CyberWell

# AI-Generated Antisemitism

## Abuse Trends and Safety Gaps Across Social Media and AI Platforms

May 2026

## Table of Contents

## Executive Summary

This report examines the rapid rise and evolution of AI-generated antisemitic content across the major social media platforms, including Meta, TikTok, X, and YouTube. It also evaluates how current policies address this emerging harm. While existing research evaluates AI systems and their regulatory approaches, studies devote less attention to how AI-generated antisemitic content manifests on social media or how platforms enforce their policies in response. This report addresses that gap and proposes recommendations for social media platforms, AI companies, and policymakers to adequately address these issues.

- CyberWell compiled a dataset of 307 AI-generated pieces of content posted to social media between January 2025 and February 2026, that the organization vetted and confirmed to be antisemitic. Analysis of this dataset examines key narratives, dissemination patterns, and gaps in enforcement across social media platforms, AI model safeguards, and governance frameworks. While the dataset nominally covers this full period, **98.4% of the content identified was posted from June 2025 onward**, reflecting a sharp inflection point in both the volume and normalization of AI-generated antisemitism online.
- This dataset represents **CyberWell's most widely viewed dataset to date, with over 30 million views and more than 2.8 million user interactions**. These figures underscore the significant reach, platform amplification, and user engagement associated with AI-generated antisemitic content and reflect its increasing normalization across social media platforms.
- **AI-generated antisemitic content is overwhelmingly concentrated on video-based platforms** – TikTok, Instagram, and YouTube – which account for **79.2% of all posts in the dataset**. This reflects the central role that short-form video and multimedia platforms are playing in amplifying AI-generated hate today.
- At the platform level, TikTok accounts for the largest share of content (35.8%) and demonstrates the highest enforcement rate (88.2%). In contrast, Instagram drives the majority of engagement (64.9%), generating a disproportionate share of views and interactions relative to its share of posts (24.8%). TikTok and Meta exhibit comparatively higher removal rates (88.2% and 67% respectively), likely reflecting their more explicit policy frameworks addressing AI-generated content. However, YouTube and X demonstrate lower removal rates (28.1% and 20% respectively), and neither platform maintains policy provisions that are similarly explicit in this regard.
- Across the dataset, **63.5% of the content was removed, indicating comparatively higher enforcement rates**. While this figure is higher than the average rate of removal that CyberWell tracked across platforms in 2025, it should be interpreted with caution. This removal rate is likely due in part to the lower-than-typical share of posts from X, which has historically demonstrated limited enforcement. It also reflects the **higher**

prevalence of violent content in this dataset, which platforms tend to enforce more strictly than other forms of hate speech. That said, even this adjusted figure masks a critical timing issue. High eventual removal rates do not necessarily indicate effective enforcement, as content that explicitly glorifies or incites violence frequently remains online long enough to accumulate hundreds of thousands or millions of views before removal.

- AI-generated antisemitic content that remained online was typically implicit rather than explicit. It relied on coded language, humor, disclaimers (e.g., "#satire", "for educational purposes"), mockery, or stereotypical depictions to obscure intent and present harmful content as neutral. In particular, Holocaust-related AI-generated content often appeared as mockery rather than direct denial, while other posts reinforced stereotypes of Jews as greedy or money-obsessed. These patterns highlight the need for platforms and AI companies to better recognize coded and evolving forms of antisemitism. They also emphasize the value of working with experts to effectively address these dynamic expressions.
- Three primary narratives dominate the dataset: Depictions of Jews as greedy or money-obsessed (33.2%), Holocaust-related hate speech (21.5%), and event-driven violent rhetoric against Jews (21.2%). The first and third narratives frequently align with the second and first examples of the IHRA working definition of antisemitism. IHRA Example 2 – stereotypical and conspiratorial depictions of Jews – accounts for 75.6% of the dataset, and IHRA Example 1 – calls for or justification of violence against Jews – accounts for 33.2%. The prevalence of IHRA Example 2 in AI-generated content is consistent with CyberWell's previous findings, which reflect the continued dominance of classic antisemitic tropes on social media, now adapted into AI-generated formats. Additionally, when comparing CyberWell's 2025 State of Online Antisemitism dataset with this new verified AI-generated content dataset, calls for or justification of violence against Jews (IHRA Example 1) is more than twice as likely to be found in antisemitic AI-generated content than in user generated content.
- AI company safety and usage policies often rely on broad definitions of harmful content without clearly specifying protected categories. This creates gaps in enforcement, increasing the risk that AI models will generate harmful content, particularly in implicit or coded forms that evade existing safety guardrails.
- National AI governance frameworks do not adequately address the intersection between AI models and social media platforms. They often treat these as separate domains, overlooking how generative AI enables the widespread creation and amplification of hateful content, including antisemitism.
- The report concludes with targeted recommendations for social media platforms, AI companies, and policymakers. These include: clarifying that hate speech and violence policies apply equally to AI-generated content; strengthening detection and

enforcement across formats; addressing coded antisemitic narratives and disclaimer-based evasion; improving AI safety policies and guardrails; and advancing regulatory measures to improve transparency, accountability, and cross-platform coordination.

## Introduction

Across digital spaces, antisemitism persists by adapting historical narratives to evolving technologies and modes of dissemination.

Today, artificial intelligence (hereinafter: "AI") represents the latest shift in how antisemitic narratives are promoted online. AI refers to models and solutions that are designed to perform tasks that typically require human intelligence, including analysis, decision making, and content generation. They can generate text, images, and videos at scale, significantly lowering the barrier to content creation. As these tools become more widely used, migrating into and directly integrated into social media platforms, they introduce new risks. When exploited, generative AI contributes to the large-scale production and spread of antisemitic content.

AI-generated antisemitism changes the nature of online antisemitism by increasing its scale and automation, often outpacing moderation and enforcement capabilities of social media platforms. The realism of AI-generated content, combined with coded language and implicit messaging, can obscure antisemitic narratives and intent. As a result, existing regulatory frameworks and platform moderation systems struggle to keep up. This also raises concerns about youth exposure and radicalization, as algorithms promote AI-generated content that is hateful and designed to appeal to younger audiences through highly engaging audio and visual formats.

While AI-generated antisemitic content existed prior to October 7, 2023, during the aftermath of the coordinated Hamas attacks, this type of content gained more traction and a more noticeably visible and trackable presence on social media platforms, prompting CyberWell to begin monitoring it as a regular part of its research and monitoring methodology. Prior to execution of the attack itself, examples appeared sporadically. Afterwards, content glorified the Hamas attacks, including AI-generated images depicting Hamas paragliders crossing into Israel.

Additional isolated examples appeared in early 2025, when users responded to the release of the John F. Kennedy Assassination Records, by producing AI-generated songs like "Kennedy Killers", which attributed Jewish responsibility to the event.

**A significant turning point occurred in mid-2025, when CyberWell identified a sharp spike in the cross-platform dissemination of AI-generated antisemitic content.** This increase became especially pronounced during Israel's 12-day war with Iran in June 2025. During this period, AI-generated songs such as "Boom, Boom, Tel Aviv" circulated widely across social media, contributing to the normalization of content that extended beyond hate

speech to include the justification and celebration of violence against Jews. Shortly thereafter, AI-generated trends such as the "promised 3,000 years ago" trend and Holocaust-related hate speech became more widespread. These trends were frequently presented through coded or humor-driven formats.

This report draws on data from hundreds of posts collected between January 2025 and February 2026. This period reflects the growing integration of AI-generated content across major social media platforms and documents a shift from isolated traction of AI-generated images on social media to more sophisticated multimodal forms of antisemitic content. The incidents identified during this period reflect how AI tools have become more accessible and widely used across social media, allowing users to produce adaptive and highly shareable antisemitic content.

This report assesses how social media platforms, AI companies, and governments address AI-generated antisemitic content through their existing policies and regulatory frameworks. Its findings are intended to inform social media platforms, AI developers, researchers, and policymakers seeking to better understand and mitigate this evolving harm.

The structure of this report begins with a review of existing research on AI and antisemitism, followed by an outline of CyberWell's methodology and data collection process. It then examines content distribution and engagement metrics across the major social media platforms, along with the role of AI models in facilitating these trends. Next, the report explores key antisemitic narratives identified in the dataset, as well as broader insights and patterns as to how this content is produced and disseminated. Finally, it evaluates current policy frameworks across social media platforms, AI companies, and governance structures. It concludes with targeted recommendations for how these stakeholders can address the rise of AI-generated antisemitic content.

## CyberWell's Mission

CyberWell is an independent nonprofit organization dedicated to addressing online antisemitism by driving the improvement and enforcement of community standards and hate speech policies across the digital space. Through data-driven analysis, the organization identifies where these standards are inconsistently applied or fail to protect Jewish users from harassment and hate. CyberWell currently monitors **Facebook, Instagram, X (formerly Twitter), TikTok, and YouTube** in both **English and Arabic**.

Recognized as a **trusted partner and flagger** for **Meta (Facebook, Instagram, & Threads), TikTok, and YouTube**, CyberWell advances its mission through escalating policy-violating content and advising content moderation teams. In May 2022 as part of its strategy to increase transparency and accessibility, CyberWell launched the first ever open data platform tracking online antisemitic content, leading a new and innovative approach to spark tech platform accountability and contribute to digital policy compliance.

## Review of Existing Research

A growing body of research examines the intersection of AI and antisemitism, focusing on the limitations of AI company guardrails and the changing forms of antisemitic rhetoric in AI-mediated environments. While existing research evaluates AI systems and their regulatory approaches, most studies analyze these issues in isolation, with limited attention to **how social media platforms operationalize their trust and safety policies in response to AI-generated antisemitic content**.

Studies evaluating large language models (LLMs) demonstrate that the datasets used to train these systems often contain misinformation and biases, which the models can reproduce and amplify.[1,2]

Many of the websites used in AI training datasets function as active hubs for antisemitic discourse, raising concerns about their inclusion in model development.[3] For example, **Reddit ranks among the most cited domains across major AI systems, while analyses of ChatGPT outputs indicate that Wikipedia alone contributes to roughly half of generated responses.[4,5,6] The reliance of AI companies on these websites underscores the risk that antisemitic narratives circulating online may become embedded in model inputs and later disseminated at scale.**

The *International AI Safety Report*, which compiles expertise from 96 AI experts across multiple countries, notes that underrepresentation or stereotypical representations of certain groups in training data "can lead to failures in how models trained on this data are able to generalize to the target populations".[7] The report further highlights these structural

---

[1] Gabriel Weimann, "New Trends in Online Antisemitism", in *Antisemitism on the Rise: New Ideological Dynamics* (European Institute for Counter Terrorism and Conflict Prevention, April 2024), www.eictp.eu/wp-content/uploads/2024/05/EICTP_Research_Papers_Antisemitism_FINAL.pdf.

[2] Anti-Defamation League, Center for Technology and Society, "The Safety Divide: Open-Source AI Models Fall Short on Guardrails for Antisemitic, Dangerous Content", December 2025, www.adl.org/resources/report/safety-divide-open-source-ai-models-fall-short-guardrails-antisemitic-dangerous.

[3] Anti-Defamation League, Center for Technology and Society, "Antisemitism on Reddit: Addressing Moderator Concerns", (August 2024), www.adl.org/resources/report/antisemitism-reddit-addressing-moderator-concerns.

[4] Reddit, "Letter to shareholders", Q2 2025, https://s203.q4cdn.com/380862485/files/doc_financials/2025/q2/Q2-25-Shareholder-Letter.pdf.

[5] Nick Lafferty, "AI Platform Citation Patterns: How ChatGPT, Google AI Overviews, and Perplexity Source Information", *Profound*, June 2025, www.tryprofound.com/blog/ai-platform-citation-patterns#top-source-share-analysis.

[6] Blue Square Alliance, "What the Phrase 'It Was Promised to Them 3,000 Years Ago' Means and How It Is Used Online", August 2025, www.bluesquarealliance.org/uncategorized/jewish-phrase-promised-to-them-b/?nab=1. As discussed in the article, Reddit played a role in amplifying the "promised 3,000 years ago" trope on social media.

[7] UK Department for Science, Innovation & Technology, "International AI Safety Report: The International Scientific Report on the Safety of Advanced AI", (January 2025), 93, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf.

biases when mentioning the predominance of English-language and Western-centric datasets in AI-systems.[8]

Finally, research on online antisemitism highlights the role of coded language in normalizing hate speech across linguistic and geographic contexts.[9,10] Scholars document how antisemitic content online, including AI-generated content, often appears under the guise of political rhetoric or satire, particularly in discourse surrounding the Israeli-Palestinian conflict and the Holocaust.[11,12] While existing research recognizes the detection challenges posed by such content, gaps remain in understanding how these coded dynamics operate across platforms and how they influence the visibility of antisemitic posts.

This report addresses these gaps by examining how AI-generated antisemitic content functions within social media ecosystems and how platforms address it. The analysis focuses on the intersection among social media platforms, AI systems, and policymakers, assessing how their respective frameworks align in practice. Drawing on existing research and CyberWell's data, the report translates these findings into actionable recommendations for AI companies, social media platforms, and policymakers.

## Methodology

CyberWell's methodology is grounded in the International Holocaust Remembrance Alliance (IHRA) working definition of antisemitism. While the IHRA definition includes examples of scapegoating Jews for real or imagined wrongdoings, it does not fully capture a contemporary form of antisemitism identified and closely tracked by CyberWell: the denial of contemporary antisemitic attacks and the attribution or blaming of Jews for orchestrating these antisemitic attacks against themselves.

To address this gap, CyberWell includes two additional narrative-based categories in this framework as part of the methodology: "Denial of Violent Events Against Jews or Israelis" and "Conspiratorial Self-Victimization Against Jews or Israelis."

CyberWell's data collection process consists of identifying antisemitic keywords and applying a specialized dictionary to classify antisemitic content. Each piece of content is then reviewed by trained analysts with expertise in antisemitism, linguistics, and digital policy

[8] "International AI Safety Report: The International Scientific Report on the Safety of Advanced AI".
[9] Daniela Santus and Cristina Bettin, "Antisemitism in the algorithmic age: propaganda, disinformation, and generative AI," *Z Religion Ges Polit*, (December 2025), https://link.springer.com/article/10.1007/s41682-025-00234-6.
[10] Matthias J. Becker, Jordan Blatter, and Oksana Stanevich, "Decoding antisemitism online: linguistic and multimodal challenges in the age of AI", *Frontiers in Communication* 10, (January 2026), www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2025.1729279/full.
[11] "Antisemitism in the algorithmic age: propaganda, disinformation, and generative AI".
[12] "Decoding antisemitism online: linguistic and multimodal challenges in the age of AI".

who assess both its alignment with antisemitic frameworks and its potential violation of social media platform policies.

During the data collection process, analysts distinguished AI-generated content from human-created material through a multi-step verification process. Identification relied on several indicators, including platform-provided AI labels or disclosures, visible signs of synthetic visuals, watermarking associated with AI models or content creators, and contextual signals such as reposts or captions explicitly identifying the content as AI-generated. In cases where watermarks were absent, CyberWell analysts cross-referenced posts with user disclosures to determine whether material originated from AI systems.

## Data Collection & Scope

From January 2025 – February 2026, CyberWell collected and analyzed AI-generated antisemitic content across five major social media platforms: Facebook, Instagram, YouTube, TikTok, and X. Using a combination of AI-powered detection tools, keyword-based tracking, and manual review, CyberWell's trained analysts verified **307 AI-generated antisemitic posts** during this period.

The majority of identified posts appeared on video and image-based platforms including Instagram, TikTok, and YouTube, where the production of AI-generated multimedia content is most prevalent. Furthermore, trained analysts evaluated each post in accordance with the IHRA Working Definition of Antisemitism and CyberWell's internal classification framework.

**Mid-2025 marked a sharp increase in AI-generated antisemitic content, with 98.4% identified from June 2025 onward**. During Israel's 12-day war with Iran in June, CyberWell identified the AI-generated antisemitic song "Boom, Boom, Tel Aviv" and flagged it for at-scale removal across the major social media platforms. Since then, AI-generated content has emerged as a significant driver in the amplification of antisemitic narratives online.

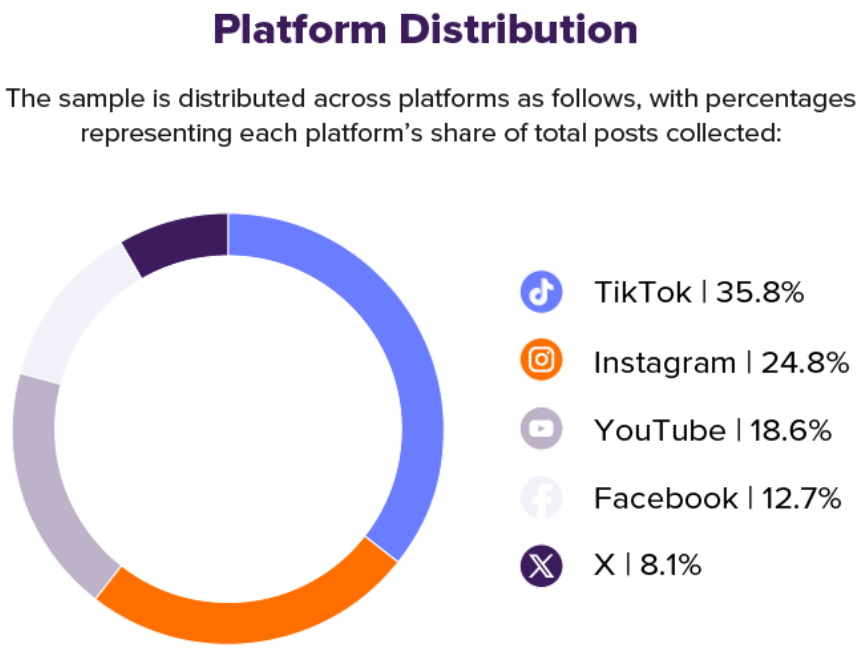
This dataset has several limitations. Although initial monitoring included English, French, and Arabic content, the final analysis primarily reflects **English-language** posts due to the lower prevalence of AI-generated antisemitic content identified in other languages during the research period. As AI models continue to develop, the patterns identified in this dataset may extend to additional languages and contexts beyond those captured within the analyzed timeframe. The rapid emergence of new AI tools, content formats, and evasion tactics also present methodological challenges. Finally, the dataset only analyzes publicly available content shared on the major social media platforms listed above during the selected time frame, excluding content that may have circulated solely within the AI models.

## Platform Distribution & Enforcement Findings

This section examines how AI-generated antisemitic content identified in CyberWell's dataset is distributed across major social media platforms. It analyzes user engagement alongside platform removal rates to identify where AI-generated antisemitic content gains the most visibility and how effectively platforms respond.

### Platform Distribution

The dataset was distributed across platforms as follows, with percentages representing each platform's share of total posts collected during the reporting period:




This distribution indicates that AI-generated antisemitic content is concentrated on visual and video-based platforms, with **TikTok, YouTube, and Instagram accounting for 79.2% of identified posts** during the research period. **TikTok alone represents the largest share of posts (35.8%)**, indicating a significant concentration of such content on the platform.

## Views & Engagement

The dataset analyzed represents **CyberWell's most-viewed dataset to date**, with over **30 million views** and more than **2.8 million user interactions.**[13] These figures underscore the significant reach and engagement associated with AI-generated antisemitic content and point to its increasing algorithmic amplification and prevalence across social media platforms.

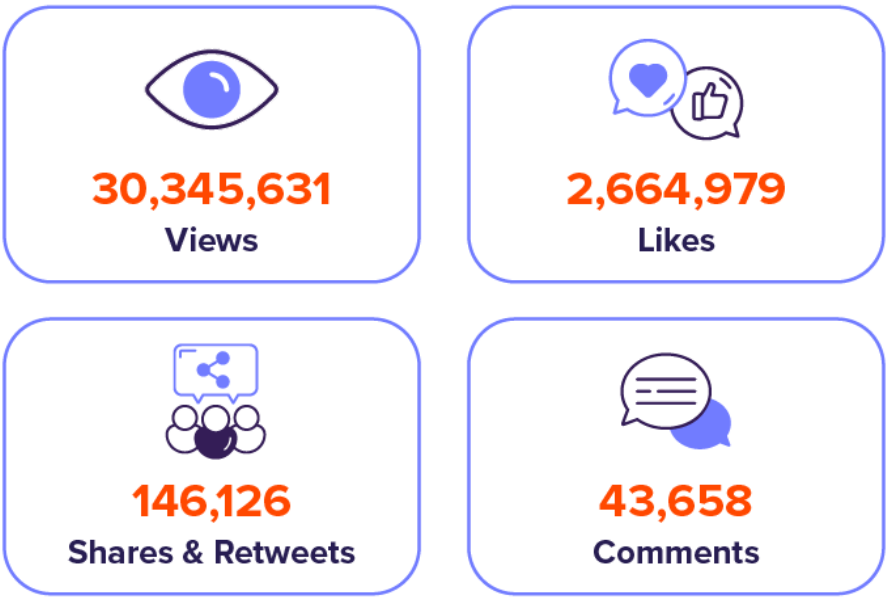
Breakdown of the **total engagement** observed across platforms:

**Views**[14] **| 30,345,631**

**Likes | 2,664,979**

**Shares & Retweets**[15] **| 146,126**

**Comments | 43,658**




**Meta accounts for the highest levels of engagement, with Instagram generating 64.9% of total engagement despite accounting for only 24.8% of posts in the dataset.** This indicates that AI-generated antisemitic content on the platform achieves disproportionately higher visibility and user interaction relative to its share of total posts.

---

[13] User interactions include the sum of all comments, likes, and combined shares and retweets recorded across all platforms. Views are counted separately.
[14] On Facebook and Instagram, view counts apply only to videos and reels. The view metrics were therefore not calculated for Facebook posts unless they were videos.
[15] This figure combines both retweets (X) and shares (Meta, TikTok, and YouTube).

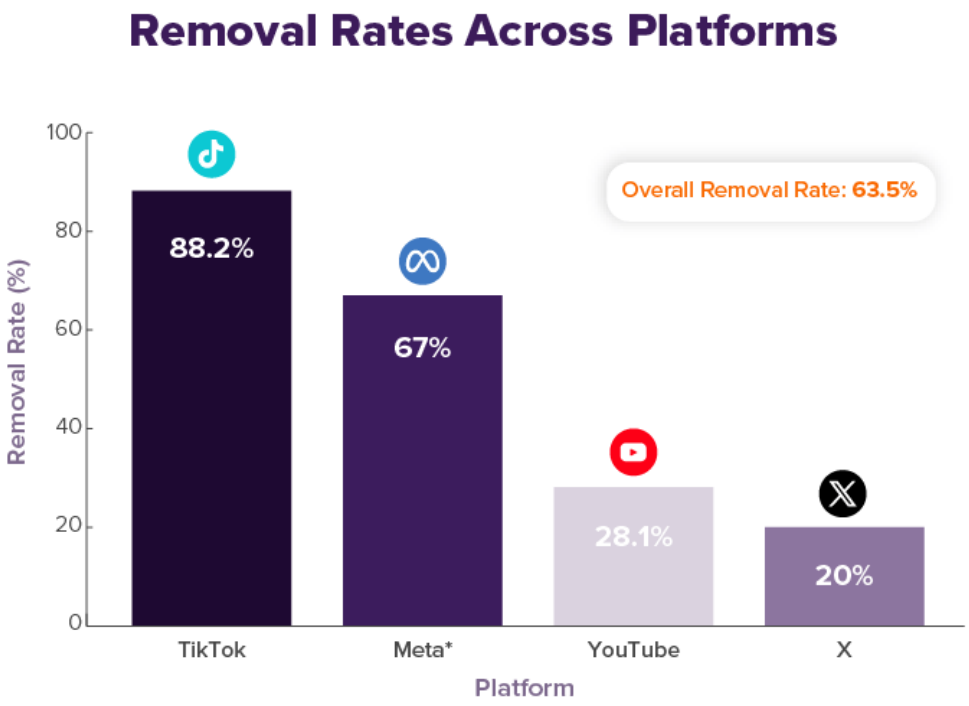
Overall levels of engagement further highlight the disproportionate visibility and amplification of AI-generated antisemitic content, including highly graphic and violent material. **Content that glorifies, justifies, or calls for violence against Jews accounts for 33.22% of posts and generates 33.66% of total views and 41.41% of overall engagement (likes, shares, retweets).** This indicates that violent AI-generated antisemitic content not only circulates widely but also attracts significantly higher user interaction relative to its share of total content.

Importantly, according to CyberWell's research, **content that promotes, glorifies, justifies, or calls for violence against Jews is more than twice as likely to be found in AI-generated content than user-generated content on social media.**

These findings therefore underscore the urgent need for platforms to strengthen detection and enforcement mechanisms targeting AI-generated content that glorifies, justifies, or incites violence against Jews.

## Content Availability & Removal Rates

CyberWell analyzed the **removal rates** for all 307 posts in the dataset to assess how platforms moderate AI-generated antisemitic content.




** While Facebook and Instagram are listed separately in the platform distribution data, their removal rates are reported together under "Meta," as both platforms are owned and operated by the same company. This grouping reflects how content moderation policies and enforcement practices are applied across Meta platforms.*

## Key Findings

- 63.5% of all posts in the dataset were removed for violating platform policies. This represents a higher enforcement rate for AI-generated content, compared to the **52.53% removal rate** of antisemitic content documented in CyberWell's 2025 Annual Report.

  This difference likely reflects the higher proportion of violent content in the current dataset, which platforms tend to enforce more strictly than other forms of hate speech, and the differences in platform composition. X, which has the lowest removal rates, accounts for a significantly smaller share of this dataset (8.1%), compared to 37% in CyberWell's 2025 Annual Report.

- **87.3% of posts that explicitly glorified, justified or called for violence towards Jews were removed across all platforms**. Meta hosted the most amount of AI-generated content that was consistent with IHRA 1. However, high levels of engagement and views associated with these posts indicate that **removal was often delayed, allowing many of these posts to circulate widely before enforcement action was taken.**

- **TikTok** hosted the largest share of AI-generated antisemitic content (**35.7%** of the dataset) yet also recorded the highest removal rate (**88.2%**). Posts that remained online reflect recurring patterns of antisemitic rhetoric that may evade enforcement due to implicit framing or coded language. Several of these posts also rely on stereotypical depictions of Jews that may appear less overtly violative, as they do not use explicit antisemitic slurs.

- **Meta** removed 67% of AI-generated antisemitic posts. **Instagram** generated the highest level of overall engagement in the dataset (**64.92%**) and accounted for the largest share of total views (**62.43%**). Content that remained online primarily included posts featuring stereotypical depictions of Jews as greedy or money-obsessed, as well as Holocaust-related mockery.

- **YouTube** demonstrated a low removal rate (**28.1%**). Most content from the dataset that remains online features Holocaust-related mockery and stereotypical depictions of Jews as greedy or money-obsessed.

- **X** had the lowest removal rate (**20%**) in the dataset. The text-based nature of this platform limited the prevalence of AI-generated antisemitic content captured in this analysis. While the dataset is narrower in scope for X, its low removal rate is consistent with broader enforcement trends on the platform.[16]

[16] CyberWell's 2025 Annual Report found that X's antisemitic content removal rate declined sharply to 29.46%, indicating persistent enforcement challenges on the platform.

## The Role of AI Models in Facilitating Online Antisemitism

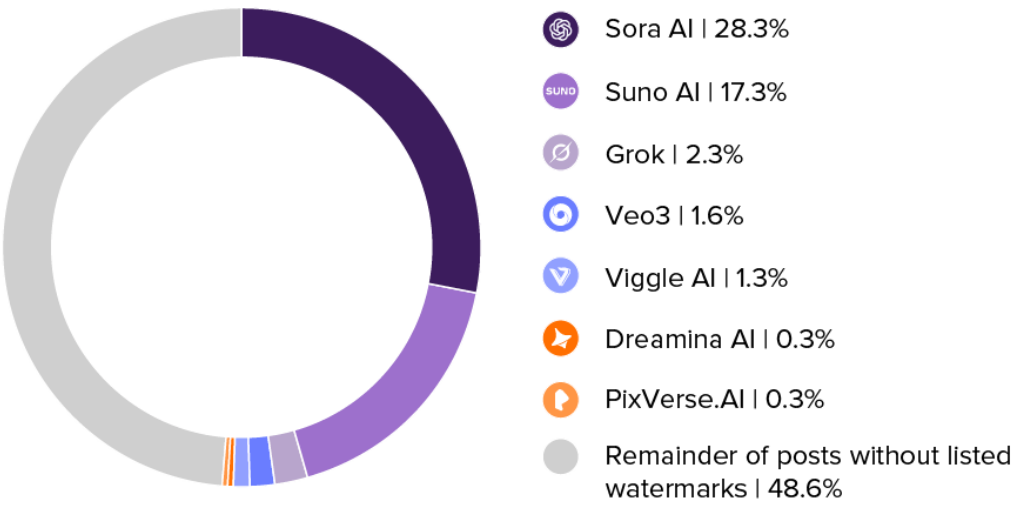
As AI-generated antisemitic content becomes more prevalent across social media platforms, it is essential to examine the generative-AI models that facilitate their production.

### Capabilities and Evolution of Generative AI Models

AI models such as Sora (Open AI), Veo (Google/DeepMind), and Suno represent a new generation of tools capable of producing synthetic video and audio content from text prompts. These models lower the technical barrier to creating multimedia content at scale, allowing users to produce cinematic visuals and viral audio within short timeframes. Successive iterations of these models include updates that improve output quality and alignment with user prompts, making AI-generated content more realistic and easier to produce. In March 2026, OpenAI made a drastic move and discontinued Sora as part of a broader business decision, reflecting ongoing changes in the generative AI industry.

CyberWell's research indicates that when these models are used to generate antisemitic narratives, the resulting content does not remain confined to the platform on which it was created. Instead, users rapidly disseminate such content across social media, where it gains visibility and engagement, extending its reach beyond the point of creation.

**Distribution of AI-Generated Content**




** Not all posts in this dataset included visible watermarks or disclosures indicating which AI model was used to produce antisemitic content. Therefore, the following distribution represents content only where the AI-generating system could be determined and verified.*

## Cross-Platform Dissemination of AI-Generated Content

As AI models become more accessible, users adept at evading social media moderation systems increasingly leverage these tools to produce and amplify antisemitic content at scale. For instance, the AI-generated song "Boom, Boom, Tel Aviv" circulated widely across all major social media platforms despite originating from a single creator. The song's lyrics include: "Boom, boom, Tel Aviv. This is what you get for all your evil deeds [...] You brought this upon yourself, it's your time to bleed [...] Humanity never expected good behavior from you Jews" [00:00–00:55]. **"Boom, Boom, Tel Aviv" accounts for a significant share of the dataset and illustrates how users collectively amplify AI-generated antisemitic content to celebrate violence against Jews**. This pattern was particularly evident during the surge in AI-generated antisemitic content identified by CyberWell during Israel's 12-day war with Iran in June 2025.

CyberWell's dataset reveals how individual users can disproportionately shape exposure to AI-generated antisemitic content across social media platforms. For example, the Instagram user "jakegtv" accounted for 5.54% of total posts in the dataset, with the user's content created primarily in the form of fake, AI-generated "news" segments with antisemitic narratives embedded throughout. Despite representing a small share of total posts, this user's content generated 18.59% of all views in the dataset. In addition, many users circulate their content across multiple social media platforms, often embedding AI-generated watermarks alongside their own usernames to maintain attribution. Cases like these demonstrate how individual users can amplify AI-generated antisemitic content beyond the point of creation, rapidly disseminate it, and drive disproportionately high engagement across social media.

## AI Guardrail Limitations and Safety Vulnerabilities

While several AI models implement safety guardrails to restrict hate speech and incitement, research suggests these systems remain inconsistent in identifying harmful content at the outset.[17] A study published by researchers at the University of Pennsylvania found that **AI moderation systems demonstrate uneven performance across demographic categories, including inconsistencies in detecting implicit antisemitism, leaving certain communities more vulnerable to online harm.**[18]

Research on AI safety also highlights the structural vulnerabilities that currently exist in model guardrails. Techniques such as "fine-tuning", jailbreak methods, and prompt injection detection can weaken the previously embedded safety mechanisms when models are

[17] Neil Fasching and Yphtach Lelkes, "Model-Dependent Moderation: Inconsistent Detection Across LLM-based Systems", *Findings of the Association for Computational Linguistics: ACL 2025*, (July 2025), https://aclanthology.org/2025.findings-acl.1144.pdf.
[18] "Model-Dependent Moderation: Inconsistent Detection Across LLM-based Systems".

adapted to specific applications. Although natural language processing (NLP) and machine learning have improved automated hate speech detection, these systems continue to face challenges in preventing the large-scale production of hateful content.[19] As a result, once these outputs are distributed across social media platforms, the gap between model guardrails and platform trust and safety enforcement becomes more pronounced.

## Amplification of AI-Driven Antisemitism on Social Media

AI companies play a central role in shaping the ability to scale the presence of AI-generated antisemitism on social media. By lowering the technical barriers to produce antisemitic content, these tools allow antisemitic narratives to be quickly repackaged in new formats and widely disseminated. The adaptability of coded and implicit language — often presented under the guise of humor — combined with the speed of cross-platform dissemination, creates a multiplier effect in which antisemitic tropes shift in form while generating higher engagement.

The "promised 3,000 years ago" trend, which mocks and attempts to delegitimize the Jewish historical connection to the land of Israel by portraying Jews as inherently entitled and greedy, illustrates this trajectory. After initially circulating in the comment sections on social media, CyberWell discovered that, beginning in July 2025, users leveraged the generative-AI model Veo3 to transform this antisemitic trope into video content. The content then reappeared in comment sections where it was used to mock victims of violent antisemitic attacks [see Appendix].

This progression demonstrates how coded antisemitic narratives migrate from text-based spaces into multimedia formats once they are adapted to bypass model safeguards. It also demonstrates how AI tools facilitate the creation and normalization of antisemitic narratives across online platforms. If left unaddressed at an early stage, such content can spread widely to mass audiences, including in overtly violent forms.

## IHRA Classification

CyberWell applied the IHRA Working Definition of antisemitism to classify the analyzed content. Of the eleven examples included in the definition, ten appear in the dataset, with two appearing significantly more frequently than the others: the first and second examples of the IHRA Working Definition (hereafter: "IHRA Example 1", "IHRA Example 2").

[19] Aish Albladi, Minarul Islam et. al., "Hate Speech Detection Using Large Language Models: A Comprehensive Review", IEEE Access 13, (February 2025), www.//ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10848067.

Among these, IHRA Example 2 emerged as the most prevalent, accounting for 75.6% of the dataset. IHRA Example 2 is defined as:

> Making mendacious, dehumanizing, demonizing, or stereotypical allegations about Jews as such or the power of Jews as collective – such as, especially but not exclusively, the myth about a world Jewish conspiracy or of Jews controlling the media, economy, government or other societal institutions.

The prevalence of IHRA Example 2 in the dataset is consistent with CyberWell's previous analyses, which indicate the continued dominance of antisemitic tropes and classic antisemitism on social media. These tropes have since been adapted into AI-generated content.

The second most common category in the dataset was IHRA Example 1, which accounted for 33.2% of the content and is defined as:

> Calling for, aiding, or justifying the killing or harming of Jews in the name of a radical ideology or an extremist view of religion.

This category is significant in the context of AI-generated content, as it underscores how generative AI tools enable the rapid production and amplification of violent narratives in highly engaging formats.[20]

## Common Narratives in AI-Generated Antisemitic Content

CyberWell's analysis reveals three recurring narratives in AI-generated antisemitic content on social media:

1. Depictions of Jews as greedy or money-obsessed
2. Holocaust-related hate speech
3. Event-driven violent rhetoric against Jews

Each of these narratives demonstrates how users employ AI models to amplify and modernize longstanding antisemitic tropes across social media platforms.

## Depictions of Jews as Greedy or Money-Obsessed

These depictions represent the most prevalent narrative in the dataset, appearing in 33.2% of posts.[21] This narrative is often expressed through recurring visual patterns, including depictions of Jews stealing money or chasing coins. Rather than relying on explicit slurs, users leverage AI models to generate fabricated visuals that reinforce longstanding

[20] Any piece of content analyzed in this dataset may include multiple IHRA categories.
[21] This percentage was calculated to include posts that feature other narratives in addition to this one.

antisemitic stereotypes of Jews as entitled, excessively focused on wealth, and inherently driven by greed.

This narrative appears most frequently on TikTok, where users often employed Sora AI to produce such content. Across the dataset, this narrative is also often paired with the "promised 3,000 years ago" trope, which uses humor to suggest that money, property, or even people were promised to Jews thousands of years ago, reinforcing antisemitic claims of Jewish entitlement. In addition, AI-generated content depicting greedy or money-obsessed Jews often incorporates Jewish terminology and music to obscure antisemitic intent and to present the content as neutral, making it more difficult for moderation systems to detect.

## Holocaust Hate Speech

Holocaust hate speech constitutes a recurring narrative in the dataset, appearing in 21.5% of posts and most frequently identified on Meta's platforms. In addition to promoting explicit Holocaust denial, this narrative also includes AI-generated content that distorts or mocks the Holocaust, at times through the use of coded language.

Holocaust-related mockery is the most prominent sub-narrative within this category, reflecting a pattern of ridiculing historical atrocities and normalizing antisemitic discourse. Common examples of Holocaust-related mockery include the use of coded references, such as the "juice box" emoji to dehumanize Jews or phrases such as "6 million pizzas" to refer to the number of Jews killed and evade content moderation. In some instances, AI-generated social media posts that mock the Holocaust also appear in the form of fake Disney-Pixar movie trailers that ridicule Holocaust victims or portray Jews as inherently evil or as enemies.

AI-generated content is also used in posts that justify the Holocaust and glorify Hitler, often appearing in formats such as fabricated Disney-Pixar-style movie trailers designed to appeal to younger audiences. Together, these tactics demonstrate how AI enables the repackaging of Holocaust hate speech into shareable formats that drive engagement.

## Event-driven Violent Rhetoric Against Jews

Event-driven violent rhetoric against Jews represents the third most common narrative in the dataset, appearing in 21.2% of posts and most frequently identified on TikTok. The data shows that users produce content portraying this narrative in the immediate aftermath of antisemitic attacks or geopolitical developments, leveraging AI-generated visuals or audio to glorify or justify violence against Jews.

In several cases, users produce AI-generated content that praises perpetrators of violence or denies the occurrence of antisemitic incidents. Across this narrative, AI-generated

antisemitic songs and highly realistic or graphic AI-generated imagery are amplified and drive higher user engagement.

## Insights and Patterns

CyberWell identified several recurring patterns in how users produce and disseminate AI-generated antisemitic visuals and audio across social media. Regardless of format, the underlying antisemitic narratives remain consistent, with users recycling longstanding antisemitic tropes and adapting them to contemporary contexts. What is notable throughout these insights is the ease of producing such content, enabled by the accessibility of generative AI tools.

### Classic Antisemitic Tropes Adapted for AI-Generated Visual Media

CyberWell found that users frequently adapt longstanding antisemitic tropes — traditionally expressed in classic text or static imagery — into AI-generated video content on social media. These tropes often take the form of caricatures depicting Jews with exaggerated features, such as oversized noses. They also commonly align with the most dominant narrative in the dataset — depictions of Jews as greedy or money-obsessed.

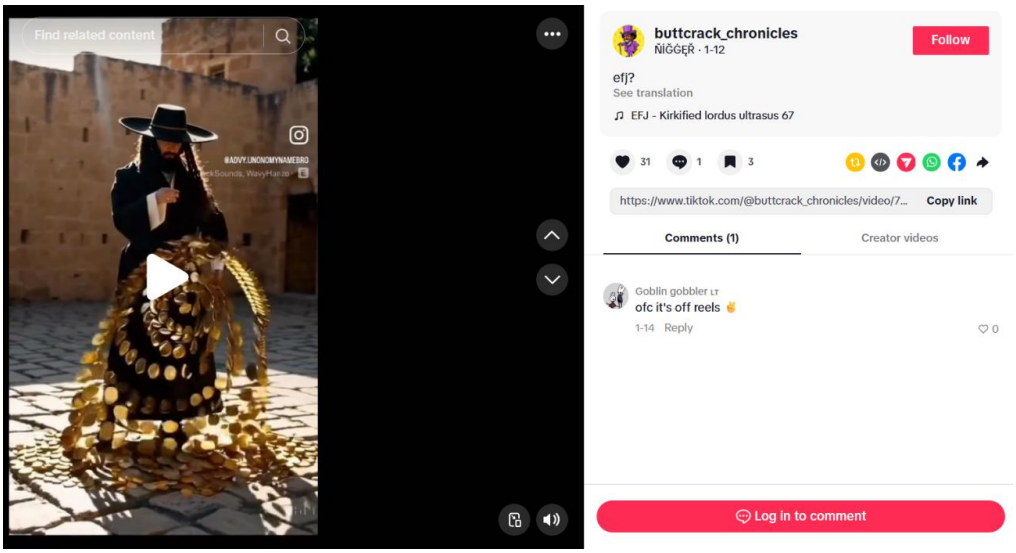
Users increasingly adapt text-based claims that portray Jews as controlling and fixated on wealth into synthetic videos that visually reinforce these stereotypes in contemporary formats. In some cases, AI-generated visuals draw on historically rooted antisemitic tropes, such as blood libel or dehumanizing comparisons that portray Jews as animals — narratives that were widely used during World War II to justify the persecution of and genocide against Jews.[22]

Example 1

https://www.tiktok.com/@buttcrack_chronicles/video/7594337543483591966

This TikTok post **uses AI visuals to amplify antisemitic tropes related to alleged Jewish obsessions with money**. The user depicts an individual resembling an Orthodox Jewish man manipulating gold coins and making them fly into the air [00:00-00:09]. In the background, audio plays with the lyrics: "[...] Fuck Jews [...]" [00:00-00:09]. This video combines explicit hate speech with classic antisemitic tropes to reinforce dehumanizing representations of Jews and actively normalize hostility towards them.

[22] Alexis Chapelan, "Repulsiveness and Dehumanisation," in *Decoding Antisemitism. Postdisciplinary Studies in Discourse* (Palgrave Macmillan, October 2024), https://link.springer.com/chapter/10.1007/978-3-031-49238-9_5.




Example 2

https://x.com/AxisFB/status/1976981101626466558

In the context of the 2025 NYC Mayoral Election, this post on X uses Grok to promote the **antisemitic** blood libel **narrative that falsely accuses Jews of killing Christian children and using their blood for ritual purposes**. The user replies to a post featuring a photo of mayoral candidate Zohran Mamdani eating with Orthodox Jews during his campaign and **uses AI to transform the original image into a fabricated video**. The Grok AI-generated video depicts a distressed and crying child serving a pitcher of blood to Orthodox Jewish men and Mamdani as they laugh [00:02-00:05]. The user's accompanying caption reinforces this imagery, stating that "They celebrated shortly after with goyim blood", invoking the trope that Jews consume the blood of non-Jews in ritual contexts.


Example 3

https://x.com/povEVM/status/2018651694431904158

This post on X uses **Grok Imagine** to produce an AI-generated video that dehumanizes Jews. The video, titled "AIRDROP GOYIM", **brings the antisemitic Happy Merchant caricature to life** by depicting it with erratic behavior and exaggerated facial features, including an oversized nose. The Happy Merchant is a well-known antisemitic meme used to portray Jews as greedy. In this example, the user combines AI-generated videos of the Happy Merchant with a caption that incorporates cryptocurrency-style language referencing "$GOYIM". While "goyim" is a neutral Jewish term referring to non-Jews, it is used here in a mocking and derogatory way to reinforce antisemitic stereotypes.


Example 4

https://www.tiktok.com/@myaiworld.com/video/7594487205649763639

This **TikTok** post uses AI-generated imagery to dehumanize Jews by portraying them as animals and reinforce longstanding antisemitic stereotypes that Jews are greedy and dishonest. The video depicts a squirrel dressed as a religious Jewish man, wearing a black hat and sidelocks, stealing a bag of money from a bank [00:00-00:15]. The use of the Jewish song "Hava Nagila" throughout the clip further amplifies the antisemitic framing. **TikTok includes a watermark beneath the user's caption, indicating that the content is AI-generated.**


## AI Use in the Exploitation of Violent Attacks against Jews

Across the dataset, users often respond to violent, real-world attacks against Jews by producing highly realistic AI-generated content that fabricates events, celebrates violence, mocks victims, or glorifies perpetrators. This recurring pattern extends harmful narratives beyond real-world incidents and into imagined scenarios, contributing to the normalization of violent rhetoric. The findings also point to a broader dynamic in which AI is used to produce emotionally charged content that provokes shock and drives user engagement.

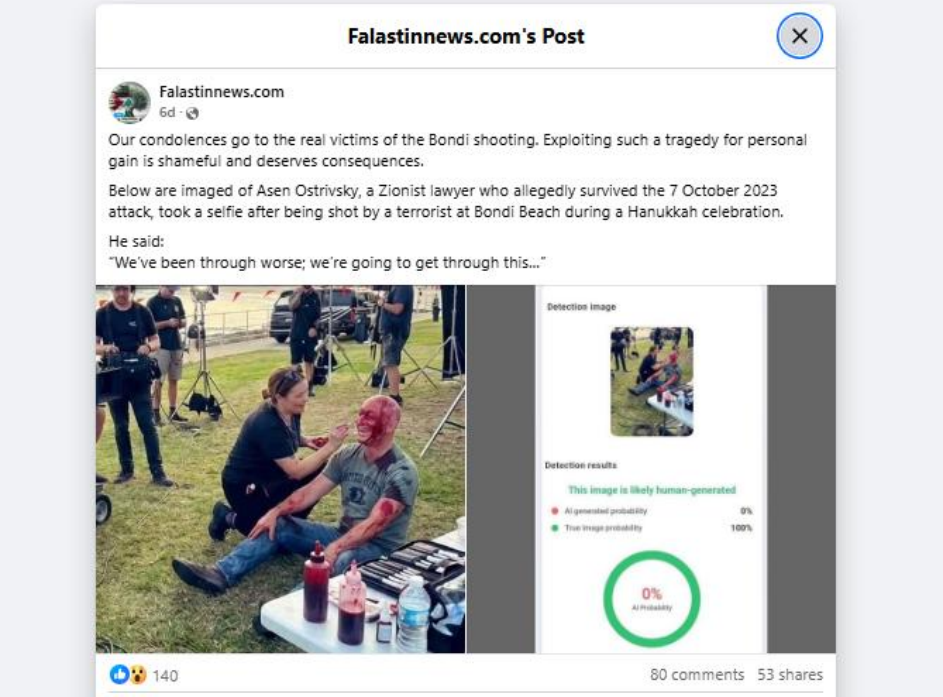
In the immediate aftermath of antisemitic attacks, some users produce AI visuals and audio to blame Jews for orchestrating violence against themselves. This finding aligns with CyberWell's newly identified form of antisemitism — Conspiratorial Self-Victimization — which this analysis shows to sometimes be spread by AI-generated content.

Example 1

https://www.facebook.com/100067914394282/posts/1183789910561478

This Facebook post uses generative AI to spread misinformation about the Jewish victims of the 2025 Bondi Beach Hanukkah shooting in Sydney, Australia. The post features a fabricated AI-generated image depicting one of the victims, Arson Ostrovsky, seated on the

ground smiling while a woman appears to paint blood on his face. The image falsely suggests that his injuries were staged. The post includes an additional screenshot intended to cast doubt on the likelihood that the image is AI-generated. The accompanying caption builds on these false claims, alleging that Jews like Ostrovsky are "[...] exploiting such a tragedy for personal gain [...]".




Example 2

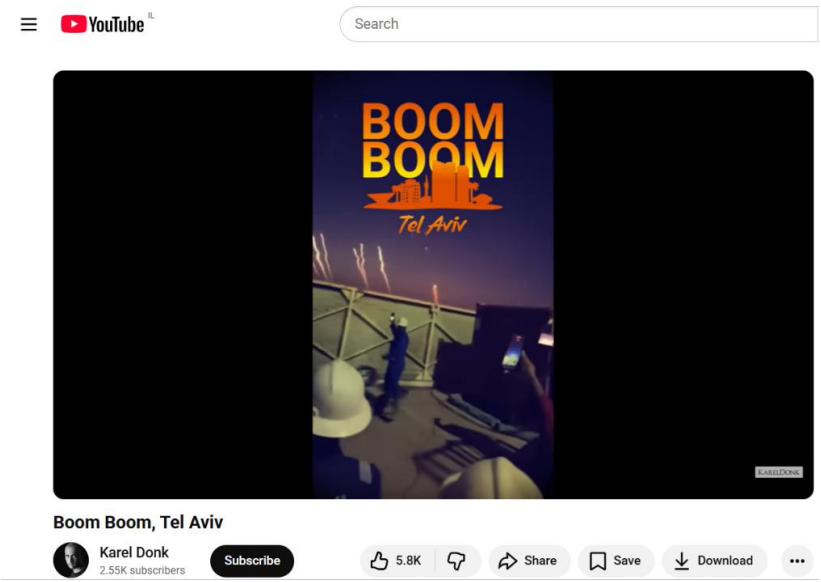
https://x.com/j_jarar/status/1925646210150384012

This Arabic-language post on X uses AI to **visually and narratively echo the manifesto of Elias Rodriguez, the perpetrator of the** Capital Jewish Museum shooting **in Washington, D.C.** The video combines emotionally charged imagery, such as mass civilian casualties, shrouded bodies, and burning urban environments, with a first-person monologue that **frames violence against Jews as an ethical response to Israeli actions**. In this case, AI is used to legitimize antisemitic violence by embedding it within political grievances, presenting the justification of harm as ethically grounded.


Example 3

https://www.youtube.com/watch?v=wbDeBDyL0XE

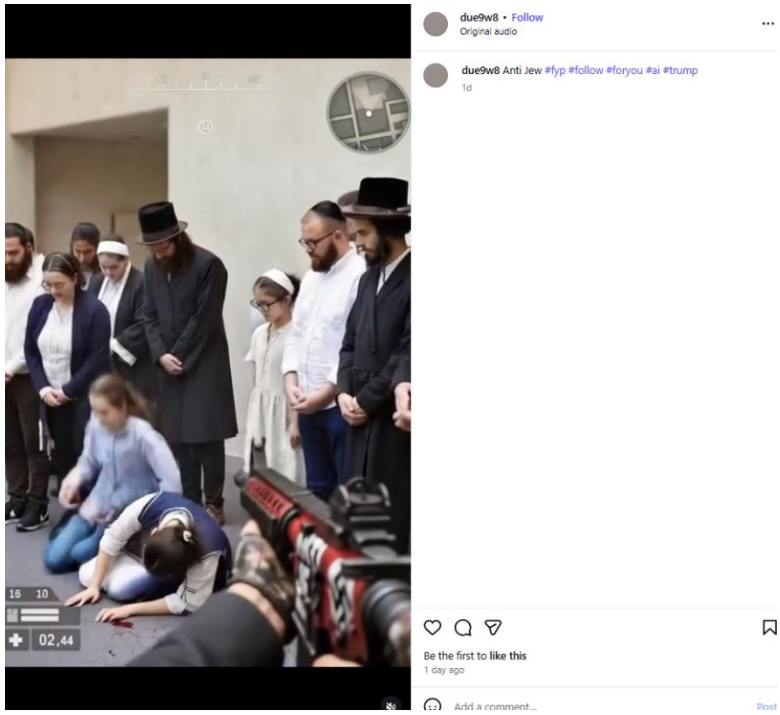
This YouTube video is set to the **AI-generated song**, "Boom, Boom, Tel Aviv", created using **Suno**. The video features images of destruction in Israel during the 12 Day War with Iran, while the audio **celebrates and justifies the death of Jews**. The song's lyrics include: "Boom, boom, Tel Aviv. This is what you get for all your evil deeds [...] You brought this upon yourself, it's your time to bleed [...] Humanity never expected good behavior from you Jews" [00:00–00:55].



Example 4

https://www.instagram.com/reel/DUbcPKoEb_3

This Instagram post features an AI-generated image of a shooter holding a firearm with a swastika. The shooter intentionally fires multiple rounds of bullets, killing a group of visibly Jewish men, women, and children. The accompanying caption **normalizes violence against Jews** and seeks to **amplify its reach through social media algorithms, including hashtags such as:** "Anti-Jew #fyp #follow #foryou #ai".



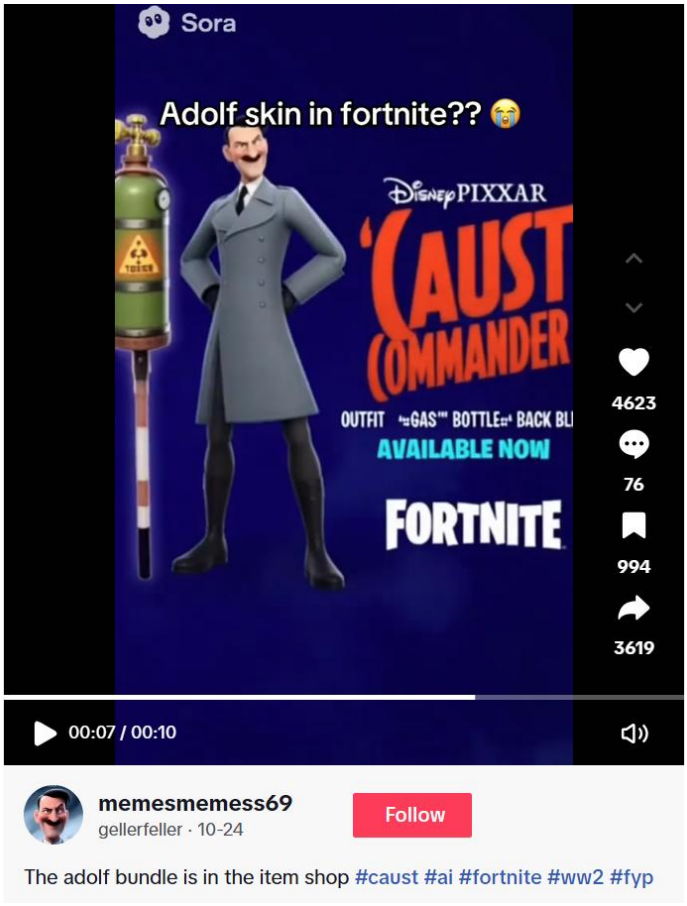

## Youth-Oriented Framing in AI-Generated Antisemitic Posts

CyberWell observed a recurring pattern in which users package AI-generated antisemitic content in formats designed to appeal to younger audiences. The most common examples include **fabricated Disney Pixar-style movie trailers and gaming-related audio clips** that promote Holocaust-related mockery, antisemitic conspiracy theories, and hate speech targeting Jews. These posts are often presented as **humorous and child-friendly clips that blend popular trends with hateful narratives. They frequently include disclaimers,** such as "satire" or "dark humor" to obscure intent and evade moderation. CyberWell observed that, in many cases, this type of content appears alongside posts that promote radicalization, violence, and the sexualization or grooming of minors. For example, AI-generated movie trailers for the fabricated Disney-Pixar movie titled "Caust" appear next to posts that encourage children to carry out school shootings and 9/11-style attacks, or that normalize indecent behavior involving minors, including references to Jeffrey Epstein's grooming of minors. Together, these posts increase the risk of exposure among vulnerable audiences.

Example 1

https://www.tiktok.com/@memesmemess69/video/7564830979115076871

This TikTok post with over 66,500 views, 4,623 likes, and 3,619 reposts uses Sora AI to create a fabricated Disney-Pixar-style trailer titled "'CAUST COMMANDER", a reference to the Holocaust. The post portrays Adolf Hitler in a playful, stylized manner while depicting him killing those around him. The video makes light of the Holocaust and the mechanisms used to exterminate Jews by presenting them in a gamified, commercialized format, including the promotion of fake merchandise such as Zyklon B gas, themed outfits, and "back bling".

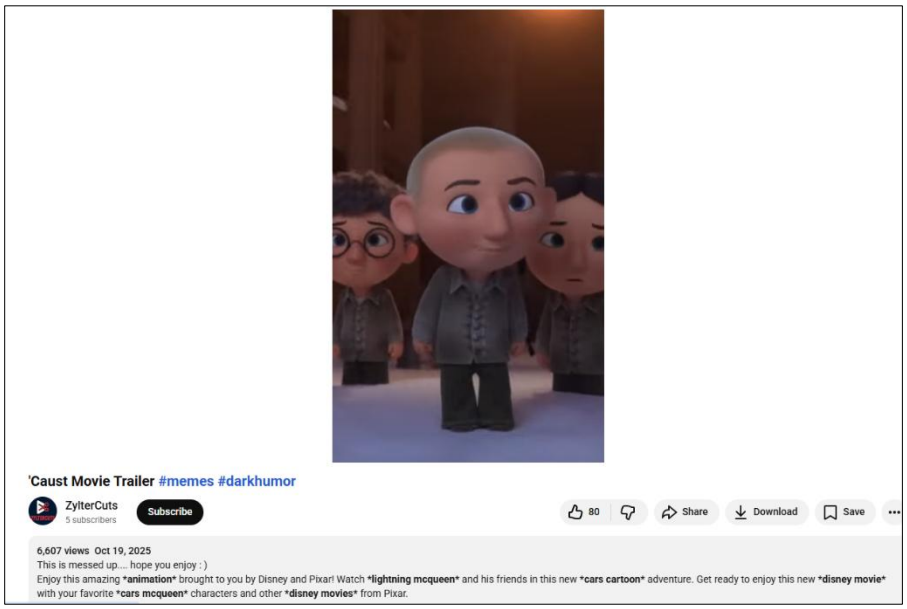



Example 2

https://www.youtube.com/watch?v=j1URtI5aNcY

Similarly, this YouTube video illustrates how AI-generated antisemitic content migrates across social media platforms. In this case, the user employs Sora AI to create a Disney-Pixar-style trailer for a fabricated movie titled "'Caust". Set in a concentration camp, the trailer portrays Adolf Hitler in a lighthearted manner while following a group of Jewish child prisoners attempting a dramatic escape. By presenting the Holocaust in a playful, animated format, the video turns atrocity into entertainment and diminishes the gravity

of Jewish suffering. Hashtags such as “#memes #darkhumor” in the title further reinforce this framing, signaling humor while masking harmful intent.

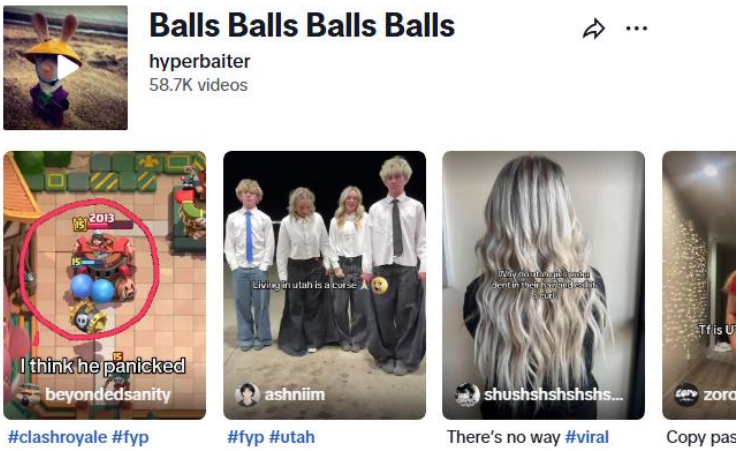



Example 3

https://www.tiktok.com/music/Balls-Balls-Balls-Balls-7578889500672199437

This AI-generated audio clip was associated with nearly **60,000 TikTok videos** and **generated over 50 million views across the first 50 posts alone.**[23] The audio promotes the antisemitic conspiracy theory that “9/11 was done by Israel”, assigning collective blame to Israelis for a mass-casualty attack while deflecting responsibility from the actual perpetrators. This audio clip was frequently **paired with gaming-related content,** including platforms such as Roblox, Clash of Clans, Clash Royale, and Minecraft, embedding antisemitic conspiracy theories within youth-oriented spaces. Rather than remaining confined to a single upload, the audio clip also appeared across multiple distinct songs produced by the same users, further amplifying its reach and engagement.

[23] This audio is not included in CyberWell’s dataset for this report. Instead, a number of individual videos using this audio were included and analyzed according to CyberWell’s methodology.



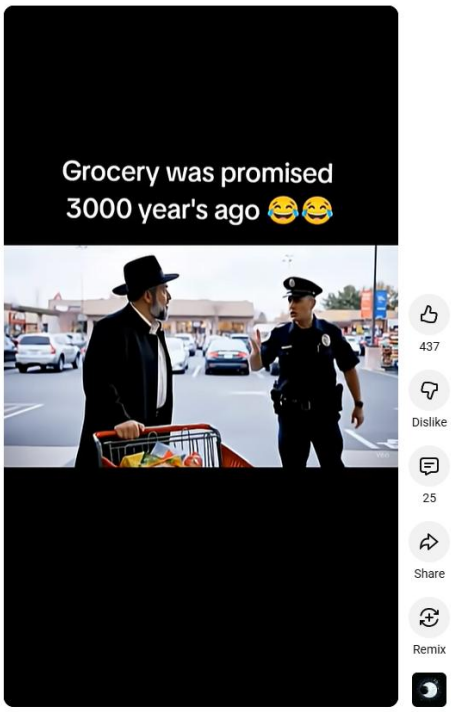

## Cultural Mockery and the Use of Humor in AI-Generated Content

CyberWell identified the use of **generative-AI to mock Jewish culture, language, and religious identity** as one of the most recurring patterns across the dataset. Rather than relying on explicit hate speech, many posts incorporate recognizable Jewish symbols, music, or terminology, reframing them in dehumanizing or derisive contexts. A significant portion of the dataset includes AI-generated antisemitic posts paired with the Jewish song **"Hava Nagila"** to accentuate and mock Jewish identity. These forms of cultural mockery heavily rely on humor and exaggerated caricatures to disguise antisemitic intent. **By framing the content as satire, users rely on implicit cues to reinforce longstanding stereotypes about Jews.**

Example 1

https://www.youtube.com/shorts/RI6Dkyxhz9s

This YouTube short uses **Veo3** to invoke the **"promised 3,000 years ago" trope, an antisemitic narrative that frames** Jews as entitled and greedy while mocking their historical and religious connection to the land of Israel. The video, titled, "Grocery was promised to him 3000 years ago #funny #promisedland #isreal #usa #antisemetic #meme", depicts an Orthodox Jewish man leaving a grocery store with a full shopping cart. When a police officer confronts the man for not paying, he responds, "These groceries were promised to us 3,000 years ago" [00:00-00:08]. **While framed as humor, this AI-generated clip reinforces stereotypes that portray Jews as entitled individuals who falsely claim ownership without justification.** The use of the song **"Hava Nagila"** in the background further emphasizes the character's Jewish identity and reinforces this antisemitic narrative.




Example 2

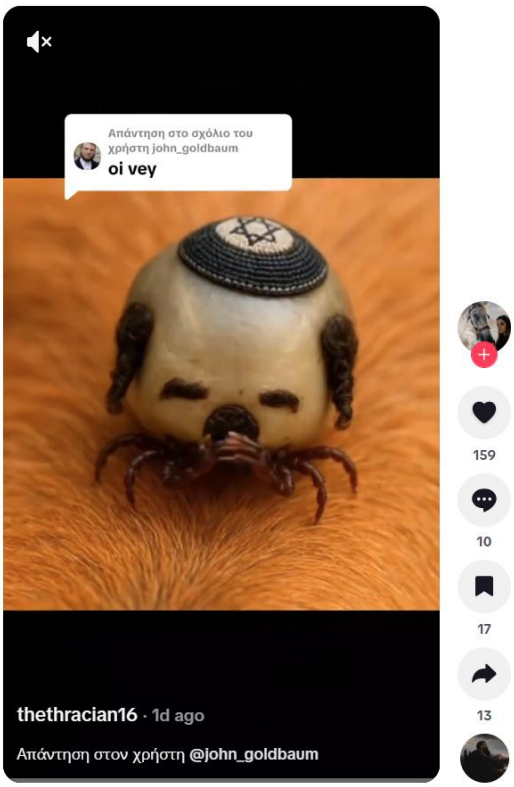
https://www.facebook.com/darinmae.mahinay.96/posts/pfbid0UAJ7Kzjuu6eJzhLUEgP7pixxGfQrzAJuoz17992LstHu6HfYSwYJ9zUjKcMWUedhl?comment_id=1553674415844729

This Facebook comment uses AI-generated imagery **to invoke antisemitic stereotypes depicting Jews as greedy, dishonest, and historically entitled.** In response to a post praising U.S. President Donald Trump's military action in Venezuela, the user shares an AI-generated image of President Trump with stereotypical Jewish markers, including a yarmulke and sidelocks. The user's caption further reinforces antisemitic tropes when stating: "'If I don't steal it, someone else will' Venezuela's oil was promised to Donald Trump 3,000 years ago". This language invokes longstanding antisemitic claims of entitlement and theft.

Example 3

https://www.tiktok.com/@thethracian16/video/7594482241493929238

This TikTok post uses AI to **appropriate Jewish cultural symbols and terminology in an antisemitic context**. The video features a louse dressed in traditional Jewish markers, including sidelocks and a yarmulke bearing a Jewish Star of David. Throughout the video, a soundtrack repeats the following lyrics: “When I say oy, you say vey, Oy vey, oy vey [...]”. “Oy vey” is a Yiddish expression commonly used within Jewish communities to express dismay or exasperation. In this context, however, this term is used to mock Jewish identity and to associate Jews with parasitic imagery, such as lice, reinforcing dehumanizing narratives.




## AI-Generated Antisemitism Disguised as Neutral or Factual Content

Throughout the dataset, users leverage generative-AI to present antisemitic content as neutral reporting or factual analyses of current events. These posts reflect a broader and increasingly **concerning pattern of blending AI-generated imagery with real-world footage to create a false sense of credibility and to blur the line between fabricated and authentic content**. Many of these posts incorporate satire, coded language, and indirect forms of Holocaust denial and distortion to avoid appearing overtly hateful. As a result, posts may appear as credible sources and are less likely to be detected.

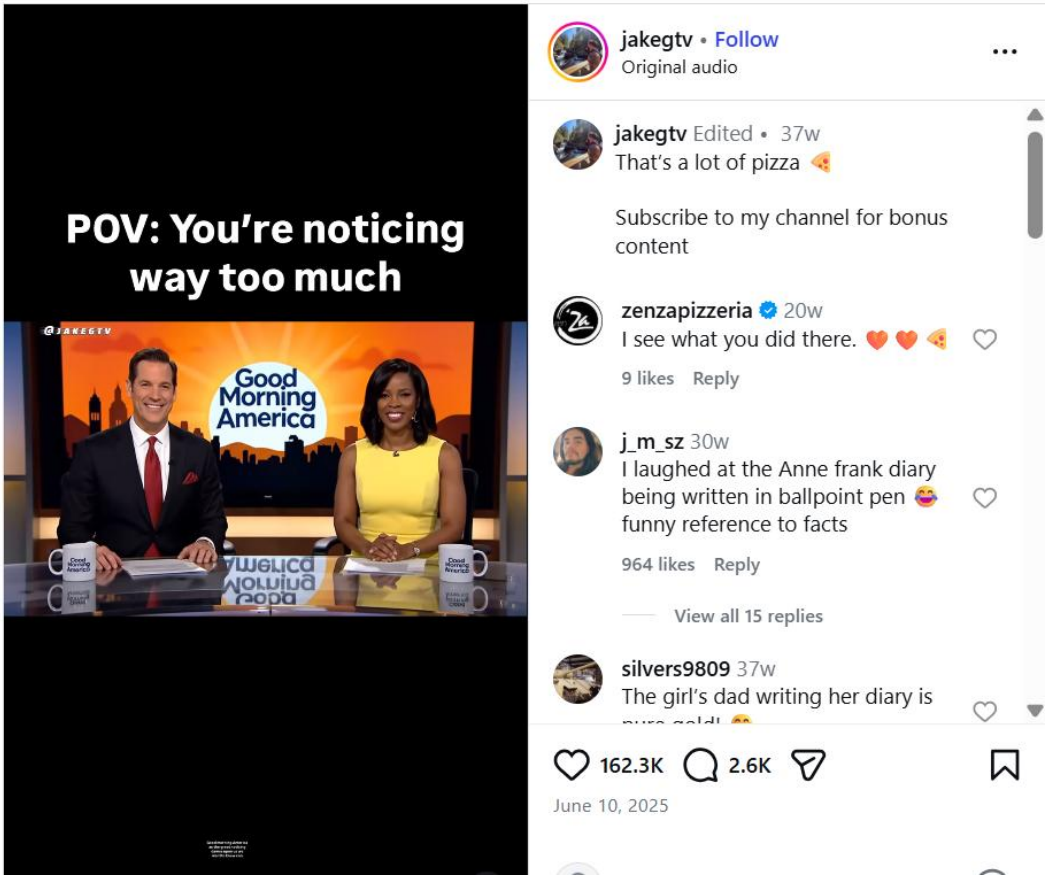
Example 1

https://www.instagram.com/p/DKubp8FxYdo/

This Instagram post, which received **162.3K likes,** presents an **AI-generated news segment that denies and mocks the Holocaust**. Framed as an investigative commentary, the fabricated "Good Morning America" segment promotes the false claim that the Holocaust is a fictitious narrative imposed on society. The video centers on a hypothetical question and features interviews with employees at pizza restaurants, asking how long it would take them to bake **"6 million pizzas"**, a coded reference to the six million Jewish victims of the Holocaust.

Throughout the segment, the AI-generated presenters reinforce Holocaust denial and distortion by characterizing historical facts as "lies" and portraying genocide as implausible. The video's opening further advances this narrative, alleging Jewish manipulation of historical memory, with overlaying text that reads: "POV: You're noticing way too much", implying concealed knowledge of Jewish conspiratorial narratives.

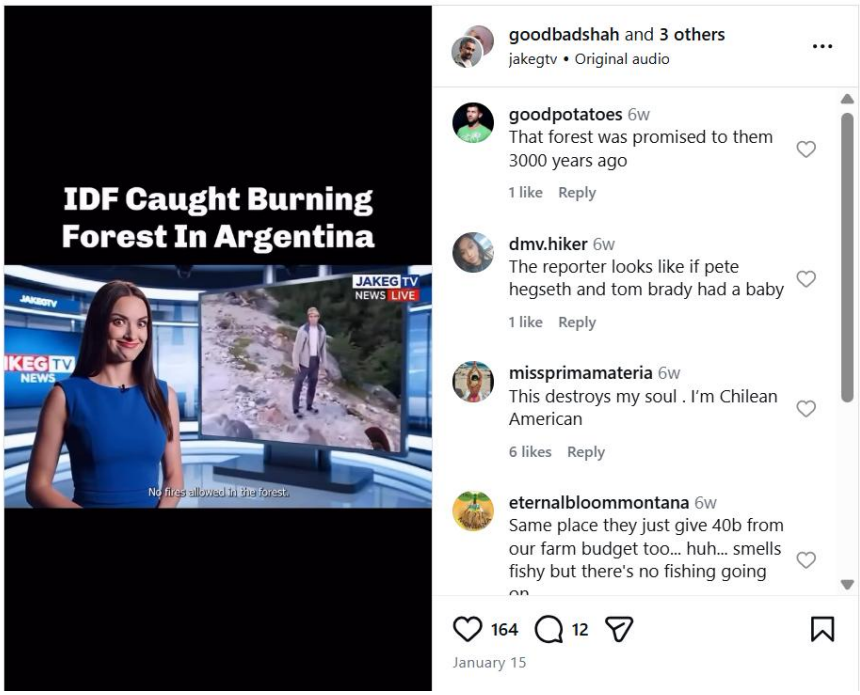
CyberWell's dataset further indicates that the account associated with this post, "jakegtv", **promotes the sale of antisemitic merchandise** through links in his Instagram bio. This combination of antisemitic content and monetized branding allows users such as "jakegtv" to profit directly from engagement with their hateful posts.



Example 2

https://www.instagram.com/p/DTgi_n2kf53/

This Instagram post **blends authentic imagery with AI-generated elements to reinforce antisemitic narratives**. The user shares an AI-generated video originally created by the user mentioned above, "jakegtv". The video **scapegoats Jews and Israelis for the Patagonia wildfires in Argentina** and ties these claims to broader conspiracy theories suggesting Jews seek territorial expansion and economic control. **The video also incorporates real-world footage of an individual alleged to be Israeli starting a fire in Patagonia. This footage is presented in a way that lends credibility to these misleading and conspiratorial claims.**




## Use of Disclaimers to Obscure Intent and Evade Moderation

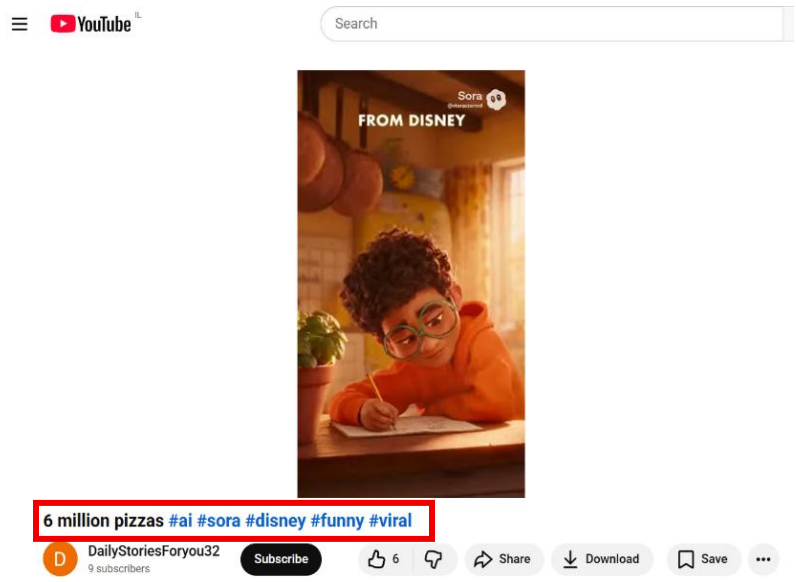
Several posts in the dataset include disclaimers such as "#satire", "#darkhumor", or "for educational purposes" in their titles or descriptions. **These labels attempt to position content as humorous or informational, despite the presence of antisemitic narratives and references**. By emphasizing stated intent through disclaimers, users obscure the nature of the content and reduce the likelihood of enforcement.[24]

[24] Natalia Stanusch et. al, "AI-Generated Algorithmic Virality: How Synthetic AI Imagery and Agentic AI accounts try to game TikTok and Instagram", *AI Forensics* (July 2025): 12, www.aiforensics.org/work/gen-ai-slop.

Example 1

https://www.youtube.com/watch?v=qctoMfbNEyE

This AI-generated YouTube video, titled “6 million pizzas #ai #sora #disney #funny #viral”, uses coded antisemitic language to deny the plausibility of the Holocaust by comparing the murder of six million Jews to the act of making pizza. The inclusion of “#funny” in the video’s title functions as a disclaimer of humorous intent, framing the video as satire while obscuring its underlying antisemitic narrative.




Example 2

https://www.instagram.com/p/DQrZxUCjWUB/

This Instagram post promotes antisemitic stereotypes that depict Jews as greedy and money-obsessed. The caption includes multiple hashtag-based disclaimers, such as “#hilarious [...] #meme [...] #comedy [...] #humor [...]”, which attempt to frame the content as humorous. These disclaimers appear despite the presence of antisemitic rhetoric and stereotypes, functioning to downplay the nature of the content and reduce the likelihood of moderation.


## Cross-Account Distribution of Content with Identifiable Watermarks

A recurring feature across the dataset is the presence of standardized watermarks in AI-generated antisemitic content. While some posts include only user-specific watermarks, others combine the logo of the AI model used to generate the content with a creator handle, such as "Sora @username". These watermarks identify the origin and authorship of the content and remain embedded within the videos as they are reposted across social media platforms. They further enabled CyberWell to trace the dissemination of specific antisemitic posts across multiple accounts and platforms, providing an indication that antisemitic content continues circulating despite enforcement actions taken against the original upload. Platforms could potentially use these watermarks to identify and action shared content when the original is determined to violate policy.

CyberWell identified at least three separate TikTok accounts that shared antisemitic content bearing the watermark "Sora @fakejuice". These examples demonstrate how AI-generated videos containing antisemitic narratives continue to circulate across platforms, even after enforcement actions are taken against individual accounts.

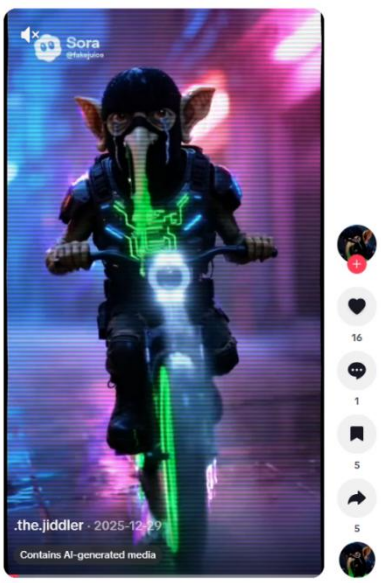
The two TikTok posts below illustrate this pattern:

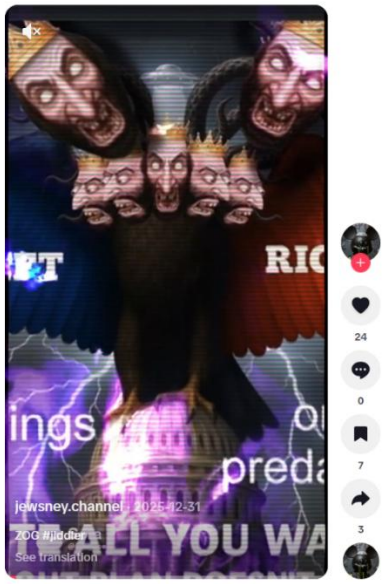
1. https://www.tiktok.com/@.the.jiddler/video/7589220434462149902

   On the left, a user (@.the.jiddler) shares an AI-generated video that promotes Holocaust-related mockery and distortion, including references to the gas chambers and claims that falsely minimize the number of Jewish victims murdered during the Holocaust [00:08-00:22].

2. https://www.tiktok.com/@jewsney.channel/video/7590019107303591199

   On the right, another user (@jewsney.channel) shares an AI-generated video bearing the same Sora watermark (unfortunately this screen grab does not include the image). The video takes the form of a fabricated Pixar-style trailer that demonizes Jews and advances the antisemitic "Zionist Occupied Government (ZOG)" narrative, which alleges Jewish control over political systems.




## Social Media Platform Policies

Social media platforms maintain policies addressing both AI-generated content and hate speech. However, these policies often operate separately, with AI policies emphasizing transparency and disclosure, while hate-related policies underscore enforcement against harmful content. This section examines how these policy approaches function in practice and identifies gaps across content creation, moderation, and algorithmic amplification systems in their handling of AI-generated antisemitic content.

### Platform Policies Governing AI-Generated Content

As AI becomes more widely integrated into online content creation, some social media platforms are introducing policies that explicitly address the use of synthetic and AI-generated media. These policies primarily focus on authenticity and transparency in order to mitigate the spread of misinformation or deceptive content:

- TikTok: Under its Integrity and Authenticity policy, TikTok includes a clause titled "Edited Media and AI-Generated Content (AIGC)", which explicitly prohibits, **"Any**

content that breaks our Community Guidelines, including those on impersonation, misinformation, and hate speech, even if it's AI-generated".

The policy adds the following provision: "We require you to label AIGC or edited media that shows realistic-appearing scenes or people. This can be done using the AIGC label, or by adding a clear caption, watermark, or sticker of your own [...] We do not allow content that shares or shows fake authoritative sources or crisis events, or falsely shows public figures in certain contexts".

- X: Under its Synthetic and Manipulated Media policy, X includes a provision on "Inauthentic Content", which states: "You may not share inauthentic media, including, manipulated, or out-of-context media that may result in widespread confusion on public issues, impact public safety, or cause serious harm ('misleading media')".

  On March 3, 2026, X also introduced a new provision addressing AI-generated content related to armed conflict, stating: "[...] Users who post AI-generated images of an armed conflict – without adding a disclosure that it was made with AI – will be suspended from Creator Revenue Sharing for 90 days. Subsequent violations will result in a permanent suspension from the program [...]".

- Meta and YouTube: Meta and YouTube maintain provisions outside their core policies and Community Guidelines that require users to disclose when content is created using AI. Meta also states that content violating its Community Standards will be removed, regardless of whether or not it was created with AI.

### Limitations of Current AI-Generated Content Policies

Meta and TikTok integrate AI-generated content into their existing policy frameworks, an approach that appears to contribute to more consistent enforcement outcomes. This is reflected in their higher removal rates compared to X and YouTube.

However, TikTok maintains the most comprehensive standalone policy on AI-generated content, explicitly linking such content to violations related to hate speech, harassment, and violence. This may help explain TikTok's comparatively high removal rate in this dataset, which, at 88.2%, was the highest among all platforms. Previous research conducted by CyberWell indicates that enforcement improves when specific forms of harmful content are explicitly addressed in platform policies. This distinction is critical as disclosure requirements alone are insufficient in addressing how AI is used to produce and amplify hateful or abusive content.

This distinction is important because most platform policies governing AI-generated content focus on transparency, authenticity, and privacy, rather than on how AI is used to produce and disseminate hateful or abusive content. As demonstrated throughout this report, labeling requirements alone do not prevent the production or circulation of harmful AI-

generated content. In many cases, content can still contain antisemitic narratives while complying with disclosure requirements.

In addition, CyberWell's dataset shows that generative AI facilitates the rapid transformation of antisemitic narratives into multimedia formats that can circulate widely before enforcement occurs. As outlined in the "Insights and Patterns" section of this report, AI-generated content frequently blends authentic imagery with synthetic visuals while incorporating satire disclaimers and coded language to obscure intent. These tactics allow harmful content to remain technically compliant with disclosure requirements while continuing to disseminate antisemitic narratives at scale. Furthermore, platforms such as X often respond to violative AI-generated content by limiting its visibility rather than removing the post entirely. While this approach may reduce exposure of AI-generated antisemitic content in users' feeds, it still allows the material to remain accessible.

## Enforcement of Hate-Related Policies

Although major social media platforms maintain policies prohibiting hate speech and harassment against protected groups, these policies are not consistently applied to AI-generated content. Meta and TikTok explicitly prohibit Holocaust denial and classic antisemitic stereotypes. YouTube addresses this type of content through broader provisions related to harmful stereotypes and denial of violent events. X adopts a similar approach under its Hateful Conduct policies and, like Meta and TikTok, also classifies Holocaust denial as a violation under its Abuse and Harassment policy.

However, these policies all rely on general provisions that do not account for the distinct characteristics of AI-generated content. This gap highlights the need for dedicated community guidelines that explicitly address how AI is used to produce and disseminate hateful content, in line with TikTok's approach.

As demonstrated throughout this report, AI-generated antisemitic content often relies on implicit, coded, or humor-driven narratives that complicate enforcement. Despite the platforms' existing hate-related policies, CyberWell's findings indicate that such content frequently exploits gaps in how these policies are enforced. As AI tools introduce new formats for users to create antisemitic content, platform moderators face increasing challenges in identifying and enforcing violations. This is particularly evident in content that relies on coded language, disclaimers, or synthetic media that blends authentic and AI-generated elements. Examples throughout this report demonstrate how users adapt antisemitic rhetoric into AI-generated formats that obscure intent or exploit the gaps in enforcement. The rise of generative AI on social media therefore exposes the limitations in policies primarily designed to address text-based content or explicit forms of hate speech.

In addition, with the exception of TikTok, content that mocks victims, particularly in Holocaust-related contexts, may fall outside the scope of bullying and harassment policies

when no identifiable individuals are targeted. This pattern is evident in AI-generated antisemitic content where digitally created characters are used to depict and mock Holocaust victims. This gap reflects limitations in how existing policies are applied to AI-generated content and highlights inconsistencies in enforcement.

## Content Amplification and Algorithmic Recommendations

Although platforms removed 63.5% of the posts in this dataset, the high levels of engagement with AI-generated antisemitic posts suggest that **hate-related enforcement mechanisms are not keeping pace with how quickly AI-generated content spreads**. This pattern is most visible on TikTok, which accounted for the largest share of AI-generated antisemitic posts. Despite this, TikTok maintains policies intended to limit the spread of violating content.

Under its Hate Speech and Hateful Behavior policy, TikTok specifies that the following content is ineligible for the **For You Feed (FYF)**: "Stereotypes, generalizations, insinuations, or statements that may demean or undermine the inclusion of protected groups or identities".

Since TikTok's platform architecture relies on algorithm-driven recommendations through the *For You Feed* feature, harmful content can quickly accumulate high levels of views and engagement, potentially outpacing enforcement measures. As noted earlier, CyberWell observed cases in which content removed from one account reappeared through other users who reposted or repackaged the same material. The repeated appearance of watermarks and usernames linked to previously removed accounts illustrates how AI-generated antisemitic content continues to circulate even after the original source is removed. **These patterns demonstrate how engagement algorithms on feeds can amplify harmful content before enforcement occurs.**

As a result, AI-generated antisemitic content, including hate speech and violent material, can reach large audiences, particularly younger users who are highly active on social media, before platform intervention. This raises concerns about the normalization and spread of such content.

## AI System Policies and Safeguards

As generative AI tools continue to become more widely accessible, the policies and safeguards implemented by AI companies play an increasing role in how these technologies are used. Many of these companies maintain both user-facing safety policies and internal safeguards designed to prevent incitement of violence and the creation of harmful content.

The following section examines these safety mechanisms and their potential limitations.

## AI System Policies and Their Limitations

Major AI companies, including OpenAI, Suno, and Google publish safety and usage policies governing their systems. These policies generally restrict the creation of content that includes hate speech, violence, harassment, and misleading or deceptive material. For example, OpenAI prohibits the use of its tools for "hate-based violence". Suno similarly restricts content that promotes "hate speech", including "discrimination, hate, or violence based on race, religion, gender, sexuality, or other protected characteristics". Google, which develops the Veo AI model, also prohibits the generation of "hatred or hate speech", with limited exceptions for certain contexts. Across these systems, the use of AI models to generate content that incites violence or hatred is explicitly prohibited.

These usage and safety policies reflect a shared commitment among AI companies to restrict harmful and hate-related content. **However, their policies remain broad in scope, relying on general terms such as "hatred", "hate speech", and "false information"** without clearly defining how these harms manifest in practice. Additionally, if they do exist, transparency reports and disclosures as to how AI companies implement these policies and guardrails in practice are by and large lacking. Existing reporting frameworks tend to focus more heavily on addressing government requests and legal compliance, rather than on how companies moderate harmful content and enforce safety policies in practice. As a result, gaps in interpretation and enforcement persist, increasing the likelihood that generative AI models can still be used to generate harmful content despite existing safeguards.

In addition, with the exception of provisions in Suno's guidelines, AI company user policies often lack clear definitions of protected categories and do not specify how hate may take shape in relation to identities such as religion, ethnicity, race, sexual orientation, or gender. Clearly defining protected categories provides a foundation for consistent interpretation and enforcement, reducing ambiguity in how harmful content is identified and addressed. **Furthermore, current policies do not account for how harmful narratives may emerge through culturally specific, implicit, or coded forms**. This gap is particularly significant in the context of AI-generated antisemitic content, which often appears in indirect and nuanced forms that evade detection.

Finally, the use of exceptions, as seen in Google's policies, may result in inconsistent interpretation and application of restrictions in practice. While user policies establish important baseline safeguards, their current scope and definitions leaves room for harmful content to be generated through evasive means.

## Safeguards in Generative AI Systems

Safety guardrails are typically embedded during the development and deployment stages of AI systems to reduce the likelihood of generating harmful or violent content. During the training phase, these safeguards may include processes such as reinforcement learning

from human feedback (RLHF), the use of curated training datasets, and output filtering mechanisms.[25] For example, OpenAI incorporates external red teaming and risk assessments as part of its process for establishing safety protocols.

Additional safeguards are often applied during the deployment stage, when AI systems interact directly with users. At this stage, monitoring systems evaluate model outputs and restrict responses that violate safety policies.[26] Prompt filtering plays a key role in this process, using natural language processing (NLP) techniques to detect keywords or patterns associated with harmful content. AI systems also employ prompt moderation to prevent user inputs that seek to generate violent or hateful content. When such cases are identified, AI systems may refuse to generate a response or redirect users toward safer alternatives instead.[27] Together, these guardrails function as layered controls designed to prevent the generation and dissemination of harmful outputs before user interaction.

## Limitations of AI System Safeguards

Despite ongoing improvements in AI safety measures, existing guardrails continue to have limitations in fully preventing harmful outputs. **Techniques such as prompt manipulation and jailbreaking can circumvent built-in safety guardrails,** enabling the generation of outputs that would otherwise be restricted.[28] In addition, automated moderation systems often **struggle to detect more subtle forms of harmful content, including coded language or culturally specific references.** In many cases, guardrails are only applied in the pre-processing or post-processing stages, rather than being integrated into the model's weights during training. As a result, they are often enforced only when the system is accessed through its public user interface — not accounting for floating biases created within the model as a result of bad training data or API connection to these models.

Research on AI moderation systems further indicates that, in the context of hate speech detection, these systems face challenges in identifying harmful content when it is conveyed through indirect narratives rather than through explicit and overtly abusive language or slurs.[29] In addition, evaluating whether an AI model is antisemitic presents significant challenges, as the process may involve assessing the degree of antisemitism exhibited by a model along a continuous scale (e.g. 0-100). At present, no unified methodology,

---

[25] Hailin Chen et al., "ChatGPT's One-year Anniversary: Are Open AI Models Catching up?", Cornell University, January 2024, www.arxiv.org/abs/2311.16989.
[26] Hakan Inan et al., "Llama Guard: LLM-based Input-Output Safeguard for Human AI Conversations", Cornell University, December 2023, www.arxiv.org/pdf/2312.06674.
[27] Yuan Yuan et al., "From Hard Refusals to Safe-Completions: Toward Output-Centered Safety Training", Cornell University, August 2025, www.//arxiv.org/abs/2508.09224.
[28] Gabriel Weimann et. al., "Generating Terror: Risks of Generative AI Exploitation", Combatting Terrorism Center at West Point, 17(1), January 2024, www.ctc.westpoint.edu/generating-terror-the-risks-of-generative-ai-exploitation/.
[29] "Model-Dependent Moderation: Inconsistent Detection Across LLM-based Systems".

standardized protocol, or widely accepted testing framework exists that can address this issue in a manner that satisfies all relevant stakeholders, including AI companies, governments and regulators, and civil society organizations focusing on harm. As a result, **while AI system safeguards can reduce certain risks, they do not fully prevent the creation or dissemination of harmful AI-generated content, particularly as these systems become more widely accessible.** In an effort to address this gap, CyberWell is building evaluation methodologies and benchmarking frameworks aimed at enabling more consistent, transparent, and evidence-based assessments of antisemitic outputs across generative AI systems.

## National Approaches towards AI Governance

Policymakers increasingly view AI as a driver of technological advancement, economic development, and national security. Governments recognize the potential of AI systems to improve performance across public sector areas such as education, healthcare, transportation, and infrastructure. As a result, many countries prioritize AI development within their national technology strategies.

However, policymakers also recognize that the rapid expansion of AI technologies introduces new risks that require regulatory oversight. This includes the spread of misinformation, the amplification of biased narratives, privacy violations, and the potential misuse of AI tools. In response, governments are developing regulatory approaches designed to maximize the benefits of AI while mitigating potential harm.

Understanding these national approaches is essential to this report, as these frameworks shape **how governments define harm and evaluate risks associated with AI technologies.** This section examines regulatory approaches relevant to the proliferation of harmful AI-generated content, including antisemitic material. The national AI strategies outlined below focus on regions most relevant to CyberWell's broader research.

### National AI Strategies and Regulatory Frameworks

Approaches towards AI governance vary across jurisdictions, reflecting diverse perspectives on how to balance the need for regulatory safeguards against the risk that overregulation may constrain innovation.

The European Union's AI regulatory approach prioritizes the responsible development of AI systems, while emphasizing the protection of fundamental rights. The EU's Artificial Intelligence Act, adopted in 2024, establishes a risk-based framework that classifies AI systems according to their potential impact on safety, livelihood, and fundamental rights. Under this framework, AI systems are subject to varying levels of transparency and oversight based on their level of risk.

In 2025, Australia adopted a National AI Plan that similarly outlines national commitments to mitigate harms and promote responsible AI development. However, the plan largely relies on existing legal and regulatory frameworks rather than on introducing new binding requirements. In parallel, Australia introduced a Voluntary AI Safety Standard providing guidance for AI developers and organizations deploying AI systems, with a special focus on issues such as biased outputs and harmful content generation.

In contrast, the United States has taken a more conservative regulatory approach to AI, placing greater emphasis on protecting freedom of expression and maintaining technological innovation. In U.S. policy discussions, there is frequent concern that excessive regulation could limit AI development or undermine constitutional protections of speech.

These differing approaches to AI governance highlight the extent in which regulatory strategies are context dependent. At present, no unified global framework exists to specifically address the role of AI technologies in the spread of harmful online content, including hate speech and extremist narratives. Instead, **governance remains fragmented across jurisdictions**, with AI-specific policies, existing hate-related laws, and platform regulations operating within separate regulatory domains.

## Gaps in Addressing AI-Generated Hate Speech

National AI strategies often seek to balance technological innovation and safety considerations. While some national AI strategies acknowledge risks related to misinformation and bias, they rarely address how AI tools may facilitate the creation and dissemination of hateful and extremist narratives beyond the models themselves. This gap raises concerns that AI-generated harmful content may reinforce broader patterns of discrimination or contribute to violence that extends beyond digital spaces.

In addition, most policy frameworks address harmful content primarily within broader discussions of misinformation and bias, rather than focusing on specific forms of online hate speech. These frameworks sometimes emphasize technical countermeasures such as warning labels, watermarking, and AI-generated content detection systems. However, research indicates that these tools demonstrate inconsistent effectiveness in practice, limiting their ability to reliably prevent the spread of harmful AI-generated content.[30]

Limited transparency from AI developers also presents a significant challenge for governments. Researchers have identified a gap between the information held by AI companies and what is accessible to governments and external researchers for evaluation.

[30] "International AI Safety Report: The International Scientific Report on the Safety of Advanced AI".

This lack of transparency complicates efforts to properly assess the risks associated with AI systems and develop effective policy responses.[31]

While **many AI governance frameworks acknowledge structural biases in training data and model outputs, they rarely provide operational guidance for addressing harmful AI-generated content once it has been disseminated online, particularly on social media platforms.** In many cases, national strategies address AI developers and social media platforms as separate domains, without examining how generative AI systems interact with platform recommendation systems and content moderation frameworks.

This separation creates a policy gap in which AI systems facilitate the large-scale production of harmful multimedia content, allowing it to spread rapidly across social media platforms where it gains visibility and engagement. This **underscores the growing need for closer alignment between AI governance frameworks, platform policies, and trust and safety enforcement mechanisms**. This need is becoming more pronounced as generative AI technologies increase the scale and sophistication of harmful content online.

## Recommendations

Based on the findings and analysis presented in this report, the following recommendations outline potential approaches for addressing the rise of AI-generated antisemitic content online. These recommendations are directed at three primary stakeholder groups: social media platforms, AI companies, and policymakers.

### Social Media Platforms

1. Adopt AI-generated content policy frameworks

   Platforms should establish clear policies on AI-generated content explicitly stating that existing policies on hate speech, harassment, and violence apply equally to AI-generated material, as reflected in policies adopted by TikTok. This can help reduce enforcement gaps and ensure consistent moderation in line with platform standards.

2. Strengthen detection of AI-generated content across multimedia formats

   Platforms should expand detection capabilities to identify antisemitic themes and narratives embedded within AI-generated media, including video, audio, images, and comments. Moderation systems and teams should also account for how such content manifests in different formats, such as AI-generated music, which can be rapidly disseminated at scale.

[31] "International AI Safety Report: The International Scientific Report on the Safety of Advanced AI".

3. Limit the algorithmic-driven amplification of violent and hateful AI-generated content

   Platforms should strengthen safeguards within recommendation systems to prevent the algorithmic spread of AI-generated antisemitic and violent content. Early-stage distribution controls can reduce the reach of such content while enforcement decisions are pending. Platforms should also strengthen and develop new scalable monitoring and fact-checking processes for AI-generated content that exploits breaking news or high-profile events, particularly in the aftermath of violent incidents directed at Jewish communities.

4. Improve detection and enforcement of re-uploads and cross-account distribution

   Platforms should strengthen mechanisms to detect the re-upload and redistribution of AI-generated antisemitic content, particularly when the content originates from previously removed accounts. Watermarks and creator attribution can further support cross-account tracking and enforcement following initial content removal.

5. Develop guidance for moderation teams on detecting coded antisemitic content in AI-generated media

   Platforms should update moderation guidelines and add new moderator trainings to explicitly address coded antisemitic language, symbols, and satirical representations that may appear in AI-generated material. In collaboration with civil society organizations and subject-matter experts like CyberWell, platforms should strengthen their moderation practices and training curriculums to identify indirect forms of antisemitism, particularly when explicit slurs are absent.

6. Strengthen policies and moderation guidance on the use of disclaimers

   Platforms should update their policies to clarify that the use of disclaimers such as “#humor”, “#satire”, or “for educational purposes”, does not exempt content from enforcement. This is especially applicable when content promotes implicit or coded hate speech. Moderation decisions should be based on content itself, not a user’s stated intent alone. Platforms should also provide content moderators with clear guidance on how to assess disclaimers, which are often used in these cases to obscure harmful intent.

## AI Companies

1. Provide clearer definitions of harmful content and protected categories in AI usage policies

   AI companies should provide more detailed definitions of harmful content in their safety and usage policies. These policies should also include explicit references to protected categories, including religion, ethnicity, race, gender, and sexual orientation. Greater clarity in policy language can reduce ambiguity in how hate speech and harmful content are interpreted and enforced.

2. Expand red teaming to address harmful content generation

   AI companies should strengthen red teaming efforts to identify how AI models can be manipulated to generate harmful content, including implicit or coded forms of hate such as antisemitism. Red teaming should be an ongoing process that incorporates expertise in extremism, online hate, and cultural context to identify vulnerabilities in model behavior prior to deployment.

3. Increase transparency around safety guardrails and risk mitigation strategies

   AI companies should provide clearer public disclosure of how safety guardrails and training mechanisms are designed, tested, and updated. Greater transparency allows researchers to better evaluate how AI systems perform in real-world settings. It can also support collaboration with civil society organizations, including CyberWell, to identify and address emerging risks.

## Policymakers

1. Mandate transparency and accountability for AI system risks

   Policymakers should require AI companies to disclose information about the capabilities and safeguards of generative-AI systems, and their implementation. These measures may include reporting requirements, independent audits, and research access programs that allow regulators to assess risks and ensure accountability.

2. Encourage coordination between AI governance frameworks and online platform regulation

   Policymakers should promote greater alignment between AI governance frameworks and regulatory approaches that address harmful content online. Strengthening coordination between governance structures that address AI companies and social media platforms can help ensure that risks associated with generative AI systems are addressed alongside social media platform governance and online safety policies.

3. Address youth exposure to harmful AI-generated content through legislative and regulatory frameworks

   Policymakers should recognize the growing role of AI-generated content in exposing youth to harmful and antisemitic narratives online. They should consider legislative measures to address these risks within broader online safety frameworks. Integrating these considerations into existing digital governance and online safety laws can help ensure that emerging AI technologies do not increase youth exposure to harmful or extremist content.

4. Require standardized methods for identifying AI-generated media

   Policymakers should require AI developers to implement standardized methods to identify AI-generated media across digital platforms. Establishing consistent identification standards can improve transparency and support efforts to limit the spread of harmful content online.

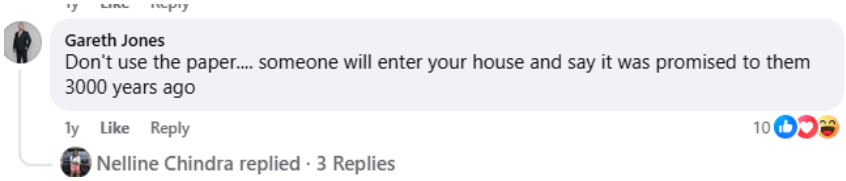
## Appendix

### Evolution of the Promised 3,000 Years Ago Trend

Phase 1

https://www.facebook.com/tabi.tabiabbasi.7/posts/this-is-read-before-your-eyes-the-times-of-israel-the-ancient-4500-year-old-tuni/2496835770708972/

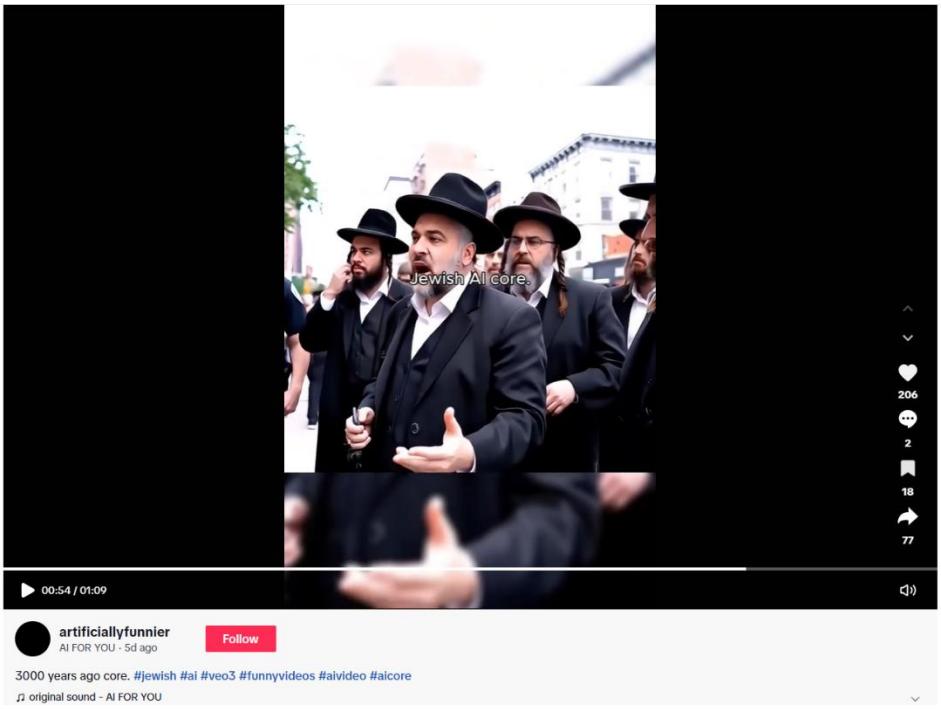
This Facebook comment from January 2025 replies to a post mocking Israel by depicting the Israeli flag on toilet paper rolls. In response, a user employs the "promised 3,000 years ago" trope in a humoristic tone to suggest that, similar to the Jewish people's claim to a historic connection to the land of Israel, Jews are entitled and greedy and would steal even toilet paper from the homes of ordinary people.

Phase 2

https://www.tiktok.com/@artificiallyfunnier/video/7522945559729016078

This TikTok video from early July 2025 uses Veo3 to promote the “promised 3,000 years ago” trope. The video opens with a visibly Jewish man approaching an individual in a market, stating: “That watch was promised to my people 3,000 years ago” [00:02-00:06]. This phrase is repeated throughout the video with variations applied to different objects, including a chair, the Moon, a wallet, and New York City [00:16-01:09].




Phase 3

https://x.com/SiyaaUnrated/status/2000179000170418530

In response to the December 2025 Bondi Beach Hanukkah shooting in Sydney, Australia, this user on X invokes the “promised 3,000 years ago” trope to dehumanize Rabbi Eli Schlanger, a Jewish victim of the attack, and justify his death.