

Monitoring and Reporting Exercise 2023/2024

/1st evaluation round



The report was prepared as part of the project:

"Securing Our Community, Preserving Our Democracy: Combating Antisemitism through Integrated Advocacy and Security Approaches" (PROTEUS project)

The project was co-funded by the European Union.

Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.



G | M | F
Transatlantic
Foundation

The project was also co-funded by the EVZ Foundation.



This report does not represent the opinion of the EVZ Foundation. The author bears responsibility for any statements contained herein.

ISBN 978-83-972977-0-8



Jewish Association Czulent ul. Św. Sebastiana 36/1, 31-051 Kraków www.czulent.pl

Kraków 2024

Support the Jewish Association Czulent 36 Swietego Sebastiana St., apt. 1 31-051 Krakow

Polish account numbers in the IBAN format: PL42 1600 1462 1880 9351 6000 0001

BIC / SWIFT code for international transfers: PPABPLPK

We need your help to protect our Jewish community and combat antisemitism



Contents

- 4 Introduction
- 5 Glossary
- 7 The legal framework
- 9 Background
- 10 Methodology of the MRE
- 12 Key figures
- 14 Removal rates
- 17 Legal grounds of the reported hateful content
- 18 Conclusions and recommendations

Introduction

Antisemitic content is present and widely disseminated on the internet, constituting a threat to democracy and public security. As highlighted in analyses and reports prepared by international organizations such as the Institute for Strategic Dialogue (ISD)^{1,2} and the Kościuszko Institute, antisemitism serves as a tool for disinformation and societal polarization.³

Monitoring of antisemitic incidents conducted by the Jewish Association Czulent further indicates that antisemitic crimes are predominantly committed online and negatively affect the sense of security of members of the Jewish community and Jewish organizations.⁴

Our goal was to identify which online services effectively remove hateful content to ensure they are safe spaces for users. We aimed to determine whether the content reported by users is evaluated and removed in the same manner as content reported by trusted flaggers. For this purpose, we conducted the Monitoring and Reporting Exercise (hereafter, the MRE or the Exercise) evaluating international online platforms and Polish IT services on the application of national and EU laws. The Exercise was carried out within the new EU digital legal framework known as Regulation (EU) 2022/2065 or the Digital Services Act (hereafter, the DSA) that aims to create a safer environment for users on the internet.

¹ Institute for Strategic Dialogue, B'nai B'rith International, Online Antisemitism: A Toolkit for Civil Society, 2022.

² Institute for Strategic Dialogue, Narratives of Hate: Post-7 October Antisemitism and Anti-Muslim Hate on Social Media, 2024.

³ Civic Resilience Initiative, Detector Media i Instytut Kościuszki, Odporność na dezinformacje w krajach trójkąta lubelskiego, 2022.

⁴ The Jewish Association Czulent, <u>Antysemickie incydenty w 2023 roku. Badanie doświadczenia społeczności żydowskiej,</u> 2024.

Glossary

The Monitoring and Reporting Exercise developed by the Jewish Association Czulent uses specific terms applied in the EU digital environment on counteracting the dissemination of illegal content online, such as hate speech. Below is the list of terms applied in the Exercise and used in this Report.

antisemitism — the Jewish Association Czulent uses the IHRA working definition of antisemitism already recognized by the Polish government in 2016 and confirmed by the Polish Ministry of Culture and National Heritage in 2021:5 "antisemitism is a certain perception of Jews, which may be expressed as hatred toward Jews. Rhetorical and physical manifestations of antisemitism are directed toward Jewish or non-Jewish individuals and/or their property, toward Jewish community institutions and religious facilities".6

content moderation — "content moderation' means the activities, whether automated or not, undertaken by providers of intermediary services, that are aimed, in particular, at detecting, identifying, and addressing illegal content or information incompatible with their terms and conditions, provided by recipients of the service, including measures taken that affect the availability, visibility, and accessibility of that illegal content or that information, such as demotion, demonetization, disabling of access to, or removal thereof, or that affect the ability of the recipients of the service to provide that information, such as the termination or suspension of a recipient's account".

DSA or Regulation 2022/2065 — Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC — Digital Services Act⁸

escalation — notices (notifications) of illegal content submitted by entities with trusted flagger or trusted partner status to very large online platforms via designated forms provided by a VLOP

hate speech — there is no legal definition of hate speech in Poland. We use the definition adopted by the Council of Europe in 1997: "(...) "hate speech" shall be understood as all forms of expression which spread, incite, promote, or justify racial hatred, xenophobia, antisemitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin." 9,10

- 5 Ministry of Culture and National Heritage, <u>Statement on the IHRA working definition of antisemitism.</u>
- 6 IHRA, IHRA working definition of antisemitism.
- 7 Regulation 2022/2065, art. 3 (t).
- 8 Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ 2022 L 277/1.
- 9 Council of Europe (1997), Recommendation No. R(97)20 of the Committee of Ministers to Member States on "hate speech".
- 10 See the new definition of hate speech adopted by the Council of Europe in 2022: Council of Europe (2022).
 Recommendation CM/Rec (2022) 16 of the Committee of Ministers to Member States on combating hate speech.

IHRA — International Holocaust Remembrance Alliance¹¹

illegal content — "any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law"¹²

intermediary services — "(...) a 'hosting' service, consisting of the storage of information provided by, and at the request of, a recipient of the service"¹³

MRE — Monitoring and Reporting Exercise

online platform — "a hosting service that, at the request of a recipient of the service, stores and disseminates information to the public, unless that activity is a minor and purely ancillary feature of another service or a minor functionality of the principal service and, for objective and technical reasons, cannot be used without that other service, and the integration of the feature or functionality into the other service is not a means to circumvent the applicability of this Regulation"¹⁴

PC — Polish Penal Code¹⁵

trusted flagger — entities recognized for their expertise in flagging illegal content. Under the DSA, the status "should be awarded by the Digital Services Coordinator of the Member State in which the applicant is established and should be recognized by all providers of online platforms within the scope" of the DSA.¹⁶

trusted partner — entities that have been granted special status by VLOPs allowing for flagging or escalating reports on illegal content or material violating VLOPs' community guidelines. The status is granted on a voluntary basis and under the platform's terms.

VLOP — a very large online platform¹⁷ including Facebook, Instagram, TikTok, X, YouTube¹⁸

VLOSE — a very large online search engine

¹¹ International Holocaust Remembrance Alliance.

¹² Ibidem, art. 3 (h).

¹³ Regulation 2022/2065, art. 3 (g) (iii).

¹⁴ Ibidem, art. 3 (i).

¹⁵ Penal Code (<u>Ustawa z dnia 6 czerwca 1997 r. - Kodeks Karny</u>), Polish Journal of Laws of 1997 n 88 item 553.

¹⁶ Ibidem, art. 22.

¹⁷ See also: European Commission, Supervision of the designated very large online platforms and search engines under DSA, 2024.

¹⁸ Regulation 2022/2065, section 5.

The legal framework

The DSA aims to harmonize the rules applicable to providers of intermediary services. It regulates their responsibilities in the EU with the objective of ensuring "a safe, predictable and trusted online environment, addressing the dissemination of illegal content online and the societal risks that the dissemination of disinformation or other content may generate." ¹⁹ The illegality of online content reflects the offline environment. In other words, it should be understood broadly and regardless of form, whether it is content, products, services, or activities. ²⁰ The DSA does not provide a clear-cut definition of what constitutes an "illegal activity" or "illegal content". It stresses that the concept of illegality should be assessed in line with EU and national laws. ²¹

Under EU Council Framework Decision 2008/913/JHA, EU Member states are obliged to penalize "public incitement to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin".²²

Polish Penal Code (PC) does not provide a separate legal definition of hate speech. However, it contains a group of legal provisions aimed at tackling the phenomenon of hate speech, which are also applicable in the online environment. These provisions prohibit, among others: making unlawful threats towards a person or a group of persons (Art. 119 of the PC), promoting a fascist or other totalitarian state system or inciting hatred (Art. 256 of the PC), or insulting a specific population or individual (Art. 257 of the PC). Nevertheless, these provisions can only be applied if the above-mentioned crimes are motivated by the national, ethnic, racial, political, or religious affiliation of an attacked individual or group. This means that only some categories of hate speech are protected (among them antisemitic hate speech). Others (like homophobic hate speech) are not protected within this special framework and require the application of general provisions (for instance, prohibiting insulting other persons). Separately, the Act on the Institute of National Remembrance prohibits denying nazi crimes and other crimes against humanity committed between 1917 and 1990 (in its Art. 55).²³

Among Polish national laws, it is also important to mention Art. 14 of the Polish Act on providing services by electronic means, which states that providers are not liable for storing illegal content if they remove it after receiving reliable information on it.²⁴

- 19 Regulation 2022/2065, recital (9).
- 20 Ibidem, recital (12).
- 21 Ibidem, recital (9) and art. 3 (h).
- 22 Art. 1 (1) (a) of the Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law, OJ 2008 L 328.
- 23 The Act on the Institute of National Remembrance, Polish Journal of Laws of 2016 items 152, 178, 677, 749.
- 24 The Act on providing services by electronic means (<u>Ustawa z dnia 18 lipca 2022 r. o świadczeniu usług drogą elektroniczną</u>), Polish Journal of Laws of 2002 no 144 item 1204.

This provision should be applicable to content moderation in Poland, especially since the DSA provisions are not fully applicable in the country. Since the end of **August 2023**, the DSA has been applied to very large online platforms (VLOPs) and very large online search engines (VLOSEs). VLOPs and VLOSEs are defined as designated platforms with more than 45 million users in the EU. **On February 17, 2024**, the DSA became fully applicable. This means that all service intermediaries, offering their services in the single market, irrespective of their place of their establishment, have to comply with the new rules.²⁵

Providers of intermediary services including very large online platforms, are not obliged to monitor for illegal activity with respect to the information that they transmit or store.²⁶ The DSA clearly states in Art. 6 of the Regulation that online service providers are not liable for storing information if they are not aware of the illegal activity or illegal content or if, after obtaining such knowledge, they act expeditiously to remove or disable access to the illegal content.²⁷

The content detection and reporting roles that play an essential role in creating a safer online environment have been granted under Art. 22 of the DSA to "trusted flaggers". The status of "trusted flaggers" is granted, upon application, by a Digital Services Coordinator of a Member State to entities having expertise in a specific area. As stated in Art. 22 (2) (a) of the Regulation, the applicant entity should demonstrate a "particular expertise and competence for the purposes of detecting, identifying and notifying illegal content". 28

At the MRE's publication date, Poland has not amended the national law to provide for the full application of the DSA. It has not designated an institution that would play the role of the Digital Services Coordinator responsible for supervising, enforcing, and monitoring the DSA, including granting, upon an application, "trusted flaggers" status to designated entities. Nevertheless, the Jewish Association Czulent has engaged in monitoring and reporting of antisemitic illegal content, applying general user mechanisms and partnering with trusted flaggers or trusted partners in submitting notifications. Despite the lack of implementation of the Digital Services Act into Polish law, Regulation 2022/2065 is binding within the European Union.

The Exercise presents the reporting and monitoring results concerning illegal online content as set up by the DSA. All reported antisemitic content violated provisions of the Polish Penal Code and met the requirements of the IHRA working definition of antisemitism.

28 Ibidem, art. 22 (2) (a).

²⁵ European Commission, The Digital Services Act.

²⁶ Regulation 2022/2065, art. 8.

²⁷ Ibidem, art. 6.

In light of the above and upon the application of national and EU laws, this Report presents major facts regarding content moderation by foreign and Polish IT companies within the DSA framework, including the number and rates of reported incidents (notices or notifications), time of assessment of notices, removal rates and legal grounds of the reported illegal hateful content. The Exercise was inspired by a yearly evaluation of the EU Code of conduct on countering illegal hate speech online.²⁹

Background

The MRE is a comparative study of the DSA application in relation to the removal of illegal online content, with a focus on antisemitism, by very large online platforms and Polish providers of intermediary services. The Exercise was conducted from November 1, 2023 to April 7, 2024.

The Jewish Association Czulent identified, collected, and reported a total of **174** illegal online incidents motivated by antisemitism to IT companies, including **129** on VLOPs and **45** on Polish online platforms. These reported illegal incidents took the form of posts, videos, and pictures and violated Polish criminal provisions. They also constituted a criminal offense as defined by EU Council Framework Decision 2008/913/JHA.³⁰

The Exercise focuses on the number of notifications (interventions), understood as reports sent to IT companies. This approach stems from the aim to evaluate the conduct of IT companies in response to the notices provided by general users and trusted flaggers or partners. Notifications should not be mistaken with the number of reported cases. This is why the total number of notifications (192) is higher than the number of identified illegal online incidents (174). This discrepancy comes from the fact that in some cases concerning Polish intermediary services, the same content was reported twice by different individual users to two Polish online platforms.

The MRE is an activity built upon detecting, identifying, and reporting illegal online content to chosen VLOPs and Polish IT companies. The reporting was carried out through a two-stage mechanism — individual users and partner organizations possessing a trusted partner status at the foreign very large online platforms.

Methodology of the MRE

- The figures presented in the Exercise are not statistically representative of the absolute prevalence and types of illegal hate speech. They are derived from the total number of notifications sent from specific accounts used in the Exercise, reflecting a methodological approach to the analysis.
- The Exercise was carried out for a period of approximately 6 weeks, from November 1, 2023 to April 7, 2024, applying measures prescribed by the Digital Services Act and parts of the methodology used in yearly evaluations of the EU Code of conduct on countering illegal hate speech online.³¹
- The MRE was conducted in line with the organization's internal workflow and relevant procedures to detect, identify, and report illegal content at 5 designated VLOPs (Facebook, Instagram, YouTube, TikTok, and X) and 6 Polish intermediary services (Agora³², wp.pl,³³ onet.pl³⁴, natemat.pl³⁵, dorzeczy.pl³⁶, wykop.pl³⁷).
- All reported content met the IHRA working definition of antisemitism and CoE's definition of hate speech.
- All detected content was evaluated by a legal counsel and was deemed to be illegal
 hate speech under the Polish Penal Code, transposing EU Council Framework
 Decision 2008/913/JHA on combating certain forms and expressions of racism and
 xenophobia by means of criminal law.
- All detected content, based on our analysis, violated the rules and guidelines set up by IT companies.
- Notifications were submitted through reporting channels available to all users, and in some cases, also via dedicated channels accessible to trusted flaggers and trusted partners, which allowed us to compare responses from IT companies within the MRE. The trusted flaggers/partners mechanism, which uses a more elaborative escalation or flagging mechanism, currently operates in relation to very large online platforms. Polish IT companies, do not currently provide reporting mechanisms by trusted flaggers.

³¹ European Commission, Code of conduct on countering illegal hate speech online.

³² Agora is a Polish media company known for operating online news outlets: gazeta.pl and wyborcza.pl.

³³ wp.pl is a Polish online platform owned by Wirtualna Polska Holding S.A.

³⁴ onet.pl is an online news portal owned by Ringier Axel Springer Polska joint venture company.

^{35 &}lt;u>natemat.pl</u> is an online news portal owned by Na Temat Group.

^{36 &}lt;u>dorzeczy.pl</u> is a right-wing online news platform owned by Orle Pióro Ltd.

³⁷ wykop.pl is a Polish social news aggregation and discussion website, similar in concept to Reddit.

In other words, Polish intermediary services use only automated reporting tools accessible to general users and the same online content submitted to them via e-mail could not be identified, for legal and evaluation purposes, as a form of escalation submitted by the same user.

The Jewish Association Czulent sent three requests to five VLOPs to join the programs for trusted partners. Only one VLOP provided an answer directing us to the DSA provisions. No VLOP provided a direct response to our request. The escalations were, therefore, provided to VLOPs via third entities.

KEY FACTS ABOUT THE WORKFLOW OF THE EXERCISE

- a. detection, identification, and aggregation of illegal online content;
- b. reporting illegal content to the online platforms through an automated reporting tool by general users or using the "notice and take down" procedure;
- **c.** verification of the time of assessment and the online availability of notices submitted by general users;
- d. escalation submission of illegal content to VLOPs by trusted flaggers or trusted partners;
- e. verification of the time of assessment and the online availability of notices submitted by trusted flaggers or trusted partners;
- f. verification of the availability of reported illegal content at the end of the MRE (closure date).

Key figures

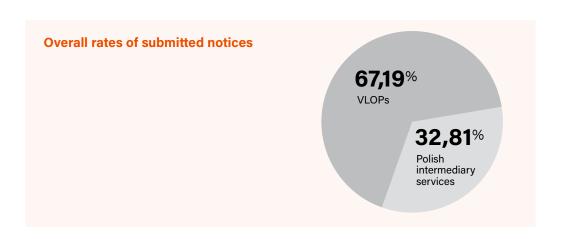
 A total of 192 notices were submitted to VLOPs and Polish intermediary services: including 129 notifications to very large online platforms and 63 notifications to Polish intermediary services. Number of submitted notifications

- A total of 129 notices were submitted to VLOPs through the automated reporting tool available to general users. In addition, 60 out of 129 interventions were escalated, meaning they were additionally submitted through the designated channels available only to entities having trusted flagger or trusted partner status for VLOPs.
- A total of 63 notices were submitted to Polish intermediary services, including
 45 notices submitted through the automated reporting tool and 18 via e-mail.
- The DSA states that notices on illegal content submitted by individuals or entities, including via automated electronic means, should be treated by providers of hosting services in a "timely, diligent, non-arbitrary and objective manner". It also emphasizes that notifications sent by trusted flaggers to very large online platforms should be taken with priority and without "undue delay". The terms "undue delay" or "timely manner" do not indicate strict time limits for the assessment of notices, thus for the purpose of the MRE we used a similar methodology to the one used in the EU Code of conduct on countering illegal hate speech online. Time limits for the responses to notices were set up for 24 hours, 48 hours, less than one week, and longer than one week.

Time of assessment of notifications

On average, it took Instagram **28,5 days**, Youtube **28,5 days**, Facebook **8,26 days**, and TikTok **7 days** to make a decision on a submitted escalated notification.

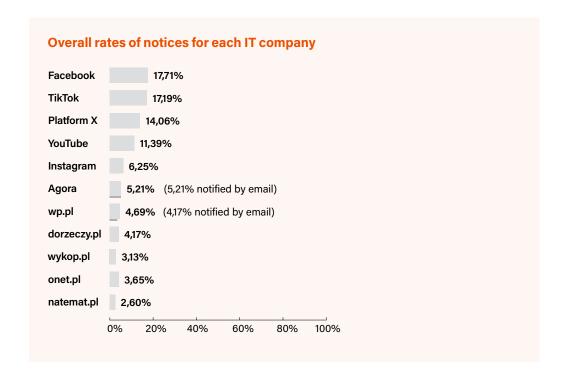
39 Ibidem, art. 22 (1).



Rates and numbers of notices to IT companies

Among VLOPs, Facebook received the largest amount of notifications (34), followed by TikTok (33), X (27), YouTube (23) and Instagram (12).

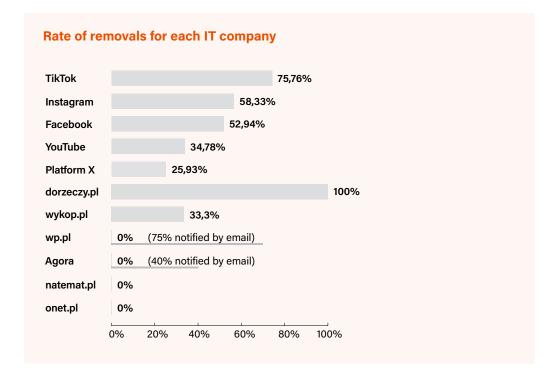
Among Polish online services, **Agora** (wyborcza.pl and gazeta.pl) received the most interventions (20), followed by **wp.pl** (17), **dorzeczy.pl** (8), **onet.pl** (7), **wykop.pl** (6) and **natemat.pl** (5).

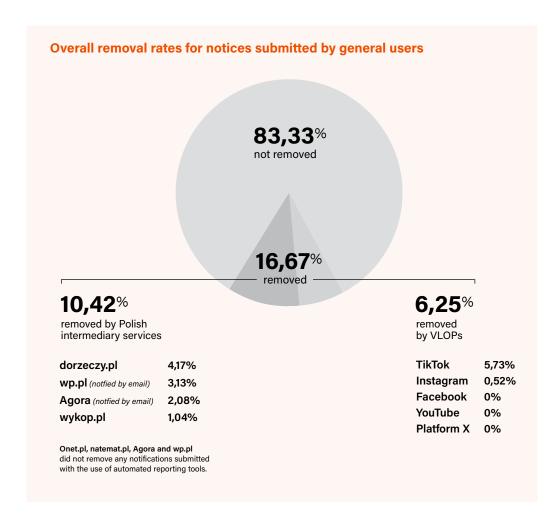


Removal rates



Availability of the reported illegal content at the end of the MRE





Results for notices submitted through an automated reporting tool by general users or "notice and take down" procedure

For VLOPs, an overall number of removed notices was 12 out of 129 interventions submitted by general users, and constituted 9,30% of the notifications made to the very large online platforms.

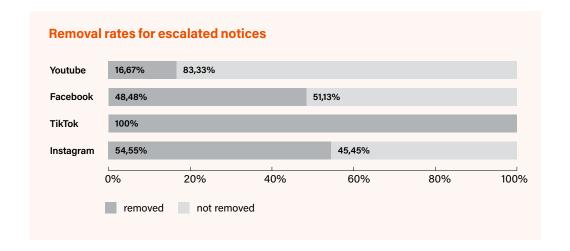
For Polish services, an overall number of removed notices was **20** out of **63** interventions submitted by general users, and constituted **31,75%** of the notifications made to Polish intermediary services.

ESCALATIONS

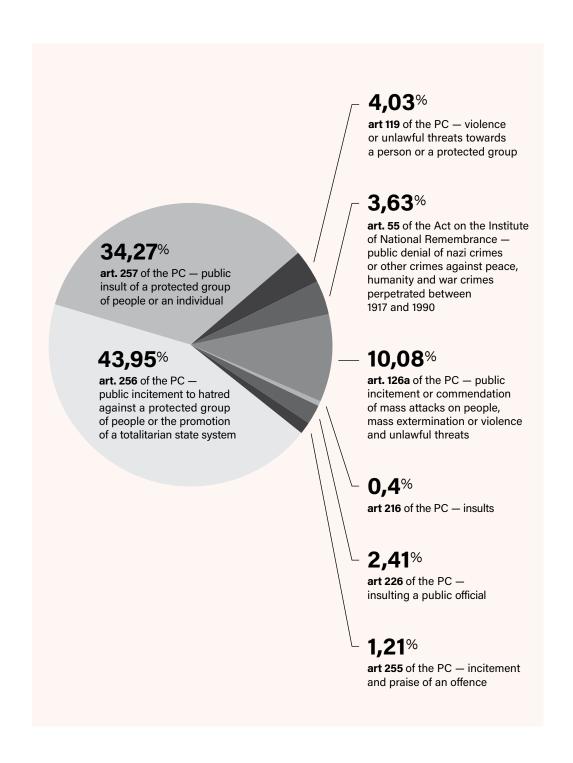
Escalations were submitted to 4 very large online platforms: Facebook, Instagram, YouTube, and TikTok. No escalated notifications were submitted to platform X due to the inability of trusted flaggers or partners to submit notices at the time of conducting the MRE.

The overall **number of removed escalated notices** was **28** out of **60** interventions submitted to VLOPs by trusted flaggers or partners. In other words, VLOPs removed **46,67%** of notices submitted by trusted flaggers or trusted partners.

Results for submissions of notices by trusted flaggers or trusted partners



Legal grounds of the reported illegal hateful content



Conclusions and recommendations

The Exercise showed that the evaluated IT companies removed only **50,39%** of reported illegal antisemitic hate speech. Many cases of notified illegal content remained online despite content moderation obligations imposed on IT companies by the Digital Services Act and although they violated the internal regulations of IT companies.

The Exercise showed a higher removal rate of illegal content when submitted by trusted flaggers or partners to very large online platforms in comparison to those submitted by general users. Similar findings can be attributed to more descriptive notifications submitted by email rather than those submitted through an automated tool to Polish intermediary services. The final results remain far from satisfactory, including the assessment time of submitted notifications, yet provide a foundation for further efforts in combating illegal content online.

Public security and discourse in the digital space are one of the fundamental and most pressing challenges of our time. Social media platforms insufficiently ensure the safety of groups particularly vulnerable to hate crimes, as well as online service users overall. Within the analyzed content, we observed disinformation aimed at polarizing audiences and creating or reinforcing fears and anxieties instigated by antisemitic conspiracy myths. The MRE showed that the reported illegal antisemitic content exhibited characteristics of intersectional discrimination, including anti-Ukrainian sentiments.

The DSA creates a new basis for efforts to combat illegality on the internet. It necessitates field expertise and strong monitoring and reporting skills by reporting entities. Its effectiveness, however, is dependent first of all on the recognition and implementation of content moderation obligations by IT companies and all stakeholders engaged in creating a safer environment online, including swift implementation of the DSA into Polish law.

RECOMMENDATIONS

We see a particular need to promptly remove illegal content that violates applicable national and EU laws. In our opinion, it is essential for IT companies to:

- refine algorithms designed to remove illegal content,
- provide regular training for Polish-speaking IT administrators to recognize antisemitism, including knowledge of the law and constantly evolving antisemitic language code,
- continue granting trusted partner status to expert entities, especially in the absence of a national Digital Services Coordinator.

In our view, the Polish legislator should:

- prioritize the establishment of a national Digital Services Coordinator, equipped with adequate human and financial resources,
- clarify the role and reporting guidelines for trusted flaggers under Polish law,
- provide trusted flaggers with genuine financial support to carry out their activities,
- establish an advisory council to the Digital Services Coordinator to facilitate cooperation, primarily between the government and organizations holding trusted flagger status.

SEE → ACT → REPORT ANTISEMITISM

Why is it important to report antisemitic incidents?

We take every experience that you perceive as antisemitic seriously—your perspective as a victim or witness is paramount in our work. We use a victim-oriented approach and focus on their needs.

Our role is unique: we record reports of incidents that do not always constitute a crime. All data you provide is confidential. You decide on their further use.

We want to inform you about support opportunities: anti-discrimination, psychosocial, and legal counselling.

How can you report antisemitic incidents?

via the website www.zglosantysemityzm.pl



by e-mail: info@zglos-antysemityzm.pl

or by phone at 12 400 00 08

(see www.zglosantysemityzm.pl for operating hours)

About the Jewish Association Czulent

The Jewish Association Czulent is an independent non-profit organization operating both nationally and internationally, primarily engaging in advocacy activities. We gather professionals from the Jewish community both in Poland and abroad.

"Czulent" undertakes comprehensive initiatives to combat antisemitism, which include developing analyses and reports on the phenomenon of antisemitism in the Visegrad Group countries. It is also involved in strategic litigation activities. It operates the platform <u>zglosantysemityzm.pl</u>, which enables the reporting of antisemitic incidents and crimes, as well as providing legal support for victims.

Our activities encompass political, social, and legal aspects, implemented through the introduction of innovative educational solutions and the building of coalitions for openness, and against antisemitism, racism, and discrimination. We collaborate with institutions, public administration, and dialogue organizations to influence social attitudes and shape Polish legislation on tolerance and combating racism. Our partners include the OSCE Office for Democratic Institutions and Human Rights (ODIHR), the American Jewish Committee Central Europe, and the National Democratic Institute.

As part of international coalitions — the European Network on Monitoring Antisemitism (ENMA), the Coalition to Counter Online Antisemitism (CCOA), the European Network for Countering Antisemitism through Education (ENCATE) and the European Network Against Racism (ENAR) — we collect and promote best practices and recommend solutions at the European level.



