Вестник РУДН. Серия: Юридические науки 2024 Т. 28. № 3. 584–603
RUDN JOURNAL OF LAW. ISSN 2313-2337 (print), ISSN 2408-9001 (online) http://journals.rudn.ru/law

https://doi.org/10.22363/2313-2337-2024-28-3-584-603
EDN: HYOTYH

Научная статья / Research Article

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

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 constrains 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

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

Введение

Проблемы модерации интернет-платформ, то есть вопросы, связанные с удалением или ограничением распространения нежелательного контента и сообщений пользователей, обостряются по мере увеличения значимости цифровизации в жизни человека. Очевидно, что удаление нежелательных массовых рассылок (спама), ограничение возможностей взаимодействия с другими пользователями тех, кто ведет себя агрессивно, рассылает сообщения с угрозами жизни и здоровью, оскорбляет с использованием ненормативной лексики, распространяет порнографию и материалы, содержащие сцены насилия правомерно, в целом необходимо, чтобы интернет-платформы не были пространством полной анархии и беззакония. Так, например, распространение порнографической продукции среди пользователей интернет-платформ с огромной аудиторией в десятки и сотни миллионов человек до сих пор представляет собой серьезную проблему, которая не нашла однозначного решения (Sorbán, 2023). Тем не менее, модерация может быть предметом злоупотребления, когда разумные и законные ограничения заменяются цензурой, т.е. удалением сообщений и контента, которые считаются интернет-платформами нежелательными по идеологическим и другим причинам (Diskin, 2023). Некоторые авторы называют данный процесс «приватизацией цензуры» в том смысле, что в современных условиях цензура исходит не только от государственных органов, но и от частных компаний (Monti, 2020). К сожалению, ввиду тяжелой геополитической ситуации одной из наиболее уязвимых групп пользователей являются российские граждане, СМИ, государственные органы, которые де-факто подвергаются массовому преследованию на иностранных интернет-платформах. Более того, в ведущих зарубежных научных журналах отмечены публикации по данной тематике, с требованием максимального ужесточения цензуры русскоязычного сегмента сети «Интернет» (далее – Сеть) под предлогом того, что российские СМИ разжигают ненависть и публикуют только недостоверную информацию (Filatova-Bilous, 2023). Данные факты не могут не вызывать тревогу и требуют усилий, направленных на регулирование правоотношений, связанных с процессами модерации интернет-платформ, идентификацию практики злоупотреблений с их стороны. Данная публикация анализирует текущий вектор развития Искусственного интеллекта (далее – ИИ) как средства модерации интернет-платформ, она нацелена на повышение качества и широты дискурса по данной проблематике, привлечение внимания к тому, какие угрозы несет применение данной технологии, оценки опыта нормотворчества в данной сфере в России и в других странах. Кроме того, в статье содержатся предложения по изменению действующего законодательства.

Искусственный интеллект как средство модерации

Решение задач по модерации контента интернет-платформами требует привлечения значительных людских и вычислительных ресурсов, так как подобные задачи связаны с обработкой огромных объемов информации. В 2020 году Марк Цукерберг

сообщал о том, что модераторы Facebook[1] проверяют более 3 миллионов запросов на удаление контента в день[2]. Огромную роль в том, что интернет-платформы успевают обрабатывать данные объемы информации в режиме реального времени играют специализированные программные комплексы, использующие элементы ИИ для обнаружения и удаления запрещенной правилами платформы информации. Многие интернет-платформы используют программные комплексы, которые недоступны для продажи третьим лицам, однако появляются и коммерческие решения для решения таких задач. Так, в 2023 году корпорация Microsoft объявила о запуске в промышленную эксплуатацию программного комплекса Azure AI Content Safety[3]. Упомянутый программный комплекс с помощью ИИ автоматизирует процесс модерации, то есть удаления, ограничения и контроля над информацией публикуемой, распространяемой и комментируемой пользователями интернет-платформ, площадок электронной коммерции и игровых сервисов. Обращает внимание заявление Microsoft Corp. о том, что в настоящее время процесс модерации является критически важным для любых форм ведения коммерческой деятельности с использованием Сети ввиду необходимости удовлетворять ожидания пользователей в части создания так называемого «безопасного онлайн-пространства» и соблюдения регуляторных требований по недопущению распространения запрещенной информации[4].

Модерация и право на свободу слова

Обращает на себя внимание тот факт, что в официальном заявлении Microsoft Corp. по поводу запуска программного комплекса Azure AI Content Safety отсутствует указание на то, как данный программный комплекс будет способствовать соблюдению законных прав пользователей Сети. Данное обстоятельство отражает сложившиеся в последнее время тенденции последовательного сокращения возможностей для реализации прав, закрепленных статьей 19 Всеобщей декларации прав человека[5], которая гласит, что каждому гарантируется свобода искать, получать и распространять информацию и идеи любыми средствами и независимо от государственных границ. Статья 19 Международного пакта о гражданских и политических правах конкретизирует данное право. Часть 2 упомянутой статьи содержит условие, согласно которому допустимы некоторые ограничения права на свободу слова — они должны быть установлены законом и являться необходимыми: а) для уважения

[1] Продукт корпорации Meta Inc. признанной экстремистской организацией в Российской Федерации. 21.03.2022 г. Тверской районный суд г. Москвы удовлетворил иск Генпрокуратуры РФ и признал деятельность соцсети Facebook и Instagram, принадлежащей Meta, экстремистской, запретив ее работу в России.

[2] John Koetsier. Report: Facebook* Makes 300,000 Content Moderation Mistakes Every Day. Режим доступа: https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/?sh=221b87854d03 (дата обращения: 16.01.2024). *21.03.2022 г. Тверской районный суд г. Москвы удовлетворил иск Генпрокуратуры РФ и признал деятельность соцсети Facebook и Instagram, принадлежащей Meta, экстремистской, запретив ее работу в России.

[3] Federico Zarfati. Introducing Azure AI Content Safety: Helping Organizations to Maintain Safe Online Spaces. Режим доступа: https://techcommunity.microsoft.com/t5/ai-cognitive-services-blog/introducing-azure-ai-content-safety-helping-organizations-to/ba-p/3825744 (дата обращения: 16.01.2024).

[4] Federico Zarfati. Introducing Azure AI Content Safety: Helping Organizations to Maintain Safe Online Spaces. Режим доступа: https://techcommunity.microsoft.com/t5/ai-cognitive-services-blog/introducing-azure-ai-content-safety-helping-organizations-to/ba-p/3825744 (дата обращения: 16.01.2024).

[5] Всеобщая декларация прав человека. Принята резолюцией 217 А (III) Генеральной Ассамблеи ООН от 10.12.1948 г. Режим доступа: https://www.un.org/ru/documents/decl_conv/declarations/declhr.shtml (дата обращения: 16.01.2024 г.).

прав и репутации других лиц; b) для охраны государственной безопасности, общественного порядка, здоровья или нравственности населения[6]. Эти основополагающие принципы отражены в статье 29 Конституции Российской Федерации, которая гарантирует свободу слова и мысли, но одновременно запрещает пропаганду или агитацию, возбуждающие социальную, расовую, национальную или религиозную ненависть и вражду, а также пропаганду социального, расового, национального, религиозного или языкового превосходства. Соседство этих положений внутри статьи 29 подчеркивает, что свободе слова всегда должно соответствовать разумное ограничение этой свободы. Отдельно отметим, что запрет цензуры установлен частью 5 данной статьи.

Фиксация недобросовестных практик интернет-платформ

На наш взгляд, вопрос о том, каковы реальные, фактически действующие гарантии защиты прав пользователей сети «Интернет» (далее — Сеть) при осуществлении модерации, является одним из ключевых для определения того, как будет развиваться Сеть в ближайшие годы в контексте массового внедрения технологий искусственного интеллекта (далее – ИИ) в различных сферах профессиональной деятельности. Особенно это касается тех сфер, где происходит обработка, анализ и распространение разного рода информации, в первую очередь это касается интернет-платформ.

Отсутствие эффективных гарантий и способов защиты прав пользователей, особенно в тех случаях, когда граждане одного государства пользуются интернет-платформами, созданными и контролируемыми в другом государстве (для России данный вопрос стоит особо остро), опосредует необходимость повышенного внимания к негативным практикам, связанным с нарушением прав граждан на доступ к информации, возможности свободно высказывать свои мысли путем размещения сообщений, распространения информации любым законным способом. С подобной проблемой сталкиваются не только в России, и нередко конфликт, который возникает из-за нежелания интернет-платформ прекратить такие практики, приводит к их блокировкам – например, платформа X (ранее Twitter) была заблокирована в Нигерии в 2021 году (Chiroma & Sule, 2022). В 2023 году краткосрочные блокировки данной платформы отмечались в Турции[7]. В 2021 году блокировки ряда интернет-платформ (Twitter, VK, Tik-Tok) в связи с неисполнением законодательства о персональных данных вводились в Узбекистане[8].

Не приветствуя блокировки Сетевых ресурсов как универсальный метод борьбы с нарушением применимого законодательства, отметим, что в ряде случаев злоупотребления со стороны интернет-платформ не оставляют иного выбора, как прибегать к ограничениям, в том числе радикального характера. В качестве примера таких вынужденных мер можно привести блокировку интернет-платформы

[6] Международный пакт о гражданских и политических правах. Принят резолюцией 2200 A (XXI) Генеральной Ассамблеи ООН от 16.12.1966 г. URL: https://www.un.org/ru/documents/decl_conv/conventions/pactpol.shtml (дата обращения: 16.01.2024 г.).

[7] Adam Satariano. Twitter Was Blocked in Turkey, Internet-Monitoring Group Says. Режим доступа: https://www.nytimes.com/2023/02/08/world/europe/turkey-earthquake-twitter-blocked.html (дата обращения: 11.01.2024).

[8] Catherine Putz. Uzbekistan Unblocks Twitter, TikTok Still Restricted. Режим доступа: https://thediplomat.com/2022/08/uzbekistan-unblocks-twitter-tiktok-still-restricted/ (дата обращения: 11.01.2024).

LinkedIn, нарушившей требования о переносе в Россию оборудования для обработки персональных данных российских пользователей, установленного ч. 5 ст. 18 Федерального закона от 27.07.2006 № 152-ФЗ «О персональных данных» (Sherstobitov, Neverov & Barinova, 2018).

Отметим, что в сложившихся внешнеполитических условиях необходимо подробное описание недобросовестных практик интернет-платформ в области модерации, совершенствование и внедрение через нормотворчество соответствующих гарантий и средств защиты, широкого внедрения в юридическую практику обжалования неправомерных действий интернет-платформ, в том числе путем увеличения числа обращений граждан и организаций за защитой своих прав в судебном порядке.

Искусственный интеллект как средство модерации интернет-платформ

Искусственный интеллект, используемый в программных комплексах для осуществления функций модерации постоянно совершенствуется и потенциально может сделать контроль за высказываниями в Сети, распространяемой информацией и контентом абсолютным и тотальным со стороны «левиафанов киберпространства» (Sirichit, 2015). В отличие от методов модерации Сетевого контента, которые требуют принятия человеком (модератором) решения по конкретному спорному случаю применения правил и ограничений той или иной интернет-платформы, теоретическая пиковая производительность специализированных систем ИИ в качестве средства модерации практически не ограничена (точнее ограничена лишь уровнем затрат на построение центров обработки данных). Иными словами, специализированные программные комплексы, использующие ИИ в качестве базового инструмента принятия решений, такие как Microsoft Azure AI Content Safety[9] и Google Content ID, смогут практически в режиме реального времени фильтровать то, что публикуют пользователи на любой интернет-платформе. Такой рост скорости модерации, когда все решения относительно возможности осуществлять одно из базовых индивидуальных прав человека, на котором зиждется демократическое государственное устройство, принимает машина, несет в себе беспрецедентный риск для дальнейшего развития и функционирования общества и государства. Отметим, что данный вывод подтверждается исследованиями, которые фиксируют значительную долю ошибок, которые допускают подобные программные комплексы. В частности, только по внутренним оценкам Facebook[10] не менее 10% решений о модерации являются ошибочными[11] (Barret, 2020). Осложняет ситуацию и то, что ради снижения затрат, интернет-платформы занимаются аутсорсингом, в результате чего окончательные решения о модерации осуществляются низкоквалифицированным персоналом, не

[9] Прим.: инструменты ИИ, встроенные в программный комплекс модерации контента Google, Facebook*, X (бывш. Twitter) и других крупных интернет-платформ обычно не имеют определенного фирменного обозначения, т.к. не предлагаются как самостоятельный рыночный продукт, однако умозаключения автора по поводу Azure AI Content Moderation в полной мере относятся и к ним. * 21.03.2022 г. Тверской районный суд г. Москвы удовлетворил иск Генпрокуратуры РФ и признал деятельность соцсети Facebook и Instagram, принадлежащей Meta, экстремистской, запретив ее работу в России.

[10] Продукт корпорации Meta Inc. признанной экстремистской организацией в Российской Федерации. 21.03.2022 г. Тверской районный суд г. Москвы удовлетворил иск Генпрокуратуры РФ и признал деятельность соцсети Facebook и Instagram, принадлежащей Meta, экстремистской, запретив ее работу в России.

[11] Barret, P.M. (2020) Who Moderates the Social Media Giants? A Call to End Outsourcing. Режим доступа: https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_content_moderation_report_final_version (дата обращения: 16.01.2024).

имеющим достаточных навыков и подготовки (Roberts, 2019). Эпидемия COVID-19 ухудшила описанные тенденции, так как в условиях недоступности защищенных рабочих мест, большинство решений о модерации принималось машинами самостоятельно. Как признали в X в марте 2020 года, «ИИ может не понимать контекст высказываний и делать ошибки»[12]. Данное признание было сделано под давлением фактов о том, что алгоритмы ИИ приняли ошибочные решения о многочисленных блокировках пользователей, которые не совершали нарушений[13]. Отметим, что помимо прямых последствий применения ИИ для осуществления модерации в виде ограничений на доступ и распространение информации отмечается эффект сужения информационных потоков человека, что получило название «эхо камеры» (echo chamber) (Kovaleva, et al., 2022), однако данный вопрос не входит в предмет рассмотрения данной статьи в силу его специфики.

Законодательство и его влияние на внедрение ИИ для целей модерации

Описанные выше тенденции требуют изучения определенного контекста, который следует рассмотреть подробнее. В частности, расширение использования программных комплексов с элементами ИИ для целей модерации является не только результатом внутреннего развития и конкуренции интернет-платформ, но и результатом давления со стороны ряда государств, продиктованном в том числе реакцией на ряд террористических актов, сопровождавшихся трансляцией случаев насилия в прямом эфире и массовым распространением такого контента (Crosset & Dupont, 2022). В частности, одним из знаковых законодательных актов такого рода, появившихся в результате реакции на упомянутые события, является Закон ФРГ от 1 сентября 2017 года «О мерах в отношении социальных сетей» (нем. Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken, NetzDG), который требует от интернет-платформ удаления противоправного контента в течение 24 часов[14].

В этом контексте обращает на себя внимание тональность дискуссии по поводу законодательных концепций модерации интернет-платформ, которая ведется в иностранной научной среде. В частности, такими зарубежными исследователями, как Jonathan Zittrain[15], Tim Wu (Wu, 2019), Rachael Griffin (Griffin, 2022), Evelyn Douek (Douek, 2021), провозглашается переход «от эпохи прав к эпохе «публичного здоровья», где ценность высказываний сопоставляется с рисками их распространения». За этим скрывается идея необходимости превращения интернет-платформ в некую «безопасную среду», что требует от них больших ограничений высказываний, чем требует текущее законодательство, а соответственно и отход от традиционной модели конституционной защиты права на свободу слова.

[12] Vijaya Gadde, Matt Derella. An update on our continuity strategy during COVID-19. Режим доступа: https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19 (дата обращения: 11.01.2024).

[13] John Koetsier. Facebook* Deleting Coronavirus Posts, Leading To Charges Of Censorship. Режим доступа: https://www.forbes.com/sites/johnkoetsier/2020/03/17/facebook-deleting-coronavirus-posts-leading-to-charges-of-censorship/?sh=3eb598255962 (дата обращения: 11.01.2024). *21.03.2022 г. Тверской районный суд г. Москвы удовлетворил иск Генпрокуратуры РФ и признал деятельность соцсети Facebook и Instagram, принадлежащей Meta, экстремистской, запретив ее работу в России.

[14] Heldt, A.P. (2018) Reading between the Lines and the Numbers: An Analysis of the First NetzDG Reports. Режим доступа: https://papers.ssrn.com/abstract=3413677 (дата обращения: 15.01.2024).

[15] Zittrain, J.L. (2019) Three Eras of Digital Governance. https://doi.org/10.2139/ssrn.3458435 (дата обращения: 15.01.2024).

Отметим, что в обоснование таких мер упомянутые исследователи ссылались на пример выборов президента США 2016 года и якобы недостаточные меры по борьбе с «российским вмешательством» в них. Таким образом, российские пользователи Сети давно находятся «под прицелом» интернет-платформ и будут преследоваться с помощью инструментов ИИ в первую очередь – значительная часть дискурса о регулировании Сети в США и ЕС сосредоточена на том, чтобы направить усилия интернет-платформ на борьбу с «российской дезинформацией и вмешательством в выборы» (Saurwein & Spencer-Smith, 2020; Francois & Lin, 2021; Geissler, et al., 2023; Weintraub & Valdivia, 2020; Hodgins, 2022). Далеко не в последнюю очередь внедрение систем ИИ активизировалось в ответ на описанные требования, что предоставляет благоприятную среду для их неконтролируемого внедрения и нарушения прав российских пользователей. Упомянутый выше германский закон NetzDG также был принят в том числе под влиянием этого дискурса (Rumyantsev, 2019).

Важно понимать, что тот объем данных, которыми в настоящее время обладают крупнейшие интернет-платформы о том, где пользователь находится, что вводит с клавиатуры, какие делает фотографии и видео, какие посещает сайты и какими пользуется приложениями, сколько он прошел шагов и какой у него пульс, позволит в самом ближайшем будущем (в течение 1–2 лет) создавать цифровые модели пользователей, которые будут включать психологический портрет, карту социальных связей и в целом предугадывать поведение человека с постоянно возрастающей точностью.

Учитывая то, что в настоящее время такими корпорациями, как Google[16] и Apple[17], разрабатываются наушники с функциями электроэнцефалограммы мозга (так называемые нейроинтерфейсы), что в перспективе позволит им собирать данные о мозговой активности, соединяя эти данные с информацией о том, где человек находится, что он в данный момент пишет, читает, говорит в слух (эти данные соберут голосовые помощники в смартфонах), понятие «свобода мысли» начинает наполняться новым содержанием, трансформироваться в задачу обеспечения свободы человеческой мысли от вполне буквального контроля цифровыми средствами.

Описанное выше не может рассматриваться исключительно как гипотетическая возможность – примеры приложений, создающих иллюзорное, не существующее информационное пространство с помощью средств ИИ уже существуют. В частности, доступное для операционных систем iOS от Apple и Google Android приложение Parallel Live Simulator создает иллюзию того, что человек ведет прямую трансляцию, которую одновременно комментируют тысячи человек, ставят лайки и оставляют реакции на видео[18]. Цель приложения в том, чтобы стимулировать тщеславие, однако технология, которая используется данным приложением, может быть без труда масштабирована для того, чтобы создавать иллюзии коммуникации для пользователей любой интернет-платформы[19].

[16] Steven Levy. This startup wants to get in your ears and watch your brain. Wired. Режим доступа: https://www.wired.com/story/nextsense-wants-to-get-in-your-ears-and-watch-your-brain/ (дата обращения: 16.01.2024).

[17] Apple's new patent shows AirPods with brain wave-detecting sensors // The Times of India. Режим доступа: https://timesofindia.indiatimes.com/gadgets-news/apples-new-patent-shows-airpods-with-brain-wave-detecting-sensors/articleshow/102278175.cms (дата обращения: 16.01.2024).

[18] Apple AppStore. Parallel Live Simulator. Режим доступа: https://apps.apple.com/us/app/parallel-live-simulator/id1534165497 (дата обращения: 15.01.2024).

[19] Прим., соответствующие угрозы будут также освещаться в других разделах статьи.

Отметим, что интернет-платформы с самого начала своей деятельности старались максимизировать сбор данных о пользователях. Шон Паркер, бывший вице-президент компании Facebook[20], в ходе одной из конференций поделился своими наблюдениями о том, что стояло за архитектурой социальной сети при ее создании в 2005 году: «Как нам поглотить как можно больше вашего времени и сознательного внимания? Это значит, что нам нужно время от времени давать вам немного дофамина, потому что кому-то понравилась фотография, или пост, или что-то еще. Лайки, комментарии, сообщения – все это дает пользователям небольшую дозу дофамина, заманивая в петлю социального одобрения»[21]. Важно учитывать то, что в 2005 году у интернет-платформ не было такого мощного инструмента, как доступные в настоящее время программные комплексы, использующие ИИ. Однако, по словам Марка Цукерберга, уже в 2024 году ИИ будет «главным приоритетом для инвестирования Facebook»[22].

Угрозы свободе слова со стороны интернет-платформ

Для того чтобы проиллюстрировать данные тенденции более подробно, следует принять во внимание такие примеры использования цифровых технологий, которые в ближайшем будущем будут объединены в синергетические цифровые системы внутри интернет-платформ для получения «ИИ-клона» любого пользователя[23].

Распознавание политической ориентации по выражению лица – исследование M. Kosinski подтверждает, что ИИ уверенно распознает политическую ориентацию человека, для чего достаточно проанализировать одну фотографию, на которой четко видно выражение лица. Средняя точность, с которой ИИ устанавливает политическую ориентацию по фото, составляет 72% – это выше, чем анализ заполненного опросного листа, включающего не менее 100 вопросов (такой опрос дает ответ с точностью 66 %) (Kosinski, 2021). Для того чтобы получить соответствующий результат, потребовалась обработка около 1 миллиона доступных в сети анкет с сайтов знакомств. Отметим, что аналогичным образом ИИ может определять и другие личностные, а также социальные характеристики, глубоко затрагивающие частную жизнь человека. В частности, еще в 2018 году тем же исследователем сообщалось, что алгоритм может корректно определять половую ориентацию мужчин в 81 % случаев и женщин в 71 % случаев при загрузке одной фотографии человека. При загрузке не менее 5 фотографий точность алгоритма повышалась до 91 % у мужчин и 83 % у

[20] Продукт корпорации Meta Inc. признанной экстремистской организацией в Российской Федерации. 21.03.2022 г. Тверской районный суд г. Москвы удовлетворил иск Генпрокуратуры РФ и признал деятельность соцсети Facebook и Instagram, принадлежащей Meta, экстремистской, запретив ее работу в России.

[21] Mike Allen. Sean Parker unloads on Facebook*: “God only knows what it's doing to our children's brains”. Режим доступа: https://www.axios.com/2017/12/15/sean-parker-unloads-on-facebook-god-only-knows-what-its-doing-to-our-childrens-brains-1513306792 (дата обращения: 16.01.2024). *21.03.2022 г. Тверской районный суд г. Москвы удовлетворил иск Генпрокуратуры РФ и признал деятельность соцсети Facebook и Instagram, принадлежащей Meta, экстремистской, запретив ее работу в России.

[22] Meta’s AI Investments Ramp Up as Reality Labs Loses Billions. Режим доступа: https://www.pymnts.com/meta/2023/meta-threads-grabs-nearly-100m-users-just-three-months-after-launch/ (дата обращения: 15.01.2024).

[23] Yoon, D. (2023) AI clones made from user data pose uncanny risks. 4 June 2023. The Conversation. Режим доступа: http://theconversation.com/ai-clones-made-from-user-data-pose-uncanny-risks-206357 (дата обращения: 15.01.2024).

женщин. Данный результат был достигнут с помощью выборки величиной 35 326 изображений (Wang & Kosinski, 2018).

Оценка общих психологических характеристик человека с помощью открытых данных социальных сетей также не представляет для IT-специалистов какой-либо сложности. В частности, проведенное D. Quercia, M. Kosinski, D. Stillwell, J. Crowcroft еще в 2011 году исследование подтвердило возможность составить психологический портрет пользователя по 5 ключевым параметрам с точностью не ниже 0.88 по шкале до 1,5 единиц на основании лишь трех показателей аккаунта любого пользователя: числа подписчиков, числа подписок и числа добавления в тематические списки (Quercia et al., 2011). Аналогичные выводы об основных психологических свойствах личности ИИ способен делать на основании одной достаточно четкой фотографии (Kachur, et al., 2020).

Как указывает корпорация Microsoft, программный комплекс Azure AI Content Safety способен идентифицировать 4 вида контента, который, как полагают создатели, должен удаляться или ограничиваться средствами автоматизации. В частности, речь идет о контенте или сообщениях, выражающих ненависть (hate speech), подстрекающих к насилию (violence), сексуально откровенных материалах (sexually explicit material) и информации, направленной на распространение членовредительства (self-harm). В принципе это основные категории контента, которые стремятся ограничивать и удалять любые интернет-платформы (Dym & Fiesler, 2020). На первый взгляд, ничего предосудительного в удалении соответствующей информации, сообщений и контента, действительно нет. Однако, по мнению автора, данный взгляд является несколько поверхностным. Оценка угроз, которые несут подобные программные комплексы, требует внимательного и детального погружения в проблематику свободы слова и самовыражения с учетом специфики процессов цифровизации, протекающих в настоящее время и характеризуемых колоссальным влиянием крупных интернет-платформ и широким внедрением ИИ.

Еще 20 лет назад профессор L. Lessig писал о сети «Интернет» как об особом пространстве, где у человека включается другая логика поведения: «Этот охват стал возможен благодаря отсутствию власти в сети: любой человек, где бы он ни был, мог публиковать [*любую информацию*] для всех, везде. Сеть дала возможность публиковать [*информацию]* без фильтрации, редактирования или ответственности. Можно было написать то, что ты хочешь, решить, подписаться или нет [*т.е. указывать свое авторство, или нет*], отправить сообщение в компьютеры по всему миру, и через несколько часов эти слова будут повсюду. Сеть убрала самые важные недостатки свободы слова в реальном мире – *границу между издателем и автором* [*курсив мой*]. В реальном мире существует «тщеславная публикация» [*модель, в которой за публикацию платит автор, а не издатель*], но только богатые могут получить охват широкой аудитории таким образом. Для остальных из нас в реальном мире можно высказаться только в том объеме, который издатели хотят дать нам» (Lessig, 1999).

Это чрезвычайно точное и важное замечание, однако эту логику можно развить и продолжить. В частности, следует отметить, что Сеть диктует совершенно иной темп информационного обмена, чем тот, что был доступен человечеству до создания технологии WWW (World wide web). До цифровизации, в эпоху печати и телевидения, информация была доступна в вечерних и утренних газетах, еженедельных или ежемесячных журналах, книгах, телевизионных программах, которые выходили в определенные промежутки времени. Соответственно, формы самовыражения, такие как возможность высказаться по каким-либо насущным вопросам, были ограничены

архитектурой средств коммуникации — обыватель мог направить письмо в редакцию газеты или телевизионной передачи, однако это была в первую очередь односторонняя связь. Тем более что размещение поступившего письма в газете или какая-то иная реакция оставались на усмотрение редакции — именно об этом и говорил L. Lessig.

Другой важной формой публичного самовыражения была и остается возможность принять участие в публичном мероприятии, собрании, встрече закрытой группы. Однако здесь речь идет, скорее, о таком конституционном праве, как право на публичные мирные собрания. Возможность выступить на каком-либо крупном собрании со множеством участников следует считать привилегией, т.к. правом высказываться в таком формате в основном пользуются лица, играющие роль докладчиков, модераторов, членов президиума. Безусловно, такие формы общественной дискуссии важны, однако они в определенном смысле зависят от внешних обстоятельств, которые не находятся под контролем человека. Как отмечал L. Lessig, Сеть дала возможность высказываться практически без ограничений, но что еще очень важно — без какого-либо контекста, внешних обстоятельств, без необходимости достигать договоренностей о публикации. Отметим, что, в свою очередь, интернет-платформы дают возможность не просто высказаться, но и мгновенно распространить свое высказывание для огромного количества пользователей — охват может достигать десятков и даже сотен миллионов просмотров за короткий промежуток времени. Важно так же учитывать, что общение в Сети мультимодально и может вести способами невозможными в реальной жизни, когда в ответ на текстовое сообщение другой пользователей отвечает картинкой, видео или ссылкой (Yushkina & Panarina, 2019). Разумеется, описанные обстоятельства породили множество негативных эффектов, например, падение качества общественной дискуссии, снятие множества табу, которые существуют в реальном мире, грубость и бесцеремонность, агрессия и буллинг в коммуникации распространились слишком широко (Baburin & Cheremnova, 2022; Nikishin & Galyashina, 2021; Zhumabekova, 2022; Romanov, et al., 2021).

Однако цензура со стороны интернет-платформ, которая частно действует под благовидным предлогом удаления «ненадлежащей информации», не может быть оправдана этими обстоятельствами, несмотря на тот факт, что в Сети есть множество видов деструктивного контента, с которым необходимо бороться (в том числе применяя ИИ) как по соображениям защиты морали, так и во исполнение законодательных запретов, таких как запрет порнографии, пропаганды вражды или розни, защиты несовершеннолетних от информации, которая может негативно повлиять на их развитие (Titov, 2021).

В подтверждение данного тезиса можно привести множество примеров цензуры со стороны интернет-платформ. В 2016 году ряд работников Facebook[24] анонимно сообщили о фактах целенаправленного ограничения распространения информации консервативного политического спектра – она намеренно удалялась из раздела «Тренды», с помощью которого пользователи могут узнать о том, что в настоящее

[24] Продукт корпорации Meta Inc. признанной экстремистской организацией в Российской Федерации. 21.03.2022 г. Тверской районный суд г. Москвы удовлетворил иск Генпрокуратуры РФ и признал деятельность соцсети Facebook и Instagram, принадлежащей Meta, экстремис 21.03.2022 г. Тверской районный суд г. Москвы удовлетворил иск Генпрокуратуры РФ и признал деятельность соцсети Facebook и Instagram, принадлежащей Meta, экстремистской, запретив ее работу в России.

время популярно. Особо отметим, что уже тогда данный раздел формировался с помощью алгоритмов, но мог подвергаться «ручному» вмешательству, нарушающему «органическую выдачу»,[25] то есть не имеющую политической предубежденности. Схожую информацию предоставили общественности бывшие модераторы Facebook[26] Райан Хартвиг и Зак МакЭлрой. В частности, они подтвердили многочисленные факты применения двойных стандартов при осуществлении модерации в зависимости от того, против какой политической группы были направлены рестриктивные меры[27]. В 2018 году компания Twitter была уличена в применении цензуры путем применения «скрытого бана» (Shadow ban) – ее суть сводится к тому, что пользователь не знает о том, что к его аккаунту применяются какие-либо запретительные меры, однако другие пользователи не видят этот аккаунт в своей новостной ленте, сообщения от аккаунта не появляются в разделе «Тренды» (Jaidka, Mukerjee & Lelkes, 2023). В частности, такая мера применялась в отношении Президента США Дональда Трампа и Ронны Макдэниэл, председателя организационного комитета республиканской партии США[28]. Отметим, что в руководстве Twitter последовательно отрицали, что данные действия были намеренными, ссылаясь на некий программный сбой[29]. Более того, руководители обеих социальных сетей последовательно отрицали факты цензуры, причем делали это под присягой перед Конгрессом США[30]. Тем не менее раскрытые Илоном Маском после приобретения компании X внутренние документы дают серьезные основания для утверждений о существовании системы цифровой цензуры со стороны крупнейших интернет-платформ, расположенных в США (Diskin, 2023).

Регулирование рекомендательных алгоритмов и модерации в России

В настоящий момент законодательство, регулирующее правоотношения по поводу рекомендательных алгоритмов, находится на ранних этапах формирования. Тем

[25] Michael Nunez. Former Facebook workers: We routinely suppressed conservative news. Режим доступа: https://gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006 (дата обращения: 16.01.2024).

[26] Продукт корпорации Meta Inc. признанной экстремистской организацией в Российской Федерации. 21.03.2022 г. Тверской районный суд г. Москвы удовлетворил иск Генпрокуратуры РФ и признал деятельность соцсети Facebook и Instagram, принадлежащей Meta, экстремис 21.03.2022 г. Тверской районный суд г. Москвы удовлетворил иск Генпрокуратуры РФ и признал деятельность соцсети Facebook и Instagram, принадлежащей Meta, экстремистской, запретив ее работу в России.

[27] Christopher Boyle. Exclusive: Inside Facebook's* content moderator bias with Ryan Hartwig and Zach McElroy, two Ex-Employees come forward to Project Veritas. Режим доступа: https://www.publishedreporter.com/2020/07/03/exclusive-inside-facebooks-content-moderator-bias-with-ryan-hartwig-and-zach-mcelroy/ (дата обращения: 16.01.2024). 21.03.2022 г. Тверской районный суд г. Москвы удовлетворил иск Генпрокуратуры РФ и признал деятельность соцсети Facebook и Instagram, принадлежащей Meta, экстремис 21.03.2022 г. Тверской районный суд г. Москвы удовлетворил иск Генпрокуратуры РФ и признал деятельность соцсети Facebook и Instagram, принадлежащей Meta, экстремистской, запретив ее работу в России.

[28] Alex Thompson. Twitter appears to have fixed "shadow ban" of prominent Republicans like the RNC chair and Trump Jr.'s spokesman. Режим доступа: https://www.vice.com/en/article/43paqq/twitter-is-shadow-banning-prominent-republicans-like-the-rnc-chair-and-trump-jrs-spokesman (дата обращения: 16.01.2024).

[29] Alex Thompson. Twitter appears to have fixed "shadow ban" of prominent Republicans like the RNC chair and Trump Jr.'s spokesman. Режим доступа: https://www.vice.com/en/article/43paqq/twitter-is-shadow-banning-prominent-republicans-like-the-rnc-chair-and-trump-jrs-spokesman (дата обращения: 16.01.2024).

[30] Kevin Roose, Cecilia Kang. Mark Zuckerberg Testifies on Facebook Before Skeptical Lawmakers. Режим доступа: https://www.nytimes.com/2018/04/10/us/politics/zuckerberg-facebook-senate-hearing.html (дата обращения: 16.01.2024).

не менее, в последнее время активизировались усилия, направленные на регулирование рекомендательных алгоритмов и модерации как самостоятельных явлений.

Освещая изменения в российском законодательстве, особое внимание следует уделить Федеральному закону от 31.07.2023 № 408-ФЗ «О внесении изменений в Федеральный закон «Об информации, информационных технологиях и о защите информации» (далее – Закон о рекомендательных алгоритмах)[31]. Данный закон вводит новое понятие «рекомендательные технологии», обязывает информационные ресурсы размещать правила их применения на русском языке в публичном доступе. Краеугольным камнем данного закона является введение ответственности для владельцев сайтов, страниц, информационных систем и программ для ЭВМ, для которых применяются рекомендательные технологии за нарушение прав и законных интересов граждан, а также за предоставление информации с нарушением законодательства (статья 10.2–2). Однако вопрос о том, что именно является нарушением законодательства, законных прав и интересов граждан в данном контексте остается неясным. Федеральный закон от 27.07.2006 № 149-ФЗ «Об информации, информационных технологиях и о защите информации» (далее – Закон об информации) содержит большое число норм об ответственности за различные нарушения в сфере распространения информации, однако их нельзя однозначно, без сомнений и длительного анализа соотнести с нарушениями в области рекомендательных технологий. Как известно, данное обстоятельство является препятствием для привлечения лица к ответственности в силу положений ч. 3 ст. 49 Конституции Российской Федерации, так, любые неустранимые сомнения в виновности толкуются в пользу обвиняемого. Отдельно отметим, что глава 13 Кодекса об административных правонарушениях Российской Федерации (далее – КоАП РФ) не содержит специального состава, который бы устанавливал ответственность за нарушения, предусмотренные статьей 10.2–2 Закона об информации. В свою очередь статья 13.50 КоАП РФ не устанавливает ответственности за нарушения прав пользователей в ходе осуществления модерации, хотя и обязывает социальные сети отчитываться о ходе рассмотрения жалоб на наличие информации, размещенной с нарушением закона.

Таким образом, вопросы защиты прав граждан при осуществлении модерации интернет-платформ не были в полной мере учтены при принятии Закона о рекомендательных алгоритмах, хотя председатель комитета Государственной Думы по информационной политике, информатизации и связи Александр Хинштейн указывал, что основной идеей законопроекта было пресечь возможность «навязывать информацию, выгодную третьим лицам, как, например, это было во время последних выборов в США [в 2020 году]»[32].

Необходимо указать, что злоупотребления интернет-платформ во время подготовки и проведения выборов президента США в 2020 году не сводились лишь к необоснованному продвижению информации, выгодной для кандидата в президента от демократической партии, Джозефа Байдена, но и намеренному удалению негативной информации о нем, в том числе путем блокировок, направленных против лиц, сообщавших негативные сведения о нем (Diskin, 2023). Примечательно, что данные

[31] Федеральный закон от 31.07.2023 № 408-ФЗ «О внесении изменений в Федеральный закон «Об информации, информационных технологиях и о защите информации». Режим доступа: http://publication.pravo.gov.ru/Document/View/0001202307310021 (дата обращения: 15.01.2024 г.).

[32] Николай Козин. Как будет работать закон о рекомендательных алгоритмах \\ Парламентская газета. Режим доступа: https://www.pnp.ru/economics/kak-budet-rabotat-zakon-o-rekomendatelnykh-algoritmakh.html (дата обращения: 15.01.2024 г.).

блокировки проводились под предлогом борьбы с недостоверной информацией, в том числе якобы распространяемой в результате проведения специальной информационной операции властями Российской Федерации[33]. Однако позже было доказано, что запрещенная к распространению ведущими интернет-платформами под предлогом распространения «русской дезинформации»[34] статья в газете New York Post о содержимом ноутбука сына президента Байдена Хантера была основана на фактах, а информация о «дезинформации» как раз и была информационной специальной операцией, организованной штабом демократической партии США[35]. Описанные обстоятельства позволяют понять, почему противодействие недобросовестным практикам интернет-платформ в области модерации имеет большое значение. Удаление какой-либо информации под предлогом того, что она является «дезинформацией, исходящей из России»[36] с помощью инструментов ИИ становится инструментом цензуры, что безусловно недопустимо как с точки зрения защиты государственного суверенитета, так и прав граждан в области распространения и получения информации. Тем не менее, Закон о рекомендательных алгоритмах, хотя и нуждается в доработке, представляет собой важный шаг в области защиты прав и законных интересов пользователей интернет-платформ, в том числе в области модерации с применением технологий ИИ, так как обязывает платформы раскрывать методы работы рекомендательных технологий.

Делая обзор отечественного законодательства регулирующее правоотношения в области модерации интернет-платформ, необходимо упомянуть Федеральный закон от 04.03.2022 № 30-ФЗ «О внесении изменений в Федеральный закон «О мерах воздействия на лиц, причастных к нарушениям основополагающих прав и свобод человека, прав и свобод граждан Российской Федерации» и статью 27 Федерального закона «О порядке выезда из Российской Федерации и въезда в Российскую Федерацию» (далее – Закон о противодействии цензуре). Как указывалось в официальном сообщении Роскомнадзора, данный закон принимался с целью «пресечения цензуры со стороны иностранных интернет-компаний в отношении российских СМИ»[37]. Согласимся с высказанным Роскомнадзором тезисом о том, что «российское право не может и не должно подменяться на территории страны правилами интернет-компаний»[38]. Тем не менее несмотря на то, что в данном случае мы имеем официальное признание факта цензуры со стороны иностранных интернет-платформ в отношении

[33] Natasha Bertrand. Hunter Biden story is Russian disinfo, dozens of former intel officials say. Режим доступа: https://www.politico.com/news/2020/10/19/hunter-biden-story-russian-disinfo-430276 (дата обращения: 15.01.2024).

[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'. Режим доступа: https://www.businessinsider.com/yoel-roth-twitter-censor-hunter-biden-laptop-story-was-mistake-2022-11 (дата обращения: 15.01.2024).

[35] Paul Gigot, Bill McGurn, Kim Strassel. The Hunter Biden Laptop Disinformation Is Exposed. Режим доступа: https://www.wsj.com/podcasts/opinion-potomac-watch/the-hunter-biden-laptop-disinformation-is-exposed/4e8baf05-447c-419e-80d8-7424827c7b52 (дата обращения: 15.01.2024).

[36] Twitter. Disclosing networks of state-linked information operations. Режим доступа: https://blog.twitter.com/en_us/topics/company/2021/disclosing-networks-of-state-linked-information-operations- (дата обращения: 15.01.2024).

[37] Роскомнадзор. В Госдуму внесен законопроект о противодействии цензуре иностранных интернет-компаний в отношении российских СМИ. Режим доступа: https://rkn.gov.ru/news/rsoc/news73178.htm (дата обращения: 18.01.2024).

[38] Роскомнадзор. В Госдуму внесен законопроект о противодействии цензуре иностранных интернет-компаний в отношении российских СМИ. Режим доступа: https://rkn.gov.ru/news/rsoc/news73178.htm (дата обращения: 18.01.2024).

российских юридических лиц, нельзя не отметить, что Закон о противодействии цензуре не дает пользователям эффективных мер по противодействию такой цензуре. Более того, он не содержит формального определения цензуры, которое бы расширяло понятие, которое содержится в статье 3 Закона РФ от 27.12.1991 № 2124–1 «О средствах массовой информации». Сам Закон о противодействии цензуре состоит из трех статей, не содержит определений, ссылок на Закон об информации, не дает физическим лицам каких-либо инструментов по защите своих прав в спорах с иностранными интернет-платформами.

Регулирование рекомендательных алгоритмов и модерации в Европейском Союзе

Значительное внимание к регулированию правоотношений в области рекомендательных алгоритмов и модерации интернет-платформ уделяется в Европейском Союзе. В частности, 19.10.2022 был принят Регламент Европейского парламента и Совета ЕС 2022/2065, сокращенно именуемый «Закон о цифровых услугах (Digital Services Act – DSA)»[39]. В области модерации интернет-платформ с применением технологий ИИ, DSA обязует информировать пользователей об использовании таких технологий (часть 1 статьи 14). Онлайн-платформы обязаны готовить ежегодные отчеты об использовании инструментов модерации, а Европейская комиссия обязана создать открытую базу данных, где будут содержаться все решения онлайн-платформ по модерации, что делает этот процесс значительно более прозрачным (часть 5 статьи 22). Тем не менее, в данном законе есть положения, которые могут иметь негативные последствия для пользователей. В частности, статья 19 DSA предусматривает назначение «доверенных информаторов» (trusted flaggers). Роль данных лиц состоит в том, чтобы сообщать об информации, нарушающей законодательство. Интернет-платформы обязаны обрабатывать жалобы, поступающие от этих лиц немедленно и эффективно. Доверенные информаторы будут назначаться координаторами, которые в свою очередь назначаются компетентными органами государств-членов ЕС (статья 38). Несмотря на указание на то, что координаторы должны быть независимыми в своей деятельности (статья 39), какой-либо ответственности за злоупотребление со стороны доверенных информаторов или координаторов не установлено, что позволит государствам ЕС воздействовать на интернет-платформы через механизм подачи жалоб от доверенных информаторов. К ним не предъявляется каких-либо специальных квалификационных требований, не установлена ответственность за их беспристрастность или злоупотребление механизмом жалоб.

Регулирование рекомендательных алгоритмов и модерации в США

В свою очередь в США, где находятся штаб-квартиры крупнейших транснациональных корпораций – владельцев интернет-платформ (таких как X Corp., Alphabet Inc., Microsoft Corp., Meta Inc. и др.), федерального законодательства, которое бы защищало права пользователей при осуществлении модерации и использовании рекомендательных алгоритмов нет, ввиду сформировавшейся в 1990-е годы концепции

[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). Режим доступа: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065 (дата обращения: 16.01.2024).

невмешательства в деятельность технологических компаний. В силу данной концепции интернет-платформы могут самостоятельно определять контент, который можно и который нельзя публиковать пользователям на основании пользовательских соглашений. В свою очередь пользователи не имеют права оспаривать соответствующие решения интернет-платформ, а все споры должны рассматриваться по правилам, установленным в пользовательском соглашении. Это отражено в положении подраздела 230 раздела 47 (Section 230 Title 47) Акта о Пристойности Телекоммуникаций 1996 года (Communications Decency Act of 1996 – CDA; далее – подраздел 230), который дает интернет-платформам статус «издателя» (publisher), что в силу устоявшегося толкования Первой поправки к Конституции США (Билль о правах) наделяет их правами, аналогичными праву СМИ определять собственную редакционную политику, т.е. публиковать какую-либо информацию исключительно на основании редакционной политики, вмешательство в которую считается цензурой. Данное понимание правоотношений между интернет-платформами и пользователями является предметом как общественной дискуссии, так и борьбы за изменение законодательства. Например, текущий собственник платформы X Илон Маск отстаивает точку зрения, что такие интернет-платформы должны быть «цифровыми публичными пространствами» в том смысле, что они должны обеспечивать права на протест и высказывание, аналогичные публичным пространствам, как это гарантируется Биллем о правах[40]. В этом смысле он является противником подраздела 230 как положения, позволяющего интернет-платформам цензурировать контент, избегая контроля со стороны государства и пользуясь иммунитетом от судебного преследования, связанного с нарушениями, которые допускаются при модерации.

Борьба за отмену подраздела 230 ведется и на законодательном уровне. По оценкам исследователей, по состоянию на 2023 год было предпринято 84 попытки отменить или существенным образом изменить подраздел 230 на уровне федерального закона, однако ни одна из них не увенчалась успехом. 32 штата предприняли попытки принять свои законы, отменяющие на местном уровне положения подраздела 230. Штаты Флорида и Техас приняли соответствующие законы, прямо запрещающие цензуру и алгоритмическую дискриминацию (злоупотребление алгоритмами для дискриминации пользователей), однако оба были признаны неконституционными федеральными судами первой инстанции (Sinnreich, et al., 2023). По состоянию на январь 2024 года соответствие обоих законов Конституции рассматривается Верховным судом США[41]. Отметим, что в сфере регулирования интернет-платформ есть и другие законодательные инициативы, например в штате Нью-Йорк (Senate Bill S4531A)[42], однако их рассмотрение выходит за рамки настоящей статьи, так как они направлены на ужесточение модерации платформ, но не содержат требований по обеспечению гарантий прав пользователей.

[40] Ezra Klein. The Great Delusion Behind Twitter. Режим доступа: https://www.nytimes.com/2022/12/11/opinion/what-twitter-can-learn-from-quakers.html (дата обращения: 16.01.2024).

[41] Amy Howe. Justices request federal government's views on Texas and Florida social-media laws. Режим доступа: https://www.scotusblog.com/2023/01/justices-request-federal-governments-views-on-texas-and-florida-social-media-laws/ (дата обращения: 16.01.2024).

[42] New York State Senate. Senate Bill S4531A. Режим доступа: https://www.nysenate.gov/legislation/bills/2021/S4531 (дата обращения: 16.01.2024).

Перед заключением

Применение технологий ИИ для осуществления модерации интернет-платформ постоянно расширяется и все чаще становится причиной нарушения прав пользователей как в силу технологических ошибок, когда ИИ не способен понять контекст, в котором делается высказывание или распространяется информация, так и в силу специфической политики, направленной на подавление распространения информации по политическим соображениям. ИИ предоставляет интернет-платформам большие возможности для контроля за размещаемым пользователями контентом. Уже отмечены случаи, когда действия по блокировке общественно важной информации и сообщений, признанные впоследствии «ошибками» со стороны интернет-платформ, совершались полностью автоматически (средствами ИИ) в реальном времени – пользователь не мог разместить информацию или поделиться ею в личных сообщениях, так как она была заранее определена интернет-платформами как «недостоверная»[43].

Граждане России часто являются целью соответствующей политики интернет-платформ, когда под предлогом борьбы с недостоверной информацией удаляются аккаунты государственных органов, общественных организаций, СМИ, физических лиц. Так, начиная с 2021 года Роскомнадзор зафиксировал 132 акта цензуры со стороны иностранных интернет-платформ в отношении российских СМИ[44]. К сожалению, внимание к защите физических лиц-пользователей со стороны отечественного законодателя явно недостаточно. В частности, признание многочисленных фактов цензуры не привело к разработке законодательного акта, который содержал бы необходимые определения, упорядочивал существующее законодательство и устанавливал систему прав пользователей и гарантий их осуществления в процессе модерации. В то же время обвинения в адрес государственных органов, общественных организаций, отдельных физических лиц в «распространении дезинформации» в основном остаются без ответа как со стороны государства, так и со стороны пострадавших лиц, ввиду отсутствия такой системы прав и гарантий. Принятый в 2022 году Закон о противодействии цензуре не достиг своей цели – цензура российских пользователей продолжается и расширяется. Так, в декабре 2022 года официальный представитель МИД России Мария Захарова заявила, что «Запад организовал против российских медиа самый масштабный в истории акт массовой тоталитарной цензуры»[45]. Безусловно, предполагать, что один, пусть даже детально проработанный законодательный акт, решит проблему такого масштаба, было бы наивно. Тем не менее, очевидно, что Закон об информации должен быть дополнен группой норм, которая бы детально закрепляла права пользователей и системой гарантий их исполнения. В этом смысле ориентиром должен выступать Закон о цифровых услугах ЕС за исключением норм о доверенных информаторах, которые автор считает инструментом для опосредованной государственной цензуры.

---

[43] Facebook* and Twitter restrict controversial New York Post story on Joe Biden. Режим доступа: https://www.theguardian.com/technology/2020/oct/14/facebook-twitter-new-york-post-hunter-biden (дата обращения: 16.01.2024). *21.03.2022 г. Тверской районный суд г. Москвы удовлетворил иск Генпрокуратуры РФ и признал деятельность соцсети Facebook и Instagram, принадлежащей Meta, экстремистской, запретив ее работу в России.

[44] Роскомнадзор выявил девять фактов цензуры в отношении российских СМИ. Режим доступа: https://regnum.ru/news/3806632 (дата обращения: 16.01.2024).

[45] Захарова заявила о масштабной цензуре Запада против российских СМИ. Режим доступа: https://regnum.ru/news/3763034 (дата обращения: 16.01.2024).

Заключение

Реакция России на описанные тенденции использования ИИ для осуществления цензуры не должна ограничиваться совершенствованием национального законодательства. В рамках председательства России в БРИКС+ необходимо обратить внимание данной организации на данную проблематику. В частности, предлагается создание специальной рабочей группы в рамках БРИКС+ по совершенствованию регулирования Сети, в том числе поставив перед ней задачу выработки мер по недопущению дискриминации граждан государств-участников со стороны интернет-платформ, с особым фокусом на наиболее крупных (very large providers), как это определено в Законе о цифровых услугах ЕС.

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.

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) Ilon 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) C*ountering 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., Галяшина 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 essense 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