Ivana Stepanović[*]

ALGORITHMIC REPRODUCTION OF HATE ON SOCIAL MEDIA

Abstract: Inspired by agent-based modelling experiments, social media strive to alter human behaviour by endorsing content that feeds our subconscious cravings to stimulate reactions or production and reproduction of digital content and, ultimately, to motivate financial transactions. YouTube dramas, dangerous TikTok challenges and hatred-fuelled comments and hashtags are instigated by profit-driven strategies and engagement-based ranking. Algorithms prefer clickable repetitive content and therefore stimulate hyperproduction of hate speech simply because it drives engagement. In this way, algorithms are regulating visibility on social media, but their settings are biased because they always adhere to the logic of the market rather than ethical standards. This rapid production of content is impossible to control or censor in real-time, and legal regulations are usually applied *post festum* after a cybercrime has already been committed. In Serbia, the main problems are the lack of specialised legislation and cybercrime prevention mechanisms, but also the awareness that the so-called 'negative comments' can be interpreted as a type of crime that can be sanctioned. This paper investigates how fast 'prosumption' (a combined process of production and consumption) leads to the proliferation of hate on social media while underlining the importance of raising awareness and improving the prevention of cybercrimes that are stemming from hate-fuelled narratives on social media.

Keywords: algorithms, hate speech, cybercrime, prosumption, social media

[*] Dr Ivana Stepanovic, Institut za kriminološka i sociološka istraživanja, Srbija i Institut za napredne studije Koszeg, Mađarska, ivana.stepanovic@iask.hu

1. Introduction

Human rights in the age of mass surveillance and algorithmic social sorting (Baruh & Popescu, 2015: 579) on social media and other platforms are no longer reduced to the protection of personal data in the traditional sense. The extraction of user data has become so refined that it threatens the essence of the self, individuality, and subjectivity on the corporeal and spiritual levels. The concept of surveillance has expanded to include „surveillance of emotional life“ (McStay, 2020: 1) and „intellectual surveillance“ that is juxtaposed with „intellectual privacy“ (Richards, 2013: 1953) because contemporary data collection and data manipulation practices threaten to eliminate the freedom of thought and drive the world into the darkest Orwellian nightmare. Since behavioural data include everything from patterns of sleeping, walking, or eating to information about thoughts and emotions read through facial expressions, breathing rhythms, eye movements and tone of voice, surveillance practices have become overly sophisticated and intrusive to the extent that they leave no room the existence of an inner world. Foucault's insights into surveillance practices of modernity have shown that even passive observation leads to the alteration of behaviour and that they function as disciplinary measures, not only in prisons which were designed to discipline convicts, but also in all institutions of modern societies, including factories, schools, or hospitals (Foucault 1995: 207). Algorithmic surveillance is only an extension of the much older system for controlling large populations, but due to its active role it is not simply an archiving machine but an automated decision-making mechanism that directly intervenes in human life (Stepanović, 2020).

From Domesday Book[1] to Facebook, categorisation, and systematisation of people through bureaucratic archives has been the method of governing large populations. With artificial intelligence, a traditional archive shapes up to become a live sorting mechanism that sources, stores, classifies and interprets the data while making significant decisions that concern people's personal or professional life and political decisions. But algorithmic surveillance can be manipulative (Darmody & Zwick, 2020; Pasquale, 2015) and lead to all types of cybercrimes. Research in psychology and agent-based modelling experiments can be used to manipulate the outcome of elections or referendums. The most notable example is the Cambridge Analytica scandal (Ward, 2018) which has revealed how social media can be used to mine data, create psychological profiles, and then target users with content tailored to shape their political opinions. This is because algorithms are in their

[1] The Domesday Book, available at: domesdaybook.co.uk (Accessed: 29.03.2022.)

nature manipulable and manipulative. Understanding the criminal aspects of algorithmic surveillance is essential to examine how social media and other online platforms undermine digital rights that go beyond the protection of personal data and how they contribute to the reproduction of hate online. Raising awareness of these processes and of the dangers of hate speech narratives can help improve the mechanisms for legal protection from various cybercrimes while the shift towards a more curated internet can avoid the trap of using the biased, flawed, and intrusive surveillance-based algorithms for hate detection and censorship.

2. Prosumption of Hate Speech

Algorithmic content sorting on social media platforms crucially affects the logic of the production, reproduction, and consumption of hate speech narratives. These narratives populate the online public spaces of the internet „repeatedly, systematically and uncontrollably“ (Castano-Pulgarin et al. 2021: 1) because they are, in a way, promoted by the organising algorithms. Initially, social media platforms had linear news feeds, which meant that all the content was visible to all the users and that it was ordered according to the time of posting. As the numbers of profiles and the amount of content started to grow rapidly, the platforms have introduced algorithms to achieve a non-linear organisation of content and „impose a quantitative logic of visibility“ (Sued et al. 2021). Programmed to create a new system of information distribution, the algorithms have shaped up to become automated decision-makers with abilities to censor and control the internet. This new algorithmic order was introduced to manage large quantities of information dispersed in the online space. The main concern is typically the capability of the so-called „algorithmic censorship“ to „exercise an unprecedented degree of control over both public and private communications“ (Cobbe, 2021: 739), but the uncontrollable reproduction of harmful content is potentially even more damaging for societies.

By manipulating what we see, algorithms have transformed a free and democratic space of the internet into a highly controlled environment where information is disseminated according to the parameters set by the platforms and services themselves. Social media have come to replace the traditional top-down mass media but only to offer a „new form of authoritarianism“ and „machine politics“ (Turner, 2019) with the help of artificial intelligence. From search engines to social media and even streaming platforms like Netflix, using surveillance-based algorithms as organising mechanisms is fully normalised. What was conceived as a simple and highly practical navigation system now operates as a mechanism of censorship that jeopardises human rights and freedoms and consequently also becomes a propaganda machine. Manipulation lies in the very core of the algorithmic organisation purely because it is a way to control the distribution of information and consequently knowledge itself.

The two most important principles of algorithmic information sorting that make it problematic are the engagement-based ranking and surveillance-based personalisation of content. Namely, by favouring engagement (or the frequency of clicks, reactions, and shares), algorithms are serving the financial needs and goals of platforms. They are market-oriented, and they tend to promote content with a high level of engagement. At the same time, platforms are collecting and processing personal information which also includes behavioural data to assess the users and offer them information tailored to their interests and habits. With the aim to utilise the data to boost sales and increase revenues, social media and all other services are essentially conceptualised as advertising platforms where all activities are in the function of the market. As a result, all personal data, including behavioural data that are collected on the platforms are commodified through the process of prosumption (Gerbaudo 2015: 81; Dyer-Witheford 2015: 92; Duffy et al 2021: 1; Fuchs 2014: 245). A combination of production and consumption, prosumption is the *modus operandi* of social media and all other platforms on the internet. Unlike the traditional media, contemporary ones are interactive because the consumer of the content is at the same time the producer, the resource, and the product itself. One of the key problems is the „abuse the internet for commercial purposes“ because algorithmic sorting leads to „inconsistent“ moderation and censorship practices (Bromell, 2022: 29). This means that do not prioritise the battle against hate speech and other forms of cybercrime but are rather encouraging mass prosumption of engageable content, regardless of whether it includes abusive and harmful narratives.

Prosumption processes guided and navigated by the algorithmic surveillance and engagement-based rankings lead to uniformed online landscapes where reproduction of the same or the similar is a norm rather than an exception. Originality is not desirable as it produces lower levels of engagement than repetitive content that is cross-referenced and anchored with well-known keywords or hashtags. Illuminating the logic and the mathematics of the algorithmic prosumption, therefore, explains the overpowering hate speech narratives. They are produced, reproduced, and replicated in such a way that they create patterns and even trends of online behaviour. Digital ethnography research shows how these narratives appear in clusters of identical or look-alike social media content as well as reactionary comments attached to it. Algorithms are set to drive engagement and financial transactions. They tend to be guided by the market needs rather than ethical standards, and their efforts to impose algorithmic hate speech detection and censorship mechanisms remain to be futile. They are undermined by the sheer complexity of linguistic content that is challenging for artificial intelligence systems (Kovács, Alonso & Saini, 2021) but also by the parallel algorithms that regulate visibility and prioritise posts that draw engagement.

In the era of instant sharing on market-driven platforms, engagement is always favoured over ethical values even if the content is focusing on hatred and involves other aspects of cybercrime. Since it is extremely challenging to control or censor hate speech on the internet in real-time, legal regulations are typically applied *post festum*, and there are no adequate preventive mechanisms. This problem is related to the „dispute in international human rights law" over the freedom of speech and the prohibition of hate speech (O'Regan, 2018: 403). In Serbia, the proliferation of hate speech on social media shows a low level of awareness of its criminal aspects. The so-called „negative comments" are usually understood as normal, even though they can often be interpreted as digital crimes that correspond to certain types of classic crimes such as bullying, stalking or even domestic violence.

3. The Curated Internet vs Algorithmic Wars

The manipulative nature of algorithmic sorting is explicitly discussed in the proposals for the two new legal regulations of the European Union, namely the Artificial Intelligence Act (AIA)[2] and the Digital Services Act (DSA)[3]. These proposals recognise that algorithmic manipulation is a realistic threat that stems from intrusive dataveillance practices related to modern digital technologies, especially artificial intelligence systems and digital platforms such as social media. These two newly formulated legislations that have not yet entered into force nevertheless reconceptualised the idea of digital privacy and online rights. Rather than focusing solely on the privacy of data, which is the case with the General Data Protection Regulation[4], the two new laws are emphasising the importance of the protection of interconnected rights and freedoms.

Agent-based modelling shows how algorithms can be used to analyse, understand, and predict behaviour, but they can also be utilised to modify it as well. Speaking about challenges of creating such simulations in the post-truth era, Sobkowicz says that „increasingly detailed data on our behaviours, and the tools to analyse it open the way not

[2] Artificial Intelligence Act, available at: eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, (Accessed 29. 03. 2022.)

[3] Digital Services Act, available at: eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A825%3AFIN, (Accessed 29. 03. 2022.)

[4] General Data Protection Regulation, available at: gdpr-info.eu, (Accessed: 29.03.2022.)

only to understand social behaviours but also to monitor (often in real-time) and manipulate behaviours, both on an individual level and for social groups“ (Sobkowicz, 2020). Due to the vast amount of personal data available on their servers, social media platforms can deploy strategies to affect human reasoning or instigate certain behaviours. Targeted marketing can have wide social implications and can involve manipulative tactics that are, or at least should be, considered illegal. However, the absence of adequate legislation that regulates the domain of artificial intelligence systems and digital platforms makes many of these digital crimes ultimately unpreventable and unpunishable. The concept of digital criminology (Stratton, Powel & Cameron, 2017) should therefore be expanded to include a broad range of issues, including, in a very wide sense, the structure and the functioning of platform algorithms that are surveillance-based and designed to allow the instant proliferation of engageable content regardless of whether it involves hate speech or not.

Currently, social media rely on algorithmic systems to organise visibility and filter out hate speech. One of the key problems with these algorithms is their increasing intrusiveness. Some of the recent research projects suggest that hate speech detection systems involve surveillance practices such as recording and analysing the facial expressions of users while they are posting content online (Montefalcon et al, 2021: 201). On the other hand, the algorithms themselves favour engagement-based ranking and are not sensitive to all forms of hate speech. They, therefore, contribute to the reproduction of hate through cyber wars on Twitter, YouTube, TikTok, Instagram and other social media.

Considering the large numbers of users who have the possibility to instantly share content or comment on it, it is virtually impossible to control activities as they happen and prevent hate speech and other types of cybercrime. In many cases, hate speech can be properly detected only retroactively. Social spaces on the internet are not by default curated spaces like traditional media where every piece of content is carefully assessed by the editors before it gets published. However, they allow the creation of clusters of smaller, curated profiles - much like the traditional media. Some of the social media channels already comply with these standards. They publish only edited and approved content and work with designated moderators who are controlling comments as well. This means that social media could potentially be redefined to provide a platform only for profiles that can guarantee the respect of specific content guidelines and cybercrime laws. However, the concept of the curated internet that is not synonymous with censored internet is not commonly discussed in the scientific community. The ideas are rather coming from platform users, content creators and entrepreneurs who are thinking of ways to organise online space crammed with user-generated content. For them, the main problem is an insurmountable amount of information online which is too much not only for humans but even for machines[5], and they see a curated internet as a solution and an alternative to algorithmic filtering and censorship.

There are two main obstacles to creating more controlled and curated social media. The first one is the core „democratic" principle, as social media were initially designed to offer everyone freedom of speech and the chance to have their voice heard. On the other hand, the level of misuse of social media and high crime rates challenge the validity and justification of this principle. As a response, social media platforms strive to find solutions to battle cybercrimes through the system of algorithmic policing of users and censorship of the uploaded content, but these practices only lead to further problems with algorithmic bias and responsibilities for algorithmic errors (O'Neil, 2016: 26). The second obstacle is the market-driven approach to financing social media. Offering free services, these platforms are funded through trading users' personal data. Data surveillance is therefore a part of their business model, and the platforms tend to be inclusive and unfiltered rather than exclusive and curated. Many of their data harvesting practices are not compliant with privacy laws or are not properly limited, regulated, and sanctioned mainly because data privacy laws fail to encapsulate the fast-paced development of surveillance mechanisms. In this sense, social media platforms are borderline illegal by design, and the concept of digital criminology should be expanded to include such practices that are not explicitly regulated by legislation but can be considered non-compliant with basic principles of international law.

The idea of curated social media would help prevent the problematic use of artificial intelligence for moderating or policing the platforms and help eradicate algorithmic wars. The so-called YouTube dramas are just one of the examples of such wars which motivate users to engage in disputes that spark between social media influencers. These social media conflicts range from localised ones such as YouTube dramas to large-scale cyber wars that can affect greater populations. In both cases, the effects can be devastating and affect human lives – whether through cyberbullying that might result in suicides or through mass manipulation that can lead to radical changes in world politics. Only systemic solutions could help reduce or eliminate the uncontrollable reproduction of hate on the internet, but small-scale changes and raising awareness of the problem can help

[5] Burke, E. (2021, January 4th) „We're stuck in a swamp of online content – how do we get out?, Silicone Republic, available at: www.siliconrepublic.com/business/curation-is-king-online-content-ai-ethics (Accessed: 29.03.2022.)

reach a consensus on how to approach the problem. In Serbia, the lack of adequate laws that target social media related cybercrimes is complemented by the lack of understanding of what constitutes these crimes.

4. Defining Online Hate Speech as a Type of Cybercrime in Serbia

As a candidate for EU membership, Serbia is continuously working on the alignment of its legislation to the normative framework of the European Union. It has ratified documents such as the International Covenant on Civil and Political Rights[6], the International Convention on the Elimination of All Forms of Racial Discrimination[7], and the European Convention on Human Rights[8]. It has also adopted a new digital privacy law that complies with the General Data Protection Regulation[9]. Additionally, Serbian laws that are regulating hate speech include the Criminal Code of the Republic of Serbia[10], Law on the Prohibition of Discrimination[11], Media Law[12], Law on Electronic Media[13], and the Law on Electronic Communications[14]. To efficiently deal with cybercrime, Serbia's Ministry of Internal Affairs has established the Department for Cybercrime and defined criminal offences that fall within this scope[15]. It specifically mentions hate speech and describes it as the action of the spread of ethnic, racial, religious, and other forms of hate online. The legal framework offers robust protection against hate speech, but implementation remains weak. Some of the most important reasons for this is the lack of knowledge of regulations and their implementation, a low level of media literacy and the lack of commonly accepted definitions of hate speech, and the internet, therefore, becomes a place where anyone can say anything without the feeling of responsibility (Ivanović, Ranđelović, 2019: 49). The low number of reported cases is disproportionate

[6] International Covenant on Civil and Political Rights, available at: treaties.un.org (Accessed: 29.03.2022.)

[7] International Convention on the Elimination of All Forms of Racial Discrimination, available at: www.ohchr.org (Accessed: 29.03.2022.)

[8] European Convention on Human Rights, available at: www.echr.coe.int (Accessed: 29.03.2022.)

[9] General Data Protection Regulation, available at: gdpr-info.eu, (Accessed: 29.03.2022.)

[10] Kirivični zakonik, Official Gazette RS, No. 85/2005, 88/2005, amendment 107/2005, amendment 111/2009, 121/2012, 104/2013, 108/2014, 94/2016, 35/2019

[11] Zakon o zabrani diskriminacije, Official Gazette RS, No. 22/2009, 52/2021

[12] Zakon o javnom informisanju i medijima, Official Gazette RS, No. 83/2014, 58/2015 and 12/2016

[13] Zakon o elektronskim medijima, Official Gazette RS, No. 83/2014, 6/2016, 129/2021

[14] Zakon o elektronskim komunikacijama, Official Gazette RS, No. 44/2010, 60/2013, decision US, 62/2014 and 95/2018

[15] Krivična dela koja obuhvata visokotehnološki kriminal, available at: http://mup.gov.rs/ (Accessed: 29.03.2022.)

to the quantities of various forms of hate speech throughout social media platforms. Digital ethnography research shows that the so-called „hate comments“ are normalised on social media while the reporting of hate speech and other types of cybercrimes is rare.

Users of social media typically rely on platforms to regulate content, and they consider as socially, morally, and legally acceptable any behaviour that is allowed on these platforms. The guidelines and policies of these platforms are often controversial because they tend to perpetuate and „encourage“ harassment, abuse and hate speech (Konikoff, 2021: 502). On the other hand, social media strive to improve their algorithms to detect and censor hate speech more efficiently while at the same time also contributing to the spread of hate speech trends in an algorithmic way. In other words, hating someone online can be, at the same time, socially unacceptable and trendy. High-profile cases of cyberhate have unveiled these mechanisms of algorithmic reproduction of similar content and dispersion of harmful social media trends. Namely, deaths of social media influencers in Serbia and the rest of the world are often placed in the context of some form of illegal online activity from dangerous challenges on TikTok to online drug dealing and cyber hate that could potentially motivate murders or suicides and other crimes. While platforms themselves with their algorithms and ethical guidelines cannot and should not be the only mechanism for the prevention of hate speech, they remain to be so even though their primary purpose is earning revenue and not crime prevention.

A higher percentage of reported hate speech cases on social media and other platforms could make the legal mechanism for the protection of individuals against it much more efficient. Explaining what constitutes the crime of hate speech to the public can help achieve this goal. Cases that get a lot of media attention can be used to raise awareness of these problems, but they can also be misused or misinterpreted and relativise the meaning of cybercrime in general. The case of the Serbian influencer Kristina Đukic who was found dead in December 2021 proves this point. The story attracted a lot of media attention, but it is questionable whether the message about cyberbullying, hate speech and other crimes was conveyed correctly. Quickly after her death, the media started reporting about the possible case of suicide linked to cyberbullying[16]. Some of them reported that this case simply uncovered the world of digital violence to wider audiences[17]. However, the cause of death was not yet established at the time and the link between the death and

[16] Vreme (2021, December, 10th) „Mediji i internet: smrt Kristine Kike Đukić“, available at: https://www.vreme.com/kolumna/4577751/, (Accessed: 29.03.2022.)

[17] Tuvić, S. (2021, December 10th) „Smrt mlade jutjuberke je razotkrila stravičan svet pretnji i vređanja među mladima na internetu“, available at: https://www.euronews.rs/srbija/drustvo/28590/smrt-mlade-jutjuberke-kike-razotkrila-stravican-svet-pretnji-i-vredanja-medu-mladima-na-internetu/vest (Accessed: 29.03.2022.)

cyberbullying was not grounded. Digital ethnography analysis of her profiles has shown that her posts on YouTube, Instagram and TikTok have been generating large quantities of hate comments prior to her death and that they have started to reduce rapidly after her death when positive comments have become dominant.

By analysing thousands of comments under her YouTube, TikTok and Instagram posts, it was possible to monitor how narratives have changed over time, prior to and after her sudden death. Before her death, her posts featured large numbers of hate comments mainly focusing on her physical look. Since she has openly talked about her aesthetic surgeries, she has started receiving comments about her breast implants and lip enhancers. Misogynous comments were mainly coming from male viewers who criticised her for her unnatural look. The algorithmic nature of social media comments is mirrored in the way they are sorted. Namely, algorithms are organising the visibility of comments in a similar way to the content. Users view comments in order of their popularity rather than in relation to the time when they were originally posted. The top-rated comments are the ones seen first because they drive more engagement. Comments are also often mimicking each other, because they tend to be repetitive just like the content itself. This is how trends in comments emerge, and the reason why many of them utilise hate speech is the potential to attract more reactions. Kristina’s videos generated thousands of negative comments such as „you are ugly“, „You are plastic“, „you should die“ etc. When she was found dead in her apartment in Belgrade in December 2021, these comments were replaced by „RIP“, „RIP angel“, „I already miss you“ and similar. Additionally, many negative comments started to disappear as a result of raised awareness of the possible relation of hate speech to cyberbullying that leads to death.

YouTube dramas are also stimulated by the algorithms because they favour content that gets the most engagement. Kristina’s drama with another influencer Bogdan Ilic has been brought up in the media as one of the reasons why Kristina allegedly committed suicide. The case has reached the special department of cybercrimes and Bogdan was invited for a hearing because of the allegations in the media, even though the YouTube drama occurred two years before the death and that accusations were made before the cause of death was established. The media in Serbia started writing about the case of suicide and speculated that the YouTube drama with Bogdan Ilic has led to suicide before it was established whether her death can be categorised as suicide or not. By drawing attention to Bogdan, the media have contributed to the rise of another flood of hate comments that were targeting him. These comments blossomed all over his social media profiles, namely on YouTube and Instagram where users were writing comments such as „You should be ashamed of yourself“, „are you happy because you killed her“, „murderer“ etc. Since many traditional media have started accusing Bogdan of his hate speech in the drama itself while implying that he popularised hate comments directed at her, he was then transformed into a new victim of cyberbullying.

As both social media and traditional media are striving to get the attention of their audiences. They depend on algorithmic politics of visibility and they create narratives that provoke an immediate reaction which is most commonly hate speech. In the context of influencer dramas such as this one, the commentary is almost reduced to „tribal" binary thinking which implies that the audiences are divided into camps, and hate is reproduced in the *circulus vitiosus* fashion: response to hate tends to be more hate (Janjić, 2020). The result of such online wars is often the „cancel culture" (Norris, 2021) or the tendency to outlaw specific behaviour, certain personalities, or even entire cultures. In the context of the war in Ukraine, these algorithmic tribalistic divisions are palpable in the battle for the narratives against Russia or, conversely, against Ukraine.

These examples clearly show how hate speech reproduces itself and contributes to polarisation on a wide range of social issues including hate speech itself. Additionally, hate speech has grown to become an important part of political discourse (Wasilewski, 2019) that is algorithmically enforced (Darius & Stephany, 2019). On a local level, the proliferation of hate speech can be reduced by educating the population about various forms of cybercrimes. On the global level, the only way to tackle this problem is to redefine the rules and regulations on how social media platforms are being used, and how algorithms are utilised to organise visibility and perform other tasks. The lack of international as well as national laws that would regulate artificial intelligence and digital services remains to be the most important problem in Serbia and worldwide.

5. Conclusion: Digital Criminology Beyond Traditional Cybercrimes

With potentially unlimited possibilities for privacy infringements, manipulation, and violation of families of human rights, contemporary communication technologies need to be systematically reassessed and regulated. Thinking about digital criminology in a broad sense means considering many of the normalised practices as illegal or at least borderline illegal. Hate speech, child pornography and revenge pornography are considered typical cybercrimes along with computer scams, identity theft and information theft or dissemination of computer viruses. Online privacy is particularly important and regulated by the General Data Protection Regulation[18] which is applicable to EU countries but also throughout the world because of its wider applications through platforms that are used globally. But because the fast-developing technologies of the interconnected world are increasingly endangering the whole network of human rights, the concept of digital criminology should be expanded to include different potentially harmful practices including the increasingly intrusive surveillance and data mining practices. Namely, cybercrime should be understood in a much broader sense.

The entire sphere of the internet remains to be underregulated and therefore allows for practices that are not in line with general principles of international law and democracy. Algorithmically reproduced hate speech is only one aspect of the criminal nature of online platforms that are systematically infringing human rights and encouraging totalitarian practices. A more regulated, curated internet would enable better control of the information flow online. AIA[19] and DSA[20] as new legal regulations are some of the recent attempts to provide a normative framework and protect interlinked human rights against unethical practices, including the right to privacy, the right of thought, consciousness, and religion, the right to freedom of opinion and expression, the right to work and to free choice of employment, the right to rest and leisure and many others listed in the Universal Declaration of Human Rights[21] and other documents.

The role of algorithms in the reproduction of hate on social media and other online platforms is crucial because they organise visibility by setting the parameters that can even predetermine the narratives. They are enclosing the users in filter bubbles and leaving them exposed to the content that caters to their preferences, habits, and desires. As a result, they, at least potentially, have a profound impact on how people interact, form opinions, and express themselves online. Their capacity for manipulation remains unaddressed because curation of the content is against the preferences of the digital market itself which profits from data transactions and therefore requires high engagement rates. Understanding criminal aspects of algorithmic social sorting and adequate legal

[18] General Data Protection Regulation, available at: gdpr-info.eu, (Accessed: 29.03.2022.)

[19] Artificial Intelligence Act, available at: eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, (Accessed 29. 03. 2022.)

[20] Digital Services Act, available at: eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A825%3AFIN, (Accessed 29. 03. 2022.)

[21] Universal Declaration of Human Rights, available at: www.un.org/en/about-us/universal-declaration-of-human-rights, (Accessed: 29.03.2022.)

regulation of this field is crucial for finding systemic solutions for resolving problems of hate speech online.

References

Baruh, L, Popescu, M. (2015) Big data analytics and the limits of privacy self-management, New Media & Society 19(4), pp. 579-596, doi:10.1177/1461444815614001

Bromell, D. (2022) Regulating Free Speech in Digital Age: Hate, Harm and the Limits of Censorship. New York: Springer

Castano-Pulgarin, S. A., Suarez-Betancur, N., Vega, L. M. T., Lopez, H. M. H. (2021) Internet, social media and online hate speech. Systematic review. Aggression and Violent Behaviour 58, pp. 1- 7, doi: 10.1016/j.avb.2021.101608

Cobbe, J. (2021) Algorithmic Censorship by Social Platforms: Power and Resistance, Philosophy & Technology 34, pp. 739-766, doi: 10.1007/s13347-020-00429-0

Darius, P., Stephany, F. (2019). „Hashjacking" the Debate: Polarisation Strategies of Germany's Political Far-Right on Twitter, Social Informatics 11864, pp. 298-308, doi: 10.1007/978-3-030-34971-4_21

Darmody, A., Zwick, D. (2020) Manipulate to empower: Hyper-relevance and the contradictions of marketing in the age of surveillance capitalism, Big Data & Society, pp. 1-12, doi: 10.1177/2053951720904112

Duffy, E. B., Pinch, A., Sannon, S., Sawey, M. (2021). The Nested Precarities of Creative Labour on Social Media, Social Media + Society 7(2) doi: 10.1177/20563051211021368.

Dyer-Witheford, N. (2015). Cyber - Proletariat: Global Labour in the Digital Wortex. London:Pluto Press.

Foucault, M. (1995) Discipline and Punish; The Birth of Prison. New York: Vintage Books

Fuchs, C. (2014). Digital Labour and Karl Marx. New York and Oxon: Routledge.

Gerbaudo, P. (2012). Tweets and the Streets: Social Media and Contemporary Activism. London and New York: Pluto Press.

Ivanović, A. R., Ranđelović, D. (2019) Sankcionisanje govora mržnje na internetu prema nacionalnoj regulativi Republike Srbije, Arhiv za pravne i društvene nauke 1(1), pp. 49-61

Janjić, S. (2020) Govor mržnje na portalima i društvenim mrežama u Srbiji. Novi Sad: Novosadska novinarska škola

Konikoff, D. (2021) Gatekeepers of toxicity: Reconceptualizing Twitter's abuse and hate speech policies, Policy Internet, 13, 502– 521, doi: 10.1002/poi3.265

Kovács, G., Alonso, P. & Saini, R. (2021) Challenges of Hate Speech Detection in Social Media, SN Computer Science 2 (95) doi: 10.1007/s42979-021-00457-3

Kirivični zakonik, Official Gazette RS, No. 85/2005, 88/2005, amendment 107/2005, amendment 111/2009, 121/2012, 104/2013, 108/2014, 94/2016, 35/2019

McStay, A. (2020) Emotional AI, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy, Big Data & Society, pp. 1-12, doi: 10.1177/2053951720904386

Montefalcon, M. D., Padilla, J. R., Paulino, J., Go, J., Rodriguez, R. L., Imperial, J. M. (2021) Understanding Facial Expression Expressing Hate from Online Short form Videos, 5th International Conference on E-Society, E-Education and E-Technology, August 2021, pp. 201 – 207

Norris, P. (2021) Cancel Culture: Myth or Reality?, Political Studies, doi: 10.1177/00323217211037023

O'Neil, C. (2016) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishing

O'Regan, C. (2018) Hate Speech Online: an (Intractable) Contemporary Challenge? Current Legal Problems 71(1), pp. 403-429, doi: 10.1093/clp/cuy012

Pasquale, F. (2015) The algorithmic self, The Hedgehog Review 17(1)

Richards, N. M. (2013) The Dangers of Surveillance, Harvard Law Review 126, pp. 1934-1965

Sobkowicz, P. (2020) Whither Now, Opinion Modellers?, Frontiers in Physics, doi: 10.3389/fphy.2020.587009

Stepanovic, I. (2020) From Traditional Bureaucracy to Algorithmic Data Processing: How Digital Technology Transforms the Concept of Surveillance, Socijalna Misao 99(2), 69-85

Stratton G, Powell A and Cameron R (2017) Crime and Justice in Digital Society: Towards a 'Digital Criminology'?, International Journal for Crime, Justice and Social Democracy 6(2), pp. 17- 33, doi: 10.5204/ijcjsd.v6i2.355

Sued, E. G., Castillo-Gonzalez, M. C., Pedraza, C., Flores-Marquez, D., Alamo, S., Oritz, M., Lugo, N., Arroyo, R. E. (2021) Vernacular Visibility and Algorithmic Resistance in the Public Expression of Latin American Feminism, Media International Australia, doi: 10.1177/1329878X211067571

Turner, F. (2019) Machine Politics: The rise of the internet and a new age of authoritarianism, Harper's Magazine, January 2019, fredturner.stanford.edu/wp-content/uploads/Turner-Machine-Politics-Harpers-Magazine-2019-01.pdf (Accessed: 29.03.2022.)

Ward, K. (2018) Social networks, the 2016 US presidential election, and Kantian ethics: applying the categorical imperative to Cambridge Analytica's behavioral microtargeting, Journal of Media Ethics, 33:3, 133-148, DOI: 10.1080/23736992.2018.1477047

Wasilewski, K. (2019) Hate speech and identity politics. An intercultural communication perspective, Przeglad Europejski 3, pp. 175-187

Zakon o zabrani diskriminacije, Official Gazette RS, No. 22/2009, 52/2021

Zakon o javnom informisanju i medijima, Official Gazette RS, No. 83/2014, 58/2015 and 12/2016

Zakon o elektronskim medijima, Official Gazette RS, No. 83/2014, 6/2016, 129/2021

Zakon o elektronskim komunikacijama, Official Gazette RS, No. 44/2010, 60/2013, decision US, 62/2014 and 95/2018

Online sources

Artificial Intelligence Act, eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, (Accessed: 29.03.2022.)

Burke, E. (2021, January 4th) We're stuck in a swamp of online content – how do we get out?, Silicone Republic, available at: www.siliconrepublic.com/business/curation-is-king-online-content-ai-ethics (Accessed: 29.03.2022.)

Digital Services Act, available at: eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A825%3AFIN, (Accessed: 29.03.2022.)

European Convention on Human Rights, available at: www.echr.coe.int (Accessed: 29.03.2022.)

General Data Protection Regulation, available at: gdpr.eu/tag/gdpr/ (Accessed: 29.03.2022.)

International Covenant on Civil and Political Rights, available at: treaties.un.org (Accessed: 29.03.2022.)

International Convention on the Elimination of All Forms of Racial Discrimination, available at: www.ohchr.org (Accessed: 29.03.2022.)

MUP RS Krivična dela koja obuhvata visokotehnološki criminal, available at: mup.gov.rs, (Accessed: 29.03.2022.)

The Domesday Book, available at: domesdaybook.co.uk (Accessed: 29.03.2022.)

Tuvić, S. (2021, December 10th) „Smrt mlade jutjuberke je razotkrila stravičan svet pretnji i vređanja među mladima na internetu", available at: https://www.euronews.rs/srbija/drustvo/28590/smrt-mlade-jutjuberke-kike-razotkrila-stravican-svet-pretnji-i-vredanja-medu-mladima-na-internetu/vest (Accessed: 29.03.2022.)

Universal Declaration of Human Rights, available at: www.un.org/en/about-us/universal-declaration-of-human-rights, (Accessed: 29.03.2022.)

Vreme (2021, December, 10th) Mediji i internet: smrt Kristine Kike Đukić, available at: https://www.vreme.com/kolumna/4577751/, (Accessed: 29.03.2022.)

***

Ivana Stepanović[*]

ALGORITAMSKA REPRODUKCIJA GOVORA MRŽNJE NA DRUŠTVENIM MREŽAMA

Apstrakt: Inspirisane simulacionim modelima zasnovanim na agentima, društvene mreže imaju za cilj da menjaju ljudsko ponašanje tako što favorizuju sadržaj koji hrani podsvesne želje i stimuliše proizvodnju i potrošnju. Drame na Jutjubu, izazovi na Tiktoku i komentari ili haštagovi inspirisani mržnjom podstaknuti su marketinškim strategijama samih društvenih mreža koje su motivisane profitom i rangiraju sadržaje prema na osnovu broja interakcija. Algoritmi društvenih mreža bolje pozicioniraju klikabilne koji su repetitivni i tako dovodi do hiperprodukcije govora mržnje samo zato što to motiviše korisnike da reaguju. Oni regulišu

[*] Ivana Stepanovic, PhD, Institute of Criminological & Sociological Research, Serbia and Institute of Advanced Studies Kőszeg, Hungary, ivana.stepanovic@iask.hu

vidljivost na mrežama ali uvek su podvrgnuti logici tržišta a ne etičkim principima. Rapidnu proizvodnju sadržaja na društvenim mrežama nemoguće je kontrolisati ili cenzurisati u realnom vremenu i zakoni se primenjuju naknadno, nakon što je već došlo do nekog oblika visokotehnološkog kriminala. U Srbiji nedostaju specijalni zakoni i mehanizmi prevencije visokotehnološkog kriminala, ali takođe nema svesti o tome da takozvani „negativni komentari“ mogu biti interpretirani kao krivično delo koje može biti sankcionisano. Ovaj rad istražuje na koji način „prozumacija“ (sjedinjeni proces proizvodnje i konzumacije) dovodi do proliferacije mržnje na društvenim mrežama i ukazuje na značaj podizanja svesti o opasnostima krivičnih dela povezanim sa govorom mržnje.

Ključne reči: algoritmi, govor mržnje, visokotehnološki kriminal, prozumacija, društvene mreže