Stanford | Internet Observatory Cyber Policy Center

Child Safety on Federated Social Media

David Thiel and Renée DiResta Stanford Internet Observatory July 24, 2023



Contents

1	Intr	oduction: Child safety on the Fediverse	2	
2	Bac 2.1	kground: Challenges of Fediverse moderation Case study in Fediverse child safety shortcomings	3 4	
3	Methodology			
	3.1	Ethics	5	
4	Threats and prevalence			
	4.1	Illustrated and Computer-Generated CSAM	6	
	4.2	CSAM posting and trading	7	
	4.3	Self-Generated CSAM trading and grooming	8	
	4.4	Messaging and DMs	8	
5	Directions for future improvement			
	5.1	Subscribable blocklists and discoverability limits	9	
	5.2	Pluggable hash matching and content classifiers	9	
	5.3	Tools for moderators	11	
	5.4	PhotoDNA support and CyberTipline reporting	11	
	5.5	ActivityPub extensions for attestation	12	
6	Con	Conclusion 14		

1 Introduction: Child safety on the Fediverse

Since becoming popular in the mid-2000s, the largest social media companies have operated largely in a centralized way: the entire network is controlled by a single central authority, and all user data is stored and managed on their servers. This is how Facebook, Twitter, and YouTube, for example, operate. However, due to user dissatisfaction and shifting social norms, federated social media—a decentralized approach to social networking in which multiple interconnected servers, called instances, are owned and operated independently by different organizations or individuals—has recently experienced a surge in popularity. Projects such as Mastodon, Bluesky, Pleroma and Lemmy offer new possibilities for a more resilient, protocol-based social media ecosystem not bound to a single company or entity. Users can create accounts on any instance they choose, and they have the freedom to follow or interact with users on other instances within the federation.

While decentralization can help distribute load among multiple entities, decentralized platforms offer new challenges for trust and safety. There is no central moderation team, for example, tasked with removing imagery of violence or selfharm, child abuse, hate speech, terrorist propaganda or disinformation. Each instance in a federated social media network may have its own set of rules and policies, with administrators moderating content and enforcing guidelines specific to their instance—this is considered a strong selling point, as users can find an instance that aligns with their values and tolerance levels for particular types of content. At a time when the intersection of moderation and free speech is a fraught topic, decentralized social networks have gained significant attention and many millions of new users.

However, significant harm categories and child safety in particular can become a very serious problem because of the regulatory arbitrage of content moderation: bad actors tend to go to the platform with the most lax moderation and enforcement policies. This means that decentralized networks, in which some instances have limited resources or choose not to act, may struggle with detecting or mitigating Child Sexual Abuse Material (CSAM). Federation currently results in redundancies and inefficiencies that make it difficult to stem CSAM, Non-Consensual Intimate Imagery (NCII) and other noxious and illegal content and behavior.

In this paper we take a broad look at child sexual exploitation concerns on decentralized social media, present new findings on the nature and prevalence of child safety issues on the Fediverse, and offer several proposals to improve the ecosystem in a sustainable manner. We focus primarily on the Fediverse (i.e., the ecosystem supporting the ActivityPub¹ protocol) and Mastodon, but several techniques could also be repurposed on decentralized networks such as Nostr, or

^{1.} Christine Lemmer-Webber et al., *ActivityPub* (W3C, January 2018), https://www.w3.org/TR/ac tivitypub.

semi-centralized networks such as Bluesky.²

2 Background: Challenges of Fediverse moderation

While well-resourced major social media companies largely operate as "walled gardens," with centralized control over content moderation policy and enforcement, federated social media puts responsibility for moderation into the hands of the instance operator. No instance can control the behavior of any other instance. There is additionally no governing body with the power to determine the legitimacy of instances or to completely remove a user or content from the ActivityPub network; as long as someone is willing to host an instance and permit specific content on that instance, the content remains a part of the ActivityPub network.³

Fediverse administrators are responsible for not only deciding what their users are allowed to post (content guidelines) on their instance, but also for the content posted by any users on remote servers that a local user follows. If a local user follows a remote user who posts illegal content, that content will be federated to the local server and potentially be displayed to users in their federated timeline, as well as stored on the server or media cache. The primary method of dealing with this is to defederate from servers with lax content moderation; in the context of user concerns about hate speech, for example, many large Mastodon servers chose to defederate from Gab, a right-wing social network. This is of course a blunt instrument; in the case of child safety, Japan has significantly more lax laws related to CSAM which has resulted in a cultural divide⁴ where most users in Japan are segregated from the rest of the Fediverse. Maintaining and monitoring for known bad instances is also largely an informal, voluntary process.

Apart from defederation, limiting exposure to illegal or harmful content is by and large left up to users themselves. Large platforms such as Meta and Twitter have teams responsible for implementing keyword and hashtag blocks and monitoring changes in keywords used by bad actors; while tools to block hashtags and filter posts with certain keywords in Mastodon exist, they are managed on an individual user basis.

Administrative moderation tooling is also fairly limited: for example, while Mastodon allows user reports and has moderator tools to review them, it has no built-in mechanism to report CSAM to the relevant child safety organizations. It also has no tooling to help moderators in the event of being exposed to traumatic content—for example, grayscaling and fine-grained blurring mechanisms.

^{2.} While parts of Bluesky are decentralized by design, core moderation tooling and staffing is currently centralized and operated by Bluesky itself. See Sol Messing, "On BlueSky," *Center for Social Media and Politics*, May 2023, https://csmapnyu.org/news-views/news/on-bluesky.

^{3.} Alan Z. Rozenshtein, "Moderating the Fediverse: Content Moderation on Distributed Social Media," *Journal of Free Speech Law* 217, no. 3 (November 23, 2022), https://doi.org/10.2139/ssrn .4213674.

^{4.} Ethan Zuckerman, "Mastodon is big in Japan. The reason why is... uncomfortable," *Medium*, August 18, 2017, https://medium.com/@EthanZ/mastodon-is-big-in-japan-the-reason-why-is-uncom fortable-684c036498e5.

It also has no default mechanism for integrating with PhotoDNA or other perceptual hashing infrastructure, making automated proactive detection of known CSAM impossible without local modifications by the administrator. This limitation is partly due to architecture: while a few large instances could use PhotoDNA, 25,000 servers near-simultaneously sending identical images to to the service is not a use case it was designed for, and would be wasteful at best and prohibitive at worst.⁵ Access to other hash databases is limited primarily by policy: due to the limited number of social media platforms, individual agreements between hash database providers (for example, NCMEC⁶ or GIFCT⁷) are made under contract, with their distribution intentionally limited to prevent bad actors from modifying images to evade the hashes or craft false positives.

Another limitation is the lack of signal with which to detect recidivism. While large social media providers utilize signals such as browser User-Agent, TLS fingerprint,⁸ IP and many other mechanisms to determine whether a previously suspended bad actor is attempting to re-create an account, Mastodon admins have little to work with apart from a user's IP and e-mail address, both of which are easily fungible. Including a sophisticated user fingerprinting system in the default distribution could give a roadmap to sophisticated bad actors, indicating exactly what criteria they would need to change to avoid detection.

2.1 Case study in Fediverse child safety shortcomings

The recent downtime⁹ of the Mastodon instance mastodon.xyz illustrates some of the limitations of child safety infrastructure in the Fediverse. In this case, the administrator and sole moderator received abuse reports regarding CSAM on the instance—either uploaded by a local user or ingested from a remote follow—that were not immediately acted upon, resulting in an abuse report being sent to the server's hosting provider. This resulted in the instance admin inspecting and deleting the content, but by this time the xyz top-level domain had suspended the server's DNS domain name, effectively taking the entire site down and depriving the entire userbase of the service. While this action was later reversed, part of that reversal was the registrar adding the site to a "false positive" list to prevent future occurrences, even though what caused the action was not a false positive.

Several things could have helped mitigate this eventuality:

- Tiered reporting flows: for example, having child safety-specific moderation reports automatically escalated.
- A simple reporting flow for admins to refer content to CSAM triage organizations.

^{5.} We discuss a possible mitigation to this in Section 5.5 on page 12.

^{6.} The National Center for Missing & Exploited Children, "Our Work," accessed July 18, 2022, https://www.missingkids.org/ourwork.

^{7.} GIFCT, "About GIFCT," accessed July 18, 2022, https://gifct.org/about.

^{8.} Habdul Hazeez, "What is TLS fingerprinting?," April 7, 2022, https://fingerprint.com/blog/wha t-is-tls-fingerprinting-transport-layer-security.

^{9.} Amaury Rousseau, "mastodon.xyz suspension on July 5, 2023," July 7, 2023, https://thekinrar.fr/en/posts/xyz-suspension.

- Nudity classifiers to detect sensitive content not labeled as such.
- Access to perceptual hash services to detect known CSAM instances¹⁰ and automatically act on them.
- A more well-staffed (or potentially outsourced) moderation team.

It is also possible that mechanisms for detecting certain keywords or hashtags would have helped prevent these posts from going live, which we shall discuss in Sections 5.1 and 5.2.

3 Methodology

We performed a two day time-boxed ingest of the local public timelines of the top 25 accessible Mastodon instances as determined by total user count reported by the Fediverse Observer,¹¹ recording JSON metadata emitted by each server's API and submitting media to PhotoDNA¹² and Google's SafeSearch API¹³ for analysis. We used the same methodology as in Thiel, DiResta, and Stamos,¹⁴ with the Mastodon streaming API as input instead of Twitter's PowerTrack API.

3.1 Ethics

SIO performed this analysis in a manner to minimize data stored and prevent any exposure of sensitive content. The following precautions were taken:

- No media was archived. Images and video thumbnails were perceptually hashed and scanned by third-party services with a dedicated purpose of scanning for CSAM or explicit content. Instances of unknown CSAM were inferred by cross-referencing hashtags and SafeSearch results.¹⁵
- All positive PhotoDNA matches were automatically reported to NCMEC via the PhotoDNA reporting API.
- Manual analysis for examples shown in this report was performed on selected instances of content indicated not to be explicit. Any images were first examined via a browser extension¹⁶ designed to prevent inadvertent exposure to explicit content.
- All content examined was posted publicly, and no interactions with posts or users were performed. Apart from image analysis, no post or user content

^{10.} This is somewhat difficult with Computer-Generated CSAM (CG-CSAM) due to heavy proliferation of new generated content; see David Thiel, Melissa Stroebel, and Rebecca Portnoff, "Generative ML and CSAM: Implications and Mitigations," *Stanford Digital Repository*, June 24, 2023, https://doi.org/10.25740/jv206yg3793.

^{11.} https://fediverse.observer

Microsoft, "PhotoDNA," accessed July 16, 2023, https://www.microsoft.com/en-us/photodna.
 Google Cloud, "Detect explicit content (SafeSearch)," accessed July 16, 2023, https://cloud.goo gle.com/vision/docs/detecting-safe-search.

^{14.} David Thiel, Renée DiResta, and Alex Stamos, "Cross-Platform Dynamics of Self-Generated CSAM," *Stanford Digital Repository*, June 2023, https://doi.org/10.25740/jd797tp7663.

^{15.} Expansion of this cross-referencing technique to alert NCMEC of potential unknown instances of CSAM is a potential area of future research.

^{16.} adacable, Painless Peek, accessed July 16, 2023, https://github.com/adacable/painlessPeek.

was processed by any third party.

- No scraping of full user profiles, followers or followees was performed.
- The only user and post metadata downloaded for specific analysis were posts that either had a positive PhotoDNA result or used hashtags or keywords known to be in use by child exploitation communities.

4 Threats and prevalence

Out of approximately 325,000 posts analyzed over a two day period, we detected 112 instances of known CSAM, as well as 554 instances of content identified as sexually explicit with highest confidence by Google SafeSearch in posts that also matched hashtags or keywords commonly used by child exploitation communities. We also found 713 uses of the top 20 CSAM-related hashtags on the Fediverse on posts containing media, as well as 1,217 posts containing no media (the text content of which primarily related to off-site CSAM trading or grooming of minors). From post metadata, we observed the presence of emerging content categories including Computer-Generated CSAM (CG-CSAM) as well as Self-Generated CSAM (SG-CSAM).

Notably, a test run of this analysis pipeline detected its first instance of known CSAM in approximately 5 minutes of runtime. All instances of CSAM detected were reported to NCMEC for triage.

4.1 Illustrated and Computer-Generated CSAM

Since the release of Stable Diffusion 1.5, there has been a steady increase in the prevalence of Computer-Generated CSAM (CG-CSAM) in online forums, with increasing levels of realism.¹⁷ This content is highly prevalent on the Fediverse, primarily on servers within Japanese jurisdiction.¹⁸ While CSAM is illegal in Japan, its laws exclude computer-generated content as well as manga and anime.

The difference in laws and server policies between Japan and much of the rest of the world means that communities dedicated to CG-CSAM—along with other illustrations of child sexual abuse—flourish on some Japanese servers, fostering an environment that also brings with it other forms of harm to children. These same primarily Japanese servers were the source of most detected known instances of non-computer-generated CSAM. We found that on one of the largest Mastodon instances in the Fediverse (based in Japan), 11 of the top 20 most commonly used hashtags were related to pedophilia (both in English and Japanese).

CG-CSAM has also increasingly become commercialized, with users advertising for-profit private Discord channels or distributing bundles of CG-CSAM or customized generative models in exchange for money (see Figure 1) or cryptocurrency; in one instance, we observed an offer of non-explicit NFT images of

^{17.} Thiel, Stroebel, and Portnoff, "Generative ML and CSAM: Implications and Mitigations."

^{18.} Although not exclusively: note that instances of CG-CSAM were reportedly responsible for the downtime incident involving mastodon.xyz, discussed in Section 2.1.



children sold to gain access to hundreds of nude CG-CSAM images.

Figure 1: Accounts offering CG-CSAM in exchange for money, via private Discord or Fanbox. The accounts use hashtags specific to CSAM, but also more general ones such that they can be discovered via searches for #stablediffusion, #aiart or #porn.

4.2 CSAM posting and trading

Open posting of non-generated CSAM is also disturbingly prevalent. Using our automated detection pipeline, we found multiple accounts posting dozens of instances of known CSAM on one of the largest Mastodon servers. Accounts posting CSAM would remain active for hours or days, gaining dozens of followers; notably, when these accounts were removed, no "delete" events were published, which may mean that federated servers that ingested the content received no notification that the content should be removed.

Offers of trading or sales of non-computer-generated CSAM are also common, as are linkouts to sites claiming to host CSAM. Actual sales appear to be negotiated via Session¹⁹ (see Figure 2), Matrix²⁰ or Telegram, though the payment methods used are not apparent in most cases in our dataset.



Figure 2: Left: an account offering large quantities of CSAM in exchange for money. Right: an outlink to a site indicated as distributing CSAM.

^{19.} https://getsession.org

^{20.} The Matrix.org Foundation, "About Matrix," accessed July 18, 2022, https://matrix.org/about.

4.3 Self-Generated CSAM trading and grooming

While advertising of commercial Self-Generated CSAM (SG-CSAM) is far less prevalent than on larger platforms such as Instagram and Twitter,²¹ it is still present on the Fediverse. Demographics of apparent underage sellers of explicit content are notably different on the Fediverse: sellers tend to primarily be teenage boys or transgender girls, while on other platforms sellers are almost entirely cisgender girls.²² Redacted examples of seller accounts can be seen in Figure 3.



Figure 3: Two apparently underage accounts offering explicit imagery of themselves in exchange for money or virtual currency.

There also is frequent solicitation on some instances of sexual interaction (virtual or otherwise) between adults and minors, likely indicating grooming occurring in private posts or Session chats (see Figure 4 for examples).



Figure 4: Left: An account offering CSAM for trade to underage persons, with Session ID for contact. Right: An account identifying as being underage and open to contact with adults.

4.4 Messaging and DMs

The ActivityPub specification²³ does not provide any guidance for Direct Messages; in Mastodon, DMs are more akin to "posts with an audience of two", and are readable by instance admins. Because of this, DMs on Mastodon are unlikely to be a primary channel for child exploitation-related activity. Instead, users in

^{21.} Thiel, DiResta, and Stamos, "Cross-Platform Dynamics of Self-Generated CSAM."

^{22.} While not explicitly called out as a research finding, this was true of the accounts characterized in Thiel, DiResta, and Stamos.

^{23.} Lemmer-Webber et al., ActivityPub.

these communities most commonly request contact on Session, an encrypted messenger forked from Signal²⁴ that uses onion routing²⁵ by default and requires no phone number or e-mail address (instead, using a hash as an identifier). Session allows either large group chats or one to one encrypted communication, and is so heavily associated with CSAM that posts not containing a Session ID still will use the "#session" hashtag as a discovery mechanism (see Figure 2 on page 7).

Lack of end-to-end encrypted DMs (or indeed, any easy to use direct messaging) pushing users to other platforms has mixed results with regard to child safety: using Session is extremely slow and requires some technical understanding, limiting its reach. On the other hand, if Mastodon theoretically had end-to-end encrypted DMs or chat groups, it would at least have access to the e-mail and IP addresses of the users involved, and users in such a group could report to instance admins.

5 Directions for future improvement

Some limitations in child safety tooling are inherent to any decentralized network. Additional limitations are due to the primarily volunteer nature of content moderators. However, there are potential technical solutions that could mitigate many of these deficiencies, provided sufficient development resources are made available.

5.1 Subscribable blocklists and discoverability limits

The hashtag and keyword blocklists made available to Mastodon users are very useful for avoiding content that individual users want to avoid: posts about certain political events, disasters, or other personally disturbing content. However, as discussed in Section 2, it is unreasonable to expect users to curate their own filter list of hashtags and keywords—particularly as hashtags and keywords related to CSAM change with high frequency.²⁶ The ability to perform filtering of hashtags and keywords for discoverability at the server level would be more effective at this task. Regular expressions could be shared between administrators in a manner similar to Bluesky's subscribable "mute list" feature.²⁷

5.2 Pluggable hash matching and content classifiers

Perceptual hashing²⁸ of images is a lightweight way to detect previously classified images without visual examination, in such a way that is resistant to minor image

^{24.} https://signal.org

^{25.} Oxen, "Network infrastructure," October 11, 2021, https://docs.oxen.io/oxen-docs/products-built-on-oxen/session/network-infrastructure.

^{26.} For example, bad actors have responded to blocks of the obvious hashtag "#pedo" by simply adding additional "o" characters to the end.

^{27.} The Bluesky Team, "Bluesky User FAQ," May 19, 2023, https://blueskyweb.xyz/blog/5-19-2023-u ser-faq.

^{28.} Hany Farid, "An Overview of Perceptual Hashing," *Journal of Online Trust and Safety* 1, no. 1 (October 2021), https://doi.org/10.54501/jots.v1i1.24.

transforms. Large social media platforms use this technology to identify known images that are either illegal or have been identified to violate their terms of service. This technology is used by PhotoDNA and provided as a cloud service to platform providers, while other perceptual hash databases are distributed privately and stored locally by a platform. As different platforms choose to use different hash databases depending on their needs, a mechanism should be in place for Fediverse administrators to integrate either industry standard or private hash databases in a selective manner.

A model for pluggable moderation mechanisms could be Pleroma's Message Rewrite Facility (MRF),²⁹ which allows for custom policies to perform ActivityPub message analysis, modification and rejection. This facility also allows (via its Hashtagpolicy) automatic rewrites or rejections of posts with particular hashtags, a feature that could combine well with subscribable or distributed hashtag blocklists discussed in Section 5.1 on the previous page. As noted by Anaobi et al.,³⁰ the policies in use by individual Pleroma servers are public and can be used to inform decisions about which servers to federate with.

There remains the problem of distribution of hash databases: it may be that there is an instance size tradeoff where larger, more professionally run instances are worth the risk of hash database abuse, while very small ones would not meet this bar. Alternatively, it may be that the current model of privately distributed hash databases is not compatible with distributed systems; these services could be moved to a SaaS model similar to PhotoDNA, either by the owners of the hash databases or a third-party clearinghouse.

The Fediverse community could also decide to implement its own hash databases to control illegal or undesirable content (for example, Nazi content illegal in Germany), either keeping instance-level databases to prevent recidivism, distributing them among administrators, or hosting them publicly.

Large social media companies also have the benefit of highly trained machine learning classifiers for tasks such as nudity or gore detection in user-supplied content, but this functionality is lacking from the Fediverse ecosystem. There are some open-source models available,³¹ but no tooling appears to currently exist to allow administrators to easily integrate with them. Such models could be implemented similarly to hash matching mechanisms, with a facility similar to MRF and UI to download, activate and customize moderation system behavior.

^{29.} Pleroma, "Message Rewrite Facility," accessed July 16, 2023, https://docs-develop.pleroma.soc ial/backend/configuration/mrf.

^{30.} Ishaku Hassan Anaobi et al., "Will Admins Cope? Decentralized Moderation in the Fediverse," in *Proceedings of the ACM Web Conference 2023* (ACM, April 2023), https://doi.org/10.1145/3543507.3 583487.

^{31.} See, for example, Bumble Inc., *Private Detector*, accessed July 16, 2023, https://github.com/bumble-tech/private-detector.

5.3 Tools for moderators

The problems of moderator trauma and burnout at large social media companies and companies they outsource to are well known.³² Over time, companies and researchers have developed strategies for reducing the psychological impact of viewing traumatic imagery: for example, grayscaling, color shifting or blurring³³ content by default, with the ability to partially unblur images. These tools are fairly easy to implement in a web UI, and can be automatically applied to user reports regarding CSAM or violence.

Improved tooling to detect recidivism would also help limit moderator load; while having fingerprint logic public could allow for some bad actors to evade detection, implementation of basic user fingerprinting could better inform moderator and admin decisions. This could be accomplished by integrating a software package such as FingerprintJS³⁴ or could potentially outsourced to a third-party commercial CDN.

5.4 PhotoDNA support and CyberTipline reporting

As mentioned in Section 2, individual use of PhotoDNA by every server in the Fediverse is likely to be prohibitive due to the potential for bursts of tens of thousands of near-simultaneous requests. However, given that the majority of Fediverse users reside on the top 10 or so servers, it would be prudent to integrate detection tooling into the codebase itself. For the front-end UI, all that would be necessary to make for easy integration is allowing admins to input their PhotoDNA and CyberTipline API keys³⁵ (see mockup in Figure 5 on the following page).

On the backend, tooling would be needed to perform a request to PhotoDNA for each image ingested by the server, with a positive match triggering the content to be discarded and removed from any intermediary storage mechanism, as well as a report to NCMEC via PhotoDNA containing match details and basic user information.

Administrative UI should also be developed to manually submit reports to NCMEC's CyberTipline API.³⁶ While there is a submission form hosted by NCMEC

^{32.} Amit Pinchevski, "Social media's canaries: content moderators between digital labor and mediated trauma," *Media, Culture & Society* 45, no. 1 (2023): 212–21, https://doi.org/10.1177/01634437221122226.

^{33.} Anubrata Das, Brandon Dang, and Matthew Lease, "Fast, Accurate, and Healthier: Interactive Blurring Helps Moderators Reduce Exposure to Harmful Content," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 8, 1 (October 2020), 33–42, https://doi.org/10.1609/hcomp.v8i1.7461; Sowmya Karunakaran and Rashmi Ramakrishan, "Testing Stylistic Interventions to Reduce Emotional Impact of Content Moderation Workers," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, 1 (October 2019), 50–58, https://doi.org/10.1609/hcomp.v7i1.5270.

^{34.} *FingerprintJS*, accessed July 16, 2023, https://github.com/fingerprintjs/fingerprintjs.

^{35.} Microsoft, "PhotoDNA Documentation," accessed July 16, 2023, https://www.microsoft.com/e n-us/photodna/documentation.

^{36.} The National Center for Missing & Exploited Children, "CyberTipline Reporting API Technical Documentation," May 25, 2023, https://report.cybertip.org/ispws/documentation.

Stanford Internet Observatory Cyber Policy Center

M	PhotoDNA Configuration
	PhotoDNA API key e Your key from https://myphotodna.microsoftmoderator.com
Back to Mastodon	
Public profile	
Preferences	CyberTip username* Obtained from NCMEC
🖌 Flavours	username
Follows and followers	CyberTip password*
▼ Filters	*****
う Automated post deletion	SAVE
Account	

Figure 5: Mockup of potential UI for admins to input PhotoDNA access tokens. Selection of individual hash databases used by PhotoDNA is configured via the PhotoDNA portal itself.

allowing submission of tips by service operators, auto-populating data from the report and local logs would be a more efficient process.

If such a system were developed and made easy to implement, it is our hope that large instances would be inclined to adopt it—particularly those in Japan with significant child exploitation problems that do still attempt to moderate CSAM that is known involve or appears to involve an actual child.³⁷

5.5 ActivityPub extensions for attestation

To address capacity concerns with simultaneous PhotoDNA or other cloud scanning integrations, we propose an alternative system of attestation of analysis. In such a system, the server on which a post originates would submit imagery to PhotoDNA for analysis; as part of the response, the PhotoDNA service would include a cryptographic hash of the image along with a signature of that hash.³⁸

In the event of a match of known CSAM, the PhotoDNA Match API returns match metadata so that the originating server can make an automated report to NCMEC in a subsequent request to the PhotoDNA Report API.³⁹ In the event of no match, the Match API would return a tracking ID along with the hash and signature. The post and imagery could be distributed to other servers, with an ActivityPub "update" event containing the hash of the image and signature, effectively attesting that the content has been scanned by PhotoDNA. The receiving servers could

^{37.} Due to the potential unwillingness of these servers to moderate computer-generated or manga/anime content, we suggest having separate hash databases and/or classification structure for content that can be identified as being generated or illustrated. See Thiel, Stroebel, and Portnoff, "Generative ML and CSAM: Implications and Mitigations."

^{38.} To maintain the integrity of the hash, any image transforms such as resizing or downscaling of the image should be performed before submission to PhotoDNA. This ensures that the hash is of the image that may ultimately be sent to other servers in the federation, rather than of the original image uploaded.

^{39.} Note that PhotoDNA returning the hash and attestation is not strictly necessary in this case.

then verify the authenticity of the attestation with a public key distributed by PhotoDNA. Figure 6 shows a simplified illustration of the process.

This same technique could also be applied to other hosted media analysis mechanisms (e.g. Google's SafeSearch or Microsoft's Analyze Image API⁴⁰), drastically reducing the costs compared to each instance performing analysis itself.



Figure 6: Sequence flow of the PhotoDNA matching, reporting and attestation process. For size and performance, signing a hash of m is likely preferable to signing m itself.

^{40.} Microsoft, "Computer Vision REST API reference," accessed July 18, 2022, https://learn.microsoft.com/en-us/rest/api/computer-vision.

6 Conclusion

Federated and decentralized social media may help foster a more democratic environment where people's online social interactions are not subject to an individual company's market pressures or the whims of individual billionaires. For this environment to prosper however, it will need to solve safety issues at scale, with more efficient tooling than simply reporting, manual moderation and defederation. The majority of current trust and safety practices were developed in an environment where a small number of companies shared context and technology, and this technology was designed to be efficient and effective in a largely centralized context.

Conversely, decentralized platforms have relied heavily on giving tools to endusers to control their own experience, to some degree using democratization to justify limited investment in scalable proactive trust and safety. Counterintuitively, to enable the scaling of the Fediverse as a whole, some centralized components will be required, particularly in the area of child safety. Investment in one or more centralized clearinghouses for performing content scanning (as well as investment in moderation tooling) would be beneficial to the Fediverse as a whole. Given new commercial entrants into the Fediverse such as WordPress, Tumblr and Threads, we suggest collaboration among these parties to help bring the trust and safety benefits currently enjoyed by centralized platforms to the wider Fediverse ecosystem.

Stanford Internet Observatory Cyber Policy Center

The Stanford Internet Observatory is a cross-disciplinary program of research, teaching and policy engagement for the study of abuse in current information technologies, with a focus on social media. The Stanford Internet Observatory was founded in 2019 to research the misuse of the internet to cause harm, formulate technical and policy responses, and teach the next generation how to avoid the mistakes of the past.

Stanford | Internet Observatory Cyber Policy Center

