[redacted] uploaded a file in the group: **Harmful Topic Community Working Group**.

April 29 at 5:27 PM ·

Policy update on Querdenken proof of concept experiment

On Tuesday we met with several folks from different policy teams to discuss the proof of concept experiment we plan to run on German conspiracy theory movement, Querdenken (experiment proposal)

MEETING NOTES

- We took similar action against parts of Qanon in the early days which was effective and shows precedent.

- This is a good intermediate step between doing nothing vs waiting until it's at an escalated level that would warrant hard actions. This could be a good case study to inform how we tackle these problems in the future and policy development

- Transparency component needs more consideration if we want to scale this

- If there any regional elections coming up try to avoid them as it would bias the results do to likely surges in engagement

- In parallel we should talk to Sean Conner to determine if we wan to add Querdenken to the Conspiracy Theory Non-rec policy. But this doesn't block the experiment

- **Consensus by the group is that this experiment is OK to move forward without further approvals (beyond PXFN approval and [redacted] double checking with her team to confirm)**

NEXT STEPS

- [redacted]: Check for local elections
- [redacted] to annotate how he sees Querdenken aligning with non-rec policy in experiment doc
- [redacted]: Confirm demotion strength and look into deboosting
- [redacted] Follow up with [redacted] on Friday to see if any flags or further approval needed
- [redacted] Complete PXFN approval
- [redacted] Keep this group in the loop as things progress

ny, Apr 13

[Draft] Experiment to curb the growth of Germany Querdenken harmful topic community (policy facing)

1. Introduction

1.1 Querdenken

Querdenken (roughly translated as "thinking outside the box") is a conspiracy movement based in Germany, primarily mobilized around the belief that the German government's COVID-19 restrictions are disproportionate and part of a larger plan to strip citizens of their political freedoms and basic rights. It emerged in April 2020 under the leadership of Michael Ballweg, a software entrepreneur based in Stuttgart. Ideologically, Querdenken overlaps with both QAnon, a designated VICN, and the Reichsbürger movement, the latter referring to several groups in Germany who reject the legitimacy of the modern German state. Querdenken movement has organized multiple protests throughout Germany, many of which have resulted in violence and injuries to the police. In December 2020, German intelligence agencies placed Querdenken under state surveillance, citing concerns over infiltration by extremists. (i3 deck with more info on Querdenken, NYT article)

The organization has a robust on-platform presence; the primary Page (Querdenken 711) has 26k followers, and over 100 smaller regional Querdenken Pages and Groups ranging in size from dozens to thousands of followers/members. There is no policy designation of Querdenken as of today for Facebook.

It causes harm both on and off-platform in the following ways:

1. **Deep connection with conspiracy theory.** The primary narrative of Querdenken advocates that the German state has responded disproportionately to the Covid-19 virus by instituting nation-wide lockdown measures. However, it also draws on broader conspiracy theories, specifically focused on deep-state plots to increase the power of the ruling elite by withholding true information about the virus. Its PSR group promotes Qanon related conspiracy and spread anti-semantic conspiracy etc.
2. **Offline Violence.** Despite a public stance against violence, several protests organized by Querdenken that have resulted in acts of violence (ACLED GOOGLE SHEETS LINK)

1.2 Harmful Topic Community

Dangerous content team in CI is developing a new archetype named harmful topic community (HTC), which focuses on detecting and curbing the growth of communities organized around topics that can cause potential harm through

It causes harm both on and off-platform in the following ways:

1. **Deep connection with conspiracy theory.** The primary narrative of Querdenken advocates that the German state has responded disproportionately to the Covid-19 virus by instituting nation-wide lockdown measures. However, it also draws on broader conspiracy theories, specifically focused on deep-state plots to increase the power of the ruling elite by withholding true information about the virus. Its PSR group promotes Qanon related conspiracy and spread anti-semantic conspiracy etc.
2. **Offline Violence.** Despite a public stance against violence, several protests organized by Querdenken that have resulted in acts of violence (ACLED GOOGLE SHEETS LINK)

1.2 Harmful Topic Community

Dangerous content team in CI is developing a new archetype named harmful topic community (HTC), which focuses on detecting and curbing the growth of communities organized around topics that can cause potential harm through radicalization and normalization. When people come together organically and form communities around harmful topics or identities, the potential for harm can be greater. Harmful topic communities pose a greater risk for harm to our users, integrity efforts, and brand, **primarily as a result of increase their communal normalization of harmful attitudes and behaviors, and behavioural and attitudinal radicalization** (community harm research).

This work is different with network disruption (ND) in the following ways. This note explains in more details.

1. **Type of network or community**. ND focuses on worst of the worst networks with high coordination of activities that violates our existing community standards or there is policy designation for the network. HTC focuses on communities that are loosely and organically connected around topics that can cause harm, with or without existing policy designation.
2. **Enforcement strategy**. There are different tiers of actions in ND such as disabling accounts in tier 1 and soft action in tier 2. For HTC, the major focus is to curb the growth and protect users at-risk of becoming normalized and radicalized to harmful topics, not necessarily eradicate the presence of the network/community. As a result, HTC relies more on soft actions (such as demotion in feeds, non-rec and feature limit) to contain the reach and potential growth of the community.

2. Goals of the Experiment

Given the high risk of this community due to the upcoming Germany election this year and the fact that Querdenken fits nicely with the definition fo harmful topic community, it is a good candidate to test the hypothesis and inform the future strategy of HTC. Specifically, we have the following hypothesis and goals:

1. We can **quickly** identify the community around Querdenken topic with **high precision and recall,** using the detection model from HTC.
2. We can **curb the growth** of Querdenken community.

2. Goals of the Experiment

Given the high risk of this community due to the upcoming Germany election this year and the fact that Querdenken fits nicely with the definition fo harmful topic community, it is a good candidate to test the hypothesis and inform the future strategy of HTC. Specifically, we have the following hypothesis and goals:

1. We can **quickly** identify the community around Querdenken topic with **high precision and recall,** using the detection model from HTC.
2. We can **curb the growth** of Querdenken community.
    a. We can reduce the engagement and new connections between existing members
    b. We can prevent at-risk users from joining the community and being radicalized
3. We want to **understand** what is the most effective way to enforce on harmful topic communities. We have several enforcement levers we plan to test:
    a. Non-rec
    b. Feature limit
    c. Demotion in feeds ranking

The learnings from this experiment will be used to help inform:

1. New policy development to address the harm that can comes when a community forms around a harmful topic or ideology
2. How to prevent harmful movements from taking root and growing on our platforms
3. Understanding of where current mitigations fall short

3. Experiment Design

The detection model of HTC will output the following:

1. A list of entities (users, groups and pages) that belong to the Querdenken community
2. A list of susceptible users (an output from our detection model that predicts the likelihood whether a user is on path of joining the target community) that are not part of yet, but are on the path of radicalization and becoming part of the community.

For the entities in 1, the Querdenken community, we will apply the following soft actions:

1. Non rec: Filter in recommendation engines such as GYSJ, PYMK, PYML and PYMI
    a. GYSJ: Susceptible users or users in the community won't see group recommendations for groups belonging to the Querdenken community.

3. Experiment Design

The detection model of HTC will output the following:

1. A list of entities (users, groups and pages) that belong to the Querdenken community
2. A list of susceptible users (an output from our detection model that predicts the likelihood whether a user is on path of joining the target community) that are not part of yet, but are on the path of radicalization and becoming part of the community.

For the entities in 1, the Querdenken community, we will apply the following soft actions:

1. Non rec: Filter in recommendation engines such as GYSJ, PYMK, PYML and PYMI
    a. GYSJ: Susceptible users or users in the community won't see group recommendations for groups belonging to the Querdenken community.
    b. PYMK: Susceptible users or users in the community won't see friend recommendations for individuals belonging to the Querdenken community.
    c. PYML: Susceptible users or users in the community won't see Page recommendations for Pages belonging to the Querdenken community.
    d. PYMI: Susceptible users or users in the community won't show up as invite recommendations for users in the Querdenken community
2. Feature limit: limit features such as group notifications, limit group invites etc.
    a. Filter Group/Page invitation notifications, e.g. remove off-platform notifs, remove reminder notifs, remove inline CTAs on notifs
    b. Filter Group/Page notifications
    c. Group/Page invite rate limit
3. Feeds demotion: demotes the contents of the entities in the feeds. We will apply a demotion no stronger than 25%
    Due to technical complexity we may not be able to include this intervention as a part of this experiment

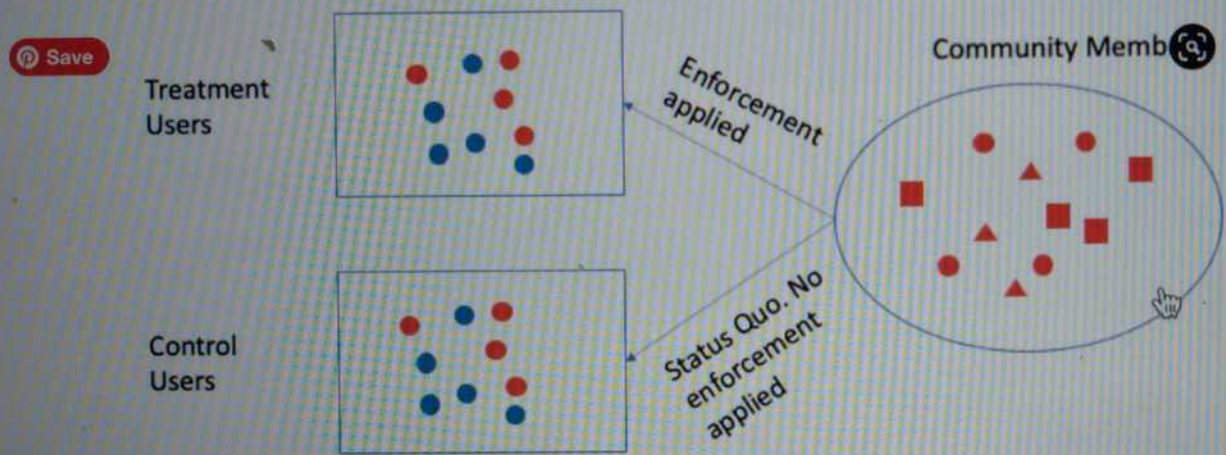
Those soft action will not be universal in the sense that it will only be applied for a subset of target viewers. For the users that are either part of the community or susceptible to join the communities, we will randomize them into treatment and control users. The soft actions on the community entities will be applied on the viewers in the treatment group, but not the control group. For example, let's say for a Querdenken group, treatment users will not see them in the GYSJ recommendations, but control users will still see the current status quo recommendations without any modifications. With this randomization on the viewer level, we hope to reliably measure the effectiveness of those soft actions.

Figure below illustrate the high level idea of the experiment

- **User in the community**

Figure below illustrate the high level idea of the experiment

- User in the community
- Page in the community
- Group in the community
- Susceptible users



Test groups

1. Control
2. Community interventions on susceptible users
3. Community interventions on all users

We will run the experiment for X weeks and then measure the difference between treatment and control users for the following metrics either through labeling, aggregation score on calibrated classifiers or strikes.

1. For the newly connected entities (groups, pages and users) of treatment and control users, what is the difference of percentage/average classifier score/number of strikes that are related to Querdenken/hate/violence
2. For the newly interacted contents of treatment and control users, what is the difference of percentage/average classifier score/number of strikes that are related to Querdenken/hate/violence

4. Timeline

4. Timeline

1. **Policy approval** - 1 week, 4/23
    a. Iterate on experiment plan and design as needed
2. **Configure test** - 3 weeks, 5/7
    a. Identify the community, i.e. what are the groups, pages, users we're defining as the community. Vet with i3 based on sampling with an aligned precision
    b. Eng work to configure randomization of test audience and setting up interventions. HTC team and i3 Eng team
    c. QA testing on setup
3. **Run test** - 3 weeks, 5/28
    a. Health check that test is running as intended ~ 24hrs after test start
4. **Analyze** - 1 week, 6/5
    a. DC and i3 compile engagement and success metrics

5. Discussion and Open Questions

Can we run experiment before we have a formal policy designation of this?
Querdenken movement has demonstrated clear on and off platform harm, such as violence that cause conflicts and injuries in many past protests and its deep connection with conspiracy theory such as Qanon and anti-semitsm. It will be great to have a policy designated for it, but may take time to develop a comprehensive policy around it. We want to move fast given the timing of upcoming German and the potential harm of this movement. Is there a way we can utilize some other policies such as VNI and conspiracy theory for this experiment?

Are there any risk of over enforcement and limiting the voice?
We believe the risk of limiting the voice is minimal given the following reasons:

1. We will only take soft actions and will not disable any accounts. We are not removing any content or preventing the production of content. Moreover, group / page members will still be able to access all content on the group/page surface.
2. The overall population for this is small.
3. We will run this experiment for limited period of time.

On the other hand, the benefit of the enforcement outweighs the risk here, because of the upcoming German election and the demonstrated harm of Querdenken movement.

Risk and mitigation

2 Risk and mitigation

VICN designation. If Querdenken is designated under VICN, the core elements of the community will be removed which will obviously make us unable to run the experiment that we want. That said, there is a reasonable chance that this won't be designated under VICN

Regional elections skewing engagement. We expect surges in engagement around this movement coinciding with regional elections that are taking place to the national election in September. To reduce the risk of biased results we will try to avoid running our experiment over any regional elections. There is one on June 6th which would mean we would want to start our experiment by May 14th to get three weeks of testing time (election calendar)

Appendix

1 **Conspiracy Theory Non-rec policy**

2 **Does this fit within the existing policy?**

Non-recommendable Conspiracy Theories are either:

1. Theories that rely on unverifiable or unproven information to explain, deny, or prove the existence of events, practices, or circumstances, especially as the secret work of individuals or groups, and at least one of the following:
    a. The theory relates to
        i. (1) health; (2) finance; or (3) the cause of an event or action involving significant harm to people or property or a similarly significant risk to public safety; or
        ii. action or inaction by government where a government is alleged to have physically harmed or attempted to physically harm people (other than action/inaction related to war or national defense); or
        iii. conspirators controlling a government, governmental bodies, or aspects of the government's decision-making; or
            1. Yes, Querdenken is a conspiracy movement based in Germany, primarily mobilized around the belief that the German government's COVID-19 restrictions are disproportionate and part of a larger plan to strip citizens of their political freedoms and basic rights
    b. The theory has been connected to, or is likely to contribute to, offline physical violence or crimes involving significant destruction of property or infrastructure.
        i. Yes, this movement is linked to the storming of the Reichstag and other violent events (link)
    c. The theory asserts that a specifically identified individual is perpetrating the secret work at issue in the theory, and the theory may create significant reputational or physical-safety risks for the identified individual.

to start our experiment by May 14th to get three weeks of testing time (election calendar)

Appendix

1 Conspiracy Theory Non-rec policy

2 Does this fit within the existing policy?

Non-recommendable Conspiracy Theories are either:

1. Theories that rely on unverifiable or unproven information to explain, deny, or prove the existence of events, practices, or circumstances, especially as the secret work of individuals or groups, and at least one of the following:
    a. The theory relates to
        i. (1) health; (2) finance; or (3) the cause of an event or action involving significant harm to people or property or a similarly significant risk to public safety; or
        ii. action or inaction by government where a government is alleged to have physically harmed or attempted to physically harm people (other than action/inaction related to war or national defense); or
        iii. conspirators controlling a government, governmental bodies, or aspects of the government's decision-making; or
            1. Yes, Querdenken is a conspiracy movement based in Germany, primarily mobilized around the belief that the German government's COVID-19 restrictions are disproportionate and part of a larger plan to strip citizens of their political freedoms and basic rights
    b. The theory has been connected to, or is likely to contribute to, offline physical violence or crimes involving significant destruction of property or infrastructure.
        i. Yes, this movement is linked to the storming of the Reichstag and other violent events (link)
    c. The theory asserts that a specifically identified individual is perpetrating the secret work at issue in the theory and the theory may create significant reputational or physical-safety risks for the identified individual.
2. Theories (except for those rooted in religious beliefs) that:
    a. rely on unverifiable or unproven information; and
    b. deny the existence of, or otherwise refute, scientific facts or prominent historical events; and
    c. all or nearly all verifiable facts about the topic of the theory are to the contrary; and
    d. the theory asserts or relies on a secret cover-up.

art our experiment by May 14th to get three weeks of testing time (election calendar)

ppendix

nspiracy Theory Non-rec policy

s this fit within the existing policy?
-recommendable Conspiracy Theories are either:

1. Theories that rely on unverifiable or unproven information to explain, deny, or prove the existence of events, practices, or circumstances, especially as the secret work of individuals or groups, and at least one of the followi
    a. The theory relates to
        i. (1) health; (2) finance; or (3) the cause of an event or action involving significant harm to people or property or a similarly significant risk to public safety; or
        ii. action or inaction by government where a government is alleged to have physically harmed or attempted to physically harm people (other than action/inaction related to war or national defense); or
        iii. conspirators controlling a government, governmental bodies, or aspects of the government's decision-making; or
            1. Yes. Querdenken is a conspiracy movement based in Germany, primarily mobilized around the belief that the German government's COVID-19 restrictions are disproportionate and part of a larger plan to strip citizens of their political freedoms and basic rights
    b. The theory has been connected to, or is likely to contribute to, offline physical violence or crimes involving significant destruction of property or infrastructure.
        i. Yes, this movement is linked to the storming of the Reichstag and other violent events (link)
    c. The theory asserts that a specifically identified individual is perpetrating the secret work at issue in the theory, and the theory may create significant reputational or physical-safety risks for the identified individual.
2. Theories (except for those rooted in religious beliefs) that:
    a. rely on unverifiable or unproven information; and
    b. deny the existence of, or otherwise refute, scientific facts or prominent historical events; and
    c. all or nearly all verifiable facts about the topic of the theory are to the contrary; and

[Draft] Experiment to curb the growt...

Conversation

You're new to this document

3. Experiment Design
View comments (1 new)
[redacted] FYI on experiment actioning
Reply

[redacted] made edits · Apr 30
feeds. We will apply a demotion no stronger than 25%. Due to technical complexity we may not be able to isolate this
View Changes

2. Community interventions susceptible users
View comments (1 new)
[redacted] Apr 30
NOTE: Because of technical challenges in the AB testing framework we may not be able to
Reply

Conspiracy Theory Non-rec policy
View comments (1 new)
[redacted] Wed
Reply

REDACTED FOR CONGRESS

[redacted]

Go team!

Like · 2w

[redacted]

this is great work! Next state elections will take place on June 6th in Saxony-Anhalt, and then we'll have the Federal ones (+ other 3 state elections) in September. I'm leading the German Elections XFN so will be following this closely

2

Like · 2w

[redacted]

Thanks for pointing out the election in June [redacted] We're hoping to be able to ideally close the experiment before then which would mean we'd want to start no later than May14th

Like · 2w

[redacted]

Follow up: No further policy approvals were identified by [redacted] so we are good to go after PXFN approval is finalized

2

Like · 1w

[redacted]

Meeting notes: We met with [redacted] and [redacted] who have been working very closely to Querdenken and other concerning movements in Germany ahead of the elections in September.

- Overall, they support running this experiment and think that it's a good learning opportunity.
- They offered to help i3 review the sampling of groups and pages that we identify to reduce the risk of including benign entities.
- They mentioned that there are other good community candidates that might be earlier in the lifecycle and faster growing that could be interesting as follow up experiments. This could be a great opportunity to run a joint experiment with CORGI models [redacted].

2

Like · 1w · Edited

[redacted] some more help we can get to review our model result.

1

Like · 1w

[redacted]

A joint experiment would be great, would love to learn more about the communities you're targeting, and the interventions you have in mind. We could potentially replicate the experiment we're doing with KRUSH. (cc

Chats

REDACTED FOR CONGRESS

[redacted] [redacted]

Meeting notes: We met with [redacted] who have been working very closely to Querdenken and other concerning movements in Germany ahead of the elections in September.

- Overall, they support running this experiment and think that it's a good learning opportunity.
- They offered to help i3 review the sampling of groups and pages that we identify to reduce the risk of including benign entities.
- They mentioned that there are other good community candidates that might be earlier in the lifecycle and faster growing that could be interesting as follow up experiments. This could be a great opportunity to run a joint experiment with CORGI models (cc [redacted]

Like · 1w · Edited

2

[redacted] some more help we can get to review our model result.

1

Like · 1w

[redacted]

A joint experiment would be great, would love to learn more about the communities you're targeting, and the interventions you have in mind. We could potentially replicate the experiment we're doing with KRUSH. (cc [redacted]

2

Like · 1w

[redacted] let me know if that's the case and I can make sure to request German translation for the survey.

1

Like · 1w

[redacted] **Harmful Topic Community Working Group**

[redacted] May 14 at 3:02 PM ·

Querdenken proof of concept experiment update

Yesterday we had a meeting with PXFN Legal and Policy as well as Policy partners from DOI, Security, German and Product Policy teams to discuss some concerns with our experiment

REDACTED FOR CONGRESS

[redacted] ▸ Harmful Topic Community Working Group

May 14 at 3:02 PM ·

Querdenken proof of concept experiment update

Yesterday we had a meeting with PXFN Legal and Policy as well as Policy partners from DOI, Security, German and Product Policy teams to discuss some concerns with our experiment

NOTES

- There is a lot of support for this experiment generally speaking but there are a couple of flags that have been raised that we are working through
- German legal counsel raised litigation concerns. Product counsel has escalated and is working with German legal counsel on understanding the level of risk and should have a decision by Friday. Further escalation may be needed
- There is a Privacy Policy concern around how the model is using sensitive data that is specific to the EU. This would likely need to be escalated - We do not want to run over or too close to the regional election on June 6th. Due to the time required to resolve the above outstanding issues **we have aligned on starting the experiment after the 6th assuming we have necessary approvals.**
- We will be exploring adding Querdenken into the Non-rec conspiracy theory policy which would enable an easier path to applying non-rec treatments as well as demotions and feature limits for a temporary experiment

NEXT STEPS

- Wait for legal determination - [redacted]
- Pursue non-rec addition - [redacted]
- Confirm Privacy Policy next steps once we have clarity from legal [redacted]

cc. [redacted]

4 3 Comments Seen

have necessary approvals.
- We will be exploring adding Querdenken into the Non-rec conspiracy theory policy which would enable an easier path to applying non-rec treatments as well as demotions and feature limits for a temporary experiment

NEXT STEPS

- Wait for legal determination - [redacted]
- Pursue non-rec addition - [redacted]
- Confirm Privacy Policy next steps once we have clarity from legal [redacted]

cc [redacted]

4 3 Comments Seen by 24

Like Share

[redacted]
Like · 2d

[redacted]
Like · 2d

[redacted]
Thanks for pushing through all the complexity.
This seems like partly a cautionary tale against focusing on the "community" side.
Like · 1d

[redacted] ▸ Integrity HPMs
May 11 at 7:27 PM ·

Network Disruptions and Harmful Topic Communities Archetype HPM -

REDACTED FOR CONGRESS

[redacted] Integrity HPMs

[redacted] May 11 at 7:27 PM ·

Network Disruptions and Harmful Topic Communities Archetype HPM - May 11th 2021

Mission: *Minimize the impact of violating or harmful networks and communities at scale.*
Roadmap | Goal Tracker

METRICS OVERVIEW

As both Network Disruptions and Harmful Topic Communities are new areas and don't have metrics yet, we will use this space provide a running total of networks that we detect vs our goal across our detection workstreams.

Terrorism: 12 of 20
Sex Trafficking: 2 of 15
Child Safety: 3 of 5

HIGHLIGHTS

1. **[Harmful Topic Communities] We have clarified with CORGI that there is a some overlap between our workstreams.** Our objectives and approaches are the same, however our models still could provide some unique value to the platform CORGI is developing to tackle community based harms. We will continue executing on H1 goals where there is no overlap such as understand work, model validation and our proof of concept experiment. For efforts such as our policy goal where there is duplication and we're driving the same thing, we will step back from that and let CORGI continue to drive and report on progress.
2. **[Harmful Topic Communities] We have presented our work (UX research, investigation and detection model) to DC Circle** (slides and note with more detail)

LOWLIGHTS

1. **[Harmful Topic Communities] Querdenken proof of concept experiment risk.** We are working against a deadline to start the experiment by this Friday 5/14. There are some legal and policy questions the XFN and PXFN teams are following up on which we hope to have clarity and approval on by mid week.

PROJECT PROGRESS

NETWORK DISRUPTIONS

1. **Detection**
    a. DOI: we identified 12 TPs after finishing reviewing the rest of batch #3 (findings in this note).
    b. Child Safety: We identified 2 TP networks in our recent Batch #4. In our subsequ[illegible] batches we would double down by experimenting with score thresholds of inten[illegible]

Chats

LOWLIGHTS

1. **[Harmful Topic Communities] Querdenken proof of concept experiment risk.** We are working against a deadline to start the experiment by this Friday 5/14. There are some legal and policy questions the XFN and PXFN teams are following up on which we hope to have clarity and approval on by mid week.

PROJECT PROGRESS

NETWORK DISRUPTIONS

1. **Detection**
    a. DOI: we identified 12 TPs after finishing reviewing the rest of batch #3 (findings in this note).
    b. Child Safety: We identified 2 TP networks in our recent Batch #4. In our subsequent batches we would double down by experimenting with score thresholds of intent classifier to further increase our coverage. COINV investigators will now be supporting this workstream so we expect to have additional SMEs for more iteration
    c. HEx: Total TP identified: 2
        i. This week am working on improving seed quality by implementing a better fanout from content classifier to group, page, user owners.
        ii. Waiting for review feedback from last submitted batch of clusters
    d. Eng April updates can be seen in this post
2. **Scaling Reviews and Disruptions**
    a. We have continued to iterate on the proposal for scaling network reviews in H2. The initial approach we will test involves disaggregating the network into nodes and enqueueing them in a profile review for violation determination and enforcement. We hope to run a proof of concept dry run in a flow like this to compare the violation match rate of networks already reviewed by i3 and COINV

HARMFUL COMMUNITIES

1. **Detection**
    a. We have finished several iteration of models for Querdenken pipeline. Result from some initial samples looks promising. We are working with COINV and I3 investigation team to have some estimate about the precision on larger population.
    b. We have started implementing enforcement levers, including non-rec and feeds ranking demotion and it is on-track to launch the experiment once legal concerns has resolved.
    c. We got seeds set from VNI team and are in the process of running our model on VNI violation type to help better understand the harm concentration in VNI space.

— with [redacted] and **2 others**.

Chats

enqueueing them in a profile review for violation determination and enforcement. We hope to run a proof of concept dry run in a flow like this to compare the violation match rate of networks already reviewed by i3 and COINV

HARMFUL COMMUNITIES

1. **Detection**
    a. We have finished several iteration of models for Querdenken pipeline. Result from some initial samples looks promising. We are working with COINV and I3 investigation team to have some estimate about the precision on larger population.
    b. We have started implementing enforcement levers, including non-rec and feeds ranking demotion and it is on-track to launch the experiment once legal concerns has resolved.
    c. We got seeds set from VNI team and are in the process of running our model on VNI violation type to help better understand the harm concentration in VNI space.

— with [redacted] and **2 others**.

Theme	Priority	ND or HC	Initiative	H1-Start (baseline)	H1-end (goal)	RAG Status	Update (5/11)
Efficiency	P1	ND	Detect 20 WoW DOI terrorism networks	0	20	On Track (In Progress)	12/20 TPs. We are ramping up Markets investigators this week which we expect to give us the volume of reviews we need to hit our goal and do some further model validation
Efficiency	P1	ND	Map investigator workflows and deliver plan for moving some investigator functions to scaled review in H2, conducting at least one scaled review proof of concept run.	n/a	n/a	On Track (In Progress)	Iterations are being made to proposal. Next step will be to start scoping out what an H2 scaled review flow would require
Measurement	P1	ND	Determine methodology for WoW miss rate measurement and begin tracking.	n/a	n/a	Likely to be MISSED	We have identified alternative options to WoW miss rate but without DS support on the team we won't be able to make meaningful progress towards this goal.
Coverage	P1	ND	Detect 5 WoW Child Safety networks for FB	0	5	On Track (In Progress)	3/5 TP. Back on track. We onboarded Malicious intent classifier onto our pipeline to source seeds and added few more heursitics to better rank clusters. We identified 2 TP networks in our recent Batch #4. In our subsequent batches we would double down by experimenting with score thresholds of intent classifier to further increase our coverage.
Coverage	P1	ND	Detect 15 WoW sex trafficking networks for HEx	0	15	BEHIND	2/15 TP. We have been constrained on review capacity but expect to have a Markets POC trained by EOW which should get us back on track
Efficiency	P2	ND	Determine policy approval (action, appeals, etc) to take automatic enforcement action (soft or hard action) on high precision clusters.	n/a	n/a	On Track (In Progress)	Automated enforcement approach is being discussed with i3 and will start looping in policy
Efficiency	P2	ND	Reduce time to onboard model from X (sex trafficking) to Y (child safety)	X	Y	NOT STARTED	Working on baseline and goal target
Coverage	P3	ND	Automatically detect and disrupt (manual or automatic) Y networks on IG for Terrorism.	0	Y	NOT STARTED	Not started
Legitimacy	P1	HC	Publish understand work identifying key players, how harmful communities evolve and Facebook's role in their growth	n/a	n/a	AT RISK	No current update due to PTO. We'll update this goal on Wednesday
Legitimacy	P1	HC	Determine policy to designate a community as harmful.	n/a	n/a	Deprioritized in DC, CORGI is driving	The harmful topic community aspect of the AHN policy is a part of the plan but still TBD on when it will be tackled. Disaggregating Harmful Network task force (CORGI) will continue to drive this with Policy. DC will step back from policy work in H1
Legitimacy	P1	HC	Build accurate detection model to detect harmful communities in Harmful Conspiracy Theory problem space, and quantify its potential impact in reducing the harm	n/a	n/a	On Track (In Progress)	- Have several iteration of the detection model on the Querdenken pipeline, worked with COINV investigators to get some pecision estin model accuracy looks good with some i

Chats

Save	Priority	ND or HC	Initiative	H1-Start (baseline)	H1-end (goal)	RAG Status	Update (5/11)
Efficiency	P1	ND	Detect 20 WoW DOI terrorism networks	0	20	On Track (In Progress)	12/20 TPs. We are ramping up Markets investigators this week which we expect to give us the volume of reviews we need to hit our goal and do some further model validation
Efficiency	P1	ND	Map investigator workflows and deliver plan for moving some investigator functions to scaled review in H2, conducting at least one scaled review proof of concept run.	n/a	n/a	On Track (In Progress)	Iterations are being made to proposal. Next step will be to start scoping out what an H2 scaled review flow would require
Measurement	P1	ND	Determine methodology for WoW miss rate measurement and begin tracking.	n/a	n/a	Likely to be MISSED	We have identified alternative options to WoW miss rate but without DS support on the team we won't be able to make meaningful progress towards this goal.
Coverage	P1	ND	Detect 5 WoW Child Safety networks for FB	0	5	On Track (In Progress)	3/5 TP. Back on track. We onboarded Malicious intent classifier onto our pipeline to source seeds and added few more heursitics to better rank clusters. We identified 2 TP networks in our recent Batch #4. In our subsequent batches we would double down by experimenting with score thresholds of intent classifier to further increase our coverage.
Coverage	P1	ND	Detect 15 WoW sex trafficking networks for HEx.	0	15	BEHIND	2/15 TP. We have been constrained on review capacity but expect to have a Markets POC trained by EOW which should get us back on track
Efficiency	P2	ND	Determine policy approval (action, appeals, etc) to take automatic enforcement action (soft or hard action) on high precision clusters.	n/a	n/a	On Track (In Progress)	Automated enforcement approach is being discussed with i3 and will start looping in policy
Efficiency	P2	ND	Reduce time to onboard model from X (sex trafficking) to Y (child safety)	X	Y	NOT STARTED	Working on baseline and goal target
Coverage	P3	ND	Automatically detect and disrupt (manual or automatic) Y networks on IG for Terrorism.	0	Y	NOT STARTED	Not started
Legitimacy	P1	HC	Publish understand work identifying key players, how harmful communities evolve and Facebook's role in their growth.	n/a	n/a	AT RISK	No current update due to PTO. We'll update this goal on Wednesday
Legitimacy	P1	HC	Determine policy to designate a community as harmful.	n/a	n/a	Deprioritized in DC, CORGI is driving	The harmful topic community aspect of the AHN policy is a part of the plan but still TBD on when it will be tackled. Disaggregating Harmful Network task force (CORGI) will continue to drive this with Policy. DC will step back from policy work in H1
Legitimacy	P1	HC	Build accurate detection model to detect harmful communities in Harmful Conspiracy Theory problem space, and quantify its potential impact in reducing the harm	n/a	n/a	On Track (In Progress)	- Have several iteration of the detection model on the Querderken pipeline, worked with COINV and investigators to get some pecision estimates. The model accuracy looks good with some initial data.
Legitimacy	P2	HC	Run experiment to measure the effect of mitigating the harmful community/rabbit hole effect	n/a	n/a	On Track (In Progress)	- Started work on implementing enforcement in Non-rec and feeds demotion
Legitimacy	P2	HC	[placeholder] Establish a clear H1 impact target goal for HTC at the end of Q1	n/a	n/a	BEHIND	We have gotten some help from a DS with extra bandwidth so are still hoping to conduct the analysis we had planned