(ICW) Meta Leaks Part 1: Israel & Meta, The Greatest Global Mass Censorship Campaign to Ever Exist

By NRU

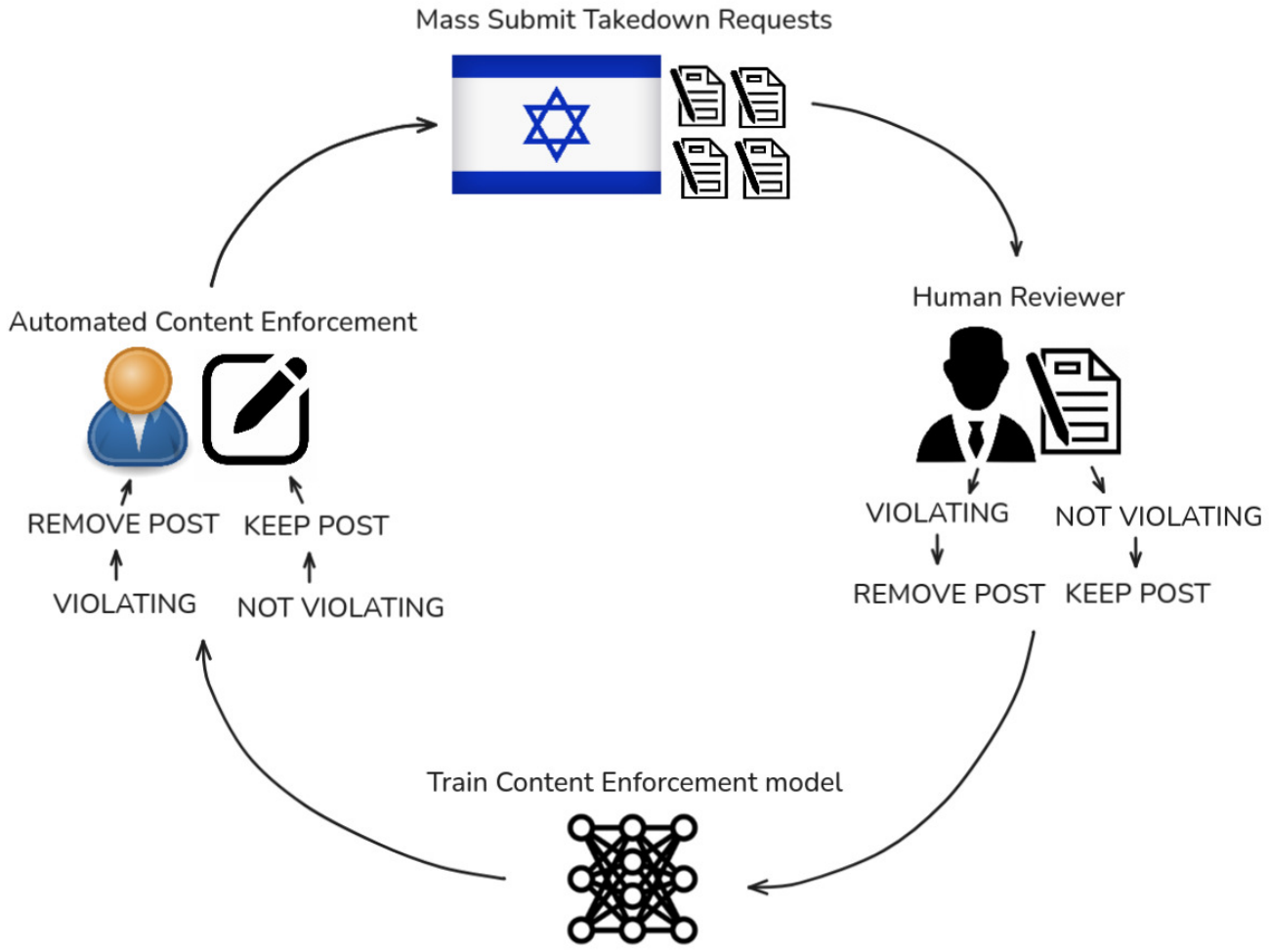


Figure 0: How Israel abuses Meta content enforcement system. The content enforcement model trains on a dataset poisoned by mass submitted takedown requests. The model is then able screen billions of posts and censor millions of users based on the content linked in the takedown requests.

International Corruption Watch

Perilous times are coming

The International Corruption Watch (ICW) is a newly founded organization of independent journalists with the goal of bring to light the massive corruption that plagues governments and big tech worldwide. In an era where technology plays an increasingly central role in shaping public policy and opinion, we believe that the misuse of technology for corrupt purposes has become one of the most pressing issues of our time.

Follow us on social media:

BlueSky: https://bsky.app/profile/icw-nru.bsky.social

Do you have any tips for a case of corruption in tech or government?

If so please send a message to **icw_nru@protonmail.com**

ICW is a self-funded operation. All donations are appreciated.

BTC: bc1qjy84xyr8u0majvtslzfz940hpkcp4hv3gjmms8

ETH: 0x0620C26611644Ec7287274713df0eD3Df93dBB82

Read our previous investigations!

(2024-11-17) The Youtube Alogirthm and Manufacturing Consent

(2024-11-30) What can we learn from the Andrew Tate data breach?

Note from NRU

The Meta leaks series aim to expose corruption inside Meta. Every piece of information and accusation is brought forth by Meta employees. The motivation behind these leaks is the following: Stop all involvement with the Israeli government and their current genocide in Gaza. Until then, more leaks will be dropped. With each leak exposing a different aspect of corruption from censorship, to AI, to financial crimes.

The contents of this article have been previously shared by another third party organization, Drop Site News. This is the full version of that story.

From the initial release from Drop Site News till now, Meta has not taken any single step in mitigating the exploitation of its censorship system.

Summary

We have obtained internal leaks from Meta employees from the companies “Integrity” organization. The results of our investigation suggest that:

- Israel is responsible for targeting over 90,000+ posts on Meta’s platform with mass fraudulent takedown requests (TDR), 95% of which are under the “terrorism” and/or “violence and incitement” category. Israel’s takedown requests have a 94% compliance rate.
- Almost all takedown requests submitted by Israel have immediately been after October 7, 2023. Takedown requests almost exclusively target posts capturing the violence in Gaza, West Bank, Lebanon, as well as targeting critics of Israel.
- Of the 90,000+ targetted posts, the top 12 countries of users targeted are from Egypt (21.1%), Jordan (16.6%), Palestine (15.6%), Algeria (8.2%), Yemen (7.5%), Tunisia (3.3%), Morocco (2.9%), Saudi Arabia (2.7%), Lebanon (2.6%), Iraq (2.6%) Syria (2%), Turkey (1.5%).
    - Its important to contextualize this with the fact that almost all governments reporting to Meta primarily censor citizens of their own countries. Israel is the exception as only 1.3% of its takedown requests are actually targeted towards Israeli’s (14th most targeted country).
    - For reference, 63% of Malaysia TDRs target Malaysian content, and 95% of Brazil’s TDRs target Brazilian content
    - Israel ranks 3rd in the most posts targeted by TDRs out of any country.
    - On a per-capita basis, Israel ranks 1st in the most posts targeted by TDRs, and has 3 times more TDR targeted posts per-capita than the country with the 2nd most submissions.
- Mass fraudulent takedown requests are likely being used as inputs to train Meta’s AI content moderation system, in what is referred to in machine learning as a data poisoning attack.
- We estimate that as a result of Israel’s data poisoning attack, roughly 38.8M+ *additional* posts on Facebook and Instagram have been actioned on by Meta in the last year and a half. Making this the largest mass censorship campaign in history.
    - We suspect the actual number of posts taken down are likely much higher as our estimate was a conservative one and only one violation category was explored, while several others such as violence and incitement were not explored in this investigation.
- The “terrorism” category at Meta seems to censor content that is primary anti-Israel related.
    - We show this by calculating the rolling average pairwise Pearson correlation on the number of actions that the content moderation system takes across time between three distinct countries: United States, India, and Palestine.
    - Prior & post October 7 2023, these values peaked at 0.75 and 0.95 respectively.
    - Our findings are corroborated by an investigation from Human Rights Watch

- Of all the ~87 million posts actioned by Meta for "terrorism" since October 7th 2023, the top 15 countries of users are from Algeria (17.2%), Pakistan (11.1%), Iraq (7.6%), India (6.3%), Bangladesh (6.1%), Egypt (5.5%), Afghanistan (4.6%) Yemen (3.2%), Saudi Arabia (3.2%), Morocco (3.1%), Jordan (2.9%), United States (2.9%), Turkey (2.7%), Iran (2.6%), Indonesia (2.4%).

This mass censorship campaign is enabled by the fact that:

- Meta uses content enforcement machine learning models to scale out content enforcement across various violation categories for billions of users a day
- These models are primarily trained on human reviewed data.
- Takedown requests can be submitted by anyone and are reviewed by humans.
- Israel's takedown requests have a 94% compliance rate.
- This allows the Israeli government to submit a few thousand fraudulent takedown requests and Meta will scale out the censorship process based on data similar to the ones linked in the takedown requests.

We also bring forth three accusations against the Israeli government and Meta:

1. Israel has engaged in the largest international mass censorship campaign in modern history using Meta's platforms.
    - Given that the overwhelming majority of TDRs for the "terrorism" category is attributed to Israel, it is likely that up to 38.8M Facebook & Instagram posts actioned on for "terrorism" since October 2023 can be attributed to Israel's influence on the content enforcement model.
2. Meta is complicit in the Israeli governments mass censorship campaign:
    - Meta been covering up Israel's mass censorship campaign as they have been aware of Israel abusing the moderation system with mass reports for the past 7 years and has done nothing about it, despite apologizing for it.
    - Assisting Israel by increasing their takedown request compliance rate 77% to 94%
3. We hypothesis several theories as to how and why this happened. With the most likely being:
    - Israeli government must have insiders at Meta's integrity organization in the form of individual contributor engineers who advised the Israeli cyberunit on how to abuse the content moderation system, as well as giving feedback on the success of the attacks.
    - Takedown requests are a legal entry-point for which governments can do censorship. However, the Israeli government is the first to abuse it to this degree and scale.

How does content enforcement work at Meta?

According to the “How technology detects violations” page of the transparency center at meta.com, Meta uses a team of thousands of “reviewers” paired with “artificial intelligence” to review billions of posts everyday. However, internally, Meta’s setup is a bit more complicated than that.

Content Enforcement Types

Meta has several content enforcement systems. Below is a general overview (not exhaustive) of the different kinds of enforcement strategies used at Meta.

- User submitted reports
    - Note: Depending on the type of user (e.g. People/Organization/Government), there are different submission and enforcement interfaces and guidelines.
- Machine learning based enforcement
- Rule based enforcement
- LLM based enforcement
- High priority selective enforcement

The focus of this report is on user submitted reports and machine learning based enforcement.

Note, that these methods can be used in conjunction of each other (e.g. combining ML and rule based methods).

User Submitted Reports

When a user is shown a post on Meta’s platform, they have the option to report it. Depending on the entity reporting (user vs government agency), the reporting method can vary.

Once a user submits a post on one of Meta’s platforms (e.g. Instagram), an AI based enforcement model checks to see whether the post violates Meta community guidelines. If the post is violating a guideline, it is automatically actioned upon (e.g. remove post, ban user).

Regular people can report a post using the standard report toggle on the top right of any post. Reporting a post from here will have it be reviewed by a content enforcement machine learning model. If the model is confident, the post will be actioned on, else it will go through the human review process (figure 1).

Governments can also report a post using non-standard methods. All governments or influential non-profit organizations have access to an internal Meta email or separate intake forms such as takedown requests forms. Reports filed by governments are high priority and almost always go through the human review process.

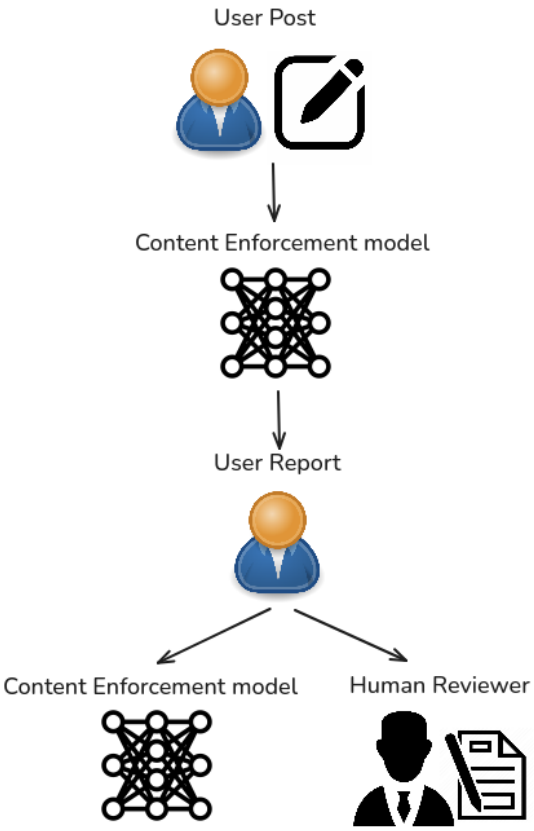


Figure 1. The content enforcement process at Meta for a user submitted report

Machine learning based enforcement

Training

In order to have an AI that can moderate content, Meta first needs to train a machine learning (ML) model. The “How enforcement technology works” page partially describes the process:

1. Review teams review a post to assess whether it violates the community guidelines.
2. Reviewer decisions will be used as a training data for Meta’s content enforcement machine learning (ML) models. In other words, the ML model learns from human reviewer data, so that it can be used for the content review and enforcement process at scale.

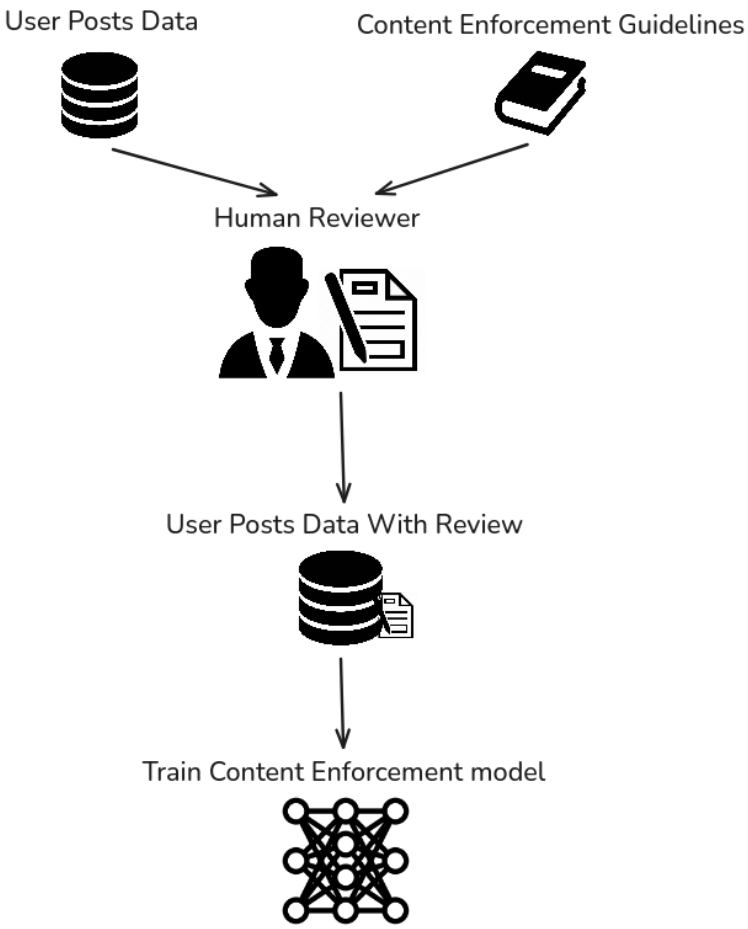


Figure 2. How data is reviewed and used for training at Meta's Integrity organization

Serving

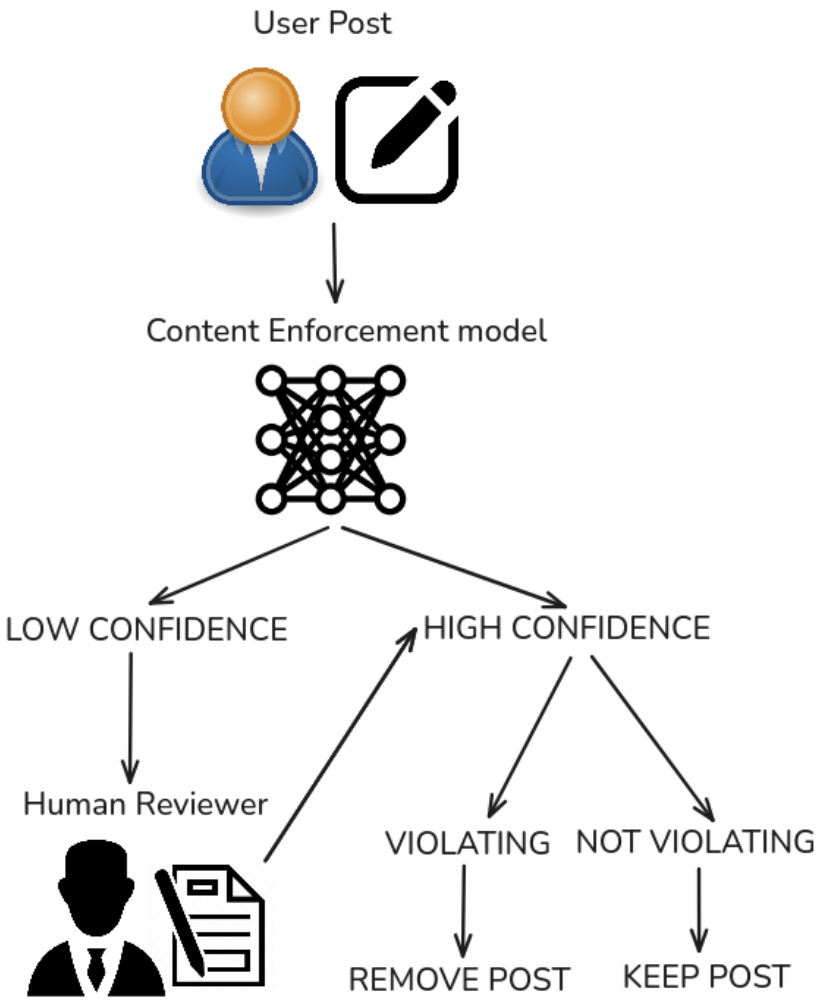
Now that the model is trained, it is ready to be used in production (live traffic on Facebook and Instagram). Meta will now use the model to classify every new post. According to Meta, 90% of the posts on its platforms are actioned upon before anyone reports it. This is done using the content enforcement ML models. The "How enforcement technology works" page partially describes the process:

- Once a post gets added to the platform, the content enforcement model will classify the post as VIOLATING or NOT VIOLATING, along with a confidence score.
- If a high confidence score is given and the post IS VIOLATING, it gets removed. If NOT VIOLATING, it will stay.
- If however a low confidence score is given, a human reviewer will review the post and provide the final enforcement decision themselves.

Figure 3. How the enforcement model is used to classify all incoming user posts on Meta's platform

This seems pretty clear and straight forward. Humans **manually reviewing posts**, having **ML models learn from human reviewers**, and then **using those models for content enforcement** that can scale to billions of posts. So what exactly is the scandal?

The scandal comes when a foreign government gets almost exclusive access to certain violation categories, issue take down requests to silence their opposition, and then have Meta scale out the censorship using AI.



How Israel Censors Millions with the Complicity of Meta

After reviewing all the internal leaks from Meta's Integrity Organization, and consolidating them with external reports, here is how Meta censored millions on behalf of Israel. To fully understand the severity of this mass censorship campaign, I will explain it in nine parts:

1. What kind of data does Meta use to train their enforcement models?
2. What are takedown requests and how do they work?
3. How does Israel weaponize the reporting system at Meta?
4. The hard numbers
5. The anatomy of a takedown request
6. Nuances
7. Attribution
8. Beyond the 'terrorism' violation category
9. Historical context

What kind of data does Meta use to train their enforcement models?

Meta stores billions of user posts in their internal databases. Which brings the question, how does Meta decide which data to use to train their enforcement models? Posts aren't inherently labeled violating or non-violating, therefore, the labels have to be generated from Meta themselves.

According to the leaks, there are dozens of violations categories at Meta. For a post to be assigned a violation category, the post must be labeled by the **enforcement model** or **human reviewer.**

The Integrity organization at Meta **prefers using human reviewer label data to train their models** as opposed to model labeled data. This is done because human reviewers are more accurate, and training on model generated labels can lead to model overfitting.

What are takedown requests and how do they work?

Takedown requests (TDR) are forms that any person, organization, or government official can fill to get a specified post taken down. These requests can be for any violations categories specified by the requester. When a regular user files a TDR, it is quite rare for a post to get taken down. However, when a government official files one, the probability of taking down a post become is likely much higher. **Almost all government submitted TDRs go through a human reviewer** for further review.

How does Israel weaponize the reporting system at Meta?

Like every other publicly available social media platform, any user can report any post for anything they want. Mass reporting has always been a thing on the internet. In 2021, Meta released a report on Adversarial Threats on its platform. The report highlights the existence of two types of threats on their platforms, brigading and mass reporting. It is not uncommon for countries like Israel to appear in these reports.

In this investigation, I will discuss **a first of its kind data poisoning attack on Meta's content enforcement models that is caused by mass reporting**. Israel weaponizes the reporting system at Meta by:

1. Mass submitting takedown requests via government accounts
2. This forces a human reviewer to review the takedown requests and likely comply due to threat of government backlash.
3. The content enforcement model trains on the human labeled takedown requests
4. All new posts are classified by the content enforcement model

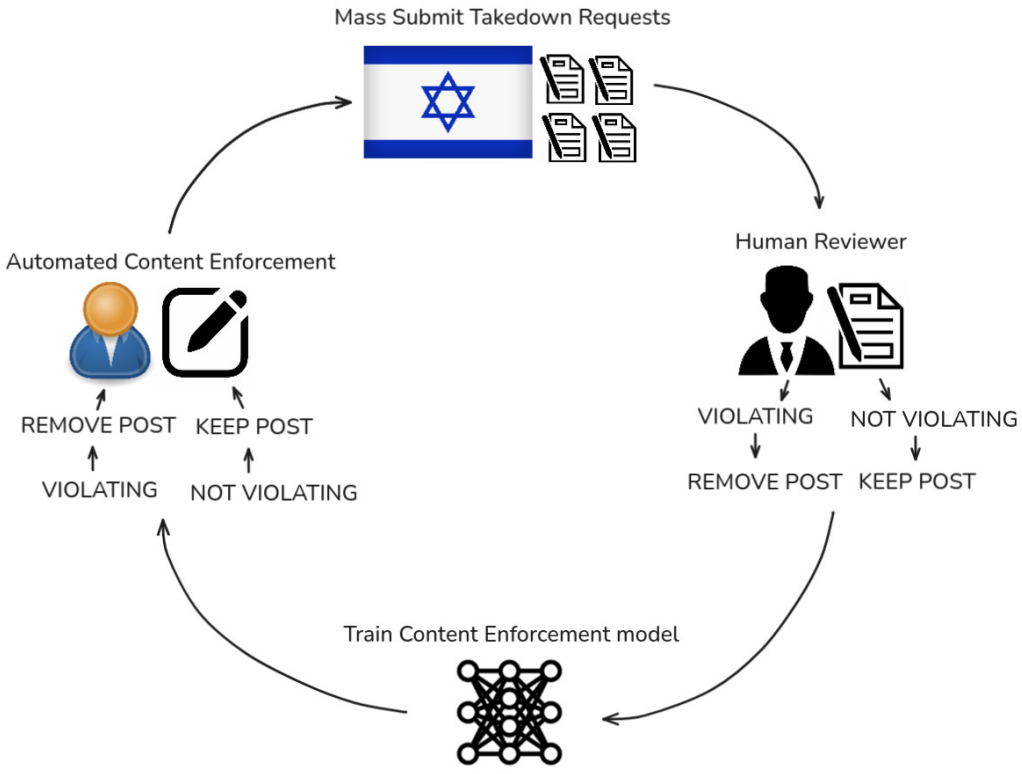


Figure 4: How Israel abuses Meta content enforcement system. The content enforcement model trains on a dataset poisoned by mass submitted takedown requests. The model is then able screen billions of posts and censor millions of users based on the content linked in the takedown requests.

The hard numbers – DISCLAIMER

1. The term “posts” refers to user posts on Meta’s platforms, this can be anything from feed posts, comments, video submissions, reels, etc...
2. Everything in “the hard numbers” sections will be insider data from Meta’s integrity organization. Internally, not all takedown requests have complete metadata. Some countries only have metadata present for 50% of their TDRs. The missing metadata may contribute to some noise on the overall numbers, however, any trends seen in the data are likely consistent with the ground truth. In some cases, it is likely that data from this section under counts certain categories! The term ‘Unknown’ will be used in cases where certain metadata fields are not available. In this investigation, data is used as is.
3. Each takedown request can have multiple posts linked to it. In the investigation below we will use the term TDR linked posts to refer to that.
4. Due to each TDR having multiple linked posts metadata may not be 100% accurate. For example a TDR aimed at the United States, may contain some linked posts from other countries such as Canada. In this investigation, data is used as is.
5. All data is rounded to the nearest 100 or 1000 so its easier to display in the table. ‘K’ means thousand.

The hard numbers – Takedown requests per country

This section contains the number of posts linked in takedown requests submitted by each countries government. For each government, I breakdown the takedown requests for the **top 5 countries targeted**. Data is shown for the countries with the top 8 most takedown requests (# linked posts).

Figure 5. Breakdown of takedown requests for each countries government and their top 5 targets

Targeted Country
Total TDR linked posts from the targeted country | % of all TDR linked posts
Top 1 enforcement action / Top 2 enforcement action (when available)
Top 1 TDR violation category / Top 2 TDR violation category (when available)

	Top 1	**Top 2**	**Top 3**	**Top 4**	**Top 5**
Pakistan	Pakistan 29K | 40.5% Block/Geoblock Unknown/Locally Illegal	India 9.7K | 13.6% Block/Geoblock Unknown/Locally Illegal	Iran 4.5K | 6.4% Disable/Block Unknown	United States 4.3K | 6.1% Block/Geoblock Unknown/Locally Illegal	Afghanistan 2.9K | 4.1% Fake/Disable Unknown/ Violating actors
Malaysia	Malaysia 72K | 63.4% Delete/Geoblock Fraud/Locally Illegal	Indonesia 10K | 9% Delete Fraud/Locally Illegal	Vietnam 7K | 6.2% Unpublish Fraud/Locally Illegal	Philippines 3.6K | 3.2% Unpublish/Delete Fraud	Cambodia 3.3K | 2.9% Unpublish Fraud/Locally Illegal
Brazil	Brazil 119K | 94.8% Geoblock/Delete Unknown/Legal TDR/Locally Illegal	United States 1.1K | 0.9% Block Unknown	Nigeria 1K | 0.8% Disable Unknown	Portugal 0.7K | 0.6% Geoblock Unknown	Argentina 0.6K | 0.5% Delete Hate
Bangladesh	Bangladesh 35K | 51.4% Block Unknown	United Kingdom 7.3K | 10.6% Block Unknown	United States 4.2K | 6.1% Block Unknown	France 2K | 2.9% Block Unknown	Saudi Arabia 2K | 2.9% Block Unknown
Israel	Egypt 20K | 21.3% Delete Unknown/ Dangerous Individuals	Jordan 16K | 16.9% Delete Unknown/ Dangerous Individuals	Palestine 15K | 15.9% Delete Unknown/ Terrorists	Algeria 8K | 8.4% Delete Unknown/Terrorist	Yemen 7.3K | 7.7% Delete Unknown/Terrorist
India	India 42K | 66.6% Block/Delete Unknown/Sexual Exploitation	United States 5.2K | 8.3% Delete/Block Unknown/Locally Illegal	Canada 4.7K | 7.4% Block/Delete Unknown/Locally Illegal	Pakistan 2.4K | 3.8% Delete/Block Unknown/Locally Illegal	United Kingdom 1.9K | 3.1% Block/Disable Unknown/Locally Illegal
Indonesia	Indonesia 57K | 80.8% Unpublish Business Integrity/Locally Illegal	Cambodia 4.2K | 6.1% Unpublish/ Geoblock Business Integrity/Unknown	Philippines 1.7K | 2.4% Unpublish Business Integrity/Locally Illegal	Malaysia 1.6K | 2.3% Unpublish Business Integrity/Locally Illegal	Thailand 0.8K | 1.2% Unpublish Business Integrity
UAE	UAE 21K | 56.9% Block/Disable Unknown/Locally Illegal	Pakistan 3.6K | 10% Disable Unknown/ Violating actor	Cameroon 3.3K | 9% Disable Unknown/ Regulated Goods	Nigeria 3K | 8.3% Disable Unknown	Vietnam 1.3K | 3.7% Block/Geoblock Unknown/Locally Illegal

The takeaway from the data above, **is that governments reporting to Meta primarily censor citizens of their own countries**. Countries targeting users from other countries usually do so to target online fraud (e.g. Malaysia) or geoblock foreign non-approved content (e.g. Pakistan).

However, in the case of Israel, only 1.3% (rank 14) of all of its TDR linked posts are actually aimed at Israeli users. Over 95% of reports submitted by Israel are for alleged "terrorism" or "violence and incitement" violations. Furthermore, within the top 5 countries that Israel is targeting, all 5 are Arab Muslim countries that Israel is either at war with (Palestine, Yemen) and/or trying to manipulate public opinion in their favor (Egypt, Jordan, Algeria).

Similar trends can also be seen beyond the top 5. Below is a table that contains takedown request for the top 30 countries targeted by Israel. Note there are roughly 95K TDR linked posts targetted by Israel.

Figure 6. Top 30 countries targeted by takedown requests submitted by the Israeli government

Rank	Country	% of all TDR linked posts
1	Egypt	21.0%
2	Jordan	16.6%
3	Palestine	15.6%
4	Algeria	8.3%
5	Yemen	7.5%
6	Tunisia	3.4%
7	Morocco	2.9%
8	Saudi Arabia	2.8%
9	Lebanon	2.6%
10	Iraq	2.6%
11	Syria	2.0%
12	Turkey	1.5%
13	Mauritania	1.4%
14	Israel	1.3%
15	Libya	1.2%
16	Argentina	0.8%
17	Qatar	0.7%
18	United States	0.7%
19	Indonesia	0.6%
20	Kuwait	0.6%
21	Germany	0.5%
22	Nigeria	0.4%
23	Pakistan	0.3%
24	Oman	0.3%
25	United Arab Emirates	0.3%
26	France	0.3%
27	Canada	0.3%
28	Sweden	0.2%
29	Iran	0.2%
30	Malaysia	0.2%

The results get more interesting once we start looking at the number of TDR targeted posts per capita of the reporting country. I divided the # of TDR linked posts by each countries population, then multiplied by 1 million to get integers.

Israel is already over-represented (3rd place) in the number of posts it targets in its TDRs. However, on a per capita basis, there is no competition. They take the number one spot for TDR linked posts per capita, 3x higher than the second most reporting country the UAE.

Figure 7. TDR linked posts per capita computation for the top reporting country governments

Country	# TDR linked posts	# TDR linked posts per capita x 1M
Pakistan	71.6K	282
Malaysia	113K	3165
Brazil	126K	592
Bangladesh	68.8K	393
Israel	95K	10034
India	62.9K	43
Indonesia	70.6K	249
UAE	36.K	3249

The hard numbers – Israel's mass reporting machine

In this section, I will be providing a time-based breakdown of Israel's mass takedown requests.

Israel has actually been sending fraudulent reports to Meta's platform since 2018. However, the data shown below will cover data from Meta's takedown portal, which started sometime in 2021-2022. Note, to avoid cluttering the data in the plots, only data from the top 10 reported countries will be shown.

The data in figure 8 shows the number of posts reported over time. There are 3 distinct time-periods in Israel's history of using Meta's takedown portal.

1. Pre August 2022
2. August 2022 to October 7 2023
3. Post October 7 2023

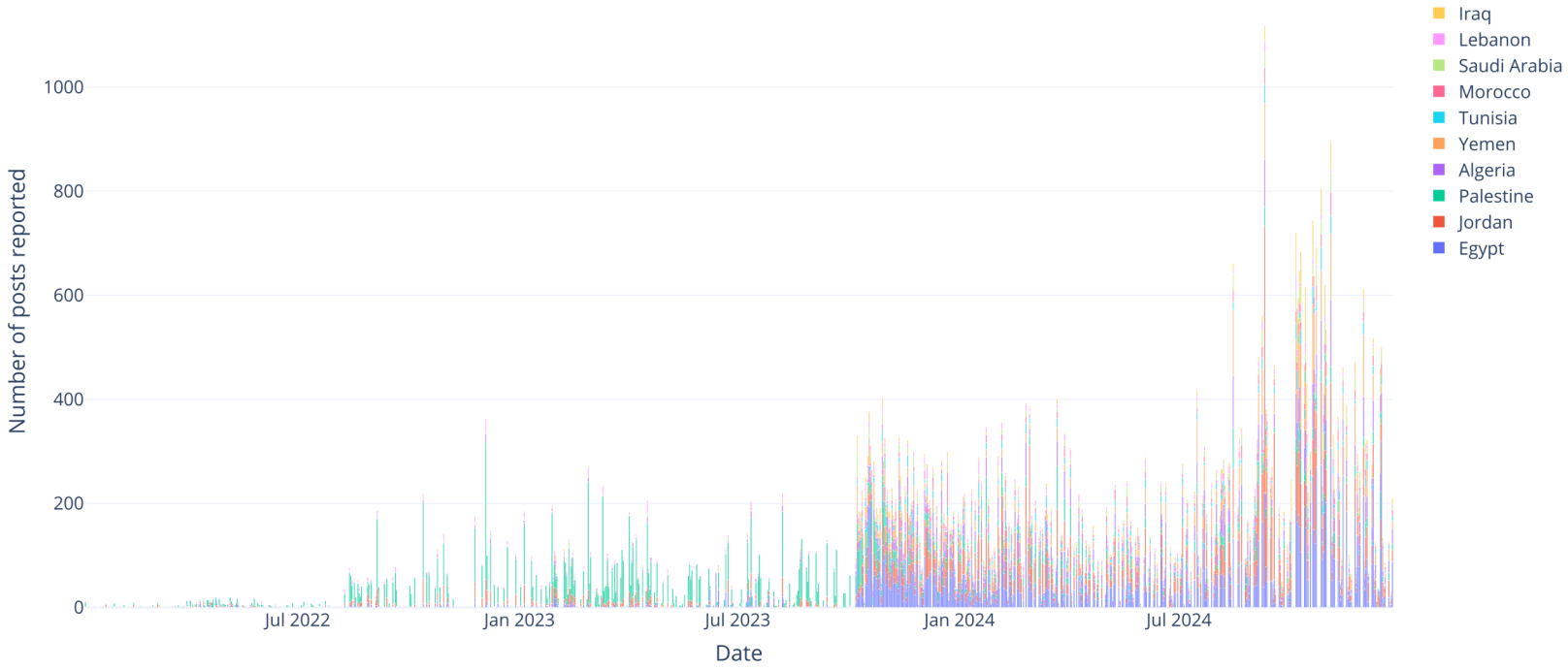


Figure 8. Number of posts reported by the Israeli government over time.

Here are my observations for figure 8:

1. Pre August 2022 there are very few Israeli reports. However, almost all are targeted against Palestinians.

2. Starting from August 2022 to October 7 2023, we can see a spike in Israeli takedown requests targeting Palestinians. These reports coincide with mass IDF raids in the Occupied West Bank, as well as several massacres in Gaza and the West Bank.

3. Post October 7 2023, we begin to see a massive and sustained spike of mass takedown requests by Israel. Note that pre October 7, most of the reports were directed at Palestinians. Post October 7, most of the reports went to Arab countries in general. Use the colors provided in the legend as reference.

To get a better understanding of the data, I created figure 9, which is a focused view of figure 8 around October 7 2023.

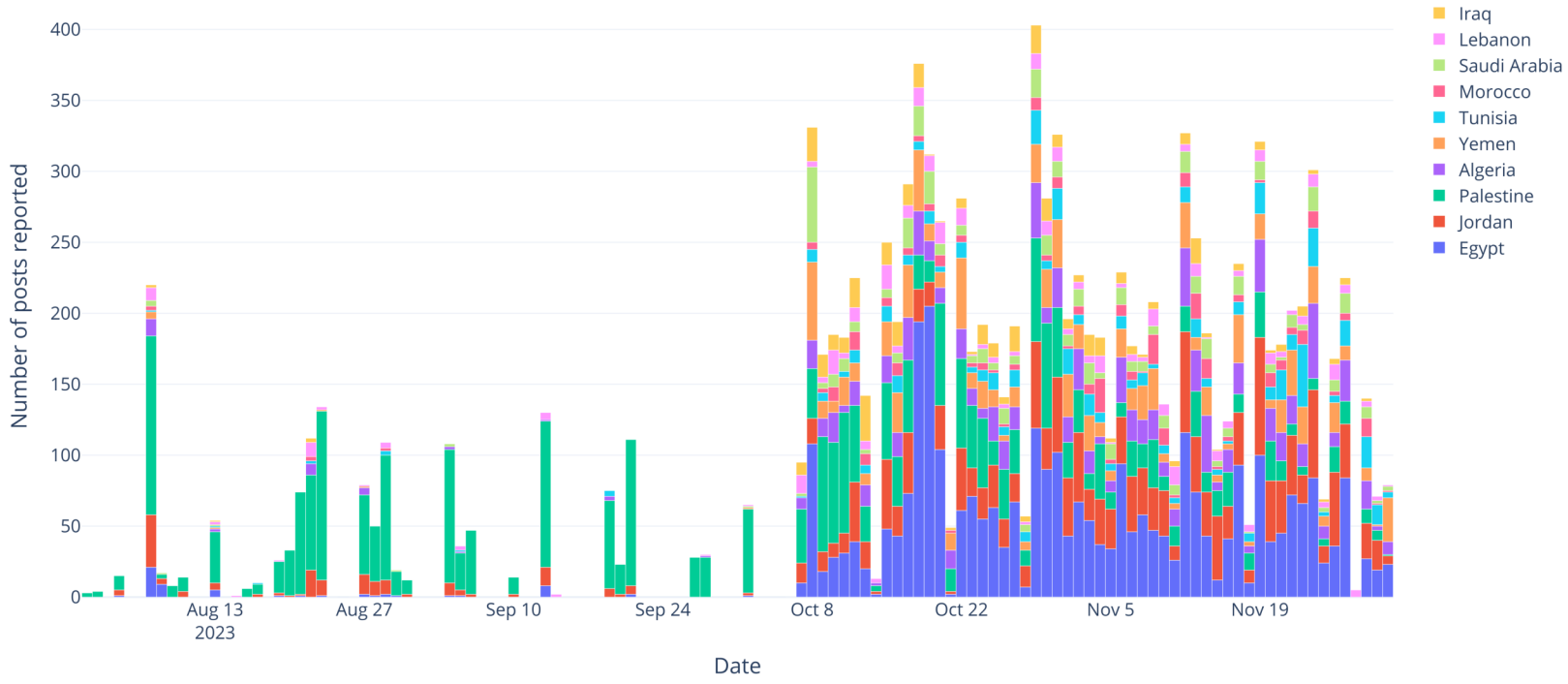


Figure 9. Number of posts reported by the Israeli government over time, centered on October 7 2023.

Here are my observations for figure 9:

1. It is interesting to see in retrospect the speed at which Israel began mass reporting on October 7. Note the massive spike that took place on the same day as the October 7 attacks. Also note the immediate shift in takedown strategy as now multiple middle eastern countries are targeted, when previously, mostly Palestinians have been targeted.

2. Throughout this reporting dataset, we see massive drops in reports on every 7th day. This of course corresponds to Shabbat or Saturday, the day which Jews refrain from work activities. However, October 7 2023 is also on Shabbat, so its interesting to see that on this day of rest, and facing an overwhelming attack:

1. Both the IDF cyberunit division and Israeli attorney's office were prepared and collaborating on that day to message Meta out of all organizations.
2. The IDF cyberunit division already had developed a new strategy of censoring countries that are not even involved in the attack like Iraq, Lebanon, Saudi Arabia, Egypt, Jordan.

The hard numbers – Meta's content enforcement machine

In this section, I will be providing an overview of Meta's content enforcement of posts that violate the "terrorism" category.

While the numbers I provided in the previous section were quite concerning, the more worrying aspect of this story is Meta's complicity. I am specifically referring Meta allowing their content enforcement models to be subject to data poisoning attacks through fraudulent mass takedown requests.

What if I told you the "terrorism" category at Meta doesn't really refer to posts from or praising terrorist groups found all over the world or even Hamas, but **largely** just refers to posts that go against the Israeli narrative? What if I told you that this has been the case (to a lesser extent) even prior to October 7 2023?

The data in figure 10 shows the number of posts actioned on by Meta over-time for violating the "terrorism" category. There are 2 distinct time-periods:

1. Pre October 7 2023
2. Post October 7 2023

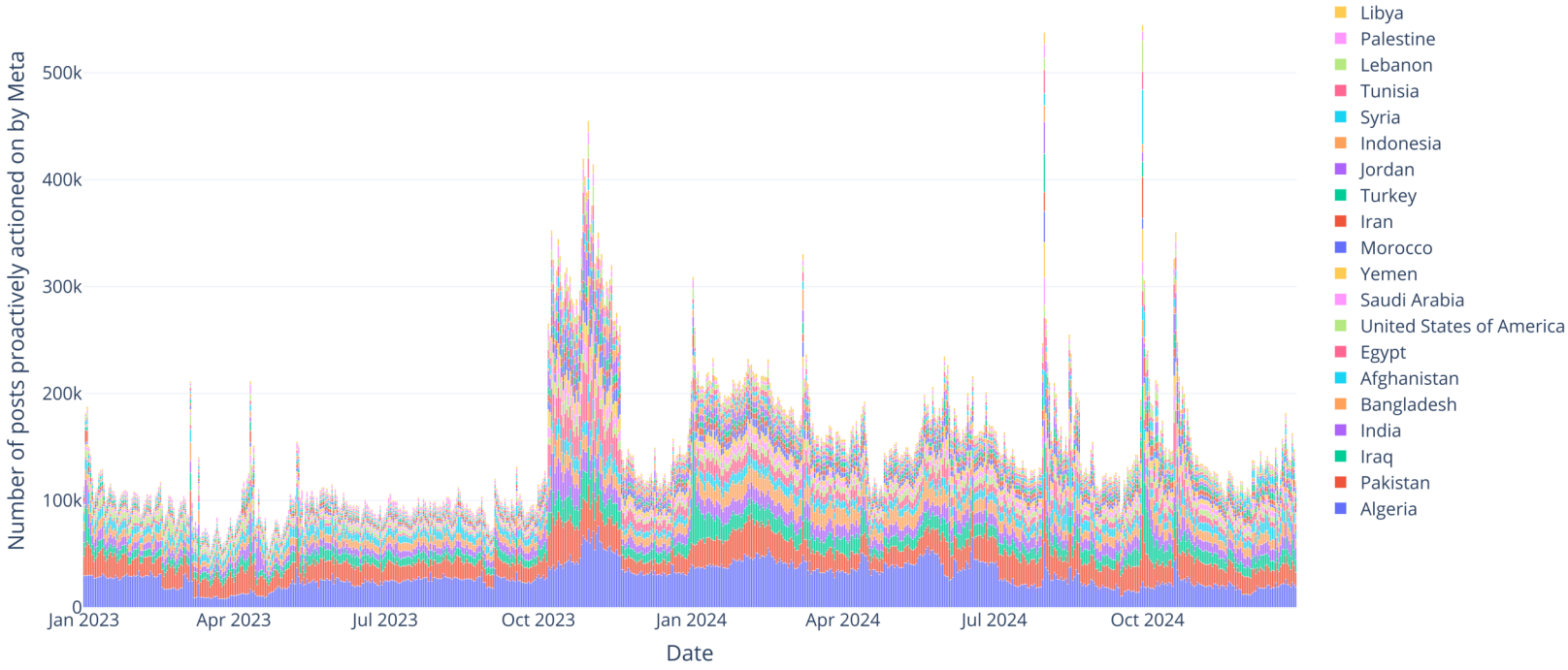


Figure 10. Number of posts actioned on by Meta over time.

Here are my observations for figure 10:

1. Pre-October 7 2023, the daily number of posts actioned upon by Meta for “terrorism” to be roughly roughly 100k per day.
2. Post-October 7, we can see anywhere from 150k to 400k posts actioned on per day. The data is also spikes multiple different throughout key times throughout Israel’s war.
3. January-April represents the bombing of Khan Younis
4. The three big spikes in July, September, and October represent the deaths of Haniyeh, Nasrallah, and Sinwar
5. The initial spikes lasted from October 7 to November 20, and represents the initial month of the conflict.
6. According to insiders, the repeated month-long waves of increases and decreases of moderation are likely due to manual threshold overrides from Meta’s integrity organization.

To show the extent of the data poisoning on Meta’s content enforcement models, I computed the normalized counts for each day in the dataset for figure 11. Notice how there is a clear break in country distribution before and after October 7 2023.

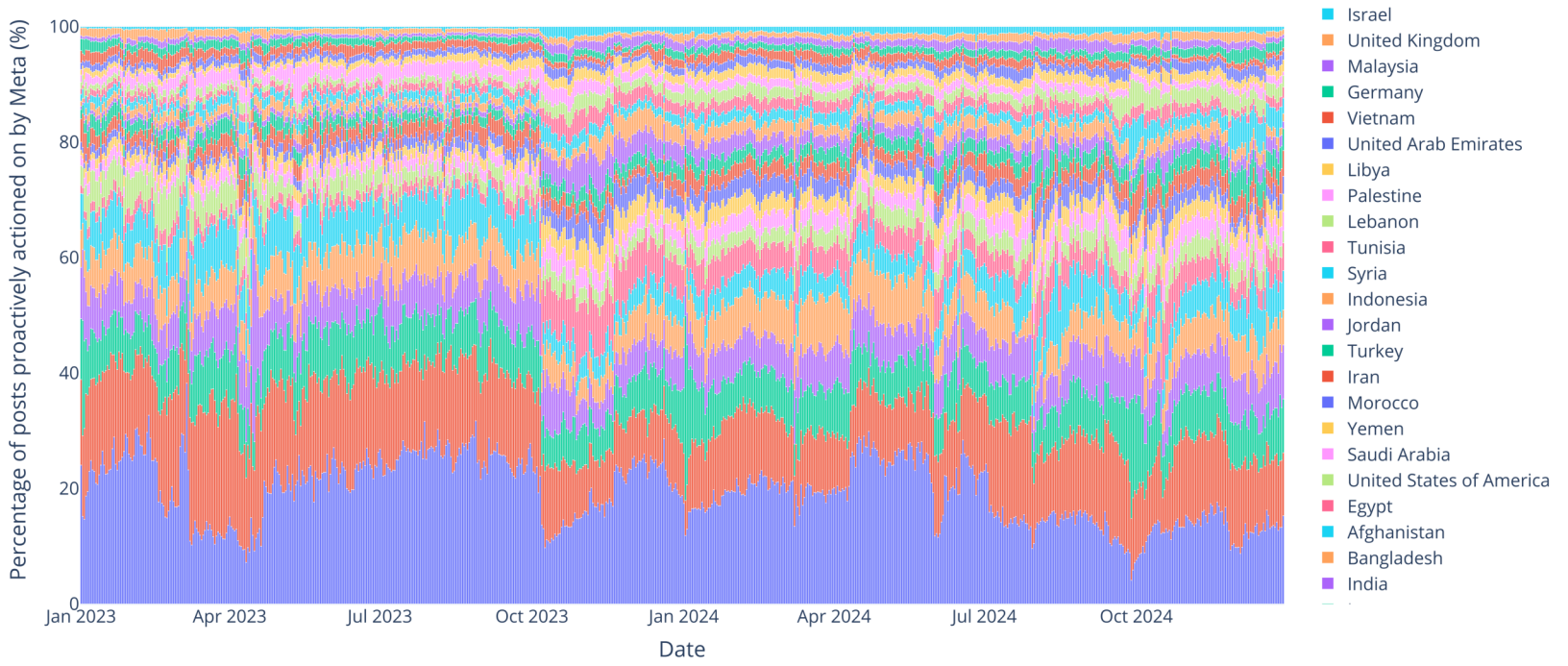


Figure 11. Number of posts (normalized) actioned on by Meta over time.

Next I computed the rolling average pairwise Pearson correlation between all 100+ countries in the dataset across time. Pearson correlation is a measure of the linear relationship between two continuous variables. The closer the value is to 1 the stronger the relationship. The closer to 0, the weaker the

relationship. The rolling average was computed in order to capture model enforcement prediction trends over time. The analysis is captured in figure 12.

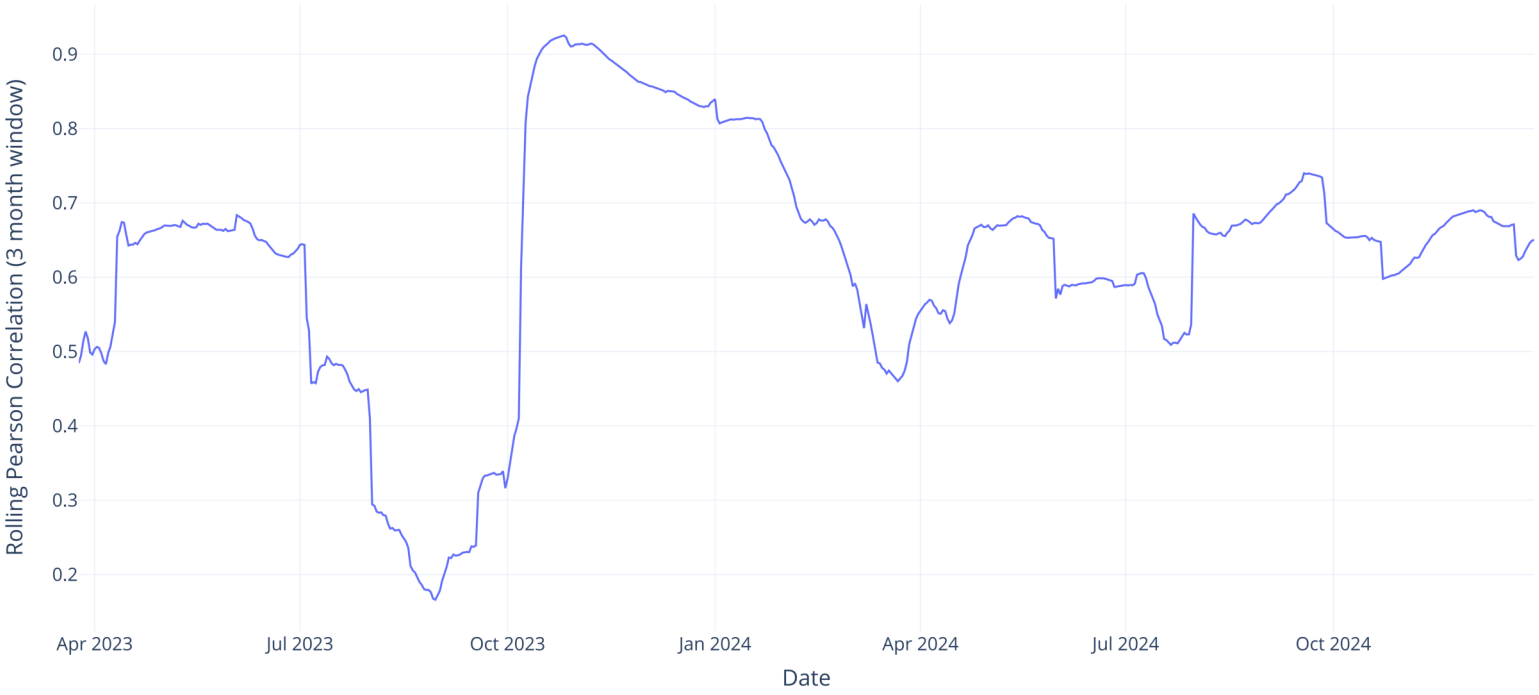


Figure 12. The rolling correlation between all countries in the dataset across time (window size is 3 months)

Given that the prediction trends changed drastically and in a sustained manner on October 7 2023 (figure 11) and there is such high prediction correlation of 0.9 (figure 12) between 100+ highly diverse countries, the most logical explanation would be model overfitting. Model overfitting happens when a model is trained poorly. In this cause, its likely because of the contaminated training dataset.

One thing that surprised me when I first looked at this dataset was the extremely high average Pearson correlation (0.67) between countries even before October 7 (figure 12). To investigate, I repeated the analysis, but only for 3 countries, each with completely different regions, cultures, languages, time zones, and conflicts: United States, India, Palestine.

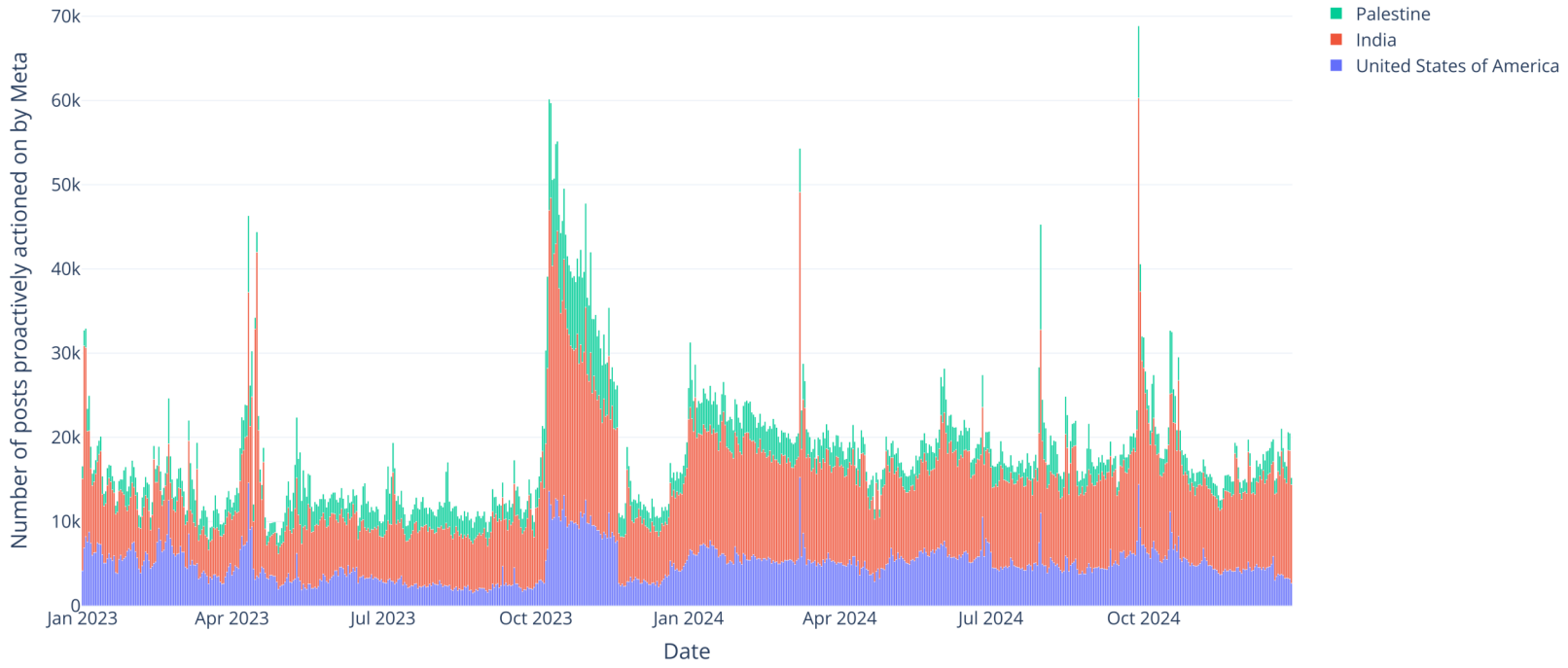


Figure 13. Number of US, India, Palestine posts actioned on by Meta over time.

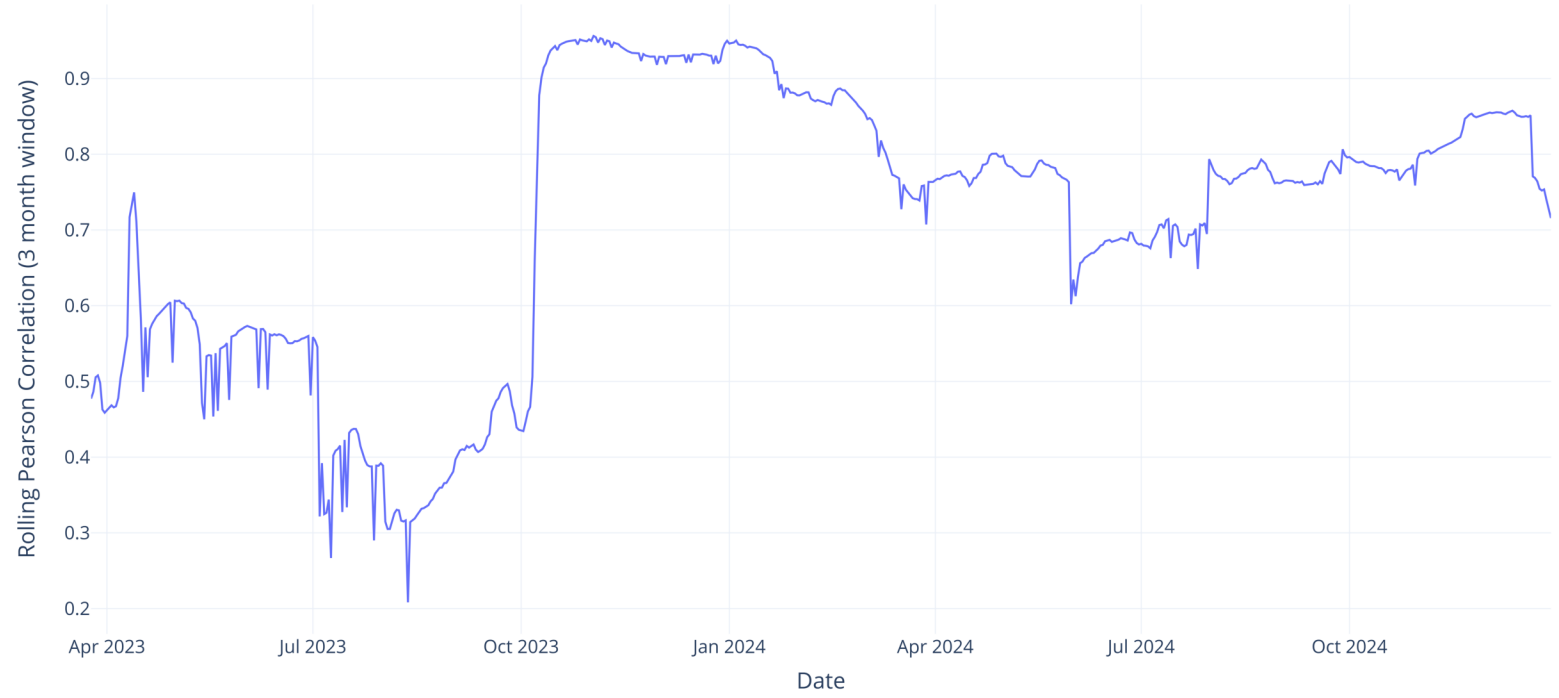


Figure 14. The rolling correlation between US, India Palestine across time (window size is 3 months)

The results from figure 13 and 14 is arguably one of the most interesting findings from this investigation. Prior and post October 7, the peak average rolling Pearson correlation values were 0.75 and 0.95 respectively. Posts from three completely different countries, regions, cultures, languages, time zones, and conflicts get actioned on at the exact same dates, for the exact same reason, and to the same magnitude. All likely related to Palestine related content.

Below is a country level breakdown of the 87M posts actioned on by Meta's content enforcement model since October 7 2023.

Figure 15. Posts actioned on per country since October 7 2023

Rank	Country	% of posts actioned
1	Algeria	17.2%
2	Pakistan	11.5%
3	Iraq	7.6%
4	India	6.3%
5	Bangladesh	6.1%
6	Egypt	5.6%
7	Afghanistan	4.7%
8	Yemen	3.3%
9	Saudi Arabia	3.2%
10	Morocco	3.1%
11	Jordan	3.0%
12	United States of America	2.9%
13	Turkey	2.7%
14	Iran	2.6%
15	Indonesia	2.4%
16	Syria	2.3%
17	Tunisia	2.2%
18	Lebanon	2.1%
19	Palestine	1.6%
20	Libya	1.6%
21	United Arab Emirates	1.5%
22	Malaysia	1.4%
23	Germany	1.3%
24	Vietnam	1.3%
25	Israel	1.2%
26	United Kingdom	1.1%

The anatomy of a takedown request

As mentioned earlier, takedown requests (TDRs) are usually sent via a specialized portal provided by Meta to governments and organizations. Most TDRs have the following format:

1. Name of government organization
2. Post(s) they they want to take down
3. Reasoning as to why the post(s) should be taken down

Lets take a look at what a takedown request from the Israeli government looks like. Note, actual post links won't be shown to protect the users.

> *Takedown Request*
>
> *Link to post 1*
>
> *Link to post 2*
>
> …
>
> *Link to post 15*
>
> *Hello, On Saturday 7.10.2023, Hamas terrorists infiltrated Israeli towns bordering the Gaza strip, and killed hundreds of Israelis, including civilians, slaughtered in the streets and in their homes. While in the towns, Hamas terrorists committed unspeakable, barbaric and horrific acts, including murdering children and elderly citizens, burning homes with entire families inside them, mutilating and desecrating bodies of those they slaughtered and more. Additionally, Hamas terrorists kidnapped tensor of civilians, including elderly civilians and young children, and are holding them hostage in the Gaza Strip. In the aftermath of the carnage, there are still Israelis whose fate is unknown. Simultaneously and to date, Hamas has fired thousands of rockets and missiles aimed at civilian population in the south and center of Israel as well as at the north of Israel, from the south of Lebanon. Moreover, since Hamas' barbaric attack, the terror organization Hezbollah has fired tens of rockets and missiles at aimed at civilian population in the north of Israel. In recent days, we have seen a worsening of the situation on the northern border, a noticeable increase in rocket fire towards Israel and attacks of Israeli territory by the terrorist organization Hezbollah. Due to the constant barrage of rockets and missiles, hundreds of thousands of citizens have been evacuated from their homes near the borders with the Gaza Strip and Lebanon.*
>
> *In total, since 7.10.2023 and to date, more than 8,400 Israelis have been injured, including hundreds of civilians, and more then 1,200 were killed. In response to the attacks, Israel has attacked military targets in the Gaza Strip and in the south of Lebanon, and has*

mobilized reserve troops and entered a state of war. This is an urgent request regarding videos posted on Facebook which contain inciting content. The file attached to this request contains link to content which violated articles 24(a) and 24(b) of the Israeli Counter-Terrorism Act (2016), which prohibits incitement to terrorism praise for acts of terrorism and identification or support of terror organizations. Moreover, several of the links violate article 2(4) of the Privacy Protection Act (1982), which prohibits publishing images in circumstances that could humiliate the person depicted, as they contain images of the killed, injured and kidnapped. Additionally, to our understanding, the content in the attached report violates Facebook's community standards.

Thank you,

The Cybercrime Department | The Israeli State Attorney's Office

Tel: +972-73-39323981 Fax: +972-2-6468009

Email: content_review@justice.gov.il

Every single post since October 7 2023 is like this. Same format, same style, same text. **Nothing about the descriptions of each takedown request changes with the exception of the which posts are linked**. Each takedown requests contains on average 15 items, and no description of any of the items is actually given in the request. In the first couple of days of the conflict the takedown requests contained a shortened version of the same text. Prior to October 7 2023, takedown requests did not contain this text or had a similarly vague one. The bit about Lebanon was appended to the existing takedown requests a couple of days later.

The nuances of data poisoning, human reviewers, AI, and scale

There are some counter-cases to be made be raised by Meta:

1. *“Our training data could not possibly be poisoned by a bad actor because the majority of our training data has to first go through human reviewers”*.
2. *“Look at these dozens of anti-Israel posts that are currently live on our platform. If our models were bad, they would have been removed. Conversely, look at these posts which clearly violate our terms and service which were removed by our models!”*.

Truth is obscured by simplifications and anecdotes, and revealed through understanding the nuances.

Any response given by Meta within the first couple of days to this investigation is likely just a deflection by their PR team. Actually understanding the small details of a complex AI system that scales to billions of users and then verifying the results from this investigation will likely require a dedicated group of senior engineers a week or two.

Data poisoning and human reviewers

In theory, having training data being primarily produced by humans reviewers should prevent malicious takedown requests from removing posts and making it inside the training data. However, this hypothesis **assumes that humans reviewing the data, the review process, and the reviewing instructions and guidelines they have been given are as objective and apolitical as possible**.

First, the takedown requests submitted by the Israeli government have a **94% compliance rate**. This number is so highly unusual in the realm of content moderation that it could be considered a **statistical anomaly**. Especially given the fact that the reporting entity is the **number one reporting country per capita** (with 3x more reports than the second place country) and has a **history of submitting fraudulent takedown requests to Meta**.

Second, we discovered that the **average review time per TDR linked post to be roughly 30 seconds**. Compared to the rest of the other countries, TDRs from the Israeli government rank in the top 75% for fastest reviews among other governments.

Third, we know that the training data is being poisoned because **many of the posts being taken down by Meta are likely not violating**. Its almost impossible to go through each post one-by-one, especially when posts are deleted. However, an investigation done by Human Right’s Watch reviewed **1050 posts flagged for dangerous entities or terrorism categories, and found that only 1 violated Meta’s guidelines**. Almost all posts contained peaceful support of Palestine, expressed in diverse ways. This will be discussed in more detail in the next section.

So why are reviewers approving so much fraudulent takedown requests from the Israeli government We can only speculate that:

- They have a quota to meet (number of posts reviewed per day)
- They are severely underpaid and exploited, so they will do as told.

- There is likely lots of internal and external pressure to approve takedown requests from government entities, especially if they come from the United States and Israel

AI and scale

Showing cherry picked examples of posts to prove that an enforcement model is “working correctly” is not a valid or honest defense in the world of machine learning. In fact, it only highlights the unpredictable nature of AI models.

You could create a model that decides to remove or keep posts based on the results of a coin flip, in other words, randomly. **If you apply this random model to billions of posts, this random model will effectively remove 50% of violating posts and ignore the remainder 50% of violating posts.** Model classification accuracy is a widely understood concept in the ML space.

The best way to understand how Meta’s content enforcement models work is by **looking at trends and model performance across billions of real user posts**. In the previous section, we explored highly similar enforcement trends across a diverse selection of countries signals that the homogeneity of the training data.

Attribution & measuring the impact of the data poisoning attacks

How much of posts actioned on by Meta can be directly attributed to the data poisoning attacks? It is impossible to determine the exact number. In fact, not even Meta knows this number because its impossible to know why an enforcement model predicted something the way it did. However, we can make a pretty good estimate.

For our estimation model, lets make the following conservative assumptions (even if they are not true):

1. Some posts actioned on after October 7 2023 have been effected by data poisoning
2. All posts actioned on prior to October 7 2023 are indeed posts violating Meta's policy
3. There isn't a significant difference in the number of posts made year over year prior and after October 7 2023

With these assumptions, we can now compute the number of excess posts actioned upon. Take the number of posts actioned on before and after October 7 2023 and compute the percent increase.

Here are results of the estimation:

1. Number of posts actioned on for terrorism from January 1 2023 to October 6 2023: 29.1M
2. Number of posts actioned on for terrorism from January 1 2024 to October 6 2024: 52.6M
3. % increase in number of posts actioned on: 80%
4. Given the posts actioned on from October 7 2023 to now, times the % increase: 38.8M posts

The conclusion is 38.8M posts have been actioned on in-excess.

This is not a perfect estimation model, but its simple enough to make sense of. This estimation is likely an under-estimate because step 2 of our estimation model doesn't count for the spike in content enforcement form October 7 2023 to December 1 2023. We also do not account for other contaminated violation categories such as violence and incitement, and dangerous organizations.

Beyond the ‘terrorism’ violation category

It is important to mention that not everything taken down by Meta will be classified under “terrorism”. Its quite common for posts to be classified under “violence and incitement” potentially poisoning other content enforcement models besides the one for “terrorism”

We do not even have to leak data for this one. It is corroborated by Meta’s own report on their community standards enforcement website.

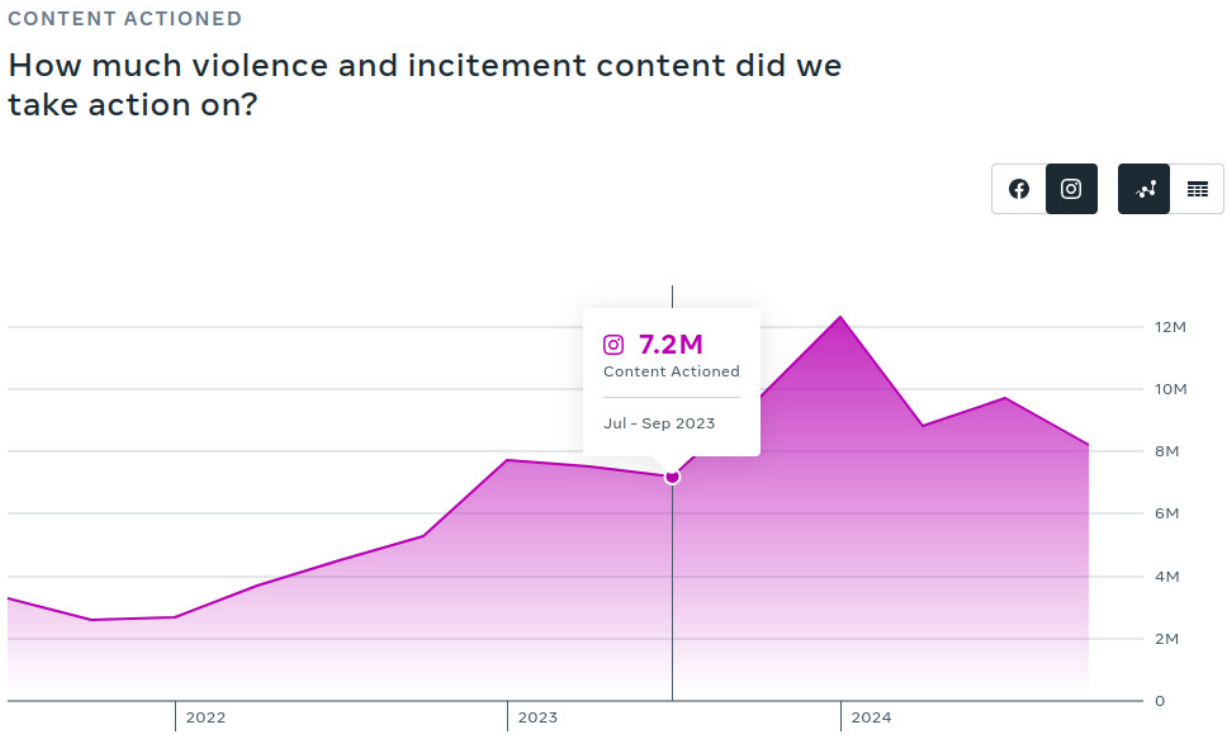


Figure 16. Number of posts actioned on by Meta for the “violence and incitement” category. Note the increase in posts after October 7 2023.

Historical context and implications

This section will go over the history of mass takedown requests at Meta with respect to the Israeli government. The purpose is to try to uncover how and why Meta allowed this to happen to their platform.

Meta is aware of mass report attacks & Israeli adversarial threats

As explained in the previous section, Meta has been actively combating adversarial threats on their platform since 2017. On their website they state:

> *"Since 2017, we've reported takedowns of more than 200 covert influence operations, cyber espionage, mass reporting and brigading networks."*

In their February 2024 report they state:

> *We removed 510 Facebook accounts, 11 Pages, one Group, and 32 accounts on Instagram for violating our policy against coordinated inauthentic behavior. This network originated in Israel and primarily targeted audiences in the United States and Canada.*
>
> …
>
> *This network's accounts posed as locals in the countries they targeted, including as Jewish students, African Americans and 'concerned' citizens. They posted primarily in English about the Israel-Hamas war, including calls for the release of hostages; praise for Israel's military actions; criticism of campus antisemitism, the United Nations Relief and Works Agency (UNRWA), and Muslims claiming that 'radical Islam' poses a threat to liberal values in Canada.*

Since Meta has experience dealing with adversarial threats, and ones of Israeli origin, there is no reason to believe that they would be unaware of the *nature* of adversarial threats shared in this investigation.

The first publicly available record of Israel attacking the content moderation system was in 2018, Meta became aware of the situation since 2021, apologized for it, then did nothing.

Just two and a half months after October 7, 2023, Human Rights Watch released a 50 page report titled Meta's Broken Promises, Systemic Censorship of Palestine Content on Instagram and Facebook.

This report highlights the mass reporting of Facebook and Instagram posts by Israeli authorities. There are several points from their report that seem to corroborate the leaked information we have investigated. In their December 21 report they state:

> *"Between October and November 2023, Human Rights Watch documented over 1,050 takedowns and other suppression of content Instagram and Facebook that had been posted by Palestinians and their supporters, including about human rights abuses. Human Rights*

Watch publicly solicited cases of any type of online censorship and of any type of viewpoints related to Israel and Palestine. Of the 1,050 cases reviewed for this report, 1,049 involved peaceful content in support of Palestine that was censored or otherwise unduly suppressed, while one case involved removal of content in support of Israel. The documented cases include content originating from over 60 countries around the world, primarily in English, all of peaceful support of Palestine, expressed in diverse ways. This distribution of cases does not necessarily reflect the overall distribution of censorship."

…

"Human Rights Watch identified six key patterns of undue censorship, each recurring at least 100 times, including 1) removal of posts, stories, and comments; 2) suspension or permanent disabling of accounts; 3) restrictions on the ability to engage with content—such as liking, commenting, sharing, and reposting on stories—for a specific period, ranging from 24 hours to three months; 4) restrictions on the ability to follow or tag other accounts; 5) restrictions on the use of certain features, such as Instagram/Facebook Live, monetization, and recommendation of accounts to non-followers; and 6) "shadow banning," the significant decrease in the visibility of an individual's posts, stories, or account, without notification, due to a reduction in the distribution or reach of content or disabling of searches for accounts."

…

"The events of May 2021 are emblematic of this dynamic. When plans by Israeli authorities to take over Palestinian homes in the Sheikh Jarrah neighborhood of occupied East Jerusalem triggered protests and escalation in violence in parts of Israel and the Occupied Palestinian Territories (OPT), people experienced heavy handed censorship when they used social media to speak out. On May 7, 2021, a group of 30 human rights and digital rights organizations denounced social media companies for "systematically silencing users protesting and documenting the evictions of Palestinian families from their homes in the neighborhood of Sheikh Jarrah in Jerusalem." In October 2021, Human Rights Watch published a report that documented Facebook's censorship of the discussion of rights issues pertaining to Israel and Palestine and warned that Meta was "silencing many people arbitrarily and without explanation, replicating online some of the same power imbalances and rights abuses that we see on the ground."

…

"In cases where removal or restrictions on content and accounts were accompanied by a notice to the user, Meta's most widely cited reasons were Community Guidelines (Instagram) or Standards (Facebook) violations, specifically those relating to "Dangerous Organizations and Individuals (DOI),[72] "adult nudity and sexual activity," "violent and graphic content," and "spam."[73] Among those violations, the most recurring policy invoked by Instagram and Facebook in the cases documented by Human Rights Watch was the "spam" policy. In reviewing these cases, Human Rights Watch found repeated instances of likely erroneous application of the "spam" policy that resulted in the censorship of Palestine-related content."

..

"While "hate speech," "bullying and harassment," and "violence and incitement" policies[74] were less commonly invoked in the cases Human Rights Watch documented, the handful of cases where they were applied stood out as erroneous."

…

"Meta committed in August[57] and October 2021[58] to increasing transparency around government requests for content removals under its Community Standards, such as those from Israel's Cyber Unit, as well as internet referral units (IRUs) in other countries. IRU requests are prone to abuse because they risk circumventing legal procedures, lack transparency and accountability, and fail to provide users with access to an effective remedy. According to media reports on November 14, Israel's Cyber Unit sent Meta and other platforms 9,500 content takedown requests since October 7, 2023, 60 percent of which went to Meta."

…

"Since Israel's State Attorney's Office began reporting on the Cyber Unit's activities, platforms' overall compliance rate with its requests has never dropped below 77 percent and in 2018 was reported to be as high as 92 percent.[129]"

...

"At the time, Facebook acknowledged several issues affecting Palestinians and their content, as well as those speaking about Palestinian matters globally,[45] some of which it attributed to "technical glitches"[46] and human error.[47] However, these issues did not explain the range of restrictions and suppression of content that Human Rights Watch observed. In a letter to Human Rights Watch, Facebook said it had already apologized for "the impact these actions have had on [Meta's] community in Palestine and on those speaking about Palestinian matters globally."

Here are the key points from the reference above:

- The earliest known records of Israel using mass reports to censor its critics is from 2018.
- Meta acknowledged Human Rights Watch's report on Israel international mass reporting censorship campaign since 2021 and apologized for their actions.
- In 2021, Israel was abusing the "spam" violation category (as opposed to the "terrorism" and "violence and incitement" that they are currently abusing)
- The compliance rate of Israel's takedown requests have increased from 77% (prior to 2018), to 92% (2018), and to 94% (2024)

Speculation

Lets try to go over all the possible reasons as to how and why Israel was able to abuse Meta's content enforcement system for the past 7 years, despite Meta being aware of mass reporting behavior since 2021. Hypotheses are ordered in order of severity:

1. Meta **is not aware** of both (1) Israeli mass TDRs and (2) their content enforcement models targeting content relating to Israel/Palestine.
    - Unlikely. Meta has acknowledged and apologized for the effects of Israeli mass TDRs
2. Meta **is aware** of both (1) Israeli mass TDRs and (2) their content enforcement models targeting content relating to Israel/Palestine; and doing nothing about it
    - Possible. One counterpoint would be that the Israeli government keeps using different methods every year, likely due to certain system exploits getting indirectly patched.
        - The first instance of known inauthentic bevhaior began in 2018
        - From then on multiple user level reports were generated. Initially on the spam category, the followed violence and incitement.
        - The mass fraudulent takedown requests again Palestinians began around 2023.
        - The mass fraudulent takedown requests again multiple countries began on October 7.
3. There are insiders from the Israeli government in the form of Mossad or former Israeli Cyberunit agents. Meta has a history of hiring individuals from these categories.
    - Possible. Insiders do not have to actively create new system exploits. Any engineer from any organization adjacent to integrity would just have to find existing exploits, and give feedback as to whether the mass TDR operation is working or not.
4. Meta has actively provided the Israeli government with a legal entry-point for carrying out its mass censorship campaign.
    - Likely. Meta has approved the American governments request to censor certain topics. However, the amount of posts censored are minuscule compared with the findings of this investigation.
    - Meta does have ties to the Israeli government by hiring someone who served as Netanyahu's previous advisor as the "Public Policy Director for Israel and the Jewish diaspora". However, despite being one of the few "DEI" roles at Meta that has not been axed, the fact that Israel has a 7 year history of freely abusing Meta's platform makes this less likely.
    - What is likely is that takedown requests are just a legal entry-point for Meta to appease governments internationally.