

[A/C Privileged] Athena Targeted Detection Launch One Pager (WIP)

Author: [HYPERLINK "mailto:[REDACTED]@google.com" \h]

Contributors: [HYPERLINK "mailto:[REDACTED]@google.com" \h]

COPPA & Athena-v3 Classifier

To comply with COPPA (Children's Online Privacy Protection Act) YouTube is required to prevent the usage of its platform by unsupervised users under the age of 13. Consequently, we require users to confirm they are at least 13 when creating a YouTube account, and remove accounts when we suspect them to be of underage (below 13) users.

There are multiple avenues (General TnS review queues, Child Safety Livestream queue, External Reports) through which YT accounts can get flagged and even possibly removed for suspicion of being underage, the most prominent being the Athena-v3 classifier.

The Athena-v3 classifier was launched on February 8, 2021. The classifier uses Golden7 features and looks at the previous 10 uploads of an account to determine the likelihood that it is operated by an underage user.

Since its launch, Athena-v3 has terminated over 35.1M channels ([HYPERLINK "https://dashboards.corp.google.com/_fb650775_d938_4a5a_99b6_7db0e19bd161" \h]) for suspicion of being underage.

How does Athena work today?

For a given channel, Athena looks at up to the last 10 uploads of a channel (audio and video characteristics only) and generates a confidence score that predicts the likelihood a channel is operated by an under 13 user. The confidence scores are then mapped to precision buckets through human evals, where the legal ops team assesses whether an account is underage or not ([HYPERLINK "?tab=t.0" \l "bookmark=id.sehz66cmo51d" \h]).

Athena operates at two distinct confidence score/precision points

1. High Precision (Target: 95% / Current Operation: 99%) - 95%+ of the accounts within this score range were determined to be underage by legal ops
 - a. Result: Auto-action - Trigger 14 day account age verification grace period
2. Low Precision / Active Learning Queue (Target: 80% / Current Operation: 80%) - 80%+ of the accounts within this score range were determined to be underage by legal ops

- a. Result: Enqueue account to legal review queue for human evaluation and policy application if determined to be likely underage
 - i. This is used as the active learning queue by the ML model

Youtube actions 300K¹ accounts per week for suspicion of being underage. The accounts are identified as a result of scanning a random .3%² of the corpus on a daily basis.

¹YT is capped at actioning 300K accounts per week because when an account is suspected to be underage it triggers a Google-managed flow called [[HYPERLINK "?tab=t.0" \l "bookmark=id.7fs06qn8htxs" \h \]](https://www.youtube.com/watch?v=7fs06qn8htxs), which relies on human reviewers for ID verification[[HYPERLINK "?tab=t.0" \l "bookmark=id.7fs06qn8htxs" \h \]](https://www.youtube.com/watch?v=7fs06qn8htxs) Given human review capacity limitations, YT is limited to actioning 300K accounts per week.

²We only scan .3% of the corpus daily because we only want to score as many accounts as we can action in a given week - 300K. YT operates under the principle that once we have knowledge of an underage account we must action it.

What criteria is used to determine if an account is underage?

The Legal Ops team has developed an extensive set of criteria ([[HYPERLINK "https://docs.google.com/presentation/d/1ZU64UOJO7zVxIavgaFntP5FSOFW77o2S3S6kKfOrhX8/edit" \l "slide=id.g2a88f03cb3_0_148" \h \]](https://docs.google.com/presentation/d/1ZU64UOJO7zVxIavgaFntP5FSOFW77o2S3S6kKfOrhX8/edit)) to assess if an account is underage. The criteria boils down to 3 core questions

1. Has the account expressed admission of being underage in a video, comments, or in metadata (channel description or video title/description) - identifying that the account is clearly owned by an underage child?
2. Is the account clearly underage given developmental indicators (specifically body development lacking any secondary sex characteristics) and behavior and contextual indicators?
3. Is it possible the account is owned by an adult - indicated through presence and acknowledgement in videos, channel metadata, adult filming behind the camera?

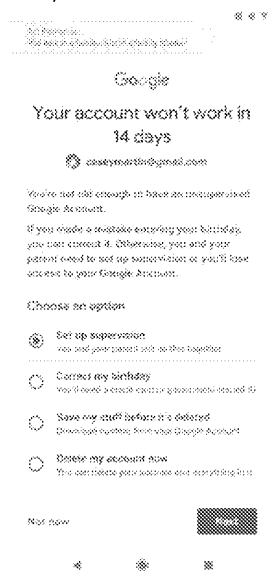
What is the user experience of an account that is suspected to be underage?

Once an account is suspected to be underage the user enters the Patronus flow which is broken into two distinct phases

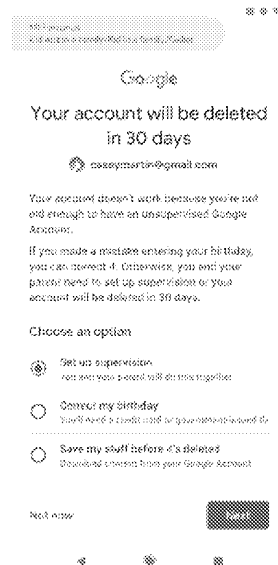
- Phase 1 (14 day grace period): During this time the user account is active and the user is presented with 4 options
 - Set up supervision:
 - The child and their parent will go through the Alaska onboarding flow on the child's device (detailed flow is described in the [[HYPERLINK](#)

"https://docs.google.com/document/d/116alazSAoZzXob1HSHU6T9z2_tvUXW-WnXK1WFdgSZY/edit" \ "bookmark=id.kul64tfuxuq7" \h }).

- Correct my birthday:
 - This assumes that the user entered the wrong age in the DOB collection screens & the user provides a government ID or a credit card as proof for their age correction.
- Takeout:
 - A regular takeout of the user data is available.
- Delete my account:
 - A regular takeout of the user data is available.
 - The user can proactively delete their account.



- Phase 2 (30 day pre-deletion): During this time the user account is no longer active and the user is presented with the same options as before excluding the “Delete my account now” selection.



What account protections do we have in place?

Athena Classifier Protections

- Only scan S6 (Aspiring) channels
- Exclude Dasher/Madison accounts
- Exclude Unicorn accounts
- Exclude accounts that have either 1) already age verified or 2) legal reviewers confirmed to be 13 or above

Athena Legal Review Protections

- Premium accounts & accounts with over 10K subs are escalated to tier 2 reviewers

Athena's Limitations

Over the last few years, TnS & Legal Ops have identified the following limitations with the current Athena setup

1. Latency: The 300K per week cap and randomized corpus scan means that active underage accounts can possibly skirt detection for a long time. Today, underage accounts that are actioned by Athena are generally on the platform for 938 days before detection - giving them plenty of time to create content and continue putting themselves and the platform at risk.

- a. Legal Ops has flagged that given this latency, often they can't action clearly underage accounts because while the content they reviewed was created before the users 13th birthday, they are now clearly old enough to be on the platform, i.e., by the time Athena flags the accounts and sends them for review the user may now be 14 but they were obviously abusing the platform for a few years before that.
2. ID Verification Loophole: A large number* of suspected underage channels flagged by Athena, ID verify and continue creating content similar to that which got them flagged in the first place. (Not addressed in this proposed Athena Targeted Detection Launch)

**Data shows that 30-50% of channels that go through the Patronus flow convert to SupeX or age verify. Historically, ~1% of Athena channels convert to SupeX.*

Proposed Change: Athena Targeted Detection

Motivation

In Q1 of 2023, the TnS Child Safety and Shorts team engaged in a joint effort to identify Child Safety risks on Shorts. The workstreams effort's centered around the following [[HYPERLINK "https://docs.google.com/presentation/d/1LQHi9QanN3BSbqQZ3uxj3l-_ZWT_cdkLTkiyB-1-7PM/edit?resourcekey=0-JtVoROeschvQCPeDwL7PDg" \l "slide=id.g215d233b1c9_1_0" \h \l](https://docs.google.com/presentation/d/1LQHi9QanN3BSbqQZ3uxj3l-_ZWT_cdkLTkiyB-1-7PM/edit?resourcekey=0-JtVoROeschvQCPeDwL7PDg)] potential user safety, legal/regulatory, brand risks*:

1. The increased participation of minors on Shorts vs. VOD (37% of Shorts contain minors <18 vs. 19% on VOD)
2. Creation Tool Shorts 2.5X more likely to be minor risky (Bambo) than VODs
3. The aggregation of classifiers misses, specifically Bambo on surfaces such as Pivots
4. The aggregation of suspected underage account content on surfaces such as Pivots
5. 90% of Creation Tool Shorts CSAM is "self-gen" vs. 40% on VOD
6. Intel desk flagging that effects were possibly impacting TnS detection

As a result of the workstream, TnS agreed to adopt the following [[HYPERLINK "https://docs.google.com/presentation/d/1MazFo2SC3hOZXxvxAQJHZ9kWXQRUCU9hUxB5eizDCNNA/edit?resourcekey=0-B18RYtBKZI36DHMa5dbmhA" \l "slide=id.g20fc5717bde_0_0" \h \l](https://docs.google.com/presentation/d/1MazFo2SC3hOZXxvxAQJHZ9kWXQRUCU9hUxB5eizDCNNA/edit?resourcekey=0-B18RYtBKZI36DHMa5dbmhA)], including the proposed Athena Targeted Detection change. The hypothesis was that a large number of underage users are on Shorts vs. VOD and the plan was to focus our limited Athena resources on removing active underage accounts participating in the Shorts ecosystem.

- Reduces total volume of underage account content
- Reduces total volume of classifier misses

Subsequently, throughout 2023 we saw other instances reinforcing our suspicions regarding underage participation on Shorts / highlighting the need for the Athena changes

- TnS & ACE CS detection analysis given that 70% of effects are Bambo risky
 - [HYPERLINK "https://docs.google.com/document/u/0/d/1s2xpOsnJdFbluPDD00ckmxomvaOtQ Nviv8uWhDgsN_0/edit" \h]
- Blocking effects pivot launch due to concerns about CS protections
 - [HYPERLINK "https://docs.google.com/document/u/0/d/1Wx3l3_PBK667aX2VAZw83ilBwiEZm vbtIzSFrac4M_A/edit?resourcekey=0-2NgIkv27s1ISmO7UIP5uCg" \h]
- Underage accounts revealing PII in Shorts-Audio trend
 - [HYPERLINK "https://buganizer.corp.google.com/issues/314396047" \h]
- Shorts Shelf featuring Bambo/near Bambo content
 - [HYPERLINK "https://docs.google.com/presentation/u/0/d/1ZN8QEcvT4zAF-19fYcBpdrJd4CHdgTNPbKCKYqbsDU/edit?resourcekey=0-7Rw2l_fE6c4zZC-FM8DcgQ" \h]
 - [HYPERLINK "https://www.wsj.com/tech/meta-instagram-video-algorithm-children-adult-sexual-content-72874155" \h]

The common theme across each of these escalations / workstreams was that if we had less underage accounts creating risky or Bambo content we could have either avoided the situation entirely or greatly mitigated the risks.

Separately, given the regulatory landscape w.r.t Child Safety we are seeing increasing attention / scrutiny to how well companies are preventing underage users from accessing their platform

- [HYPERLINK "https://www.nytimes.com/2023/11/25/technology/instagram-meta-children-privacy.html" \h]
- [HYPERLINK "https://www.nytimes.com/2023/04/04/business/media/tiktok-fine-uk-kids-privacy.html" \h]

Athena Targeted Detection Details

The two key choices made when designing Athena Targeted Detection were the following:

- 1) Triggering Athena on upload¹ / livestream creation as opposed to a corpus scan
 - a) Newer content gains more views and is in turn more discoverable. Scanning and removing active accounts will better rid the ecosystem of content that puts the users/platform at the most risk
- 2) Triggering Athena on Shorts/Livestream SCT uploads as opposed to all uploads
 - a) Given the scanning capacity limitations of Athena it makes the most sense to focus our resources where the problem is the most acute.

¹ Public or Unlisted videos only - every 24 hours Athena pulls the list of uploaded videos and the channels associated with them, ignoring videos that are Private at that instance

	Current Setup	Proposed Launch Athena Targeted Detection
Scanning Trigger	Corpus Scan	On-Upload* <i>*Upload of video or Livestream using the SCT</i>
Scanning Coverage	.3%* of corpus per day <i>*Random selection</i>	25%* of SCT uploaders / livestreamers per week <i>*Random selection</i>
Weekly Actioned Accounts	300K	450K* <i>*Worked with Birthday Correction Flow team to secure additional capacity</i>

Date	Stage description	Done?
2023-11-15	Rollout MiniReibe enqueues based on Paced Classification (50% A/B).	Yes
2023-11-27	Roll out athena_v3_underage_suspend at 300 actions / day (0.5%)	Yes
2023-12-05	Recalibrate athena_v3_underage_suspend and ramp up to 840 [HYPERLINK "https://monitoring.corp.google.com/explorer?qe_title=Construx%20matche s&q_names=Rule%20Matches%20(QPS)%20for%20athena_v3_underage _suspend&duration=259200&refresh=3600&expr=CpQCCpECCo4CEIMK EG1vbmFyY2Y2uQm9yZ1Rhc2sSPwo9L3ZpZGVvL2xlZ29zL2R1cGxvL3Ru c19ydWxlcY9kZXRIY3Rpb25fc291cmNIX3ZlcnNpb25fbWF0Y2Y1hlcxpbcGlib 3JnX3VzZXISTggCEkp5b3V0dWJlXRUcy1jb21tZW50cy1kdXBsb3x5b3V 0dWJlXRUcy1rYXBsYS1kdXBsb3x5b3V0dWJlXRUcy1ydWxlcY1kdXBsb xo8ChBtZXRYaWM6cnVsZV9uYWU1IEglIARikZGV0ZWNoaW9uL2F0aGVu YV92M191bmlRlcmFnZV9zdXNwZW5kGhwKDW1ldHJpYzpzZGF0dXMS wgBEGdNQVRDSEVEChAiDiIKCAIqBAIAowUwAjaCCj06OwgBEGclAzoD CNgEGi4KGm1ldHJpYzpkZXRIY3Rpb25fc291cmNIX2lkChBtZXRYaWM6c nVsZV9uYWU1&percentiles=0&utc_end=0" \h] (1%).	Yes

TnS <launch approval>		
2023-12-12	Ramp up active learning enqueues from 50% to 100% Paced (and fully retire enqueues from Batch feed).	Yes
2023-12-12	Roll out athena_v3_underage_suspend to 5k actions / day (~8%). Decrease Athena Batch auto-actions from 43k to 38k/day.	Yes
2023-12-14	Roll out athena_v3_underage_suspend to 6.5k actions / day (~8%) ¹ . Decrease Athena Batch auto-actions from 43k to 36.5k/day.	Yes
Pending EMCO launch approval		
January	Ramp up auto-actions to 22k and ramp down Athena Batch to 21k (50%).	
January TBC	Ramp up auto-actions to 43k. Disable Athena Batch.	
January TBC	Ramp up auto-actions to 64k / day.	

Expected Impact of Change

TnS Metrics

- P0: Decrease # of days detected underage account on platform by X%
- P1: Reduce Bambo trigger rate on Short Camera Tool by X%
 - To be updated: TnS working on a Bambo based safety metric in H1 2024
- P1: Reduce Livestream VVR by X%
- P2: Reduce CS Shorts VVR by X%
 - Policies likely impacted
 - 1032: Innocently posted minor nudity
 - 1004: Minor provocative behavior

Shorts Ecosystem Impact

How many users would be impacted by this change?

Relevant Data Points

- 25% of channels that create Shorts Camera Tool based Shorts/livestream are not [[HYPERLINK "?tab=t.0" \ "bookmark=id.9k8p085azfva" \h](#)].

¹ Increase athena_v3_underage_suspend rollout from 15% to 100%. After this, further rollout will be controlled by ramping up the daily Paced Classification volume.

- Of the eligible channels, early results show that ~20% of those, score high enough to be auto-actioned by Athena
- Given Athena capacity limitations, we will scan a random 1.5M (19%) of eligible Shorts uploaders/livestreamers per week
- 30-50% of channels that go through the Patronus flow convert to SupeX or age verify. Historically, ~1% of Athena channels convert to SupeX.

Criteria	Classification Volume (weekly Channels)
Shorts (Camera Creation Tool)	6.3M
Shorts (all)	30.3M
Livestreams (Cameo)	1.7M
Livestreams (all)	38M
All	100M
All, without [HYPERLINK "http://google3/learning/tfx/users/yt_video_abuse/athena/prod/plx/scoreable_channels.sql"]	234M

Estimated impact

1-2 Months after full launch, Athena Targeted Detection will likely remove 7.5% - 10.5% of Shorts Camera Tool users.

Who are these users?

They are all s6

They create this sort of content -

% Bambo

Channel/Content examples

Commented [1]: @adajain@google.com
@benoitd@google.com do we have any estimates how much this will impact Live?

FAQ

What age verification options do other platforms offer?

Instagram

[HYPERLINK "<https://help.instagram.com/271237319690904>" \h]

[HYPERLINK "<https://about.fb.com/news/2022/06/new-ways-to-verify-age-on-instagram/>" \h]

- ID Verification

- One from this group

- Birth certificate
- Driver's license
- Passport
- Marriage certificate
- Official name change paperwork
- Personal or vehicle insurance card
- Non-driver's government ID (example: disability, SNAP or national ID card)
- Green card, residence permit or immigration papers
- Tribal identification or status card
- Voter ID card
- Family certificate
- Visa
- National age card
- Immigration registration card
- Tax identification card

- Two from this group

- Bank statement
- Transit card
- Check
- Credit card
- Employment verification
- Library card
- Mail
- Magazine subscription stub
- Medical record
- Membership ID (example: a pension card, union membership, work ID, professional ID)
- Paycheck stub
- Permit
- School ID card

- School record
- Social Security card
- Utility bill
- Yearbook photo (actual scan or photograph of the page in your yearbook)
- Company loyalty card
- Contract
- Family registry
- Diploma
- Religious documents
- Certificate of registration for accreditation or professional
- Professional license card
- Polling card
- Health insurance
- Address proof card
- Social welfare card
- Video selfie
- Social vouching

TikTok

<https://support.tiktok.com/en/safety-hc/account-and-user-safety/underage-appeals-on-tiktok>

- Selfie with government issued ID
- Photo with parent/guardian
- Credit card authorization
- Selfie based facial age estimation (if clearly over 18)