

FILED UNDER SEAL

EXHIBIT 25

From: Launch Manager [REDACTED]@fb.com]
Sent: 8/24/2022 9:21:20 AM
To: [REDACTED]@fb.com]
Subject: [Privacy Decision] Next steps for Privacy Review L1079249PRV3



Next steps for Privacy Review L1079249PRV3

Hello [REDACTED],

Thank you for accepting the Privacy Decision for **Integrity Relational Understanding Service (I-RUS) - Privacy - RUS embeddings for Child Safety Bad Actor detection**.

There's more work to do before you can launch. See below for next steps. If you're new to this part of the process or need a refresher, visit the [Evidence Submission wiki](#) for more details.

Your Privacy Decision includes **13 mitigations** which must be verified by 6/29/22. Here's what happens next:

1. [REDACTED], who you selected earlier as the Engineering Owner, has received a task to verify and provide evidence for each mitigation.

Privacy Review: Integrity Relational Understanding Service (I-RUS) - Privacy - RUS embeddings for Child Safety Bad Actor detection

6/29/22 · Reporting & Review, AI Integrity · L1079249PRV3



Privacy Decision Summary

Privacy Review

Description

SCOPE

Surface: Facebook only

Experiment time-box: 6 weeks

Country: Global (excluding the EU and UK)

User "Behavioural" Data: Actor ID, User ID, Post ID, Content ID, Likes, Reactions, Comments, Shares, Follows, Groups, Pages, Relationships on the platform connecting users to content (Note: No inference is made for off-platform relationships or type of relationship) - collected from the 90 days prior to labelling.

BACKGROUND

Violations in child safety are some of the most egregious on our platform and for the larger society. However, with end-to-end encryption (e2ee) across private surfaces, the child safety team will lose most of the signals used to detect actual violations. Addressing this requires a change of approach, from detecting in the moment of danger, to pre-empting danger that a minor is at risk. Most of the prevalence in the problem space is driven by a small set of bad actors, and establishing the intent of actions is critical to ensure accurate and proportionate enforcement. For example, the child safety team has found that 70-80% of accounts sending CEI do not intend to exploit or harm children, and that we can identify this intent using encryption-resilient signals.

The Child Safety team built an Intent Model for Bad Actor Detection, which captures the signals mentioned in L1084219PRV. However, the most important and challenging part of the model is to implement social context graph information on the actor into features that the model can use. This is achieved by aggregating various kinds of attributes of graph fanouts from the user as features, which is an expensive process of manual feature engineering. RUS embeddings capture the social context surrounding FB entities like users, posts, etc. The hypothesis is that the network signals captured in the social context of the RUS embedding will improve the model while making the process of capturing this information more efficient.

CURRENT APPROACH

The current Child Safety Intent Classification model captures this information by aggregating various kinds of attributes of fanouts data from the user as features, which is an expensive process of manual feature engineering.

PROPOSED APPROACH

L1079249PRV approved the development of the Relational Understanding Survey (RUS) embeddings to understand user relationships on the platform. The output of the embedding is a score that determines the nature of the relationship between two entities and degree of confidence in the relationship. The Interactions Integrity - Child Safety team would like to experiment and explore how the RUS embedding could support the Child Safety Intent Classification ML model (L1084219PRV, L1074754PRV). The RUS embeddings will be used to plug in the social graph information more easily and efficiently at scale.

The processing will follow the following data flow:

1) The team will access existing data sets in order to create new RUS embeddings that will then be used to train the Intent classifier model.

2) Specifically, the team will pull/ingest the data sets for different violator groups and control groups, as follows:

- Users disabled in strategic network disruptions (approx. 15k), including users who disabled back in October 2021.
- Users who engaged in CSAM solicitation (approx. 7.5k)
- Users who engaged in IIC with young teens (Approx. 3k)
- Users labeled as malicious in our Intent review queue and from other manual processes (Approx. 1k)
- Users who have adult sexual content violations, but no known child safety violations (Approx. 38k)
- Users who have fake account violations, but no known child safety violations (Approx. 38k)
- Randomised FB users (Approx. 76k)

3) The team will analyse user behaviour from the past 90 days prior to the date of labelling* in order to create the RUS embeddings for all violation types. These interactions can be fetched based on FBID for the last 90d, but beyond 90d we need to fetch data via RID route (thru RID<>FBID mapping).

4) Following analysis of the above-mentioned user behaviour, the RUS embeddings will be used to create new features that will feed into and train the Child Safety Intent classifier.

*date of labelling = date on which the user was determined to show certain characteristics for malicious intent towards children

EXAMPLE FLOW:

- User A was disabled from CS related violations in Oct 2021
- The CS team will review user data to create the RUS embeddings, dating back to 90 days prior to the date of labelling
- The data for this activity is analysed by the CS team
- The data is further labelled and used to train the Intent classifier model
- All data for this activity is retained for a maximum of 90 days

CLASSIFIER TRAINING DATA (FROM RUS EMBEDDINGS DATA)

The RUS embeddings will be used to train the Child Safety Intent classifier.

PRIVACY CONSIDERATIONS

- The team will run the above experiment for a maximum of 6 weeks. Once the experiment is completed, the embedding will be removed from the classifier feature set.
- If the experiment is successful, full launch will go through PXFN review.
- The data processing for this activity will not include Messenger or IG Direct data.
- Retention of RUS embeddings are limited to 90 days after date of embedding calculation i.e. the date point on which the embedding was actually created/formulated from.
- The lookback period for the data to construct the RUS embedding is limited to 90 day prior to date of labelling.
- The team will not use deleted users' data (if a disabled user has requested for account deletion, they will not be included in this experiment).

Mitigations

1. This activity will process user data for a maximum of 75k benign users selected at random from MAP
2. Human annotation must be conducted on FB premises by a FB employee or supervised contractor who has undergone appropriate training.

3. If the labellers find any new CEI that we haven't already flagged it will be escalated to the safety team
4. User Deletions will be respected (to include but not limited to post deletion (images and text), audience downgrades) except in the case of confirmed child safety violations whereby data stored against an RID may be retained for up to 365 days.
5. Will implement access controls in Hive so that only individuals with relevant business purposes can access the data.
6. The experiment results will be kept for a maximum of 90 days or less.
7. Once the experiment is completed, the embeddings will be removed from the classifier feature set.
8. Behaviour data collected to construct RUS embeddings will be limited (in all cases) to the previous 90 days prior to date of initial Child-Safety labelling OR 90 days prior to the date on which they were selected (applicable to benign users)
9. Not to use EU or UK data
10. RUS embeddings will be retained for no longer than 90 days
11. RUS embeddings will not make inferences to off-platform relationships or types of relationships
12. For users disabled in SNDs (who were disabled in October 2021), behaviour data collected to construct embeddings will be limited to data collected in the 90 days prior to October 2021 and no earlier data sets will be used.
13. All datasets (except for those disabled in SNDs), data collection will be limited to users who have been labelled as violating in the 90 days prior to the date of the RUS embedding creation date.

[Click to configure Launch Manager Review notification settings.](#)