- What are the **design** visions that engineers, designers, and other technology workers imagine will impact behavior when they are building these systems (ie., that will “afford” certain technology-mediated activities within their platforms)?
- What **evaluations** do technology workers conduct and measure to understand if users will adopt those activities… or if actors will take other actions to create abusive behaviors/content?
- What **features** are ultimately produced, launched, and scaled within technology platforms? Which features are successful? Where do they fail because actors manipulate them to produce abusive behaviors/content?

DEF responses

- What are the **design** visions that engineers, designers, and other technology workers imagine will impact behavior when they are building these systems (ie., that will “afford” certain technology-mediated activities within their platforms)?
- What **evaluations** do technology workers conduct and measure to understand if users will adopt those activities… or if actors will take other actions to create abusive behaviors/content?
- What **features** are ultimately produced, launched, and scaled within technology platforms? Which features are successful? Where do they fail because actors manipulate them to produce abusive behaviors/content?

AI as a surface

Generative AI tools are not just another distribution channel — they are **creation surfaces**

- The platform *is* the tool that produces harmful content
- AI-generated CSAM: NCMEC received 440,419 reports in H1 2025 vs. 6,835 in all of 2024
- LoRA models can create realistic deepfakes of specific children from ~20 photos in 15 minutes
- Every surface category in the table above now has an AI-generated content problem

This changes the ABC framework: the “Actor” may be the AI system itself, not just the user

Roles & Lifecycle

Key Trust & Safety Teams

Content Policy: Responsible for developing content policies that outline what content is allowed on a platform.

- Reflects company values and user sensibilities
- Aims to comply with legal and regulatory requirements
- Protects community voice
- Provides recommendations to leadership on policy situations

Typical Roles: Content Policy Manager, Policy Analyst, Knowledge Management, Public Policy Manager

Key Trust & Safety Teams

Content Design and Strategy: Identifies optimal strategy for user-facing content and develops effective language to communicate with users.

- User-education material
- Help center articles
- Product interventions

Typical Roles: Content Strategist, Content Designer

Key Trust & Safety Teams

Data Science and Analytics: Build measurement methods to understand policy violations and their impact.

- Analyze the impact of content moderation and abuse detection efforts
- Predict policy violation trends through data analysis
- Develop creative ways to address adversarial behavior

Typical Roles: Data Scientist, Data Engineer, Data Analyst

Key Trust & Safety Teams

Engineering: Build and maintain technical systems for content moderation and policy enforcement.

- Develop ML models to scale/automate policy enforcement
- Build systems for user-facing reporting options (e.g., DMCA takedowns)
- Create internal review tools and technical infrastructure

Typical Roles: Software Engineer, Security Engineer

Key Trust & Safety Teams

Legal: Manage responses to official requests and advise on legal risks.

- Respond to requests from law enforcement and regulatory bodies
- Proactively identify and mitigate potential legal issues
- Advise product teams on legal issues for existing and planned products

Typical Roles: General Counsel, Cybersecurity Law and Investigations, Regional Experts, Subject Matter Experts (copyright, privacy, etc.)

Key Trust & Safety Teams

Law Enforcement Response and Compliance: Process legal requests and handle sensitive escalations.

- Review and assess legal requests from law enforcement
- Respond to sensitive or urgent escalations
- Coordinate with internal and external partners to assist people in crisis

Typical Roles: Law Enforcement Response Analyst, Investigations Analyst, Incident Response Analyst

Key Trust & Safety Teams

Operations: Handle day-to-day moderation and develop scalable processes.

- Manage content moderation professionals or vendor relationships
- Oversee quality assurance, training, and workflow management
- Develop review protocols and manage crisis/incident response

Typical Roles: Project Manager, Program Manager, Vendor Manager, Analyst, Investigator, Specialist

Key Trust & Safety Teams

Product Policy: Develop and refine principles and policies specific to particular products.

- Introduce policy nuances for individual products (e.g., Ads)
- Counsel internal product teams on policy considerations
- Provide practical product strategies across multiple jurisdictions

Typical Roles: Product Policy Manager, Product Policy Associate

Key Trust & Safety Teams

Product Management: Drive strategy, vision, and execution for preventing policy violations.

- Partner with cross-functional teams on policy areas or focus areas
- Develop strategies to address specific issues (e.g., hate speech)
- Execute strategies through collaboration with other teams

Typical Roles: Product Manager

Key Trust & Safety Teams

Public Policy and Communications: Build partnerships and manage stakeholder relations.

- Maintain relationships with NGOs, governments, and regulatory bodies
- Advise on regional public policy matters
- Design campaigns to shape public and political opinion

Typical Roles: Public Policy Manager

Key Trust & Safety Teams

Sales and Advertiser Support: Address advertisers' concerns about policy violations.

- Help brands avoid appearing alongside objectionable content
- Communicate content safety measures to advertising partners
- Balance commercial interests with platform safety requirements

Typical Roles: Client Partners, Industry Managers, Vertical Leads

Key Trust & Safety Teams

Threat Discovery and Research: Investigate networks of abuse and research bad actor behavior.

- Research emerging patterns of abuse
- Analyze networks of coordinated malicious actors
- Collaborate with internal teams and external parties (e.g., law enforcement)

Typical Roles: Abuse Investigators, Threat Intelligence Investigators

The State of Trust and Safety in 2025–2026

The T&S profession is in crisis:

- Major platforms **gutted T&S teams** in 2024–2025: Google cut ~1/3 of Jigsaw, Meta shifted 700 staff to Reels growth while refusing child-protection hires
- At TrustCon 2025, TSPA Executive Director Charlotte Willner: *"so many things are not getting better"*
- No major T&S leader publicly condemned the rollbacks (Platformer investigation)
- Regulatory obligations are **expanding** (EU DSA, UK Online Safety Act) even as teams shrink

Students entering this field face a paradox: the work has never been more important, and the industry commitment has never been more uncertain.

AI is Replacing Human Moderators

In March 2026, Meta announced AI-driven enforcement systems replacing third-party human moderation vendors

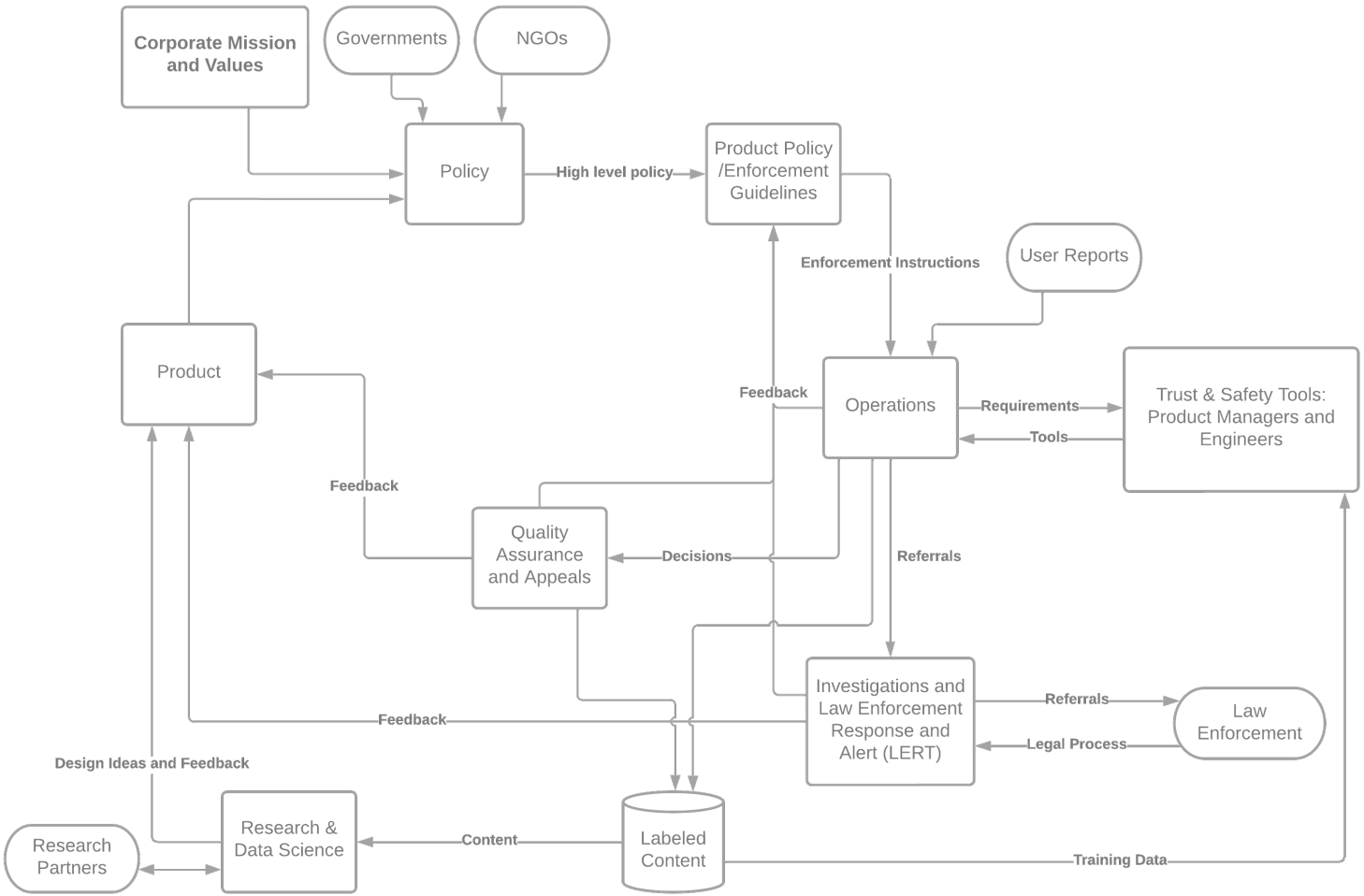
- One system detects 5,000 phishing attempts daily
- Handles terrorism, CSAM, drugs, fraud, and scams

But the Oversight Board flagged ongoing problems:

- AI struggles with context, cultural nuance, and language complexity
- Over-removal of legitimate content (satire, journalism, protest)
- Under-detection of harmful content using evasion techniques
- Content moderation market: $11.63B (2025) → projected $23.2B (2030)

The Operations role is shifting from “manage vendor relationships” to “manage AI pipelines with human review for edge cases”

Trust and safety lifecycle



Word of the Day: chinchilla

Measurement

Why measure abuse?

Without proper measurement, abuse fighting is shooting in the dark.

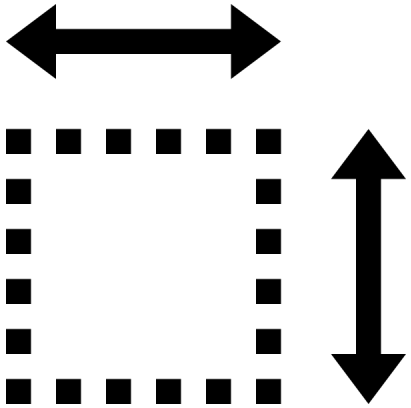


Figure 1: Evaluate the size of abuse

Figure 2: Evaluate the efficiency of abuse mitigations

Measurement for internal management

T&S is perceived by many executives as a "cost center" and drag on product

- It takes money from the company through disabling ads, slowing development, blocking accounts, reducing engagement.

A key use of measurement is to support investment in T&S work

- Goals for T&S: user trust, civil communication, healthy interactions, sense of safety
- Measurement can be part of demonstrating that these goals are additive to the business

Measurement to judge effectiveness

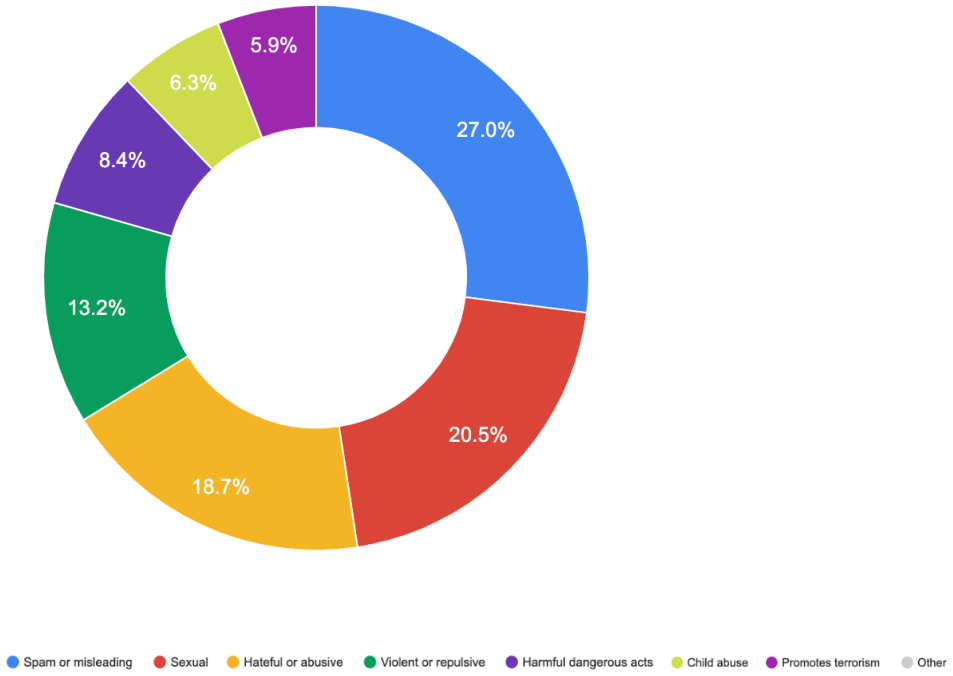
How can we know T&S is doing a good job? That users trust the platform and feel safe?

- User surveys (e.g. Edelman Trust Barometer, safety surveys)
- Challenge: these surveys are subjective, and sometimes noisy.
- What data can be more subjective?

Measuring impact

Abuse leads to poor user experience, has impact on user trust and in some cases leads to user harm.

Human flags by flagging reason (YouTube Transparency Report)



Measuring impact

Top user harms that results from a data breach

Potential harm	Breakdown	N
Identity theft	52%	287
Leak of personal information	25%	138
Monetary loss	9%	50
Loss of access to personal information	5%	28
Phone being monitored by hackers	3%	17
Computer being infected with virus	3%	17
Spam being sent out from your account	2%	11
Other	1%	4
No harm	< 1%	2

Measurement Methods

Internal Methods

- Collecting and measuring reports and actions
- Random sampling and tagging
- Behavioral metrics

External Methods

- Crawling and scraping
- Crowdsourcing
- User metrics
- User experience studies
- Prosecution statistics (for criminal actions)

Case Study: Community Notes vs. Fact-Checking

In January 2025, Meta replaced its third-party fact-checking program with Community Notes

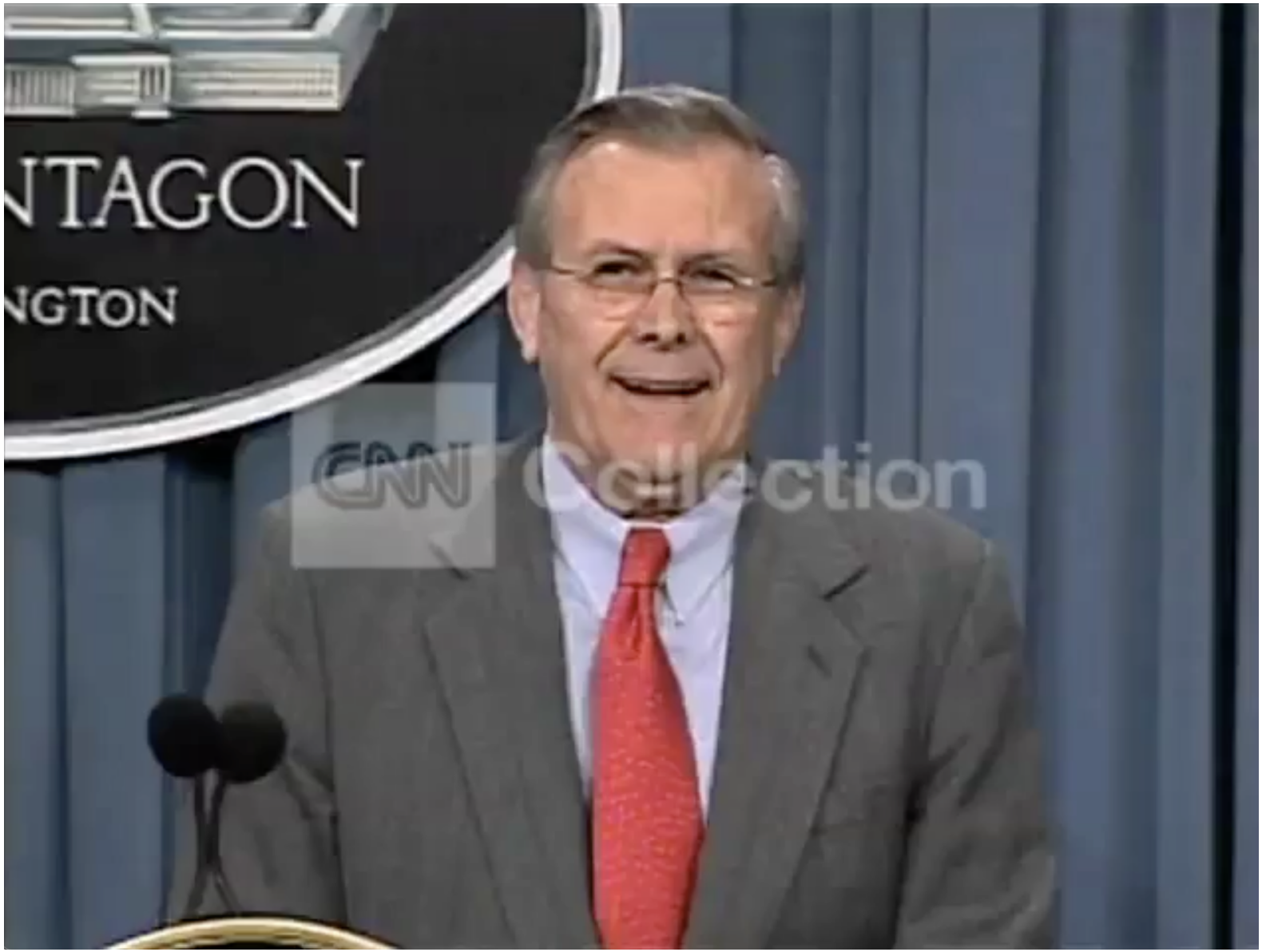
The measurement gap is stark:

	Community Notes (U.S.)	Professional Fact-C
Labels in 6 months	~900	~35 million
Avg. time to appear	65.7 hours	Hours
Coverage	Low	Broad

- In March 2026, the Oversight Board warned Community Notes are *"not a proper substitute"* for fact-checking
- Notes arrive well after content reaches peak visibility
- Serious risks in conflict zones, elections, and repressive regimes

What does this tell us about measuring the effectiveness of content moderation?

Unknown Known Unknowns



AI-Generated CSAM: A Measurement Crisis

AI-generated CSAM reports to NCMEC:

- **2024**: 6,835 reports
- **H1 2025**: 440,419 reports

IWF found a **26,385% increase** in AI-generated CSAM videos (13 in 2024 → 3,443 in 2025)

This illustrates the **Unknown Unknown** problem:

- Existing measurement frameworks were built for *distributed* content, not *generated* content
- Hash-matching (PhotoDNA, CSAI Match) cannot detect AI-generated material — it has no known hash
- New detection methods must be developed for content that has never existed before

False positives and false negatives

Let us take the example of a binary email spam classifier, in which each email can be either spam or not spam.

Confusion Matrix

	Pred: Real	Pred: Spam
Real	True Neg (TN)	False Pos (FP)
Spam	False Neg (FN)	True Pos (TP)

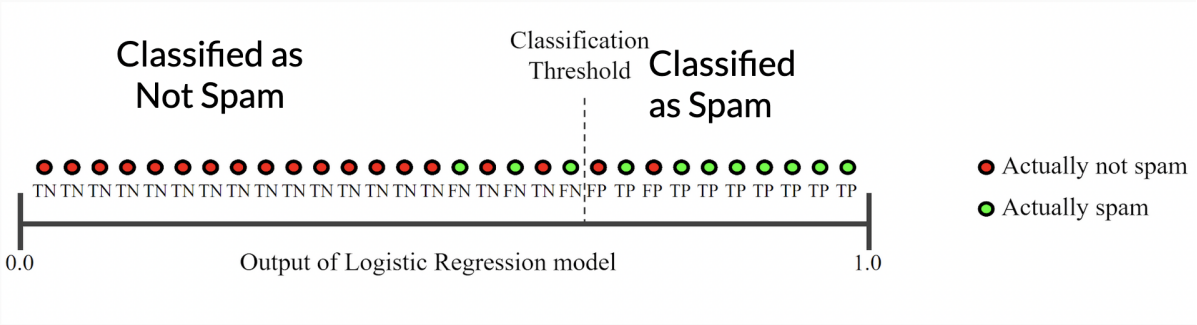
Precision and recall: a constant tug of war

Precision = True Positive/(True Positive + False Positive)
The proportion of positive identifications that are correct

Recall = True Positive/(True Positive + False Negative)
The proportion of actual positives that were identified correctly

Accuracy = Number of Correct Predictions/Total Predictions
The fraction of predictions the model got right

Prevalence = False Negative/Total
The proportion of bad content that got past your system



Precision/Recall in Practice: AI Moderation at Meta

Meta's March 2026 shift to AI-first moderation is a live experiment in this tradeoff:

High recall, reasonable precision:

- Terrorism, CSAM — known patterns, hash databases, low ambiguity
- 5,000 phishing attempts caught daily by one system

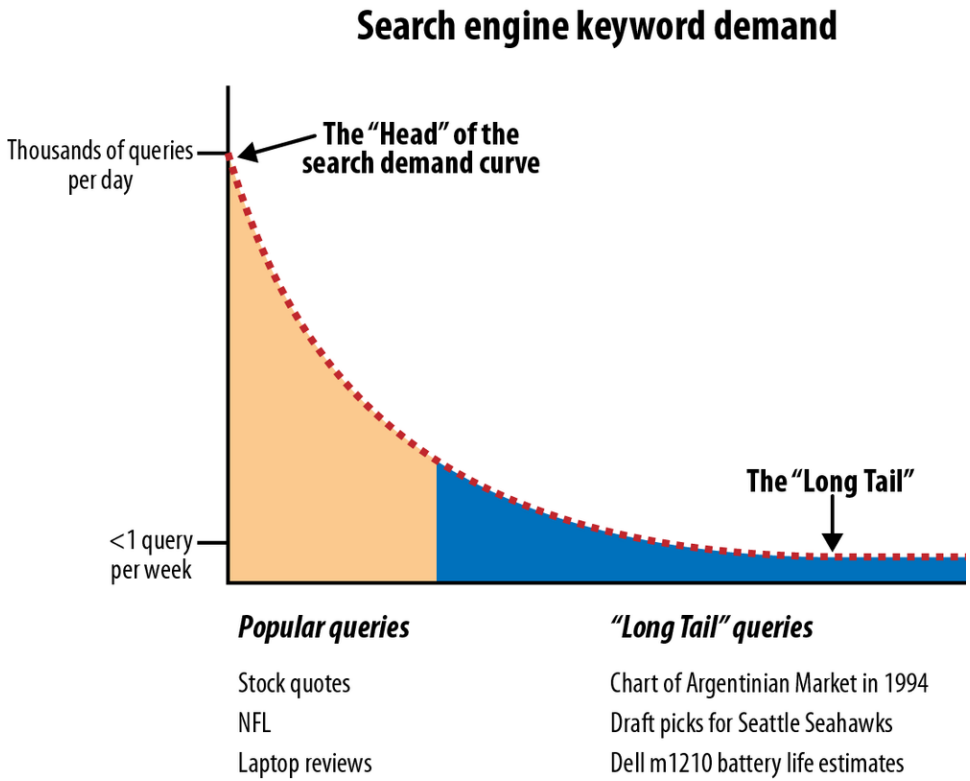
High recall, low precision:

- Political speech, satire, protest imagery
- Oversight Board documented over-removal of Palestinian, Armenian, and Ethiopian content
- Journalists and activists disproportionately affected

Low recall:

- Novel evasion techniques, coded language, cultural context
- AI-generated harmful content with no prior training examples

Problem of long-tails



The majority of online searches comprise unique rather than popular queries.

- Videos that get less than 10 viewers
- Blog post that have zero followers
- Websites that show up on the 10th page of the search results

"Two basic principles of management, and regulation, and life, are:

You get what you measure.

The thing that you measure will get gamed."

- Matt Levine, Bloomberg

Metric Perversity

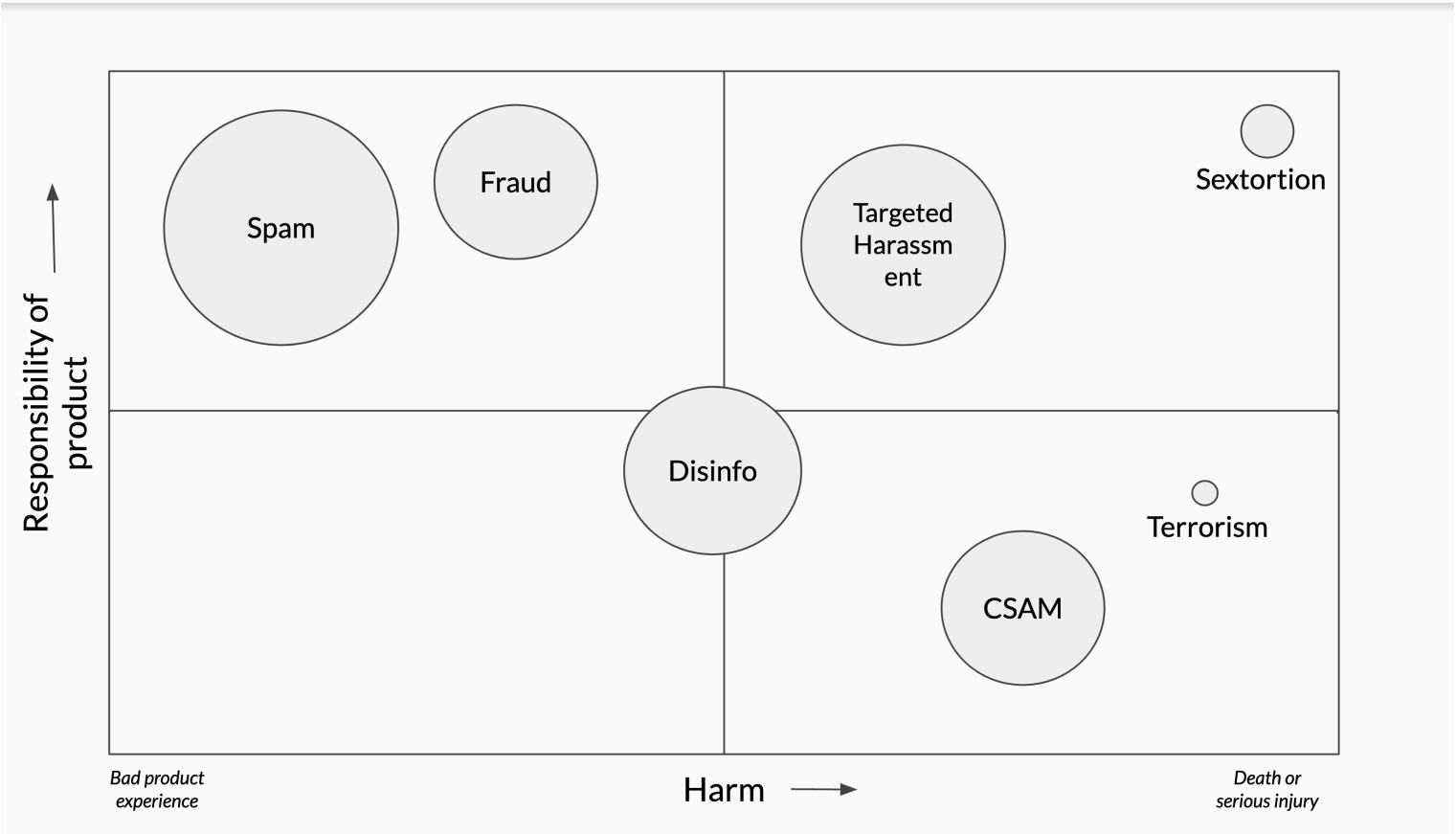
Metrics are supposed to help you make good decisions

- Are your metrics respectful?
- Are they perverse?
- Will they cope with rare, important failures?
- Are they robust to data collection issues?

Many safety problems come from bad metrics

Making Decisions and Tradeoffs

Using Measurements to Make Decisions



Prioritization in Trust and Safety

The **level** of harm associated with an area of abuse

+

The **prevalence** of the abuse

+

The product's **responsibility** for enabling such abuse

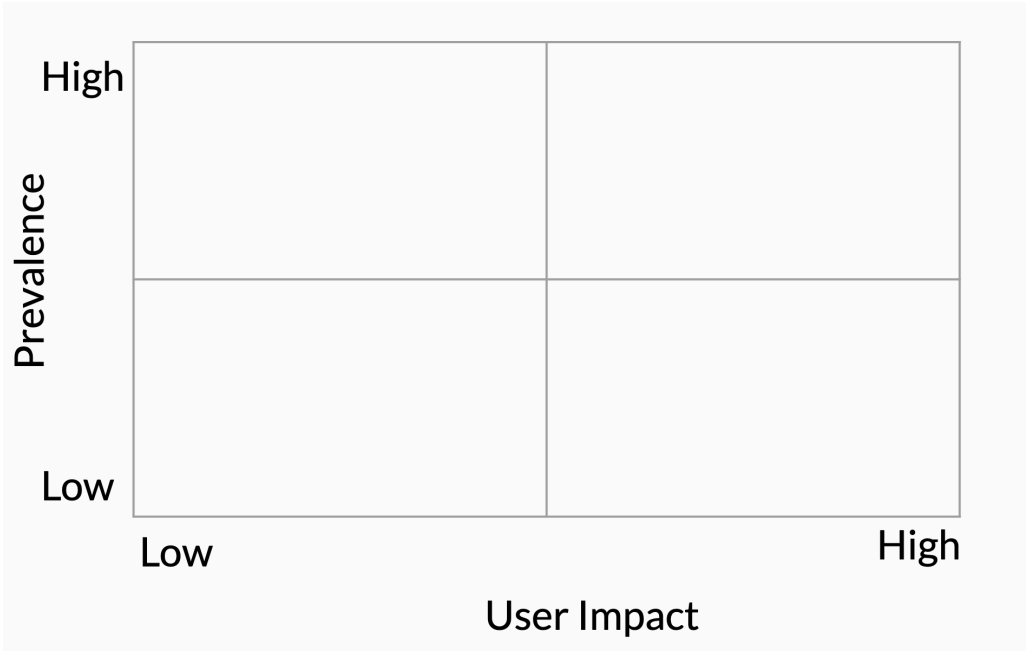
=

Prioritization

ACTIVITY - Prioritization

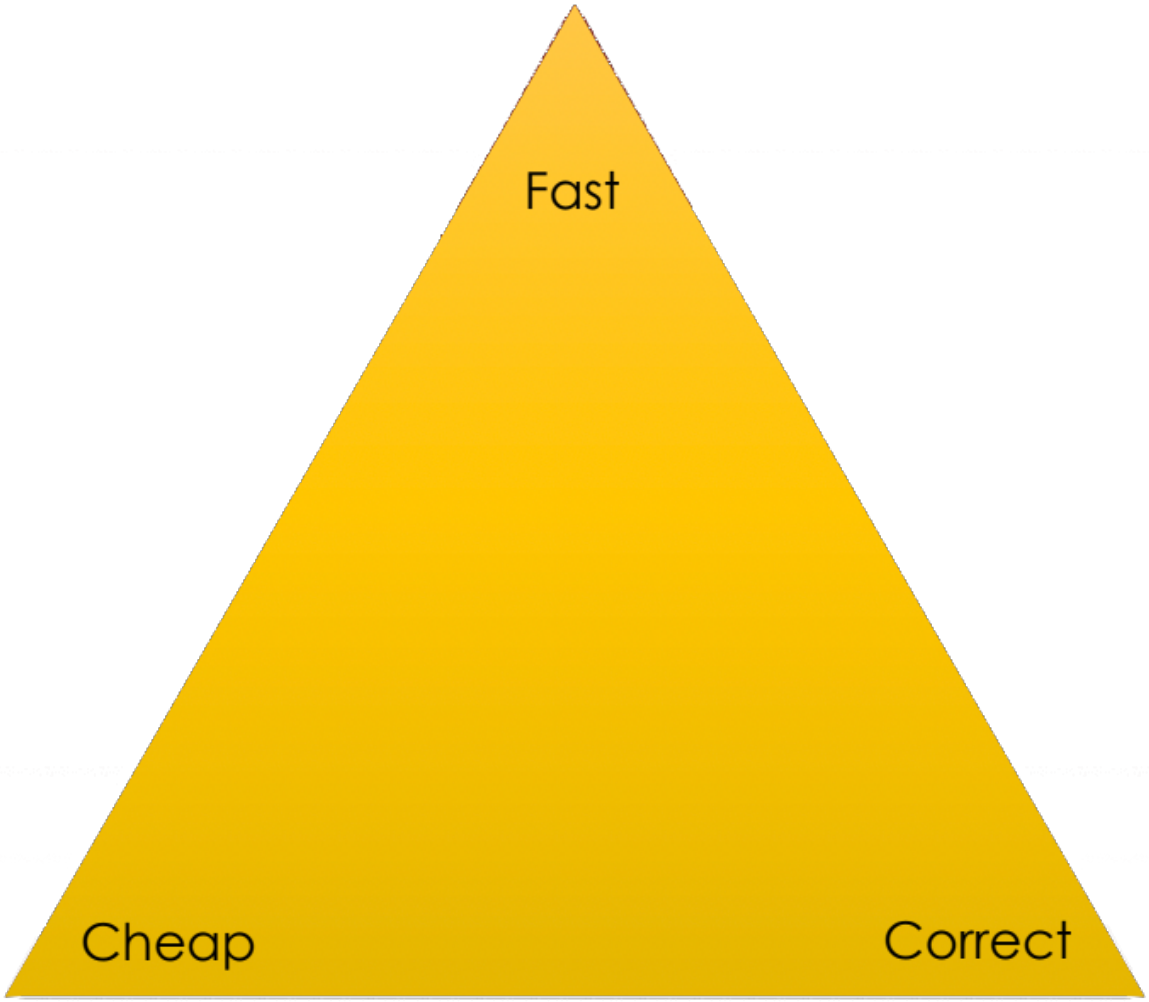
1. Hateful comments on a blog posts
2. Violent images on a photo sharing platforms
3. Suicide how to videos on a video sharing platforms
4. Clickbait ads on websites
5. False stories from low-quality news sources
6. Offensive words on reviews of Apps on a mobile store
7. AI-generated deepfake CSAM on an image generation platform

Based on your understanding of the abuse types listed above, place them on the matrix below

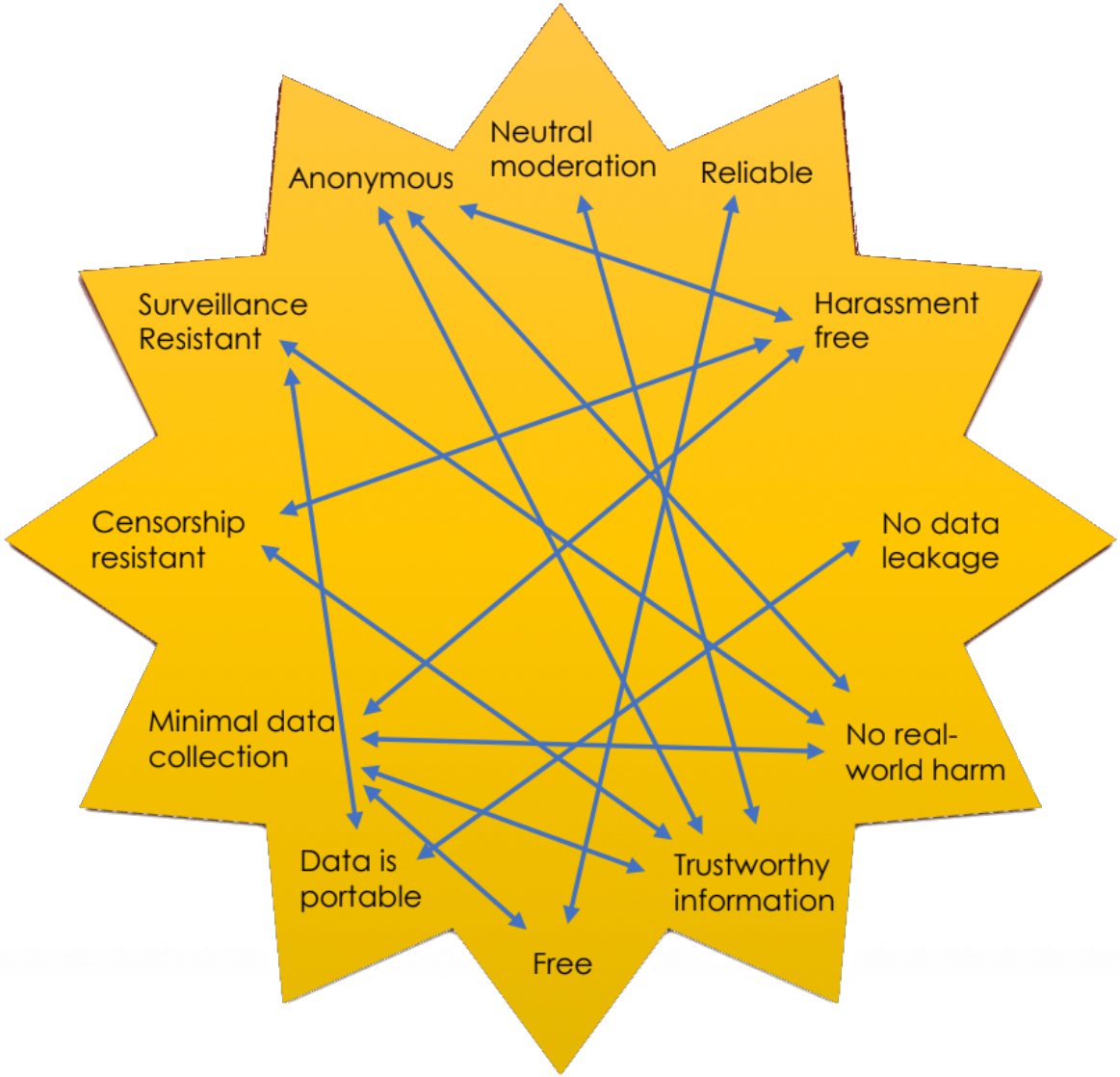


Tradeoffs

The Iron Triangle of Engineering



The Stamos Tridecagram of Difficult Tradeoffs



Landmark: Platforms Liable for Design (March 2026)

On March 25, 2026, a California jury found **Meta (70%) and YouTube (30%) liable** for addictive platform design

- $6 million in compensatory and punitive damages
- Jury found both companies were negligent in design, knew designs were dangerous, and failed to warn users
- TikTok and Snapchat settled before trial
- Hundreds of similar cases are pending

This reshapes the tradeoffs discussion: courts are now holding platforms liable for **product design decisions**, not just individual content moderation choices

Regulatory Enforcement is Real

EU Digital Services Act: Fined X (Twitter) €120M in December 2025

- Misleading blue checkmarks implying trustworthiness
- Poor advertising transparency
- Restricted researcher data access

UK Online Safety Act: Active enforcement since July 2025

- Ofcom rejected self-declaration as age verification
- Fined AVS Group £1M for missing age checks
- Penalties up to £18M or 10% of global turnover

Compliance is no longer optional — it is a cost of doing business in major markets

Inherent Tradeoffs

- Under-Enforcement vs. Over-Enforcement
  - E.g. privacy v. anti-abuse scanning
  - E.g. designating political terrorist groups
- Legal Obligations vs. Awful but Lawful
- Compliance Costs vs. Market Access
  - EU DSA, UK OSA, and state-level laws create a patchwork of obligations
  - Failure to comply now carries real financial penalties
- Every company has finite resources

Questions?