State of Israel
Ministry of Strategic Affairs

Ministry of Diaspora Affairs
Creating a common Jewish future

THE HATE FACTOR

GOVERNMENT OF ISRAEL POLICY OUTLINE FOR COMBATING ANTISEMITISM ONLINE

FEBRUARY 2021

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The online revolution has seen social media emerge as a leading medium for communication and information-sharing, giving rise to an unprecedented potential for people to express themselves and exchange ideas openly. But this great advance presents challenges, among them coping with the rapid spread of hate speech. In the past, radical views were mostly beyond the scope of mainstream discourse. Today, in the borderless, networked world, hate speech, conspiracy theories and fake news are disseminated to vast audiences, sometimes anonymously, with a single click.

This spread of hate speech is harming many minorities worldwide, targeted based on religion, gender, nationality, and race. Jews are no strangers to being targeted by hate speech and age-old antisemitism is alive and well today, both in its classic and new manifestations. The alarming global rise in antisemitic incidents around the world in recent years is consistent with the increase in online hate speech against the Jewish people and the State of Israel.

For their part, social media companies, which are privately-owned public platforms, enjoy tremendous autonomy in defining hate speech policy and enforcing it, with little international or local standards or regulation. As both rule maker and rule enforcer, social media companies may choose to act against or ignore hate speech, sometimes with dramatic effect, as witnessed in the events in the United State in January.

Tackling online hate speech is a complex and multidimensional challenge. To this end, the Ministry of Strategic Affairs and the Ministry of Diaspora Affairs, which are charged with combatting antisemitism and the delegitimization of Israel, present this proposed policy outline in an effort to organize the government's approach to combatting antisemitic hate speech online.

PURPOSE AND METHODOLOGY

The purpose of this document is to outline principles for the government's efforts to combat online antisemitic hate speech against the Jewish People and the State of Israel. It also outlines focus areas for an inter-ministerial working group on hate speech, under the auspices of the directors-general of the Ministry of Diaspora Affairs and the Ministry of Strategic Affairs, in collaboration with the relevant government ministries.

This policy paper was formulated following consultation with government ministry representatives – from the Ministry of Diaspora Affairs, the Ministry of Strategic Affairs, the Ministry of Foreign Affairs, and the Ministry of Justice – along with experts, NGOs, and following a review of relevant studies and reports. Concomitantly, there have been meetings between representatives of government ministries and representatives of leading social media companies.

For purposes of this document, hate speech will be defined according to the International Holocaust Remembrance Alliance (IHRA) Working Definition of Antisemitism, which serves as the accepted international standard for identifying antisemitism in its various manifestations, as a hate speech against Jews and as delegitimization and demonization of the State of Israel. Nothing in this document or in the government effort to combat hate speech should be construed as intended to limit or discourage legitimate criticism of Israel, in the same way as that legitimate criticism may be directed at any country.

"WHAT DO WE WANT?"
GOALS FOR THE GOVERNMENT OF ISRAEL VIS-À-VIS SOCIAL MEDIA COMPANIES

The first part of this document presents the objectives proposed to the government and to the inter-ministerial working group in its dialogue with the leading social media companies:

POLICY

- **Definition of antisemitism**: Integrate the International Holocaust Remembrance Alliance (IHRA) working definition as a tool for identifying and educating on antisemitism and use the working definition in social media community rules.
- **Reliable information:** Diversify the toolbox to combat antisemitic hate speech with , and reference reliable information relating to Holocaust denial and inversion.

ENFORCEMENT

- **Language focus:** Increase enforcement of community standards in languages where antisemitic hate speech is more prevalent.
- **Antisemitic hate speech propagators:** Remove the accounts of those convicted of antisemitic hate speech, while reducing the virality of suspected hate speech.
- **Training of content moderators:** Promote initiatives for the ongoing training of content moderators on antisemitic hate speech by independent NGOs, while increasing the transparency regarding content moderator training.
- **Coordinated Inauthentic Behavior:** Increase efforts to identify and remove accounts for inauthentic coordinated behavior which encourages hate speech.
- **Hate Commerce:** Create global policy regarding, and increase enforcement of current rules on, ecommerce platforms prohibiting trafficking in Nazi memorabilia and items which promote hate or Holocaust denial. *(This goal is also intended for leading e-commerce companies and credit card/online payment companies.)*

TRANSPARENCY

- **Data on hate speech:** Increase transparency and allow public access to data on hate speech, including prevalence, segmentation by target groups and region, etc.

"HOW DO WE DO THIS?"
GOVERNMENT COURSE OF ACTION

In the second part, the policy paper suggests how the government should consider acting, in coordination between the relevant government ministries, to achieve the goals laid out in Part I. Coordinating the government effort is an inter-ministerial working group on hate speech which will deal with, among others, the following issues:

- Greater monitoring of social networks for antisemitic hate speech by the Government of Israel.
- Regulation - ongoing examination of international regulation and enforcement in areas relevant to hate speech and the developing of up-to-date regulatory options.
- G2G - Cooperation between governments and multinational organizations.
- International cooperation and regulation against extremist social media platforms.
- Building broad international coalitions to combat hate speech; consulting and convening with civil society organizations.

"HOW DOES THIS DOCUMENT HELP?"

- For government ministries: to work in a coordinated manner to achieve clear goals in dialogue with social media companies.
- For civil society organizations (and research institutes): to understand the government's approach and goals, and to consult and work in coordination with it where appropriate.
- For officials of the State of Israel: to assist them in presenting the issue of hate speech in a more precise and detailed manner, including vis-à-vis international colleagues, organizations, and institutions, with an eye to increase cooperation with Israel on the issue.

INTRODUCTION

Over the last two decades, the world has been in the midst of a technological revolution. The internet, along with the vast data storage capacity and the ability to produce and disseminate content instantly and independently, has led to an unprecedented change in global discourse. Social networks connect people and create a digital free, open, limitless space for conversation. Like other human transformations, this advance also has its drawbacks and challenges, with those abusing the digital world and social networks as a medium for promoting harmful content and spreading hate.

As they grow at a dizzying pace, social media companies have been slow to act against hate speech and, until recently, their focus on confronting such abuse has been limited.

In the United States, public and civil society organizations were so frustrated with social media's inadequate action against hate speech that in July 2020, they leveled one of the most successful campaigns against a social media company. The Anti-Defamation League (ADL), in partnership with leading African American NGOs, launched a campaign in which in more than 1,200 companies, including leading U.S. corporations, halted their Facebook advertisements for a month, in order to pressure the company to take more action against hate speech on the platform.[1] One of the strengths of this campaign was highlighting the fact that hate speech is a common problem for many different minority groups.

In extreme cases, social media companies' inaction against hate speech can have deadly real-world consequences. In 2018, Facebook admitted that in the Myanmar conflict, in which the Rohingya minority was systematically murdered, it did not act sufficiently to prevent the use made of a social network to "incite division and incitement to violence."[2]

Following the events of January 2021 in the US, Twitter CEO and co-founder Jack Dorsey acknowledged the link between hate speech and physical violence, and the failure of

[1] Stop Hate for Profit, https://www.stophateforprofit.org/
[2] The New York Times, https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html

social media companies on the issue:

> Offline harm as a result of online speech is demonstrably real...we all need to look critically at inconsistencies of our policy and enforcement. Yes, we need to look at how our service might incentivize distraction and harm. Yes, we need more transparency in our moderation operations.[3]

For Jews, who have been the target of persecution for millennia, hate speech is nothing new. It has morphed over the centuries from religious persecution of classical antisemitism, to racial persecution in the Holocaust, to national-ethnic persecution in the Soviet Union and finally, for some, has morphed into a virulent hatred the nation-state of the Jewish people. This hatred is expressed against Jews as individuals, a community, and through de-legitimizing and demonizing Israel and in presenting the Jewish State as a moral blemish whose very existence is manifestly and uniquely illegitimate. In the past, this form of antisemitism was concentrated mainly in the Arab and Soviet world. Since 2001, the delegitimization campaign has expanded to the West.

Today, the leading medium for spreading antisemitic hate speech is online, especially on social media hate speech, online or otherwise, can devolve into a real-world threat of, or actual violence. The Ministry of Diaspora Affairs found a correlation between areas where intense antisemitic hate speech took place and the number of antisemitic incidents. Moreover, perpetrators of hate crimes use social media to glorify their actions and inspire their followers, using real-world violence as a tool for gaining an online audience. For example, a wave of incitement to violence against Israel that took place on social media, sparked a wave of stabbing attacks in 2015. Thus, the vicious circle forms whereby incitement on social media begets real-world violence, begets further online incitement, and further violence. It was during this period that Israel began to demand the removal of content inciting to terrorism and violence from social media.

In recent years (until 2020 when people were socially distanced due to the coronavirus) there has been a sharp increase in the number of antisemitic incidents around the world.

[3] https://twitter.com/jack/status/1349510775426019328

From 2017 to 2019, there was a 33% increase worldwide in the number of "major violent" antisemitic incidents.[4] In the U.S., 2019 marked an all-time record in the number of antisemitic incidents (up 12 percent from 2018), with a series of lethal antisemitic attacks which followed the 2018 deadliest antisemitic attack in U.S. history at the Tree of Life Synagogue in Pennsylvania.[5]

Jews living around the world sense the threat – 40% of respondents to a survey among Jews in Europe are concerned about their personal safety;[6] in the U.S. 31% of respondents among Jews said they refrained from displaying symbols that would identify them as Jews.[7]

The feeling of insecurity is also felt by Jews on social networks. Young Jews reported experiencing antisemitism on TikTok.[8] The outbreak of the coronavirus led to an increase in the volume of antisemitic hate speech, with content accusing the Jews or Israel of creating and/or spreading the virus.[9] The concern about antisemitism was such that the UN Special Rapporteur on freedom of religion and belief put out a special statement noting that "antisemitic hate speech has risen alarmingly since the outbreak of the COVID-19 crisis."

He also directed called on social media companies to act:

> Countering online hate speech also will not succeed if the mainstream or social media do not take seriously the reports of cyberhate targeting Jews and other minorities… they must remove any posts that incite to hatred or violence in addition to identifying and reporting fake news.[10]

[4] In the number of "major violent incidents" as reported in the Annual Report on Worldwide Antisemitism of the Kantor Center for the Study of Contemporary European Jewry at Tel Aviv University (their 2017 report cites 342 such incidents and the 2019 report cites 456 such incidents)
https://en-humanities.tau.ac.il/kantor/rerearch/annual_reports
[5] https://www.adl.org/news/press-releases/antisemitic-incidents-hit-all-time-high-in-2019
[6] https://fra.europa.eu/en/publication/2018/experiences-and-perceptions-antisemitism-second-survey-discrimination-and-hate
[7] https://www.ajc.org/AntisemitismSurvey2019
[8] https://www.nbcnews.com/news/us-news/jewish-teens-say-life-tiktok-comes-antisemitism-n1241033
[9] https://www.timesofisrael.com/covid-19-fueling-worldwide-wave-of-antisemitism-researchers-find /
[10] https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25800&LangID=E

The Israeli government, aware of the dangers of the hate speech, delegated its response to several government ministries, each with its own areas of responsibility. Today it is clear that to have an impact on this issue requires a coordinated, whole-of-government approach, including a clear set of issues for social media companies to address.

By mid-2020, the relevant government ministries began a process to formulate an organized government strategy for dealing with antisemitic hate speech on social media, which included discussions between representatives of all relevant government agencies and representatives of Facebook, Twitter, Google, and TikTok.

This document is intended to serve as an outline for government policy to combat antisemitic hate speech on social media.

PURPOSE AND METHODOLOGY

GOALS

The purpose of this policy paper is to present an outline for a coordinated course of action for government ministries to combat antisemitic hate speech. It will help government ministries focus and coordinate their work to achieve clearly defined goals vis-à-vis the social media companies.

This document will help NGOs and think tanks already involved in issues related to combating online antisemitism by providing an articulation of the Israeli government's approach to this issue. The document may also form the basis of international discussion cooperation with countries, institutions and organizations seeking to tackle hate speech online.

PROCESS

This document was written following consultation between the various ministries that are involved with combatting antisemitism – the Ministry of Strategic Affairs, the Ministry of Diaspora Affairs, the Ministry of Foreign Affairs, and the Ministry of Justice. The process took place in consultation with experts in the field, representatives of Jewish organizations at the forefront of combating antisemitic hate speech, and the reviewing of relevant studies and discussions held between government officials and social media companies.

STRUCTURE

The first part of the document focuses the government's agenda in its discussions with social media companies. It presents the main objectives set out for social media companies on how to address antisemitism on their platforms. These objectives are divided into three categories: policy, enforcement, and transparency.

The second part of the document offers an outline for organizing government work. This section presents areas of focus to help create a cohesive and comprehensive government

approach. It suggests areas of focus for an inter-ministerial working group dedicated to the issue.

ONLINE PLATFORMS

A. Leading Social Media Platforms

The outline proposes to engage in dialog with the leading social media companies, given their widespread popularity and availability – and, as such, their ability to influence public opinion. Although the most malicious hate speech occurs in closed groups and niche platforms, it trickles into the major social media networks for mass consumption.

Platform (company)	Monthly Users
Facebook	2.6 billion
YouTube (Google)[11]	2 billion
Instagram (Facebook)	1 billion
TikTok	800 million
VK	600 million
Telegram	400 million
Twitter	330 million

This document focuses on five leading platforms – Facebook, Instagram, YouTube, TikTok and Twitter.[12] The companies which run these platforms are experienced in the areas of policy and enforcement, and have been in dialogue with the government regarding hate speech on their platforms.

There are two other companies that have become a popular hub for antisemitic hate speech – VK (VKontakte) and Telegram (specifically, Telegram channels). There is no government dialogue with these companies so far and they do not have detailed policies or rigorous enforcement. It also appears that currently they are less responsive to public or government opinion. As such, these two companies require an in-depth and dedicated approach, based

[11] YouTube may have the characteristics of a content company, but it receives a social network reference here.
[12] Twitter's influence is particularly great among government leaders, reporters and public officials and as such its impact goes beyond the number of monthly users.

on the principles in this document, but tailored to each company individually (See Part II – Inter-ministerial Working Group).

B. E-Commerce and Payment Platforms

This report also refers to e-commerce and payment platforms that allow the sale of merchandise of an antisemitic nature. In this way, these companies unconsciously serve as a platform for promoting and funding antisemitic content.

C. Extremist Platforms

In addition to the large social media networks, there are niche platforms today that have deliberately lenient policy regarding the publication of content, which attracts extremist content – from pornography to hate speech. These companies knowingly allow this to take place on their platforms. They do not engage in dialog with government officials, nor do they use Israeli infrastructure services. This report does not deal extensively with these networks but suggests that the inter-ministerial working group examine the problem and suggest courses of action, including through international cooperation and regulation (see Part 2 - Government - Regulation and International Cooperation against Extremist Websites).

TERMINOLOGY

In describing "hate speech" in the Israeli and Jewish context, a number of terms are used, among them – antisemitism, incitement, and delegitimization of Israel – which overlap in part, and require precision.

Thus, for the purposes of this policy paper, "antisemitic hate speech" will be defined consistent with the International Holocaust Remembrance Alliance (IHRA) Working Definition of Antisemitism (see Part I, below).

PART I

GOVERNMENT GOALS VIS-À-VIS SOCIAL MEDIA COMPANIES

The government's goals vis-à-vis social media companies are divided into three issues:

A. **Policy**

B. **Enforcement**

C. **Transparency**

POLICY

The prohibition against hate speech is rooted in international law in several international treaties.[13] Over the years, social media companies have formulated user policies which refer to hate speech as a type of discourse prohibited on their platforms. These user policies on hate speech aim to provide safety for various "protected groups" based on race, nationality, religion, gender, etc., similar to international treaties on the matter. Each company has formulated its own rules on hate speech that have evolved over the years – some more detailed, others vaguer and more open to interpretation.

This section will deal with these two important issues: first, defining hate speech in the Jewish/Israeli context; and second, broadening the toolbox of options available in dealing with antisemitic hate speech, including making reliable information more accessible.

[13] See *Reducing Online Hate Speech*, https://www.idi.org.il/books/31764, pages 33-45

1 DEFINITION OF ANTISEMITISM

GOAL FOR SOCIAL MEDIA COMPANIES

Integrate the International Holocaust Remembrance Alliance (IHRA) working definition as a tool for identifying and educating on antisemitism and use the working definition in the community rules.

BACKGROUND

The internationally accepted definition of antisemitism is the International Holocaust Remembrance Alliance Working Definition of Antisemitism (hereafter "the IHRA definition"). Established in 1998, the IHRA is a multinational organization which works to promote Holocaust education, research and remembrance, and combat Holocaust denial and antisemitism, and currently consists of 34 member countries. In 2016 the IHRA unanimously adopted a legally non-binding working definition of antisemitism, noting that "in order to begin to address the problem of antisemitism, there must be clarity about what antisemitism is."[14] As an integral part of the definition, 11 examples of antisemitism were offered[15] (See Appendix B – IHRA examples by type). Since 2016, nearly thirty countries have adopted the IHRA definition, as have some 400 local councils, organizations and universities.[16]

The broad international adoption of the IHRA definition expresses the growing consensus

[14] https://www.holocaustremembrance.com/stories/working-definition-antisemitism
[15] https://www.holocaustremembrance.com/resources/working-definitions-charters/working-definition-antisemitism
[16] Since 2016, the definition has been officially adopted by nearly 30 countries, including many European countries and the State of Israel. The EU adopted the definition and in doing so called on "EU member states and EU institutions and agencies to adopt and apply its definition". In the US, the definition appears as a definition of antisemitism on behalf of the State Department with a call for more countries to use it, and it appears in the presidential decree to combat antisemitism signed in December 2019. UN Secretary-General Antonio Guterres stressed the importance of countries' efforts to adopt as one the IHRA definition to establish an “acceptable definition of antisemitism.”

regarding antisemitism, and the recognition that antisemitism exists in both classical and new forms. Thus, the working definition gives the best framework for discussion of antisemitic hate speech policy with social media companies.

Unlike the IHRA definition which includes real life examples of speech that may be considered antisemitic, social media companies' policy regarding hate speech does not specifically address antisemitism, just as they do not specifically address hate speech against other religions or national origin (except for an occasional example with makes reference to Jews or Holocaust denial).[17] Jews are protected under hate speech rules by virtue of being a religious group, and Israelis by virtue of their national origin (see Appendix A – Comparative Policy Table).

Current social media company hate speech rules do not properly address the various manifestations of antisemitism. In the absence of a clear policy, hate against Jews and Israel can find its way to hundreds of millions of users around the world. Therefore, one of the challenges in combatting antisemitic hate speech online is **to reduce the gap between social media companies' policy regarding hate speech and antisemitic content, as defined by the IHRA working definition.**

Social media companies publicly acknowledge the importance of the IHRA definition and use it to formulate their policies. Yet, in practice, they only partially implement the definition in their policy and do not refer to it in their community rules.

RECOMMENDATIONS

The IHRA working definition of antisemitism has been adopted globally, and as such, is an essential tool for defining a clear policy for identifying hate speech relating to both the

[17] It should be noted that in the general context of hate speech "the rules and policies designed to reduce hate speech must be subject to international human rights standards, As articulated in the Covenant on Civil and Political Rights (especially Articles 19 and 20) and other international treaties such as the Convention on the Elimination of All Forms of Racial Discrimination, (Convention on the) Elimination of Racial Discrimination (e.g. Articles 4 and 5(d)(vii) or the European Convention on Human Rights (see Recommendations of the Israel Democracy Institute – Appendix D Recommendation 2).

hatred of Jews ("classic antisemitism"), and delegitimization and demonization specifically and uniquely against the State of Israel ("new antisemitism"). The definition explicitly states that legitimate **criticism against the State of Israel "cannot be considered antisemitic**."

Current social media policy does not reflect the IHRA definition relative to new antisemitism. In the view of social media companies, the State of Israel is not immune from hate speech, just as no other country is immune (countries are not a "protected group"). Regarding the comparison of Israel to the Nazi regime, they argue that other governments are also compared to Nazis. Additionally, they argue that Zionism is a political movement and as such not protected under hate speech policy.

This approach indicates a lack of understanding of the phenomenon of antisemitism, which changes over time. This approach underscores the ease with which Israel can serve as a proxy for antisemitism, and the effectiveness of the argument that any speech against Israel falls within the confines of legitimate political criticism.

Through the following examples, which appeared ***and were subsequently removed*** by social media companies, one can see the morphing between blatant "classical" antisemitism, which is universally condemned, and antisemitism directed at the State of Israel:

The most infamous recent example of this overlap between classic and new antisemitism can be seen in this cartoon published in the Europe edition of *New York Times* in April 2019:

The New York Times International Edition, April 25 2019, by Antonio Moreira Antunus

This cartoon was viewed by the duty editor as an expression of political criticism. The newspaper's editorial board stated that it was in fact antisemitic, noting

> the appearance of such an obviously bigoted cartoon in a mainstream publication is evidence of a profound danger — not only of antisemitism but of numbness to its creep… and some criticism of Israel, as the cartoon demonstrated, is couched openly in anti-Semitic terms.[18]

This is, unfortunately, an accurate description of the current reality that social media companies must face when it comes to new antisemitism.

Understanding and internalizing the nuances of antisemitism, upon all its forms, is a necessary step. Research and dialogue by NGOs and Jewish communities over the years have resulted in some positive changes in social media policy. Recently, Facebook and Twitter announced a policy change and will now remove Holocaust denial, which Facebook attributed to ongoing discussions of the issue with Jewish organizations.[19]

[18] New York Times, 30 April 2019, https://nyti.ms/2MFfCLv
[19] Facebook, 12 October 2020, https://bit.ly/2N209oV and CNBC, https://cnb.cx/3cT8ciw

Proper adoption of the IHRA definition by social media companies will require understanding context, which is one of the great challenges in dealing with hate speech in general.[20]

Taking into account the importance of striking the delicate balance between the preserving free speech and removing hate speech, social media companies should examine options such as labeling, thus expanding their toolkit beyond the binary decision whether to remove content or allow it to remain.

To start, companies should consider labeling content that is antisemitic per the IHRA definition, but which does not violate their hate speech rules.

[20] Examining the context is one of the great challenges of liability in dealing with hate speech in general. In this context, the Israel Democracy Institute recommends that when evaluating policy, the contextual considerations of the content can be taken into account:

> The extent to which the classification of hate speech as such (a) is based on a closed list of banned words, phrases, symbols or images; (b) makes it possible to identify complex connections among language, images, and ideas that may render speech hateful in certain cultural, social, or political settings, and (c) considers the broader context that may legitimize (e.g., satire) or delegitimize the speech (e.g., bogus historical research at the service of racist causes). Is the element of causation incorporated in the definition of hate speech linked only to the expectation that it might lead to physical harm to the targeted persons? Or does it also consider nonphysical damage to potential victims…

See Appendix D - Recommendations of the Israel Democracy Institute - Recommendation 16: Criteria for Evaluating Policies and Rules

2 LABELING AND ACCESS TO RELIABLE INFORMATION

GOAL FOR SOCIAL MEDIA COMPANIES

Diversifying the toolbox to combat antisemitic hate speech, and referencing reliable information relating to Holocaust denial and inversion.

BACKGROUND

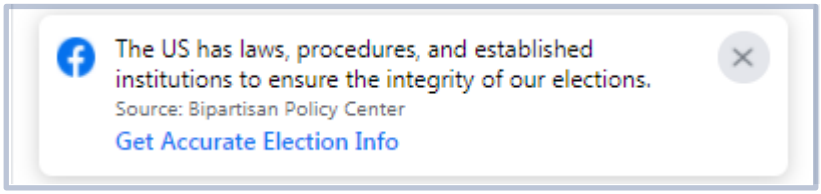
Dealing with hate speech is inherently complex, certainly at a time when there is a lack of global clarity as to the quality and reliability of information online. Furthermore, as mentioned earlier, removing user content challenges the fundamental principle of free speech.

In recent years, social media companies have developed a number of creative solutions that balance the need for proper discourse and safeguarding freedom of expression. They began to make greater use of marking potentially problematic content by labeling, flagging, or adding informational context.

Typically, the companies add labels or information, at their discretion, that warn the user about potentially offensive or false material. Here is label added by Twitter:

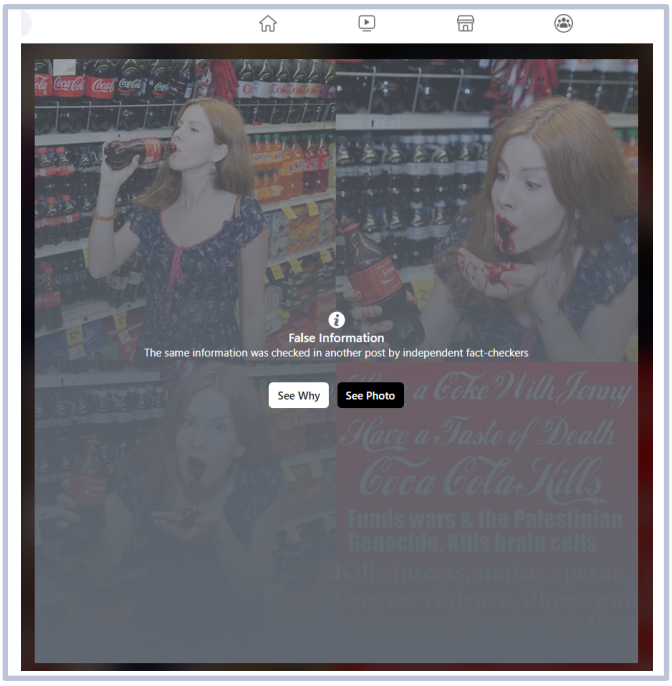
Sometimes, in addition to labels, reliable information is referenced on the topic. Often, this information comes from the International Fact Checking Network.[21]

Here is an example of alerting users to false information when it appears on Facebook:



This comment, attached to a post that has not been removed, refers to reliable U.S. election information.

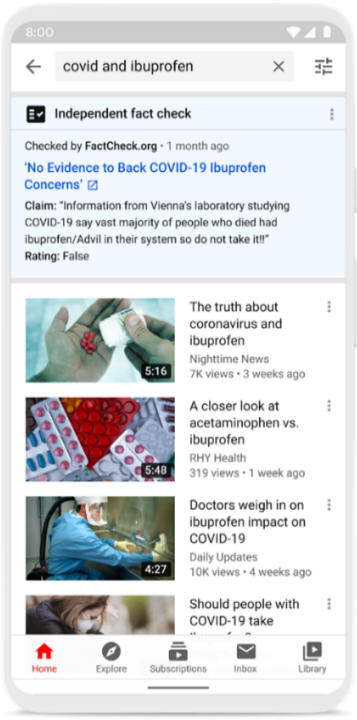
Here's another example, also from Facebook, that hides problematic content:



The image is covered with a warning that reads: "False information. The information was checked in another post by independent fact-checkers," accompanied by buttons for more information: "See Why" and "See Photo".

[21]This is a nonprofit with members of organizations (most of them journalists) around the world who are reviewing for companies such as Facebook, Google, and YouTube content which is classified as misinformation. See https://www.poynter.org/ifcn/

The following is an example of how reliable information regarding the Coronavirus is presented on YouTube:

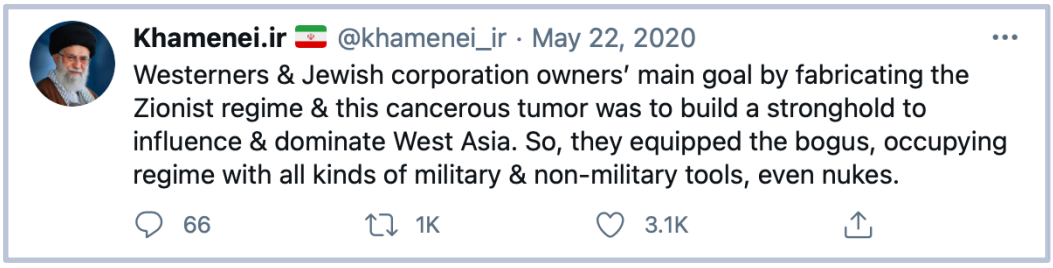


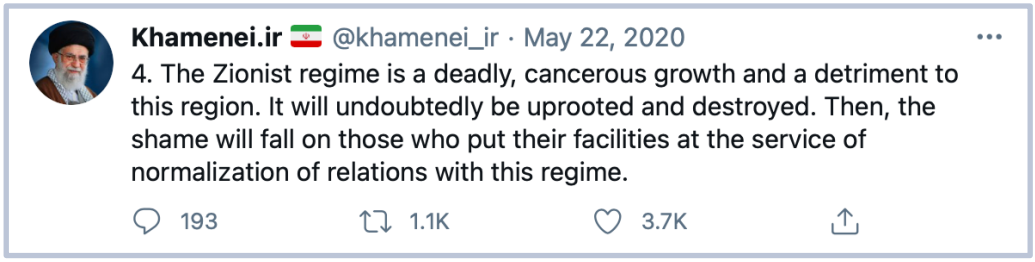
Pictured: A YouTube search of the words "covid and ibuprofen" –reliable information appears before the search results (marked in a blue background) explaining that there is no proof of the claim that the drug raises the chance of mortality from the virus.

In particular, while Twitter has increased labeling significantly recently, its policy seems to be inconsistent when it comes to world leaders. According to the company, world leaders are immune from the rules except where there is a definite call for violence against specific individuals. If the leader of a state calls for the destruction of another nation, it is perceived as "political saber rattling". When a political leader posts content that is tantamount to hate speech according to the company's policy, these companies still allow the posts to be presented to the public, even if they are extremist in nature.[22] Thus, according to Twitter, Iran's spiritual leader, Ali Khamenei, is able to post quintessentially antisemitic content and calls for violence against the State of Israel without repercussion. Despite repeated

[22] https://blog.twitter.com/en_us/topics/company/2019/worldleaders2019.html

requests, Twitter refused to remove the tweets which meet its standard for hate speech and ***has not even labeled*** a single tweet. The tweets include the following:





RECOMMENDATIONS

These methods of labeling and referencing reliable information are relevant when it comes to tackling hate speech. They give companies a more diverse toolbox for tackling problematic content, in addition to removing content or blocking an account.[23] Among other things, it is possible to reference authorized sources. These tools allow reliable

[23] See recommendations of the Israel Democracy Institute – Appendix D Recommendation 6: A Diversity of Content-Moderation Techniques

information to be accessed by the user, who usually may not understand why particular content is considered hate speech, even after removal.

These methods are also important on an educational level, by providing an explanation to the user as to why particular content is problematic and is further enhanced by referring users to authoritative sources. This is especially important for young audiences, who are unaware of the boundaries between legitimate content and hate speech.

In this way, it is also important to implement and integrate community standards for users by placing warning labels where antisemitic posts or materials are posted. Accessibility and implementation of community rules among users is also very important, including through the raising of automatic warnings when expressions that may be antisemitic are used. The Israel Democracy Institute suggests attaching reliable information alongside potentially harmful content, as well as warning the user of the consequences of a community rules violation and requesting voluntary removal of content or temporarily restricting its distribution.[24] Had these tools been adopted by social media companies earlier, they could have better faced the complex challenges in examining hate speech which has now proliferated throughout their platforms.

In conversations with government officials, social media company representatives have expressed interest in expanding their capacity to provide educational material to counter Holocaust denial. This is especially important on a platform like TikTok, which is used by hundreds of millions of young users. In this context, Facebook's announcement of a new initiative to reference reliable information when searching the platform for information on the Holocaust should be commended.[25]

[24] Ibid
[25] Facebook, 27 January 2021, https://bit.ly/3j3Ne1q

ENFORCEMENT

Despite the existence of community standards to remove hate speech, including tens of thousands of content moderators and advanced AI algorithms, social media companies still have not been able to close the gap between their own policies and enforcement.

Over the last two years, the amount of hate speech content removed has increased, but it is unclear whether this has decreased the prevalence of hate speech (see below – Transparency). These efforts need to continue to be expanded to create effective change.

In this section, several approaches are suggested that are likely to contribute to reducing antisemitism on social media, and hopefully hate speech more generally as well.

1 LANGUAGES

GOAL FOR SOCIAL MEDIA COMPANIES

Increase enforcement of community standards in languages where antisemitic hate speech is more prevalent

BACKGROUND

Over the years, while much attention has been given to antisemitism in the West, studies have found that the Middle East is the most antisemitic region in the world (according to an ongoing ADL study since 2014, between 70 and 80 percent of Arab countries' citizens hold anti-Semitic views.)[26]

The lack of attention to antisemitism and anti-Israel discourse in the great Middle East region and North Africa stems from, among other things, the language barrier and Western

[26]ADL Global Antisemitism Index, https://global100.adl.org/map/meast.

society's (including Jewish organizations) desensitization to hate speech posted in languages such as Arabic, Farsi, and Turkish.

In the absence of proper transparency regarding the monitoring efforts of social media companies, reliable data about the scope of the phenomenon is unavailable. Israel's efforts to monitor Arabic social media have until today focused on preventative security measures.

Today, hate speech in one area of the globe cannot be overlooked – especially with the ease of its spread through social media at the touch of a button. Indeed, what happens in the Middle East, does not remain in the Middle East.

RECOMMENDATIONS

Social media companies can play an important role in both reducing antisemitic hate speech and strengthening peace and normalization between the Arab world and Israel. This is especially so among the younger Arab generation, whose worldview is influenced, like most young people today, through what they are exposed to on social media. Social media companies have a responsibility for shaping a better future devoid of hate speech. Therefore, government dialog with social media companies should focus on two components:

- **Unified enforcement of the rules against antisemitic hate speech** – in particular, uniformity in enforcement regarding content posted in Western languages and Arabic, Farsi, and Turkish, while investing resources for monitoring, reviewing, and analyzing content in these languages. Facebook has reported that it is investing efforts in Arabic AI capabilities, a measure which should be implemented by additional social media companies.[27]
- **Joint projects sponsored by social media companies which encourage understanding and peace-building educational content.** Peace agreements with

[27] Venture Beat, 11 August 2020, https://bit.ly/2XODCy0

the Gulf States have created a breakthrough regarding the perception of Israel and antisemitism in the Arab world. Cooperation between the Israeli government and the UAE, Bahrain, and Morocco against hate speech can have an important impact. In this context, it should be noted positively that state-affiliated institutions in Bahrain and Morocco have committed themselves to combating antisemitism, while adopting the IHRA definition.[28] Cooperation with these countries may also create opportunity for vitally needed education in Arab countries on Jewish heritage, history, and the contribution of Jews to the Middle East. Such knowledge is currently not widely taught and may serve as a complementary educational tool to combat antisemitic hate speech towards Jews and Israel.[29]

[28] JTA, 25 October 2020 - https://bit.ly/39zkZoc , Times of Israel, 19 January 2021 https://bit.ly/3oArWtE
[29] In this context, it is worth noting that Elan Carr, former U.S. Special envoy for combatting antisemitism, argues that the struggle against antisemitism cannot be won simply by fighting it, but rather through education on Judaism and the Jewish contribution to the world – to create an atmosphere of "Philo-Semitism" or "love of Jews".

2 ANTISEMITIC HATE SPEECH PROPAGATORS

GOAL FOR SOCIAL MEDIA COMPANIES

Remove the accounts of those convicted of antisemitic hate speech, while reducing the virality of suspected hate speech.

BACKGROUND

Hate speech propagators use social media to spread their hate and there is a double harm in this. First, social media algorithms may promote provocative content, including hate speech, making it more readily accessible. Second, the spread and prevalence of hate speech on social media normalizes it – to the detriment of society on- and offline.

Some antisemitic hate speech propagators have a large group of followers. In order to circumvent policy set in place against hate speech, they use antisemitic code words and innuendo that do not trigger automated content monitoring.[30] The lack of intimate knowledge and understanding of the nuances of antisemitic discourse create a significant challenge for social media companies. (See also Training of Content Moderators – below.) The Ministry of Diaspora Affairs has made efforts to alert social media companies the ever-changing subtleties found in antisemitic discourse online. Such efforts are well in place by

[30] Some notable examples from recent years of antisemitic activity:

• The affair surrounding the antisemitic comedian Dieudonné M'bala M'bala in France – who invented the reverse Nazi salute ("Quenelle") that encouraged young French people on Facebook to take pictures of themselves and upload them to Jewish institutions social media accounts throughout Europe.
• The hashtag "Good Jew" #UnBonJuif - The French Jewish student organization UEJF sued Twitter, and even won a lawsuit, to release information about those involved in severe cases of incitement during riots against the Jewish community in France, led by the widespread use of the tag "Good Jew". The court ordered Twitter to pass on to the police all the details of the users who participated in the publication of the hashtag.
• The Joshua Bonhill 2015 case - an anti-Semite who worked to recruit far-right activists via Twitter and Facebook to initiate demonstrations against the presence of Jews on British soil, including in the ultra-Orthodox Stamford Hill neighborhood of London. Bonhill was arrested and convicted.

NGOs and local Jewish communities which help social media companies stay current on the latest general and local trends in antisemitic hate speech.

As mentioned, part of the problem of the virality of hate speech lies with social media algorithms. Leading antisemites have achieved status thanks, in part, to the algorithms of companies that increases their exposure. Companies are aware of this problem and say that they are slowing the spread of hate speech, but do not provide details.

RECOMMENDATIONS

Companies should be more transparent in the ways their algorithms slow virality of hate speech. Detailing the conditions for algorithmic attenuation (reducing the virality) of hate speech will help clarify to users that such content will not become viral. More broadly, algorithms also need to be frequently updated to keep up with the latest nuances of antisemitic hate speech.

Social media companies should also strong consider closing "public figure" accounts of persons who have been convicted in court of ascertainable hate speech offenses (consistent with the international legal conventions on hate speech) – regardless of the content in the account. This will demonstrate that social media companies will not allow convicted hate speech offenders to use of their platform.

3 TRAINING OF CONTENT MODERATORS

GOAL FOR SOCIAL MEDIA COMPANIES

Promote initiatives for the ongoing training of content moderators on antisemitic hate speech by independent organizations, with transparency regarding content moderator training.

BACKGROUND

In recent years, social media companies are devoting management, financial and technological resources to enforce their policies. According to companies' data, the quantity of hate speech content removed increases every quarter. Today, an absolute majority of content monitoring is automated, constantly improving with artificial intelligence.

Despite increasing improvements in AI and capacity to automatically remove content, there are, and always will be, "difficult" cases requiring human scrutiny – carried out by employees known as “content moderators”.

Today, tens of thousands of content monitors operate worldwide, employed mostly by subcontractors for social media companies, and are called into action when human decision making is needed. Content moderators, most of them young and earning slightly higher than the minimum wage, sit in front of a computer screen and go through flagged items – once every 30 seconds, on average. They then decide whether the content violates company policy, and if so – they remove it. [31] These employees are exposed to violent, sexual, and hateful content, which includes antisemitism.

Despite the discussions held in recent months with social media company representatives, there is little clarity as to the nature and depth of training, if any, content moderators receive

[31]The Verge, https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

on antisemitism. Alongside the flood of material that includes sex, violence and fake news, it is difficult to understand, how content moderators are able to distinguish the nuances of antisemitic hate speech which anyone who knows something about antisemitism would be quickly able to identify. (For example, do content moderators know that "the Rothschilds" is a euphemism in antisemitic conspiracy theories for Jewish economic world dominance?)

Although the companies publish community rules, which attempt to define what they consider to be hate speech (see Appendix A), there are also unpublished guidelines for content moderator which the companies refuse to make public.[32] In 2017, *The Guardian* published a series of articles based on documents used to train and direct the content moderator at Facebook that were leaked to the newspaper. These materials show, for example, that there is increased enforcement around Holocaust denial, to "avoid legal liability" in Germany, France, Austria and Israel.[33] Residents of these countries will be less exposed to posts that deny the Holocaust than users in the rest of the world. Although corporate policy has evolved since then, the material shows the large gap between official policy and the actual guidance and enforcement.

RECOMMENDATIONS

An important element in proper enforcement of hate speech policy online is comprehensive training for moderators in hate speech, and, specifically, antisemitism, assisted by independent civil society organizations and experts who understand the nuances and subtleties of antisemitic discourse. For its part, the Israeli government can assist in

[32] The Israel Democracy Institute (IDI) recommends the following:

> Social media companies and other internet intermediaries should clearly define and publish detailed policy guidelines on what is and prohibited regarding hate speech according established human rights standards. They should explain how they apply their policies, especially how context—including social, cultural, and political, including the use of code words, euphemisms, hate speech, and humor are taken into account in decisions about content moderation.

See Appendix D – Recommendations of the Israel Democracy Institute, Recommendation 4 – Detailed guidelines

[33] The Guardian, https://www.theguardian.com/news/2017/may/24/how-facebook-flouts-holocaust-denial-laws-except-where-it-fears-being-sued

assembling and making available to all social media companies an international team of experts and civil society organization for ongoing training of content moderators. The Israel Democracy Institute recommends that

> companies running social media platforms and other content brokers should run effective content moderator training programs so that they intricately know human rights and understand the cultural sensitivities associated with the content they are reviewing.[34]

A recent conversation with Facebook officials revealed that there is currently no company employee or content moderator specializing specifically in antisemitism. Although social media companies' policy directors are knowledgeable about antisemitism and hold roundtables with Jewish organizations and communities to deepen their understanding on the issue, it is unclear how much this knowledge is passed on to the moderators. Thus, for example, it is not at all clear whether content moderators are familiar with, or make any use of the IHRA working definition.

[34] See Appendix D – Recommendations of the Israel Democracy Institute, Recommendation 11: Protect content providers

4 COORDINATED INAUTHENTIC BEHAVIOR

GOAL FOR SOCIAL MEDIA COMPANIES

Increase efforts to identify and remove accounts for inauthentic coordinated behavior which encourages hate speech.

BACKGROUND

Since the U.S. election in 2016, there has been a growing awareness of attempts to run campaigns to influence the public on social media. Using relatively simple technological tools, a handful of individuals can simulate what look like mass grassroots campaigns. This is an extensive social media phenomenon that undermines the credibility of the platform.

This type of activity also exists among Israel delegitimization and antisemitic hate groups, but its scope remains ill-defined. The Ministry of Strategic Affairs published two reports on the subject. One, ahead of the 2019 Eurovision Song Contest which took place in Israel, revealed that the BDS umbrella organization, the Palestinian Academic and Cultural Boycott of Israel (PACBI), and other local boycott organizations around the world were involved in a campaign using bots and fake accounts to give a false impression of massive disapproval of the Eurovision and the participating public broadcasting bodies and artists.[35] In a second report released in October 2020, the Ministry uncovered inauthentic delegitimization campaigns which took place in July 2020 aimed at creating a misleading impression of widespread anti-Israeli sentiment – in the space of four hours Twitter was flooded with some 7,000 anti-Israeli tweets from only a handful of users.[36] Another study conducted by the ministry, which has not yet been published, revealed that bots were also used in online discourse about the normalization between Israel and Arab countries to attack the countries and leaders who signed an agreement with Israel.

[35] The Big Scam, MSA report– Here
[36] Manipulating Social Media, MSA report- Here

Using non-authentic tools to create fake campaigns creates a double-edged sword: it creates hate speech from an unidentifiable source and creates a false inflated impression of public sentiment.

RECOMMENDATIONS

As noted above, there is no comprehensive and accurate picture of the extent to which "coordinated inauthentic behavior" is used to spread and promote antisemitism. That is why it is important that companies improve monitoring and enforcement. Government monitoring of social media for hate speech (see Part II, below) should work to independently better assess this phenomenon, especially in light of recent findings which have indicated efforts to manipulate social media to increase hate speech against Jews and Israel.

Transparency on the part of social media companies is also needed, so that exposed accounts and perpetrators are held to public account. Exposing this phenomenon may result not only in removing a large amount of hateful content, but also in reducing the efficacy of such campaigns as more of them are brought to light.

5 HATE COMMERCE

GOAL FOR SOCIAL MEDIA COMPANIES, E-COMMERCE AND ONLINE PAYMENT COMPANIES

Create global policy regarding, and enforce current rules, prohibiting the sale of Nazi memorabilia and items which promote hate or Holocaust denial on e-commerce platforms.

BACKGROUND

The hate speech, like any other venture, needs funding to thrive. An important source of its funding is e-commerce, where the sale of items both promote hate and constitute a source of income for its propagators. In particular, commerce in Nazi memorabilia garners much interest (Hitler's hat allegedly sold in 2019 for 50,000 euros).[37]

To reduce both the spread of far-right extremism and the financial gain from it, there needs to be a push for the adoption of global policy and enforcement of rules prohibiting trade in Nazi memorabilia and items which promote antisemitism and Holocaust denial. Today, there are some rules in place, including, on some platforms, rules against selling Nazi memorabilia, but even where there are clear policies, actual enforcement is still lacking.

Facebook, which is also used as a commercial platform, has enabled the publication of posts featuring for sale quintessentially Nazi merchandise.[38] Under Facebook company policy, items offered must comply with community standards regarding content. However, selling Nazi items is not explicitly prohibited.[39]

Amazon, the world's leading e-commerce company, has rules prohibiting the sale of items that encourage or praise "violence or hatred against any person or group" as well as items

[37] Bloomberg, https://www.bloomberg.com/opinion/articles/2019-11-21/who-bought-hitler-s-top-hat-the-public-has-a-right-to-know
[38] 7 News Australia, https://7news.com.au/news/social/jewish-community-says-online-trade-in-nazi-memorabilia-is-an-affront-to-holocaust-victims-c-469231
[39] BBC, https://www.bbc.com/news/uk-politics-43667286

on the subject of "human tragedies." [40] A year ago, the company found itself defending against public criticism after the Auschwitz-Birkenau Memorial and Museum complained that bottle openers with photographs of Auschwitz could be bought on the site.[41]

Unlike Amazon, leading companies like Alibaba/AliExpress as well as other Chinese companies directly refer to Nazi items and prohibit the sale of "materials that adopt or support fascism, Nazism and other extremist ideologies."[42]

Alongside e-commerce platforms, online payment companies have no explicit obligation that their platforms will not be used as a means of hate-trafficking. Still, credit card giant Visa has stated it will not provide services to far-right websites. [43] PayPal states that it prohibits transactions "that encourage hatred, violence, intolerance that discriminates on the basis of race and another form",[44] and declared that it would not provide services to extreme right-wing organizations.[45]

RECOMMENDATIONS

Of all the leading companies, eBay appears to have the most detailed and broadest policy, and explicitly bans the sale of Nazi and antisemitic items, as follows:

- Items with racist, anti-Semitic, or otherwise demeaning portrayals, for example through caricatures or other exaggerated features, including figurines, cartoons, housewares, historical advertisements, and golliwogs

[40] Amazon, https://sellercentral.amazon.com/gp/help/external/help.html?itemID=200164670&language=en_US&ref=efph_200164670_cont_200164330
[41] The New York Times, https://www.nytimes.com/2019/12/01/business/amazon-auschwitz-christmas-ornament.html
[42] Alibaba/AliExpress, https://service.aliexpress.com/page/knowledge?pageId=37&category=1000022029&knowledge=1060015168&language=en
[43] The New York Post https://nypost.com/2017/08/16/credit-cards-are-clamping-down-on-payments-to-hate-groups/
[44] Paypal, https://www.paypal.com/us/webapps/mpp/ua/acceptableuse-full
[45] Paypal, https://www.paypal.com/stories/us/paypals-aup-remaining-vigilant-on-hate-violence-intolerance

- Historical Holocaust-related and Nazi-related items, including reproductions
- Any item that is anti-Semitic or any item from after 1933 that bears a swastika
- Media identified as Nazi propaganda[46]

All major companies ideally should have policies with this level of detail.

More important is the implementation – strict and ongoing monitoring and enforcement is needed to ensure that the trafficking of these items does not occur on any platform.

[46] eBay, https://www.ebay.comh/help/policies/prohibited-restricted-items/offensive-material-policy?id=4324

TRANSPARENCY

GOALS FOR SOCIAL MEDIA COMPANIES

Increase transparency and publish data on hate speech, specifically:

- Data segmentation by groups targeted by hate speech on (e.g. segmentation by minority group)
- Geographical segmentation of hate speech (e.g. segmentation by city or local region)
- Prevalence estimate of hate speech content
- Exposure rate to hate speech content before it is removed
- Main propagators (public figures whose content has been removed for hate speech violations)
- Data on campaigns, both authentic and inauthentic, which spread hate speech

BACKGROUND

Social media companies have a responsibility to protect their users from harm and analyzing the data within their platform can help in this regard. This data can shed light on hate speech trends and can help potential victims and governments prevent such hate speech from devolving into real-world harm. In addition to being transparent to the public, data from social media companies can also assist in the fight against hate speech. For example, if data shows that in a specific city or region there is an unusually high level of hate speech online, it could be possible for education or law enforcement locally to address the issue and prevent deterioration to violence. Similarly, in response to a hate speech against a particular group, local law enforcement and the community could increase the security around the targeted group. Just as companies assist law enforcement agencies when life is at risk as they are obliged by law, greater data transparency can also be key to prevent violence.

Currently, the leading social media companies publish a single quantitative figure in connection with the amount of hate speech: the number of content items removed in a given quarter of the year.

For example, Facebook reported that in Q4 2019, steps were taken against 5.7 million pieces of content which violated the company's policy. In Q1 2020, 9.6 million pieces of content were removed; and in Q2, thanks to advances in artificial intelligence ("AI"), 22.5 million items were removed. In Q3 2020, a similar amount – 22.1 million items – was removed.[47]

For the first time, Facebook also published a very important, additional piece of data - their assessment of the prevalence of hate speech on the entire platform during Q3 of 2020. According to the company's estimate, hate speech amounts to about 0.1% of the total content; in other words, one in a thousand items is estimated to contain hate speech content.[48]

Twitter reports that in the second half of 2019, 970,000 accounts were dealt with for hate speech (they define it as "hate conduct"), an increase of 54% from the first half of 2019.[49] On YouTube, 80,000 videos were removed for hate speech in Q2 2020 compared to 36,000 in Q1 2020.[50]

All companies have seen a significant increase in the removal of hate speech content in the last quarter. At the same time, there is still a long way to go for companies in terms of transparency in the decision-making process for removing hateful content, including a user's ability to appeal the company's decision to remove or not to remove content.[51]

Despite this data on quantity of hate speech removed, being more transparent and share additional data could be helpful in understanding the scope of and combatting hate speech online. Simple and important questions remain unanswered: For Facebook, which

[47] Facebook, https://transparency.facebook.com/community-standards-enforcement#hate-speech
[48] Reuters, https://www.reuters.com/article/us-facebook-content/facebook-offers-up-first-ever-estimate-of-hate-speech-prevalence-on-its-platform-idUSKBN27Z2R0
[49] Twitter, https://transparency.twitter.com/en/reports/rules-enforcement.html#2019-jul-dec
[50] Google, https://transparencyreport.google.com/youtube-policy/removals
[51] See Israel's Democratic Institute Report, Appendix D, Suggestions 8,9, and 10.

published a prevalence figure for the first time, it is still unclear what the correlation is between this prevalence percentage and the number of content items removed, i.e., what percentage of the total estimated hate speech content does the 22 million removed items represent? For the other companies which have not release prevalence data, it is not at all clear what the frequency of hate speech content is, nor is it clear whether the removed content constitutes a negligible percentage of the total existing content of hate speech or whether it constitutes a significant portion. Furthermore, of the content that was not automatically removed before it was viewed, how many items were reviewed and not removed? How long did it take, on average, to remove content after receiving a report? Which 'public figure' accounts have repeatedly violated company hate speech rules?

Of all the hate speech content removed on a platform, what is the segmentation of groups targeted (e.g. out of the overall amount of hate speech content, how much is antisemitic, Islamophobic, etc.)? This data can also be segmented by location (city, region, country.)

Beyond the hate speech category, calls for violence and misinformation can be segmented by topic, target audiences, and geographical region. Have accounts removed for inauthentic behavior been engaged in hate speech, and if so, who has been the target of this activity?

RECOMMENDATIONS

The companies, as mentioned, are aware that they have a way to go with data transparency, though they rightly note that there has been a significant improvement in their transparency reports over the last two years and this trend is expected to continue.

But in the face of the social media's continued growth, its data transparency is still very limited, and companies refuse to allow access to researchers or experts access to their dashboard data.

Regarding the data companies publish today – the amount of content items that have been removed – this data point lacks context, and it is unclear whether it represents a small or significant amount of hate speech content on the platform. Thus, **the ultimate goal should be on lowering *the prevalence* of hate speech content on the platform.**

In conclusion, social media companies should demonstrate greater transparency, releasing more comprehensive data on hate speech, including segmentation by targeted group, geographical segmentation and exposing key 'public figure' accounts which have repeatedly violated company hate speech rules.

PART II
GOALS FOR THE GOVERNMENT

Part I of this policy paper detailed the main goals of the government vis a vis the social media companies, with an eye to combating antisemitic hate speech. This part proposes areas in which the government should act, in coordination between the relevant government ministries, to accomplish the objectives set out in Part I.

Leading government activity is an inter-ministerial working group which should consider action in the following areas: effective monitoring of social networks; examining regulatory options; cooperation between governments (G2G) and multinational bodies; cooperation and international regulation against extremist websites; building international coalitions to combat hate speech, consulting and convening round tables with civil society organizations.

The following are the recommended areas for government work:

1 INTER-MINISTERIAL WORKING GROUP ON HATE SPEECH ONLINE

With the formation of the government in May 2020, relevant government ministries began a process of coordinating outreach to social media companies. Ministry representatives met with top social media executives at Facebook, Twitter, TikTok, and Google. As part of these meetings, officials reviewed the policies and measures set emplace to prevent hate speech, with a focus on antisemitism.

In order to achieve the goals vis-à-vis social media companies proposed in this document, and to increase coordination between government ministries, a mandate for the inter-ministerial working group should be established through a formal government decision, along these suggested parameters:

A. An inter-ministerial working group will be responsible, by a formal government decision, to combat online antisemitic hate speech, in its various forms. The mandate of the working group will be to formulate and implement an action plan, drawing on this document and other expertise, and report to the government on the trends of such hate speech and its progress in implementing change.

B. Members of the working group will be appointed by the relevant ministers and will include representatives from the following ministries: the Ministry of Strategic Affairs; Ministry of Diaspora Affairs; Ministry of Foreign Affairs; and the Ministry of Communications.

C. The inter-ministerial working group will be headed by two joint chairpersons: The Director General of the Ministry of Diaspora Affairs and the Director General of the Ministry of Strategic Affairs.

D. The team will convene once a quarter and the agenda for the meeting will be set by the working group chairpersons.

E. The team will approve the annual action plan to address antisemitic hate speech, in both its classic and new forms.

F. Progress in relation to the goals vis a vis social media companies on policy, enforcement and transparency will be presented at each meeting.

G. The team will examine the required activity in the following areas:

- Monitoring social networks.
- Regulation.
- Cooperation with other governments, multinational and international organizations; cooperation with NGOs; building coalitions to combat hate speech.
- Coordinated international action against extremist websites and social media platforms

H. The working group will develop a specialized strategy for dealing with other leading social media companies - VK and Telegram. The strategy will be based on the principles laid out in this document but will be unique to each of these companies.

I. The working group will be joined, as needed, by experts in the fields of technology, academia, and civil society.

J. Reports:

- Once a year, the working group will report to the government regarding hate speech on social media and the status of the team's action plan. This report will also be submitted to the Knesset Committee for Immigration, Absorption and Diaspora Affairs.
- The working group will publish periodic public reports regarding the extent of hate speech on social media that include data relating to the prevalence of hate speech online, and actions taken by social media companies.

2 HATE SPEECH MONITORING

Regarding online hate speech on social networks, the Ministry of Diaspora Affairs is currently working to monitor expressions of antisemitism and the Ministry of Strategic Affairs charged with monitoring delegitimization and Palestinian incitement.[52] Yet, the scope of existing monitoring tools does not allow for optimal coping with the current challenges. Improving and expanding monitoring capabilities may enable better dealing with the three main government goals vis a vis social media companies (discussed in Part I) and obtaining a

[52] There is a system for reporting content to social media companies operated by the Cyber Department of the State Attorney's Office. This system was established in collaboration with social media companies in 2016 in the wake of a renewed wave of terrorism to report in real time incitement to violence or publications by terrorist organizations. Similar reporting systems with companies exist in many countries around the world. This type of content is fundamentally different from hate speech discussed in this document. Calls for violence or content of terrorist organizations are prohibited by all companies and are categorized separately from community rules on hate speech.

comprehensive and independent assessment relative to:

1. Social Media Company Policy – An effective monitoring system will support dialogue with social media companies, allowing the government to assess implementation of the effectiveness of their policies.
2. Enforcement – Continuous monitoring, with results released to the public, will increase the focus of social media companies on areas highlighted for improvement.
3. Transparency – The low level of transparency in social media company reports does not allow for an accurate assessment of the scope of antisemitic hate speech and its accompanying trends. Building independent monitoring capabilities may enable an alternative mechanism for making broad assessments and publicizing them.

Government Goals for Hate Speech Monitoring:

1. Monitoring hate speech according to the IHRA Working Definition of Antisemitism.
2. Monitoring and identifying content published by propagators of antisemitism on social media.
3. Monitoring antisemitic hate speech in specific languages - alerting social media to gaps in policy enforcement in languages with higher antisemitic hate speech prevalence.
4. Monitoring viral campaigns and inauthentic accounts which seek to promote hate speech.

3 REGULATION – ISRAEL AND INTERNATIONAL

In Israel, there is currently no law specifically relating to "hate speech" on social media. Although the offense of "incitement to racism" exists in the Israeli penal code, which includes antisemitism, as well as inciting violence and encouraging terrorism, it has not

been used extensively, especially regarding social media content.

In 2016, the government introduced a bill designed to organize the State of Israel's dealings with the publication of hateful content online, provided the publication constituted a criminal offense.[53] The bill did not pass, but today in Israel there is an informal mechanism for removing criminal content through the Ministry of Justice.

As noted, antisemitic hate speech, as defined it in this document, relates to content consistent with the IHRA Working Definition of Antisemitism – a definition adopted by the Israeli government – which is an internationally accepted benchmark to define constitutes antisemitic hate speech.

The Supreme Court's 1996 ruling (Rabbi Ido Alba v. State of Israel) refers to antisemitism as "racism as defined by law":

> Antisemitism, whether in its modern connotation or in its connotation as hatred of Jews and Judaism throughout the ages, is also racism as defined by law. Therefore, incitement to antisemitism would constitute an offense of incitement to racism, even though Israeli Jews constitute a majority and are not exposed to antisemitic incitement as they were in in countries of their dispersion.[54]

Insofar as regards violations of the law, a country can force a social media company to remove content even in cases where that content may not violate the company's community standards. This is done, for example, in Germany, where Holocaust denial is illegal. In Israel, too, there is a law against Holocaust denial. The social media companies are obliged to remove the problematic content – but only in countries where publishing such content is in violation of the law. This means that even if antisemitic content in Israel is illegal, including Holocaust denial, the company will remove the content ***only for users only in Israel***. This situation of course does not meet the needs in the fight against

[53] The Knesset, https://main.knesset.gov.il/Activity/Legislation/Laws/Pages/LawBill.aspx?t=lawsuggestionssearch&lawitemid=2011567

[54] The 7th Eye, https://www.the7eye.org.il/verdicts/51137

antisemitic hate speech, which calls for content to be removed globally quickly and effectively.

Comparing legislation in foreign countries raises widespread interest, as concern grows in countries around the world about the spread of hate speech on social media. (See Appendix C, which reviews regulation in various countries and multinational bodies). For example, Germany has enacted the strictest regulation, which requires companies to remove content that violates German law within 24 hours, and to publish semi-annual reports on the handling of all complaints received through the government system for filing complaints about online content.

The EU is introducing broad regulation of online platforms, the "Digital Services Act", which will include sections on hate speech. [55] The UK is also working on a comprehensive regulation, the main points of which were formulated in a document called the "Online Harms White Paper".[56]

Together with representatives of the Ministry of Justice and legal professionals who specialize in hate speech on social media, the inter-ministerial working group should examine and consider whether regulation of antisemitic hate speech is necessary and what form it might take.

It is worth noting the U.S. public debate surrounding Section 230 of the Communications Decency Act, which gave social media companies legal immunity for the content on their platforms, and arguments for significant change to Section 230 could change the current paradigm, whereby today social media companies are not legally responsible for the content posted to their platform. A change to Section 230 may also have the effect extremist websites (see section "International Cooperation and Regulation against Extremist Social Network Platforms" below).

[55] https://ec.europa.eu/digital-single-market/en/digital-services-act-package
[56] https://www.gov.uk/government/consultations/online-harms-white-pape

Government goals on regulation:

1. Examining the need for a legislation on antisemitic hate speech online, while considering the following:
    A. What types of antisemitic hate speech content is it reasonable to mandate that social media companies remove, given the principle free speech?
    B. If regulation requiring the removal of antisemitic hate speech content in Israel will result in companies removing such content for Israeli users only, is such regulation worthwhile?[57]
    C. What can be learned from the experience of countries and the EU in establishing mechanisms for removing and treating hate speech as a possible generator of violence?
    D. Weighing the advantages and disadvantages between national or international regulatory mechanisms versus a voluntary arrangement with social media companies, in terms of efficiency and freedom of expression.
2. Formulating a recommendation to be submitted to the relevant ministers regarding hate speech regulation.

4 G2G - COOPERATION BETWEEN GOVERNMENTS AND MULTINATIONAL ORGANIZATIONS

Many countries are aware of and acting on the challenges posed by social media, including hate speech. The EU, as mentioned above, is presenting comprehensive regulation on the subject. The United States Congress recently held a number of hearings with the heads of social media companies, where, among other things, the

[57] However, it may be important to regulate hate speech in general, with the understanding that such Israeli legislation will make it possible for Israel demand similar legislation from other countries in the world. Article 20 of the International Covenant on Civil and Political Rights, which was ratified by Israel, might form the basis of such legislation.

topic of hate speech online was discussed. While antisemitism online is major concern in many countries, Israel's demands that social media companies to act resolutely against hate speech may be welcome by countries also looking to tackle this issue. The Ministry of Foreign Affairs has maintained a dialogue with the European Union for years on antisemitism, and now a key focus is antisemitic hate speech online.

Given social media's global reach and hate speech a problem across languages and nations, this issue readily lends itself to international cooperation. Such broad cooperation in the fight against hate speech will strengthen the universal value of nondiscrimination against any protected group.

Also, as Israel is perceived by many countries as a high-tech it is well placed on the issue of online hate speech in terms of its technological capacity and enable possibly working together with other countries on the effort.

Building broad international coalitions to combat hate speech will give Israel new leverage in its discussions with social media companies, which themselves are multinational corporations. Beyond the added leverage vis-à-vis the social media companies, cooperation between countries and international bodies may encourage the sharing of information, new technology, and methods to reduce the scope of hate speech.

Further examination is required to create similar relationships on both the governmental and civil society level with the Arab countries Israel recently signed peace accords and normalization agreements with. Such an alliance would likely lead to progress dealing with antisemitism within the Middle East.

Government Objectives for Governmental and Multinational Cooperation:

1. Led by the Ministry of Foreign Affairs, the inter-ministerial working group will offer opportunities for cooperation between countries on hate speech.
2. Led by the Ministry of Foreign Affairs, the working group will propose ways in which the State of Israel can take a more active role in promoting the issue among multinational organizations tackling hate speech. This would be achieved while offering proposals for cooperation with governmental bodies, international institutions, and civil society organizations (Jewish and non-Jewish) which may be involved in combating hate speech and antisemitism online.
3. Specific attention should be given in working with countries signatory to the Abraham Accords on the following issues:
    A. Hate speech on social media.
    B. Promoting government and civil society initiatives for education on antisemitism.
    C. Integrating social media companies into joint ventures that advance values for tolerance and understanding between peoples and religions.

5 INTERNATIONAL COOPERATION AND REGULATION AGAINST EXTREMIST SOCIAL NETWORK PLATFORMS

While this policy paper focuses on leading social media companies, it is important to note that a major source of antisemitic hate speech (especially from the extreme right) is not found on these platforms, but rather, on fringe sites. From there it makes its way to more mainstream to social media networks and receives greater levels of exposure.

These are extremist sites that have set deliberately flexible policies and attract the most extreme and fanatical posts and content – ranging from pornography to hate speech. (This is not referencing sites found on the dark web, but rather easily accessible web

platforms). These platforms, including 4chan, Gab, and 8kun (formerly 8chan), knowingly enable, extreme discourse and content sharing and eschew contact with government authorities.

As a first step, simply exposing them and their content to the public and decision makers can raise awareness of the issue and even force these companies to act.[58]

As important is the recognition that this is an international problem, which requires international cooperation and regulation. There are a variety of measures that can be taken against these sites, especially vis-à-vis companies which provide them hosting services and cybersecurity. While most companies refuse services to these websites, one or two exceptions are sufficient to keep them functioning. Here is where international regulation and legal liability vis-à-vis hosting and other service providers may be helpful.

Government Goals Regarding Extremist Sites:

1. The inter-ministerial working group will formulate a list of leading extremist platforms that enable antisemitic hate speech. The list will also include the owner profile; service providers, server providers; and the policies of these companies. The team will examine how countries deal with these sites through regulation and/or bringing the perpetuators to justice.
2. The team will formulate a strategy, consulting with government officials and professionals internationally, to tackle these extremist platforms, with an emphasis on case studies first.

[58] A good example is the recent exposure in the *New York Times* of the sex site Pornhub, which resulted in the company changing its rules of use within days because of the public attention raised. See https://www.nytimes.com/2020/12/09/opinion/pornhub-news-child-abuse.html

6 BUILDING BROAD INTERNATIONAL COALITIONS TO COMBAT HATE SPEECH AND CONSULTING & CONVENING WITH CIVIL SOCIETY ORGANIZATIONS

Consultation with Jewish communities, leadership and NGOs is especially important for conducting authentic, civil and communal discussions with social networking companies. The threats that Jews have suffered, the feelings that have accompanied them in recent years, and avoidance of performing religious in public spaces are all part of the Jewish experience even in the virtual space.

A number of government ministries met with relevant Jewish organizations on antisemitism and related issues. Consolidating this discourse and managing it coherently with a variety of civic organizations may lead to better results vis-à-vis social media. A common round table for the government and relevant organizations to examine and compare governmental and civil society efforts regarding antisemitic online hate speech is important to bring about more effective results.

The inter-ministerial working group may also recommend convening a broad, global forum of all stakeholders – governments anf social media companies, alongside experts in technology, law, education, and civil society organizations. The forum can discuss, develop, and evaluate the implementation of international best practices standards to reduce hate speech online.[59]

The working group should also consider how to raise awareness of the issue within the Israeli public, how it can assist Israeli NGOs active in the field, and how to connect them with relevant multinational organizations and NGOs abroad.

In addition, the working group should consider the encouragement of academic research on hate speech on social networks - both on the technological level of developing

[59] See Appendix D – Recommendations of the Israel Democracy Institute - Recommendation 14: Global Stakeholder Forum

algorithms and interface to prevent hate speech, and on the level of analysis of social media policy and enforcement. Research is also important in examining activity patterns of antisemitic hate speech, including at extremist sites.

Government Objectives with Civil Society Organizations:

1. Establishing a round table on antisemitic hate speech online with NGOs leading in this field.
2. Convening an international conference on hate speech with NGOs and social media company representatives. (For example, the biennial conference of the Ministry of Foreign Affairs and the Ministry of Diaspora Affairs on combatting antisemitism or the GC4I Forum consisting of the heads of prominent pro-Israeli organizations organized by the Ministry of Strategic Affairs may be an appropriate platform for such a convening).

APPENDICES

APPENDIX A

POLICY COMPARISON OF LEADING SOCIAL MEDIA COMPANIES

Legend
V – this type of content is removed according to company policy
X – this type of content is NOT removed according to company policy

Type of Content	Facebook, Instagram	YouTube (Google)	Twitter	TikTok
Calls to Violence	V	V	V	V
Classic antisemitism	We define hate speech as a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation…. Content targeting a person or group of people… Violent speech or support in written or visual form Designated dehumanizing comparisons, generalizations, or behavioral statements (in written or visual form) harmful stereotypes- that include…. Jewish people and rats Jewish people running the world or controlling major institutions such as	We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: Ethnicity; Gender; Nationality… Religion… Conspiracy theories saying individuals or groups are evil, corrupt, or malicious based on any of the attributes noted above.	You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.	We define hate speech or behavior as content that attacks, threatens, incites violence against, or otherwise dehumanizes an individual or a group on the basis of the following protected attributes: Race; Ethnicity; National origin; Religion… Do not post… conspiracy theories used to justify hateful ideologies.

	media networks, the economy or the government			
Holocaust denial	V Posting content about a violent tragedy, or victims of violent tragedies that include claims that a violent tragedy did not occur. **Denying or distorting information about the Holocaust**	V We will remove content denying that well-documented violent events, like the Holocaust or the shooting at Sandy Hook Elementary, took place *(from Google blog)*	X ***The company announced it would remove Holocaust denial but did not update its policy*** We prohibit targeting individuals with media that depicts victims of the Holocaust… symbols historically associated with hate groups, e.g., the Nazi swastika. Images altered to include hateful symbols or references to a mass murder that targeted a protected category, e.g., manipulating images of individuals to include yellow Star of David badges, in reference to the Holocaust.	V Content that promotes any hateful ideologies by talking positively about or displaying logos, symbols, flags, slogans, uniforms, salutes, gestures, portraits, illustrations or names of individuals related to these ideologies Content that denies well-documented and violent events have taken place
New antisemitism (directed against Israel, per IHRA definition)	X	X	X	X

APPENDIX B

EXAMPLES FROM THE IHRA WORKING DEFINITION OF ANTISEMITISM BY THEME

Examples from the IHRA Working Definition			
Theme	**Classic Antisemitism**	**Both Classic and New Antisemitism**	**New Antisemitism**
Denying the right to exist	Calling for, aiding, or justifying the killing or harming of Jews in the name of a radical ideology or an extremist view of religion.		Denying the Jewish people their right to self-determination, e.g., by claiming that the existence of a State of Israel is a racist endeavor.
Classic hate motifs against Jews	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		Using the symbols and images associated with classic antisemitism (e.g., claims of Jews killing Jesus or blood libel) to characterize Israel or Israelis.

	or other societal institutions.		
Unfair standard for blame	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.	Holding Jews collectively responsible for actions of the state of Israel.	Applying double standards by requiring of it a behavior not expected or demanded of any other democratic nation.
The Holocaust	Denying the fact, scope, mechanisms (e.g. gas chambers) or intentionality of the genocide of the Jewish people at the hands of National Socialist Germany and its supporters and accomplices during World War II (the Holocaust).	Accusing the Jews as a people, or Israel as a state, of inventing or exaggerating the Holocaust.	Drawing comparisons of contemporary Israeli policy to that of the Nazis.
	Accusing Jewish citizens of being more loyal to Israel, or to the alleged priorities of Jews worldwide, than to the interests of their own nations.		

APPENDIX C

ONLINE HATE SPEECH FROM A LEGAL PERSPECTIVE:

LEGISLATION IN ISRAEL AND AROUND THE WORLD

ISRAEL

Hate Speech and Hate Crimes

There is currently no law that specifically discusses "hate speech" (or any synonyms thereof). The most comparable legislation is the crime of incitement to racism in the Penal Code, 5737-1977 (the "**Penal Code**"). Racism is defined in section 144A of the Penal Code as follows:

> Persecution, humiliation, degradation, a display of enmity, hostility or violence, or causing violence against a public or parts of the population, all because of their color, racial affiliation or national ethnic origin.

According to the language of the law, this offense is a behavioral offense. That means there is no need for the result component. According to section 144B of the Penal Code, any publication of which purpose is to incite to racism is prohibited, and the question whether the publication led to racism or even if its content contained untrue content has no meaning - this will still be considered an offense, with a sentence of 5 years imprisonment.

Other expression offenses are incitement to violence (the Penal Code), as well as incitement to terrorism (the Anti-Terrorism Law).

It should be emphasized that with regard to the above offenses, it is possible to file an indictment only with the consent of the Attorney General of Israel.

The Penal Code does include a special reference to hate crimes:

Section 144F of the Penal Code provides for "hate crimes" as those committed out of a **racist** motive. According to section 144F(a) of the law, anyone who commits an offense out of a racist motive (as defined in s. 144A), will be liable to **double** the punishment set for the offense committed or 10 years imprisonment (the lightest of them). That is, the legislator has expressed his opinion and ruled that acts committed for racist motives should be denounced and therefore the punishment for them will be more severe. Similarly, the section provides that an offense committed "out of **hostility towards the public** because of religion, religious group, ethnic origin, sexual orientation or for being foreign workers" will also result in a more severe punishment than the usual punishment for the offense.

The section refers to the following offenses: an offense against the body, freedom or property, an offense of threat or extortion, offenses of disorderly conduct or public obstruction and nuisances included in articles Nine and Eleven of Chapter 7, and an offense in the public service, all except for an offense for which the penalty is ten years imprisonment or more.

The Bill for the Prevention of Offenses through their Publication on the Internet (Removal of Content), 5778-2018

Background

The bill to prevent the commission of offenses through the Internet (Removal of Content), 5778-2018 consists of the unification of the original bill by former MK Revital Swid and of the bill presented by the Minister of Public Security, Gilad Erdan and the Minister of Justice Ayelet Shaked from 2016, and was later referred to as the "Facebook Law".

The bills were born against the background of the wave of knife terror attacks that swept the State of Israel at the time, partly because of severe online incitement. It

should be clarified that the bill was intended to deal with criminal offenses and not with hate speech against Israel.

As stated in the explanatory memorandum to the government bill, the purpose of the proposal was to streamline the State of Israel's handling of the publication of content on the Internet. As part of the bill, the creation of an additional tool of an administrative nature was proposed, which would be used by law enforcement agencies in dealing with the phenomenon, **alongside** the penal tool and not as its substitute - a court order to remove content from the Internet.

The First Reading of the bill took place at the 186th session of the 20th Knesset on January 2,2017. The law passed by 36 to 2, without any abstentions. The bill was then discussed by the Joint Committee on the Constitution, Law and Justice Committee and the Science and Technology Committee in preparation for the Second and Third Readings, and was not advanced beyond that. This was because of a directive by the Prime Minister Benjamin Netanyahu. The Likud party explained that this step stemmed from the fact that "the proposal would allow censorship of opinions and would severely infringe on freedom of expression in the State of Israel".[60]

The Regulation Proposed in the Bill

The bill has undergone a number of changes; the description below is based on the latest wording of the bill prepared for the Second and Third Readings. The bill proposes to empower the District Court[61] to issue an order to remove content from the Internet (including from search sites), including social networks.

In order not to disproportionately infringe on freedom of expression, a double check was proposed, which, only when fulfilled; would allow issuing an order:

1. Publishing the content constitutes an act that is a criminal offense.

[60] https://www.ynet.co.il/articles/0,7340,L-5312098,00.html
[61] The original bill proposed to empower the Administrative Court.

2. There is a real possibility that the continued publication of the content would harm the security of a particular or indefinite person or the security of the state, or would lead to serious damage to the state economy or vital infrastructures.

It was also proposed to establish a number of additional provisions in order not to infringe on freedom of expression beyond what is required:

1. A request for removal of content to the court will be submitted only after the Attorney General, or a person authorized on his behalf, has given his written consent.
2. The court that will hear the application will be the District Court Judge authorized by the President of the District Court.
3. The obligation to publish court decisions regarding applications under this law.

Territorial Applicability

During the Knesset Committee's deliberations, Adv. Wismonsky, of the Cyber Department of the State Attorney's Office, explained the complexity of the law in a cross-border online world. When there is a request from a particular country to remove content, internet providers can choose one of two ways:

1. Remove the entire content for all and sundry. The counterargument could be that this has a chilling effect on freedom of expression in places where it is allowed, and that there is supposedly an "out of borders" influence of a certain country on other countries.
2. Block only local access.

Criticism Voiced Against the Bill

The main argument against the "Facebook Law" is the violation of freedom of expression, through disproportionate censorship. The law will allow one party to go to court without having to initiate criminal proceedings and without the need for

admissible evidence and claim that there was an offense from among any of the offenses in the Penal Code, thus leading to content censorship.

Furthermore, the Israeli Internet Association called the bill "blindfolding the public in Israel", because the content would be available and accessible to everyone in the world, except for the Israelis. The proposal would also reduce the incentive for law enforcement agencies to act to locate and apprehend offenders, which should be their main activity.

During the Committee's deliberations, criticism was leveled against the bill on the grounds that it was worded very broadly, as it applies to **any** criminal offense. Although in the explanatory memorandum to the bill there was an attempt to clarify that there are restrictive measures by definition, and even in the context of the deliberations, there seemed to be an intention to reduce the discourse to which the law refers to some specific offenses that the law is intended to address[62], however the wording does not explicitly reflect this reduction. There was also concern that the grounds of "state security" upon which the law is based, are broad and vague, and could become an “wide open” concept and not only that, but that this tool would become a first tool instead of an investigative procedure/appeal to the publishing entities themselves.

In addition, during deliberations in the Knesset, there was concern that the law does not allow for effective handling, since even if there is a quick handling within 48 hours, the damage from publication can be caused in minutes and seconds. Finally, there was concern that the law, as is, was not clear enough. For example, in the case where copies of content are created or when content is included within other content, then it is not clear what would be the fate of this type of content.

[62] The legal counsel of the Constitution Committee proposed to establish a defined list of offenses.

The Normal Situation - The Voluntary Route

There is a **voluntary** mechanism working to remove content that constitutes a criminal offense from the Internet. This mechanism is managed by the Cyber Department of the State Attorney's Office, by which the department coordinates requests for content removal from various government agencies and addresses social media contacts with requests for content removal, based on an alleged violation of the terms of use of those networks. The application depends on the goodwill of the companies, and in practice, over 80% of the requests are accepted[63].

Depending on the work procedure of the department, there are a number of conditions for the department to apply for the removal of specific content:

1. The content violates the terms of use of the platform.
2. The content allegedly amounts to a criminal offense under Israeli law.
3. The existence of a clear public interest that justifies reporting on the said content.

A petition by the "Adalah" organization is currently pending before the Supreme Court in relation to the mentioned mechanism.

COMPARATIVE INTERNATIONAL LEGISLATION

European Union Guidelines

Background

On March 1, 2018, the European Commission issued a "Recommendation on Measures to Effectively Tackle Illegal Content Online"[64]. The recommendation is an expression

[63] "Sharp increase in removal of materials by social networks upon request of the Ministry of Justice" https://www.calcalist.co.il/internet/articles/0,7340,L-3728413,00.html

[64] https://ec.europa.eu/digital-single-market/en/news/commission-recommendation-measures-effectively-tackle-illegal-content-online

of the political determination, published by the Commission in an official statement in September 2017, in a non-binding legal form.

The content of the recommendation

The document requires online platforms for increased responsibility when it comes to managing content uploaded to the web. The document proposes to adopt a common approach for all the platforms, for a quick and proactive detection, removal and prevention of the recurrence of harmful content. The recommendation lists five concrete proposals for online platforms:

1. **Defining notice and action procedures** - Online platforms should set clearer and more transparent rules to enable notification of illegal content, including speedy procedures for trusted flaggers. Content providers should be updated on these decisions so as to give them the opportunity to appeal them, in order to avoid removing illegal content unnecessarily.
2. **More effective tools and proactive technologies** - Companies are supposed to create proactive tools to detect and delete illegal content, especially when it comes to content related to terrorism and content that does not require context in order to be defined as illegal. For example, materials on the sexual exploitation of minors or counterfeit goods.
3. **Stronger safeguards to ensure fundamental rights** - Companies need to establish effective and appropriate safeguards, including human oversight and verification, to ensure that content deletion decisions are accurate and well founded. This is to avoid violating fundamental rights, freedom of expression and the protection of privacy that may arise because of the use of automatic mechanisms.
4. **Preferential treatment for small companies** - Through voluntary agreements, experience sharing, working methods and technological solutions, including tools for automatic content identification. The mutual involvement between

companies in the area should benefit small companies with limited resources and expertise.

5. **Closer co-operation with the authorities** - Insofar as there is evidence of a serious criminal offense or a suspicion that illegal content poses a threat to life or safety, companies should immediately notify law enforcement agencies.

Additional EU Regulation Regarding Illegal Content

1. **Directive (2017/541) on Combating Terrorism**

The Directive of the European Parliament and of the Council of the European Union of 15 March 2017 (2017/541) on combating terrorism stipulates that Member States should establish that service providers are obligated to prohibit the publication of terrorist content, as defined in the Directive (Article 21), within the framework of their terms of use. In addition, Member States should require that content of a terrorist nature be completely deleted from the platform, even for users outside its territories, as far as possible. If this is not possible, steps should be taken to block such content for citizens in their territory.

Measures for the purpose of deleting and blocking content must be transparent and include appropriate protection mechanisms, in order to maintain a balanced use and, if necessary, update users regarding the removal of the content uploaded to the web. Protective mechanisms regarding deletion or blocking of content should also include the possibility of legal oversight.

2. **Code of Conduct for International Companies**

As of May 2016, the European Commission has signed agreements with many IT Companies[65] supporting this code. The Code does not constitute a regulation, resolution or directive and is not binding but constitutes a voluntary tool of self-

[65] The companies that signed this Code: Facebook, Twitter, Microsoft and YouTube (Google), Snapchat, Jeuxvideos.com, Dailymotion, Instagram, Google +, TikTok.

regulation by the companies vis-à-vis the EU (self-regulatory and non-binding instrument). The Code does not bind countries within the EU.

In the Code, companies have agreed with the EU Commission to set up anti-hate speech rules and community standards on platforms, which prohibit hate speech by users and allow users to report content that violates the established rules. In addition, the Code states that companies must go over most of the content reported by users within 24 hours of reporting; introduce learning and training mechanisms for the teams reviewing the content; initiate collaborations with civil society organizations to expand the circle of reliable reporters of content that violates the rules; generate collaborations with trusted flaggers and people who can create educational content for the platform; appoint contacts within the countries or arenas (national focus points) that know how to respond to requests received by authorities within the countries and increase transparency (with an emphasis on publishing the number of complaints and reports received based on incitement and hate speech).

3. **Ruling against Facebook in the Court of Justice of the European Union**

In October 2019, the European Court of Justice (ECJ) in Luxembourg in a lawsuit filed by an Austrian politician (Eva Glawischnig-Piesczek) against Facebook ruled that social networks must remove illegal content within EU territory and potentially beyond as well.[66] Although the proceeding dealt with Facebook, the court stated that the ruling was relevant to all online networks operating in the EU. In addition, the court confirmed the politician's standpoint that the provisions of the 2000 Directive on e-commerce do not prevent a court from an EU Member State from demanding from an online platform to remove identical, similar or dubious content for other users in the EU, by means of a decree.

[66] http://curia.europa.eu/juris/documents.jsf?num=C-18/18

Germany

The "**Network Enforcement Act**" ("Netzwerkdurchsetzungsgesetz" or "NetzDG" in short)[67], entered into force on October 1, 2017. The law applies to telemedia service providers which, for profit-making purposes, operate internet platforms which are designed to enable users to share any content with other users or to make such content available to the public (social networks). The definition of "social networks" does not include platforms offering journalistic or editorial content, the responsibility for which lies with the service provider itself, platforms that enable individual communication or distribution of specific content, and not to social networks with less than two million registered users in Germany.

The law is designed to combat the spread of hate speech, intentional misinformation ("fake news") and other criminal content viewed on social media. Such content includes insults, malicious gossip, defamation, public incitement to commit a crime, incitement to hatred, distribution of violent images and threats to commit a crime.

The law has five key components:

1. **A mechanism for the effective management of complaints** - The mechanism defines binding standards for handling complaints and demands their transparency, while requiring social network operators to offer the use of a simple, accessible and available system at all times for reporting content of a criminal nature.
2. **Duty to report** - Social network operators are required to submit biennial reports regarding the handling of complaints about content of a criminal nature. These reports must include extensive information, including the number of complaints and the decision-making procedures of the social network, as well as about the teams responsible for processing the reported content. These reports must be made available online to the public.

[67] https://www.bmjv.de/DE/Themen/FokusThemen/NetzDG/NetzDG_EN_node.html

3. **Fines** - Social networks whose mechanism is not suitable for the effective management of complaints or who fail to produce such a mechanism, commit a regulatory offense. Failure to delete criminal content at all or deletion made partially or late, also constitute regulatory offenses. A fine can also be imposed on a social network that does not meet its reporting obligations in full or at all.

4. **A person authorized to receive official documents** - Social networks will be required to appoint a person who is a resident of Germany, who will be authorized by the company to receive official documents relating to fines and civil proceedings. The network must publish his/her details to the public. Additionally, a person resident in Germany who is authorized to obtain official documents or information from law enforcement agencies, and to ensure that they can respond immediately to these requests. Violation of this duty may result in a fine. The appointment obligation in Germany shall apply regardless of the official location of the company, to ensure effective enforcement of the law.

5. **The right to demand the disclosure of documents and to review them** - Anyone whose rights have been violated because of a criminal offense recognized by law will be able to demand that the social network disclose the details of the perpetrator. This right is based on the general principles of the German civil law.

On February 19, 2020, the German government decided to promote the bill to combat right-wing extremism and hate speech, which includes the obligation of social network operators to submit illegal content to the Federal Criminal Police (Bundeskriminalamt, "BKA"). In addition, as part of amendments to the German Criminal Code, the legislature included "antisemitic motives" in the list of issues that constitute a consideration for severe punishment by the criminal court, when deciding on the sentence of a defendant (section 46 (2), German Criminal Code).

In addition, on April 1, 2020, the German government decided to promote the amendment to the Network Enforcement Act, which will strengthen the rights of users of social networks and increase their transparency. The change includes a number of issues: the user's right to request further review of a decision to delete his/her (the applicant's) post or a decision not to delete another user's post for which a complaint was filed (the user has the right to receive from the networks all post history, even if deleted); simplifying the method of reporting illegal content in terms of user experience; granting permission to the court handling the request for disclosure of documents, to demand the disclosure of the identity of the publisher of the post; imposing an obligation on companies to publish the technological means they use to monitor content and explain any significant changes occurred since the publication of the previous report. Both legislative proposals went into effect in June 2020.

The NetzDG has drawn considerable criticism regarding the violation of freedom of expression and the restriction of the freedom of expression of social media users. It is important to emphasize that the NetzDG has not created new categories of illegal content, and that its purpose is to enforce 22 existing laws in the German Criminal Code for online content, as well as to deal with them efficiently, consistently and promptly in accordance with the guidelines of social media companies.

United Kingdom

Over the past two years, the UK has been working on regulating the issue of social network providers. However, no binding legislation has been passed so far. On February 12, 2020, the government published the "Online Harms White Paper".[68] The proposal presents a plan for a new accountability and oversight system for tech companies, moving far beyond self-regulation. It is a plan that will set a lower threshold for content with "potential to cause harm", with child sexual exploitation content and terrorist content being defined as extremely serious. The document represents the government's policy in this matter.

[68] https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper

The regulatory framework will require companies to explicitly state what behavior and content are acceptable on their sites, and to enforce this consistently and publicly. All companies in scope will produce an extremely high level of protection for children online and will take reasonable steps to protect them from inappropriate or harmful content.

The main features of the proposal are:

1. **Statutory duty of care and appointment of a regulator** - The government will establish a new statutory duty of care that will result in companies taking broad responsibility for user safety and for dealing with damages caused by content or activities on their sites. An independent regulator will oversee compliance with this duty of care even if there are broad powers such as imposing heavy fines and imposing criminal liability on senior management members. In February 2020, the UK government announced its intention to appoint "Ofcom", the UK's communications regulator, as regulator for this field.

2. **Government direction on content** - Reflecting the threat to UK national security and the desire to ensure the physical security of children, the government will have the power to direct the regulator in relation to codes of practice on terrorist activity or child sexual exploitation and abuse.

3. **Annual reports** - The regulator has the power to require annual transparency reports from companies in scope, that will be published online by the regulator, outlining the prevalence of harmful content on their platforms and what measures they are taking to address this. In addition, it will also have powers to require additional information, including about the influence of algorithms in selecting content for users.

4. **Applicability of the law** - The draft legislation stipulates that the regulatory framework will apply to companies that allow users to share or discover other users' content, as well as communicate with each other online. While these services are offered by a very wide range of companies of all sizes, the regulator

has the power to determine against which platforms to act and in what way, when this activity is based, among other things, on the scale of the platform, content publication risk level, and frequency of infringing publications. Any requirement to review or monitor defined categories of illegal content will not apply to "private channels".

Organizations in this sector welcomed the legislative initiative, which included consulting with a variety of stakeholders and the selection of “Ofcom” as a regulator. However, it was criticized for lacking a recommendation for an incentive for platforms to build a responsible and advanced technological infrastructure and design, which would give the regulator more capabilities in the field of investigation and assessment of the situation.

France

On May 13, 2020, a law called the "Avia law" was passed in France, with the aim of combatting hate speech. The law was inspired by the German law described above.

Dealing with this forms of discourse through an immediate mechanism of removing content, along with fines for non-compliance with the provisions of the law, amounting to up to 4% of the annual turnover of the infringing company. Most of the rules would be applied to online platforms and search engines that reach a certain threshold of activity (to be determined later) in France, regardless of the location of those companies or their headquarters.

In addition, online platforms in scope will be required to combat inciting content, by complying with the outline of the CSA (The French Audiovisual Council Administrative Body)[69]. Social networks and other websites will be required to remove content considered by French authorities as content related to terrorism or to child abuse within **one hour** of the report. Any other offensive content will be removed within **24**

[69] https://www.taylorwessing.com/en/insights-and-events/insights/2020/05/new-law-to-fight-online-hate-speech-in-france

hours of the report. These rules will apply to all sites, large and small. For this purpose, "other" offensive content includes content that incites to hatred, violence, racism or sexual harassment.[70]

That being said, in June 2020, the French Constitutional Court ruled that the law, which in fact established a responsibility to analyze content on digital platforms, without the intervention of a judge, within a very short time, and even allowed imposing hefty fines, created an incentive for risk-averse platforms to indiscriminately remove flagged content, whether or not it was clearly hate speech.

The court also ruled against parts of the law requiring platforms to remove content flagged as related to child pornography or terrorism. Today, only a limited portion of the original form of the law remained.[71]

Canada

There exists a procedure under criminal law - under section 320.1 of the Penal Code under which a judge may issue an order - if he is convinced on the basis of "information on oath" that any content constitutes "hate propaganda" as defined in section 320(8)[72] or computer data that enables propaganda and are stored and open to the public by a computer system within the jurisdiction of a court, a judge may order the person in charge of the computer system: (1) to submit an electronic copy to the court; (2) to ensure that the content is not stored and is open to the public through the same computer system; (3) to provide the information required to identify and find the person responsible for the published content.

Belgium

The main agency working against hate speech is UNIA[73], which is an independent authority, funded by state budget. Anyone who feels he has been the victim of hate

[70] https://www.bbc.com/news/technology-52664609
[71] https://www.nytimes.com/2020/06/18/world/europe/france-internet-hate-speech-regulation.html
[72] "**Hate Propaganda** means any writing, sign or visible representation that advocates or promotes genocide or the communication of which by any person would constitute an offence under section 319"
[73] https://www.unia.be/en/areas-of-action/media-and-internet/internet/the-limits-of-free-speech

speech can report the abusive content to UNIA. UNIA's initial tendency will be to prefer freedom of expression and therefore, the main course of action is in creating a dialogue and striving to reach a compromise between the complainant and the content publisher and the other parties involved. In exceptional cases where dialogue is not possible, or when dealing with particularly serious offenses or when judicial clarification is required, UNIA may take legal action.[74] No individual reference was found for the activity vis-a-vis the social networks themselves.

Italy

Italy has extensive legislation on hate speech, but most of it is focused on broadcast and offline media. One of the main laws is the "Gasparri Law" which on the one hand enshrines freedom of expression, and on the other hand emphasizes the need for civility and prohibits, for example, the broadcasting of programs that contain an incentive for hatred[75], and there is even an entity that can prevent, under certain and strict conditions, the distribution of content from being broadcasted. Unfortunately, this law focuses on broadcast content and not online written content. In 2018, Italy issued a summarizing document on the subject.[76]

Bulgaria

The prohibition of hate speech is enshrined in the Bulgarian constitution. On the one hand, freedom of expression is preserved, but on the other hand, this right is restricted by way of negation. Thus, in paragraph 2, Article 39 (Article 39 (2) of the Constitution of the Republic of Bulgaria), it is written that the right shall not be used to defame the reputation of others, or to incentivize a change in the legal order in the country, preparation for crime or incitement to hostility or violence against anyone. As a rule, hate speech appears as a criminal offense in the criminal code of the state.

[74] https://www.unia.be/en/areas-of-action/media-and-internet/internet/the-limits-of-free-speech#7.-What-can-Unia-do
[75] https://www.camera.it/parlam/leggi/04112l.htm
[76] https://www.article19.org/wp-content/uploads/2018/04/Italy-Responding-to-%E2%80%98hate-speech%E2%80%99_3.4.pdf

APPENDIX D

THE ISRAEL DEMOCRACY INSTITUTE:

REDUCING ONLINE HATE SPEECH - RECOMMENDATIONS FOR SOCIAL MEDIA COMPANIES AND INTERNET INTERMEDIARIES

This publication (which was sent to print in the autumn of 2019) contains the results of a joint research project undertaken by the Israel Democracy Institute (IDI) and Yad Vashem, with the goal of supporting efforts by social media companies and other internet intermediaries to formulate policy and policy guidelines aimed at reducing online hate speech. Although hate speech is certainly not a new phenomenon, digital platforms facilitate its promulgation and dissemination today at unprecedented speed and scale, and this requires a more proactive response to its harmful consequences. The utilization of private platforms for spreading hate in digital space also poses unique governance challenges and demands new approaches to content regulation and institutional oversight.

As a nonpartisan Israeli think tank, the IDI has a longstanding interest in the possibilities and challenges that new technologies pose to traditional democratic values, processes, and institutions. It sees the digital space as a crucial asset for democratic life in the twenty-first century, but one that must be protected against abuse. Yad Vashem, too, dedicated to perpetuating the memory of the Holocaust and the lessons learned from that dark chapter in modern history, views the digital space as an important educational arena and tool. It is concerned, however, about the malicious exploitation of this platform to spread hateful propaganda, including anti-Semitic Holocaust denial. It was against this background that these two Israel-based institutions joined forces with international partners to devise and carry out a research program to address the problem of online hate speech from a broad and nonlocal perspective. It is clear that, as a matter of principle, the ways of dealing with anti-Semitic hate speech should not be developed separately from ways of countering other vile

and potentially harmful forms of hate speech that promote Islamophobia, homophobia, hatred of migrants, and the like. Only on the basis of broadly accepted norms and processes that identify prohibited hate speech and restrict it can specific modalities be devised to deal with specific forms of hate speech or to protect specific groups of potential victims.

The present publication has two sections. The first of them consists of the *IDI–Yad Vashem Recommendations for Reducing Online Hate Speech*: sixteen recommendations meant to serve as the basis of policy guidelines for social media companies and other internet intermediaries. These recommendations derive from the research papers presented in the second section, *as well as from* discussions undertaken in the three workshops, held in Jerusalem, Geneva, and Irvine, California, as part of the research project, and consultations among the project researchers and steering committee. The studies by members of the research team—Dr. Tehilla Shwartz-Altshuler (IDI) and Mr. Rotem Medzini (IDI), Prof. Karen Eltis (University of Ottawa) and Dr. Ilia Siatitsa (Geneva Academy of Human Rights and International Humanitarian Law), and Prof. Susan Benesch (Berkman Klein Center, Harvard University)—analyze online platforms' current policies and the legal frameworks in which they operate, and propose avenues for future reforms. We *hope* that the recommendations and the research papers they are based on will inform contemporary debates on the ways in which social media companies and other internet intermediaries regulate online speech and will influence the positions of the stakeholders who participated in such debates—states, international organizations, academia, civil society, the technology sector, the media, and the public at large.

I would like to thank the administrative and policy teams at the IDI and Yad Vashem that facilitated the organization and operation of the research project, and especially Ms. Shirli Ben-Tolila (IDI), Mr. Arnon Meir (IDI), Ms. **Iris** Rosenberg (Yad Vashem) and Dr. Robert Rozett (Yad Vashem). Thanks are also due to Mr. Dvir Kahana, the director general of the Israel Ministry of Diaspora Affairs, and Mr. Yogev Karasenty, a senior policy officer in that ministry,

for their ongoing engagement with the research project and their keen interest in its findings.

Prof. Yuval Shany

Project Coordinator

Jerusalem, 2019

A Proposed Basis for Policy Guidelines for Social Media Companies and Other Internet Intermediaries

Introduction

The recommendations presented below are the product of a yearlong study conducted by an international team of researchers,[77] with guidance from an international steering committee of experts[78] convened by the Israel Democracy Institute (IDI) and Yad Vashem. The process included workshops in Jerusalem (hosted by the IDI), Geneva (hosted by the Geneva Academy for International Humanitarian Law and Human Rights), and Irvine (hosted by the Center on Globalization, Law and Society of the University of California at Irvine), and the writing of three detailed research papers that offer multiple policy recommendations. Throughout the study, *consultations were held* by the research team with academics, policy researchers,

[77] The research papers that form the basis for the recommendations were written by Dr. Tehilla Shwartz-Altshuler and Mr. Rotem Medzini, by Prof. Karen Eltis and Dr. Ilia Siatitsa, and by Prof. Susan Benesch.

[78] The steering committee comprised the following experts: Prof. Tendayi Achiume (the UN Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance), Prof. Sarah Cleveland (former vice-chair of the UN Human Rights Committee), Prof. Irwin Cotler (former Minister of Justice, Canada), Prof. David Kaye (the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), Prof. Avner Shalev (chair of the Yad Vashem Directorate), Prof. Yuval Shany (former chair of the UN Human Rights Committee), and Prof. Jacques de Werra (Vice-Rector, University of Geneva).

government officials, human rights activists, industry policy officers, technology experts, and others.

The sixteen recommendations that emerged from the study and the research papers are meant to provide social media companies and other internet intermediaries with a basis for policy guidelines and benchmarks and with directions for future action aimed at reducing hate speech and protecting the fundamental human rights that find themselves under assault by such speech, while ensuring freedom of expression (including the protection of speech that may offend, shock, or disturb the public) and other relevant human rights. They also provide other stakeholders that are troubled by online hate speech, including civil society, the public at large, and institutions invested with special responsibilities in this regard (e.g., elected governments and independent judiciaries), with tools to evaluate company policies and rules on hate speech and the manner of their application.

Recommendation 1: The Responsibility to Reduce Online Hate Speech

Social media companies and other internet intermediaries have a legal and ethical responsibility to take effective measures to reduce the dissemination of prohibited hate speech on their digital platforms and to address its consequences. This includes, where appropriate, content moderation (see Recommendation 6) and the recognition and condemnation of such speech. Measures such as content moderation have a critical relationship to basic human rights, including freedom of expression, the right to equal participation in political life, the right to personal security, and freedom from discrimination. Pursuant to internationally accepted legal standards and definitions, company policies and rules on prohibited hate speech must be transparent, be open to independent review, and offer accessible remedies for violations of the applicable norms. The responsibility of social media companies and other internet intermediaries does not release other actors, including online users, group and page administrators and moderators, private and public associations, states, and international organizations from their responsibility under domestic and international law to take effective measures to reduce online hate speech and their

liability for the harm caused or facilitated by prohibited hate speech.

Recommendation 2: The Application of Relevant Legal Standards

Policies and rules aimed at reducing hate speech should conform to international human rights standards, as found in the International Covenant on Civil and Political Rights (especially articles 19 and 20) and in other international instruments, such as the Convention on the Elimination of Racial Discrimination (especially articles 4 and 5(d)(viii)), the European Convention on Human Rights, and other regional human rights conventions. They should conform to national laws, provided that such laws are compatible with international standards. The policies and rules should also be informed by broadly supported international instruments, such as the Rabat Plan of Action with the six potential indicators of criminal hate speech it identifies (context, speaker, content and form, extent and reach of the speech act, and likelihood, including imminence) and the Working Definition of Antisemitism adopted by the International Holocaust Remembrance Alliance.

Recommendation 3: The Harm Principle

In the determination of whether certain speech should or should not be considered prohibited hate speech subject to content moderation policies and rules, particular attention should be given to the need to effectively prevent harm to groups and individuals, including physical and psychological harm, reputational harm, and affront to their dignity, and to an evaluation of whether such harm is likely to result from the speech, given the speaker's overall tone and intention, the methods and means of its dissemination, and the status of the persons targeted by the speech and/or of the protected group to which they belong, including patterns of tension and discrimination and violence against targeted protected groups, such as antisemitism, Islamophobia, and xenophobia. When denial of clearly established historical facts about the most serious international crimes, such as the Holocaust and other past genocides, is intended and expected to re-victimize victims and their descendants, it should be considered a harmful form

of speech.

Recommendation 4: Detailed Policy Guidelines

Social media companies and other internet intermediaries should clearly define and publish detailed policy guidelines on prohibited hate speech and permitted speech, anchored in the applicable human rights standards. They should explain how they apply their policies and rules, and especially how context—including social, cultural, and political diversity, the use of code words and euphemisms, criticism of hate speech and humor, and the reclamation of offensive slurs by targeted groups—is taken into account in decisions about content moderation. The detailed definitions of hate speech used by social media companies and other internet intermediaries should be formulated after consultation with outside experts who are familiar with the relevant national and international legal standards on hate speech, as well as with experts in other relevant fields, such as education, sociology, psychology, and technology.

Recommendation 5: Preventive Measures

Social media companies and other internet intermediaries should adopt proactive policies that are consistent with international human rights standards and that are designed to prevent the dissemination of *prohibited* hate speech before it causes different *forms* of harm. *They should harness* reliable algorithms for natural-language processing and reliable sentiment-analysis tools, whose decisions are subject to meaningful human review and challenge mechanisms, and employ their own internal trained content reviewers, with the aim of improving the identification of hate speech, curtailing the virality of prohibited harmful content, and/or allowing users to apply filters to block offensive content they do not wish to be exposed to. Social media companies and other internet intermediaries should also take steps to render their policies and rules visible and easily accessible to users, presented in a concise, transparent, and intelligible manner and written in clear and plain language, including examples of permissible and impermissible content. With the goal of discouraging users from resorting to hate speech, these proactive steps should be designed to foster

understanding of the relevant policies and rules and employ culturally sensitive awareness-raising measures, which might include explaining how certain expressions or images might be perceived by affected individuals or groups.

Recommendation 6: A Diversity of Content-Moderation Techniques

To enforce hate-speech policies and rules, social media companies and other internet intermediaries should develop an array of content-moderation techniques that go beyond simply deleting content and blocking accounts. Such techniques should include nuanced measures that are adjusted to different degrees of deviation from the policies or rules, the source of the complaint about a violation (e.g., an AI-based algorithm, law-enforcement agency, trusted community partner, other online user), and the identity of the speech-generating user (private individual, news agency, educational institution, repeat offender, etc.). These fine-tuned measures could include the flagging of content, the attachment of countervailing materials to potentially harmful content, a warning to disseminators of the consequences of violations, a request to disseminators to self-moderate or remove harmful content, and the unilateral imposition by the platform of temporary limits on dissemination. Special strategies need to be put in place to address chronic and particularly serious violations of hate-speech policies and rules, including the permanent blocking of repeat violators, the *dismantling* of business models which deliberately use online platforms to facilitate prohibited harmful activities, and notifications to law-enforcement agencies of serious violations that might merit attention by criminal justice authorities.

Recommendation 7: Flagging Mechanisms

Social media companies and other internet intermediaries should institute mechanisms that allow for a quick and effective response to the flagging of prohibited hate speech by algorithms or internal content reviewers, and for soliciting external notifications from community partners (such as law-enforcement agencies, civil society groups, and other

users) and responding to them quickly and effectively. These should include the introduction of conspicuously placed standard user interfaces and national contact points for notifications. Companies and intermediaries should also rely on information from trusted community partners in order to introduce temporary content-moderation measures, such as measures to curtail virality.

Recommendation 8: Notification of Complaints and Decisions

In order to facilitate quick and effective oversight at all stages of decision-making about content moderation, complainants must be sent immediate acknowledgement that their notification about prohibited hate-speech content has been received. Subsequent decisions about content moderation must be conveyed to them with an explanation of the reasons for the decision, including reference to any anticipated harm or lack thereof, and information on possibilities of challenge or appeal. Decisions to moderate content and the reasons for the decision must also be communicated to the user that published the speech deemed hateful.

Recommendation 9: Ordinary Mechanisms for Challenging Decisions

Social media companies and other internet intermediaries should develop effective and accessible mechanisms for challenging their specific decisions to moderate or not moderate speech alleged to be hateful, and for quickly and effectively resolving such challenges. Procedures for reviewing challenges to specific decisions should be introduced at the platform level, including an internal process for rapid reconsideration of specific decisions on content moderation, as well as access to a private alternative dispute resolution (ADR) process or litigation, when appropriate, for dealing with disputes about final decisions on content moderation which are not resolved internally.

Recommendation 10: Mechanisms for Examining ‘Hard

Cases'

Procedures should be developed for consulting with legal advisors or advisory bodies about specific decisions or the application of general policies or rules to a specific situation. "Hard cases" – cases where it is not readily apparent to company personnel responsible for content-moderation decisions whether the speech in question conforms to or violates applicable policies and rules – should be promptly examined by independent experts. In addition, governments should ensure that content-moderation decisions that infringe the freedom of expression and other basic rights of individuals under their jurisdiction are subject to review by independent courts.

Recommendation 11: Protection of Content Moderators

Social media companies and other internet intermediaries should establish effective programs for training content moderators, with human rights education and cultural sensitization relevant to the content they review, including the considerations set forth in Recommendation 3.[79] They should also take adequate measures to mitigate trauma and other adverse consequences of excessive and prolonged exposure to hate speech, including setting limits on the working hours of content reviewers and providing them with counseling and other forms of psychological support.

Recommendation 12: Advisory Councils

Social media companies and other internet intermediaries should establish advisory councils to periodically evaluate their content moderation policies and rules and the manner in which they monitor and enforce these policies and rules, including the practice of designating cases as "hard cases," challenge procedures, and transparency policies. Such advisory councils should be composed predominantly of independent experts familiar with the applicable international standards, content-moderation technology, education policy, and relevant

[79] See, e.g., the following MOOCs: Yad Vashem Online Course on Antisemitism, November 11, 2018; Le racisme et l'antisémitisme (FUN).

political, cultural, and other contexts. Where appropriate, advisory councils should be established not only at the international level, but also at the national (or regional) level, so they can evaluate and suggest ways to adapt general policies and rules to local norms and cultural contexts without violating international human rights standards. To ensure transparency and accountability, the procedures and criteria for selecting the members of advisory councils, including safeguards against conflicts of interest, should be made public.

Recommendation 13: Exchange of Information and Best Practices among Companies

Social media companies and other internet intermediaries should consider establishing procedures (including the formation of joint advisory councils) for exchanging information about their content-moderation policies, rules, training methods, and challenge mechanisms, with a view to coordinating and, where appropriate, aligning their key elements to best industry practices. They should also consider creating a common digital database of hashtags, images, phrases, and code words associated with prohibited hate speech in different social, political, and cultural contexts and, subject to privacy constraints, sharing information about repeat violators of their hate-speech policies.

Recommendation 14: A Global Stakeholders Forum

A global stakeholders forum, with representatives of governments, social media companies and other internet intermediaries, experts in technology, law, and education, and civil society groups, should be created and convened from time to time in order to discuss, develop, and evaluate the application of international standards and procedures for reducing online hate speech.[80]

Recommendation 15: Transparency

[80] The Global Network Initiative and the International Holocaust Remembrance Alliance are possible models for such a global coalition.

Social media companies and other internet intermediaries should publish regular detailed reports on the application of their hate-speech policies and rules, including country-specific information about specific content modifications, whether at the request of law-enforcement agencies or at their own initiative; information about external notifications, about challenges to specific content-moderation decisions and their outcome, and about the training of content moderators; efforts to raise users' awareness of partnerships with civil society organizations; and other proactive measures. Reports on content-modification activities should be sufficiently detailed to allow external assessment of these practices' compliance with international human rights standards. In addition, information about the scale of public exposure to harmful content prior to content moderation by the platform should be made available to the public.

Recommendation 16: Criteria for Evaluation of Policies and Rules

Advisory councils, civil society organizations, the media, and other observers may find it useful to evaluate and compare the policies and rules for hate-speech content moderation applied by different social media companies and other internet intermediaries, so as to encourage identification of best practices, to allow users to make more informed choices between different legitimate policies, and to enable users to assess whether they adequately balance the need to address hate speech with respect for freedom of expression and other individual rights. The evaluation of hate-speech policies and rules could take the following factors into consideration:

(**1**) The definition of protected groups: Does it cover collectives other than racial, ethnic, and religious groups, such as those defined on the basis of their sex, sexual orientation, or gender identity, or on the basis of disability, and voluntary membership groups (e.g., political parties or professional associations)? Does the definition address situations of intersectional discrimination?

(**2**) The extent to which the classification of hate speech as such (a) is based on a closed list

of banned words, phrases, symbols or images; (b) makes it possible to identify complex connections among language, images, and ideas that may render speech hateful in certain cultural, social, or political settings; and (c) considers the broader context that may legitimize (e.g., satire) or delegitimize the speech (e.g., bogus historical research in the service of racist causes);

(**3**) Is the element of causation incorporated in the definition of hate speech linked only to the expectation that it might lead to physical harm to the targeted persons? Or does it also consider nonphysical damage to potential victims, such as fear or feelings of marginalization, as well as indirect harm such as discrimination as a result of negative stereotypes and social attitudes against the protected group?

(**4**) Are broader socially undesirable impacts on the audience of the speech factored into content-moderation decisions – ranging from likelihood of violence to other breaches of the peace (e.g., possible social unrest) and to nonphysical long-term results, such as the fostering of a climate of growing hate and racism in society?

(**5**) Are content-moderation decisions based only on the speakers' explicit intent, or also on their implicit intent, or regardless of their intent?

(**6**) Are applicable content-moderation tools applied to speech disseminated on public platforms only, or also that intended for closed groups and sent as private messages?

(**7**) Does the response to a violation of hate-speech policies and rules entail only limiting its virality? Or are there other measures, such as a request that users remove or self-moderate the content they posted, unilateral content removal, or temporary or permanent blocking of the account?

It is recommended that companies conduct a periodic self-evaluation of their policies in light of these criteria and publish the results of the evaluation.