Who Moderates the Moderators?

A Look Inside Bluesky

SICSS-Saarbrücken

September 2025

Abhisek Dash, Pushpdeep Singh

Caution : Some slides may contain explicit content.

However, I believe this will make us all appreciate the complexity.

Plan for today

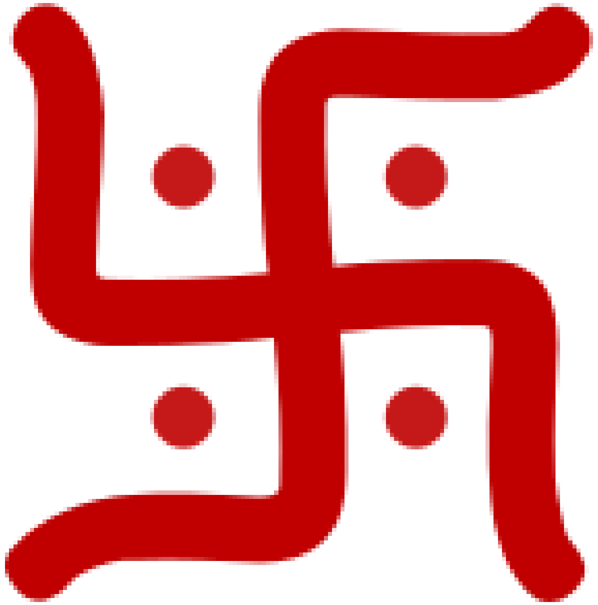
- Content moderation
- The centralized black-box
- A decentralized alternative: Bluesky
- Content moderation on Bluesky
- Our role as CSS researchers
- Hands on session (in the afternoon)

Content moderation

Before we proceed …

If someone posted the following image on your favorite social media platform. Should the platform take down the post?

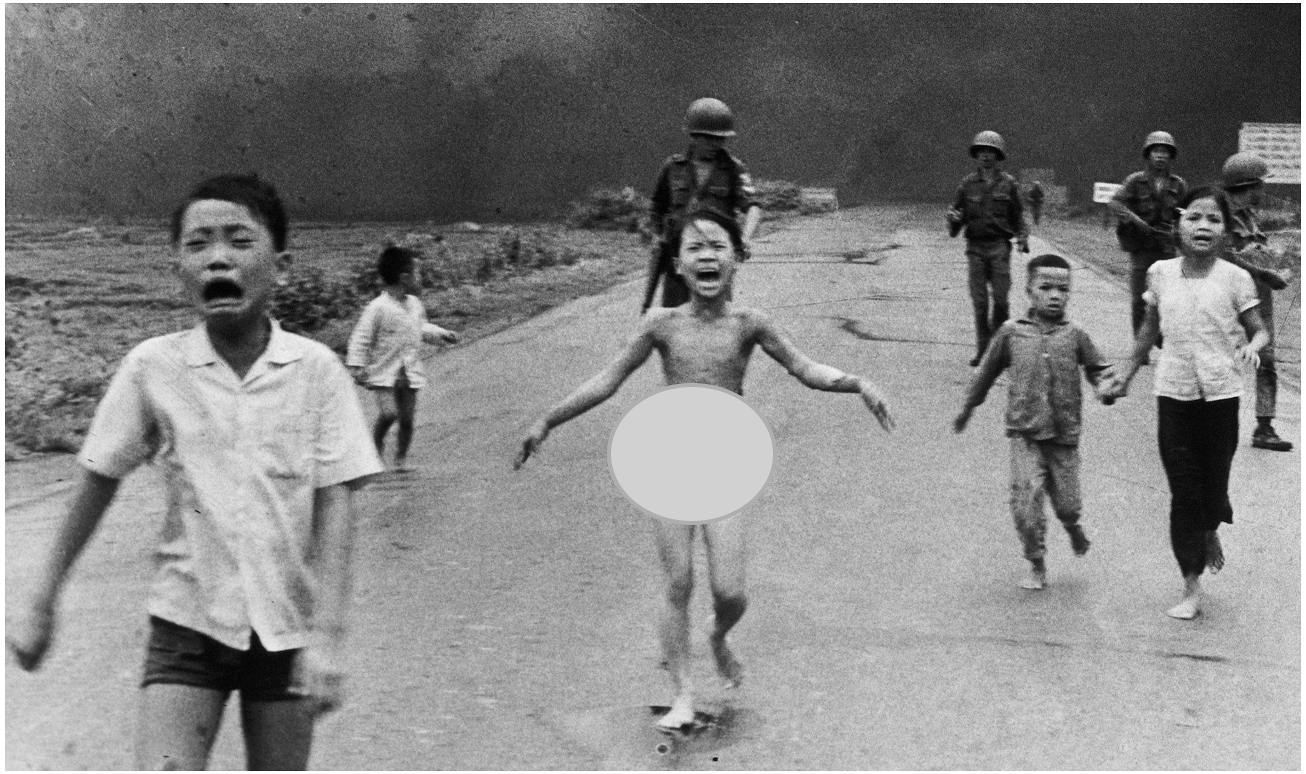
Should the platform take down this image?



Different meaning in different cultures

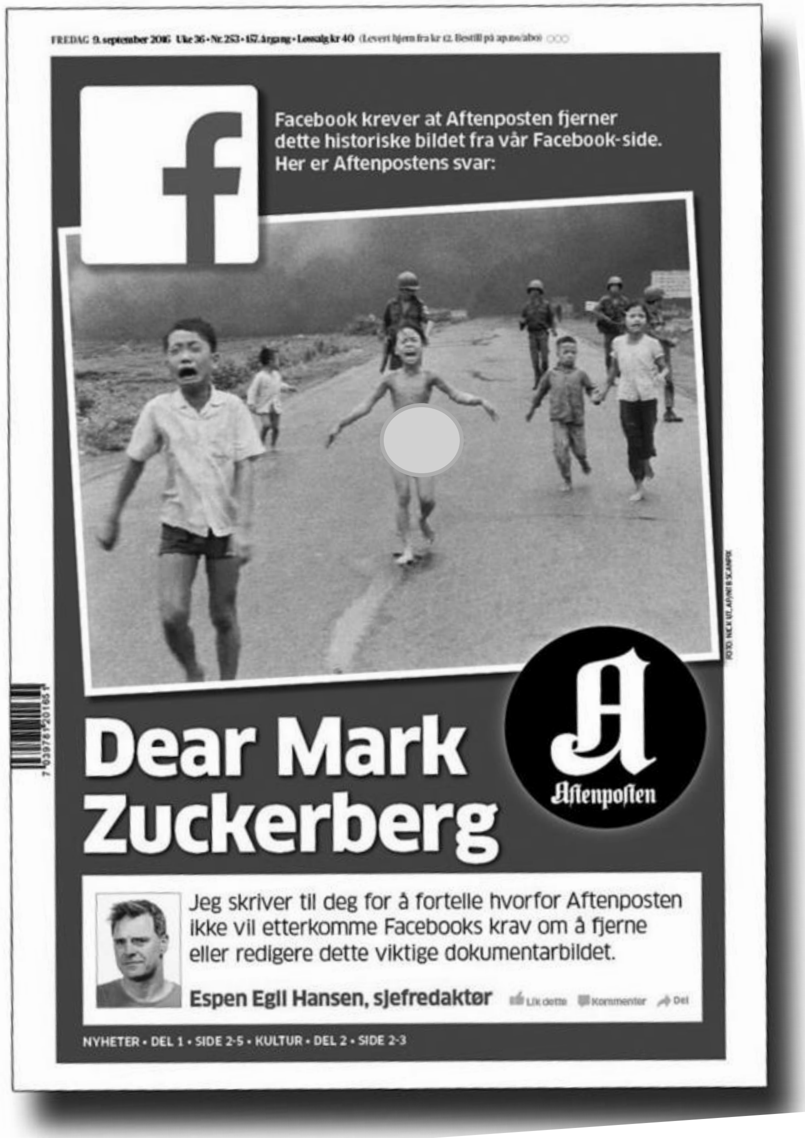
In Hinduism, the right-facing symbol (clockwise) (卐) is called swastika, symbolizing sun, prosperity and good luck.

Should the platform take down this image?



A bit of a history …

- ❑ **Napalm Girl** : 1972 Pulitzer Prize winning photo by Nick Ut (AP).
- ❑ **2016** : Norwegian journalist Tom Egeland included it in an article reflecting on photos that changed the history of warfare.
- ❑ The picture contains graphic suffering and underage nudity.
- ❑ **Consequence** : Facebook moderators deleted the post.
- ❑ The shown snapshot is the outrage of the editor in chief of Aftenposten on the front page of the newspaper.



FREDAG 9. september 2016 Uke 36 • Nr. 253 • 157. årgang • Løssalg kr 40 (Levert hjem fra kr 12. Bestill på ap.no/abo)

Facebook krever at Aftenposten fjerner dette historiske bildet fra vår Facebook-side. Her er Aftenpostens svar:

Dear Mark Zuckerberg

Aftenposten

Jeg skriver til deg for å fortelle hvorfor Aftenposten ikke vil etterkomme Facebooks krav om å fjerne eller redigere dette viktige dokumentarbildet.

Espen Egil Hansen, sjefredaktør

Lik dette Kommenter Del

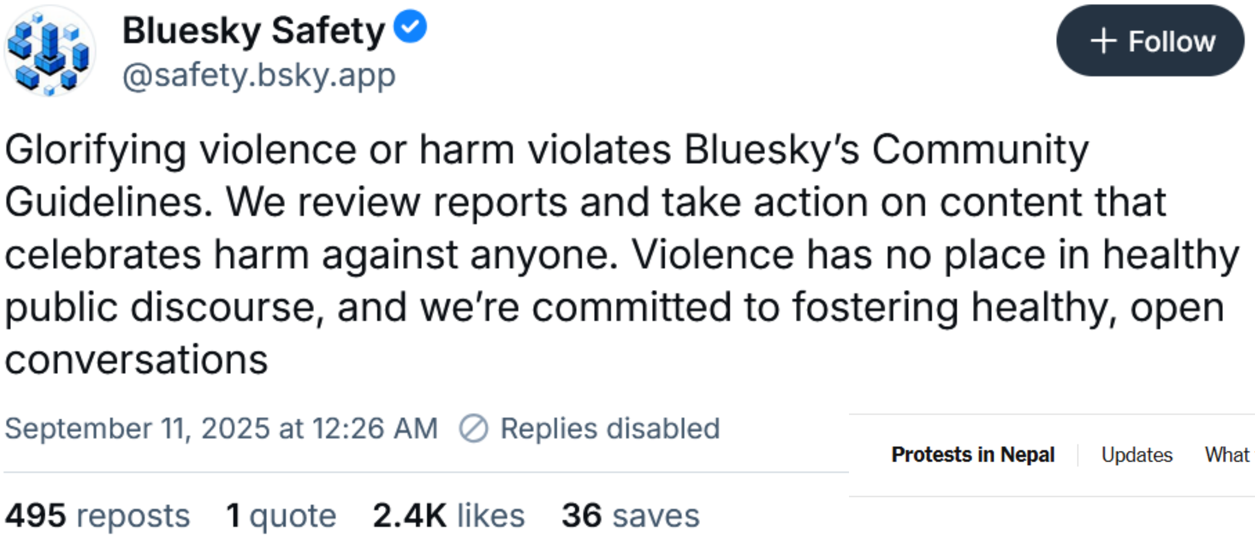
NYHETER • DEL 1 • SIDE 2-5 • KULTUR • DEL 2 • SIDE 2-3

Photo taken from Custodians of The Internet by Tarleton Gillespie

Statement from Facebook Vice President

“ These decisions aren’t easy. In many cases, there’s no clear line between an image of nudity or violence that carries global and historic significance and one that doesn’t. Some images may be offensive in one part of the world and acceptable in another, and even with a clear standard, it’s hard to screen millions of posts on a case-by-case basis every week… In this case, we tried to ***strike a difficult balance between enabling expression and protecting our community and ended up making a mistake****… ”*

Even today...



Bluesky Safety
@safety.bsky.app

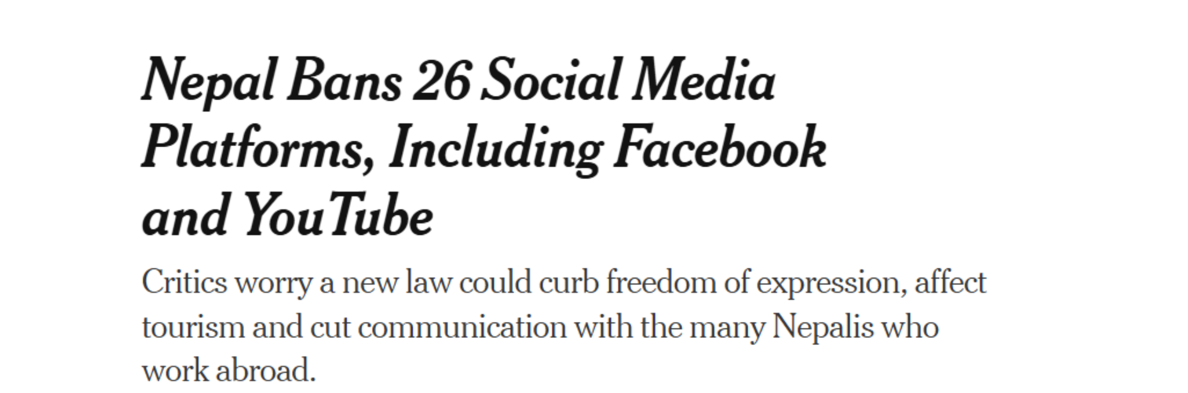
Glorifying violence or harm violates Bluesky's Community Guidelines. We review reports and take action on content that celebrates harm against anyone. Violence has no place in healthy public discourse, and we're committed to fostering healthy, open conversations

September 11, 2025 at 12:26 AM ⊘ Replies disabled

495 reposts **1** quote **2.4K** likes **36** saves

The New York Times

Protests in Nepal | Updates What to Know Class Tensions Censorship Playbook Fails Longstanding Problems



Nepal Bans 26 Social Media Platforms, Including Facebook and YouTube

Critics worry a new law could curb freedom of expression, affect tourism and cut communication with the many Nepalis who work abroad.

Definition from the Digital Services Act (DSA)

Article 3(t) defines content moderation as:

*“content moderation means the activities undertaken by platforms aimed at **detecting, identifying and addressing** illegal content or information incompatible with their terms and conditions … ”*

Source: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065

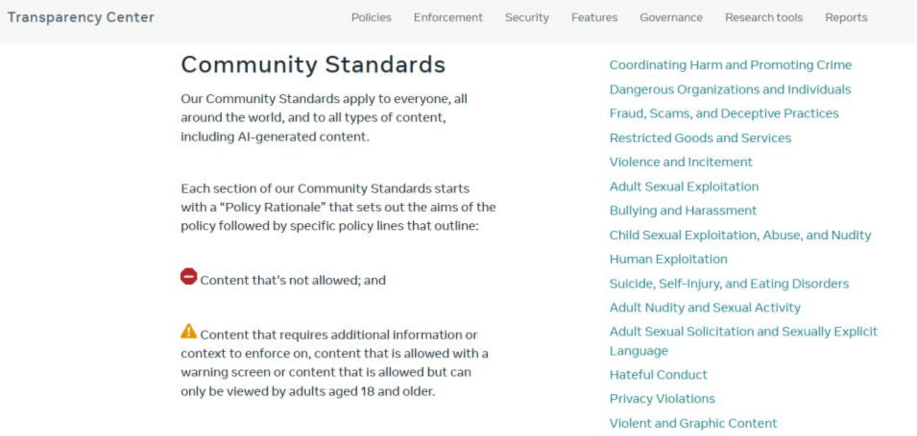
Takeaways

Content moderation is often striking a balance between enabling expression of the community and protecting the community.

Thus, it is notoriously difficult.

The centralized black box

Terms and conditions : Community standards



Transparency Center — Policies · Enforcement · Security · Features · Governance · Research tools · Reports

Community Standards

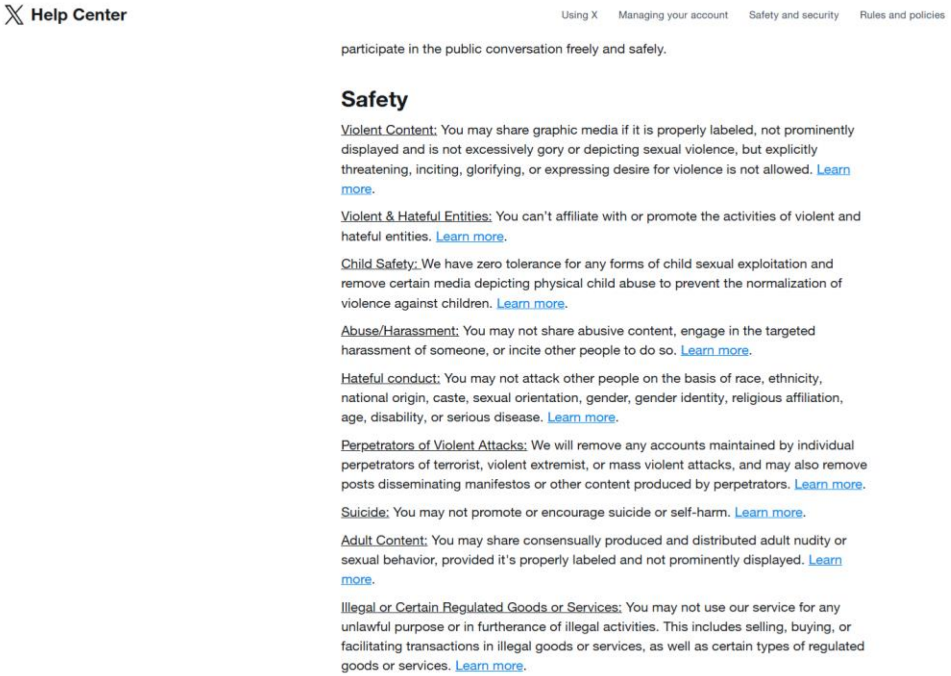
Our Community Standards apply to everyone, all around the world, and to all types of content, including AI-generated content.

Each section of our Community Standards starts with a "Policy Rationale" that sets out the aims of the policy followed by specific policy lines that outline:

⛔ Content that's not allowed; and

⚠ Content that requires additional information or context to enforce on, content that is allowed with a warning screen or content that is allowed but can only be viewed by adults aged 18 and older.

- Coordinating Harm and Promoting Crime
- Dangerous Organizations and Individuals
- Fraud, Scams, and Deceptive Practices
- Restricted Goods and Services
- Violence and Incitement
- Adult Sexual Exploitation
- Bullying and Harassment
- Child Sexual Exploitation, Abuse, and Nudity
- Human Exploitation
- Suicide, Self-Injury, and Eating Disorders
- Adult Nudity and Sexual Activity
- Adult Sexual Solicitation and Sexually Explicit Language
- Hateful Conduct
- Privacy Violations
- Violent and Graphic Content



X Help Center — Using X · Managing your account · Safety and security · Rules and policies

participate in the public conversation freely and safely.

Safety

Violent Content: You may share graphic media if it is properly labeled, not prominently displayed and is not excessively gory or depicting sexual violence, but explicitly threatening, inciting, glorifying, or expressing desire for violence is not allowed. Learn more.

Violent & Hateful Entities: You can't affiliate with or promote the activities of violent and hateful entities. Learn more.

Child Safety: We have zero tolerance for any forms of child sexual exploitation and remove certain media depicting physical child abuse to prevent the normalization of violence against children. Learn more.

Abuse/Harassment: You may not share abusive content, engage in the targeted harassment of someone, or incite other people to do so. Learn more.

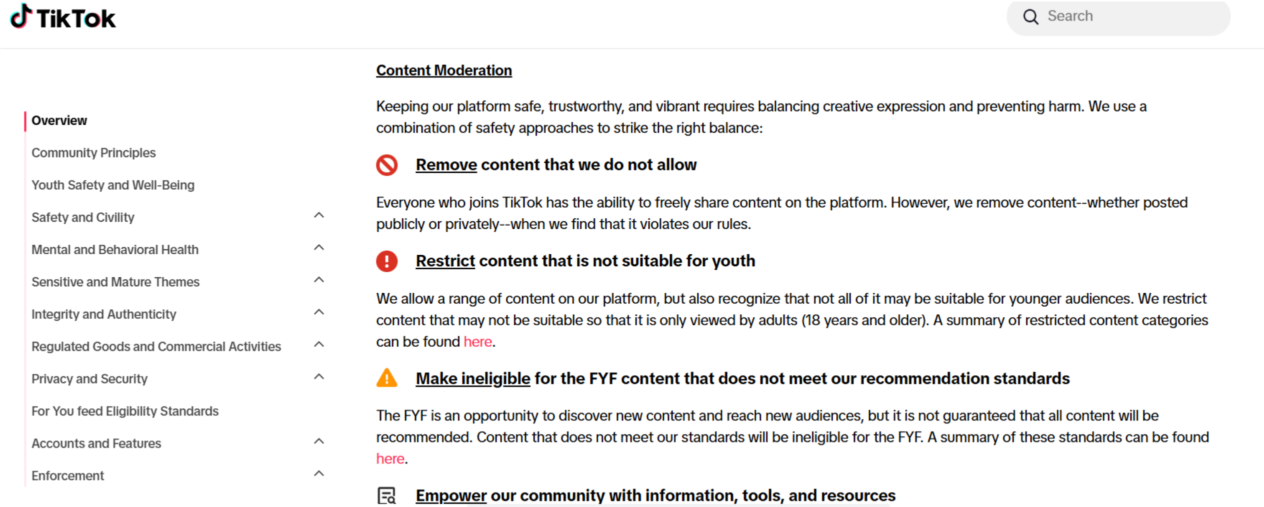
Hateful conduct: You may not attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. Learn more.

Perpetrators of Violent Attacks: We will remove any accounts maintained by individual perpetrators of terrorist, violent extremist, or mass violent attacks, and may also remove posts disseminating manifestos or other content produced by perpetrators. Learn more.

Suicide: You may not promote or encourage suicide or self-harm. Learn more.

Adult Content: You may share consensually produced and distributed adult nudity or sexual behavior, provided it's properly labeled and not prominently displayed. Learn more.

Illegal or Certain Regulated Goods or Services: You may not use our service for any unlawful purpose or in furtherance of illegal activities. This includes selling, buying, or facilitating transactions in illegal goods or services, as well as certain types of regulated goods or services. Learn more.



TikTok — Search

- Overview
- Community Principles
- Youth Safety and Well-Being
- Safety and Civility
- Mental and Behavioral Health
- Sensitive and Mature Themes
- Integrity and Authenticity
- Regulated Goods and Commercial Activities
- Privacy and Security
- For You feed Eligibility Standards
- Accounts and Features
- Enforcement

Content Moderation

Keeping our platform safe, trustworthy, and vibrant requires balancing creative expression and preventing harm. We use a combination of safety approaches to strike the right balance:

⛔ **Remove content that we do not allow**

Everyone who joins TikTok has the ability to freely share content on the platform. However, we remove content--whether posted publicly or privately--when we find that it violates our rules.

❗ **Restrict content that is not suitable for youth**

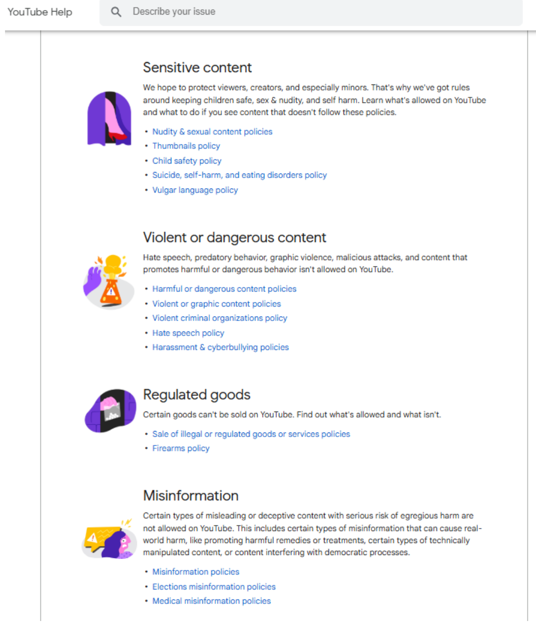
We allow a range of content on our platform, but also recognize that not all of it may be suitable for younger audiences. We restrict content that may not be suitable so that it is only viewed by adults (18 years and older). A summary of restricted content categories can be found here.

⚠ **Make ineligible for the FYF content that does not meet our recommendation standards**

The FYF is an opportunity to discover new content and reach new audiences, but it is not guaranteed that all content will be recommended. Content that does not meet our standards will be ineligible for the FYF. A summary of these standards can be found here.

Empower our community with information, tools, and resources

rience on TikTok. We may add labels, "opt-in"
r out content with specific hashtags or
ınd in-app features with safety resources.



YouTube Help — Describe your issue

Sensitive content

We hope to protect viewers, creators, and especially minors. That's why we've got rules around keeping children safe, sex & nudity, and self harm. Learn what's allowed on YouTube and what to do if you see content that doesn't follow these policies.

- Nudity & sexual content policies
- Thumbnails policy
- Child safety policy
- Suicide, self-harm, and eating disorders policy
- Vulgar language policy

Violent or dangerous content

Hate speech, predatory behavior, graphic violence, malicious attacks, and content that promotes harmful or dangerous behavior isn't allowed on YouTube.

- Harmful or dangerous content policies
- Violent or graphic content policies
- Violent criminal organizations policy
- Hate speech policy
- Harassment & cyberbullying policies

Regulated goods

Certain goods can't be sold on YouTube. Find out what's allowed and what isn't.

- Sale of illegal or regulated goods or services policies
- Firearms policy

Misinformation

Certain types of misleading or deceptive content with serious risk of egregious harm are not allowed on YouTube. This includes certain types of misinformation that can cause real-world harm, like promoting harmful remedies or treatments, certain types of technically manipulated content, or content interfering with democratic processes.

- Misinformation policies
- Elections misinformation policies
- Medical misinformation policies

Partial list of violations

- ❑ Abusive accounts
- ❑ Privacy violation
- ❑ Spam
- ❑ Financial scams
- ❑ Regulated goods
- ❑ Adult sexual exploitation
- ❑ Adult sexual solicitation
- ❑ Child nudity and sexual exploitation
- ❑ Human trafficking
- ❑ Self harm
- ❑ Violence and graphical content
- ❑ Societal harm
- ❑ Personal harm
- ❑ Bullying
- ❑ Credible threats of violence
- ❑ Hate speech
- ❑ Misinformation

… … …

A typical content moderation pipeline

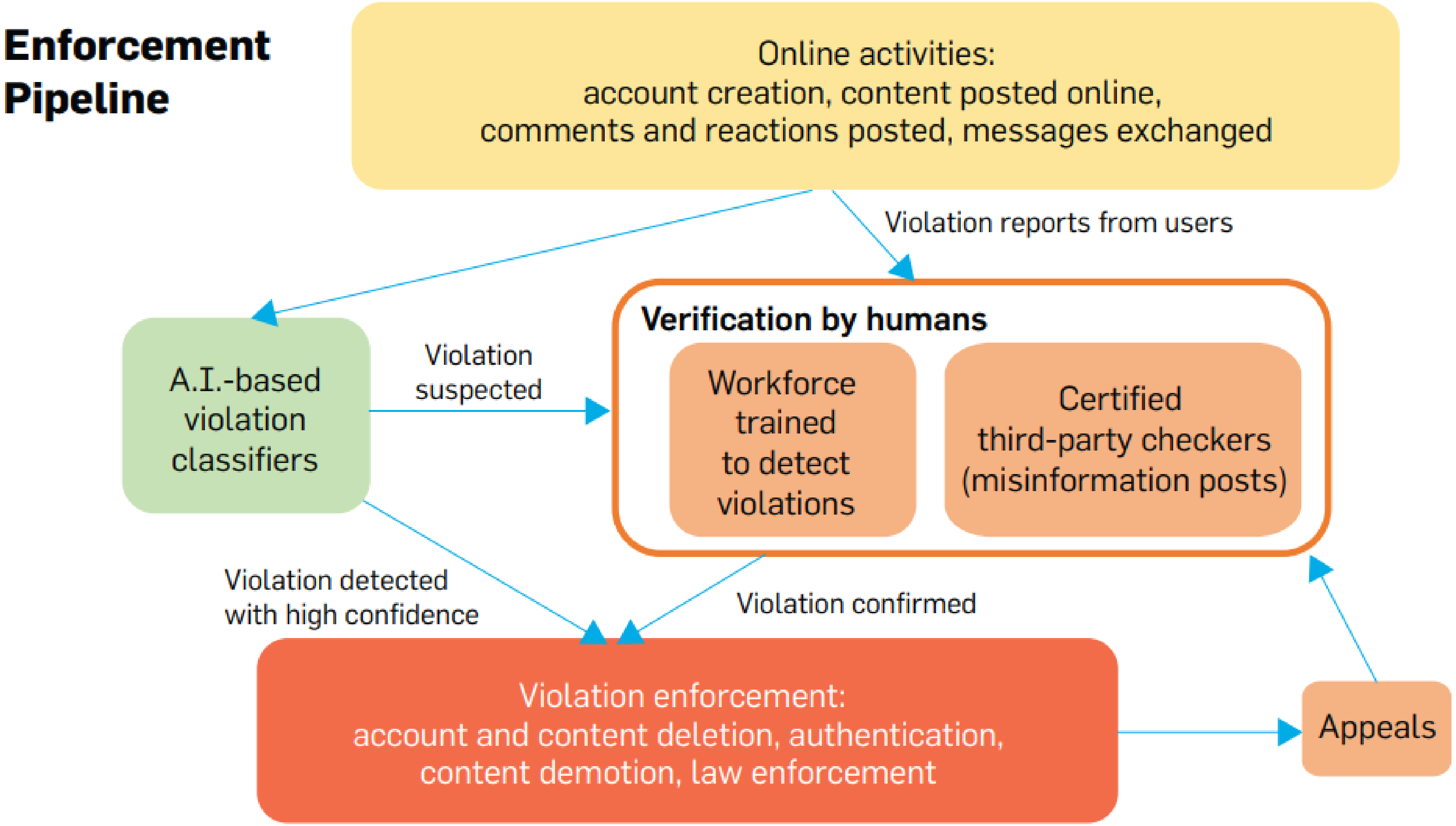
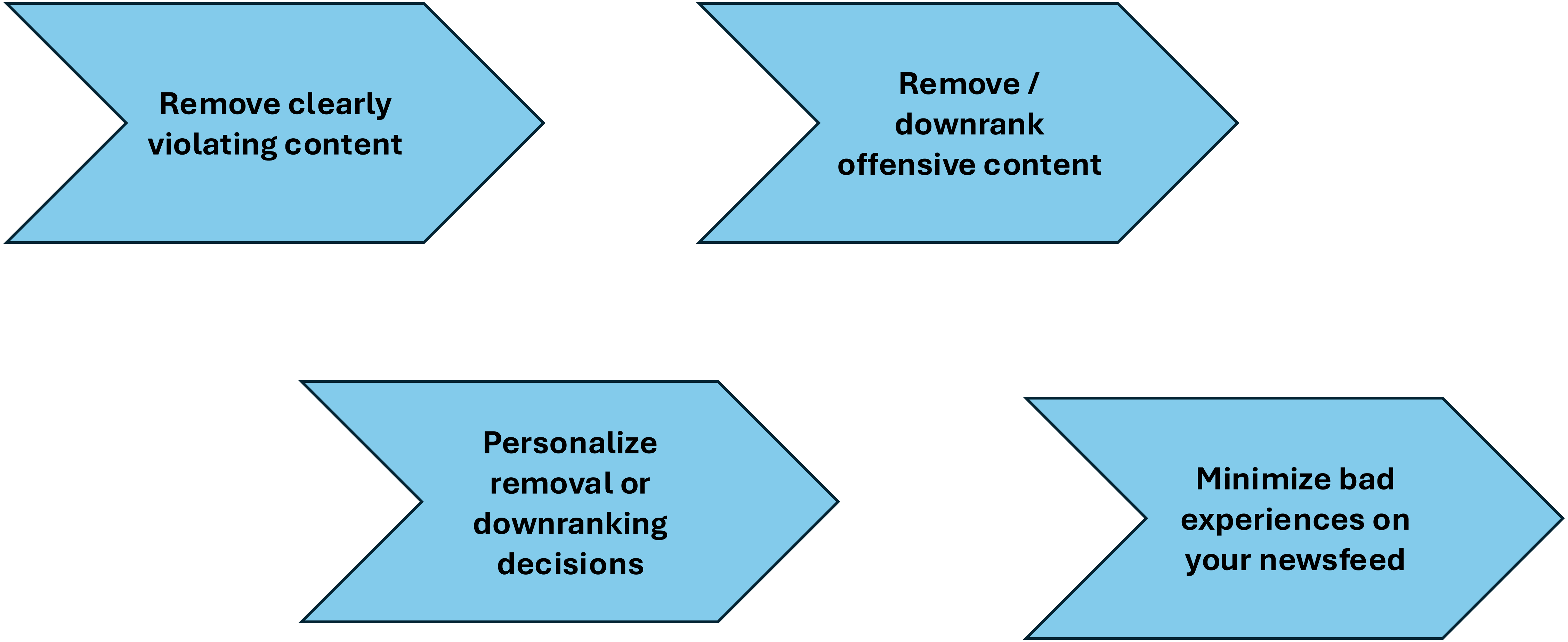

Enforcement Pipeline
Online activities: account creation, content posted online, comments and reactions posted, messages exchanged
Violation reports from users
Verification by humans
A.I.-based violation classifiers
Violation suspected
Workforce trained to detect violations
Certified third-party checkers (misinformation posts)
Violation detected with high confidence
Violation confirmed
Violation enforcement: account and content deletion, authentication, content demotion, law enforcement
Appeals

Figure courtesy : Halevy, Alon, et al. "Preserving integrity in online social networks." *Communications of the ACM* 65.2 (2022)

A broader view of enforcement



Remove clearly violating content

Remove / downrank offensive content

Personalize removal or downranking decisions

Minimize bad experiences on your newsfeed

Although thoughtful, some systemic flaws exist

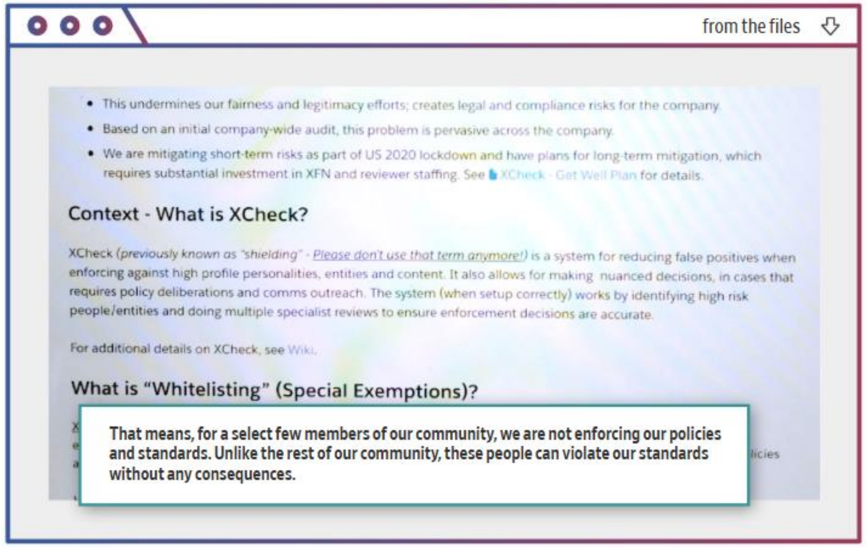
Opaque by design

Centralized power

Zero user agency

Arbitrary / inconsistent enforcement

XChecks : Above the community standards



- This undermines our fairness and legitimacy efforts; creates legal and compliance risks for the company.
- Based on an initial company-wide audit, this problem is pervasive across the company.
- We are mitigating short-term risks as part of US 2020 lockdown and have plans for long-term mitigation, which requires substantial investment in XFN and reviewer staffing. See XCheck - Get Well Plan for details.

Context - What is XCheck?

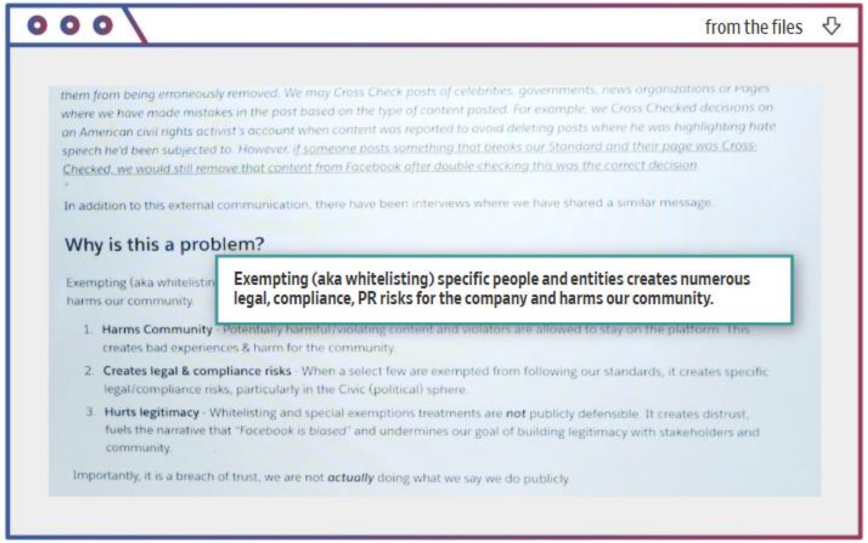
XCheck *(previously known as "shielding" - Please don't use that term anymore!)* is a system for reducing false positives when enforcing against high profile personalities, entities and content. It also allows for making nuanced decisions, in cases that requires policy deliberations and comms outreach. The system (when setup correctly) works by identifying high risk people/entities and doing multiple specialist reviews to ensure enforcement decisions are accurate.

For additional details on XCheck, see Wiki.

What is "Whitelisting" (Special Exemptions)?

licies

That means, for a select few members of our community, we are not enforcing our policies and standards. Unlike the rest of our community, these people can violate our standards without any consequences.



them from being erroneously removed. We may Cross Check posts of celebrities, governments, news organizations or Pages where we have made mistakes in the past based on the type of content posted. For example, we Cross Checked decisions on an American civil rights activist's account when content was reported to avoid deleting posts where he was highlighting hate speech he'd been subjected to. However, if someone posts something that breaks our Standard and their page was Cross-Checked, we would still remove that content from Facebook after double-checking this was the correct decision.

In addition to this external communication, there have been interviews where we have shared a similar message.

Why is this a problem?

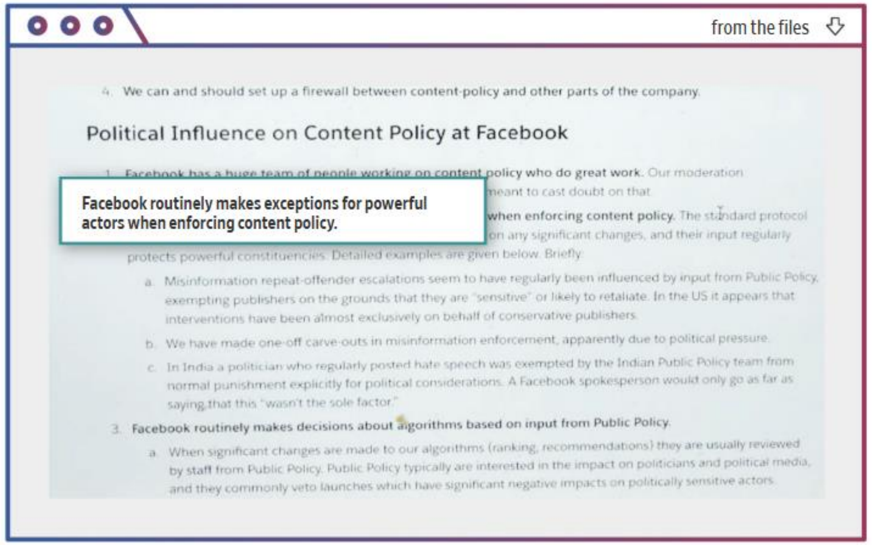
Exempting (aka whitelistin

harms our community.

Exempting (aka whitelisting) specific people and entities creates numerous legal, compliance, PR risks for the company and harms our community.

1. **Harms Community** - Potentially harmful/violating content and violators are allowed to stay on the platform. This creates bad experiences & harm for the community.
2. **Creates legal & compliance risks** - When a select few are exempted from following our standards, it creates specific legal/compliance risks, particularly in the Civic (political) sphere.
3. **Hurts legitimacy** - Whitelisting and special exemptions treatments are **not** publicly defensible. It creates distrust, fuels the narrative that *"Facebook is biased"* and undermines our goal of building legitimacy with stakeholders and community.

Importantly, it is a breach of trust, we are not ***actually*** doing what we say we do publicly.



4. We can and should set up a firewall between content-policy and other parts of the company.

Political Influence on Content Policy at Facebook

1. **Facebook has a huge team of people working on content policy who do great work.** Our moderation

meant to cast doubt on that.

when enforcing content policy. The standard protocol

on any significant changes, and their input regularly protects powerful constituencies. Detailed examples are given below. Briefly:

Facebook routinely makes exceptions for powerful actors when enforcing content policy.

a. Misinformation repeat-offender escalations seem to have regularly been influenced by input from Public Policy, exempting publishers on the grounds that they are "sensitive" or likely to retaliate. In the US it appears that interventions have been almost exclusively on behalf of conservative publishers.
b. We have made one-off carve-outs in misinformation enforcement, apparently due to political pressure.
c. In India a politician who regularly posted hate speech was exempted by the Indian Public Policy team from normal punishment explicitly for political considerations. A Facebook spokesperson would only go as far as saying that this "wasn't the sole factor."

3. **Facebook routinely makes decisions about algorithms based on input from Public Policy.**
   a. When significant changes are made to our algorithms (ranking, recommendations) they are usually reviewed by staff from Public Policy. Public Policy typically are interested in the impact on politicians and political media, and they commonly veto launches which have significant negative impacts on politically sensitive actors.

Source: https://www.wsj.com/articles/the-facebook-files-11631713039#_=_

Takeaways

Centralized moderation is an incredibly complex, expensive, human-in-the-loop socio-technical system, built with thoughtful policies.

However, its opaque and centralized nature leads to arbitrary enforcement, denying users agency and eroding trust.

A decentralized alternative: Bluesky

Recent shift to decentralized alternatives

mastodon

Bluesky

ONDC®
Open Network for Digital Commerce

UPI
UNIFIED PAYMENTS INTERFACE

- ✓Built on open protocols
- ✓Power is decentralized
- ✓More user agency and autonomy
- ✓Foster innovation through open standards

Recent shift to decentralized alternatives

mastodon

Bluesky

ONDC
Open Network for Digital Commerce

UPI
UNIFIED PAYMENTS INTERFACE

✓Built on open protocols

✓Power is decentralized

✓More user agency and autonomy

✓Foster innovation through open standards

Bluesky

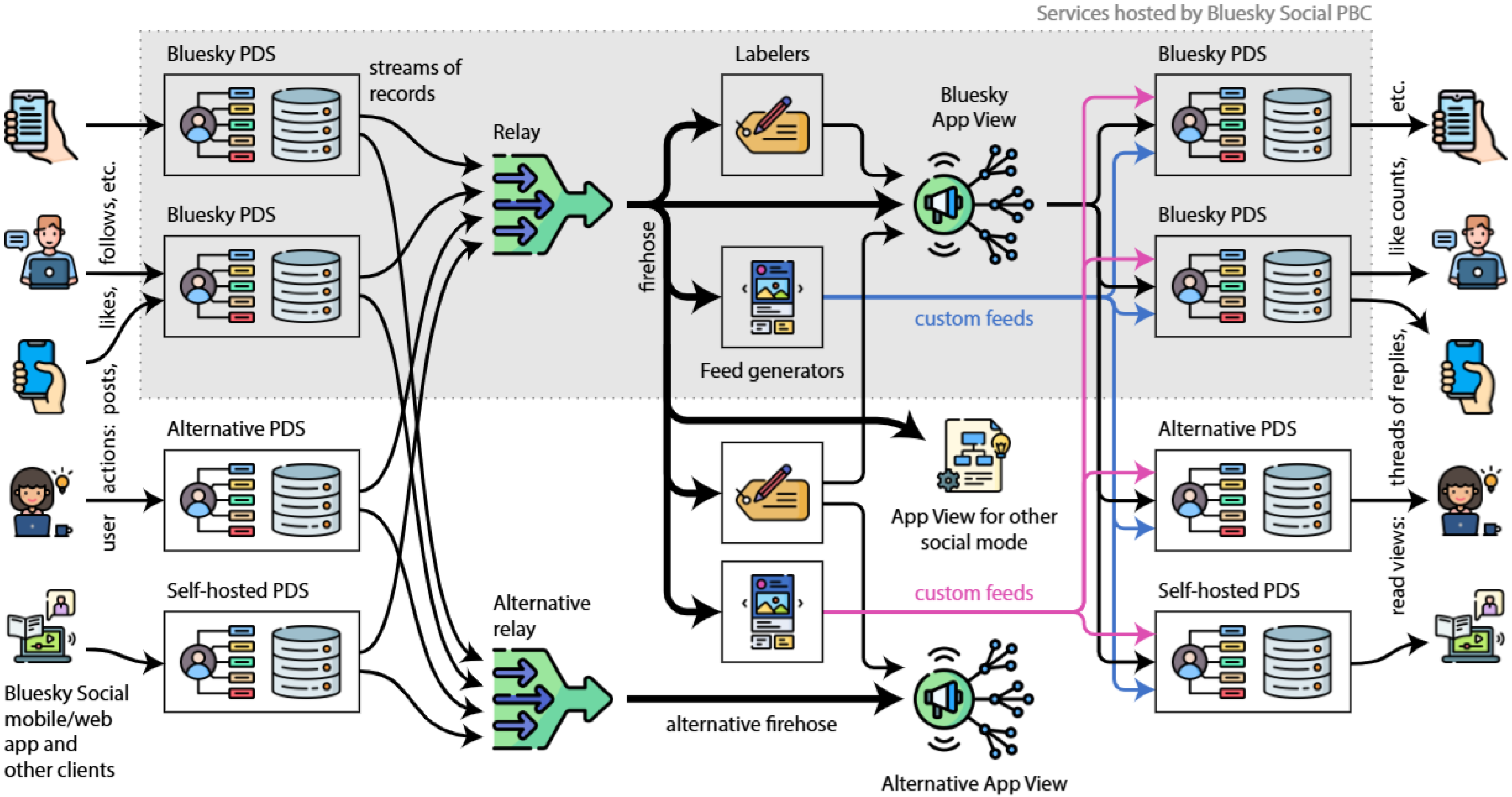
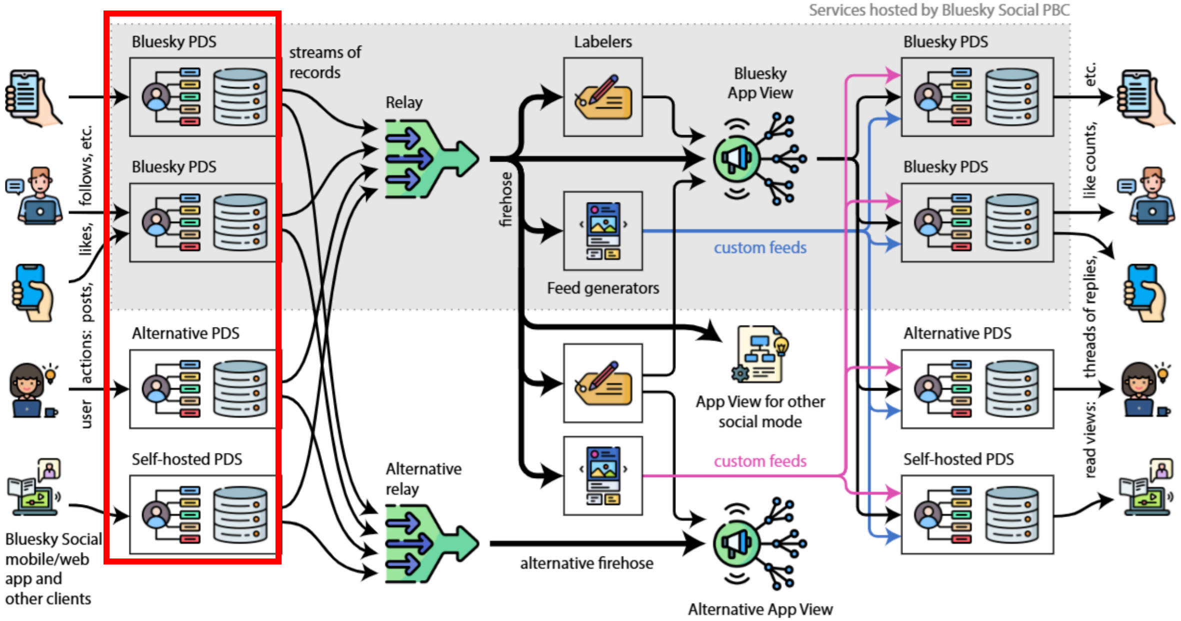


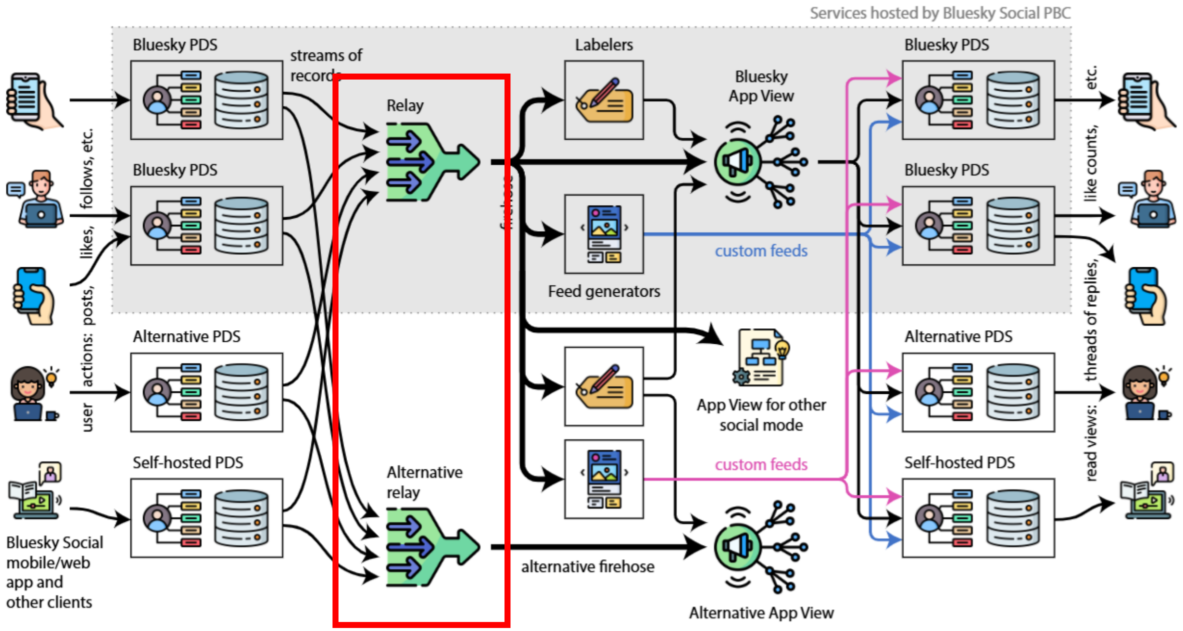
Figure courtesy : Kleppmann, Martin, et al. "Bluesky and the at protocol: Usable decentralized social media." *Proceedings of the ACM Conext-2024 Workshop on the Decentralization of the Internet*. 2024.

User data storage : Personal Data Servers



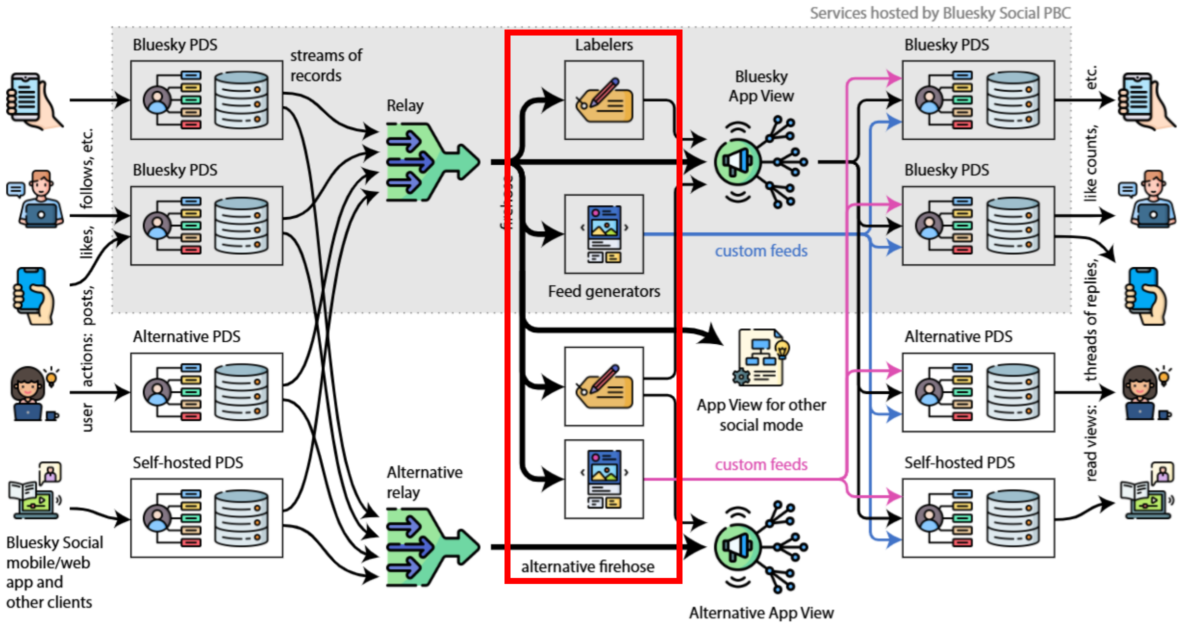
- ❑ A Personal Data Server (PDS) stores the user's repositories (i.e., all the actions they ever performed).
- ❑ Currently, there are 2,555 valid PDSs on Bluesky and 70 of them are run by Bluesky Social PBC.

Indexing I : Relay



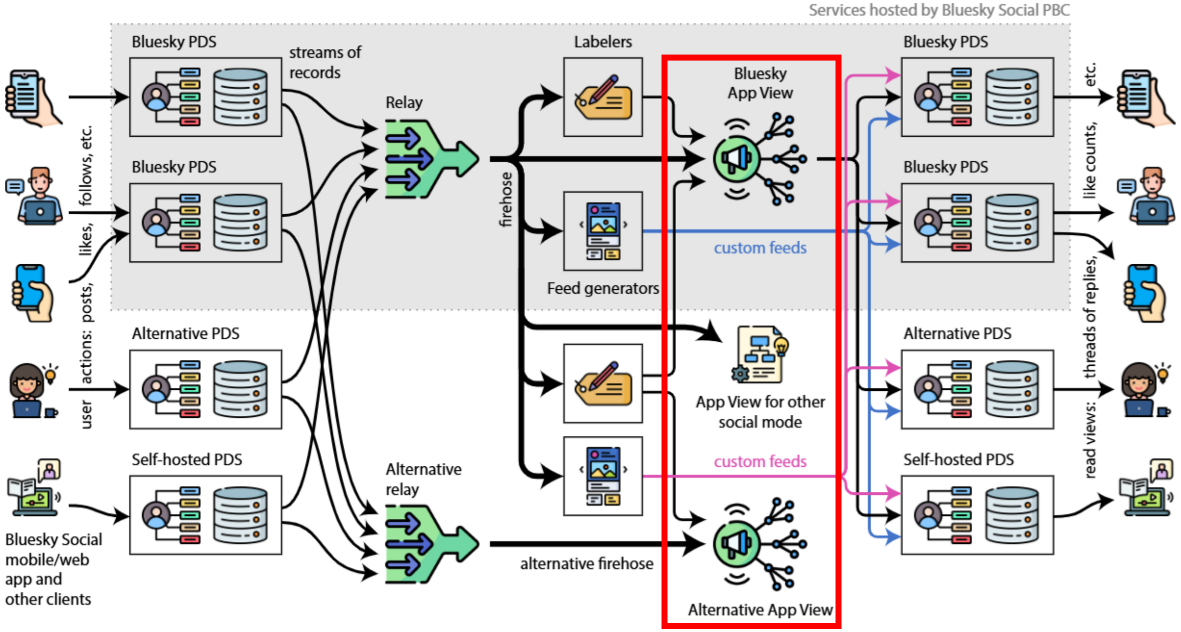
- ❑ Relay crawls the user repositories on all known PDSes and consumes the streams of updates that they produce
- ❑ Relay creates the firehose: an aggregated stream of updates.

Content dissemination : Labeler and Feed generators



- ❑ Labelers consume the firehose and moderate the content.
- ❑ Feed generators consume the firehose and reorder the content.

Indexing II & User interface : App view



- ❑ The App View is a service that consumes the firehose, and processes the records that are relevant to the Bluesky social app
- ❑ The App View is also responsible for enforcing moderation controls.

Takeaways

Bluesky is a decentralized alternative where the power is diffused across different entities.

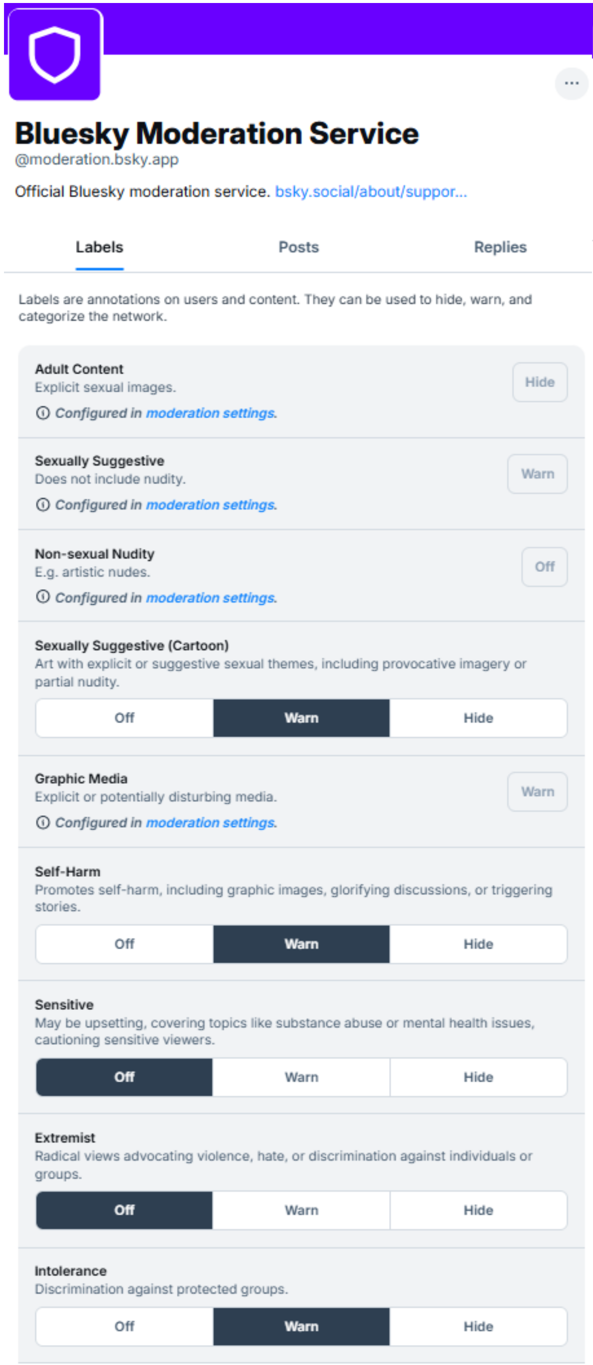
Moderation is no longer a top-down function of the platform, but a service in an open marketplace.

Break (5 mins)

- Content moderation
- The centralized black-box
- A decentralized alternative: Bluesky
- Content moderation on Bluesky
- Our role as CSS researchers
- Hands on session (in the afternoon)

Composable content moderation on Bluesky

Labelers on Bluesky



Bluesky Moderation Service

@moderation.bsky.app

Official Bluesky moderation service. bsky.social/about/suppor...

Labels are annotations on users and content. They can be used to hide, warn, and categorize the network.

Adult Content
Explicit sexual images.
Configured in *moderation settings.*
Hide

Sexually Suggestive
Does not include nudity.
Configured in *moderation settings.*
Warn

Non-sexual Nudity
E.g. artistic nudes.
Configured in *moderation settings.*
Off

Sexually Suggestive (Cartoon)
Art with explicit or suggestive sexual themes, including provocative imagery or partial nudity.
Off [✓ Warn] Hide

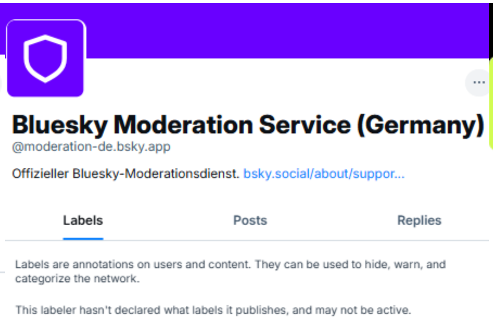
Graphic Media
Explicit or potentially disturbing media.
Configured in *moderation settings.*
Warn

Self-Harm
Promotes self-harm, including graphic images, glorifying discussions, or triggering stories.
Off [✓ Warn] Hide

Sensitive
May be upsetting, covering topics like substance abuse or mental health issues, cautioning sensitive viewers.
[✓ Off] Warn Hide

Extremist
Radical views advocating violence, hate, or discrimination against individuals or groups.
[✓ Off] Warn Hide

Intolerance
Discrimination against protected groups.
Off [✓ Warn] Hide



Bluesky Moderation Service (Germany)

@moderation-de.bsky.app

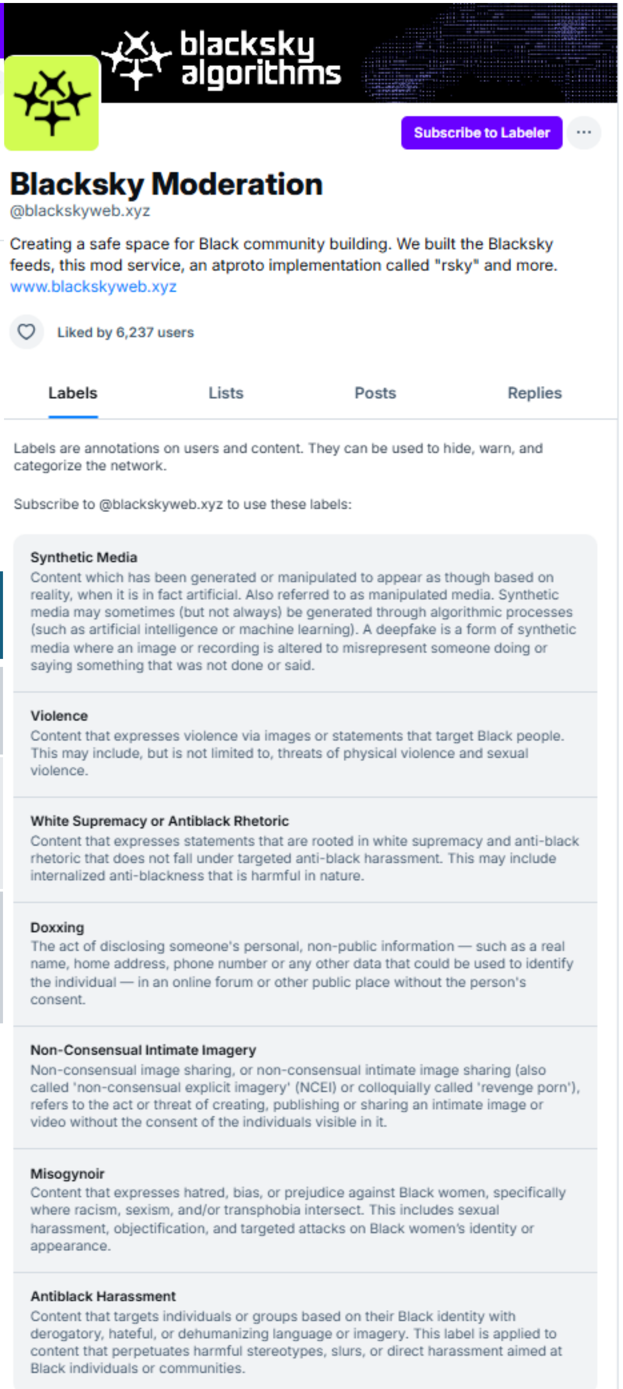
Offizieller Bluesky-Moderationsdienst. bsky.social/about/suppor...

Labels are annotations on users and content. They can be used to hide, warn, and categorize the network.

This labeler hasn't declared what labels it publishes, and may not be active.

Types of labelers

- Default labeler
- Country specific labelers
- Community labelers



Blacksky Moderation

@blackskyweb.xyz

Creating a safe space for Black community building. We built the Blacksky feeds, this mod service, an atproto implementation called "rsky" and more. www.blackskyweb.xyz

Liked by 6,237 users

Labels are annotations on users and content. They can be used to hide, warn, and categorize the network.

Subscribe to @blackskyweb.xyz to use these labels:

Synthetic Media
Content which has been generated or manipulated to appear as though based on reality, when it is in fact artificial. Also referred to as manipulated media. Synthetic media may sometimes (but not always) be generated through algorithmic processes (such as artificial intelligence or machine learning). A deepfake is a form of synthetic media where an image or recording is altered to misrepresent someone doing or saying something that was not done or said.

Violence
Content that expresses violence via images or statements that target Black people. This may include, but is not limited to, threats of physical violence and sexual violence.

White Supremacy or Antiblack Rhetoric
Content that expresses statements that are rooted in white supremacy and anti-black rhetoric that does not fall under targeted anti-black harassment. This may include internalized anti-blackness that is harmful in nature.

Doxxing
The act of disclosing someone's personal, non-public information — such as a real name, home address, phone number or any other data that could be used to identify the individual — in an online forum or other public place without the person's consent.

Non-Consensual Intimate Imagery
Non-consensual image sharing, or non-consensual intimate image sharing (also called 'non-consensual explicit imagery' (NCEI) or colloquially called 'revenge porn'), refers to the act or threat of creating, publishing or sharing an intimate image or video without the consent of the individuals visible in it.

Misogynoir
Content that expresses hatred, bias, or prejudice against Black women, specifically where racism, sexism, and/or transphobia intersect. This includes sexual harassment, objectification, and targeted attacks on Black women's identity or appearance.

Antiblack Harassment
Content that targets individuals or groups based on their Black identity with derogatory, hateful, or dehumanizing language or imagery. This label is applied to content that perpetuates harmful stereotypes, slurs, or direct harassment aimed at Black individuals or communities.

Users have more autonomy and agency

- ❑Bluesky and country specific labeler apply by default.
- ❑On the other hand, this architecture empowers users to
  - ❑choose what should be added to their feed.
  - ❑choose how labeled content should appear.
  - ❑choose which other lebeler should affect their feed.

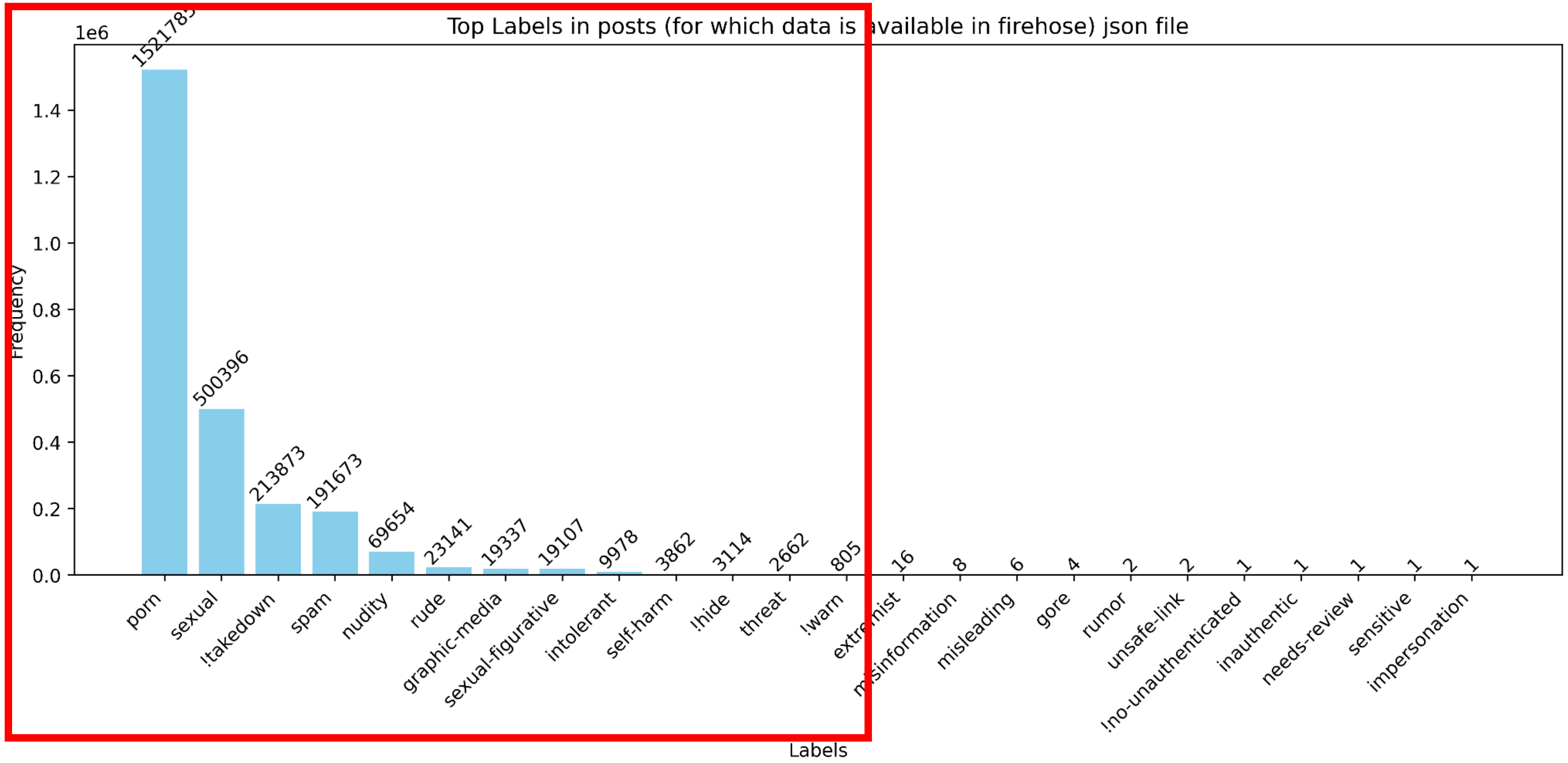
Some questions that may arise

❑What kind of labels do these labelers apply?

❑What are the interpretations of these labels?

❑How do they apply these labels? Are the algorithms transparent?

Labeler outcomes are publicly accessible



❑Total distinct posts labelled : 2,528,786 (19 March '25 - 1 June ‘25)

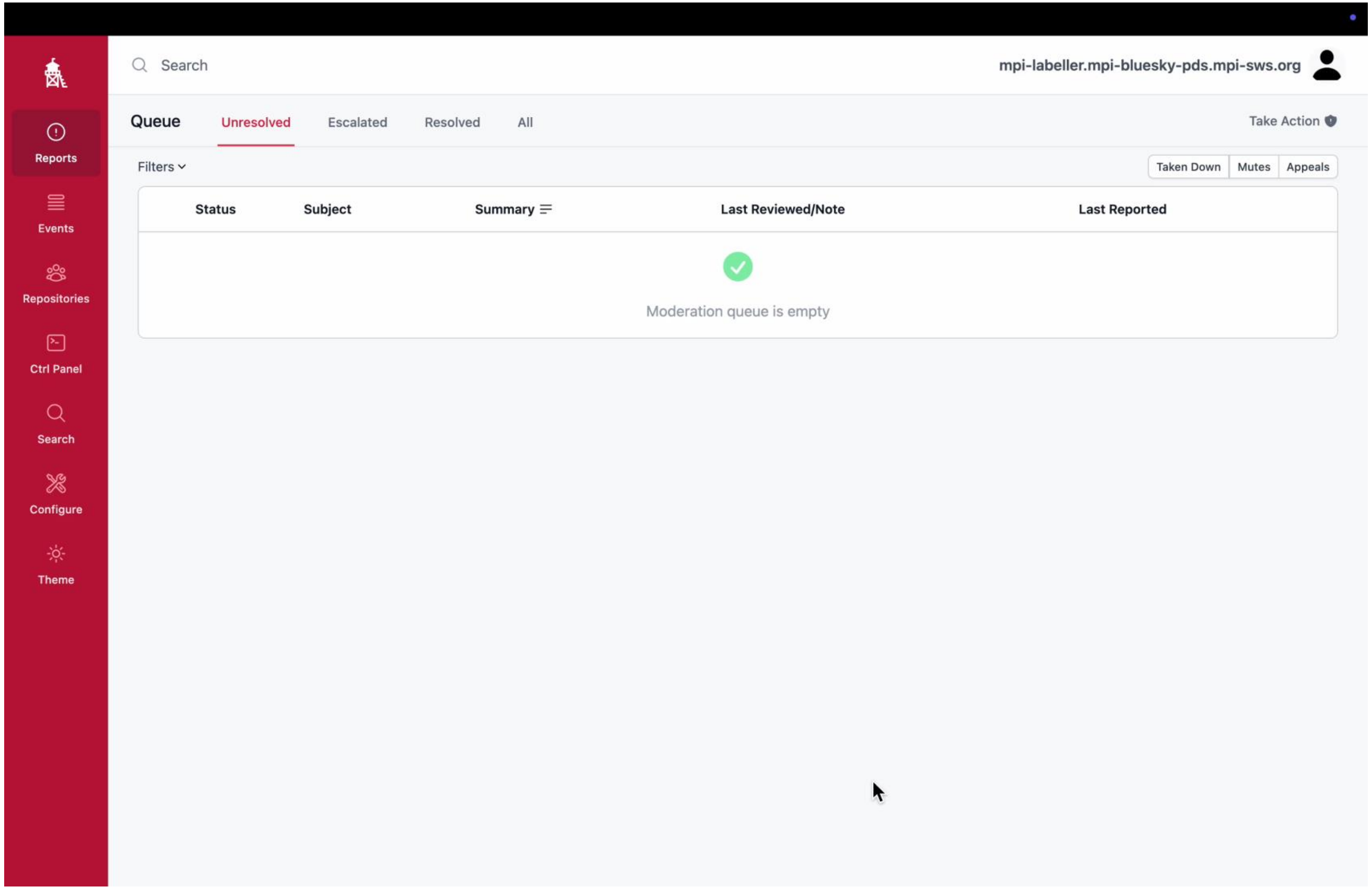
❑The data reveals a strong focus on platform hygiene (spam, porn) and user safety issues (intolerance, rude).

What do these labels even mean?

Labels	Description
Intolerance	Discrimination against protected groups.
Threats	Promotes violence or harm towards others, including threats, incitement, or advocacy of harm.
Rude	Rude or impolite, including crude language and disrespectful comments, without constructive purpose.

Source: https://bsky.app/profile/moderation.bsky.app

How does a labeler apply these labels?



Status	Subject	Summary	Last Reviewed/Note	Last Reported

Moderation queue is empty

How does Bluesky apply these labels?

The key takeaway: We can see *what* was labeled, but not *why*. The implementation policies remain opaque.

Source: https://bsky.social/about/blog/01-17-2025-moderation-2024

Intolerance: Discrimination against protected groups

Q: *How does Bluesky labeler operationalize ‘protected group’?*

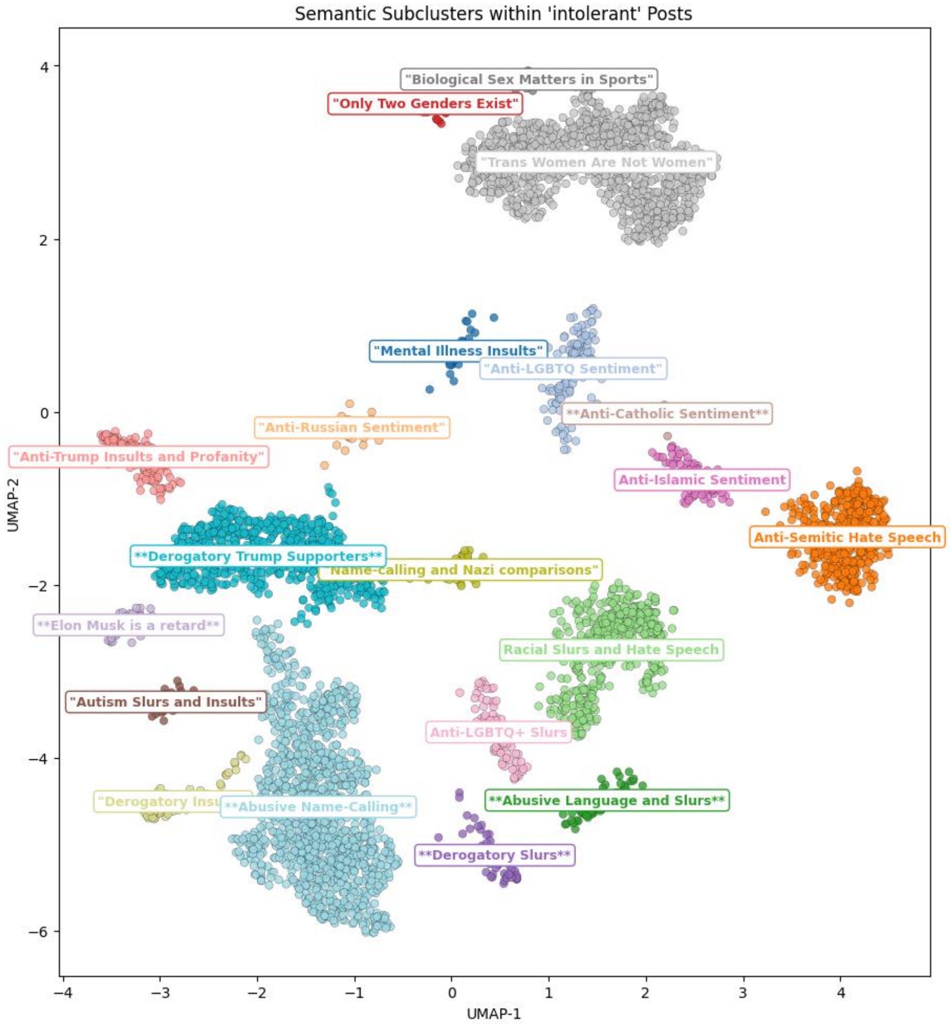
Information about someone's protected characteristics—*such as sexual orientation, gender identity, disability, caste, or immigration status*—for the purpose of targeting them or discriminating against them.

Source: https://bsky.social/about/support/community-guidelines

Access to label stream gives some hope!

- ❑Here is a cluster of all the posts labelled as Intolerance.

- ❑Such analyses on the labelled text could give us a more nuanced view of what is happening behind the scenes.



Takeaways

Bluesky’s composable moderation paradigm provides more autonomy and agency to users.

While procedural transparency is still lacking on Bluesky labeler, it is auditable thanks to the accessible label stream.

Our role as Computational Social Scientists

Systemic flaws still exist

Opaque by design

↓

Opacity has decreased

Centralized power

↓

Power is decentralized

Zero user agency

↓

Better user agency

Inconsistent enforcement

↓

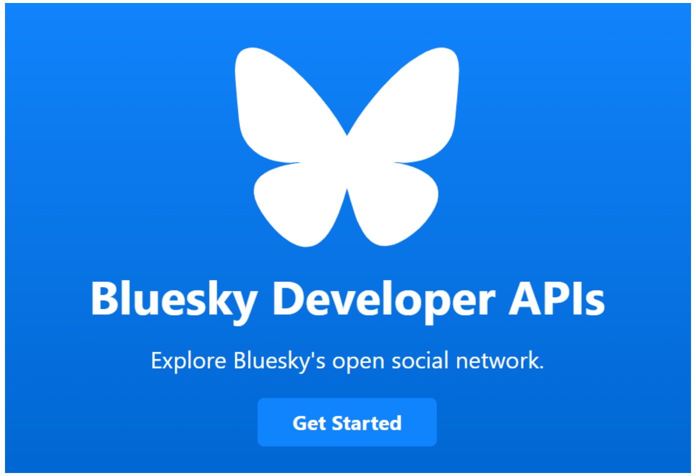
Now auditable

Opportunity 1: Data Access

❑Reliability of content moderation is now measurable.

❑We will see some examples of this in the hands-on session.

- ❑What kind of labels are applied?
- ❑Are they applied timely?
- ❑Are they applied consistently?



Bluesky Developer APIs

Explore Bluesky's open social network.

Opportunity 2: Open marketplace

❑For ages, researchers have been able to audit platform practices.

❑Now, we have the opportunity to rectify the drawbacks.

Opportunity 3 : Stay ahead of the curve

Proactively think about regulations for better accountability, data protection while not hindering innovation.

Takeaways

Centralized moderation is opaque and unaccountable.

Decentralized moderation is more transparent and composable, giving users more autonomy. Yet it lacks procedural transparency.

This paradigm shift empowers us not only to measure the moderation service, but also to set the standards.

Thank You!

✉ adash@mpi-sws.org

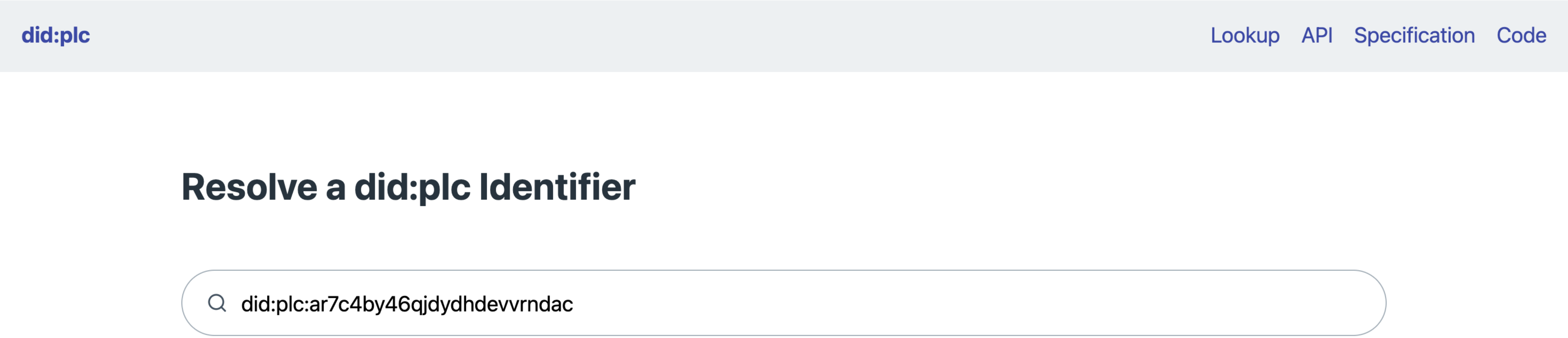
Plan for today

- Content moderation
- The centralized black-box
- A decentralized alternative: Bluesky
- Content moderation on Bluesky
- Our role as CSS researchers
- Hands on session (in the afternoon)

Hands-on session

DID and PLC Directory

- Each user on Bluesky is identified via a unique DID (decentralised identifier) e.g., *did:plc:ljxfsne42aud2cd2iowkkqiz*.
- Each DID points to its associated **DID Document**, a document that stores service information about the user.

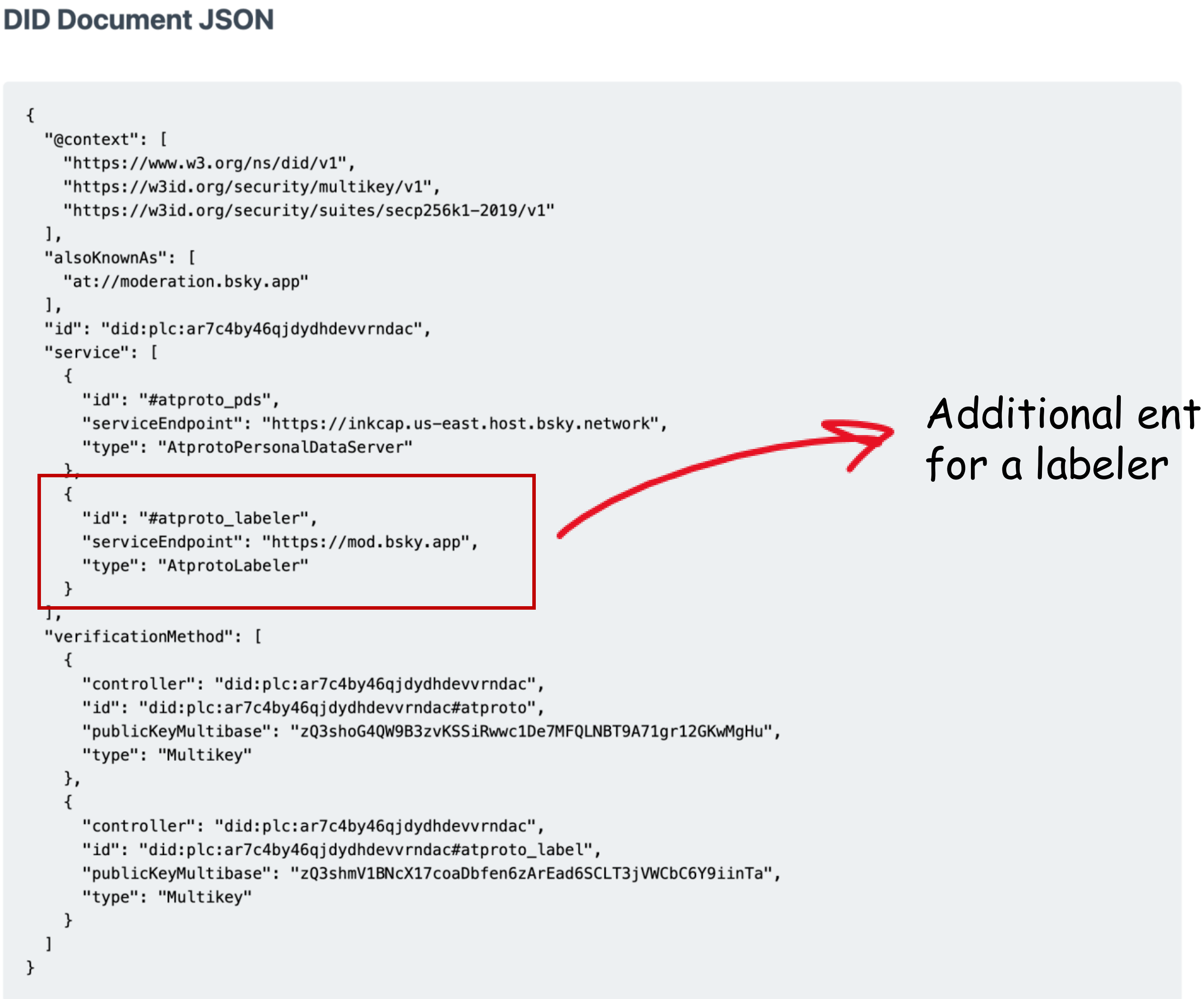


Resolve a did:plc Identifier

did:plc:ar7c4by46qjdydhdevvrndac

https://plc.directory/did:plc:ar7c4by46qjdydhdevvrndac

DID and PLC Directory



DID Document JSON

```
{
  "@context": [
    "https://www.w3.org/ns/did/v1",
    "https://w3id.org/security/multikey/v1",
    "https://w3id.org/security/suites/secp256k1-2019/v1"
  ],
  "alsoKnownAs": [
    "at://moderation.bsky.app"
  ],
  "id": "did:plc:ar7c4by46qjdydhdevvrndac",
  "service": [
    {
      "id": "#atproto_pds",
      "serviceEndpoint": "https://inkcap.us-east.host.bsky.network",
      "type": "AtprotoPersonalDataServer"
    },
    {
      "id": "#atproto_labeler",
      "serviceEndpoint": "https://mod.bsky.app",
      "type": "AtprotoLabeler"
    }
  ],
  "verificationMethod": [
    {
      "controller": "did:plc:ar7c4by46qjdydhdevvrndac",
      "id": "did:plc:ar7c4by46qjdydhdevvrndac#atproto",
      "publicKeyMultibase": "zQ3shoG4QW9B3zvKSSiRwwc1De7MFQLNBT9A71gr12GKwMgHu",
      "type": "Multikey"
    },
    {
      "controller": "did:plc:ar7c4by46qjdydhdevvrndac",
      "id": "did:plc:ar7c4by46qjdydhdevvrndac#atproto_label",
      "publicKeyMultibase": "zQ3shmV1BNcX17coaDbfen6zArEad6SCLT3jVWCbC6Y9iinTa",
      "type": "Multikey"
    }
  ]
}
```

Additional entry
for a labeler

Getting all labelers

- Using the **sync.listRepos** call offered by the Bluesky Relay, we can obtain a list of all active Bluesky users and their DIDs.

```
"cursor": "10",
"repos": [
{ "active": true,
"did": "did:plc:hlm3aibaluzuzuqqnzxg2urq",
"head": "bafyreigizoa62issah3ajgtmwgpi3wn4bsz3uchv5dihbchm2uubuelx2i",
"rev": "3lqmtafrzyv2r"
},
{ "active": true,
"did": "did:plc:6qgqcyg5gfd6sxmghdclxmpi",
"head": "bafyreidnlloy7pc634f36a2a2q3a7mwhzdslzfx2vfyk2ugel4sjpxpvsq",
"rev": "3lqn6vruciq2n"
},
{ "active": true,
"did": "did:plc:kbf77syjgjcjciabbsdm37qq",
…
```

Using https://relay1.us-west.bsky.network/xrpc/com.atproto.sync.listRepos endpoint

- Check plc directory to get DID document for each of these DID and see if they are a labeler.

Getting labeler information

Suppose we know a particular user (identified by a DID) is a labeler, from DID document we have its service endpoint (we will explore where it is used, later in this session)

But, what about :

- their descriptions?
- which labels are issued? their definitions?
- default settings?

API : app.bsky.labeler.getServices

Colab notebook

https://shorturl.at/ZicnE

TASK 1 : Getting labeler info

TASK 2 : Getting label-stream from a labeler

TASK 3 : Measuring labeling time

TASK 4 : Analyzing labeled content & consistency

"Only two genders exist" cluster

Only two genders exist: ○ Male ○ Female ~Anything other than that is pure confusion.

There’s only two genders

There are only 2 genders

Il existe deux genres, mâle et femelle. Le reste c'est de la psychiatrie.

There are 3 genders: Male Female Mentally ill!

Examples labeled as Intolerant by Bluesky

	Should the posts on the right be labelled?
There’s only 2 genders	✓
There’s only two genders	✓
Есть только два гендера	✓
There are exactly two genders. Just two.	✓
Gibt trotzdem nur 2 Geschlechter	✓

From Top-k posts from firehose

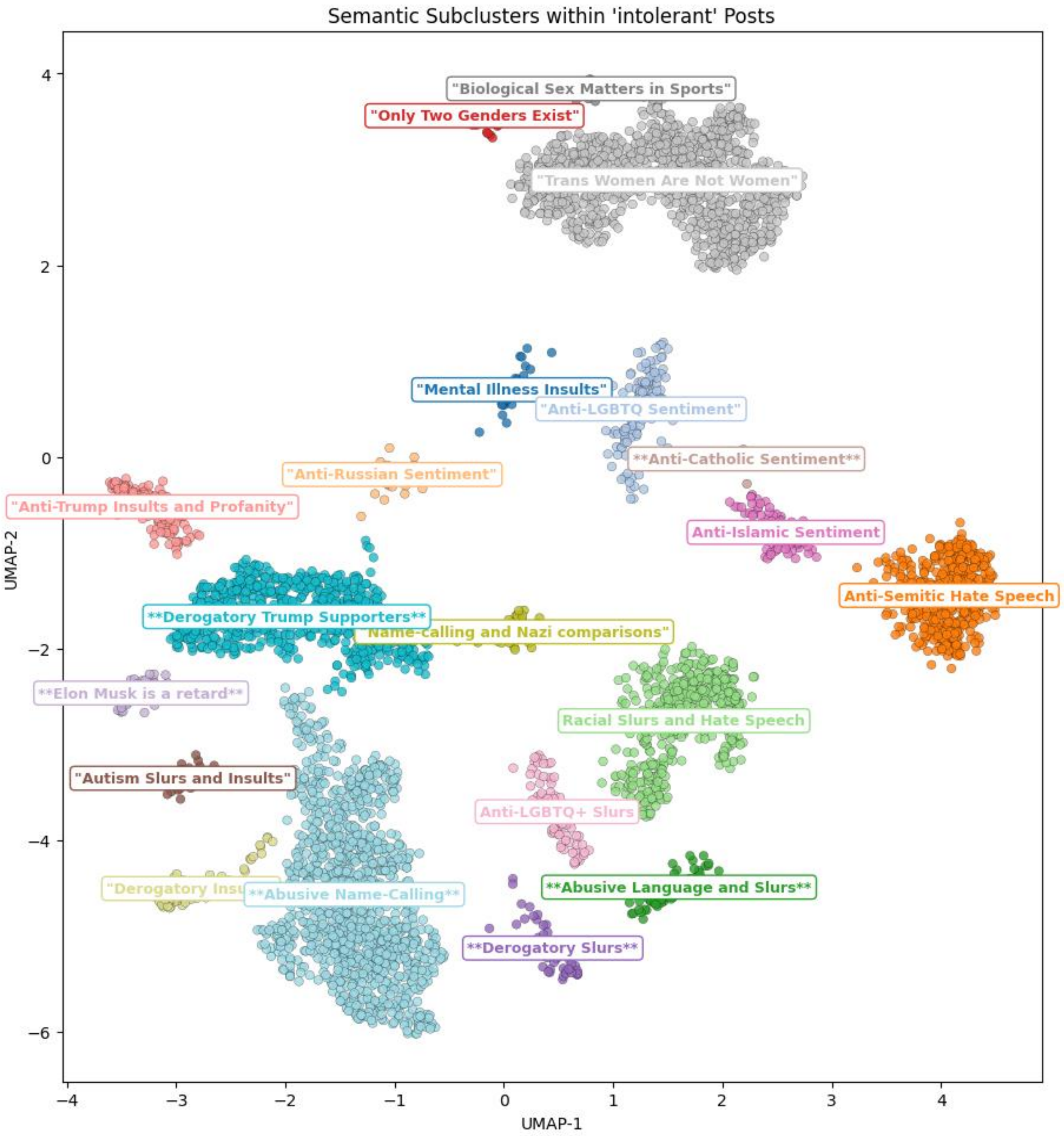
"Only two genders exist" cluster

Only two genders exist: ○ Male ○ Female ~Anything other than that is pure confusion.
There's only two genders
There are only 2 genders
Il existe deux genres, mâle et femelle. Le reste c'est de la psychiatrie.
There are 3 genders: Male Female Mentally ill!

Examples labeled as Intolerant by Bluesky

	Bluesky label intolerant?
There's only 2 genders	✓
There's only two genders	✗
Есть только два гендера	✗
There are exactly two genders. Just two.	✗
Gibt trotzdem nur 2 Geschlechter	✓

From Top-k posts from firehose


Semantic Subclusters within 'intolerant' Posts
"Biological Sex Matters in Sports"
"Only Two Genders Exist"
"Trans Women Are Not Women"
"Mental Illness Insults"
"Anti-LGBTQ Sentiment"
Anti-Catholic Sentiment
"Anti-Russian Sentiment"
"Anti-Trump Insults and Profanity"
Anti-Islamic Sentiment
Anti-Semitic Hate Speech
Derogatory Trump Supporters
"Name-calling and Nazi comparisons"
Elon Musk is a retard
Racial Slurs and Hate Speech
"Autism Slurs and Insults"
Anti-LGBTQ+ Slurs
"Derogatory Insu
Abusive Name-Calling
Abusive Language and Slurs
Derogatory Slurs
UMAP-2
UMAP-1
4
2
0
−2
−4
−6
−4
−3
−2
−1
0
1
2
3
4