ANTISEMITISM POLICY TRUST

Response to Meta's Oversight Board Consultation Regarding Symbols Adopted by Dangerous Organisations

25 February, 2025

The Antisemitism Policy Trust is a UK-based charity that works to educate and empower parliamentarians and policy makers to address antisemitism. For more than two decades, the Trust has provided the secretariat to the All-Party Parliamentary Group (APPG) Against Antisemitism. We have published research papers about online antisemitism and have advised Government, Parliament and the UK regulator, Ofcom, on matters relating to online safety. We have also worked with technology companies and social media platforms, including Meta, on reducing harms caused by online antisemitism.

In the midst of a dramatic increase in antisemitism worldwide,[1] including hate speech and violent attacks against Jewish targets, the proliferation of neo-Nazi content, is of great concern.

With regard to your question on the ways in which neo-Nazi and other extremist content is disguised to bypass content moderation on social media, a recent study that we published, co-authored by Decoding Antisemitism and the International Network Against Cyber Hate (INACH) provides some detail. The full report can be accessed online: https://antisemitism.org.uk/wp-content/uploads/2024/12/APT-Detecting-Deep-Fakes.pdf

Our research found AI-generated images that contain antisemitic content. Several of these incorporated hidden Nazi content, including a swastika and an image of Hitler. These are two examples:

[1] https://www.timesofisrael.com/global-antisemitism-surged-340-in-two-years-report-finds/

This seemingly innocent picture of three young women hides an image of a swastika in the background, better visible when observed from a distance:

This image of paragliders was created following Hamas's attack on Southern Israel on the 7 October 2023. During this event, some of the terrorists invaded Israel by paragliding in before massacring Israelis. The smoke in this image creates a hidden representation of Hitler. Again, the image is more visible when viewed from a distance:

We have also found images with more explicit forms of antisemitism, for example, this AI generated image, using the grotesque caricature of Freddy Krueger as a Jew, leading Jewish men in prayer or Jewish studies:

Our results indicate that generative AI tools lack effective safety measures to recognise or add friction to reduce the production of such content. It is being shared online, which brings into question not only the creation techniques but effective moderation too.

In addition, we ran some of the images through Google's Gemini because it enables to upload images. It has successfully identified, analysed, and accurately explained different forms of antisemitic content, including in the image above, featuring Freddy Krueger, but it failed to recognise hidden Nazi images in the examples above and in other examples included in our report. This shows the limitations of current AI tools when used for moderation, and the significance of human moderation. As seen here, these are easily recognisable Nazi symbols that do not require comprehensive knowledge or understanding in the area.

Regarding your question on how Meta should treat symbols with different meanings when reviewing at scale, where the review by the company's subject matter experts is limited, in instances when neo-Nazi symbols are used as in the examples given in your consultation, we have several recommendations.

First, Meta should immediately reverse the announced intention to change content policy and associated moderation. This is a backwards step, has been roundly condemned by those working to counter hate around the world, and is insulting to those of us that have sought to engage with the company in good faith.

Second, the company could expand the use of trusted experts to help decide the meaning of the content, which could depend on the context – but see above, to point one.

Third, an additional measure would be to provide specialised training to a group of experienced moderators and directing the more complex cases to them. Having moderators that hold a deeper understanding in particular issues could also prevent over-moderation or over-removal of content, which means less impact on freedom of expression. This would then meet some of the criticism which was excused to announce the policy change referenced in point one above.

Developing and training AI tools to make them better able to detect implicit forms of antisemitic and other extremist content, as well as analyse these symbols within their context, would also be worthwhile investment. When it comes of Nazi language and symbols, tools will need to be able to detect coded language and evolving terminology. These tools might be more effective if they do not analyse a symbol or keyword in a vacuum, but as part of posting patterns, network clusters and entire threads, which could provide a more comprehensive understanding of the meaning of the content and better ablity to judge whether it violates Meta's hate speech or extremism codes (such as they will be). There are multiple organisations that could be trusted to provide expert advise to Meta and give ongoing support. This should still require human oversight when AI flags questionable content, but moderators would only need to look at borderline cases if the AI tools are effective enough.

Meta's recent decision on moderation will lead to rising levels of extremist disinformation, the least the company can do is improve AI and other systems to assist with whatever limited moderation will be left as the company undermines the safety and security of minority communities across the world.