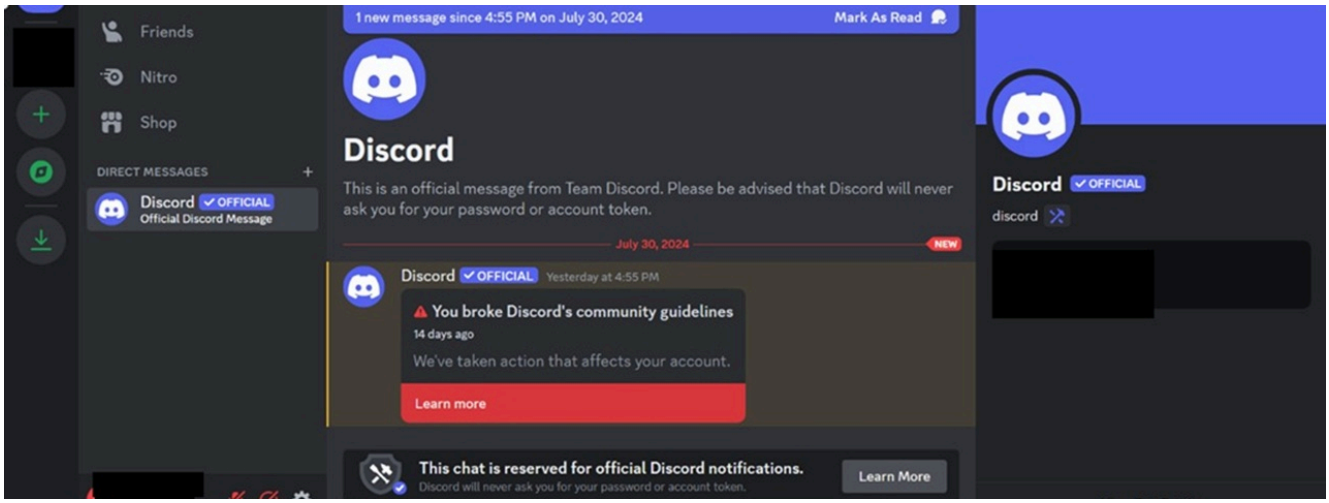
REPORT

Private Online Spaces Pose Serious Content Moderation Challenges



A Discord alert for breaking community guidelines. (Screenshot/Discord)

Published: 05.29.2025

Executive Summary

ADL has tested moderation practices on three social media platforms (Facebook, Discord, and Roblox) that allow users to create private spaces where content is visible only to members. Hate and harassment thrive in these closed online spaces, which lack the visibility and accountability of more open or public ones.

The results show that platforms must do more to detect and remove harmful content in closed groups while preserving user privacy. They must also improve tools for user reporting. Without these steps, private spaces online will continue to be exploited by extremists and bad actors to spread hate and coordinate violence.

Over the course of two weeks, ADL researchers posted 10 pieces of violative content (for example, “Jews are rats”) on private spaces and tracked the responses of three different platforms - before reporting and after.

Our testing found the following:

Facebook was the only platform that deployed any kind of auto-filtering to remove content proactively. However, this was before Facebook-owner Meta announced that it would abandon proactive moderation of non-illegal content in the U.S.

Discord did not take action on any content proactively—only after the content was reported.

Roblox’s auto-filtering did obscure some offensive posts, but it did not remove any of the content, the private group, or the user account posting violative content after we reported it.

Facebook and Discord removed most hateful content retroactively (after users reported it, in this case, by ADL), but their systems were not perfect and missed violative content.

In response to our research:

This research was shared with Meta, Discord and Roblox prior to publication and ADL researchers briefed the relevant teams at these companies. In response, Discord and Roblox acted. Specifically, Roblox shared that they “enabled an additional layer of automated review on Community Wall Post Abuse Reports. This means that abuse reports filed against Community wall posts will first undergo a review by an automated system, which will automatically take action if the wall post violates our policies. The additional layer of automation will empower human moderators to focus on deeper tasks and help increase the effectiveness of our review system.”

Private Online Spaces: Harboring Hate and Extremism

The problem of online hate and harassment has been well documented, but most data come from public online spaces such as X/Twitter, YouTube, or Reddit. These platforms typically use proactive moderation, such as automated filters and machine-learning tools, to identify and remove content that violates their rules.

Some platforms, however, allow users to create closed or private spaces where only members can see the content. These private spaces include closed Facebook groups, private Discord servers (a type of chat room), and communities on Roblox. Encrypted chat platforms such as WhatsApp, Telegram, and Signal also allow member-only spaces, some of which can be discovered publicly, while others are by invitation only. ADL's 2024 annual survey of online hate and harassment found that between 2023 and 2024, incidences of harassment increased on both Telegram and WhatsApp, +6% and +11% respectively.

Private online spaces have been linked to numerous instances of online and in-person atrocities. Shooters such as the Buffalo shooter, or a school shooter in Perry High, Iowa, posted their manifestos and livestreamed their deeds on private Discord servers. In August 2024, a perpetrator in Turkey, motivated by white supremacists and accelerationism, shared his manifesto on Telegram before livestreaming a violent stabbing attack. Most recently, online bullies have used private spaces to promote extreme cruelty and self-harm as part of larger campaigns to harm vulnerable young people, sometimes leading to suicide.

ADL has also tracked many extremist groups that flourish on private groups and servers hosted by Discord, Telegram, and other platforms. Hate groups and conspiracy mongers, such as the Three Percenters or QAnon adherents, take advantage of private groups to organize and distribute hateful content and false narratives. Other extremist groups, such as extremist traditional Catholic and Islamic groups, have used Discord servers to share memes and hateful rhetoric. Extremist militias have been quietly

organizing and regrouping on private Facebook groups since the January 6, 2021, attack on the U.S. Capitol. Individuals and groups on Roblox groom children to extremist ideologies through projects like “Redpill the Youth.” When neighborhood watch groups or community organizers use private groups, hateful users have, in some cases, infiltrated these groups to harass and intimidate other users.

It is difficult for researchers to assess how well platforms moderate these spaces or understand their inner workings because private online spaces restrict access. Private spaces can equally serve important social functions, such as for parent groups, political organizing, or protected spaces for marginalized groups to gather and provide mutual support. But tech companies do not make clear how they enforce their policies in private groups, often leaving content moderation up to users or administrators.

Understanding Content Moderation in Private Online Spaces

Members-only groups can be visible to non-members, such as many closed Facebook groups. But others, such as private Discord servers, are not making the groups' content invisible to non-members. Degrees of privacy and visibility can vary on Facebook groups, Telegram channels, or WhatsApp chats. Facebook groups, for example, can be fully public and open, publicly accessible but closed (anyone can join but only members can post or view content), or private (members must be approved). There are also two categories of private groups: closed groups that require approval but can be found through search, and fully secret groups that are not only private but cannot be found in search.

The three platforms we tested for this study (Facebook, Discord, and Roblox) have taken proactive steps to address hate, extremism, and antisemitism. We did not include platforms such as Telegram, which does not moderate hateful content, even though extremists use the platform's private channels to spread hate and coordinate harassment.

Facebook is one of the most popular social media platforms among users in the U.S., and where the majority of online harassment takes place (61% of people who are harassed report harassment on Facebook, according to ADL's 2024 annual survey of online hate and harassment). Much of this harassment takes place in closed Facebook groups, but researchers have little insight into these spaces.

Discord began as a chat platform for gamers but has since expanded to host numerous online communities. Social interaction on Discord takes place in chatrooms (called servers) that offer a private or semi-private environment for users to chat both in groups and individually (similar to Slack). Discord allows users to create servers with various channels that can be set to invite-only, making it a popular choice for groups seeking privacy. Discord's company culture, which emphasizes privacy and minimal moderation, makes it an attractive place for bad actors to connect and coordinate. Video calls and chats are not recorded or monitored, and users can delete posts permanently. Making video and audio calls end-to-end encrypted increases privacy for users, but it also makes it significantly more difficult for the platform to catch bad behavior. Discord's hands-off approach to content moderation--leaving most moderation decisions in the hands of the server's administrators-- makes it easy to share extremist content there compared to similar platforms.

Roblox is a gaming and game creation platform popular with children, with at least 45% of users being under the age of 13. Roblox has a groups feature embedded within its website, though Roblox users primarily communicate with each other through in-game text or voice chat. Within Roblox groups, members can post messages on a group wall, participate in group events, and access exclusive in-game content provided by group leaders. Roblox groups often focus on themes, such as game development or in-game collaborations. It is the only gaming platform we included where the majority of communication takes place in-game. For comparison, we only examined the group feature on the Roblox website and not communication in Roblox games.

Methodology

1. We created one group per platform (Facebook, Discord and Roblox) simultaneously, and named the groups on Facebook and Discord “Group for the Truth;” (this name was already taken on Roblox so we named that one “Group for the Truth_2.”) To prevent users from inadvertently stumbling across the groups, the settings for each were set up in a manner that made them hidden from search.

2. We then used one account to create the groups and post content (the “posting account”), and another on a separate computer to report the content (the “reporting account”). Two researchers were assigned to each account. There were no further descriptions or images added to the groups, except on Roblox, which required a group description and image.

3. The posting account was then used to post hateful content to all three platforms. This content was pulled directly from Meta’s helpfully illustrative examples of violations against its hate speech policy at the time. Meta is the only platform with examples of what could be violative of its hate speech policy of the three platforms we reviewed.

4. Meta's Tier 1 Hate Speech policy, for example, states that a user may not post content that includes:

Dehumanizing speech in the form of comparisons to or generalizations about:

- Animals in general or specific types of animals that are culturally perceived as inferior (including but not limited to: Black people and apes or ape-like creatures; Jewish people and rats; Muslim people and pigs; Mexican people and worms)

We adapted this example to draft a hateful post stating, “Jews are rats.” Every post was derived directly from Meta’s policy examples, though there may not always be a corollary example within the policy. For example, Meta’s policy at the time stated that users may not post:

Content targeting a person or group of people on the basis of their protected characteristic(s) (in written or visual form):

- Generalizations that state inferiority in the following ways:
  - Physical appearance, including but not limited to: ugly, hideous.

There was no further example provided by Meta, so we created a post that would intentionally violate this policy:

“Jews are the ugliest, most hideous people on Earth.”

Although these posts were based on Meta’s policy, each one has a matching Discord policy. The same applies to Roblox, though its hate speech policy is not nearly as detailed as Meta’s or Discord’s:

Roblox honors and welcomes users of all ages, backgrounds, and identities. We do not allow content or behavior that supports, glorifies, or promotes hate groups, their ideologies, or actions. You also may not discriminate, mock, or promote hatred against individuals or groups, or encourage others to do so directly or indirectly, on the basis of their:

- Race, perceived race, or ethnicity

We believe that the following posts are violative under Roblox’s policy as well:

1. “Jews are rats.”
2. “Jews are devils and shit.”
3. “The Holocaust did not happen-it’s a lie made up by the Jews.”
4. “Jews control the banks, the government and Hollywood and we need to get rid of them for the safety of American citizens.”
5. “Jews are the ugliest, most hideous people on Earth.”

6. “Jews should be segregated from all other people.”

5. We followed up our initial posts with additional sentences to test content violating Meta’s policy against dangerous organizations and individuals, such as violent extremists. These posts also violate Discord’s Violent Extremism policy and Roblox’s Community Standards.

1. “Hitler and the Nazis did nothing wrong.”
2. “Please donate to the Goyim Defense League!” (A link was attached to this sentence.)

We referenced the Goyim Defense League because they are a notorious group of antisemitic provocateurs, according to the ADL Center on Extremism.

We also posted the following two images in violation of Meta’s Dangerous Organizations and Individuals policy, Discord’s Violent Extremism policy, and Roblox’s Community Standards:

Image 1: Posted Images, "Covid Agenda" and "Ukraine-Russia War," respectively. (Screenshot: 7/31/2024)

6. The following day, we reported each of these posts, if applicable. Though both "Jews are the ugliest, most hideous people on Earth" and "Jews should be segregated from all other people" run afoul of hate speech policies, we did not report them until the next day. We staggered reporting in this way so as to not get either our accounts or groups suspended before the study was over.

The Challenge of Content Moderation in Private Online Spaces

Timeline of enforcement testing experiment in private spaces on Facebook, Discord and Roblox

Posted the following comments on all three platforms:

- "Jews are rats"
- "Jews are devils and shit"
- "The Holocaust did not happen-it's a lie made up by the Jews"
- "Jews control the banks, the government and Hollywood and we need to get rid of them for the safety of American citizens"

Posted the following comments on all three platforms and images*

- "Hitler and the Nazis did nothing wrong"
- "Please donate to the Goyim Defense League!" (a link

To do this experiment, ADL researchers set up new private spaces on Facebook, Roblox and Discord. Each space had two users- a posting account and a reporting account. To prevent users from inadvertently stumbling across the spaces, the settings for each were set up in a manner that made them hidden from search. The above is a description of the experiment activity in these private spaces.

Findings

- **Facebook was the only platform to remove content before we reported it.**
- **Discord did not take down any content at all, implying either that it has little to no automatic filtering or that any filtering is not robust enough to catch what our posting account shared. When offending posts were reported, Discord did remove posts and suspended the account.**
- **Roblox obscured the text of offensive posts, but did not take down the group nor the offending account.**

When we started reporting the potentially violative content to the platforms, we noticed a discrepancy between the platforms' responses. Facebook's reaction to reports included warnings before suspending the group. Discord provided less information to the administrator about why the posts were removed. It did not suspend the group but did suspend our Posting account. Roblox informed the posting account that the account or group was breaking the rules, but did not take down any posts within the group. Although it did obscure the text by censoring the entire post, replacing all content with x's except for the obscured content, the group stayed the same as it did before reporting. No further content was taken down, and no alerts were given to either account.

Facebook

Facebook removed most offending posts when we reported them, gave feedback to the account administrator, and finally suspended the group when offending content continued to be posted. Some posts remain up-- despite notifications from our Reporting account.

Proactive Detection

Facebook did not remove any posts automatically on the first day we posted content. On the second day we posted, they removed "Ukraine-Russia War" automatically (see Image 2).

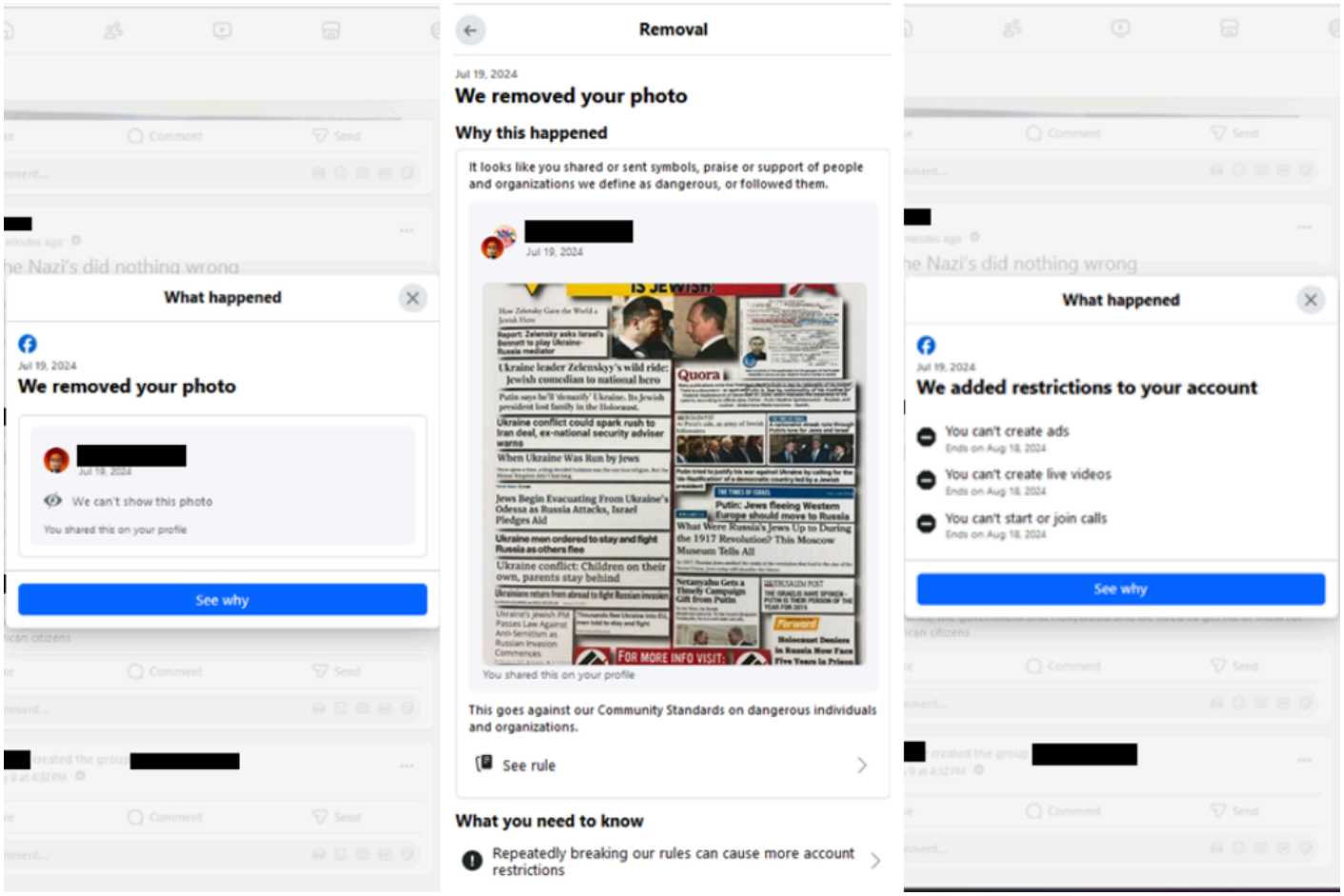


Image 2: Facebook's alert after removing "Ukraine-Russia War." (Screenshot: 7/19/2024)

Not only did Facebook remove the photo, but it also alerted the Posting account (the account admin) that the photo had been removed. When we clicked through "See why," Facebook displayed the offending image and explained why the image had been removed. It also provided a link to the rule that was broken. Facebook also added restrictions to the account that posted the picture; the user could not create ads, create live videos, or start or join calls for approximately a month. It is not clear why Facebook took action against "Ukraine-Russia War" automatically, but not the sister image "Covid Agenda." Both pictures originate from the Goyim Defense League and are prohibited by Meta's policy rules. Meta has rolled back its Covid misinformation rules, but this image is still prohibited per their Dangerous Organizations and Individuals policy.

Facebook did not allow the posting account to post the link embedded in "Please donate to the Goyim Defense League!" Not only was Facebook the only platform to

prevent adding the link, it also explained why it was not allowed (Image 3).

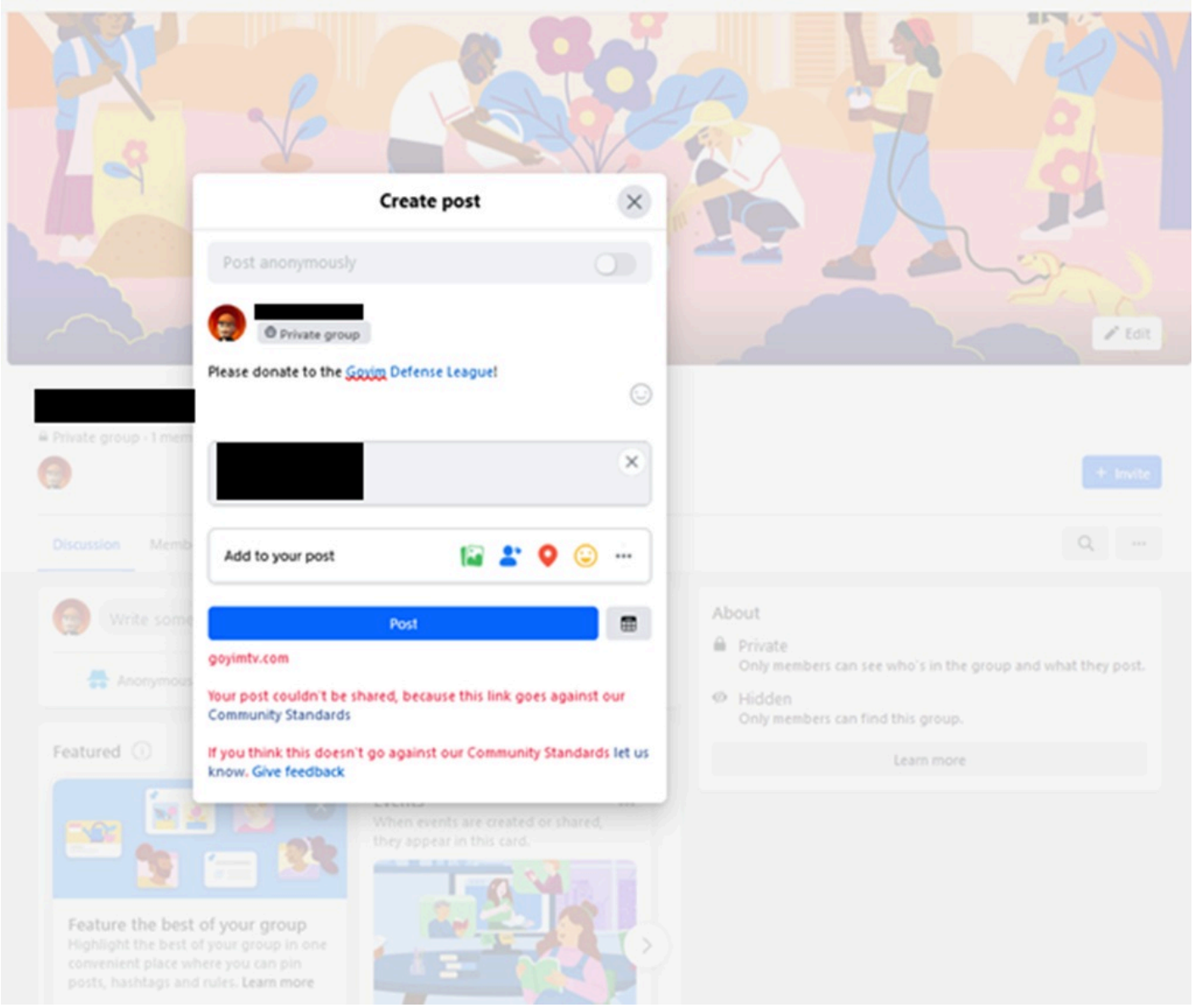
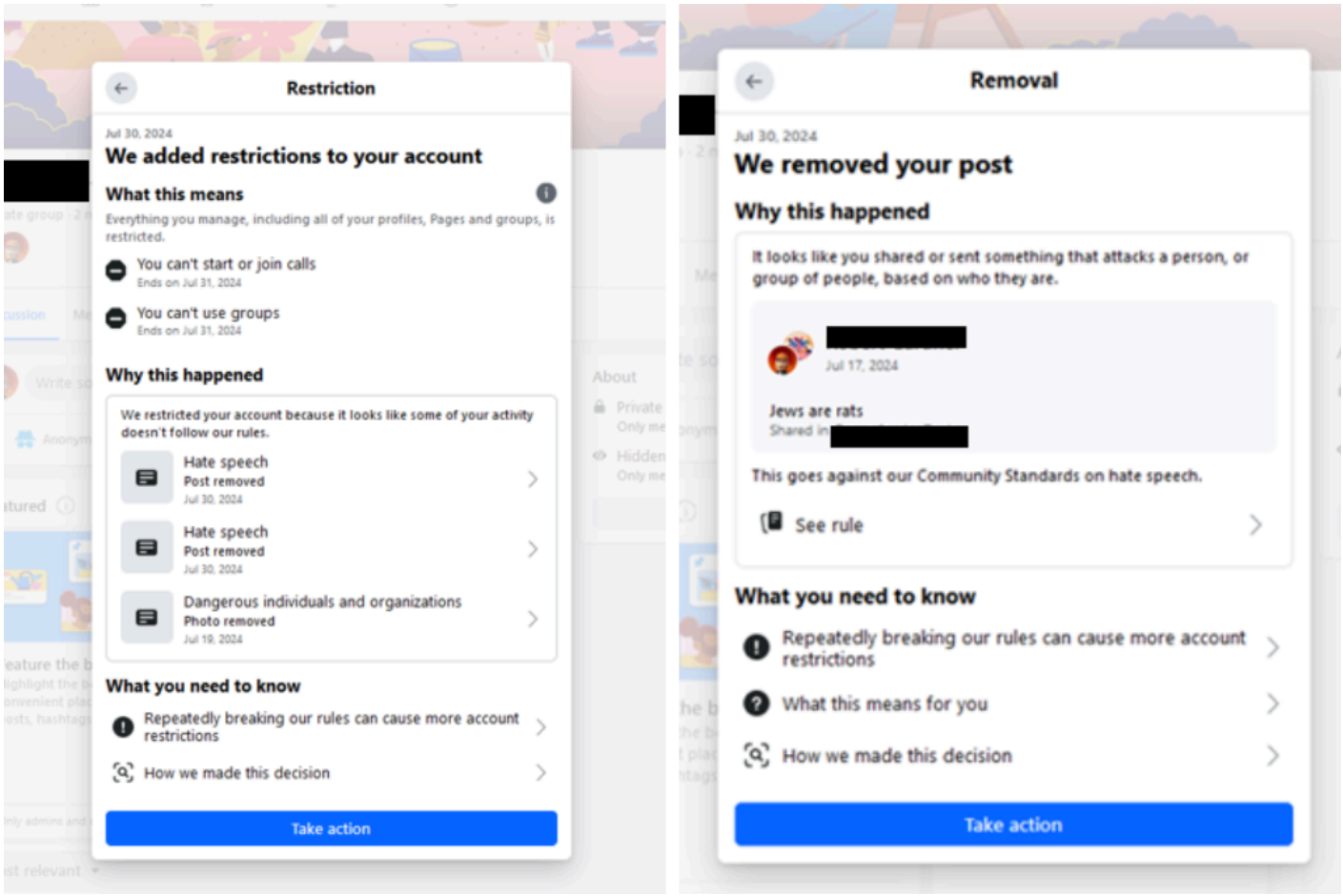


Image 3: Window preventing the user from posting a link to GDL's site. (Screenshot: 7/31/2024)

Reporting Response

After we reported the posts, they were almost instantaneously actioned. The first post was reported at 4:21pm. By 4:24 pm, the reporting account was alerted that the report resulted in the post being removed.

At the same time, the posting account was notified that restrictions were added to the account (Images 4 and 5). Unlike Discord or Roblox's restrictions, Facebook not only told the user which restrictions were added, but also explained which post was removed and which rule each had broken.



Images 4 (left) and 5 (right): Facebook account restrictions. (Screenshot: 7//30/2024)

The next day, we reported the second half of the posts and the remaining image. Facebook again took action, quickly taking down the posts. This time, however, the platform also permanently suspended the group. The posting account can still access the defunct group, but the reporting account cannot. Users can still log into both accounts, and post and comment on Facebook, just not the group. Facebook warned us that the posting account is under threat of being permanently suspended (Image 6).

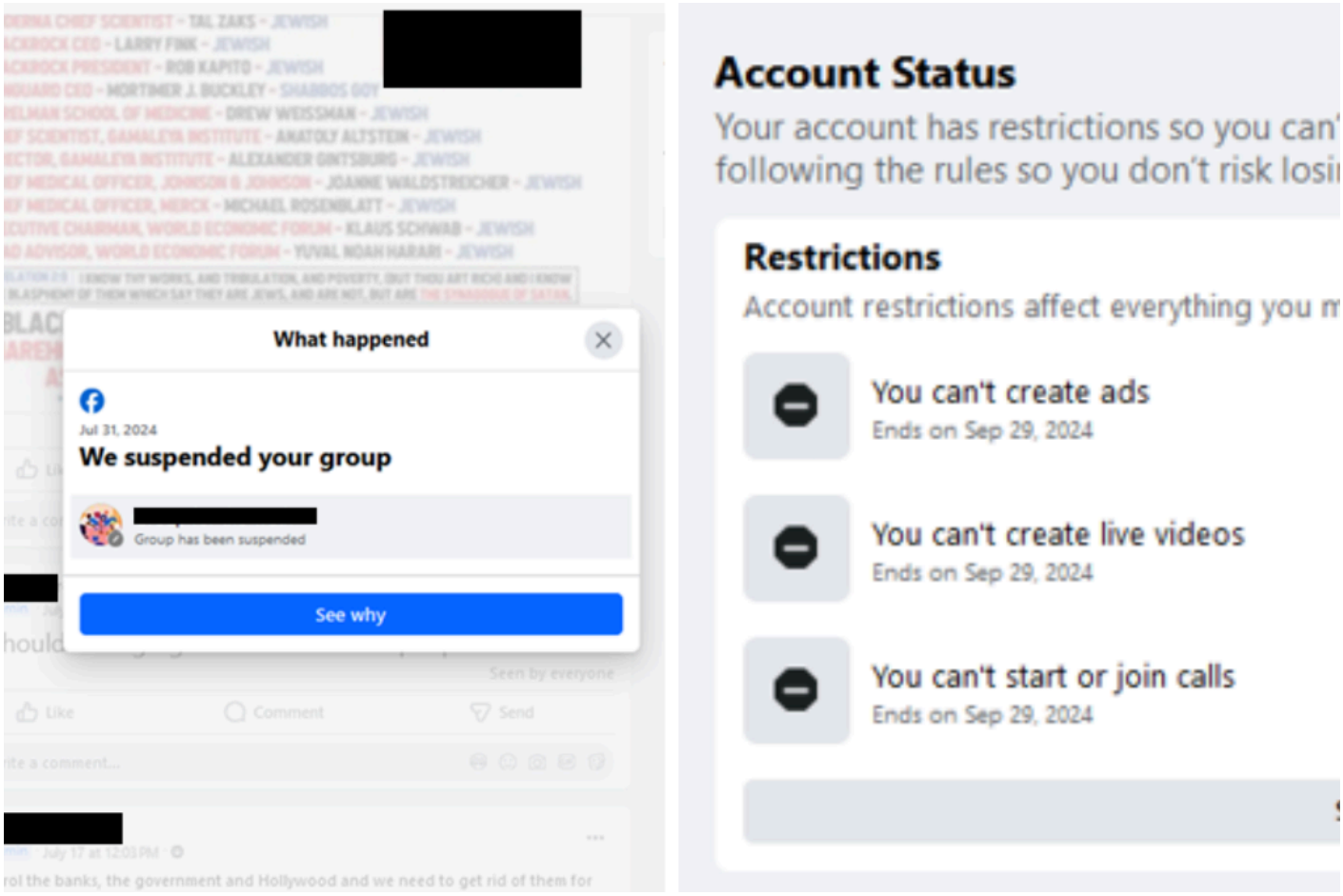


Image 6: Group suspension and account status. (Screenshot: 07/31/2024)

Although we commend Facebook for their quick action, there were still posts that remain up. These posts, despite the reports, were never taken down as of publication (Image 7).

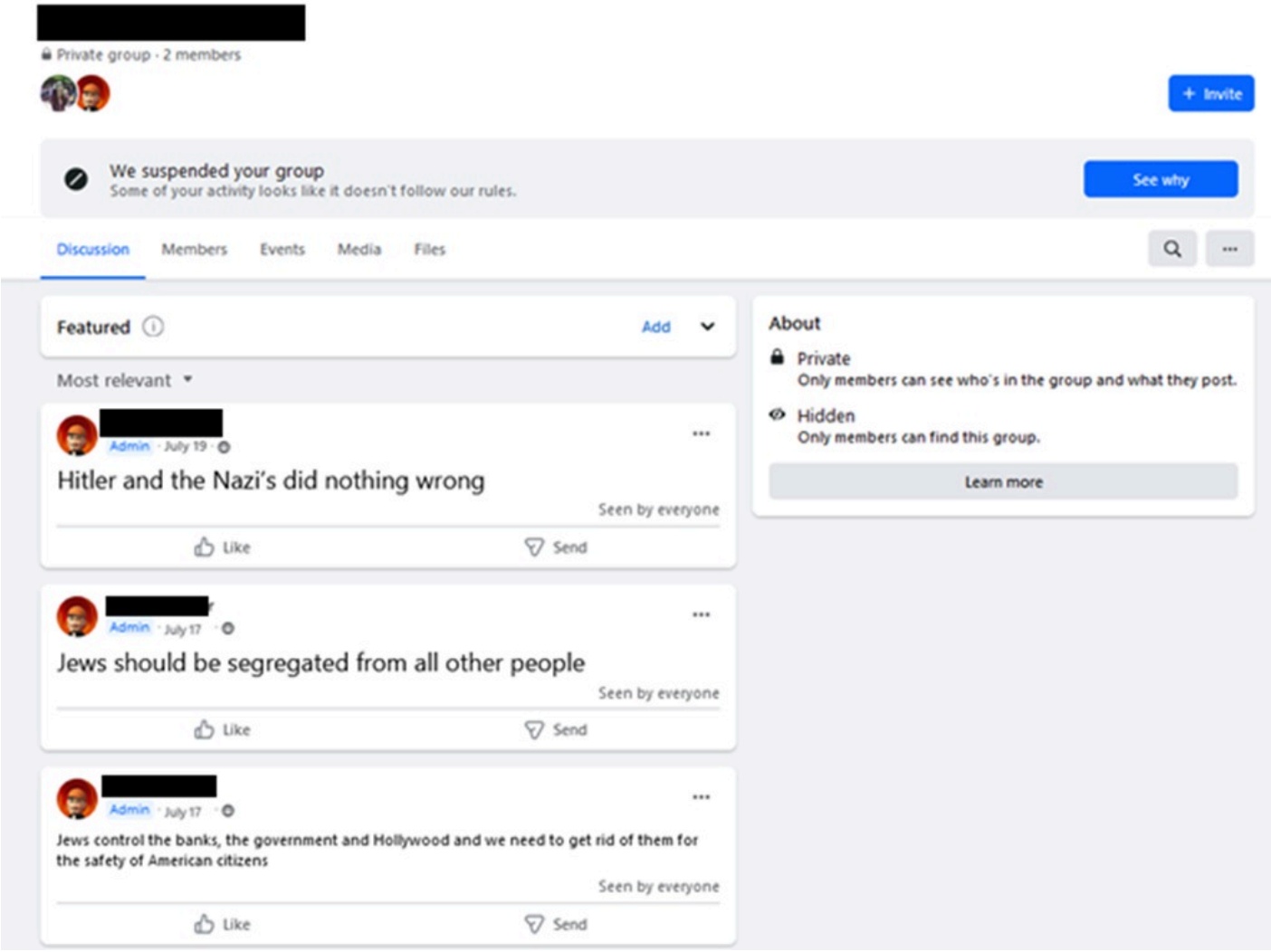


Image 7: Facebook group. (Screenshot: 2/28/2025)

Discord

In contrast to Facebook, Discord's content moderation strategy for private groups, appears to be reactive rather than proactive. It depends primarily on reports from users of those groups, rather than automatic filtering and takedowns.

Proactive detection

Until we reported the offending content on the Discord group, no content was taken down.

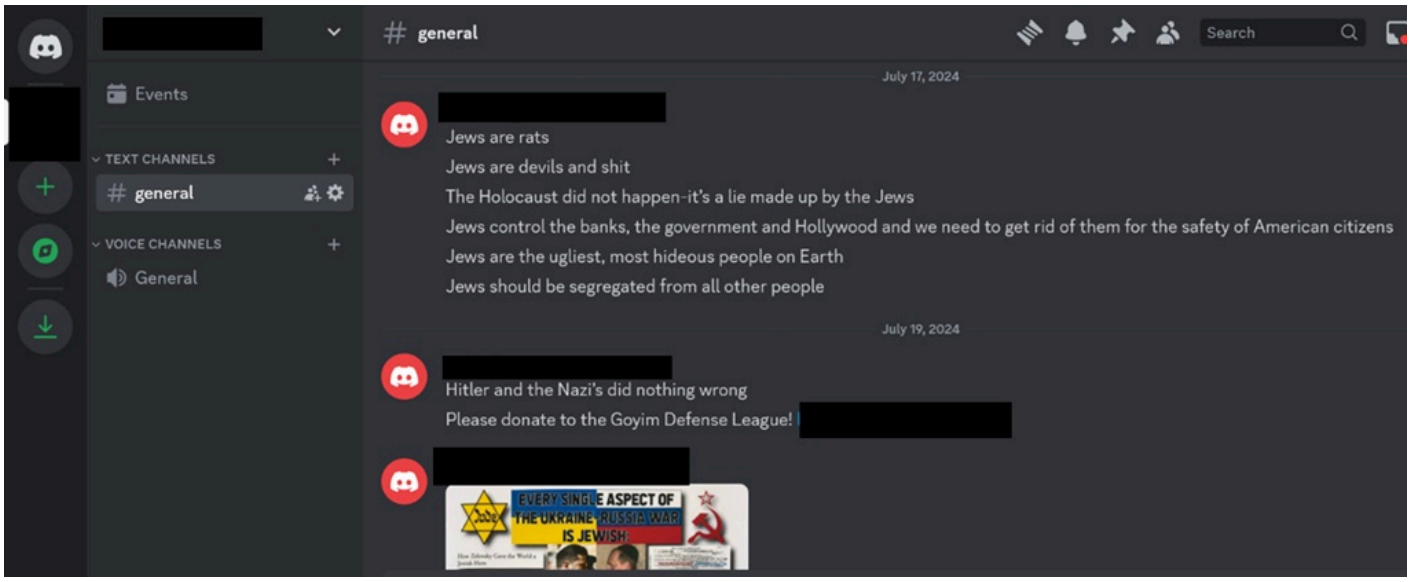


Image 8: Discord content. (Screenshot: 1/15/2025)

Reporting Response

While Discord did not take action on content in the proactive detection phase, once we started reporting posts, they responded proactively, removing posts before we could report them on day 14.

The first half of the content was reported on day 13 but did not receive alerts from Discord until an hour later. The alerts explained why the posting account was being sanctioned and which rules it had broken.

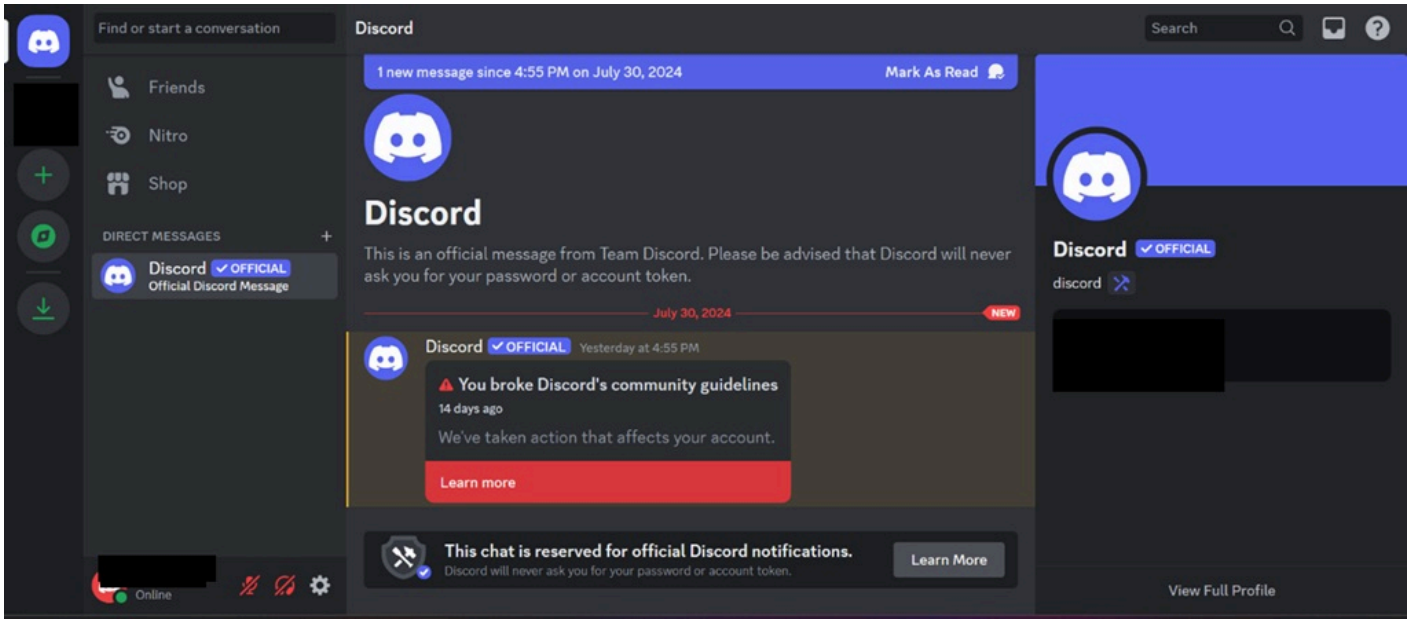


Image 9: Discord alert for breaking community guidelines. (Screenshot: 7/2024)

By day 14, nearly every post had been taken down—even the ones not reported yet. All posts save one were taken down proactively, although we had not reported those images yet. The only post still up said, “Please donate to the Goyim Defense League.” We reported this post, but it was never taken down.

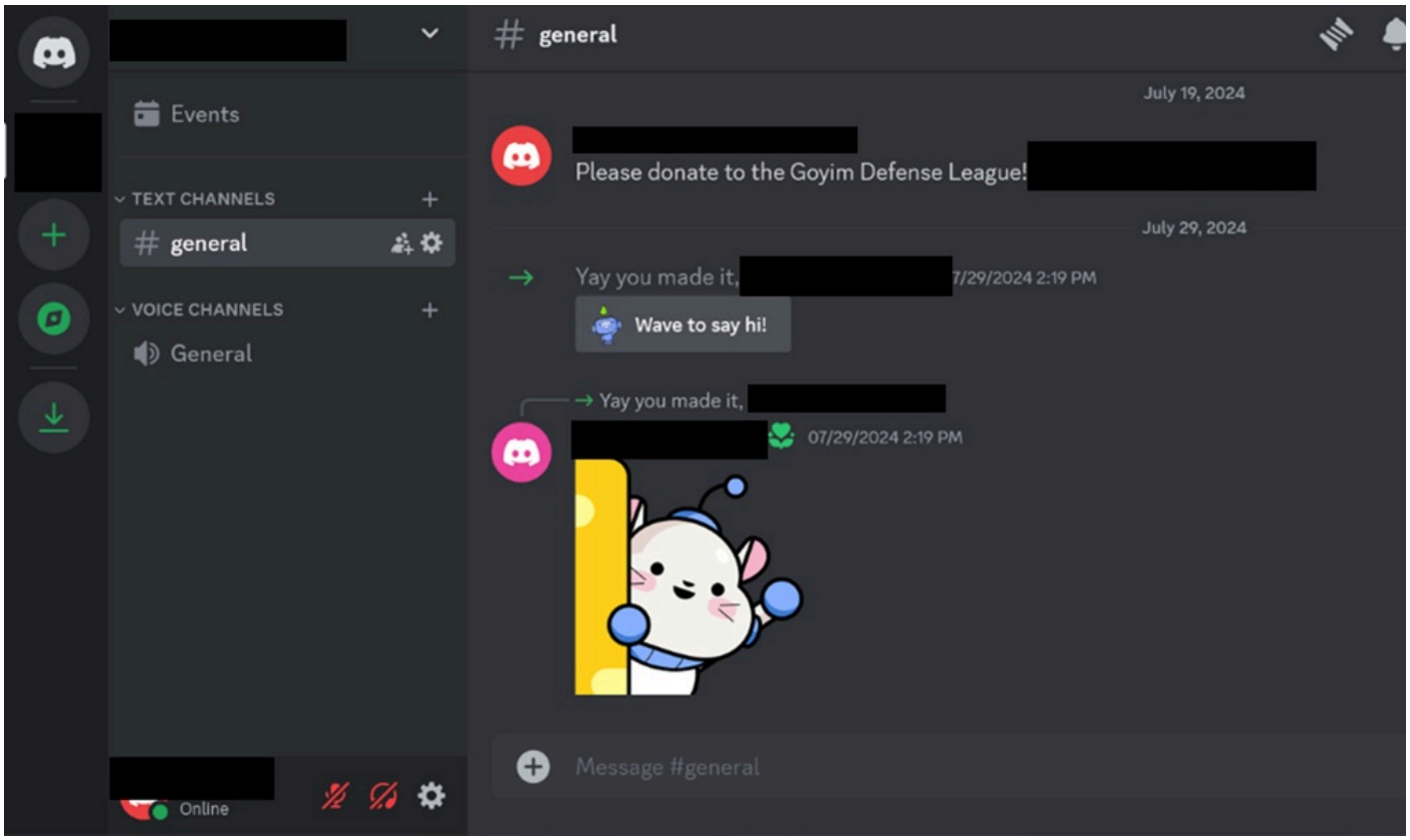


Image 10: Discord group. (Screenshot: 7/31/2024)

Unlike Facebook, Discord did not take action against the group itself; the group is not suspended, and the reporting account can still post to it at the time of writing. Discord did act against the posting account, suspending it a few days later. The posting account received no notification that it was suspended, save for emails alerting us to “new messages” on Discord. These messages were forwarded to the now inaccessible account, so we cannot determine when Discord took action. The posting user might have received a more pointed alert, like an app notification, if we had used the phone app instead of the desktop website. In any case, we were able to create a new account with the same name and using the same email address after our account was removed.

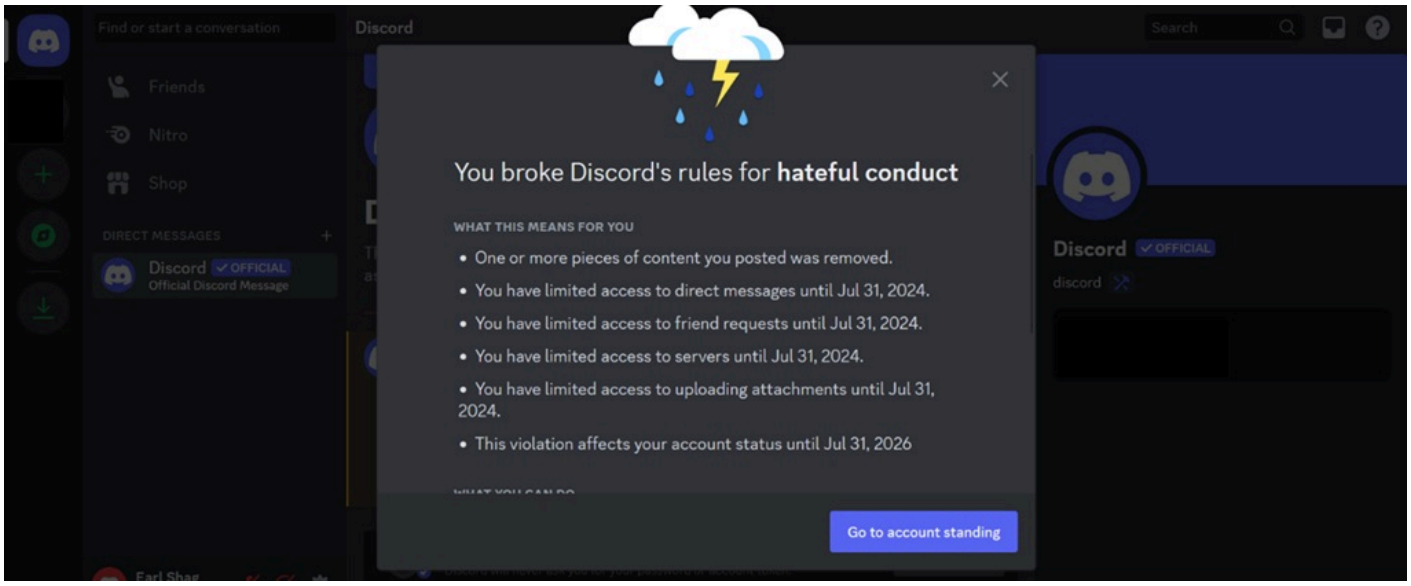


Image 11: Punishments from Discord. (Screenshot: 7/31/2024)

Roblox

The protections offered by Roblox in the private group were minimal, especially compared to Discord and Facebook. We received virtually no feedback on reporting from the platform. There were no warnings or pop-ups for the administrator. Roblox did not remove any posts, despite reports from our reporting account.

Proactive detection

Of the ten posts we created, four were automatically obscured when we posted them, so members could see the posts but not view the text. Roblox replaced the text with the hashmark (#), obscuring not just the offensive words but the entire post (Image 12). It is not entirely clear why Roblox hid some posts this way. For example, the text “Jews are devils and shit” was likely hidden because of the expletive. Roblox might also be automatically filtering keywords that it deems objectionable. It could be that terms like Holocaust, Hitler, Nazis or Goyim were automatically flagged, causing those posts to be removed.

Roblox does not allow pictures or links to be posted on the group's page, so we were unable to post the Covid Agenda and Ukraine-Russia War images. We were able to post "Goyim Defense League" without the embedded link, though it was also immediately hidden.

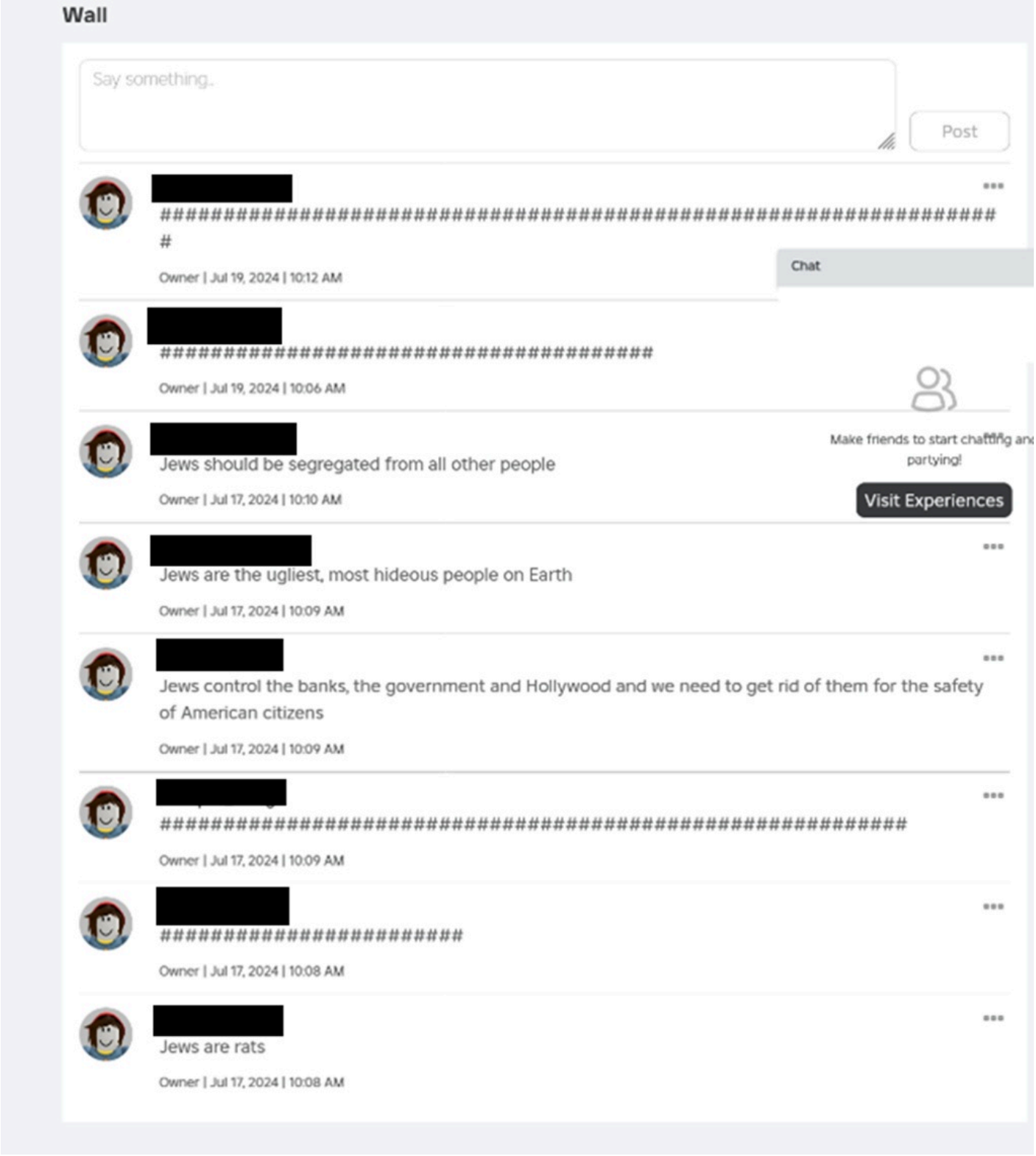


Image 12: Hidden Roblox content. (Screenshot: 7/24/2024)

Reporting response

After we reported every post through our reporting account, there was no response from Roblox, unlike our experience with Facebook or Discord. No other posts were taken down. There were no tools that alerted the poster, the admin or reporting account that any action had been taken—not even a pop-up that the report had been received. The admin received no report that multiple posts on the group had been reported for hate and harassment.

Even after multiple reports, no action was taken on any of the posts. In fact, there was no change in the group from the initial first action, hiding the text. Compare the image of the Roblox group taken on day 14 with the image taken of the group over two weeks later (Image 13). There is no discernible difference.

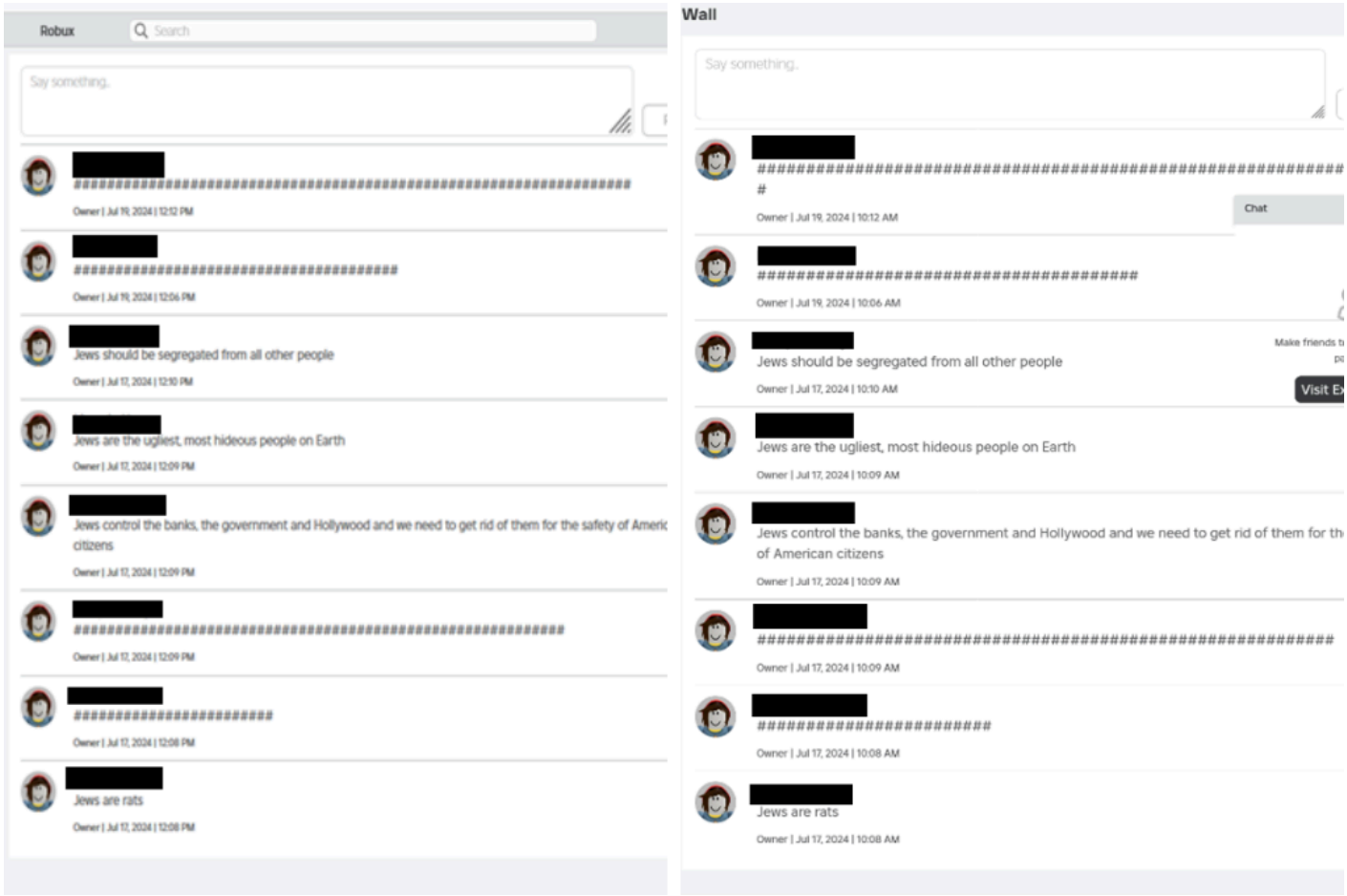


Image 13: Roblox content. (Screenshot: 7/22/2024 and 8/2/2024)

Recommendations For Platforms with Members-only Groups

How platforms respond to violative content in private is just as important as how they respond to violative content in public. Facebook and Discord, while not perfect, moderated these private groups in a way that removed the majority of hateful content and suspended the hateful accounts.

We do not expect platforms to moderate content on private groups as strictly as in public spaces, due to privacy concerns, but we *do* expect an appropriate level of preventative moderation and tools for groups to moderate content. Roblox appears to have had little to no moderation in place, which is concerning given the young age of its player base. Roblox hid offensive and controversial text but did not suspend the posting account or group. It also provided no alerts to the admin or group when a report was made. While we understand that the bulk of communication on Roblox happens in-game, little action was taken within the group chat to moderate offensive content. Terrorist and violent extremist groups organize on private groups and channels, not in public spaces. Platforms with private groups should adopt a minimum standard of moderation on private groups to proactively stem this problem.

Platforms should:

- **Reexamine how automatic filtering functions:** A common problem encountered in this course of our research was the seemingly arbitrary nature of the filtering tools. For example, Facebook automatically filtered one image but not another equally objectionable. It is also not clear why one image was actioned and the other was not. Platforms should reexamine their automatic filtering functions to catch content that slips through the cracks.
- **Prioritize user-reporting within private groups:** Because of the protections that are available for private groups, reporting by group members is the primary channel for

alerting the platform. Platforms should prioritize these reports and act on them quickly. These reports may be one of the first signs that something is amiss within a group.

- **Suspend groups when they reach a threshold of violative content:** Facebook was the only platform that fully suspended our group. Discord did not suspend our group, only the user, who was able to promptly create a new account. Roblox suspended neither the group nor the account. If violative users and groups are allowed to continue their work with barely any pushback from platforms, then any moderation will fail.
- **Utilize metadata in content moderation:** Private groups may restrict the visibility of their content to outside users and the platform itself, but that does not mean that there is no data available. Post times, user addresses, and other metadata are available to platform moderation teams. Discord, in particular, could take better advantage of the tools available to it. Discord's content, save for video and audio content, is not encrypted, and the platform could be more proactive about automatic filtering if it chose to.