



<https://doi.org/10.22363/2313-2337-2024-28-3-584-603>

EDN: HYOTYH


Research Article / Научная статья

The use of Artificial Intelligence technologies by internet platforms for the purposes of censorship

Evgeni I. Diskin  

National Research University “Higher School of Economics” (HSE),

Moscow, Russian Federation

 ediskin@hse.ru

Abstract. The issue of abuse in regulating content moderation on internet platforms using artificial intelligence technologies is relatively new in legal science and practice. Regulatory frameworks in this area are still evolving, and enforcement practices have yet to be fully established. The author employs formal-legal, comparative-legal, historical methods and legal modeling to analyze the negative consequences of using software systems with artificial intelligence elements for user content moderation. By examining various technological solutions utilized by internet platforms for data collection and processing, the article highlights a potential threat to citizens’ rights to access and share information if legal relations governing content moderation with artificial intelligence are not significantly enhanced. It explores evidence suggesting that in the absence of regulatory constraints transparency requirements, internet platforms may begin censorship by removing content, based on their own criteria, even if it does not violate laws or platform guidelines. The author argues that unchecked actions by internet platforms could restrict individuals and political entities from expressing their views, posing a significant threat to democratic principles. By examining Russian, EU and US laws alongside current trends in internet platforms operations, the article concludes that the existing legal frameworks are inadequate and calls for legislative oversight and control over technologies used for content moderation, algorithms, and artificial intelligence applications.

Key words: moderation, internet platforms, information law, censorship, digital rights, Artificial Intelligence

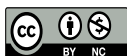
Conflict of interest. The author declares no conflict of interest.

Funding. The article was prepared during the research within the framework of the Basic Research Program of the National Research University “Higher School of Economics” (HSE).

Received: 11th August 2023

Accepted: 15th July 2024

© Diskin E.I., 2024




This work is licensed under a Creative Commons Attribution 4.0 International License
<https://creativecommons.org/licenses/by-nc/4.0/legalcode>

For citation:

Diskin, E.I. (2024) The use of Artificial Intelligence technologies by internet platforms for the purposes of censorship. *RUDN Journal of Law*. 28 (3), 584–603. <https://doi.org/10.22363/2313-2337-2024-28-3-584-603>

Применение технологий искусственного интеллекта при осуществлении цензуры со стороны интернет-платформ

Е.И. Дискин  

Национальный исследовательский университет «Высшая школа экономики»
(НИУ ВШЭ), г. Москва, Российская Федерация
 ediskin@hse.ru

Аннотация. Проблема злоупотреблений при осуществлении модерации интернет-платформ с помощью технологий искусственного интеллекта в значительной степени нова для юридической науки и практики, нормативное регулирование данной сферы только формируется, правоприменительная практика пока не сложилась. Автор, используя формально-юридический, сравнительно-правовой, исторический методы и метод правового моделирования, анализирует негативные последствия применения программных комплексов с элементами искусственного интеллекта для осуществления модерации пользовательского контента. Исследуя некоторые технологические решения, с помощью которых интернет-платформы собирают и обрабатывают значительные массивы разнообразных данных о пользователях, автор указывает на потенциальную угрозу правам граждан на поиск и распространение информации в том случае, если правоотношения в области модерации контента с помощью искусственного интеллекта не будут законодательно регулироваться на качественно новом уровне. Исследуется ряд фактов, которые подтверждают тезис о том, что интернет-платформы, в условиях отсутствия нормативных ограничений, требований к прозрачности их деятельности, контроля со стороны государства и общества начинают осуществлять цензуру путем удаления мнений, высказываний и информации, которые могут ими рассматриваться как нежелательные, при том, что данные высказывания не нарушают законодательства и правила платформы. Автор приходит к выводу о том, что интернет-платформы наращивают возможности по контролю за высказываниями, играющими существенную роль в поддержании информационного баланса, т.е. возможности различных политических сил доносить свою точку зрения до широкой общественности, что в свою очередь представляет собой серьезную угрозу демократическому правопорядку. По результатам анализа российского законодательства и законодательства стран ЕС и США, с учетом сложившихся тенденций в работе интернет-платформ, сделан вывод о несовершенстве существующего правового регулирования, необходимости законодательного ограничения и контроля подобных технологий, применительно к осуществлению модерации, использованию алгоритмов и искусственного интеллекта.

Ключевые слова: модерация, интернет-платформы, законодательство об информации, цензура, права интернет-пользователей, искусственный интеллект

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Финансирование. Статья подготовлена в ходе исследования в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ).

Поступила в редакцию: 11 августа 2023 г.

Принята к печати: 15 июля 2024 г.

Для цитирования:

Дискин Е.И. Применение технологий искусственного интеллекта при осуществлении цензуры со стороны интернет-платформ // RUDN Journal of Law. 2024. Т. 28. № 3. С. 584–603. <https://doi.org/10.22363/2313-2337-2024-28-3-584-603>

Introduction

The challenges of moderating internet platforms, particularly the issues related to the removal or restricting the unwanted content and user messages are becoming increasingly prominent as digitalization plays a more significant role in society. It is evident that the removal of unsolicited emails (spam), limitation of interactions with users displaying aggressive behavior, sending threatening messages, engaging in offensive conduct, distributing pornography or sharing violent materials is essential to prevent internet platforms from descending into an environment of chaos and lawlessness. For example, the distribution of pornographic content to the vast audiences present on internet platforms remains a significant unresolved issue (Sorban, 2023). However, moderation can be misused when legitimate restrictions are replaced by censorship, leading to the removal of messages and content deemed undesirable by internet platforms for ideological and other non-justifiable reasons (Diskin, 2023). Some authors refer to this phenomenon as the “privatization of censorship” highlighting how censorship in contemporary times is not solely enforced by governmental bodies but also by private companies (Monti, 2020). Regrettably, due to the challenging geopolitical climate, one of the most vulnerable groups of users comprises Russian citizens, mass media and state bodies, who are de facto subjected to widespread persecution on foreign internet platforms. Moreover, prominent foreign scientific journals have published articles on this subject, advocating for increased censorship of the Russian-speaking segment of the internet under the pretext that Russian media outlets fuel hatred and disseminate false information (Filatova-Bilous, 2023). These facts are alarming and underscore the need for measures to regulate legal relations related to internet platform moderation processes and identification of violations. This article examines the current trajectory of Artificial Intelligence (hereinafter referred to as AI) development as a tool for moderating internet platforms. The work is aimed at enhancing the depth and quality of discussions on this matter, highlighting the potential risks posed by this technology, and evaluating the regulatory practices in this field in the Russian Federation and other countries. The article presents recommendations for amending existing legislation.

Artificial intelligence as a moderation tool

Addressing the tasks associated with content moderation on internet platforms requires a substantial allocation of human and computational resources, particularly in managing vast amounts of data. In 2020, Mark Zuckerberg

disclosed that Facebook moderators¹ handle over 3 million requests for content removal daily². Specialized software systems leveraging artificial intelligence are pivotal in processing large volumes of information in real-time to identify and eliminate prohibited content on the platform. While some internet platforms use proprietary software unavailable to third parties, there are also commercial solutions tailored for such purposes. In 2023, Microsoft Corporation introduced a software tool named Azure AI Content Safety³, designed to enhance content moderation processes. This AI-powered software system automates the moderation process, which includes removing, restricting, and monitoring information posted, shared, and commented on by users of internet platforms, e-commerce sites and gaming services. Microsoft Corporation emphasizes that moderation is currently critical for any form of commercial activity on the Web due to the necessity to meet user expectations regarding the creation of a so-called “safe online space” and compliance with regulatory requirements in preventing the dissemination of prohibited information⁴.

Moderation and the right to freedom of speech

It is noteworthy that in the official statement released by Microsoft Corporation regarding the launch of the Azure AI Content Safety software system does not specify how this software complex will contribute to upholding the legal rights of internet users. This reflects a concerning trend of diminishing opportunities for individuals to exercise the rights enshrined in Article 19 of the Universal Declaration of Human Rights⁵, which grants everyone the freedom to seek, receive and disseminate information and ideas through any means regardless of frontiers. Article 19 of the International Covenant on Civil and Political Rights further specifies this right. Paragraph 2 of this Article stipulates that limitations on freedom of speech are permissible only if they are prescribed by law and necessary to: a) respect the rights and reputation of others;

¹ Product of Meta Inc., which is recognized as an extremist organization in the Russian Federation. On March 21, 2022, the Tverskoy District Court of Moscow satisfied a lawsuit filed by the Prosecutor General’s Office of the Russian Federation and recognized the activity of the social network Facebook and Instagram, owned by Meta, as extremist, banning its operation in Russia.

² John Koetsier. Report: Facebook* Makes 300,000 Content Moderation Mistakes Every Day. Available at: <https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/?sh=221b87854d03> [Accessed 16th January 2024]. * On March 21, 2022, the Tverskoy District Court of Moscow satisfied a lawsuit filed by the Prosecutor General’s Office of the Russian Federation and recognized the activity of the social network Facebook and Instagram, owned by Meta, as extremist, banning its operation in Russia.

³ Federico Zarfati. Introducing Azure AI Content Safety: Helping Organizations to Maintain Safe Online Spaces. Available at: <https://techcommunity.microsoft.com/t5/ai-cognitive-services-blog/introducing-azure-ai-content-safety-helping-organizations-to/ba-p/3825744> [Accessed 16th January 2024].

⁴ Federico Zarfati. Introducing Azure AI Content Safety: Helping Organizations to Maintain Safe Online Spaces. Available at: <https://techcommunity.microsoft.com/t5/ai-cognitive-services-blog/introducing-azure-ai-content-safety-helping-organizations-to/ba-p/3825744> [Accessed 16th January 2024].

⁵ Universal Declaration of Human Rights. Adopted by General Assembly resolution 217 A (III) of 10.12.1948. Available at: https://www.un.org/ru/documents/decl_conv/declarations/declhr.shtml [Accessed 16th January 2024].

b) protect national security, public order, health, or morals⁶. These fundamental principles are reflected in Article 29 of the Constitution of the Russian Federation, which guarantees freedom of speech and thought while also prohibiting propaganda or agitation that incites social, racial, national or religious hatred and enmity, as well as promoting social, racial, national, religious, or linguistic superiority. The proximity of these provisions within Article 29 emphasizes that freedom of expression must always be subject to reasonable restriction of that freedom. Additionally, we note that the prohibition of censorship is established by paragraph 5 of this Article.

Addressing unfair practices by online platforms

In our opinion, the issue of what are the genuine, effective guarantees protecting the rights of internet users in the context of moderation implementation is key to determining the future development of the internet, especially with the widespread integration of AI across various professional domains. This is particularly critical in sectors where information is processed, analyzed and disseminated, notably on internet platforms.

The lack of effective safeguards and mechanisms to protect users' rights, particularly evident when individuals from one country use internet platforms established and overseen by another country (a particularly acute issue for Russia), highlights the necessity of focusing on combating negative practices that infringe upon individuals' rights to access information, freely express their opinions by posting messages, and distribute information through legal means. This problem is not unique to Russia; conflicts often arise when internet platforms refuse to address such issues, leading to their suspension as seen when platform X (formerly Twitter) was blocked in Nigeria in 2021 (Chiroma & Sule, 2022). Temporary suspensions of the platform were also observed in Turkey⁷ in 2023. In 2021, Uzbekistan implemented the blocking of several internet platforms (Twitter, VK, Tik-Tok) due to non-compliance with personal data legislation⁸.

While not endorsing the blocking of internet resources as a universal approach to addressing legal violations, it is important to highlight that in certain instances, the misconduct of internet platforms may compel authorities to implement restrictive measures, even of a severe nature. A case in point is the blocking of the LinkedIn, which failed to comply with the stipulation outlined in paragraph 5 Article 18 of Federal Law No. 152-FZ dated 27.07.2006 On Personal Data, requiring the transfer of equipment for

⁶ International Covenant on Civil and Political Rights. Adopted by General Assembly resolution 2200 A (XXI) of 16.12.1966. Available at: https://www.un.org/ru/documents/decl_conv/conventions/pactpol.shtml [Accessed 16th January 2024].

⁷ Adam Satariano. Twitter Was Blocked in Turkey, Internet-Monitoring Group Says. Available at: <https://www.nytimes.com/2023/02/08/world/europe/turkey-earthquake-twitter-blocked.html> [Accessed 11th January 2024].

⁸ Catherine Putz. Uzbekistan Unblocks Twitter, TikTok Still Restricted. Available at: <https://thediplomat.com/2022/08/uzbekistan-unblocks-twitter-tiktok-still-restricted/> [Accessed 11th January 2024].

processing the personal data of Russian users to Russia (Sherstobitov, Neverov & Barinova, 2018).

Current foreign policy landscape necessitates a comprehensive delineation of unfair practices carried out by internet platforms in the field of moderation. This entails enhancing and implementing regulations that ensure adequate safeguards and redress mechanisms. It also involves widespread integration of challenging illegal actions of internet platforms into legal frameworks, thereby increasing the volume of appeals lodged by individuals and organizations seeking protection of their rights through judicial procedure.

Artificial intelligence as a tool for moderating internet platforms

The use of artificial intelligence in software systems for moderating functions is continually advancing and holds the potential to enforce control over expressions on the web, disseminated information, and content in an absolute manner, as depicted by the concept of “Leviathans of Cyberspace” (Sirichit, 2015). Unlike traditional internet content moderation methods, which rely on human moderators to decide on controversial cases based on platform rules and restrictions, the theoretical capabilities of specialized AI systems for moderation are virtually limitless (constrained primarily by the costs of establishing data centers). In essence, specialized software packages utilizing AI as a basic decision-making tool, such as Microsoft Azure AI Content Safety⁹ and Google Content ID represent a significant advancement in this field. They are capable of filtering the content users publish on any internet platform almost in real time. However, such an increase in the speed of moderation, where machines make all decisions regarding the exercise of one of the fundamental individual human rights on which the democratic state is based, poses unprecedented risks for the further development and functioning of society and the state. It is worth noting that this conclusion is supported by studies that document a significant proportion of errors made by such software systems. Specifically, according to Facebook’s internal assessments¹⁰ at least 10% of moderation decisions are inaccurate¹¹. This situation is further complicated by the fact that in order to reduce costs, internet platforms often outsource such tasks, resulting in final moderation decisions being made by low-skilled staff with inadequate training and skills (Roberts, 2019). The Covid-19 epidemic exacerbated these trends, as in the absence of secure jobs, a majority of moderation decisions were left to machines.

⁹ Note: AI tools integrated into the content moderation systems of major internet platforms like Google, Facebook*, and X Corp. (formerly Twitter) typically do not have specific brand names as they are not marketed as stand-alone products. However, the author’s conclusions regarding Azure AI Content Moderation are applicable to these platforms as well. * On March 21, 2022, the Tverskoy District Court of Moscow satisfied a lawsuit filed by the Prosecutor General’s Office of the Russian Federation and recognized the activity of the social network Facebook and Instagram, owned by Meta, as extremist, banning its operation in Russia.

¹⁰ On March 21, 2022, the Tverskoy District Court of Moscow satisfied a lawsuit filed by the Prosecutor General’s Office of the Russian Federation and recognized the activity of the social network Facebook and Instagram, owned by Meta, as extremist, banning its operation in Russia.

¹¹ Barret, P.M. (2020) Who Moderates the Social Media Giants? A Call to End Outsourcing. Available at: https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_content_moderation_report_final_version [Accessed 16th January 2024].

X Corp. acknowledged in March 2020 that “AI may not grasp the context of statements and can make errors”¹². This acknowledgment came in response to instances where AI algorithms erroneously blocked users who had not violated any rules¹³. It is important to highlight that aside from the direct implications of AI use in moderation, such as restricting access to and dissemination of information, there is also the “echo chamber” effect which refers to narrowing the information flows (Kovaleva, et al., 2022). However, this particular topic is not being explored in this article due to its specific nature.

Legislation and its impact on AI implementation for moderation purposes

The trends described above highlight a context that needs further examination. The increased use of AI software for moderation is not solely a product of advancement and competition among internet platforms but is also influenced by pressure from various states, in response to a series of terrorist attacks involving the live and widespread dissemination of violence (Crosset & Dupont, 2022). One significant legislative measure that emerged in reaction to these events is the Act on Improving Law Enforcement in Social Networks (Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken, NetzDG) of the Federal Republic of Germany dated September 1, 2017. This act mandates that online platforms remove illegal content within 24 hours¹⁴.

In this context, the tone of the discussion on legislative concepts for moderating internet platforms in the foreign academic spheres is particularly noteworthy. Foreign researchers such as Jonathan Zittrain¹⁵, Tim Wu (Wu, 2019), Rachael Griffin (Griffin, 2022), Evelyn Douek (Douek, 2021) advocate for a shift “from an era of rights to an era of “public health”, where the value of expressions is weighted against the risks of their dissemination”. This implies transforming internet platforms into a “safe space”, necessitating more restricted freedom of speech than current legislation requirements, thereby departing from the traditional model of constitutional protection of the right to freedom of speech.

We acknowledge that researchers referred to the example of the 2016 US presidential election and purportedly inadequate measures to address “Russian intervention” as supporting grounds for such interventions. Consequently, Russian

¹² Vijaya Gadde, Matt Derella. An update on our continuity strategy during COVID-19. Available at: https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19 [Accessed 11th January 2024].

¹³ John Koetsier. Facebook* Deleting Coronavirus Posts, Leading to Charges of Censorship. Available at: <https://www.forbes.com/sites/johnkoetsier/2020/03/17/facebook-deleting-coronavirus-posts-leading-to-charges-of-censorship/?sh=3eb598255962> [Accessed 11th January 2024]. *On March 21, 2022, the Tverskoy District Court of Moscow satisfied a lawsuit filed by the Prosecutor General’s Office of the Russian Federation and recognized the activity of the social network Facebook and Instagram, owned by Meta, as extremist, banning its operation in Russia.

¹⁴ Heldt, A.P. (2018) Reading between the Lines and the Numbers: An Analysis of the First NetzDG Reports. Available at: <https://papers.ssrn.com/abstract=3413677> [Accessed 16 January 2024].

¹⁵ Zittrain, J.L. (2019) Three Eras of Digital Governance. <https://doi.org/10.2139/ssrn.3458435> [Accessed: 15th January 2024].

internet users have long been under scrutiny by internet platforms and are likely to be targeted for prosecution with the help of AI tools primarily. A considerable portion of the discourse on internet regulation in the US and the EU is focused on directing the efforts of the internet platforms to counter “Russian disinformation and interference in elections” (Saurwein & Spencer-Smith, 2020; Francois & Lin, 2021; Geissler, et al., 2023; Weintraub & Valdivia, 2020; Hodgins, 2022). Notably, the implementation of AI systems has escalated in response to these requirements, creating a favorable environment for their unchecked deployment and potentially infringing on the rights of Russian users. The above-mentioned German law NetzDG, was also enacted under the influence of this discourse (Rumyantsev, 2019).

It is important to understand that the vast amount of data currently held by the largest internet platforms, including user location, keystroke data, photos and videos, visited websites and used apps, steps taken, and pulse rate, will enable the creation of digital user models in the very near future (within 1-2 years). These models will encompass a psychological profile, social connections map and increasingly accurate predictions of human behavior.

With corporations like Google¹⁶ and Apple¹⁷ currently developing headphones with brain electroencephalogram functions, also known as neurointerfaces, that will eventually enable them to gather data on brain activity and link it to information regarding a person’s whereabouts, activities such as writing, reading, and speaking (captured by voice assistants in smartphones), the concept of “freedom of thought” takes on new dimensions. It is evolving into the challenge of safeguarding human thought from literal control by digital means.

The scenario outlined above is not merely hypothetical; applications that generate a simulated, non-existent information environment using AI already exist. For instance, the Parallel Live Simulator app, accessible on Apple iOS and Google Android operating systems, fabricates the impression that an individual is live-streaming, attracting simultaneous comments by thousands of users, receiving likes, and eliciting reactions to videos¹⁸. While the app’s objective is to incite vanity, the technology employed by this application could readily be expanded to fabricate interactive illusions for users of any internet platform¹⁹.

It is worth noting that from its inception, internet platforms have strived to maximize the collection of user data. Sean Parker, the former vice president of Facebook,²⁰ shared his insights about the underlying motives of social network architecture when it was established in 2005: “How do we consume as much of your time and conscious attention as possible? That means we need to give you a little dopamine hit once in a while, because

¹⁶ Steven Levy. This startup wants to get in your ears and watch your brain. *Wired*. Available at: <https://www.wired.com/story/nextsense-wants-to-get-in-your-ears-and-watch-your-brain/> [Accessed 16th January 2024].

¹⁷ Apple’s new patent shows AirPods with brain wave-detecting sensors. *The Times of India*. Available at: <https://timesofindia.indiatimes.com/gadgets-news/apples-new-patent-shows-airpods-with-brain-wave-detecting-sensors/articleshow/102278175.cms> [Accessed 16th January 2024].

¹⁸ Apple AppStore. Parallel Live Simulator. Available at: <https://apps.apple.com/us/app/parallel-live-simulator/id1534165497> [Accessed 15 January 2024].

¹⁹ Note: relevant threats will also be covered in other sections of the article.

²⁰ Product of Meta Inc., which is recognized as an extremist organization in the Russian Federation.

someone liked or commented on a photo or a post or whatever. And that's going to get you to contribute more content, and that's going to get you more likes and comments. All this gives users a little dopamine hit, luring them into a social-validation feedback loop"²¹. It is crucial to acknowledge that in 2005, Internet platforms did not possess the advanced tools like AI-based software that are available today. However, according to Mark Zuckerberg in 2024, AI will be "the main priority for Facebook investment"²².

Threats to freedom of expression from Internet platforms

To further delve into these trends, it is essential to consider examples of digital technologies that will soon be amalgamated into synergistic digital systems within internet platforms to create an "AI-clone" of any user²³.

Recognition of political orientation by facial expression is feasible; research conducted by M. Kosinski confirms that AI can determine a person's political orientation by analyzing a single photo that clearly displays facial expressions. The AI achieves an average accuracy rate of 72% in identifying political orientation from a photo, surpassing the accuracy of a completed questionnaire with a minimum 100 questions (which provides a 66% accuracy rate) (Kosinski, 2021). This outcome required processing around 1 million online dating questionnaires. It is noteworthy that AI can also identify other personal as well as social characteristics that significantly influence an individual's privacy. Particularly, as early as 2018, the same researcher disclosed that the algorithm could accurately predict the sexual orientation of men in 81% of cases and women in 71% of cases through analyzing a single photo. With at least five photos, the algorithm's accuracy increased to 91% for men and 83% for women. These findings were based on a sample of 35,326 images (Wang & Kosinski, 2018).

Estimating the general psychological characteristics of an individual using openly available social network data poses no challenge for IT-specialists. For instance, a study by D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft dating back to 2011 affirmed the capability to construct a user profile with five key parameters, achieving accuracy of not less than 0.88 on a scale of 1.5 units based solely on three indicators of any user account: the number of followers, the number of accounts followed, and the number of additions to thematic lists (Quercia, et al., 2011). Likewise, AI can draw similar

²¹ Mike Allen. Sean Parker unloads on Facebook*: "God only knows what it's doing to our children's brains". Available at: <https://www.axios.com/2017/12/15/sean-parker-unloads-on-facebook-god-only-knows-what-its-doing-to-our-childrens-brains-1513306792> [Accessed 16th January 2024]. *On March 21, 2022, the Tverskoy District Court of Moscow satisfied a lawsuit filed by the Prosecutor General's Office of the Russian Federation and recognized the activity of the social network Facebook and Instagram, owned by Meta, as extremist, banning its operation in Russia.

²² Meta's AI Investments Ramp Up as Reality Labs Loses Billions. Available at: <https://www.pymnts.com/meta/2023/meta-threads-grabs-nearly-100m-users-just-three-months-after-launch/> [Accessed 15 January 2024].

²³ Yoon, D. (2023) AI clones made from user data pose uncanny risks. 4 June 2023. The Conversation. Available at: <http://theconversation.com/ai-clones-made-from-user-data-pose-uncanny-risks-206357> [Accessed: 15th January 2024].

conclusions about the basic psychological properties of an individual from a sufficiently clear photograph (Kachur, et al., 2020).

According to Microsoft, Azure AI Content Safety is capable of identifying four types of content that creators believe should be removed or restricted through automation. These categories include content or messages expressing hatred (hate speech), inciting violence, sexually explicit material, and information promoting self-harm. These are generally the primary types of content that all internet platforms aim to restrict and remove (Dym & Fiesler, 2020). Initially, deleting relevant information, messages and content may seem unobjectionable. However, the author suggests that this perspective is somewhat superficial. Evaluating the threats posed by such software packages requires a thorough and detailed examination of the issues surrounding freedom of speech and expression, considering the specifics of ongoing digitalization processes characterized by the enormous influence of major internet AI platforms and their widespread implementation.

Twenty years ago, Professor L. Lessig discussed the internet as a unique space where individuals engage in a different logic of behavior. He wrote that this coverage was made possible by the lack of control within the network: anyone, from any location, could publish any information for everyone, everywhere. The network allowed publishing information without censorship, editing or accountability. You could express whatever you desired, choose whether to attribute your authorship or remain anonymous, send a message across global computers, and within hours, those words would be accessible everywhere. The network eliminated the primary limitations of free speech in the physical world, the distinction between the publisher and the author. In physical realm, there exists a “vanity publication” model (where the author, rather than the publisher, covers publication costs), but only the affluent can reach a broad audience this way. For the rest of us in the physical world, our ability to express ourselves is confined to what publishers are willing to provide (Lessig, 1999).

This is an incredibly precise and vital observation, and it can certainly be expanded upon. It is essential to recognize that the internet sets a significantly different pace for exchange of information compared to the means available to humanity before the advent of World Wide Web technology. In the pre-digitalization era, characterized by print media and television, information was disseminated through evening and morning newspapers, weekly or monthly magazines, books, and scheduled television programs. Consequently, modes of expression, such as the ability to voice opinions on current matters, were constrained by the communication tools available; individuals could send letters to newspapers or television stations, but it was primarily a one-way communication. Moreover, the decision on whether to publish the received letter in the newspaper or another form of response was subject to the discretion of the editorial team, as highlighted by L. Lessig.

Another significant form of public expression was and continues to be the ability to participate in public events, gatherings, or closed group meetings. However, this aligns more with a constitutional right, such as the right to public peaceful assembly. The opportunity to speak at a large gathering with numerous participants is typically considered a privilege. The right to speak in such settings is primarily exercised by

individuals who hold roles as speakers, moderators, or members of the organizing team. While these traditional forms of public discourse are indeed important, they are somewhat reliant on external factors beyond individual control. As highlighted by L. Lessig, the internet provided a platform for nearly unrestricted speech, crucially without the constraints of context or external circumstances, and without the need to negotiate terms for publication. It is crucial to acknowledge that internet platforms not only offer the opportunity to express oneself but also facilitate the instantaneous dissemination of messages to a vast audience, reaching tens or even hundreds of millions of views within a short timeframe. Furthermore, it is important to bear in mind that online communication is multimodal, allowing the expressions that are impossible in real-life interactions; for instance, users can respond to a text message with a picture, video or link (Yushkina & Panarina, 2019). However, these circumstances have given rise to numerous negative consequences, including a decline in the quality of public discourse, the erosion of many real-world taboos, as well as the widespread proliferation of rudeness, arrogance, aggression and cyberbullying in communication (Baburin & Cheremnova, 2022; Nikishin & Galyashina, 2021; Zhumabekova, 2022; Romanov, et al., 2021).

Nonetheless, censorship by internet platforms, often justified under the guise of removing “inappropriate information”, cannot be fully justified by these challenges. While the internet does indeed host various forms of destructive content that must be addressed (including through the use of AI) for moral reasons and to adhere to legislative prohibitions such as those against pornography, the promotion of hatred or discord, and the protection of minors from content that could negatively affect their development (Titov, 2021).

There are numerous instances of censorship by internet platforms that support this argument. For example, in 2016, several Facebook employees²⁴ anonymously reported deliberate restrictions on the dissemination of information from the conservative political spectrum; the content intentionally excluded from the “Trends” section, a feature through which users could discover currently popular topics. While this section was algorithmically generated, it was susceptible to “manual” intervention, compromising the integrity of “organic” search results²⁵ that should have been free from political bias. Former Facebook moderators²⁶ Ryan Hartwig and Zach McElroy also publicly disclosed similar information, confirming instances of double standards in

²⁴ Product of Meta Inc., which is recognized as an extremist organization in the Russian Federation. On March 21, 2022, the Tverskoy District Court of Moscow satisfied a lawsuit filed by the Prosecutor General’s Office of the Russian Federation and recognized the activity of the social network Facebook and Instagram, owned by Meta, as extremist, banning its operation in Russia.

²⁵ Michael Nunez. Former Facebook* workers: We routinely suppressed conservative news. Available at: <https://gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006> [Accessed 16th January 2024]. *On March 21, 2022, the Tverskoy District Court of Moscow satisfied a lawsuit filed by the Prosecutor General’s Office of the Russian Federation and recognized the activity of the social network Facebook and Instagram, owned by Meta, as extremist, banning its operation in Russia.

²⁶ On March 21, 2022, the Tverskoy District Court of Moscow satisfied a lawsuit filed by the Prosecutor General’s Office of the Russian Federation and recognized the activity of the social network Facebook and Instagram, owned by Meta, as extremist, banning its operation in Russia.

content moderation based on the targeted political affiliation²⁷. In 2018, Twitter was exposed for utilizing censorship through the implementation of a “Shadow ban” technique. This practice involved imposing restrictions on a user’s account without their knowledge; as a result, other users could not see the affected account in their newsfeed, and messages from that account were omitted from the “Trends” section (Jaidka, Mukerjee & Lelkes, 2023). Such measure was implemented on the US President Donald Trump and Ronna McDaniel, chair of the Republican National Committee²⁸. It is noteworthy that Twitter officers continuously denied any intentional actions, referring to some software glitches²⁹. Moreover, executives from both social media platforms consistently refuted allegations of censorship, even testifying before the US Congress³⁰. Nevertheless, the internal documents disclosed by Elon Musk following the acquisition by X Corp. provide compelling evidence to support claims about the existence of a system of digital censorship by major US-based internet platforms (Diskin, 2023).

Regulation of recommendatory algorithms and moderation in Russia

Currently, legislation governing the legal relationship with recommendatory algorithms is in its early stages of development. However, efforts have recently been intensified to regulate recommendatory algorithms and moderation as key factors.

Special attention should be given to the Federal Law No. 408-FZ dated 31.07.2023 On Amendments to the Federal Law On Information, Information Technologies and Information Protection (hereinafter referred to as the *Law on Recommendatory Algorithms*)³¹. This law introduces a new concept of “recommendatory technologies” and requires information resources to publicly establish rules for their application in the Russian language. The cornerstone of the law is the introduction of liability for

²⁷ Christopher Boyle. Exclusive: Inside Facebook’s* content moderator bias with Ryan Hartwig and Zach McElroy, two Ex-Employees come forward to Project Veritas. Available at: <https://www.publishedreporter.com/2020/07/03/exclusive-inside-facebooks-content-moderator-bias-with-ryan-hartwig-and-zach-mcelroy/> [Accessed 16th January 2024]. *On March 21, 2022, the Tverskoy District Court of Moscow satisfied a lawsuit filed by the Prosecutor General’s Office of the Russian Federation and recognized the activity of the social network Facebook and Instagram, owned by Meta, as extremist, banning its operation in Russia.

²⁸ Alex Thompson. Twitter appears to have fixed “shadow ban” of prominent Republicans like the RNC chair and Trump Jr.’s spokesman. Available at: <https://www.vice.com/en/article/43paqq/twitter-is-shadow-banning-prominent-republicans-like-the-rnc-chair-and-trump-jrs-spokesman> [Accessed 16th January 2024].

²⁹ Alex Thompson. Twitter appears to have fixed “shadow ban” of prominent Republicans like the RNC chair and Trump Jr.’s spokesman. Available at: <https://www.vice.com/en/article/43paqq/twitter-is-shadow-banning-prominent-republicans-like-the-rnc-chair-and-trump-jrs-spokesman> [Accessed 16th January 2024].

³⁰ Kevin Roose, Cecilia Kang. Mark Zuckerberg Testifies on Facebook* Before Skeptical Lawmakers. Available at: <https://www.nytimes.com/2018/04/10/us/politics/zuckerberg-facebook-senate-hearing.html> [Accessed 16th January 2024]. * On March 21, 2022, the Tverskoy District Court of Moscow satisfied a lawsuit filed by the Prosecutor General’s Office of the Russian Federation and recognized the activity of the social network Facebook and Instagram, owned by Meta, as extremist, banning its operation in Russia.

³¹ The Federal Law No. 408 dated 31.07.2023 On Information, Information Technologies and Information Protection. Available at: <http://publication.pravo.gov.ru/Document/View/0001202307310021> (Accessed 15 of January 2024).

website owners, pages, information systems, and software using recommendatory technologies in violation of citizens' rights and legitimate interests, as well as providing information contrary to the law (Article 10.2-2). However, the definition of what constitutes a violation of the law, legal rights, and citizens' interests in that context remains unclear. The Federal Law No. 149 dated 27.07.2006 On Information, Information Technologies and Information Protection (hereinafter referred to as the *Information Law*) contains numerous rules on liability for various violations in information dissemination, but these rules cannot unequivocally, without doubt, and through long-term analysis be attributed to violations in the realm of recommendatory technologies. These circumstances present a challenge in prosecuting individuals, due to the provisions of paragraph 3 Article 49 of the Constitution of the Russian Federation: any reasonable doubt of guilt is construed in favor of the accused. It is important to note that Chapter 13 of the Code of Administrative Offences of the Russian Federation (hereinafter referred to as the *Code of Administrative Offences*) does not contain a special provision outlining the responsibility for violations under Article 10.2-2 of the Information Law. Furthermore, Article 13.50 of the Code of Administrative Offences does not establish liability for infringements of users' rights during moderation, although it does require social networks to report on their responses to complaints regarding the presence of unlawfully posted information.

The safeguarding of citizens' rights during moderation by internet platforms was not adequately considered during the enactment of the Recommendation Algorithms Law. However, Alexander Khinsein, the Chairman of the State Duma Committee on Information Policy, Informatization and Communication, stated that the primary goal of the bill was to prevent the dissemination of “information advantageous to third-parties, as witnessed during the previous election in the US [in 2020]”³².

It is important to highlight that the misuse of platforms during the preparation and conduct of the 2020 US presidential election was not limited to the unjustified promotion of favorable information for Democratic presidential candidate Joseph Biden. They also deliberately eliminated negative information about him, going as far as banning users who reported negative information about him (Diskin, 2023). These bans were justified under the guise of combating misinformation, some of which allegedly originated from a supposed information operation by Russian authorities³³. However, it was later revealed that the bans imposed by major internet platforms regarding the “Russian disinformation”³⁴ stemmed from an article published in the

³² Николай Козин. Как будет работать закон о рекомендательных алгоритмах. Парламентская газета. Available at: <https://www.pnp.ru/economics/kak-budet-rabotat-zakon-o-rekomendatelnykh-algoritmakh.html> [Accessed 15th January 2024].

³³ Natasha Bertrand. Hunter Biden story is Russian disinfo, dozens of former intel officials say. Available at: <https://www.politico.com/news/2020/10/19/hunter-biden-story-russian-disinfo-430276> [Accessed 15th January 2024].

³⁴ Kelsey Vlamis. Twitter's former trust and safety chief said it was a mistake to censor the Hunter Biden laptop story: 'We didn't know what to believe'. Available at: <https://www.businessinsider.com/yoel-roth-twitter-censor-hunter-biden-laptop-story-was-mistake-2022-11> [Accessed 15th January 2024].

New York Post about the contents of President Biden’s son Hunter’s laptop, which was verified as factual. The discussion on “disinformation” turned out to be part of an information operation orchestrated by the US Democratic Party headquarters³⁵. The mentioned circumstances provide insight into the significance of opposing unfair practices of internet platforms in moderation. The arbitrary deletion of information citing it as “disinformation originating from Russia”³⁶ through AI tools is unacceptable in terms of safeguarding state sovereignty and citizens’ rights to access and share information. Nevertheless, the Recommendation Algorithms Law despite needing amendments represents a crucial advancement in protecting the rights and legitimate interests of internet platform users, particularly in the realm of moderation utilizing AI technologies, by mandating transparency in the operation of recommendatory technologies.

When examining domestic legislation governing rights in the field of internet platform moderation, it is important to mention Federal Act No. 30-FZ dated 04.03.2022 On Amendments to the Federal Law On Measures to Influence Persons Involved in Violations of Fundamental Human Rights and Freedoms, Rights and Freedoms of Citizens of the Russian Federation and Article 27 of the Federal Law On the Procedure for Exiting and Entering the Russian Federation (hereinafter referred to as the *Law on Censorship Counteraction*). According to Roskomnadzor’s official communication, this law was adopted with the aim of “suppressing censorship by foreign internet companies against Russian media”³⁷. We agree with Roskomnadzor’s thesis that “Russian law cannot and should not be replaced on the territory of the country by the rules of internet companies”³⁸. However, despite the acknowledgment of censorship by foreign internet platforms against Russian legal entities, it should be noted that the Law on Censorship Counteraction does not provide users with effective measures against such censorship. Furthermore, it lacks a formal definition of censorship, which would expand upon the concept in Article 3 of the Russian Federation Law No. 2124-1 dated 27.12.1991 On Mass Media. The Law on Censorship Counteraction itself consists of three articles, contains no definitions or references to the Law on Information, and does not provide individuals with tools to protect their rights in disputes with foreign internet platforms.

³⁵ Paul Gigot, Bill McGurn, Kim Strassel. The Hunter Biden Laptop Disinformation Is Exposed. Available at: <https://www.wsj.com/podcasts/opinion-potomac-watch/the-hunter-biden-laptop-disinformation-is-exposed/4e8baf05-447c-419e-80d8-7424827c7b52> [Accessed 15th January 2024].

³⁶ Twitter. Disclosing networks of state-linked information operations. Available at: https://blog.twitter.com/en_us/topics/company/2021/disclosing-networks-of-state-linked-information-operations- [Accessed 15th January 2024].

³⁷ Roskomnadzor. A bill to counteract censorship of foreign Internet companies against Russian media is submitted to the State Duma. Available at: <https://rkn.gov.ru/news/rsoc/news73178.htm> [Accessed 15th January 2024].

³⁸ Roskomnadzor. A bill to counteract censorship of foreign Internet companies against Russian media is submitted to the State Duma. Available at: <https://rkn.gov.ru/news/rsoc/news73178.htm> [Accessed 15th January 2024].

Regulation of recommendatory algorithms and moderation in European Union

The European Union pays significant attention to regulating legal relations in the field of recommendatory algorithms and the moderation of internet platforms. On October 19, 2022, European Parliament and EU Council adopted Regulation 2022/2065 known as the Digital Services Act (DSA)³⁹. The DSA imposes an obligation to inform users about utilization of AI technologies for moderation on internet platforms (paragraph 1, Article 14). Online platforms must compile annual reports on the use of moderation tools, and the European Commission is mandated to establish an open database containing all moderation solutions on online platforms, enhancing transparency in the process (paragraph 5, Article 22). Nevertheless, there are provisions in the Act that could potentially negatively impact users. Article 19 of the DSA introduces the concept of trusted flaggers, whose responsibility is to report content that breaches the law. Online platforms are required to promptly and effectively address complaints from these individuals. Trusted flaggers will be appointed by coordinators designated by the competent authorities of EU Member States (Article 38). While it is stipulated that coordinators operate independently (Article 39), there is no clear liability framework in place for potential abuse by trusted flaggers or coordinators, thus allowing EU Member States to potentially influence the complaint mechanism through trusted flaggers. The Act does not establish specific accountability measures for ensuring impartiality or abuse of the complaints system.

Regulation of the recommendatory algorithms in USA

In the United States, where the headquarters of the largest transnational corporations-owners of internet platforms (such as X Corp., Alphabet Inc., Microsoft Corp., Meta Inc. etc.) are located there is no federal legislative system in place to protect users' rights during content moderation and the use of recommendatory algorithms. This is due to the concept of non-interference in the activities of technology companies that emerged in the 1990s. Under this concept, internet platforms can independently decide on the content that can and cannot be published to users through user agreements. Consequently, users do not have the right to challenge the decisions of internet platforms, and all disputes are expected to be resolved in accordance with the rules set out in the user agreement. This situation is reflected in Section 230 Title 47 of the Communications Decency Act 1996 (CDA 1996, hereinafter referred to as *Section 230*), which grants internet platforms the status of a “publisher” based on a well-established interpretation of the First Amendment to the United States Constitution (Bill of Rights). This interpretation provides internet platforms with similar rights as traditional media in determining their editorial policies, enabling them to publish information based solely on their editorial decisions without interference, which would be considered censorship. This legal understanding regarding the relationship between

³⁹ Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act). Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065> [Accessed 16th January 2024].

internet platforms and users is the subject of public debate and efforts to amend legislation. For example, Elon Musk, the current owner of platform X, supports the idea that such internet platforms should be considered “digital public spaces” where individuals have the right to protest and express themselves similarly to public spaces protected by the Bill of Rights⁴⁰. Musk has criticized Section 230 for allowing internet platforms to censor content, evade state control, and benefit from immunities from prosecution for moderation violations.

Efforts to repeal Section 230 are actively underway at the legislative level. By 2023, researchers estimated that there had been 84 attempts to repeal or significantly amend subsection 230 at the federal level, but none had been successful. Moreover, thirty-two states have attempted to pass their own laws aimed at repealing Section 230 at the local level. States like Florida and Texas have enacted legislation explicitly banning censorship and algorithmic discrimination (the misuse of algorithms to discriminate against users). However, both states’ laws were deemed unconstitutional by federal courts of first instance (Sinnreich, et al., 2023). As of January 2024, the constitutionality of these laws is being reviewed by the US Supreme Court⁴¹. It is important to note that there are other legislative initiatives related to the regulation of internet platforms, such as in the state of New York (Senate Bill S4531A)⁴². However, these initiatives are not within the scope of this article as they focus on improving platform moderation without including provisions to safeguard user rights.

Before conclusion

The use of AI technologies for moderating internet platforms is continually expanding and increasingly encroaching on users’ rights due to technological errors. These errors occur when AI fails to comprehend the context in which a statement or information is made or disseminated. Moreover, specific policies aimed at suppressing the spread of information on political grounds exacerbate the issue. AI grants internet platforms significant capabilities to regulate user-posted content. Instances have already occurred where actions to block socially significant information and messages, later acknowledged as “errors” by internet platforms, were entirely automatic in real-time, implemented by AI. As a result, users were prevented from sharing or disseminating information as deemed “unreliable” by these platforms⁴³.

⁴⁰ Ezra Klein. The Great Delusion Behind Twitter. Available at: <https://www.nytimes.com/2022/12/11/opinion/what-twitter-can-learn-from-quakers.html> [Accessed 16th January 2024].

⁴¹ Amy Howe. Justices request federal government’s views on Texas and Florida social-media laws. Available at: <https://www.scotusblog.com/2023/01/justices-request-federal-governments-views-on-texas-and-florida-social-media-laws/> [Accessed 16th January 2024].

⁴² New York State Senate. Senate Bill S4531A. Available at: <https://www.nysenate.gov/legislation/bills/2021/S4531> [Accessed 16th January 2024].

⁴³ Facebook* and Twitter restrict controversial New York Post story on Joe Biden. Available at: <https://www.theguardian.com/technology/2020/oct/14/facebook-twitter-new-york-post-hunter-biden> [Accessed 16th January 2024]. * On March 21, 2022, the Tverskoy District Court of Moscow satisfied a lawsuit filed by the Prosecutor General’s Office of the Russian Federation and recognized the activity of the social network Facebook and Instagram, owned by Meta, as extremist, banning its operation in Russia.

Citizens of Russia frequently find themselves targeted by the policies of internet platforms, when accounts belonging to state entities, public organizations, mass media and individuals are deleted under the guise of combating unreliable information. Since 2021, Roskomnadzor has documented 132 instances of censorship by foreign internet platforms against Russian media⁴⁴. Unfortunately, legislators' focus on safeguarding the rights of individual users is conspicuously inadequate. Despite the recognition of numerous instances of censorship, there has been no progress in developing legislation that would provide clear definitions, streamline existing legislation, and establish a framework of user rights and guarantees during the moderation process. Moreover, accusations of “disseminating disinformation” directed at state bodies, public organizations, and individuals largely go unanswered, by both the state and the affected parties, due to the absence of a robust system of rights and protections. The Law on Censorship Counteraction, enacted in 2022, has not effectively curbed censorship of Russian users, which continues to persist and escalate. In December 2022, Russian Foreign Ministry spokeswoman Maria Zakharova accused the West of orchestrating the most significant act of mass totalitarian censorship in history against Russian media⁴⁵. It is acknowledged that addressing a problem of this magnitude cannot be solved by a single piece of legislation, no matter how well crafted. However, the information Act must be complemented by a comprehensive set of regulations outlining users' rights and a system of guarantees for their enforcement. In this regard, the EU Digital Services Act serves as a model, with the exception of the provisions on trusted sources, which some perceive as a tool for indirect state censorship.

Conclusion

Russia's reaction to the emerging trends of AI-driven censorship must not solely focus on enhancing national legislation. As part of Russia's chairmanship of BRICS+, attention should be directed towards this issue. A proposal is made to establish a special ad hoc working group within the framework of BRICS+ to enhance internet regulation. This group would be tasked with developing measures to prevent discrimination against citizens of member states by internet platforms, particularly focusing on major providers as defined in the EU Digital Services Act.

References / Список литературы

- Baburin, V.V. & Cheremnova, N.A. (2022) Harassment in the information space (cyberbullying) and victimization of juveniles and minors. *Altai Law Journal*. (3), 54–59. (in Russian).
Бабурин В.В., Черемнова Н.А. Травля в информационном пространстве кибербуллинг и виктимность малолетних и несовершеннолетних лиц // Алтайский юридический вестник. 2022. № 3. С. 54–59.

⁴⁴ Roskomnadzor has identified nine cases of censorship against Russian media. Available at: <https://regnum.ru/news/3806632> [Accessed 15th January 2024].

⁴⁵ Zakharova spoke about large-scale Western censorship against Russian media. Available at: <https://regnum.ru/news/3763034> [Accessed 15th January 2024].

- Chiroma, I. & Sule, I. (2022) ‘Twitting to Suspend Twitter’ – Social Media Censorship in Nigeria: Possibilities, Realities and Legalities. *Scholars International Journal of Law, Crime and Justice*. (5), 202–210. <https://doi.org/10.36348/sijlcj.2022.v05i06.004>
- Crosset, V. & Dupont, B. (2022) Cognitive assemblages: The entangled nature of algorithmic content moderation. *Big Data & Society*. 9(2). <https://doi.org/10.1177/20539517221143361>
- Diskin, E. (2023) Elon Musk and crusade against digital censorship. *Law*. (1), 106–109. <http://doi.org/10.37239/0869-4400-2023-20-1-95-109> (in Russian).
Дискин Е.И. Илон Маск и крестовый поход против цифровой цензуры // Закон. 2023. № 1. С. 106–109. <https://doi.org/10.37239/0869-4400-2023-20-1-95-109>
- Douek, E. (2021) Governing online speech: From ‘posts-as-trumps’ to proportionality and probability. *Columbia Law Review*. 121(3), 759–833.
- Dym, B. & Fiesler, C. (2020) Social Norm Vulnerability and its Consequences for Privacy and Safety in an Online Community. *Proceedings of the ACM on Human-Computer Interaction*. 4(CSCW2), 1–24. <https://doi.org/10.1145/3415226>
- Filatova-Bilous, N. (2023) Content moderation in times of war: testing state and self-regulation, contract and human rights law in search of optimal solutions. *International Journal of Law and Information Technology*. 31(1), 46–74. <https://doi.org/10.1093/ijlit/eaad015>.
- Francois, C. & Lin, H. (2021) The strategic surprise of Russian information operations on social media in 2016 in the United States: mapping a blind spot. *Journal of Cyber Policy*. 6(1), 9–30. <https://doi.org/10.1080/23738871.2021.1950196>
- Geissler, D., Bär, D., Pröllochs, N. & Feuerriegel, S. (2023) Russian propaganda on social media during the 2022 invasion of Ukraine. *EPJ Data Science*. 12(35). <https://doi.org/10.1140/epjds/s13688-023-00414-5>
- Griffin, R. (2022) New school speech regulation as a regulatory strategy against hate speech on social media: The case of Germany’s NetzDG. *Telecommunications Policy*. 46(9), 102411. doi:10.1016/j.telpol.2022.102411
- Hodgins, J.M. (2022) *Countering Russian Subversion during Federal Elections in Canada in the Era of Social Media and Fake News*. ITSS Verona Magazine. 1(1). Режим доступа: https://www.researchgate.net/publication/361532403_Policy_Paper_Countering_Russian_Subversion_during_Federal_Elections_in_Canada_in_the_Era_of_Social_Media_and_Fake_News [Accessed 16 January 2024].
- Jaidka, K., Mukerjee, S. & Lelkes, Y. (2023) Silenced on social media: the gatekeeping functions of shadowbans in the American Twitterverse. *Journal of Communication*. 73(2), 163–178. <https://doi.org/10.1038/s41598-020-65358-6>
- Kachur, A., Osin, E., Davydov, D., Shutilov, K. & Novokshonov, A. (2020) Assessing the Big Five personality traits using real-life static facial images. *Scientific Reports*. (10), 8487. <https://doi.org/10.1038/s41598-020-65358-6>
- Kosinski, Mi. (2021) Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports*. 11(100). <https://doi.org/10.1038/s41598-020-79310-1>
- Kovaleva, N.N., Anisimova, A.S., Tugusheva, Y.M. & Danilova, M.A. (2022) Artificial Intelligence and Social Media: Self-Regulation and Government Control. *European Proceedings of Social and Behavioural Sciences*. State and Law in the Context of Modern Challenges. <https://doi.org/10.15405/epsbs.2022.01.56>
- Lessig, L. (1999) *Code and other laws of cyberspace*. 1st edition. New York, USA., Basic Books Inc. Режим доступа: <https://lessig.org/images/resources/1999-Code.pdf> [Accessed 16th January 2024].
- Monti, M. (2020) The EU Code of Practice on Disinformation and the Risk of the Privatization of Censorship. In: Giusti, S., & Piras, E. (eds.). *Democracy and Fake News: Information Manipulation and Post-Truth Politics*. 1st ed. London, Routledge. pp. 214–225. <https://doi.org/10.4324/9781003037385>

- Nikishin, V.D. & Galyashina, E.I. (2021) *Destructive Speech Behavior in the Digital Environment: Factors that Determine the Negative Impact on the User's Worldview*. Lex Russica. 74(6), 79–94. <https://doi.org/10.17803/1729-5920.2021.175.6.079-094> (in Russian).
Никишин В.Д., Галышина Е.И. Деструктивное речевое поведение в цифровой среде: факторы, детерминирующие негативное воздействие на мировоззрение пользователя // Lex Russica (Русский закон). 2021. № 6. С. 79–94. <https://doi.org/10.17803/1729-5920.2021.175.6.079-094>
- Quercia, D., Kosinski, M., Stillwell, D. & Crowcroft, J. (2011) Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, Boston, MA, USA. October 2011. pp. 180–185. <https://doi.org/10.1109/PASSAT/SocialCom.2011.26>
- Roberts, S.T. (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media*. Illustrated edition. New Haven, Yale University Press. <https://doi.org/10.2307/j.ctvhrcz0v>.
- Romanov, A.A., Levashova, A.V., Sidenko, K.O. & Mikhailova, E.A. (2021) *The essence of bullying as a social and legal phenomenon*. Matters of Russian and International Law. 11(1A), 98–106. <https://doi.org/10.34670/AR.2020.59.72.014> (in Russian).
Романов А.А., Левашова А.В., Сиденко К.О., Михайлова Е.А. Сущность буллинга как социально-правового явления // Вопросы российского и международного права. 2021. № 11. С. 98–106. <https://doi.org/10.34670/AR.2020.59.72.014>
- Rumyantsev, A. (2019) The German law on social networks: regulatory fixing of technological straggling. *Sravnitel'noe konstitucionnoe obozrenie*. 3(130), 27–53. <https://doi.org/10.21128/1812-7126-2019-3-27-53>. (in Russian).
Румянцев А.Г. Немецкий закон о социальных сетях: нормативное закрепление технологического отставания // Сравнительное конституционное обозрение. 2019. № 3(130). С. 27–53. <https://doi.org/10.21128/1812-7126-2019-3-27-53>
- Saurwein, F. & Spencer-Smith, C. (2020) Combating Disinformation on Social Media: Multilevel Governance and Distributed Accountability in Europe. *Digital Journalism*. 8(6), 820–841. <https://doi.org/10.1080/21670811.2020.1765401>
- Sherstobitov, A., Neverov, K. & Barinova, E. (2018) Roskomnadzor vs. strategic and institutional limits of good public governance (case of blockage of online-resources in Russia). *South-Russian Journal of Social Sciences*. 19(3), 163–176. <https://doi.org/10.31429/26190567-19-3-163-176> (in Russian).
Шерстобитов А.С., Неверов К.А., Баринова Е.А. Роскомнадзор против. Стратегические и институциональные ограничения качества публичного управления (на опыте блокировок онлайн-ресурсов в России) // Южно-российский журнал социальных наук. 2018. Т. 19. № 3. С. 163–176. <https://doi.org/10.31429/26190567-19-3-163-176>
- Sinnreich, A., Sanchez, M., Perry, N. & Aufderheide, P. (2023) Performative Media Policy: Section 230's Evolution from Regulatory Statute to Loyalty Oath. *Communication Law and Policy*. 27(3–4), 167–186. <https://doi.org/10.1080/10811680.2022.2136472>
- Sirichit, M. (2015) Censorship by Intermediary and Moral Rights: Strengthening Authors' Control Over the Online Expressions Through the Right of Respect and Integrity. *Journal of Law, Technology and Public Policy and Methaya Sirichit*. 1(3), 54–159. Режим доступа: <https://www.semanticscholar.org/paper/Censorship-by-Intermediary-and-Moral-Rights%3A-Over-Sirichit/fe9c2543d07630482816c6657b3f8ffe70d353af> [Accessed 16th January 2024].
- Sorbán, K. (2023) An elephant in the room—EU policy gaps in the regulation of moderating illegal sexual content on video-sharing platforms. *International Journal of Law and Information Technology*. 31(3), 171–185. <https://doi.org/10.1093/ijlit/eaad024>

- Titov, A.A. (2021) Novelities in the Russian legislation regulating the counteraction to Internet censorship. *Bulletin of the University of the Prosecutor's Office of the Russian Federation*. 3(83), 76–79. (in Russian).
Титов А.А. Новеллы в российском законодательстве, регламентирующие противодействие интернет-цензуре // Вестник Университета прокуратуры Российской Федерации. 2021. № 3(83). С. 76–79.
- Wang, Y. & Kosinski, M. (2018) Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*. 114(2), 246–257. <https://doi.org/10.1037/pspa0000098>
- Weintraub, E.L. & Valdivia, C.A. (2020) Strike and Share: Combatting Foreign Influence Campaigns on Social Media. *Ohio State Technology Law Journal*. 16(2), 701–721.
- Wu, T. (2019) Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems. *Columbia Law Review*. (119), 2001–2028. Режим доступа: https://scholarship.law.columbia.edu/faculty_scholarship/2598 [Accessed 16th January 2024].
- Yushkina, N.A. & Panarina, M.A. (2019) Features of the discursive environment as a source for creating meaning in online communication (using the example of social networks). *Digital Sociology*. 2(2), 25–33. <https://doi.org/10.26425/2658-347X-2019-2-25-33> (in Russian).
Юшкина Н.А., Панарина М.А. Особенности дискурсивной среды как источник создания смысла в онлайн-коммуникации (на примере социальных сетей) // Цифровая социология. 2019. № 2(2). С. 25–33.
- Zhumabekova, K. (2022) Topical issues of protecting children from cyberbullying. *Journal of actual problems of jurisprudence*. 104(4), 98–103. <https://doi.org/10.26577/JAPJ.2022.v104.i4.011>

Сведения об авторе:

Дискин Евгений Иосифович – кандидат юридических наук, ведущий научный сотрудник Международной лаборатории цифровой трансформации в государственном управлении ИГМУ, Национальный исследовательский университет «Высшая школа экономики», 101000, Российская Федерация, г. Москва, ул. Мясницкая, д. 20

ORCID: 0000-0001-9259-9820; ResearcherID: ediskin; SPIN-код: 4364-0208

e-mail: ediskin@hse.ru

About the author:

Evgenii I. Diskin – Candidate of Legal Sciences, Leading research fellow of the International Laboratory of Digital Transformation in the State Administration of IGMU, National Research University “Higher School of Economics”, Russian Federation, 101000, Myasnitskaya 20, Moscow, Russian Federation

ORCID: 0000-0001-9259-9820; ResearcherID: ediskin; SPIN-code: 4364-0208

e-mail: ediskin@hse.ru