Yale MACMILLAN CENTER
Genocide Studies Program

Mass Atrocities in the Digital Era Initiative (MADE) *Working Paper No. 4*

February 2023

Addressing the security risks of anti-Roma hate speech on social media platforms

Pavlina Pavlova

This paper examines online hate speech and the associated security risks by focusing on user-generated content (UGC) targeting Roma and related moderation standards, tools, processes, and practices. The Romani people have experienced systemic racism, discrimination, and hostility across countries. These negative attitudes are perpetuated, broadcast, and intensified in the online space in the form of hateful and racist speech or incitement to violence and genocide. The paper illustrates cases of anti-Roma narratives and their translation into the online realm while mapping their harmful impacts on the Romani communities and individuals. These observations are instrumentalized to examine the challenges and tensions that platforms encounter in moderating online content. It is proposed that AI-based detection tools are integral to tackling hate speech, but due to the highly contextual nature of hateful content and its differentiated risks, they are unfit for being the exclusive means for decision-making. Content moderation models can be effective when implemented holistically with each layer extending the security and compensating for the limitations of the other. Social media providers must follow a victim-sensitive approach to tackle the asymmetric threats that hate speech presents to minority, marginalized, and other vulnerable groups.

Keywords: Roma, hate speech, human security, content moderation, algorithmic models

Author
Pavlina Pavlova

Affiliation
CyberPeace Institute; formerly the Organization for Security and Co-operation in Europe (OSCE), Office for Democratic Institutions and Human Rights (ODIHR), Contact Point for Roma and Sinti Issues (CPRSI)

Position
Public Policy Advisor

Contact details
E-mail: pavl.pavlova@gmail.com
Phone: +421 919 027 059

TABLE OF CONTENTS

Introduction **1**

1. Anti-Roma hate speech online: context, form, and narratives **2**

2. The security risks of anti-Roma hate speech online for targeted individuals and communities **4**

a. The harmful impacts of online hate speech on victims and society 4
b. From online hate speech to offline hate crime 5

3. Challenges and tensions of content moderation models countering online hate speech **7**

a. Rules and guidelines addressing online hate speech 7
b. Artificial intelligence (AI) models for detecting and removing hate speech content 8
c. Human moderators, users reporting, and de-ranking content 10

4. Recommendations **11**
5. Conclusion **12**

References **13**

Introduction

The Internet has changed the ways we communicate, organize, and exchange information. The online space offers accessible platforms for self-expression, providing means to access and share ideas, engage, and mobilize. But as people moved to social media a series of problems arose there, too. While there have been broad discussions around the impacts of online platforms on democracy, especially on elections interference and free speech (Bradshaw & Howard 2020; Sander, 2021; Douek 2020; Persily & Tucker, 2020), their implications for certain minority groups and marginalized communities—and the associated security risks—have yet to be investigated within the framework of platform governance and content moderation. This knowledge gap hinders effective responses to online hate speech. As a result, hateful and racist expressions and incitement to hatred, violence, or even genocide against the minority, marginalized, and other vulnerable groups continue to be largely unaddressed.

Hate takes many forms. It can be masked as rational opinion or justified expression or demonstrate and openly dehumanizing or inciting language. There is no international legal definition of hate speech and the classification of what precisely constitutes hateful expressions remains disputed (UN, 2019; Banks, 2011; Citron, 2009; Heinze, 2016; Herz & Molnar, 2012). This is due to the complexity and versatility of hate speech, the conceptualization of which can be conducted from the standpoint of legal, linguistic, psychological, sociological, security, and many other approaches. Consequently, hate speech refers to a heterogeneous set of manifestations, ranging from criminalized speech to speech that is offensive and disturbing but not unlawful (Gagliardone et at., 2014). Acting on hate speech can be raised as finding an uneasy balance between two fundamental rights in conflict—freedom of speech that protects opinions stated in public and the prohibition of discrimination that protects persons and communities from harmful, humiliating, or depreciating behaviors in relation to their protected characteristics. In content moderation, social media providers calibrate conflicting demands and set the red lines on which statements need to be limited or removed. However, user-generated content (UGC) that remains present on the platforms despite violating their rules and guidelines reveals a problematic and piecemeal operationalization.

This paper illustrates how decisions on online content can impact the security of Roma in Europe and why tackling anti-Roma hate speech on social media platforms can inform content moderation globally. Moderating online content is understood as the organized practice of screening UGC posted on social media to determine the appropriateness of the content for a particular platform, locality, context, and jurisdiction (Roberts, 2017). “Roma” is applied as an umbrella term for various groups, some of which use the self-designation Roma while others use different ethnonyms—self-designations as well as external designations, including Arlije, Calé, Gurbet, Kaale, Kalderaš, Lovara, Manuš, Sepečides, Sinti, and Ursari (CoE). Roma have been selected as a minority group that finds itself on the sharp end of content moderation decisions. This is due to the vulnerability, marginalization, and pervasive discrimination that Romani communities historically experienced. The outlined security risks related to online hate speech employ the concept of human security developed in the UNDP Human Development Report (1994). The UNDP report gives a broad definition of human security, describing it as the condition of safety from several categories of threats, including economic, food, health, environmental, personal, community, and political security, and comprising the two elements of freedom from fear and freedom from want. This approach to human security is further narrowed down for analytical purposes to freedom from fear related mainly to threats and physical harm (Liotta and Owen, 2006). However, it is important to consider the broad and systemic impact that hate speech and both explicit and latent racism pose for the victims and society (UNTFHS, 2009).

The research is positioned within the wider field of critical discourse studies. It aims to analyze anti-Roma hate speech narratives on social media platforms and related security implications for the targeted

communities using discourse analysis. The identified narratives are supported by case studies collected from social media platforms using the method of qualitative content analysis. To avoid any posts that would be identifiable and linked to their authors, UGC is outlined but not cited, unless the authors are public figures. The analysis focuses on UGC, mainly posts and comments, present on major social media platforms. The case studies have been selected to illustrate the common issues and overarching tendencies in anti-Roma hate speech online. The identified narratives are placed into a broader framework of challenges and tensions that platforms encounter when moderating online content. This case-based research design is instrumental for outlining the impact of content moderation decisions on the security of marginalized groups; however, robust qualitative and quantitative research is needed to further test the conclusions. The recommendations propose that applying available resources in layers while focusing on vulnerable communities and high-risk situations can provide a safer online environment. Models for content moderation should employ a victim-sensitive approach to streamline and improve responses to the security risks that hate speech presents for vulnerable communities.

Section 1 investigates anti-Roma narratives, their context, and form and includes illustrative UGC examples. Section 2 highlights the direct and indirect security impacts and harm to the affected people. It provides examples of triggering events and high-risk environments with the potential for escalations that must be prioritized in content moderation decisions in order to create a safer online ecosystem. Section 3 analyses content moderation standards, tools, processes, and practices, and proposes steps for their improvement. Section 4 presents recommendations based on the prior analysis and argues for stronger safeguards to protect groups that are disproportionally affected.

1. Anti-Roma hate speech online: context, form, and narratives

The Roma are Europe's largest ethnic minority with about 10-12 million people living on the continent (European Commission). Despite their long European history, Roma communities have been traditionally perceived as not belonging or participating in the majority society—supporting the notion of Europe's "quintessential minority" that faces prejudices, exclusion, and discrimination across countries. The level of hostility differs depending on the region, with stronger negative views present in areas with large Romani communities (Dizdarevič, 2020; Kyslova, et al., 2020; Wike et al., 2019; Hamelmann et al., 2017; Warmisham, 2016). This correlation is important for addressing the real-life harm of online hate speech as physical attacks on Romani men and women are more likely to take place in countries such as Bulgaria, Slovakia, Romania, and Hungary, where, in terms of their share of the total population, most Roma reside (European Commission, 2021). Recent years have seen a rising wave of anti-Roma attitudes in the context of the increasing prominence of far-right groups and growing xenophobic discourse (Vojtová, 2020). The COVID-19 pandemic further reinforced this trend. Human rights organizations (ODIHR, 2020b; OHCHR, 2020a) and national equality bodies (Hale, 2020) reported an alarming rise of anti-Roma sentiment in several countries, where Roma people were alleged of not complying with the quarantine requirements, spreading, or even causing coronavirus infections (ODIHR, 2020a). These unfavorable views are perpetuated, broadcast, and intensified online as hateful or racist statements, accompanied by acts of racism, intolerance, and hate crimes (OSCE, 2021).

Hate speech spans from individual assertions to organized and political statements, which bring hateful manifestations to the online mainstream (Hamelmann et al., 2017). Anti-Roma statements in the online environment consist of similar elements as offline hate—from stigmatization by ethnicity, reinforcing negative stereotypes, supporting prejudices and misconceptions, and circulating hoaxes to using insulting and derogatory statements, emphasizing ostracization of marginalized groups, escalating tensions or even inciting hatred, violence, and genocide (Kyslova, et al., 2020). Hate speech takes different forms depending on its

historical, socio-economic, political, and linguistic backgrounds. The most pervasive narratives can be clustered in several groups, such as criminalization, welfare chauvinism, and dehumanization (Hamelmann et al., 2018). Using the image of what Kyuchokov (2015) calls the "conceptual gypsy" and creating a shared mainstream identity serves the goal of justifying the unequal treatment of Romani men and women. They come from and lead to the exclusion of Roma from society, often mixed with explicit incitements to violence and calls for their removal from a particular country (Dizdarevič, 2020). These online manifestations of hate against Roma illustrate just how alive is the underlying notion of "gypsy threat narrative" in Europe (Hamelmann et al., 2017).

Common depictions assign Roma criminal and thievish attributes (Vojtová, 2020). They are presented as violators of rules and offenders. A crime from a single person is seen as the responsibility of the whole community. The trope of Roma as "dirty", "beggars", "thugs" and "thieves" and a criminalized minority is persistent (Hamelmann et al., 2017). They are equated to liars and alleged of making their living by fraud. Statements such as *"What inadaptable?! On the contrary, they adapt quickly and know well where and how to get what they want"* have been identified as examples of UGC displaying this narrative. Stereotypes about Roma present them as uneducated and "uncivilized" people, whose members are undeserving to stay in the respective country. This narrative is often reinforced by references to their foreign origin (*"Egyptians from a ghetto"*) (Kyslova et al., 2020; Siapera et al., 2018).

Welfare chauvinism is central to anti-Roma narratives. The Romani people are characterized as parasites and swindlers who take advantage of social benefits (*"for 60 years, gypsies have only advantages and not one of them works"*). The alleged abuse of the welfare system is coupled with racializing such social phenomena as nomadism, begging, vagrancy, unemployment, or a low level of education that are treated as traditional and inherent to Romani culture (*"they will raise children according to themselves, so they will teach them to steal"*) (Hamelmann et al., 2017; Warmisham, 2016). Roma are presented as "non-integratable" despite the public resources spent on "their cause" and not participating in the majority society. The constructed inability to socially adapt as a whole community is reflected in the term "inadaptables" (*"the expression 'inadaptables' still doesn't fit. Why not tell the truth. It is a national minority that has lived in nature for thousands of years"*) (Vojtová, 2020). Roma women are exposed to multiple and intersecting discrimination—compounding hate through misogyny and anti-Roma attitudes (Dizdarevič, 2020). Among others, they face a recurring narrative of having high birth rates to obtain welfare benefits and allowances (*"10 children, not to touch work in life"*), which is also commonly used against other minority groups (Warmisham, 2016).

The concept of "double standards" is regularly reflected in hoaxes, misinformation, and disinformation materials. (Dizdarevič, 2020). Such posts and comments allege that Roma receive preferential treatment from the authorities (Vojtová, 2020). For example, in 2019, a widely circulated and commented piece of disinformation claimed that Roma were given firewood for free during the winter (Snidl, 2019). In the spring 2020, the false information that Roma in forcibly quarantined settlements during the COVID-19 pandemic received free alcohol and food from the state was rapidly shared on social media accompanied by hateful and racist reactions (Web Noviny, 2020). Fake and de-contextualized news that is prominently recycled online can result in collective punishment being meted out indiscriminately against Roma as a whole (ERRC, 2019).

Roma people are subjected to dehumanizing language—being compared to "trash", "animals", "apes", "vermin", "pests" and "parasites" that take advantage of the "decent majority". Framed as "second-class" or downright "inferior citizens" and perceived as a "threat to order and a burden on society", Roma continue to experience the image that has been historically used to legitimize oppression, deportation, and extermination (*"such monkeys should be killed"*) (Vojtová, 2020; Kyslova et al., 2020; Siapera et al., 2018).

Dehumanization as a process played a key role in the Roma genocide during the Second World War by denying humanity to groups of people and constructing an image that they are not belonging to the same human or national community as the majority population (“*for 700 years, the gypsy minority has been integrating into society ... that says a lot”*). Such posts claim genetic or other inferiority of Romani people, using mockery (*“Do you know how to stop a gypsy from bleeding? No? That’s good.”*) or supporting Roma Holocaust (Hamelmann et al., 2018). For instance, content on social media comparing Roma with vermin led to calls that the problem should be “treated accordingly” while other slurs were accompanied by calls for their forced sterilization (*“segregation and sterilisation of these primates is the only solution”*) (Enarsson & Lindgren, 2019; Dizdarevič, 2020). The references to the Nazi regime are strongly present in posts explicitly calling for violence against Roma. References to what happened during that period are used as threats, for example, by presenting gas chambers as “a suitable solution for Roma” (*“only gas applies to this”*). Calls that Roma should be “beaten up”, “burned” or put into “concentration camps” are inciting violence by evoking genocide (Hamelmann et al., 2017; Dizdarevič, 2020).

A dangerous tendency is to portray Roma issues as harming the safety and security of the mainstream society, as part of which they are treated as “abnormal citizens”, unable to fit into mainstream society. Consequently, anti-Roma statements become gradually normalized, legitimized, and socially acceptable, instead of recognizing such hate speech as racism (Warmisham, 2016). Framing “Roma issues” as a threat prepares the ground for suggesting that measures taken against them are necessary for public safety. The emergence of what Van Baar (2014) calls “reasonable anti-Roma sentiment” is based on an argument that the majority society is rightfully entitled to act against Roma because they present a threat (*“Gypsies attack us - not us them! They act like arrogant vermin!”; “dead gypsy, good gypsy”*). Such statements can translate into concrete actions undertaken by individuals or hate groups, but also by local authorities and the state (Warmisham, 2016). The cases of police interventions that disproportionately targeted Roma neighborhoods during the COVID-19 pandemic illustrate how the narrative of “dirty and disobeying” Roma re-emerges with renewed vitality in times of crisis (Rorke, 2020; Rorke et al., 2020).

2. The security risks of anti-Roma hate speech online for targeted individuals and communities

a. The harmful impacts of online hate speech on victims and society

Hate speech ranges from individual statements to manufactured and organized actions of those who for political or other reasons profit from increasing intolerance and hatred (Hamelmann et al., 2017). It is spread by users sharing and commenting on posts concerning Roma—clustered around several issues and events such as welfare benefits, criminal actions, clashes, and police interventions. Hate speech can also be found under posts portraying Roma in positive roles. For example, multi-ethnic pupils, including Roma, and teaching staff from an elementary school in the Czech Republic were attacked by hateful comments and calls for killing the schoolchildren after their photo had been posted on an online platform (Malek, 2017). Based on the qualitative content analysis of UGC, anti-Roma hate is often propagated by accounts with affiliations to right-wing extremism and adherents of neo-Nazi ideology, but there is also a growing tendency of public officials to employ openly racist rhetoric. Politicians at the local and national levels equate Roma with poverty, social issues, and crime in their online posts, especially around elections. For example, a Czech politician and SPD Party Leader Tomio Okamura in his post claimed that Roma call for his death because he wants to end the exploitation of their social support, adding *“please support the SPD in the election to end the abuse of benefits by the non-adaptable…”* (Okamura, 2021a). In another post one month later, he said: *“The SPD is fighting against the misuse of benefits by non-adaptables. A decent working citizen is in the 1st place for us!”* (Okamura, 2021b)–essentially pitting Roma against the majority population.

Public figures can contribute to making hateful views socially acceptable and prompting further online hate (Hamelmann et al., 2017; Dizdarevič, 2020). Comments declaring that the police are incompetent in sustaining public order and that the state authorities prevent investigating criminal behavior are particularly dangerous. Such reactions incite people, for instance, to take "matters into their own hands" with the call to "restore the social order". They delegitimize state institutions and create a justification for violent actions. In a reaction to a crime allegedly committed by a Roma man, Slovak right-wing parliamentarian Milan Mazurek's post included the following appeal: *"I strongly recommend all people to get a gun license. The state simply cannot protect you…"* (Benčík, 2020). On another occasion, Italian politician Matteo Salvini posted: *"Stay calm, you dirty gypsy woman, stay calm, the bulldozers will soon be there"* (Salvini, 2019). This type of inflammatory statement has the potential of instigating real-life harm (OSCE, 2003).

Hateful User-generated content has a dual effect. Firstly, its amplifying effect allows for constructing, manufacturing, and popularizing content that is easily accessible online, as individual posts or in the comments section. Secondly, the echo-chamber effect causes the users primarily to access and engage with content that aligns with their views. This can lead to further radicalization by confirming and reinforcing pre-existent bias and even encouraging people to violence (Paladino, 2018). Hate speech online correspondingly sends two types of messages—one aimed at the target group to stereotype, ridicule, diminish or dehumanize its members. The other message seeks to find support for the hateful claims and build a group mentality by making a distinction between who is inside and who stays outside of the desired society (PRISM Project, 2016). Overall, social media platforms can encourage hateful speech by providing accessible means for communication and organizing and simultaneously limiting exposure to contrasting views (Keipi et al., 2016). In this sense, hate speech both divides and unites at the same time (Gagliardone, 2015).

Exposure to hate material is associated with a series of harmful impacts on the victims (Lee & Leets, 2002; Näsi et al., 2014; Foxman et al., 2013). Perpetuated stereotypes contribute to forming a biased picture of the Roma minority and reinforce discrimination which further deteriorates the economic and social status of Roma and results in weakened social cohesion. Distorted and degraded images of a heterogeneous group of people as a negatively perceived homogeneous group affect its members, their self-confidence, and their self-concept. Biased characteristics applied to the entire minority support negative stereotypes and prejudices, lead to psychological and emotional harm to people, further ostracize and intimidate members of the targeted community, and have the potential to increase interethnic tensions. Discourses stereotyping, dehumanizing, and denigrating Roma are pervasive and play a significant role in their exclusion and a chilling effect—limiting the online participation of Roma and normalizing anti-Roma attitudes, which has a negative psychological and psychosocial effect on the communities (Siapera et al., 2018; Quarmby et al., 2020; Dizdarevič, 2020). The display of online hate creates a distinct form of victimization, as abuse is purposely aimed at a collective identity and erodes trust in society and institutions (Näsi et al., 2014; Oksanen et al., 2014). Members of the targeted group are left to attempt to navigate systems that exclude them from the safety to which they are entitled. They are expected to live their lives, access public services, and raise their children in an atmosphere poisoned by hateful and defamatory speech (PRISM Project, 2016). Considering the structural inequalities that Roma encounter and the possible repercussions of online hate speech for their communities, already indirect harm raises ethical and legal questions as to whether the dissemination of hate material should be allowed in society (Waldron, 2012).

b. **From online hate speech to offline hate crime**

A mounting number of attacks on minorities has raised concerns about the link between inflammatory speech online and violent acts offline. As far back as 2006, the OSCE Ministerial Council Decision 4/03 on

Tolerance and Non-discrimination acknowledged that hate crimes can be fueled by racist and other hateful content on the Internet. In 2020, the UN Special Rapporteur on minorities issues accused the propagation of hate speech through social media of contributing directly to the rise of hate crimes against minorities and called for this "poisoning of minds" online to be acknowledged and confronted. *"The more hate speech is widespread, the more it becomes part of the mainstream and creates a permissive and toxic environment where calls for violence against the 'hated' group, usually a minority, become normalized. This propagation of hate against minorities online must be stopped"* (OHCHR, 2020b). Violence attributed to online hate speech has increased worldwide as hate groups use social networks to attract followers, organize, and inspire acts of violence. At its most extreme, hate speech online contributed to violence ranging from lynching to ethnic cleansing, such as the declared role of social media in Myanmar's genocide (Douek, 2018; Laub, 2019).

Online hate speech that coincides with triggering events or emerging conflicts has an increased potential of being translated into offline harm. Such situations include one-off events that touch upon social or interethnic tensions or those explicitly thematizing Roma issues. A high engagement has been triggered by local events, either by certain policy measures, new emerging camps, cases of community violence, police brutality, or in connection to the COVID-19 pandemic and quarantine measures (Siapera et al., 2018; Hamelmann et al., 2017; Miškolci et al., 2020). The climate for hate speech is especially conducive in situations where the political stakes are high, such as during elections, when mounting tensions create conditions for calls to violence (Gagliardone et al., 2015). According to Olha Vesnianka (2021), a Ukrainian media expert and human rights activist, hostility against Romani communities usually begins with a report of an incident involving Roma people. The initial report is followed by negative stereotypes being perpetuated—some reported as observations, others merely speculative. This escalates into hate speech inferring violence and, on some occasions, offline crimes. As outlined above, racially motivated violence against Roma takes place predominantly in countries with large Roma populations. Furthermore, as some Romani communities live in settlements that are spatially segregated, it makes them a vulnerable target for community violence (ERRC, 2020; Tomičić, 2018).

The attacks on Roma temporary settlements in 2018 were preceded by anti-Roma articles in online publications followed by calls for violence against Roma. According to Olha Vesnianka, *"the following paradigm was observed in 2018—the right-wing local group publishes xenophobic statements about the Roma, and shortly afterward comes an offline reaction. This was also the case in Ternopil, Ukraine. There is a growing level of hostility online after such a post, there is support in the comments, the post becomes known and circulated, and then there is an attack. Some information was not verified, and it even seemed that it was created on purpose to justify the attack"*. This notion has been supported by the European Roma Rights Centre (ERRC 2019; ERRC 2018) asserting that the group violence against Romani communities in Ukraine was to a large degree a result of hate speech propagated online and through far-right subcultures. According to an ERRC expert, Jonathan Lee (2022), Ukrainian pogroms included at least seven attacks on Romani communities starting in April 2018, which were marked by offline violence that went hand-in-hand with a strong social media presence from far-right groups. There was an element of theater to most of these attacks which were live-streamed on social media. Warnings were posted on an online platform prior to the events, including ultimatums to leave the territory or face the consequences. After the attacks, messages were sent to their followers declaring the area "cleansed". The visual element to incite hatred and support online extremism was integral to these incidents. A similar case took place in Italy in April 2019, when several Romani families were placed in an emergency social shelter on the outskirts of Rome. Neo-Nazi movement CasaPound took the initial online hatred into the streets by inciting a riot. A mob set fire to cars outside the social shelter housing Romani families, prevented firefighters from accessing the blaze, and trampled food rations assigned to Roma while chanting: *"Those bastards must burn, let the gypsies die of hunger"* (Rorke, 2019a).

The incendiary nature of disinformation was demonstrated in a Paris suburb in March 2019, when an armed mob attacked a Romani community living in makeshift houses. The mob set fire to their vehicles, attacked the camp, and chased after a group of Roma in the nearby suburb—ultimately forcing them to hide in order to escape the violence. The incident was one of the twenty-five attacks against Romani communities sparked by a series of viral hoax videos claiming to show Roma in white vans abducting children. People spread false information on social media and messaging apps, posting pictures of the vans that were supposedly being used, and alerting others to be vigilant against *"Gypsies coming to steal kids"*. Such cases of violence against several different Romani communities with elements of a coordinated campaign show how the old hoaxed about Roma as child-snatchers regularly resurfaces in Europe today (Vitale, 2019; Rorke, 2019b).

The causality between online hate speech and offline hate crime has not been conclusively proven, but these and similar attacks indicate a link between the events where hate speech on social media plays a part in radicalizing and mobilizing parts of society. Types of events that carry higher risks tend to cluster in time and increase with triggering events and emerging conflicts (King & Sutton, 2013; PRISM Project, 2016). This calls into question the thresholds for tolerable hateful expression on social media as the online platforms keep hosting content that is dehumanizing and inciting—and against their public rules and standards on content moderation. Understanding the differentiated impacts and the level of harm to minority, marginalized, and other vulnerable groups when addressing hateful content can navigate the discussion toward a proactive exercise of mitigating the associated risks (Waldron, 2012).

3. Challenges and tensions of content moderation models countering online hate speech

a. Rules and guidelines addressing online hate speech

The public role of social media evolved faster than the safeguards protecting their users and the wider public. The responses addressing these policy, legal, and regulatory gaps have been asymmetric in their impact (Howard et al., 2019; Gorwa, 2019, Helberger et al., 2018). Major online platforms are private companies and as such enforce their internal rules, such as community guidelines and other standards, to determine what constitutes hateful content that should be reduced or removed. These guidelines largely prohibit hate speech, bullying and harassment, and violence and incitement. Hate speech is generally understood as a direct attack against people that share protected characteristics including race, ethnicity, and national origin. In enforcing their rules, companies rely on a combination of artificial intelligence (AI), user reporting, and human moderators to enforce their rules on permitted content. However, in practice, the social media landscape continues to be populated with posts and comments that go against these standards, proving that the current moderation systems are insufficient and prone to errors (Siapera et al., 2018; Tobin et al., 2017).

Social media operate across jurisdictions and their respective national frameworks include various interpretations of hate speech based on their legal, political, historical, cultural, and societal traditions. Significant differences can be found between the US legal framework, where the companies originated, and the European legal tradition. In the US, speech is protected unless it fits into several narrowly set categories for unlawful speech, such as speech with a demonstrable nexus between the expression in question and a substantial harm risk. Countries in Europe are at large more reactive to hate speech. Their legislative frameworks commonly prohibit certain forms of speech based on the content itself even in the absence of direct and explicit threats of violence, but significant variations remain (Siapera et al., 2018). There have been

several policy initiatives in regard to hate speech and its online forms, for instance, the Rabat Plan of Action outlining contextual considerations for the threshold of hate speech. These considerations include the context of the statement, the speaker's position or status, the intent to incite the audience against the target group, the content, and form of the statement, the extent of its dissemination, and the likelihood of harm, including imminence. The Council of Europe also issued important policy guidance for the European context and notably the Additional Protocol to the Convention on Cybercrime addressing racist and other materials promoting or inciting hatred, discrimination, or violence based on protected characteristics, including ethnic origin.

The case law of the European Court of Human Rights (ECtHR) as well focused on the many contextual factors of hate speech, such as the wider discourse in which a statement was made, the author, aim, value, and accuracy of the statement (Enarsson & Lindgren, 2019). The court further identified through its case law several groups, including Roma, that require special protection. In this framework, the court has stated that *"as a result of their turbulent history and constant uprooting the Roma have become a specific type of disadvantaged and vulnerable minority"* and that Roma people, therefore, require special protection (D.H. and others v Czech Republic, 2007). This approach can be considered victim sensitive as it acknowledges that some groups are disproportionately affected and aims to adequately balance the response. However, a differentiated approach that would be sensitive to the conflicting societal and security contexts that are relevant for local populations comes across the private actors' aim to provide a uniform product based on a global standard (Karanicolas, 2020). Social media platforms, therefore, find it difficult to reconcile their operational objectives with the increasing amount of hate speech circulating online.

b. Artificial intelligence (AI) models for detecting and removing hate speech content

Considering the sheer amount of online content, social media platforms increasingly rely on technical solutions to scale the moderation process (Walsh, 2020). Automated hash-matching and predictive machine learning tools, which are common algorithmic moderation systems, are being deployed to detect hate speech (Gorwa et al., 2020b; Ullmann & Tomalin, 2019). But algorithmic decision-making cannot avoid the fact that deciding what constitutes hate speech or harmful speech involves a value judgment. This judgment needs to consider multiple factors that are problematic to reduce to code—including societal norms, political situation, tensions, triggers, and values that may be shifting over time. These are all relevant for labeling anti-Roma content and evaluating the probability and extent of harm. This calculation needs to be done *ex-ante* through the design of the algorithm (Douek, 2018). AI models also encounter a series of technical, impact-related, and accountability problems. Algorithms are mostly pattern matchers that are vulnerable to manipulation and can perpetuate bias (Aubernach, 2015; Knight, 2017; Vidgen et al., 2020). The use of automated techniques also exacerbates the accountability deficit if the decisions are made without a human in the loop of the process (Gorwa et al., 2020b; Vidgen et al., 2020). A key area of concern is the opacity of algorithms—besides the end goals to which it is programmed, its behavior depends on multiple factors, such as the data it is trained on and how the data is labeled (European Parliament, 2019). When content moderation is performed exclusively via algorithms, the safeguards of human oversight and accountability are removed from the process (Gorwa et al., 2020a; Horwitz, 2021; Khan 2019).

Algorithmic language detection has raised concerns about the performance, robustness, generalizability, and fairness of these models (Waseem et al., 2018; Vidgen et al., 2019; Caselli et al. 2020; Mishra et al., 2019; Poletto et al., 2020). AI systems need to encompass a wide range of contextual and linguistic nuances to determine whether a particular word, phrase, or sentence is acceptable (Gorwa, 2020). Moreover, terms that have established connotations can be used by other groups with an opaque meaning (Cobbe, 2021; York

& McSherry, 2019; Gorwa et al., 2020b). For example, the ethnonym "gypsy" and its derivatives are the most productive words among the content marked as biased or hateful. "Antigypsyism" even refers to the explicit racism and xenophobia which is directed at Romani people. At the same time, the word has been reclaimed by the Roma people to show pride in their culture. Programming AI to reduce or delete posts that include the ethnonym "gypsy" as hate speech could therefore suppress minority voices. Different instances show how words used in dehumanizing meanings such as "monkeys," "trash," "vermin," or "parasites"— are also used in situations with neutral or innocuous meanings, making them difficult to be detected by algorithmic systems (Douek, 2018).

Hate speech is often not clear. It can escape the formalized language by including spelling mistakes, missing or adding some vowels or consonants or leaving blank spaces. AI models can be potentially circumvented by using mocking and sarcastic language. For example, the comment *"Try saying the word 'Roma' backward. Did you get the same thing as me?"* would be difficult to trace. The context in the Czech language is that "Rom", meaning Roma, reversed is "Mor", meaning plague. Problems can therefore arise if the content moderation is not adapted to local languages (Laub, 2019). Hate speech is also dynamic and evolving, using tropes and slang, being evasive, vague, or ambiguous (Siapera et al., 2018). For instance, in a comment *"Don't scream monkeys too much, I guess something will come soon!!!! And it will not be social benefits!!!!!"* the fact that it was written under a report about a new support program for a Romani settlement, coupled with references to animals, latent threats, and social benefits narratives, provides the evidence of this text being hateful. Writing these nuances into a code that will be efficient but not cause large collateral damage will demand a substantial improvement of algorithmic models for detecting hate speech.

Overinclusive detection of hate speech can paradoxically limit minority voices by deleting content created by human rights defenders, journalists, and activists, for example when reporting on Roma-related events, fact-checking disinformation, or portraying positive examples from Romani lives (Asher-Schapiro, 2017; Ghaffary, 2019). The platforms need to calibrate how much collateral damage against legitimate voices is acceptable, versus the amount of harmful content that is likely to escape detection. To use the precautionary principle, if we are unsure of the consequences of some process, we should err on the side of restriction for fear of an unknown disaster. Determining whether the risk of inaction is greater than the preventive action depends on numerous factors in which the local context is of key importance (Schauer, 2009). For example, the calculation of an error in areas with marginalized and segregated Roma communities will be different from places where Roma are dispersed as part of multi-ethnic societies. A lower threshold on hate speech is needed in cases of triggering events—such as the emergence of new camps, interethnic violence, or police brutality (Thompson & Woodger, 2020; Mendel, 2018; BBC, 2021). In such circumstances, the risk of insufficient preventive action can have real-life consequences in the form of violence, harm, and threats to human security.

Anti-discrimination provisions applied at speed can lead to censorship, especially if enforced inconsistently and without means to reverse these decisions (Dizdarevič, 2020; Guynn, 2019). But platforms have tools for balancing them with complaint mechanisms that allow for the reassessment and possible restoration of posts and accounts. However, for now, users who want to contest the decisions have limited options in regard to an appeal process (Tobin et al., 2017). Decisions behind the removal of specific content can be invisible to users or state only vague justification (Głowacka, 2021). Considering the influence these decisions have on free speech, platforms should increase the transparency of these processes and balance content removal with accessible means of appeal. To ensure that decisions to remove content are accurate, well-founded, and in line with the human rights standards—especially when automated models are used—companies should put in place effective and appropriate safeguards including independent verification and oversight (Siapera et al., 2018).

c. Human moderators, users reporting, and de-ranking content

Automated decision-making should at least partially include human judgment (Duarte et al., 2017; Siapera et al., 2018; Vidgen et al., 2020). Human moderators can provide context awareness, such as how the content can be perceived in different demographics, cultures, and geographies. For example, automated systems flag content for review by human moderators instead of automatically removing it. Still, the design of the gatekeeper algorithm remains highly consequential as moderators have only a few seconds to decide on individual posts or comments (Douek, 2018). Moderators can also apply the rules inconsistently (Tobin et al. 2017). Humans themselves have a strong bias, which may be particularly relevant for communities viewed as unfavorably as Roma. For instance, there has been a case when social media platforms contracted a company with alleged links to far-right groups in Ukraine (Starchenko, 2020). Moderators themselves may also not have the necessary language understanding and local awareness, required to recognize between nuanced or language-specific cases (Gorwa et al., 2020a). Companies must prioritize investments to engage diverse groups of qualified people and to improve the tools and support mechanisms that can enhance their work and inform their considerations of the risks connected to certain content. The platforms can also limit the scope of the fact-checking projects per team, which would allow them to train people on niche issues and content forensics (Stewart, 2021; Douek, 2018; Laub, 2019). This is an acute problem as moderators often work in substandard conditions as social media companies do not evenly allocate resources between the many markets they serve (Lang, 2021; Foxglove, 2021). Additionally, reviewing content takes a toll on people because of its high volume and the trauma that comes from sifting through disturbing posts and the moderators must be provided free of charge and accessible mental support (Cobbe, 2021; Newton, 2019; Reuters, 2019).

Other mechanisms can further improve systems for flagging harmful content. Many sites have dedicated end-users to this task such as “trusted flaggers” and similar programs that allow a fast-track review of content based on a company’s guidelines. It is an invite-only instrument that can be used by individuals and non-governmental organizations (NGOs) with identified expertise as well as government agencies. Trusted flaggers can help to find and label misinformation, disinformation campaigns, and hoaxes, but this framework has also serious limitations. This reactive model can be slow to act in situations of emerging tensions and high-risk events. Such mechanisms can be also misused with malicious conduct to overwhelm the reporting system (Douek, 2018). Still, partnerships between social media and NGOs can inform content decisions, while at the same time, raising awareness among social media providers, helping to understand the harm inflicted by racist material, and increasing their sense of public responsibility (PRISM Project 2016; Siapera et al., 2018; Susarla, 2021). The companies must be more transparent about the way they handle complaints and publish reports and statistics including disaggregated data about their decisions and operations, and such inconsistencies need to be examined and addressed (PRISM Project, 2016).

Content moderation discourse predominantly concentrates on the binary decisions on individual content that can be removed or kept on the platform. However, what information the users are exposed to and engage with is critical and this can be regulated. Platforms can decide to amplify or add friction to selected content, and these decisions are particularly central in high-risk situations. When heightened social tension is detected moderation can promote reliable information and reduce the distribution and virality of potentially harmful content so it circulates at a lower speed. This way, the content that has been identified as likely to violate rules on hate speech can be limited while the possible collateral damage is reduced. In this scenario, the more inflammatory the content is assessed to be, the less distribution it receives (Douek, 2021). The controlled lower engagement has been employed, for example, in the preparation for the trial of Derek Chauvin (Bickert, 2021; Roose, 2020). If the platforms can discourage hate speech and incitement to violence on special occasions, they can do so in other circumstances, too, in life-threatening examples such as organized mob

violence taking on Roma communities across Europe. Additionally, in situations of rising tensions with the potential of escalating violence, social media can restrict censorship decisions to local users through geo-blocking (Douek, 2018c). Measures that de-prioritize engagement, however, work against the companies' business models based on selling advertisements. For this reason, their algorithms are set to maximize viewing times, which can inadvertently promote extremist content and violence (Herrman, 2019).

4. Recommendations

Algorithmic detection models are integral to tackling hate speech, but they have considerable limitations for detecting linguistically nuanced and context-dependent cases. Instead of an over-reliance on a single moderating solution, the online environment can be built safer if it is approached holistically, with each layer extending the security and counterweighing the limitations of the other. While many tools are already in place, how they enforce the rules, where they set the thresholds, how they are developed to support users, and whether they are set to prioritize countering hate speech or higher engagement are all factors that determine their use and impact. Their potential efficiency is hindered by the lack of attention to the differentiated risks and harm stemming from hateful content, limited cooperation and information sharing between the companies and relevant stakeholders, and pervasive transparency and accountability deficiencies.

Primarily, social media must prioritize the detection of the most harmful and inflammatory segments of UGC as these have an imperative impact on people's safety and security. Such content includes retaliatory hate speech, disinformation, and dehumanizing language, incitement to violence directed at vulnerable communities, and racist rhetoric in situations with present triggering events or in other high-risk environments in regard to the community or interethnic violence. The focus should be on identifying coordinated harm and movements, applying a spectrum of counter-interventions to slow the visibility and availability of hateful content, and introducing robust safeguards for ensuring the protection of affected groups. There is an urgent need for a more thorough study of the impact of content moderation on different geographical, political, racial, ethnic, national, and cultural communities as well as an increased understanding of how diverse communities experience certain content or behavior and methods of measuring harms inflicting on segments of populations. Such detailed investigations remain limited, also due to the secrecy about algorithms and the absence of disaggregated data reporting on hate speech content and its removal.

Content moderation frameworks must include the communities that have their rights and security needs disproportionately affected compared to average users (Karanicolas, 2020). Guidelines for hate speech moderation set by the companies need to consider the specific circumstances of vulnerable groups as their vulnerability impacts the relative severity of harm against them, regardless of the content or violation (Scheuerman et al., 2021). The companies should employ a victim-sensitive approach to content moderation that acknowledges, respects, and draws on the experiences and views of groups and individuals whose security is being threatened (Brown, 2020). The practices should be assessed based on the impact they have on different segments of the population in different contexts. While scalable models for content moderation are highly sought after by companies with global operations, these models cannot capture the entire diversity of their users under a single moderation structure.

The platforms that impact globally must start moderating locally by recognising and acting on the differences and sensitivities across local contexts. A minimum standard across a wider population that does not consider the specific threats to vulnerable communities can cause their members will be less protected. An inspiration can be drawn from the example of the ECtHR ruling that identified Roma as a minority that due to its vulnerability requires special protection. Such considerations are critical in the context of security risks associated with online hate speech when certain disadvantages or vulnerability requires distinct

safeguards. Building a safer social media ecosystem inevitably includes trade-offs and it is important to identify and understand them. Improved content moderation models would not completely prevent users from encountering harm or restriction, but they would increase the levels of protection. Examining and responding to these compromises is key to ensuring that positive steps are retained (Rajendra-Nicolucci & Zuckerman, 2021).

Social media have acquired a pivotal role in setting red lines and parameters for speech online. Yet, as private companies, they are not bound by human rights law unless it is translated into legally binding provisions (HRC, 2018; Douek, 2018b; Taddeo & Floridi, 2016). Regulating private actors with public responsibilities has proven remarkably difficult. The companies advocate for self-regulation on the promise of adhering to their ethical guidelines. In practice, online platforms do not have independent oversight, which would guarantee compliance with these standards. As declared by Floridi (2019), the era of self-regulation is over. Regulatory initiatives are being prominently revamped by the European Union (EU) where legislation such as the Digital Services Act includes legal compliance and fines (Milmo, 2022; Floridi, 2021). Whether the stimulus to regulate harmful content comes from the private actors themselves or from regulation brought upon them by states or international bodies, the companies and other stakeholders alike must deepen their understanding of the multiple and interconnected problems in the online environment. This necessitates improved cooperation between the social media companies on one side, and the institutions, civil society organizations, academia, and other experts on the other. The online platforms must be able to upgrade their services permanently and continually with the purpose of assuring a safe and inclusive online space (PECAO, 2022).

5. Conclusion

The weak economic and social status of Romani men and women continues against the backdrop of their persistent stereotyping, discrimination, and ostracization. Recent years have seen a rising wave of anti-Roma rhetoric, especially in the context of the increasing prominence of far-right groups and xenophobic discourse. The coronavirus pandemic further intensified the stigmatization of Romani people, reinforcing the narrative according to which they present a threat to the majority population. The derogatory and often dehumanizing language is amplified in the online space with the potential of evolving into a negative spiral of hateful expressions and inflammatory statements that foster more hate or eventuate into physical harm. Such occasions demonstrate that the companies' decisions on the content available on their platforms can have literal life-and-death consequences.

Social media can provide a safer online environment—if they are context-aware, specific in their standards, transparent and accountable, conscious about their public responsibility, considerate of their impact, inclusive of the views and needs of those who are disproportionally impacted by their operations, and anticipatory about the associated and potential risks. Rather than solely relying on advanced automated solutions that remain imperfect for detecting contextual meanings and language nuances central to decisions on hateful expressions, models of content moderation should be applied holistically with each layer extending the security and compensating for the limitations of the other. The companies should be guided by a victim-sensitive approach and put in place robust safeguards to protect communities that are targeted or adversely affected by the content on their online platforms. Hate speech is not confined to Roma in Europe—it is a global problem. And the example of Romani communities representing some of the most marginalized groups provides important guidance for content moderation in addressing the asymmetric threats that hate speech presents to vulnerable populations.

References

Asher-Schapiro, A. (2017). YouTube and Facebook Are Removing Evidence of Atrocities, Jeopardizing Cases Against War Criminals. *The Intercept*. https://theintercept.com/2017/11/02/war-crimes-youtube-facebook-syria-rohingya/.

Aubernach, D. (2015). The Code We Can't Control. *Slate*. http://www.slate.com/articles/technology/bitwise/2015/01/black_box_society_by_frank_pasquale_a_chilling_vision_of_how_big_data_has.html.

Banks, J. (2011). European Regulation of Cross-Border Hate Speech in Cyberspace: The Limits of Legislation. *European Journal of Crime, Criminal Law and Criminal Justice*, 19, p. 1-13.

BBC. (2021). Roma street death was 'No Czech Floyd' say police. https://www.bbc.com/news/world-europe-57566697.

Benčík, J. (2020). Chce si Mazurek dať repete? *Dennik N*. https://dennikn.sk/blog/1929652/chce-si-mazurek-dat-repete/?fbclid=IwAR098BrWjkq_156cULQcgMbwPMN1YrX2zzyix83LIc4ifqScbLrCh62vlVs.

Bickert, M. (2021). Preparing for a Verdict in the Trial of Derek Chauvin. *Facebook*. https://about.fb.com/news/2021/04/preparing-for-a-verdict-in-the-trial-of-derek-chauvin/.

Bradshaw, S. & Howard, P. N. (2020). Oxford Internet Institute, University of Oxford. The global disinformation order: 2019 global inventory of organised social media manipulation. *Computational Propaganda Research Project*. https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf.

Brown, A. (2020). Models of governance of online hate speech. *Council of Europe*. https://rm.coe.int/models-of-governance-of-online-hate-speech/16809e671d.

Caselli, T., Basile, V., Mitrovic. J., Kartoziya, I. & Granitzer, M. (2020). I Feel Of- fended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language. *Proceedings of the 12th Conference on Language Re-sources and Evaluation* (LREC), p. 6193–6202.

Citron, D. K. (2009). Cyber Civil Rights. *Boston Law Review*, 89, p. 61-125.

Cobbe, J. (2021). Algorithmic Censorship by Social Platforms: Power and Resistance. Philos. Technol. 34, 739–766. https://doi.org/10.1007/s13347-020-00429-0.

Council of Europe (CoE). (1953). European Convention for the Protection of Human Rights and Fundamental Freedoms.

Council of Europe (CoE). (2003). Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems. https://rm.coe.int/168008160f.

Council of Europe (CoE). Factsheets on roma History: General introduction. https://rm.coe.int/factsheets-on-romani-history-general-introduction/16808b18e9

D.H and others v Czech Republic (App no 57325/00) 13 November 2007 § 182. https://www.echr.coe.int/documents/fs_roma_eng.pdf.

Dizdarevič, S. M. (2020). Re-Act Comparative report on the phenomena of online antigypsyism. *Project Re-ACT*. https://react.inach.net/wp-content/uploads/2020/10/Re-Act-Comparative-report-on-the-phenomena-of-online-antigypsyism.pdf.

Douek, E. (2018a). Facebook's Role in the Genocide in Myanmar: New Reporting Complicates the Narrative. *Lawfare Blog*. https://www.lawfareblog.com/facebooks-role-genocide-myanmar-new-reporting-complicates-narrative.

Douek, E. (2018b). U.N. Special Rapporteur's Latest Report on Online Content Regulation Calls for 'Human Rights by Default'. *Lawfare Blog*. https://www.lawfareblog.com/un-special-rapporteurs-latest-report-online-content-regulation-calls-human-rights-default.

Douek, E. (2018c). Zuckerberg's New Hate Speech Plan: Out with the Court and In with the Code. *Lawfare Blog*. https://www.lawfareblog.com/zuckerbergs-new-hate-speech-plan-out-court-and-code.

Douek, E. (2019). Why Facebook's 'Values' Update Matters. *Lawfare Blog*. https://www.lawfareblog.com/why-facebooks-values-update-matters.

Douek, E. (2020) Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability. 121 COLUM. L. REV. 759 (2021). http://dx.doi.org/10.2139/ssrn.3679607.

Douek, E. (2021). What Facebook Did for Chauvin's Trial Should Happen All the Time. *The Atlantic*. https://www.theatlantic.com/ideas/archive/2021/04/facebook-should-dial-down-toxicity-much-more-often/618653/.

Duarte N., Llanso E., & Loup A. (2017). Mixed Messages? The Limits of Automated Social Media Content Analysis. Washington, DC: *Center for Democracy & Technology*. https://perma.cc/NC9B-HYKX.

Enarsson T., & Lindgren. S. (2019). Free speech or hate speech? A legal analysis of the discourse about Roma on Twitter. *Information & Communications Technology Law*, 28:1, p. 1-18. doi:10.1080/13600834.2018.1494415.

European Commission. *Roma equality, inclusion and participation in the EU.* https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/roma-eu/roma-equality-inclusion-and-participation-eu_en.

European Commission. (2020). *EU Roma strategic framework for equality, inclusion and participation for 2020 – 2030.* https://ec.europa.eu/info/sites/default/files/union_of_equality_eu_roma_strategic_framework_for_equality_inclusion_and_participation_en.pdf.

European Commission. (2021). *Roma Inclusion in Bulgaria, Hungary, Romania, Slovakia.* https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/roma-eu/roma-inclusion-eu-country.

European Parliament. (2019). *A Governance framework for algorithmic accountability and transparency.* https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf.

European Roma Rights Centre (ERRC). (2019). *Mob Justice: Collective Punishment against Roma in Europe.* https://issuu.com/romarightsjournal/docs/mob-justice-collective-punishment-a.

Floridi, L. (2019). Translating Principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, *32*(2), 185–193. https://doi.org/10.1007/s13347-019-00354-x.

Floridi, L. (2021). The End of an Era: from Self-Regulation to Hard Law for the Digital Industry. *Philos. Technol.* 34, 619–622. https://doi.org/10.1007/s13347-021-00493-0.

Foxglove. (2021). *Open letter from 60 content moderators to Facebook calling for proper mental health support and an end to outsourcing.* https://www.foxglove.org.uk/2021/07/25/open-letter-from-60-content-moderators-to-facebook/.

Foxman, A. H., & Wolf, C. (2013). *Viral Hate: Containing Its Spread on the Internet*. London: Palgrave Macmillan.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). Countering Hate Speech. *UNESCO*, Countering Hate Speech. http://unesdoc.unesco.org/images/0023/002332/233231e.pdf.

Gagliardone, I., Patel, A., Pohjonen, M. (2014). Mapping and Analysing Hate Speech Online: Opportunities and Challenges for Ethiopia. *SSRN Electronic Journal*. doi:10.2139/ssrn.2601792.

Ghaffary, S. (2019). The algorithms that detect hate speech online are biased against black people. *Vox.* https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter.

Głowacka D. (2021). Fighting for Free Speech Online: SIN vs Facebook. *Digital Freedom Fund*. https://digitalfreedomfund.org/fighting-for-free-speech-online-sin-vs-facebook/.

Google. About the YouTube Trusted Flagger program. https://support.google.com/youtube/answer/7554338?hl=en.

Gorwa, R. (2019). The platform governance triangle: conceptualising the informal regulation of online content. *Internet Policy Review*, [online] 8(2). https://policyreview.info/articles/analysis/platform-governance-triangle-conceptualising-informal-regulation-online-content.

Gorwa, R., Binns R., & Katzenbach C. (2020a). Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance. *Big Data & Society*. doi:10.1177/2053951719897945.

Gorwa, R., Reuben B., and Christian K. (2020b). Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance. *Big Data & Society*. doi:10.1177/2053951719897945.

Guynn J. (2019). Facebook while black: Users call it getting 'Zucked,' say talking about racism is censored as hate speech. *US Today*. https://eu.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/.

Hale, S. (2020). COVID-19: National Equality Bodies Report Impacts on Equality and Share Their Responses to the Coronavirus Pandemic. *European Network of Equality Bodies (EQUINET)*. https://equineteurope.org/2020/covid-19-national-equality-bodies-report-impacts-on-equality-and-share-their-responses-to-the-coronavirus-pandemic/.

Hamelmann, M. (Ed.). (2018). Antigypsyism on the Internet, *sCan project*. http://scan-project.eu/wp-content/uploads/scan-antigypsyism.pdf.

Hamelmann, M., Lhopitault, C., & Schadauer, A. (2017). Manifestations of Online Hate Speech: Reports on antisemitic, antiziganistic, homophobic and anti-Muslim Hate Speech. *International Network Against Cyber Hate (INACH)*. https://www.inach.net/wp-content/uploads/Manifestations_of_online_hate_speech-short-final.pdf.

Heinze, E. (2016). *Hate Speech and Democratic Citizenship*. Oxford: Oxford University Press.

Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *The Information Society*, 34(1), 1-14. https://doi.org/10.1080/01972243.2017.1391913

Herrman, J. (2019). How Secrecy Fuels Facebook Paranoia. The New Your Times. https://www.nytimes.com/2019/01/16/magazine/facebook-election-analytics.html.

Herz, M. & Molnar, P. (2012). The Content and Context of Hate Speech: Rethinking Regulation and Responses. Cambridge: Cambridge University Press.

Hoffmann, S., Taylor E., & Bradshaw, S. (2019) The Market of Disinformation. *Oxford Technology & Elections Commission.*

Horwitz, J. (2021). Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt. *Wall Street Journal*. https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353.

Human Rights Council of the United Nations (HRC). (2018). Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, A/HRC/38/35. https://digitallibrary.un.org/record/1631686/usage?ln=en.

Karanicolas M. (2020). The Countries Where Democracy Is Most Fragile Are Test Subjects for Platforms' Content Moderation Policies. *Slate*. https://slate.com/technology/2020/11/global-south-facebook-misinformation-content-moderation-policies.html.

Keipi, T., Näsi, M., Oksanen, A., & Räsänen, P. (2016). *Online Hate and Harmful Content: Cross-national perspectives*. Taylor & Francis. ISBN 9780367876968. https://library.oapen.org/handle/20.500.12657/22350.

Khan, M. (2019). More 'hate speech' being removed from social media. *Financial Times*. https://www.ft.com/content/868f9d96-27bc-11e9-a5ab-ff8ef2b976c7.

King, R.D., & Sutton G. M. (2013). High Times for Hate Crime: Explaining the Temporal Clustering of Hate Motivated Offending. Criminology 51(4), p. 871-894.

Knight, W. (2017). Biased Algorithms Are Everywhere, and No One Seems to Care. *Technology Review*. https://www.technologyreview.com/s/608248/biased-algorithms-are-everywhere-and-no-one-seems-to-care/.

Kyslova, O., Kuzina, I., & Dyrda. I. (2020). Hate Speech Against the Roma Minority on Ukrainian Web Space. *Ideology and Politics*. doi:10.36169/2227-6068.2020.01.00021.

Kyuchokov, H. (2015) Preface. In: Selling, Jan; et. al.: *Antiziganism. What's in a Word?*. Cambridge Scholars Publishing.

Lang, K. (2021). The activist lawyers taking on the government. *The Times*. https://www.thetimes.co.uk/article/the-activist-lawyers-taking-on-the-government-n2pbt0nzr.

Laub, Z. (2019). Hate Speech on Social Media: Global Comparisons. *Council on Foreign Relations*. https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons.

Lee, E., & Leets, L. (2002). Persuasive storytelling by hate groups online: Examining its effects on adolescents. *American Behavioral Scientist*, 45(6), p. 927–957. doi:10.1177/0002764202045006003.

Lee, J. (2022). Personal interview. 28 March 2022.

Liotta, P.H. & Owen, T. (2006). Why Human Security? *The Whitehead Journal of Diplomacy and International Relations*. 7 (1): 37-54.

Malek, P. (2017). Nenávist k cizincům odnesly děti na památeční fotografii prvňáčků. *Teplicky Dennik*. https://teplicky.denik.cz/zpravy_region/nenavist-k-cizincum-odnesly-deti-na-pamatecni-fotografii-prvnacku-20171106.html.

Mendel, I. (2018). Attacks on Roma Force Ukraine to Confront an Old Ethnic Enmity. *The New Your Times*. https://www.nytimes.com/2018/07/21/world/europe/ukraine-roma-attacks.html.

Milmo, D. (2022). EU agrees rules to force big tech to rein in illegal content or face huge fines. *The Guardian.* https://www.theguardian.com/world/2022/apr/23/eu-agrees-rules-to-force-big-tech-to-rein-in-content-or-face-huge-fines.

Mishra, P., Yannakoudakis, H., & Shutova, E. (2019). *Tackling online abuse: A survey of automated abuse detection methods.* arXiv:1908.06024v2, p. 1–17.

Miškolci, J., Kováčová L., & Rigová E. (2020). Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia. *Social Science Computer Review*, 38, no. 2, p. 128–46. doi:10.1177/0894439318791786.

Näsi, M., Räsänen, P., Oksanen, A., Hawdon, J., Keipi, T., & Holkeri, E. (2014). Association between online harassment and exposure to harmful online content: A cross-national comparison between the United States and Finland. *Computers in Human Behavior*, 41 (December), p. 137–145. doi: 10.1016/j.chb.2014.09.019.

Newton, C. (2019). The Trauma Floor: The secret lives of Facebook moderators in America. *The Verge.* https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

OHCHR. (2020a). *UN Special Rapporteur on minority issues: International Roma Day – 8 April 2020: "UN expert urges political action to promote equality and non-discrimination during the COVID-19 crisis".* https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25782&LangID=E.

OHCHR. (2020b). *"UN expert denounces the propagation of hate speech through social media".* https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25626&LangID=E.

Okamura (2021a). Facebook. https://www.facebook.com/tomio.cz/videos/815940885763376/

Okamura (2021b). Twitter. 6 October 2021. https://twitter.com/tomio_cz/status/1445862810227982337.

Oksanen, A., J. Hawdon, E. Holkeri, M. Näsi, and P. Räsänen. (2014). *Exposure to Online Hate among Young Social Media Users.* In Soul of Society: Emerald Group Publishing Limited.

Organization for Security and Co-operation in Europe (OSCE). (2006). *Decision No. 13/06 on Combating Intolerance and Discrimination and Promoting Mutual Respect and Understanding.*

OSCE Office for Democratic Institutions and Human rights (ODIHR). (2021). *OSCE/ODIHR Hate Crime Reporting, Overview of Incidents.* https://hatecrime.osce.org.

OSCE Office for Democratic Institutions and Human rights (ODIHR). (2020a). *OSCE Human Dimension Commitments and State Responses to the Covid-19 Pandemic.* https://www.osce.org/files/f/documents/e/c/457567_0.pdf.

OSCE Office for Democratic Institutions and Human rights (ODIHR). (2020b). *Roma and Sinti in the Media: 2020 Monitoring.* https://www.osce.org/odihr/473904?fbclid=IwAR3P0088F_Vt2sA1x71v8rMM-2eysoLqaZt2Kx8LTicNGvcs7rE9VqyZSb0.

Paladino B. (2018). Democracy disconnected: Social media's caustic influence on Southeast Asia's fragile republics. *Brookings.* https://www.brookings.edu/research/democracy-disconnected-social-medias-caustic-influence-on-southeast-asias-fragile-republics/.

PECAO. (2022). PECAO Synthesis Report on Antigypsyist Online Hate Speech. *ERGO Network.* https://ergonetwork.org/wp-content/uploads/2022/04/ERGO-Synthesysreportweb-FINAL.pdf.

Persily, N., & Tucker, J. (Eds.). (2020). *Social Media and Democracy: The State of the Field, Prospects for Reform* (SSRC Anxieties of Democracy). Cambridge: Cambridge University Press. doi:10.1017/9781108890960.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, p. 1–47.

PRISM Project. (2016). *Backgrounds, Experiences and Responses to Online Hate Speech: A Comparative Cross-Country Analysis.* https://sosracismo.eu/wp-content/uploads/2016/07/Backgrounds-Experiences-and-Responses-to-Online-Hate-Speech.pdf.

Quarmby, K. (2020). Hate targeted at Gypsy, Traveller and Roma linked to rise in suicides – report. *The Guardian.* https://www.theguardian.com/world/2020/dec/10/hate-targeted-at-gypsy-traveller-and-roma-linked-to-rise-in-suicides-report.

Rajendra-Nicolucci, C. & Zuckerman, E. (2021). An Illustrated Field Guide to Social Media. *Knight First Amendment Institute.* Columbia University. https://knightcolumbia.org/blog/an-illustrated-field-guide-to-social-media.

Reuters. (2019). Beheadings, Suicide Attempts, Porn: Why Facebook Moderators in India Are Traumatised. *Huffington Post.* https://www.huffingtonpost.in/entry/beheadings-suicide-attempts-porn-why-facebook-moderators-in-india-are-traumatised_in_5c793417e4b087c2f2954fb2.

Rikke F. J., (2021). A Human Rights-Based Approach to Social Media Platforms. Berkley Forum. https://berkleycenter.georgetown.edu/responses/a-human-rights-based-approach-to-social-media-platforms.

Roberts S.T. (2017). *Content Moderation.* In: Schintler L., McNeely C. (eds) Encyclopedia of Big Data. Springer, Cham. doi:10.1007/978-3-319-32001-4_44-1.

Roose, K. (2020). On Election Day, Facebook and Twitter Did Better by Making Their Products Worse. *The New Your Times.* https://www.nytimes.com/2020/11/05/technology/facebook-twitter-election.html.

Rorke B. (2018). Anti-Roma Pogroms in Ukraine: On C14 and Tolerating Terror. *European Roma Rights Centre (ERRC).* http://www.errc.org/news/anti-roma-pogroms-in-ukraine-on-c14-and-tolerating-terror.

Rorke, B. (2019a). Smells like Fascism II: new outbreak of anti-Roma mob violence in Italy. *European Roma Rights Centre (ERRC).* http://www.errc.org/news/smells-like-fascism-ii-new-outbreak-of-anti-roma-mob-violence-in-italy.

Rorke B. (2019b). A spectre is haunting Europe – spike in anti-Roma pogroms as EU election campaigns kick off. *European Roma Rights Centre (ERRC).* http://www.errc.org/news/a-spectre-is-haunting-europe---spike-in-anti-roma-pogroms-as-eu-election-campaigns-kick-off.

Rorke, B. (2020). More toxic than covid? The politics of anti-Roma racism in Bulgaria. *European Roma Rights Centre (ERRC)*. http://www.errc.org/news/more-toxic-than-covid-the-politics-of-anti-roma-racism-in-bulgaria.

Rorke B. & Lee, J. (2020). Roma Rights in the Time of Covid. European Roma Rights Centre (ERRC). http://www.errc.org/uploads/upload_en/file/5265_file1_roma-rights-in-the-time-of-covid..pdf.

Sander, B. (2021). Democratic Disruption in the Age of Social Media: Between Marketized and Structural Conceptions of Human Rights Law. *European Journal of International Law*. Volume 32, Issue 1, February 2021. P. 159–193. https://doi.org/10.1093/ejil/chab022.

Salvini, M. (2019). Twitter. 1 August 2019. https://twitter.com/matteosalvinimi/status/1156911065025916932.

Schauer, F. (2009). Is It Better to Be Safe than Sorry?: Free Speech and the Precautionary Principle, 36, Pepp. L. Rev. Iss. 2. https://digitalcommons.pepperdine.edu/plr/vol36/iss2/3.

Scheuerman, M. K., Jiang, J. A., Fiesler, C. & Brubaker, J. R. (2021). A Framework of Severity for Harmful Content, [online]. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 368 (October 2021), 33 pages. https://doi.org/10.1145/3479512.

Siapera E., Moreo E., & Zhou J. (2018). Hate Track Tracking and Monitoring Racist Speech Online. *Irish Human Rights and Equality Commission*. https://www.ihrec.ie/app/uploads/2018/11/HateTrack-Tracking-and-Monitoring-Racist-Hate-Speech-Online.pdf.

Siapera, E., Viejo-Otero, P., & Moreo, E. (2017). *Hate Speech: Genealogies, Tensions and Contentions*, paper presented at the Association of Internet Researchers (AoIR) conference, 19-21 October, University of Tartu, Estonia.

Snidl, V. (2019) Facebook post from 4 February 2019. https://www.facebook.com/vladimir.snidl/posts/10218831460516412.

Starchenko A. (2020). *Zaborona vs. StopFake: what is hiding behind Ukraine's ongoing media conflict? New Eastern Europe.* https://neweasterneurope.eu/2020/08/03/zaborona-vs-stopfake-what-is-hiding-behind-ukraines-ongoing-media-conflict/.

Statcounter. (2021). *Social Media Stats.* https://gs.statcounter.com/social-media-stats.

Susarla, A. (2021). If Big Tech has the will, here are ways research shows self-regulation can work. *The Conversation*. https://theconversation.com/if-big-tech-has-the-will-here-are-ways-research-shows-self-regulation-can-work-154248.

Taddeo, M., and L. Floridi. 2016. The Debate on the Moral Responsibilities of Online Service Providers. *Science and Engineering Ethics* 22 (6): 1575–1603.

Thompson, N., & Woodger, D. (2020). I hope the river floods: Online hate speech towards Gypsy, Roma and Traveller communities. *British Journal of Community Justice*, 16(1), p. 41-63. ISSN 1475-0279. https://mmuperu.co.uk/bjcj/wp-content/uploads/sites/2/2020/09/BJCJ_Thompson_and_Woodger_2020.pdf.

Tobin, A., Varner, M., & Angwin, J. (2017). Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up. *ProPublica.* https://www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes.

Tomičić, A. (2018). Racial Cities: Governance and the segregation of Romani people in urban Europe, Transnational Social Review. 8:3, 346-349, DOI: 10.1080/21931674.2018.1502927.

Ullmann, S. & Tomalin, M. (2019). Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology*, p. 1-12. https://link.springer.com/article/10.1007/s10676-019-09516-z.

United Nations (UN). (2019). *United Nations Strategy and Plan of Action on Hate Speech.* https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml.

United Nations Development Programme (UNDP). (1994). *Human Development Report 1994: New Dimensions of Human Security*. Oxford: Oxford University Press.

United Nations Trust Fund for Human Security (UNTFHS). (2009). *Human Security in Theory and Practice*, New York: United Nations.

Van Baar, H. (2014). *The Securitization of Gypsies, Travellers and Roma in Europe: Context, Critique, Challenges.* Keynote speech delivered at New Scotland Yard. London, UK. Organized by IDRICS, Bucks New University and The University of Warwick. https://www.academia.edu/10862181/The_Securitization_of_Gypsies_Travellers_and_Roma_in_Europe_Context_Critique_Challenges_2014_.

Vesnianka, O. (2021). Personal interview. 18 May 2021.

Vidgen, B. Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. *Proceedings of the Third Workshop on Abusive Language Online (ACL)*, p. 80–93.

Vidgen, B., Thrush, T., Waseem, Z., & Kiela. D. (2020). Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. CoRR abs/2012.15761. https://arxiv.org/abs/2012.15761.

Vitale, T. (2019). Why are Roma being attacked in France? *The Conversation.* https://theconversation.com/why-are-roma-people-being-attacked-in-france-115030

Votavová, J. (2020). Re-Act Analytical paper on debunking online antigypsyism. *Project Re-ACT.* https://react.inach.net/wp-content/uploads/2020/10/Re-Act-Analytical-paper-on-debunking-online-antigypsyism.pdf.

Waldron, J. (2012). *The Harm in Hate Speech.* Harvard University Press.

Walsh. E. (2020). Facebook spent the equivalent of 319 years labeling or removing false and misleading content posted in the US in 2020. *Business Insider.* https://www.businessinsider.com/facebook-spent-319-years-labeling-misleading-content-posted-us-2021-9.

Warmisham, J. (2016). The situation of Roma and Travellers in the context of rising extremism, xenophobia and the refugee crisis in Europe. *Congress of Local and Regional Authorities.* https://rm.coe.int/1680718bfd.

Waseem, Z., Thorne, J., & Bingel, J. (2018). *Bridging the gaps: Multitask learning for domain transfer of hate speech detection.* In Jennifer Golbeck, editor, Online Harassment, p. 29–55. Springer International Publishing, Cham.

Web Noviny. (2020). *Rómovia v karanténe dostali od štátu zadarmo alkohol a potraviny. Je to hoax, ktorý zdieľali tisícky ľudí.* https://www.webnoviny.sk/romovia-v-karantene-dostali-zadarmo-alkohol-a- potraviny-hoax-zdielali-tisicky-ludi/.

Wike, R., Poushter, J., Silver, L., Devlin, K., Fetterolf, J., Castillo, A., & Huang., C. (2019). European Public Opinion Three Decades After the Fall of Communism. Chapter 6.: Minority Groups. *Pew Research Centre*. https://www.pewresearch.org/global/2019/10/14/minority-groups/.

World Federation of Advertisers (WFA). (2020). *WFA and platforms make major progress to address harmful content.* https://wfanet.org/knowledge/item/2020/09/23/WFA-and-platforms-make-major-progress-to-address-harmful-content.

York, J., & McSherry, C. (2019). Content moderation is broken. Let us count the ways. *Electronic Frontier Foundation Blog.* https://perma.cc/7FA6- WD6Z.