Policy Advisory Opinion 2023-09

Regarding Meta's Moderation Policy for Holocaust Denial Case 2023-022-IG-UA

Comment Submitted September 14, 2023, by Tal-Or Cohen Montemayor on behalf of CyberWell Ltd. (CC)

Executive Summary

Antisemitism, including hatred rooted in Holocaust denial and distortion, is running rampant online leading to real world consequences, including harassment and bullying, and even acts of violence. According to a survey conducted by the American Jewish Council, 38% of Jews take steps to hide their Jewish identity online out of fear of being targeted because of their identity online.

In this Policy Advisory Opinion, CyberWell submits data collected on the topic of Holocaust Denial and Distortion from Meta platforms (Facebook and Instagram) and offers recommendations for further improvement for the consideration of Meta's Oversight Board. It is incumbent upon social media platforms to ensure the safety of their users, and to adequately enforce policies meant to provide protection.

In this dataset, CyberWell detected **134 posts denying and distorting the events of the Holocaust**. After reporting to Meta platforms, **only 20% of this data set was removed.**

CyberWell further provides a case study of one example of grotesque Holocaust denial found on Facebook – blaming Jews for inflating the number of victims to gain sympathy. **One such post garnered over 200,000 views.**

CyberWell urges the Oversight Board to uphold Meta's decision to remove the post in question for case 2023-022-IG-UA and to consider the additional recommendations to ensure that Meta's policy against Holocaust hate speech is enforced more effectively.

Introduction to CyberWell

CyberWell is a nonprofit organization dedicated to eradicating online antisemitism through driving the enforcement and improvement of community standards and hate speech policies across social media platforms. Through data, we aim to identify where policies are

not being enforced and where they fail to protect Jewish users from harassment and hate. We have a vested interest in providing guidance on Case 2023-022-IG-UA, as Holocaust denial, distortion, and misinformation is one of the most prominent expressions of antisemitism.

Holocaust Denial and Distortion

A significant contributing factor to this alarming trend is the power of social media, through which geographic boundaries blur and misinformation spreads rapidly. As more people rely on social media for information on everyday topics, from politics and current events to cultural trends and celebrity news, it has become easier than ever to spread hatred and misinformation in general, and specifically against the Jewish community. Far too often "fake news" gains traction fast, leading consumers to believe false information with serious consequences. Unfortunately, misinformation regarding the Holocaust is no different.

Holocaust denial and distortion can be found on every social media platform in numerous languages. The more content that is published denying and distorting the Holocaust, the greater the ignorance and misunderstanding on the subject. As the last generation of Holocaust victims dies out, studies show that young people are shockingly uninformed about the horrors of WWII while, in parallel, hate crimes against Jews around the world are on the rise. As a society, we cannot remain indifferent as one of the most horrific genocides in human history continues to be trivialized, minimized, and thrown in the faces of the victims and their descendants. This leads directly to Jew-hatred and threatens the safety of Jews both in the virtual space and the real world. Due to these dangers, it is critical to monitor social media content for Holocaust denial and distortion and remove it with haste.

Meta Policy on Holocaust Distortion

Meta's Commitment to Community Standards

Meta set an important content moderation standard for addressing antisemitism online by publicly stating their commitment to combating Holocaust denial and distortion. On October 12, 2020, Meta updated its hate speech policy to state: "we are updating our hate speech policy to prohibit **any content** that **denies** or **distorts** the Holocaust".

Community Standards

> Hate Speech, Tier 1: *"Harmful stereotypes historically linked to intimidation, exclusion, or violence on the basis of a protected characteristic, such as Blackface;* ***Holocaust denial****; claims that Jewish people control financial, political, or media institutions […]"*

Bullying and Harassment,* Tier 1: ***"Everyone is protected from […] Claims that individuals are lying about being a victim of a violent tragedy*** *or terrorist attack, including claims that they: Acting or pretending to be a victim of a specific event"*

**Though this community standard refers to personal bullying, the right to be protected from claims that they are lying about the Holocaust, including the scope of the "Final Solution", also refers to Holocaust victims and survivors personally and collectively.*

CyberWell's Methodology

CyberWell's methodology is as follows: use of keywords to identify antisemitic content--> applying specialized dictionary based in the International Holocaust Remembrance Alliance's working definition of antisemitism (IHRA) --> two rounds of human review. Our professional analysts are trained in the fields of antisemitism, linguistics, and digital policy.

Regarding Case 2023-022-IG-UA, CyberWell's technology identified **134 posts in English and Arabic**, published on Meta's social media platforms (Facebook and Instagram), which were categorized as antisemitic according to the IHRA working definition examples focusing on Holocaust denial and distortion:

Example 4: Denying the fact, scope, mechanisms (e.g., gas chambers) or intentionality of the genocide of the Jewish people at the hands of National Socialist Germany and its supporters and accomplices during World War II (the Holocaust)

Example 5: Accusing the Jews as a people, or Israel as a state, of inventing or exaggerating the Holocaust

Rate of Removal

As of September 10, 2023 | Only 20% of the total analyzed dataset was removed

The entire dataset was reported by CyberWell directly to Facebook and Instagram.

Holocaust Distortion & Misinformation

Similar to the post discussed in this case, CyberWell identified a disgusting trend calling into question the scope of the Holocaust, specifically by rejecting the assertion that 6 million Jews were killed, through disputing the logistics of cremating 6 million people in five years. Users subscribing to this belief use the terms "pizzas" or "cookies" to refer to the dead bodies of Jews burned in cremation ovens after being killed and removed from gas chambers.

Trend Example

Link: https://www.facebook.com/100087848326497/videos/574958031487286/

Views: 200,000

Video: This animated video shows two people making pizzas and discussing an order for 6 million. The video's goal is to present the number 6 million as fictitious and impossible and to assert that Jews are in a conspiracy to exaggerate the events of the Holocaust.

Voice & Text: *"Listen up, that guy wants 6 million pizzas.* ***6 million pizzas? We only have four ovens****, boss. How are we going to pull that off? I don't know, but we got to try. Maybe we could just tell him that we made the pizzas, ok? We wouldn't lie about making the pizzas, right? We will say we made the pizzas, but we'll need to make the organizations to enforce the fact that we made the pizzas. Good idea, boss.* ***And we should also make a crime to even question if we made the pizzas and we'll need some TV channels and constant Hollywood films to remind everyone that we made the pizzas****. Yeah, good idea. We'll also need some pictures of a few pizzas. Take some pictures of these ones.* ***We'll also need a few filmstrips of pizzas. Let's pile them on top of each other and bury them****. Film a bit, but not too much, just enough so that the fact that we made the pizzas becomes engrained in the culture".*

Image comment on the right: A screenshot of an interview from the Jordanian TV channel Yarmouk with the caption: "There were 600,000 at the most. They added a zero".

Policy Enforcement During COVID-19

It was alarming to learn of the use of **automatic *closure* policies during COVID-19**. CyberWell welcomes and recognizes the need for effective AI-based tools for ***review*** of hate content to address reported hate speech at scale, especially during an emergent situation. However, for the Jewish population, which is experiencing historically high levels of disproportionate hate crimes against its communities, automatic and wrongful ***closure***

of hate speech reports leaves users feeling unsafe at the very least, and can pose a real risk to the community when that hate speech is violent or threatening at the worst. This risk is not hypothetical, as antisemitism, including violent conspiracy theories around Jewish power or cabals, tend to spike during tumultuous times of economic instability or pandemics.

COVID-19 was no exception. As the world went remote via video conferencing, online antisemitic harassment escalated, including a new tactic known as Zoombombing. Furthermore, there was a slew of online antisemitic content blaming Jews for the manufacturing or deliberate spreading of the coronavirus. Unfortunately, hate crimes against Jews have only increased since pandemic enforcement measures ended. We must consider the possibility that the automatic closure of hate speech reports on Meta's platforms enabled the perpetuation of anti-Jewish conspiracy theories - leading to lasting real-world harm.

CyberWell's research shows that **Meta's day-to-day content moderation infrastructure removes an average of 25% of reported antisemitic content** (Facebook – 15%; Instagram 34%, in the 12 months prior to this submission). This average removal rate is also reflected in the average removal rate yielded for this **Advisory Opinion's Holocaust Hate Speech dataset – only 20%.** Begging the question - is automatic closure an appropriate solution for addressing hate speech when the reports made by users flagging antisemitism are ignored or wrongfully unactioned 75% of the time? If automatic closures are necessary, the dataset of automatically closed hate speech reports must be studied to maximize enforcement of community standards in the future and to further understand the nature and spread of hate speech behavior in times of crisis. Additional options could include:

- Partnering directly with Trusted Partner nonprofits of minority communities and interest groups to handle the influx of hate speech reports during emergent situations.
- Providing access to the dataset of automatically closed reports for hate researchers and nonprofit organizations to vet and study in order to make recommendations to the content review and policy teams for future use.

In the spirit of partnership and transparency, CyberWell asks the Oversight Board to encourage Meta to share openly with stakeholders and partners:

- How many reports were automatically closed during the COVID-19 pandemic?
- How many of the automatically closed reports perpetuated the idea that Jews were responsible for COVID-19?
- Have similar automation closures been used since the COVID-19 pandemic?
- Has a policy been set to determine when automatic closures should be activated?

Policy Recommendations

- Approve and support Meta's decision to remove the post in question.
- Enforce Meta's current Community Standards in general and the tiers related to Holocaust denial and distortion in particular.
- Reduce the gaps in rates of removal for Holocaust denial and distortion in English and Arabic.
- Monitor Holocaust denial, distortion, and misinformation in all the available content dimensions - text, image, video, voiceover - in the posts and the comment section.
- Flag the combination of terms **"six/6 million"** AND **food metaphors** referred to **Jews or to the Final Solution**; **"six/6 million never happened"** as a high probability of hate speech.
- Reduce the time between receiving a report of hateful content and the final removal of the content.
- During emergent situations, partnering with Trusted Partner nonprofits of minority communities and interest groups to handle the influx of hate speech reports rather than introducing blanket automatic closures.
- Provide access to the dataset of automatically closed reports for hate researchers and nonprofit organizations to vet, study, and make subsequent recommendations.
- Introduce an indication of when a report is reviewed by automation. We believe the experience of having a report rejected by automated technology will be less hurtful for users than being wrongfully rejected by a human.

Appendix of Examples: "6 Million" as Holocaust Denial & Distortion

1. https://www.facebook.com/100087848326497/videos/574958031487286/
2. https://www.facebook.com/aaron.tubbs.355/posts/pfbid021xu7Z868CZ8wK9rF9fsJDDLrF7Ef2H7hBBrxke53EHGRyj7yDLGitVUwu926TEo8l
3. https://www.facebook.com/100078081847870/videos/316291300558654/
4. https://www.facebook.com/jacob.lynch.739978/posts/pfbid0k36EQknZ7zNViNMTC4oxJN9rHxCX1oGNoGQSFRRcX3nN2gLQtvA75KieAMupqGfNl
5. https://www.facebook.com/TheDukeOfMemes2/posts/pfbid0n1DTSYV3GTgWNytKM9Zj1ZWkW87rRziGb6oUx1vdrEUGfdThkhDqsGTWD64ipJoKl