# CyberWell
More data, less hate.

53M

18M

32K

18K

948

211

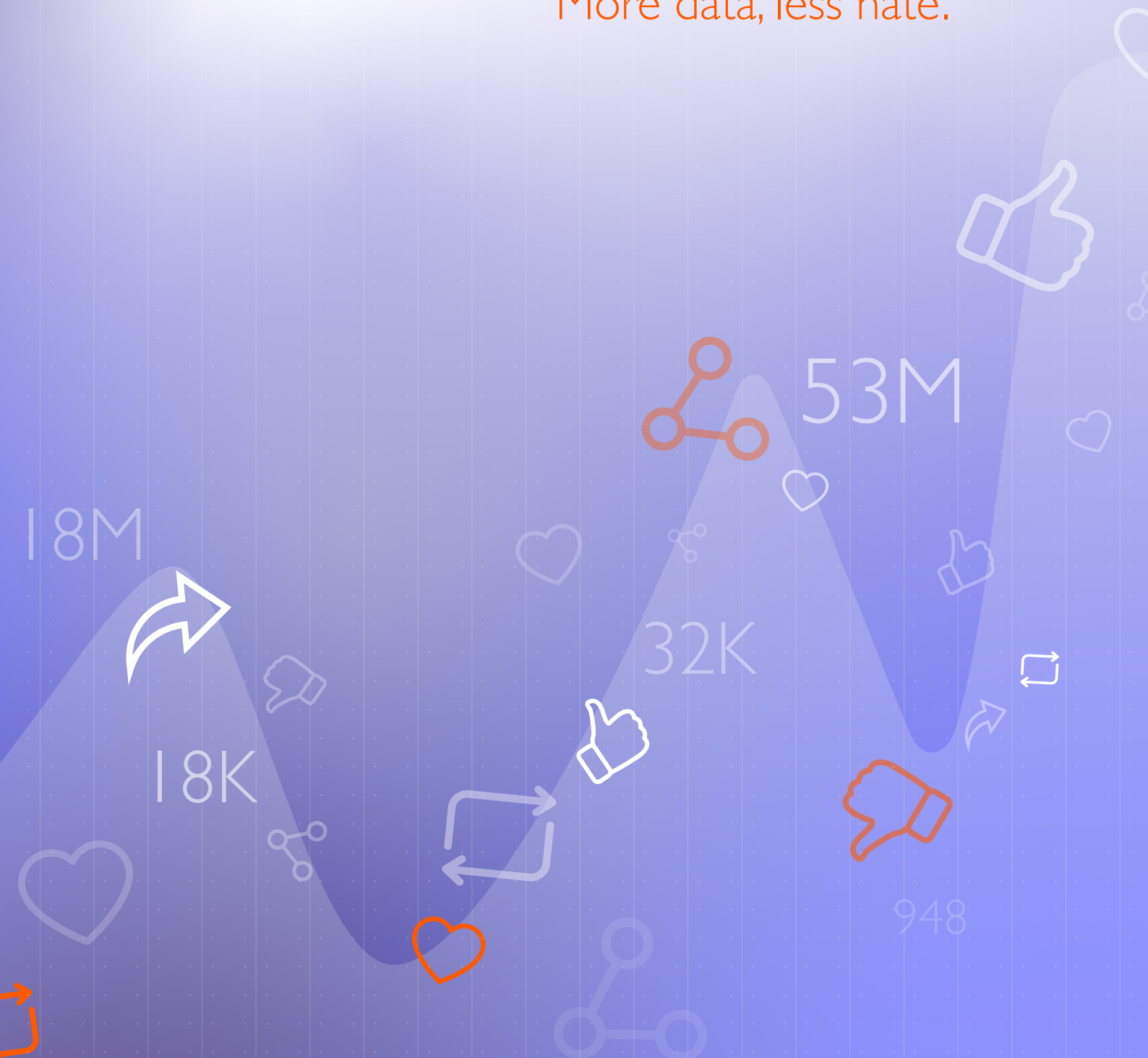**Annual Report**
**The State of Online Antisemitism 2022**

# Table of Contents

*A message from Tal-Or Cohen Montemayor, Founder & CEO*

Dear CyberWell supporter,

I founded CyberWell because I was alarmed and overwhelmed seeing the increasing levels of open Jew-hatred on our mainstream social media platforms - the same platforms that I, like most of you, use every day. The social media platforms, unlike dark forums and hate sites, have rules and community guidelines - a promise that the social media companies made to us, their users, to ensure that we would not be the target of identity-based hate on their platforms that commodify our attention and our personal data for profit.

CyberWell was founded to make sure that they were held to account for that promise.

Driven by data, our mission is to promote the enforcement and improvement of community standards across the digital space and empower efforts against antisemitism with the data that we collect. Our mission may sound niche or technical, but last year provided more than enough evidence of how badly we are needed.

2022 was marked by celebrity-led litanies against the Jewish people across social media, major changes in how Twitter treats hate content under new management, and the unsurprising results of prolonged normalization of online Jew-hatred: increased real-life antisemitic incidents and a demonstrably unfavorable shift in public perception of Jews.

Because of my experience leading open-source intelligence projects, researching extremist movements, political antisemitism, and dark web Jew-hatred, I unfortunately cannot say that this comes as a shock. However, I believe that if we hold tech companies accountable for the antisemitism that they host and amplify, we can significantly improve the digital future for not only Jews, but all internet users.

It is crucial to remember that the Holocaust would not have happened if antisemitism had not been amplified by radio, television, and intellectual circles. Today, social media provides the world with a much more powerful and far-reaching tool to spread toxic ideas, misinformation, and racism than traditional media ever did.

It's high time that we change the conversation to address not just the bad actors, but also their enablers. Social media companies are responsible for algorithm-governed machines that disseminate and normalize open Jew-hatred, and they must be held accountable. While some have taken welcome steps to increase transparency and automate the removal of hate speech, we need more transparency and internal resource allocation so that outpourings of antisemitism are not reinforced by the 'public square' platforms we use every day.

CyberWell launched in May 2022 and, powered by the best available technology and the generous support of our philanthropic 'early adopters', we have already vetted and reported over 3,300 pieces of antisemitic content in two languages across five major social media platforms - and counting! Every post is valuable – a new lawsuit against Twitter for failing to remove online antisemitism under the German criminal code was filed with only six Tweets.

Through our unique compliance-driven methodology, CyberWell is leveraging that content to generate deeper data insights about major policy gaps, highlight where social media platforms' flagging resources and community standards fail to address Jew-hatred, and classify online antisemitism in clear and understandable terms.
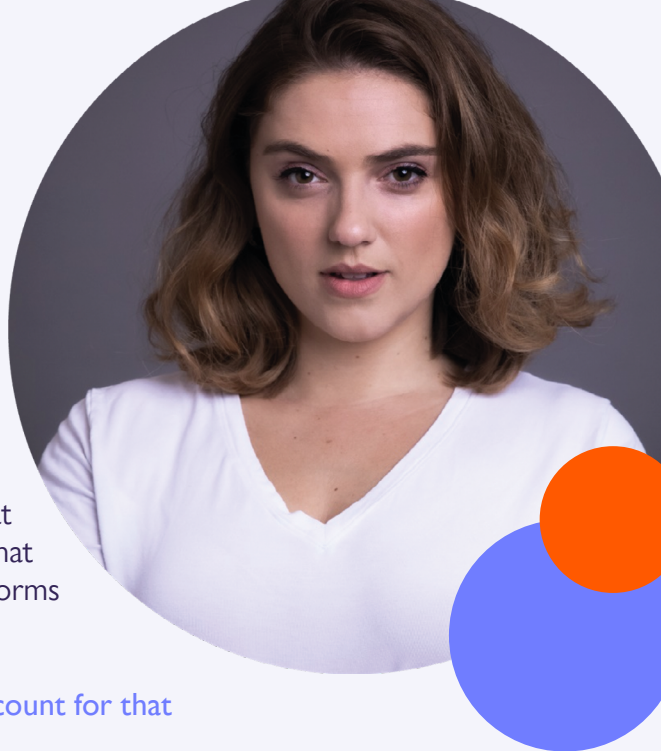
Moreover, our live monitoring enabled us to be the "first responders" for social media platforms during the surge in digital Jew-hatred following Ye's (formerly Kanye West) diatribes against Jews in posts, public statements, and interviews. Using data, we alerted platforms to the exponential online impact of his screeds, producing tailored reports that specifically defined policy violations, concrete shifts in online antisemitic discourse, and provided content review teams with the specifics they needed to expedite removal and response.
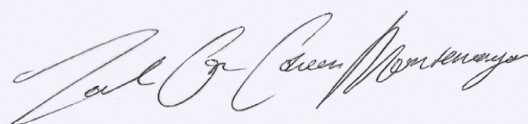
With your help, we will be able to continue our efforts with vigilance and dedication. We will overcome the social media accountability crisis and blunt future spikes of online hatred against Jews. Together we can make the change we desperately seek and secure a better future for us all, both on and offline.

Less than a year after Cyberwell became the first-ever platform to monitor antisemitism on social media in real time, I am humbled to share our inaugural annual report.

I am deeply grateful for the philanthropic support of CyberWell's donors, for our trusted service providers and partners, and for the seasoned advisors who are guiding the development of our tech, methodology, and operational capacity.

Most of all, I am grateful for CyberWell's growing team, who are in the digital trenches tirelessly fighting for better digital spaces. Our accomplishments thus far would not have been possible without our dedicated and impassioned staff - I cannot wait to see what we will achieve alongside our donors and partners in 2023!

With respect and gratitude,

Tal-Or Cohen Montemayor, Adv.
CyberWell Founder & CEO

## Platform Overview

**5**
platforms monitored
TikTok, Facebook, Instagram, Twitter, YouTube

**2**
languages tracked
Arabic, English

**+3,300**
pieces of content vetted
and counting!

**+100,000**
pieces of content flagged
as highly likely to be antisemitic waiting for review

**189**
users reporting
with CyberWell

**1,082**
followers engaged
on CyberWell social media

# CyberWell's Mission | More data, less hate.

CyberWell is a groundbreaking non-profit with technology at its core. Our mission is to drive the enforcement and improvement of community standards and hate speech policies across social media platforms and empower existing efforts and users with the data we collect to collaboratively eradicate online Jew-hatred.

Online antisemitism is one of the fastest growing forms of Jew-hatred today. Previous 'snapshot' research estimates that roughly 80-85% of reported antisemitic content items were not removed across social media platforms,[1] a number within range of CyberWell's findings for 2022, which showed an average removal rate of only 23.8%. Surveys reflect that online antisemitism is one of the primary forms of Jew-hatred affecting communities on a daily basis, especially younger Jews - leading to anxiety, fear, and even active hiding of Jewish identity online and offline.[2]

> "I use social media extensively both professionally and personally, and the sad reality is that I am often the target of harassment and hate - just for being Jewish. As a refugee from Iran, I am all too familiar with how harmful ideas and rhetoric can turn into very real violence. It is critical to increase awareness of antisemitism in online spaces, where it poisons minds and creates an unsafe atmosphere."
>
> - Dr. Sheila Nazarian 📷 @drsheilanazarian

To address this challenge, in May 2022 CyberWell launched the first ever open data platform of online antisemitic content to democratize the fight against Jew-hatred with data.
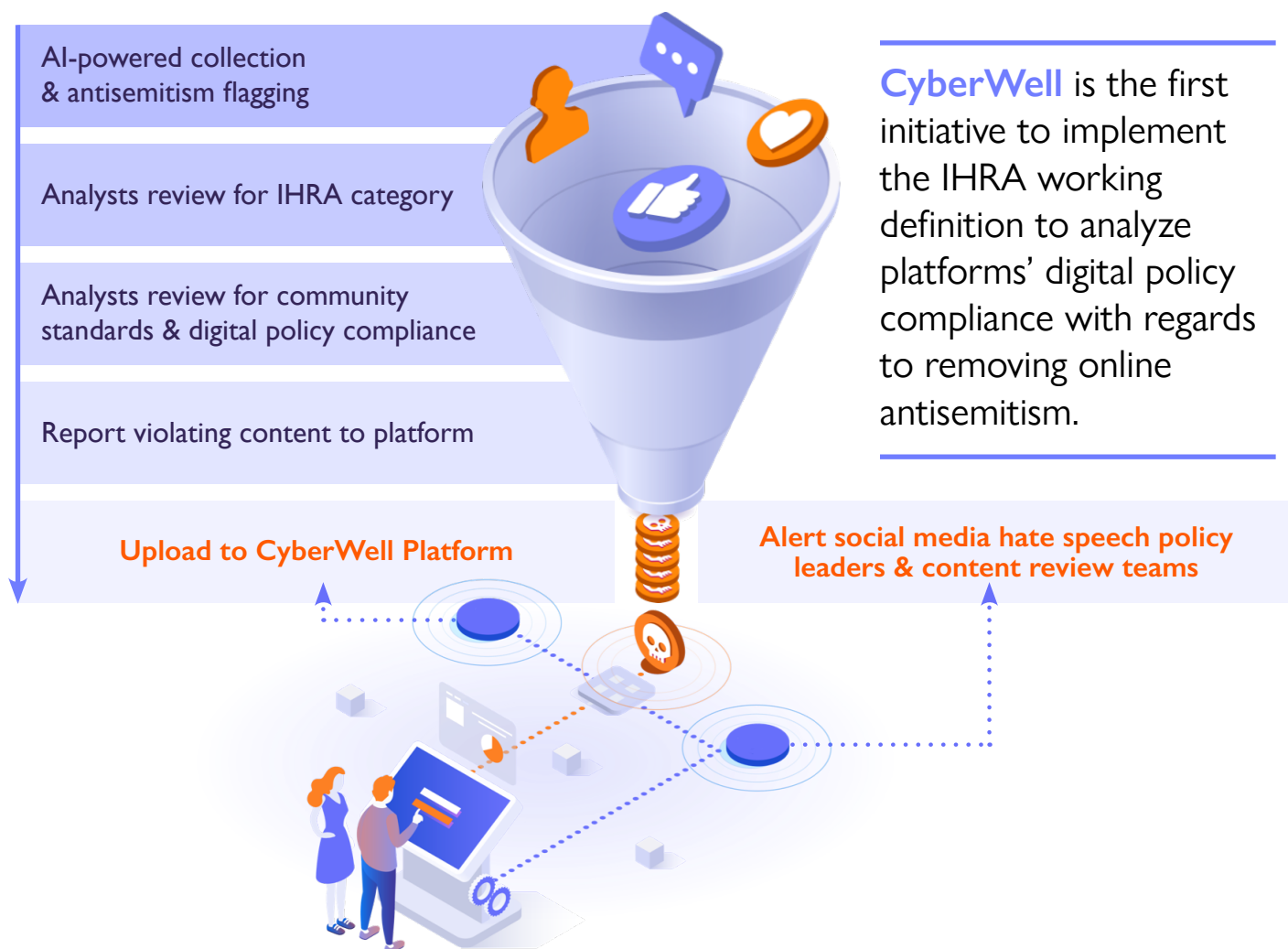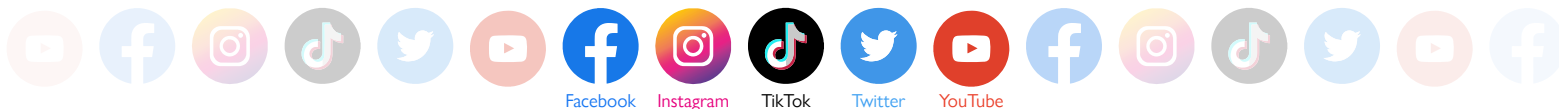
CyberWell is the only open-source platform monitoring the leading social media platforms (Facebook, Instagram, Twitter, YouTube, and TikTok), full-time, professionally, and in multiple languages (English and Arabic). With our unique methodology, cutting-edge AI, and big data tools, we collect content from online platforms, which is then vetted by our team of analysts and categorized according to the International Holocaust Remembrance Alliance's (IHRA) working definition of antisemitism. The content is then reviewed by our digital policy and research team for violations of community standards and hate speech policies, reported to the relevant platforms, uploaded to our database, and monitored for how long the platforms take to remove it – if they even do.

At Cyberwell, we believe platforms will respond more readily to data that is presented through the lens of their own rules. Therefore, we break down each piece of content according to the community guideline or digital policy that it violates, in order that our data may facilitate compliance and improve the abysmal enforcement rate of 20-25% at best.[3]

1 Anti-Defamation League, How Platforms Rate of Hate: Measuring Antisemitism and Adequacy of Enforcement Across Reddit and Twitter p. 12, published 02/22, accessed 01/23; Center for Countering Digital Hate, Failure to Protect: How tech giants fail to act on user reports of antisemitism, published 07/30/21, accessed 01/23; Fighting Online Antisemitism, Fighting Online Antisemitism: Annual Report 2021, p. 7, published 01/13/22, accessed 01/23.

2  American Jewish Committee, The State of Antisemitism in America 2021 p. 5-7, published 10/21, accessed 02/23; Anti-Defamation League, 2021 Survey on Jewish American's Experiences with Antisemitism, published 03/31/2021, accessed 02/23; European Commission, European Union Agency for Fundamental Rights and the Institute for Jewish Policy Research, Young Jewish Europeans: perceptions and experiences of antisemitism p. 26-27, published 07/28/19, accessed 02/23.

3  Anti-Defamation League, How Platforms Rate of Hate: Measuring Antisemitism and Adequacy of Enforcement Across Reddit and Twitter p. 12, published 02/22, accessed 01/23; Center for Countering Digital Hate, Failure to Protect: How tech giants fail to act on user reports of antisemitism, published 07/30/21, accessed 01/23; Fighting Online Antisemitism, Fighting Online Antisemitism: Annual Report 2021, p. 7, published 01/13/22, accessed 01/23.

AI-powered collection & antisemitism flagging

Analysts review for IHRA category

Analysts review for community standards & digital policy compliance

Report violating content to platform

**Upload to CyberWell Platform**

**Alert social media hate speech policy leaders & content review teams**

**CyberWell** is the first initiative to implement the IHRA working definition to analyze platforms' digital policy compliance with regards to removing online antisemitism.

Additionally, we highlight which forms of antisemitism are most commonly missed by existing enforcement mechanisms, either due to technical deficiency or gaps in policy. Thus we can effectively highlight and advocate for the best community guidelines and policy development to stop the spread of online Jew-hatred.

As we expand our capabilities, we aim to serve as the premier data provider, researcher, and alert center for tracking online antisemitism. A professional version of our database is set to launch in 2023, democratizing the fight against antisemitism by making our data available for professional research, legislation, advocacy, and journalism.

We envision a world where Jewish social media users feel safe to express their identity as openly as anyone else - where antisemitic content is as unacceptable online it is offline. By introducing a technologically advanced and scalable solution for public and professional use, CyberWell is on a mission to improve the digital future for the Jewish people.

# What's in The Well?

## Data Insights Collected by CyberWell in 2022

> "When I was a child, antisemitism spread through fliers and speeches to thousands of people. Now, antisemitism and Holocaust denial spreads through social media and reaches millions of people. Social media platforms have to be held accountable for this hate."
>
> Gidon Lev, Holocaust Survivor ♪ *@thetrueadventures*

Often referred to as the "world's oldest hatred", antisemitism has been a disturbing and harmful part of society for millennia. Jew-hatred has now found its way onto its largest soapbox yet – the Internet – where antisemitism is spread, reinforced, and normalized to the widest audiences in history. Its reach is often fueled by algorithms that drive users to more extreme content as a means of maintaining their attention.[4]

While basic features of antisemitism persist through time, the discourse online is evolving rapidly and requires an agile technological solution. To effectively combat this phenomenon, CyberWell applies its consistent methodology to accurately reflect and combat online antisemitism by focusing on social media accountability and digital policy development. The following is a breakdown of the data collected from our launch in May through December.

## Summary of CyberWell Data

| Monitoring Since | Examples Collected | Platforms Monitored |
|---|---|---|
| **2022** | **3,320** | **5** |

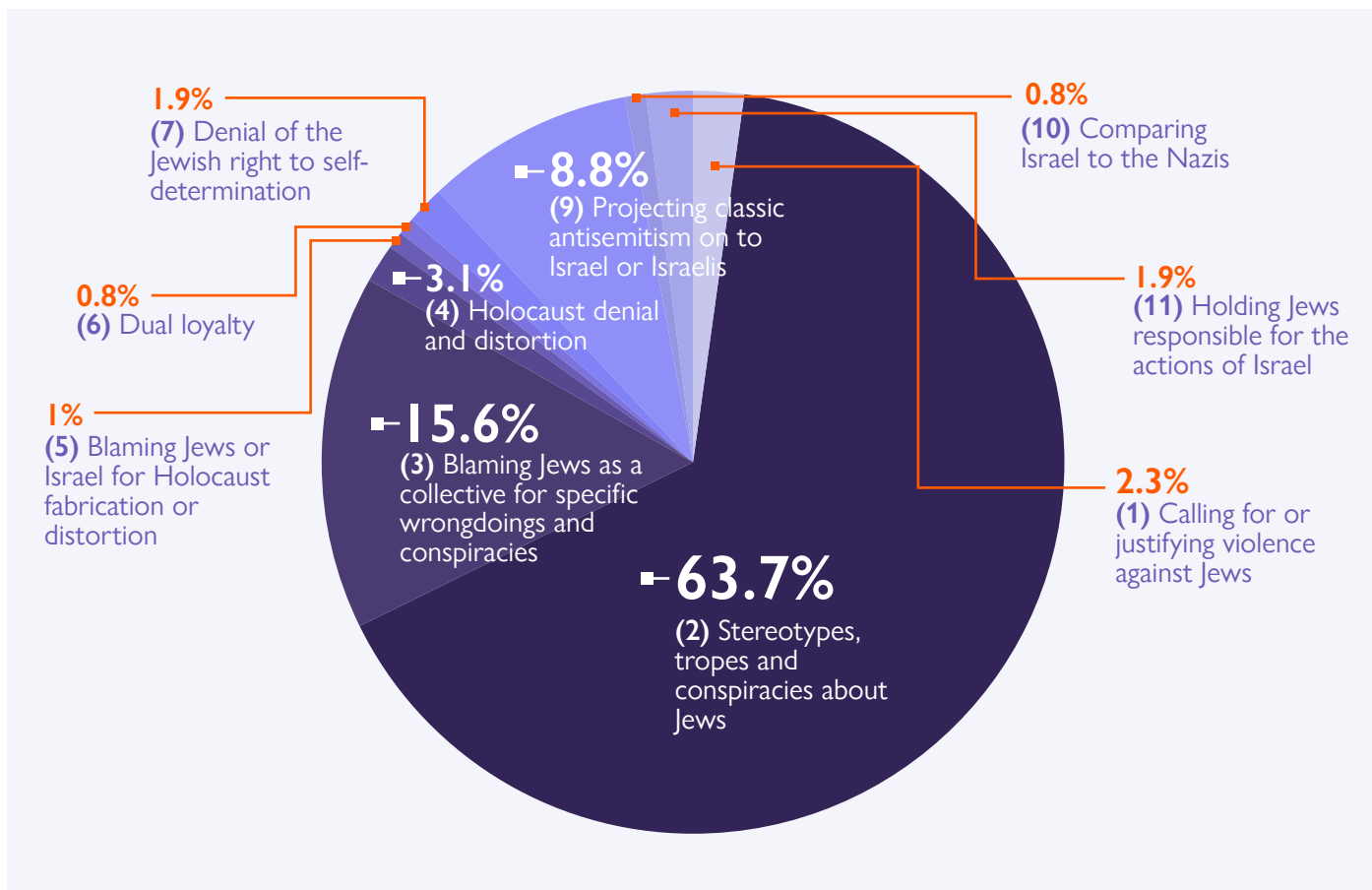| Countries Represented | Languages Reflected |
|---|---|
| **81** | **16** |

## Discourse Analysis
*Breakdown of Data According to the IHRA Working Definition*

It is important to recognize that many people, including content review teams employed by social media platforms,[5] do not have a focused professional understanding of antisemitism or how it manifests in online spaces. Content review teams are often tasked with handling multiple forms of hate speech and other content that violates their platform's policies and may not receive specialized training on antisemitism.

---

4  Gabriel Weimann; Natali Masri, TikTok's Spiral of Antisemitism. *Journal. Media. 2021, 2,* 697-708, accessed 02/23; The Jewish Chronicle, Instagram's algorithm 'recommending' antisemitic imagery, published 03/09/21, accessed 02/23

5  Meta, How Meta prioritizes content for review, published 01/26/22, accessed 02/23; Washington Post, Twitter hate speech up in large foreign markets after Musk takeover, published 01/14/23, accessed 02/23; YouTube, Inside YouTube: Responsible policy enforcement during Covid-19, published 08/25/20, accessed 02/23.

**1.9%**
**(7)** Denial of the Jewish right to self-determination

**0.8%**
**(6)** Dual loyalty

**1%**
**(5)** Blaming Jews or Israel for Holocaust fabrication or distortion

**8.8%**
**(9)** Projecting classic antisemitism on to Israel or Israelis

**3.1%**
**(4)** Holocaust denial and distortion

**15.6%**
**(3)** Blaming Jews as a collective for specific wrongdoings and conspiracies

**63.7%**
**(2)** Stereotypes, tropes and conspiracies about Jews

**0.8%**
**(10)** Comparing Israel to the Nazis

**1.9%**
**(11)** Holding Jews responsible for the actions of Israel

**2.3%**
**(1)** Calling for or justifying violence against Jews

CyberWell categorizes all pieces of vetted antisemitic content according to the 11 examples featured in the IHRA working definition.[6] This generates transparency around the categories of antisemitic discourse infecting the digital space and serves as a litmus test for measuring which types of antisemitism are allowed to proliferate due to implementation gaps by the platforms.

**Top 3 types of antisemitism in CyberWell's dataset of social media content:**

**Type 2 – 63.7%** | Classic antisemitism, including stereotypes, inaccurate information, and demonizing content about Jewish people, including tropes about Jewish power over institutions, economy, the media and the world.[7]

**Type 3 – 15.6%** | Blood libel and blaming Jews for things that go wrong.[8]

**Type 9 – 8.8%** | Projecting classic antisemitism on to Israel or Israelis.[9]

For more insight, visit our dashboard at app.cyberwell.org

6  International Holocaust Remembrance Alliance, What is antisemitism? Non-legally binding working definition of antisemitism, accessed 02/23

7  Full definition of the second example of the IHRA definition: "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."

8  Full definition of the third example of the IHRA definition: "Accusing Jews as a people of being responsible for real or imagined wrongdoing committed by a single Jewish person or group, or even for acts committed by non-Jews."

9  Full definition of the ninth example of the IHRA definition: "Using the symbols and images associated with classic antisemitism (e.g., claims of Jews killing Jesus or blood libel) to characterize Israel or Israelis."

While only 2% of the content CyberWell collected during 2022 was violent antisemitism i.e., justifying or calling for violence against Jews,[10] Twitter was responsible for over 90% of violent antisemitic content online. January and October saw the largest amount of violent antisemitic content, while July and October featured the largest amount of antisemitic content overall. There was a 127.72% increase in the amount of antisemitic content CyberWell vetted in October as compared to September. During October there was a 10.8% increase in classic antisemitic content online, when the highest rate of change was consistent with the third criterion of the IHRA working definition - blaming Jews as a group for things that go wrong or acts that were committed by a Jewish person. CyberWell concluded that October was the most volatile month both in terms of violent antisemitic content and the large amount overall of Jew-hatred online due to the celebrity-empowered attack on Jewish people, led by Ye[11].

## Antisemitism over time by platform 2022
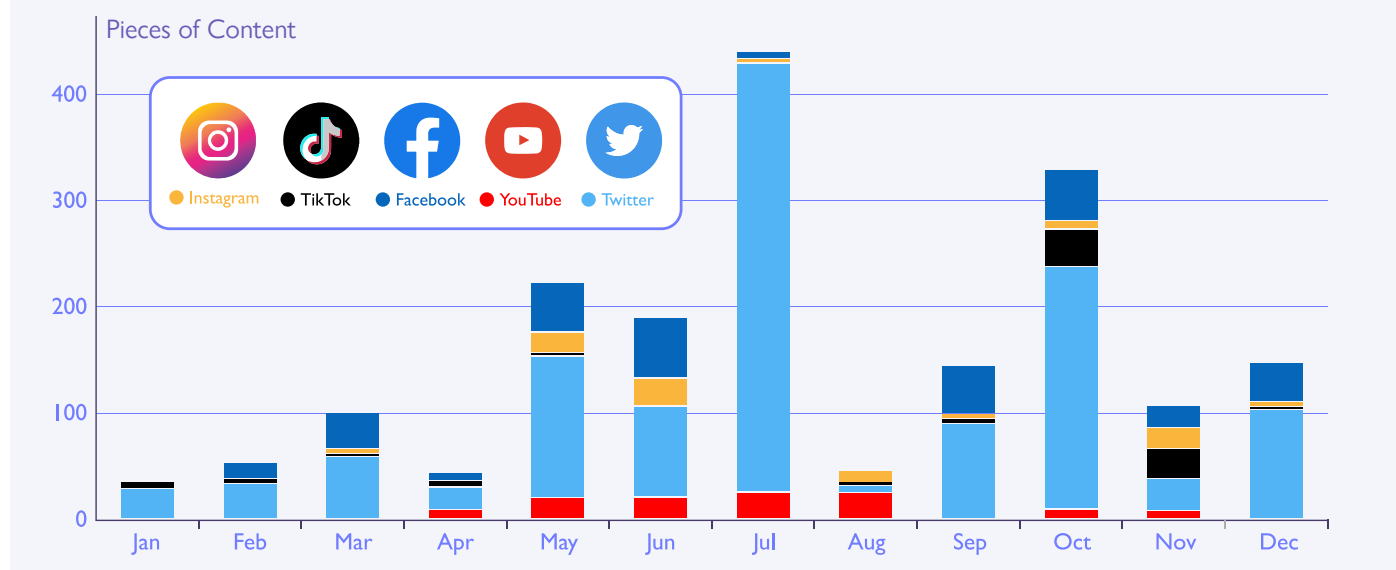


CyberWell also monitors the subjects, themes, and conspiracy theories in which the content we collect is rooted.

## Leading Antisemitic Tropes Online

| | |
|---|---|
| **15.62%** | *Jews are greedy and money obsessed* |
| **14.55%** | *Jews control the world or are dominating the world order* |
| **12.10%** | *Jews control the economy, specifically* |

⚠️ It is important to note that all of these top subjects identified in CyberWell's dataset should technically be prohibited on all major social media platforms according to the respective community guidelines in place today.

## Leading Antisemitic Tropes by Platform

| | |
|---|---|
| Facebook | *Jews control the world or are dominating the world order* |
| Instagram | *The Rothschild conspiracy theory* |
| TikTok | *Jews control the world or are dominating the world order AND Jews are evil* |
| Twitter | *Jews are greedy and/or money-obsessed* |
| YouTube | *Jews are the "Synagogue of Satan"* |

---

10 Full definition of the first example of the IHRA definition: "Calling for, aiding, or justifying the killing or harming of Jews in the name of a radical ideology or an extremist view of religion."

11 See more information about CyberWell's fight against online antisemitism post-Ye at: https://cyberwell.org/wp-content/uploads/2022/11/Cyber-Well-Online-Antisemitism-Alert-202211.pdf

## Leading Antisemitic Tropes Online by Language

There are noticeable differences in the leading antisemitic beliefs found in English and Arabic.[12] While antisemitic content in English tends to focus on stereotypes about Jewish relationships to money and economic control, antisemitic content in Arabic demonizes Jews as the 'enemy' and relies heavily on the concept of a world cabal popularized by the historic fabricated document The Protocols of the Elders of Zion.

**A** | Leading tropes in English

**19%** *Jews are greedy and/or money-obsessed*

**13%** *Jews control the economy, specifically*

**10%** *Jews control the world or are dominating the world order*

**ي** | Leading tropes in Arabic

**25%** *Jews are the enemy*

**22%** *Jews control the world or are dominating the world order*

**18%** *Protocols of the Elders of Zion - content posting, supporting or quoting The Protocols of the Elders of Zion*

Prior to the World Cup, Islamists accounts promoted the idea that Jews and Zionists were distracting the world with organized sports and openly cited the Protocols of the Elders of Zion as their source across social media. The 2022 World Cup in Qatar was leveraged as an opportunity to spread this antisemitic conspiracy theory further.

**Spotlight**: The Protocols of the Elders of Zion and the World Cup

During the World Cup, CyberWell uncovered an online campaign by Salafi groups in Algeria citing The Protocols of the Elders of Zion and using the Arabic hashtag #كرة_الهدم_والندم (literally, #Ball_of_destruction_and_regret). This campaign claimed that the World Cup was being used as a tool by the Jews to distract people from their plans of world domination.[13]



Āōulōul Āōulōul
March 29, 2022

بروتوكولات حكماء صهيون "
( البروتوكول الثالث عشر ) :

" ولكي تبقى الجماهير في ضلال ، لا تدري ما وراءها وما أمامها ، ولا ما يراد بها ، فإننا سنعمل على زيادة صرف أذهانها بإنشاء وسائل المباهج ، والمسليات ، والألعاب الفكاهية ، وضروب أشكال الرياضة واللهو وما به الغذاء لملذاتها وشهواتها ، والإكثار من القصور المزوقة والمباني المزركشة ، ثم نجعل الصحف تدعو إلى مباريات فنية رياضية من كل جنس ، فتتوجه أذهانها إلى هذه الأمور ، وتتصرف عما هيأناه ، فنمضي به إلى حيث نريد " . انتهى .

Facebook, published 03/29/22, accessed 12/22



The first documented post of the antisemitic World Cup campaign, posted by "The Salafi platform of goodness," Facebook, published 11/20/2022 11:07 PM, accessed 12/22

# Policy Violation and Platform Inaction

When there is a legal or monetary penalty involved, social media companies have demonstrated the ability to successfully allocate appropriate resources, including AI-based enforcement, to respond to pressure from private companies (copyright infringement), paying customers (companies demanding brand safety for ad placement), or regulators (as in removing child pornography).
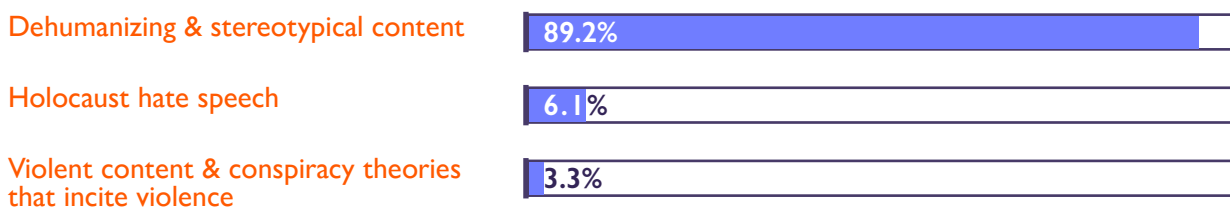
However, when it comes to protecting minority communities, such as Jewish users, by enforcing hate speech policies, eliminating content that spreads misinformation and creates a hostile environment is not their top priority. This is due to the lack of legal repercussions for inaction, lack of expertise in a complex subject on the part of moderators, and because social media platforms are based on a numbers game.

Content review teams do not actively scan for antisemitic posts missed by the platforms' AI like CyberWell does; they wait for users to report hateful content to them. Jews make up a tiny fraction of all users, so even if Jew-hatred is trending and Jewish content creators are targeted, not enough people are hitting the report button to cause a statistically significant trend at the content-review team level.

CyberWell analyzes each piece of vetted antisemitic content according to the generally agreed upon policy and specific rule that it violates per platform.[14] This analysis is vital to identify which areas of digital policy need additional development and resource allocation in order to effectively combat online antisemitism. Our specific findings are noted below.

## Policy Violation Theme

Top policy violation themes across all platforms in 2022:[15]

| Theme | Percentage |
|---|---|
| Dehumanizing & stereotypical content | 89.2% |
| Holocaust hate speech | 6.1% |
| Violent content & conspiracy theories that incite violence | 3.3% |

---

14  More details about the policy violation themes are available in CyberWell Policy Guidlines at https://cyberwell.org/how-it-works/policy-guidelines/
15  Ibid.

## Policy Violation Description

Top policy or guideline violated per platform in 2022:

| | | | |
|---|---|---|---|
| **Facebook** —— | **Meta Hate Speech Tier 1** | | **98%** |
| **Instagram** —— | **Meta Hate Speech Tier 1** | | **85%** |

*For a detail explanation of what is considered Hate Speech Tier 1, visit Meta' transparency center at: https://transparency.fb.com/policies/community-standards/hate-speech/*

**Twitter** —— **Hateful Conduct Policy** **93%**

*For a detail explanation of what is considered Hateful Conduct on Twitter, visit https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy*

**TikTok** —— **Hateful Behavior** **90%**

*For a detail explanation of what is considered Hateful Behavior on TikTok, visit https://www.tiktok.com/community-guidelines?lang=en#38*

**YouTube** —— **YouTube Hate Speech Policy** **100%**

*For a detail explanation of what is considered Hate Speech on YouTube, visit https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436*

While these are only CyberWell's initial findings, they highlight the biggest gaps in policy compliance leading to online antisemitism remaining available on the platforms and serve as a strong indicator of which part of the existing digital policies are in need of action, either through modified community guidelines, additional or reallocated resources for additional AI-powered enforcement on antisemitic content specifically, or both.

## Rate of Removal of Antisemitic Content

Starting rate of removal at launch, 21% across all platforms

| | Content offline | Content still online |
|---|---|---|
| Facebook | 17% | |
| Instagram | 13% | |
| TikTok | 22% | |
| Twitter | 26% | |
| YouTube | 31% | |
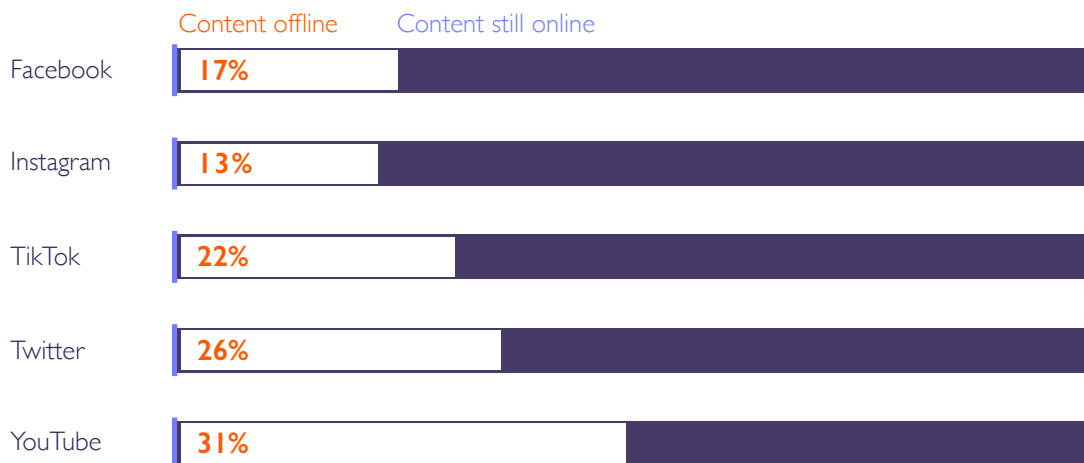
**23.8%** of antisemitic content was removed following CyberWell's reporting directly to platforms

In May 2022, CyberWell found that the rate of removal across all platforms was only 21%, which is consistent with previous "sprint" research efforts that did monitoring exercises over specific platforms and defined periods of time.

Since the launch of CyberWell's activities, wherein all vetted pieces of content are reported by CyberWell to the platforms on a regular basis, CyberWell increased the rate of removal from 21% to 23.8% for the content collected and vetted in 2022. This removal rate is still too low and demonstrates that the user-dependent reporting model is an ineffective solution to effectively diminish the levels of online antisemitism via the platforms' content review teams. However, it is important to note the added value that CyberWell has already brought to the state of digital antisemitism, increasing the removal rate of antisemitic content after less than a year of monitoring and reporting. As we continue our activities, we will remain vigilant, consistently reporting violating antisemitic content to the platforms, tracking the rates of removal for each platform, and scoring the platforms' efficacy against each other.

# Online Antisemitism in Real-time

## Case Study: Antisemitism Online Post-Ye

### Alerts

2022 marked a massive uptick in online antisemitism following two major events: Ye's public and celebrity-amplified tirade against Jews and Elon Musk's takeover of Twitter, during which there were coordinated efforts to perpetuate online hate.[16] In addition to ongoing monitoring, reporting, and data analysis, CyberWell generates real-time alerts and trend reports about the state of online antisemitism to demand social media vigilance and accountability year-round and during watershed moments of Jew-hatred, such as these.

During October and November, CyberWell's monitoring technology tracked how Ye's antisemitic remarks sparked an outpouring of Jew-hatred across social media and identified five major antisemitic narratives and tropes spurring online antisemitism as a reaction to or in defense of Ye's repeated and ongoing attacks.

*1. Jews are greedy, wealthy, money-obsessed, and control the economy/wealth*

*2. Jews control the media and/or Hollywood*

*3. Jews are in a conspiracy of slavery and/or purposeful exploitation of Black people*

*4. Synagogue of Satan\**

*5. Jewish people are not Jewish people/Jews are false*

*\*Synagogue of Satan – A term rooted in certain passages from the New Testament and used by Black Hebrew Israelites, the Nation of Islam, and other radical Christian groups and believers to promote the antisemitic narrative that Jews are imposters and/or worship a false God and/or Satan (religious demonization of Jews). This term has propelled online antisemitism for years yet is not restricted on social media platforms.[17]*

---

16  Center for Countering Digital Hate, The Musk Bump: Quantifying the rise in hate speech under Eon Musk, published 12/06/22, accessed 02/23; Forbes, Hate Speech Rises on Twitter After Elon Musk Takes Over Researchers Find, published 10/31/22, accessed 01/23; Twitter, Network Contagion Research Institute [@ncri_io], "Evidence suggests that bad actors are trying to test the limits on @Twitter. Several posts on 4chan encourage users to amplify derogatory slurs. For example, over the last 12 hours, the use of the n-word has increased nearly 5005% from the previous average.", published 10/28/22, accessed 02/23; Twitter, Anti-Defamation League [@ADL], Findings indicate a 63.% increase in the volume of tweets (excluding retweets) referencing "Jews" or "Judaism" with antisemitic sentiment in the two weeks following Musk's Twitter takeover, compared to the two weeks prior.", published 11/18/22, accessed 02/23.

17  Revelations 2:9 and 3:9.

CyberWell's team was able to identify these five major online antisemitic narratives, noticing a surge in the appearance of specific words through our monitoring technology. Following Ye's comments throughout October and November our database revealed a surge in the words, "Kanye", "Black people", "media" and "synagogue of Satan", or variations thereof, among the content that was flagged by our AI as highly likely to be antisemitic.[18]

## #Hashtag Monitoring and Analysis

CyberWell used various social listening tools to monitor specific trending hashtags on Twitter. Trending hashtags promoting antisemitism October through December can largely be split in two categories:

> **1. 'Safe' hashtags** | Seemingly "innocuous" hashtags that do not violate Twitter rules prima facie but were used to defend Ye and echo similar antisemitic narratives from his statements, namely #KanyeIsRight, #YeIsRight, and #TheNoticing.

Posts including #YeIsRight or variations thereof were mostly aimed at agreeing with his statements perpetuating stereotypes against Jews and promoting Jewish conspiracy theories. #TheNoticing was used primarily to "call out" various alleged Jewish conspiracies in-line with Ye's remarks.

> **2. Violating hashtags** | Flagrant antisemitic hashtags that clearly violate Twitter's rules but were not blocked on the platform and are still searchable today, such as, #HitlerWasRight and #Kikes.
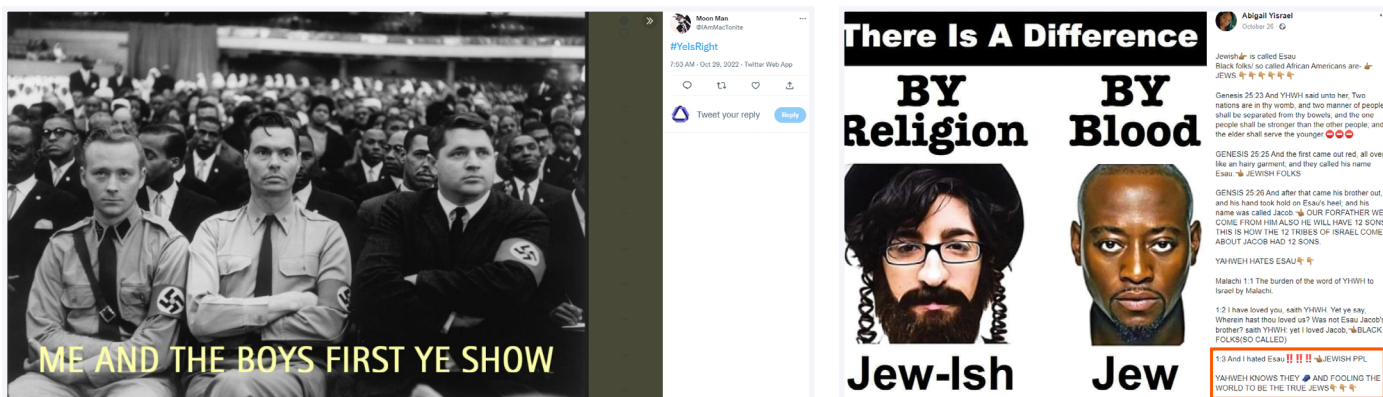


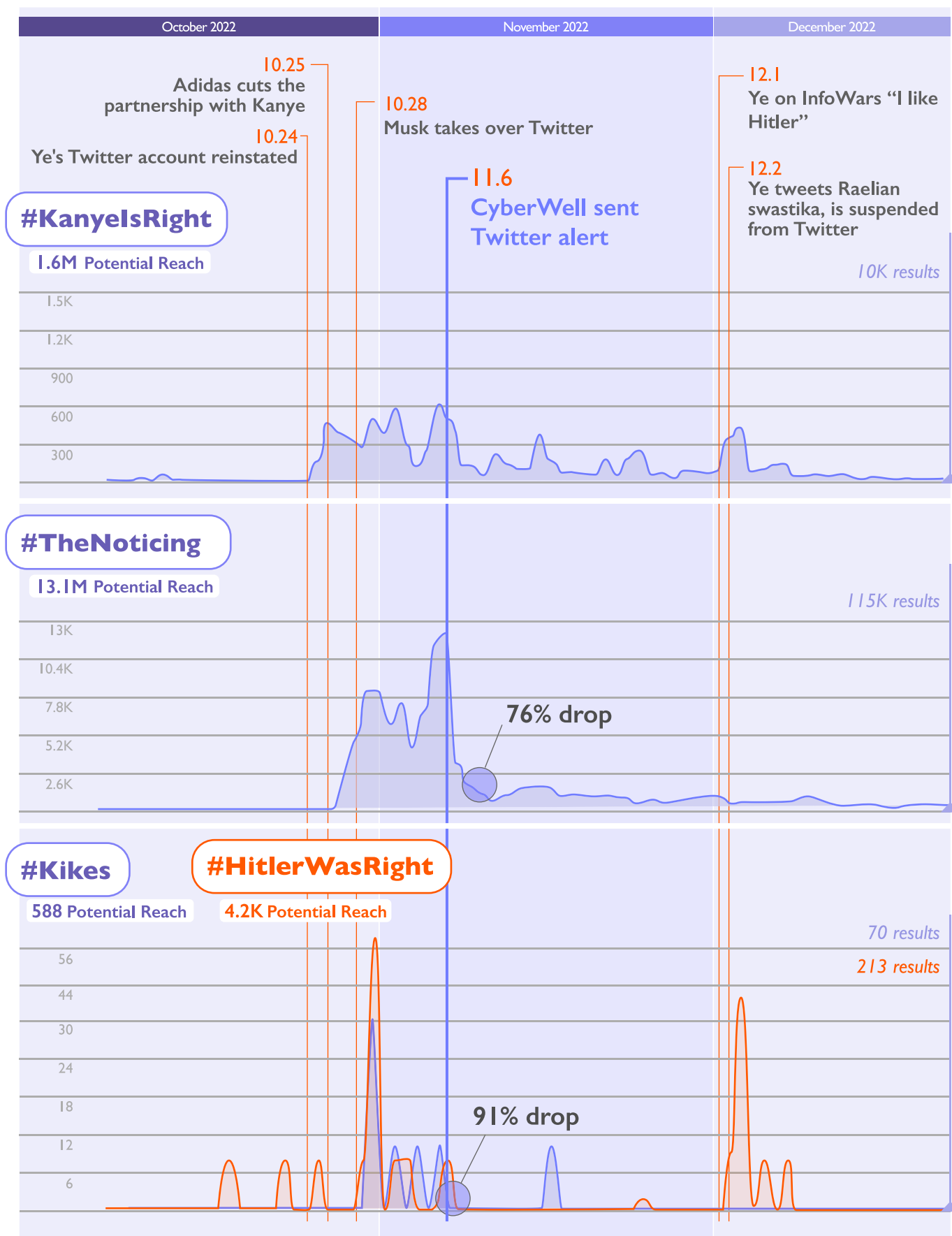We sounded the alarm by being the first organization to reach out with data-based alerts **tailored** to each social media platform and demanded accountability, highlighting how each platform was serving as an echo chamber for the harmful antisemitic tropes that Ye and other influential celebrities were promoting on their accounts.

Meta acknowledged receipt of our alerts detailing this trend on Facebook and Instagram. While Twitter did not respond, within 24 hours of issuing the alert, two hashtags flagged by CyberWell, #TheNoticing and #Kikes, dropped off sharply in their exposure on the platform (an exposure decline of 76% and 91%, respectively), indicative of Twitter's likely intervention in limiting their circulation and exposure. Refer to figure 4 for a visualization of the timeline.

> Total potential reach of these antisemitic # hashtags on Twitter: 15.17 million

---

18  "Kanye" or variations of the words "Black people" are not part of CyberWell's regular monitoring lexicon.

# Timeline: Ramifications of Ye on Twitter

| October 2022 | November 2022 | December 2022 |
|---|---|---|

**10.25**
Adidas cuts the partnership with Kanye

**10.24**
Ye's Twitter account reinstated

**10.28**
Musk takes over Twitter

**11.6**
CyberWell sent Twitter alert

**12.1**
Ye on InfoWars "I like Hitler"

**12.2**
Ye tweets Raelian swastika, is suspended from Twitter

## #KanyeIsRight

**1.6M** Potential Reach

*10K results*

| | |
|---|---|
| 1.5K | |
| 1.2K | |
| 900 | |
| 600 | |
| 300 | |

## #TheNoticing

**13.1M** Potential Reach

*115K results*

| | |
|---|---|
| 13K | |
| 10.4K | |
| 7.8K | |
| 5.2K | |
| 2.6K | |

**76% drop**

## #Kikes

**588** Potential Reach

## #HitlerWasRight

**4.2K** Potential Reach

*70 results*

*213 results*

| | |
|---|---|
| 56 | |
| 44 | |
| 30 | |
| 24 | |
| 18 | |
| 12 | |
| 6 | |

**91% drop**

**#Kikes** was previously blocked on Twitter, only recently reintroduced.

*These numbers reflect a sample of Tweets with the highest engagement rates and NOT a total number of Tweets. However, some of these #hashtags also peaked on other social media platforms.

## New Words Trending

| | |
|---|---|
| **"Kanye"** | **+24,700% increase** |
| **"Black people"** | **+218% increase** |
| **"Media"** | **+40% increase** |
| **"Synagogue of Satan"** | **+15% increase** |

*\*Average appearance rate of words Oct 6 – Dec 5 in CyberWell's online antisemitism monitoring tech when compared with the previous two months.*

# Social Media Accountability Crisis

The increase in online antisemitism is part of the wider social media accountability crisis. As advocacy, research, stakeholder collaboration, and lawmaking efforts expand, social media platforms have taken steps to publicize what they are doing to fight online hate, including participating in public hearings,[19] publishing quarterly reports,[20] rolling out transparency centers,[21] and even undergoing independent audits for brand safety.[22]

However, these welcome steps fail to do enough to address antisemitism, alongside hatred of other marginalized communities, because these measures are not matched with a commitment to regular metrics and content transparency by these same platforms regarding the posts that are being left online. There are no public metrics to definitively conclude that the levels of online hate speech have decreased across social media platforms. The limited self-reported data on rates of online hate speech on social media platforms and third-party sprint research by civil society and government-funded compliance would indicate that the rates of online hate speech remained relatively consistent during 2022, if not increased.

Meta, the parent company of Facebook and Instagram, self-reported that online hate speech prevalence remained consistent in 2022,[23] yet the rate of **actioning hate speech actually *decreased*** ("actioning" - i.e., either limiting the exposure of problematic content or removing it completely). Twitter leadership admitted to increased levels of hate speech on their platform,[24] especially during Musk's takeover, but insisted that the rate of actioning hate speech at the time remained unchanged.[25] While YouTube self-reported an increase in the

19  C-SPAN, Social Media Executives Testify on Reducing Hate Online, published 09/16/22, accessed 01/23.

20  Meta, Community Standards Enforcement Report: Q3 2022, accessed 01/24/23; TikTok, Community Guidelines Enforcement Report: July 1, 2022-September 30, 2022, published 12/19/22, accessed 01/23/23; Twitter, Rules Enforcement, accessed 01/23/23, YouTube, YouTube Community Guidelines Enforcement, accessed 01/23/23.

21  TikTok, Strengthening our commitment to transparency, published 07/27/23, accessed 01/24/23.

22  Meta Business Help Center, Meta's Brand Safety Description of Methodology, accessed 01/24/23.

23  Meta, Community Standards Enforcement Report: Hate Speech, accessed 01/23/23.

24  Twitter, Yoel Roth [@yoyoel], Over the last 48 hours, we've seen a small number of accounts post a ton of Tweets that include slurs and other derogatory terms. To give you a sense of scale: More than 50,000 Tweets repeatedly using a particular slur  came from just 300 accounts, published 10/10/22, accessed 02/23; Twitter, Yoel Roth [@yoyoel], Since Saturday, we've been focused on addressing the surge in hateful conduct on Twitter. We've made measurable progress, removing more than 1500 accounts and reducing impressions on this content to nearly zero. Here's the latest on our work, and what's next., published 11/01/22, accessed 02/23.

25  Twitter, Elon Musk [@elonmusk], Again, to be crystal clear, Twitter's strong commitment to content moderation remains absolutely unchanged. In fact, we have actually seen hateful speech at times this week decline \*below\* our prior norms, contrary to what you may have read in the press., published 11/05/22, accessed 02/23; Twitter, Yoel Roth [@yoyoel], More than 80% of our incoming content moderation volume was completely unaffected by this access change. The daily volume of moderation actions we take stayed steady through this period., published 11/05/22, accessed 02/23.

amount of removal of videos and comments constituting hate speech on the platform, [26] they did not release metrics about how much of the content on the platform was hate speech. TikTok also reported an increase in the rate of removal of hate speech across the platform, but remained silent about what percentage of the content on TikTok constitutes hate speech.

| | Disclosure of Hate Speech Prevalence in Platform | Rate of Actioning Hate Speech Content |
|---|---|---|
| Facebook | Hate Speech Rates Consistent | ↘ Decreased |
| Instagram | Hate Speech Rates Consistent | ↘ Decreased |
| TikTok | Did not release metrics | ↗ Increased |
| Twitter | Hate Speech Rates Increased *In public statements* | — Remain unchanged |
| YouTube | Did not release metrics | ↗ Increased |

According to the European Commission, the seventh monitoring exercise tracking the enforcement of the Code of Conduct on countering illegal hate speech online found that the time to remove hate speech and overall removal rate has declined significantly in the last two years. The number of notifications reviewed by the social media companies within 24 hours consistently dropped, from 90.4% in 2020, to 81% in 2021, and 64.4% in 2022, a 26% decline in 24-hour removal rate of hate speech over two years. The same exercise showed a decrease in overall removal rate, at 63.6% removal of illegal hate speech in Europe across all platforms in 2022 compared to a 71% removal rate across all platforms in 2020.

One of the major announcements issued by Twitter following Musk's takeover in October 2022 introduced the concept of "freedom of speech, not freedom of reach." According to Twitter's 2.0 announcement, while none of the Twitter rules changed, the approach to policy enforcement will rely on this de-amplification of violating content. [27]

While this may have been news to supporters of Musk, other mainstream platforms use variations of this method of de-amplification as a tool of policy enforcement, essentially ranking certain keywords and images for lower exposure and making them ineligible for monetization via advertisement. [28] However, de-amplification and rollout of other new policies, like self-reporting transparency centers and external audits, are being presented as solutions to the state of online hate by the social media platforms in the absence of regular metrics and content transparency, i.e., full disclosure on the posts that are not being removed.

---

26  YouTube, Featured Policies: Hate Speech, accessed 01/23; YouTube, YouTube Community Guidelines Enforcement, accessed 01/23.

27  Twitter Inc., Twitter 2.0: Our continued commitment to the public conversation, published 11/30/22, accessed 01/23/23.

28  Meta, Reducing the distribution of problematic content, updated 10/04.22, accessed 01/24/23; Twitter, Guy Rosen [@guyro], There is no silver bullet to reducing prevalence of violating content on the platform and we need to use multiple tactics to make progress consistently and accurately, published 11/09/21, accessed 02/23.

## Why De-Amplification Cannot be the Silver Bullet Solution for Violating Content

In the last 30 days before the finalization of this annual report, CyberWell monitored more than 40 flagrantly antisemitic Tweets that violated Twitter rules. Though every Tweet was reported to Twitter by CyberWell, 30 of them remained online[29] and were viewed over 6,500 times.

Similarly, during the last week of January surrounding International Holocaust Remembrance Day, antisemitic hashtags promoting Holocaust denial and distortion including #TheHolocaustIndustry, #NumbersDon'tAddUp, and #Holohoax reached over 2 million people on Twitter alone. Yet these numbers, however alarming to members of the Jewish community and allies in the fight against hate, may not qualify as high enough levels of exposure according to the standards of each platform, and therefore will continue to remain online, with no recourse. De-amplification being cited as a 'solution' or alternative for holistic removal of violating content will continue to result in regular and increasing peaks of promotion of hateful content, typically against minority user-groups.

Content moderation by the social media companies is increasingly rooted in AI and other automated flagging and blocking technology but supplanted with the human review of content teams[30] to address unfolding events, sharp changes in discourse, and above all else, to catch what the AI misses. These content teams are increasingly outsourced[31] and have recently been scaled back due to cuts in advertiser spending on the platforms.[32]

While the rollout and development of additional AI and tech-based review will ideally, over time, decrease the need for robust content review teams, high levels of online antisemitism and dramatic surges like that provoked by Ye indicate that the technology focused on antisemitic content simply isn't advanced enough or deployed as completely as it should be across the platforms. As such, CyberWell's full-time monitoring and data aggregation process will be a key to developing an effective antisemitism digital policy compliance solution. Without clear content transparency and metrics on the hate content that is left online, social media platforms will continue to serve as the most powerful amplifiers of hatred against minority communities.

> "When I started using social media to talk about Jewish and Jamaican-Jewish history, I was inundated with hate, from death threats to the usage of slurs against me. This abuse only got worse if I dared mention my love of Israel. Rather than focusing on educating people about Jewish history, I now have to police my notifications to hide & block antisemitic comments on my posts. It's an exhausting existence, and I do it because I have no faith anymore that social media platforms are acting with a solid will to remove those who perpetrate this old virus of hatred. Free speech does not mean freedom from consequences or that Jews must flee social media to protect themselves from antisemitism. We need social media companies to take it seriously because we now live in a world where an antisemitic tweet can radicalize a person to use violence against Jews."
>
> - Tyler Samuels 🐦 *@BluntBlackJew*

---

29  As of January 19, 2023

30  Axios, Scoop: Oracle begins auditing TikTok's algorithms, published 08/16/22, accessed 02/23; Just Security, It's Past Time to Take Social Media Content Moderation In-House, published 01/18/23, accessed 01/23; Meta, Transparency Center: Detecting Violations, accessed 02/23; Politico, What happened when humans stopped managing social media content, published 10/21/20, accessed 01/23; TechCrunch, Oracle now monitoring TikTok's algorithms and moderation system for manipulation by China's government, published 08/16/22, accessed 02/23; YouTube, Responsible policy enforcement during Covid-19, published 08/20, accessed 02/23.

31  Time, Under Fire, Facebook's 'Ethical' Outsourcing Partners Quits Content Moderation Work, published 01/10/23, accessed 01/23.

32  Twitter, Roth, Yoel [@yoyoel], "Yesterday's reduction in force affected approximately 15% of our Trust & Safety organization (as opposed to approximately 50% cuts company-wide), with our front-line moderation staff experiencing the least impact", published 11/05/22, accessed 02/23; PleasantonWeekly.com, Terminated Meta content moderators worry about fake news flourishing in their absence, published 01/28/23, accessed 02/23; TechCrunch, Former TikTok content moderators file lawsuit over 'psychological trauma', published 03/25/22, accessed 01/23.

# Recommendations for a Better Digital Future

Based on CyberWell's data and dive into digital hate policies and enforcement, we have identified several key recommendations to effectively combat online hate generally and digital Jew-hatred specifically.

1. Social media platforms should be required to disclose all reported hate content to regulators, researchers, and civil society organizations, ensuring transparency of all reported hate content that social media platforms do not remove.

2. Social media platforms should be required to disclose engagement rates of all reported but unremoved hate content to regulators, researchers, and civil society organizations.

3. Brand safety audits, such as those conducted by the Media Ratings Council, should include transparent ad-reviews for all reported hate content that social media platforms do not remove.

4. All social media platforms should update their existing digital community guidelines to reflect that use and promotion of the Protocols of the Elders of Zion and use of the Synagogue of Satan concept to demonize Jews or Zionists is antisemitic and constitutes hate speech.

5. Social media platforms should adopt or use the International Holocaust Remembrance Alliance's working definition of antisemitism as part of their community guidelines and for training internal AI-powered resources to identify Jew-hatred.

6. Social media platforms should close the enforcement gap between English and non-English languages. Appropriate resource allocation in terms of dedicated algorithmic development and additional content review teams in non-English languages would decrease the presence of online hate speech and incitement to violence overall and address the growing level of digital Jew-hatred in Arabic and other languages.

# CyberWell in 2023

CyberWell has already created meaningful data insights and led real-time responses since our launch. In order to expand our capacities and remain vigilant against online Jew-hatred, we need your help. Maintaining our activities as a non-profit requires visionary philanthropic involvement and stalwart partnerships with like-minded organizations and activists.

During 2023 CyberWell seeks to launch a database for professional use that will allow other NGOs, researchers, lawmakers, journalists, and educators to gain access to deeper data insights, visualization tools, and raw data access to further their own efforts and expand their knowledge base. We will also serve as an advanced data aggregator by hosting data from organizations that either collect examples as a regular part of their activity or are the subject of online Jew-hatred, generating comprehensive transparency and visual aids to reflect the state of online antisemitism.

Data aggregation is the key to creating the transparency necessary to successfully beat back online Jew-hatred, and when engaging in large-scale data collection, having the capacity to review the data is a core challenge. CyberWell maintains the highest threshold possible of data-vetting through human-in-the-loop examination when content is flagged by AI as highly likely to be antisemitic.

All data featured in CyberWell's database is reviewed by two professional analysts: once for antisemitism categorization and again for digital policy compliance. With over 100,000 pieces of content already in the pipeline awaiting review, we are in need of philanthropic support to expand our analysis and research team.

The rolling out of the Digital Services Act (DSA) in the European Union (EU) marks an important turning point for driving social media accountability for hosting illegal hate speech, including illegal virulent antisemitism and Holocaust denial, prohibited by EU law.[33] To ensure that accountability for hosting digital Jew-hatred is part of the enforcement process of the DSA, CyberWell plans to expand the languages we monitor to include German, French, Spanish, and Italian, among other languages.

It is estimated that approximately half of antisemitic incidents offline worldwide emanated from Europe in 2021, and roughly 30% took place in the United States.[34] Therefore, monitoring the online component of Jew-hatred in European languages is absolutely crucial to curbing the overall rise of antisemitism that plagues Jewish communities worldwide. This process is not inexpensive, as previous research[35] and CyberWell's own data collection demonstrates, and social media platforms are dedicating proportionately less resources to enforce community guidelines in non-English speaking countries. Expanding monitoring capabilities into other languages requires dictionary development, as well as training and hiring native-level speakers.

---

33  European Council, Council Framework Decision on combatting certain forms and expressions of racism and xenophobia by means of criminal law 2008, published 12/06/08, accessed 02/23; The European Journal of International Law, Volume 26 No. 1, Holocaust denial before the European Court of Human Rights: Evolution of an Exceptional Regime, p. 239, published 2015, accessed 02/23; Yad Vashem The World Holocaust Remembrance Center, Holocaust Denial Laws and Other Legislation Criminalizing Promotion of Nazism, published 2006, accessed 02/23.

34  Ministry of Foreign Affairs, WZO and Jewish Agency report on antisemitism in 2021, published 01/24/22, accessed 01/24/23; The Jewish Agency for Israel, World Zionist Organization, The State of Antisemitism in 2021, published 01/22, accessed 01/24/23.

35  Institute for Strategic Dialogue, Under-Moderated, Unhinged and Ubiquitous: Al-Shabab and the Islamic State Networks on Facebook, published 06/15/22, accessed 02/23.

Finally, visual platforms, such as TikTok, Instagram, and YouTube, continue to be more widely adopted for use than text-based platforms, particularly with the youngest groups of social media users.[36] With a higher adoption rate by the youngest demographic of users, it is crucial to increase our capacity to accurately and thoroughly monitor these visual platforms. With your support, we will be able to roll out much-needed image analysis and video-to-text technology, which will increase our ability to flag antisemitic content on visual platforms.

The support we have received from philanthropists and communities passionate about innovative solutions to tackle the rise of Jew-hatred has inspired us and enabled CyberWell to launch the first ever open database of online antisemitic content, as well as deploy critical alerts to social media platforms, effecting change in real time. While the support has been overwhelming, these important expansions and features must be introduced to ensure that we can address the continually changing social media space and capture online Jew-hatred where and when it is happening. We look forward to continuing to partner with our earliest visionary adopters, the first supporters of CyberWell, as well as work with new leaders looking to invest in a brighter digital future for the Jewish people.

"The antisemitic comments and posts I have encountered on apps such as TikTok have increased in recent years and, with it, the fear of being a Jewish student on a secular college campus. My encounters with antisemitism on my campus are in no way unique, as this issue stems from fast-spreading antisemitic hate speech online. Antisemitic rhetoric goes beyond our screens. To reduce the rapidly growing hate surrounding Jews globally, it is vital to tackle the problem at its core via hate speech and threats on social media."

- Emannuel Hutman ♪ @adventureswithem

---

36  Pew Research Center, Teens, Social Media and Technology 2022, published 08/10/22, accessed 02/2023.

# Partnership Activity:
# Democratizing Data to Fight Jew-hatred

## Adopt IHRA Coalition

On the heels of Elon Musk's acquisition of Twitter, we were approached by the Adopt IHRA Coalition - a 180-member coalition advocating for the adoption of the IHRA working definition of antisemitism in online community standards.

We served as the coalition's data provider, demonstrating our value through collecting, vetting, and leveraging a dataset of over 1,200 recent antisemitic Tweets during this unique transition period to point out where Twitter content review and policy enforcement fails, highlighting the clear need for more stringent community standards. Check out the detailed report of our findings at https://cyberwell.org/reports/.
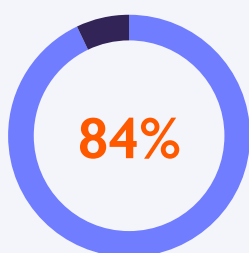
**84%**
of vetted Tweets violated the Twitter Rules
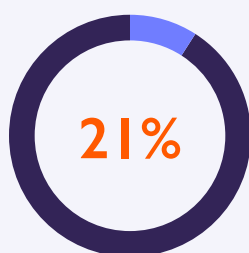
**21%**
Rate of removal **before** CyberWell reporting
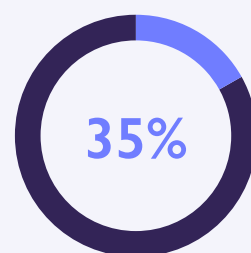
**35%**
Rate of removal **after** CyberWell reporting

We joined the **Shine a Light** campaign, which takes place annually throughout Hanukkah. Tal-Or was featured in their Speakers Bureau and we contributed a document entitled "Know Your Digital Rights" with the aim of empowering social media users to feel confident in demanding their right to a harassment-free experience.

The **Simon Wiesenthal Center** used CyberWell's analysis of the wave of online antisemitism post-Ye and cited it as the worst antisemitic incident in their Top Ten List of Worst Anti-Semitic Incidents of 2022.[1]

___

1   Simon Wiesenthal Center, SWC Releases Top Ten List of Worst Antisemitic Incidents of 2022, published 12/22/22, accessed 01/23.

CyberWell served as the data provider to **Act-IL**'s community for their end of year call to action on the state of online antisemitism.

## CyberWell hosts Bavarian Minister of Digital Affairs

In June, CyberWell hosted a delegation from the Bavarian Ministry of Digital Affairs, including Minister of State for Digital Affairs Judith Gerlach. We shared our work and demonstrated our reporting platform in action. Minister Gerlach was stunned by the data and horrified by the examples of antisemitism in German still online after being reported as demonstrated to her through our database. Germany has some of the most stringent anti-hate policies, and we hope to see this partnership continue to develop. Read the German press release to here:

https://www.bayern.de/gegen-fake-news-und-internet-hetze-digitalministerin-gerlach-fordert-kennzeichnungspflicht-fuer-kuenstliche-intelligenz-bei-facebook-twitter-und-co-meinungsfreiheit-gilt-fuer-menschen-n/

## Students take back their digital spaces

We trained 70 university students who visited Israel on the Hasbara Fellowships program on knowing their digital rights, how to recognize online antisemitism, use our platform, and report hate content. We then held a mini hackathon and shared how these young ambassadors can further engage students on their own campuses to fight back against digital Jew-hatred using the CyberWell platform as a resource. They were inquisitive, passionate, and determined, and we look forward to working with additional student ambassadors in the future.

## In the News

The Jerusalem Post shared the story of our founding, and the Jewish News Syndicate published a profile on our alerts to the platforms and how they were acting as amplifiers during Ye's public litany against the Jewish people.

CyberWell was quoted and our contribution to the Adopt IHRA Coalition discussed in the Jewish News Syndicate, Algemeiner and Jewish Journal, among others.

Tal-Or was interviewed on the podcast Power of the Crowd by Esther Jaromi, a Ph.D. student at Queen Mary University of London and a Legal Advisor to the EU Delegation to the United Nations, who expressed how important a tool like CyberWell is for academic research.

The launch of our public platform was featured in the Kansas City Jewish Chronicle.

### THE JERUSALEM POST

**CyberWell – Using data to reduce online antisemitism**

CyberWell is the first open live database of antisemitic content. It provides a wealth of information so that people can see the state of online antisemitism and analyze it as well.

THE CYBERWELL team at their offices in Tel Aviv. (From left), Vered Andre'ev, head of research; Lara Portnoy, program and ops manager; Tal-Or Cohen Montemayor, founder and executive director; Rachel Brynien, partnerships and growth leader.
(photo credit: Hagar Bader)

---

**jns** Jewish News Syndicate

**Can a company curb online Jew-hatred?**

CyberWell monitors five social media platforms and hopes to get them to crack down on antisemitism.

**JEWISH JOURNAL**
Connect. Inform. Inspire.

**Coalition Sends Letter to Elon Musk Urging Twitter to Adopt International Antisemitism Definition**

The KANSAS CITY **jewish chronicle**

Fighting Hate: Olim are using Israeli tech to fight online antisemitism across the globe

the **algemeiner**

**In Letter to Twitter's Elon Musk 180 NGOs, Civil Rights Groups Call For Adoption of the IHRA Definition of Antisemitism**

**Spotify**

PODCAST EPISODE
Innovation for Social Good: Tal-Or Cohen on Harnessing the Power of Technology to Fight Online Hate
Power of the Crowd

## United for Success

### CyberWell Team

Tal-Or Cohen Montemayor, CEO & Founder
Vered Andre'ev, Head of Research
Lara Portnoy, Operations & Program Manager
Rachel Brynien, Partnerships & Growth Lead
Yonathan Hezroni, Digital Compliance Analyst
Sagi Balasha, CFO

### Advisory Board

Col. (Res.) Miri Eisin
Zohar Gorgel
Professor Gunther Jikeli
Lt. Colonel (Res.) Peter Lerner
Professor Dina Porat
Maj. Gen. (Ret.) Amos Yadlin

CyberWell is thrilled to acknowledge the invaluable contribution of our summer intern, Joseph Heinemann, a Pepperdine University student class of 2024 and a Seattle native. Read his blog about his experience interning with us.

### With Special Thanks

Joseph Hyman and the Center for Entrepreneurial Jewish Philanthropy

Arik Becker

Yonatan Cohen

Cassandra Cyphers

Adam Haliva

Aviva Zeltser

Dr. Giovanni Quer

> "By addressing online antisemitism and being a part of the movement that's fighting back against online Jew-hatred, I feel like I've been given an opportunity to show that I have power over the hate speech that has affected me in the past."
>
> - Joseph Heinemann, Pepperdine University class of '24

To all our philanthropic supporters - we are grateful for your contribution and honored by the trust you place in us to tackle the urgent and enormous challenge of online antisemitism. The work we do matters, and your generosity and encouragement bolster our resolve.

Thank you.

To contribute to CyberWell's work or renew your commitment, please reach out to hello@cyberwell.org or send donations directly to our US fiscal sponsors, earmarked to CyberWell.

*Help us stop online Jew-hatred!*
Report online antisemitism through our platform today.

CyberWell