В ФОКУСЕ – **ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ** И **СВОБОДА СЛОВА**

Рекомендации по разработке политики

ОБСЕ
Представитель по вопросам свободы СМИ

Настоящая публикация подготовлена в рамках проекта «В фокусе - искусственный интеллект и свобода слова» (#SAIFE).

В документе представлены мнения, умозаключения, интерпретации, рекомендации и выводы авторов, которые не обязательно отражают официальную позицию ОБСЕ и/или ее государств-участников.

6a Wallnerstrasse

1010 Vienna, Austria

Тел.: +43-1-514-36-68-00

Эл. почта: pm-fom@osce.org

https://www.osce.org/fom/ai-free-speech

ISBN: 978-92-9234-743-7

Представитель ОБСЕ по вопросам свободы СМИ
В фокусе – искусственный интеллект и свобода слова
Рекомендации по разработке политики

Авторы

Элиска Пиркова, Маттиас Кеттеманн, Марлена Висняк, Мартин Шейнин, Эмми Бевензее, Кэти Пентни, Лорна Вудс, Люсьен Хайц, Бояна Костич, Криштина Розгони, Холли Сарджеант, Юлия Хаас и Владан Йолер

Редакторы

Дениз Вагнер и Юлия Хаас

Эксперты

Дженнифер Адамс, Сьюзи Алегре, Аша Аллен, Андреас Маркманн Андреассен, Николетт Ашоди, Джеф Аслоос, Жозефина Баллон, Жоан Барата, Надя Белларди, Сьюзен Бенеш, Гай Бергер, Фредерик Зюйдервеен Боргезиус, Ирина Бороган, Джонатан Брайт, Эльда Броги, Эми Бруйе, Джоанна Дж. Брайсон, Пит Бурнап, Камилла Бустани, Игнасио Талегон Кампоамор, Майя Капелло, Марсело Дахер, Анита Данка, Николас Диакопулос, Айжамал Джаныбаева, Лейла Догруэль, Мария Донде, Сеад Дзигал, Франческа Фануччи, Марк Фумагалли, Максимилиан Ганц, Яна Гайдошова, Майя Индира Ганеш, Ляля Гайе, Брэнди Гёркинк, Арзу Гейбулла, Мишель Гилман, Надин Гогу, Габриэль Гиллемин, Рустам Гулов, Бен Хейс, Натали Хельбергер, Джорджия Холмер, Андреа Хубер, Каролина Иваньска, Сэм Джефферс, Эллиот Джонс, Паскаль Юргенс, Агнес Каарлеп, Фредерике Кальтеунер, Кари Карппинен, Сьюзен Керр, Бенджамин Килле, Юджин Ким, Вольфганг Кляйнвахтер, Беата Климкевич, Джордже Кривокапич, Лубош Куклиш, Андрей Кулешов, Джоанна Кулеша, Коллин Курре, Сьюзен Ландау, Пэдди Леерссен, Эмма Ллансо, Джеймс Макларен, Жуан Карлос Магальяэс, Самвел Мартиросян, Эстель Массе, Кайл Мэтью, Элеонора Мария Маццоли, Тарлах Макгонагл, Марко Милосавлевич, Мира Милошевич, Ива Ненадич, Мариэлза Оливейра, Ребекка Овердорф, Ройя Пакзад, Седжал Пармар, Патрик Пеннинккс, Джон Пенни, Карлос Перес-Маэстро, Эмилия Петреска-Каменьярова, Андрей Петровски, Кортни Радш, Отабек Рашидов, Джудит Раухофер, Дэвид Райхель, Мориц Ризевик, Катица Родригес, Ася Рокша-Зубчевич, Бьянка Шонбергер, Кристофер Шварц, Лиза Зайдль, Муртаза Шайх, Джат Сингх, Ваня Шкорич, Андрей Солдатов, Мария-Луиза Стаси, Николас Сюзор, Дамиан Тамбини, Дханарадж Такур, Гульнура Торалиева, Макс ван Друнен, Виталий Васильченко, Франциско Вера, Кристина Воко, Диана Влад Кальсич, Бен Вагнер, Дуглас Уэйк, Хилари Уотсон, Агнешка Вавжик, Весна Вессенауэр и Андрей Цвиттер.

Кураторы

УВКПЧ ООН, ЮНЕСКО, Совет Европы, Европейская аудиовизуальная обсерватория, Европейская комиссия, Агентство Европейского союза по основным правам, Европейский вещательный союз, ОБСЕ (Секретариат, БДИПЧ, ВКНМ)

Корректор
Том Поппер

Дизайн и макет
Пено Мишоян

Переводчик
Виктория Гаспарова

Содержание

Предисловие

Уважаемые читатели!

С радостью представляю Вашему вниманию публикацию нашего Бюро, посвященную искусственному интеллекту и свободе выражения мнения (SAIFE). Эта публикация является результатом двухлетних исследований и нескольких экспертных семинаров, объединивших знания более чем ста самых известных ученых и практиков, работающих в области свободы медиа, прав человека, технологий и безопасности.

В 2022 году мандату Представителя ОБСЕ по вопросам свободы СМИ исполняется 25 лет. В 1997 году, когда был основан этот институт, лишь у 1,7 процента населения Земли был доступ к интернету, а цифровые технологии, позволявшие осуществлять связь через интернет, были новинкой и внушали уверенный оптимизм.

Двадцать пять лет спустя число людей, имеющих доступ к интернету в регионе ОБСЕ, превысило 80 процентов. Этот грандиозный рост числа пользователей самым благоприятным образом сказался на свободе выражения мнения, свободном обмене информацией и возможности искать, получать и передавать любую информацию и идеи по всему миру, не взирая на границы и расстояния.

Это имело решающее значение для экономического, общественного и политического участия населения, демократизации, развития образования и здравоохранения, обеспечения подотчетности власти, а также для выявления военных преступлений и других нарушений прав человека. В то же время это породило массовую слежку, киберпреступность и распространение незаконного и деструктивного контента в интернете.

Управление неизмеримыми объемами информации в интернете стало невозможным без использования технологий машинного обучения и других форм искусственного интеллекта (ИИ). Технологии ИИ становятся основными инструментами для формирования и анализа контента в интернете; они используются для принятия решений о том, какой контент

удалить, какой приоритизировать или кому его рекомендовать. Эти операции осуществляются системами, которые разрабатываются и внедряются рядом онлайн-платформ – так называемыми «привратниками» цифрового мира.

Это влиятельные компании, способные формировать и регулировать политический и общественный дискурс. Несомненно, то, как курируется и модерируется информация в интернете, оказывает прямое и значительное влияние на мир, стабильность и всеобъемлющую безопасность в глобальном масштабе. С большой властью должна приходить и большая ответственность. Тем не менее, эти новые «привратники» и их методы работы развиваются со скоростью, опережающей любые правовые или нормативные рамки для использования ИИ с целью формирования нашего информационного пространства в интернете.

Мы оказались на перепутье.

Государства-участники ОБСЕ должны объединиться для поиска многосторонних решений проблем, возникших в их общем информационном пространстве, поставив права человека во главу угла при разработке и внедрении ИИ для курирования и модерации контента в глобальной сети.

Эти проблемы носят далеко идущий характер, и решения могут быть найдены только благодаря действиям множества заинтересованных сторон. Что касается проблем, связанных со свободой СМИ и свободой выражения мнения, надеюсь, что эта публикация поможет государствам-участникам ОБСЕ, политикам, ученым и специалистам в области медиа как в регионе, так и за его пределами, понять, как можно совместно развивать такие гарантии прав человека в рамках своих национальных, региональных и международных возможностей.

Декабрь 2021 года.

Тереза Рибейру,
Представитель ОБСЕ по вопросам свободы СМИ.

Основные рекомендации в адрес государств-участников ОБСЕ

1. Защищать и поощрять свободу выражения мнения и другие **права человека в качестве центрального аспекта** стратегий и политики в области ИИ.

2. Сохранять и развивать **интернет** как пространство для **демократического участия** и представительства, а также **плюрализма медиа**.

3. Разработать **основанную на фактах** и базирующуюся на **инклюзивных процессах** политику с целью реагирования на вызовы, стоящие перед свободой мнения, свободой информации и свободой выражения мнений.

4. Содействовать соблюдению **Руководящих принципов предпринимательской деятельности в аспекте прав человека ООН**, с тем чтобы не допустить придание первостепенной важности извлечению максимальной прибыли в ущерб правам человека и демократическим ценностям.

5. Обязать онлайн-платформы проводить **надлежащую проверку соблюдения прав человека**, в том числе посредством оценки воздействия на права человека (ОВПЧ) их политики управления контентом и автоматизированного принятия решений, а также применяемой ими практики работы, такой как сбор данных, целевая реклама и дизайн интерфейса.

6. Обеспечить **ясность, объяснимость и доступность** использования ИИ для модерации контента, курирования контента и целевой рекламы.

7. Принять меры для того, чтобы работа по защите прав человека не перекладывалась полностью на сторонних исполнителей или автоматизированные процессы, а также обеспечить прозрачность деятельности любых **публично-частных партнерств**.

8. Принять жесткие рамки для обеспечения **прозрачности**, в том числе путем введения обязательных всесторонних отчетов, содержащих подробную информацию об использовании ИИ.

9. Обеспечить наличие **надежных механизмов защиты** от цензуры и слежки, таких как проверка человеком и независимый механизм обжалования (апеляции).

10. Гарантировать строгую **подотчетность**, в том числе посредством **независимого надзора** и **аудита**, особенно в отношении соблюдения прав человека и недискриминации.

11. Обеспечить соблюдение **права на неприкосновенность частной жизни** и защиту данных, в том числе путем ограничения рекламы на основе слежки и обеспечения максимальной прозрачности и свободы действий пользователей при осуществлении деятельности, связанной с отслеживанием и профилированием.

12. Содействовать развитию медийной и **цифровой грамотности**, а также **расширению прав и возможностей пользователей, их самостоятельности и контролю** над управлением контентом и использованием их данных, в том числе путем предоставления возможности отказаться от любого автоматизированного принятия решений

13. Принять меры по устранению неравного и **монополизированного влияния на рынок** и содействию плюрализму, технологическим и медийным инновациям.

14. Осуществлять **многостороннее взаимодействие** с целью гарантировать соблюдение прав человека при разработке и внедрении ИИ для курирования и модерации контента в интернете.

Введение: Соблюдение принципов Хельсинкского Заключительного акта в цифровую эпоху

Не так давно исполнилось 45 лет со дня подписания Хельсинкского Заключительного акта 1975 года. Этот итоговый документ первого саммита глав государств и правительств СБСЕ стал краеугольным камнем политического порядка в Европе. Государства Восточной и Западной Европы согласовали десять принципов, которыми они будут руководствоваться в своем поведении, включая уважение суверенного равенства и взаимное уважение прав человека и основных свобод, закрепленное в Принципе VII. Хельсинкский Заключительный акт также содержит обязательства по сотрудничеству между государствами, включая научно-техническое сотрудничество. В нем даже упоминаются компьютерные технологии и признается необходимость сотрудничества, особенно в отношении развития «систем телекоммуникаций и информации; техники, связанной с ЭВМ и телекоммуникациями, включая их применение в системах управления, производственных процессах, для автоматизации, при изучении экономических проблем, в научно-исследовательских работах и для сбора, обработки и распространения информации».[1]

Сегодня как никогда необходимо сотрудничество и многосторонние подходы, а новые участники, определяющие способы обработки, распространения и курирования информации, требуют новых нормативных подходов к решению проблем прав человека в современном информационном ландшафте. В то время как государства несут основную ответственность за соблюдение, защиту и реализацию прав человека, интернет-посредники, и особенно несколько доминирующих платформ социальных сетей,[2] оказывают все большее влияние на реализацию

1 Заключительный акт Совещания по безопасности и сотрудничеству в Европе, Хельсинки, <https://www.osce.org/helsinki-final-act>.

2 Онлайн-платформы выполняют широкий спектр функций, включая хранение и распространение информации. К таким платформам относятся социальные сети, поисковые системы, рекламные сети и электронные торговые площадки. Данная публикация посвящена онлайн-платформам, которые в основном предназначены для облегчения взаимодействия людей в интернете путем предоставления пространства для общения. Одни платформы в основном размещают и курируют контент, другие дополнительно способствуют цифровой торговле. Платформы, которые в основном

этих прав. В интернете можно наблюдать новый квазинормативный порядок, который бросает вызов традиционным представлениям о нормативности.[3] В современном цифровом мире осуществление свободы выражения мнения все чаще регулируется в рамках частных, гибридных и общественных пространств, которые формируются частными компаниями, государствами и пользователями в различных, крайне асимметричных властных отношениях. Более того, эти онлайн-экосистемы проложили путь новым формам управления выражением мнения, включая те, что осуществляются при помощи алгоритмов и искусственного интеллекта (ИИ). Чаще всего квазинормативные стандарты интернет-посредников повсеместно определяют объем и интенсивность регулирования свободы выражения мнения. Такое управление контентом, как правило, осуществляется без всякого общественного контроля и часто путем непрозрачного автоматизированного принятия решений в определенном масштабе, без каких-либо гарантий соблюдения международной системы прав человека.

Использование автоматизации в управлении контентом еще больше усугубляет многие существующие проблемы с правами человека в интернете и одновременно порождает новые. В целом, инструменты ИИ широко используются для модерации и курирования пользовательского контента, а также для предоставления персонализированной рекламы. Автоматизированные и основанные на ИИ инструменты, применяемые для модерации и курирования онлайн-контента, находятся в центре научных и политических дебатов. Частные субъекты и разработчики политики часто представляют ИИ в качестве панацеи, которая уже может или через несколько лет сможет решить очень сложные вопросы, связанные с распространением и тиражированием потенциально незаконного или деструктивного контента. Однако проактивная и автоматизированная идентификация, обнаружение и удаление онлайн-контента несут в себе системные риски. Такие инструменты на основе ИИ обычно внедряются доминирующими частными субъектами, часто по требованию государства, либо напрямую, через юридически обязательные законодательные рамки, либо косвенно, через усиление давления на посредников, чтобы заставить их «делать больше». Кроме

предназначены для взаимодействия между людьми, в том числе в коммерческих целях, обычно называются платформами социальных сетей.

3 Matthias C. Kettemann, The Normative Order of the Internet (Oxford: OUP, 2020).

того, использование автоматизации и ИИ для курирования контента и, таким образом, продвижения одной информации за счет другой, исходя из внутренней, ориентированной на получение прибыли политики посредников, также несет в себе системные риски. Некоторые из этих рисков проистекают из автоматизированных систем принятия решений, непосредственно связанных с основанными на слежке бизнес-моделями очень крупных интернет-посредников.

Множество групп по защите гражданских прав уже много лет бьют тревогу, указывая на продолжающиеся нарушения прав человека в результате непрозрачного автоматизированного принятия решений. Интернет-посредники, такие как платформы социальных сетей, приобрели важное значение для частного взаимодействия и публичного дискурса, однако они управляются алгоритмами, которые определяют доступ людей к информации, а значит, и процесс формирования общественного мнения. Преобладающие бизнес-модели наиболее влиятельных интернет-посредников основаны на слежке за пользователями, зачастую используя их психологическую уязвимость и другие человеческие слабости. Построенные на основе массового сбора и анализа пользовательских данных, эти бизнес-модели являются частью рыночной экосистемы, которую профессор Гарвардского университета Шошана Зубофф назвала «капитализмом слежки».[4] Факты свидетельствуют о том, что основанные на слежке бизнес-модели привели к искажению нашей информационной среды, тем самым вступив с противоречие с принципами плюрализма, разнообразия, демократии и принятия решений. Недавние разоблачительные откровения Фрэнсис Хауген только подтвердили эти предположения, подчеркнув необходимость создания государствами модели управления платформами, ориентированной на права человека. Видя необходимость обеспечения защиты прав человека, многие призывают к усилению государственного регулирования. Однако регулирование деятельности интернет-посредников, особенно использования ими искусственного интеллекта и алгоритмических систем с целью снижения их социальных рисков, является сложной и многогранной задачей.

4 Ranking Digital Rights, It's the Business Model: How Big Tech's Profit Machine is Distorting the Public Sphere and Threatening Democracy (2021).

В целом, существует крайне малое количество положительных примеров управления контентом с соблюдением прав человека, а некоторые добровольные обязательства по усилению защиты этих прав, взятые на себя интернет-посредниками, в конечном итоге оказались недостаточными. Поэтому настало время перейти к принципам создания онлайн-экосистем, ориентированных на соблюдение прав человека, и в этой связи внести свой вклад в поддержание принципов Хельсинкского Заключительного акта в цифровую эпоху. Подобный акт мог бы вновь объединить тех, кто рассматривает интернет как продолжение своих национальных границ, и тех, кто готов проводить политику, в большей степени ориентированную на права человека. Таким образом, он подтвердил бы саму суть ОБСЕ, согласно которой права человека являются неотъемлемой частью всеобъемлющей безопасности, как в интернете, так и вне его.

Цель проекта «В фокусе искусственный интеллект и свобода слова» (SAIFE) состоит в том, чтобы предоставить государствам-участникам ОБСЕ рекомендации по выполнению их позитивного обязательства по защите прав человека при разработке нормативных мер для реагирования на новые вызовы, стоящие перед правом на свободу выражения мнения в цифровую эпоху. В рамках проекта были организованы четыре экспертных семинара с целью выявления фактического и прогнозируемого негативного воздействия автоматизированных и основанных на ИИ методов обнаружения, оценки, курирования и персонализации онлайн-контента на права человека отдельных лиц. В ходе семинаров особое внимание уделялось индивидуальному праву на свободу выражения и свободу мнения, а также правам на уровне общества, включая свободу медиа. По итогам семинаров был подготовлен ряд ориентированных на права человека рекомендаций с целью определения мер для проявления должной осмотрительности в вопросах прав человека и процедурных гарантий для устранения индивидуальных рисков и рисков для общества, возникающих в результате необоснованного использования ИИ в процессе управления контентом.

Структура и краткое описание документа

В рамках проекта SAIFE, в первой половине 2021 года Бюро Представителя ОБСЕ по вопросам свободы СМИ совместно с организацией Access Now организовало четыре экспертных семинара, посвященных рассмотрению и анализу основных проблем, создаваемых системами ИИ в области прав человека, в частности, прав на свободу выражения мнения, свободу и плюрализм медиа. Семинары были посвящены четырем основным темам:

- **Модерация контента – безопасность**
  инструменты на основе ИИ, применяемые при модерации контента для обнаружения и оценки незаконного контента в интернете, включая угрозы безопасности, экстремистский и террористический контент.

- **Модерация контента – язык ненависти**
  инструменты на основе ИИ, используемые для обнаружения и оценки потенциально деструктивного, но легального контента, с особым акцентом на ненавистническую риторику в интернете и алгоритмическую дискриминационную предвзятость.

- **Курирование контента – плюрализм медиа**
  инструменты на основе ИИ, предназначенные для курирования и персонализации интернет-контента, с акцентом на системы рекомендации контента и их влияние на плюрализм медиа.

- **Курирование контента – слежка**
  инструменты на базе ИИ, используемые в рекламе на основе слежки, и их связь с курированием контента посредством профилирования частных лиц и прогнозирования их поведения.

В данном документе содержатся основные выводы экспертных семинаров, а также политические рекомендации, адресованные государствам-участникам ОБСЕ, при этом признается необходимость многостороннего подхода для эффективного и устойчивого решения сложных проблем, которые модерация и курирование контента создают для свободы выражения мнения. В ходе семинаров были сформулированы рекомендации для государств-участников ОБСЕ, которые были рассмотрены известными

экспертами в области свободы выражения мнения, плюрализма медиа и искусственного интеллекта. Публикация основана на докладах по итогам каждого семинара, которые были подготовлены совместно председателем, назначенным руководить работой отдельных экспертных групп, докладчиками соответствующего экспертного семинара и Элишкой Пирковой – представительницей партнерской организации Access Now. Подготовка докладов включала консультации со всеми экспертами и наблюдателями, участвовавшими в соответствующих семинарах. Структура документа соответствует тематическим областям, рассмотренным каждой экспертной группой. Он состоит из четырех отдельных разделов, каждый из которых содержит ориентированные на права человека политические рекомендации, адресованные государствам-участникам ОБСЕ.

Использование ИИ в модерации контента

Результаты первых двух экспертных семинаров, посвященных использованию ИИ в модерации контента для борьбы с незаконным и потенциально деструктивным контентом, таким как язык ненависти, были объединены в один общий раздел. Этот раздел содержит ряд политических рекомендаций, призванных содействовать предотвращению негативного воздействия инструментов ИИ, используемых при модерации контента, на право искать, получать и распространять любого рода информацию и идеи.

Модерация контента – безопасность

инструменты на основе ИИ, применяемые при модерации контента для обнаружения и оценки незаконного контента в интернете, включая угрозы безопасности, экстремистский и террористический контент

Одна из двух рабочих групп, занимавшихся вопросами модерации контента, сосредоточилась на автоматизированных системах и системах на основе ИИ, используемых для обнаружения и принятия мер в отношении незаконного контента и аккаунтов, связанных с его распространением. Эта практика включает в себя технологии фильтрации и сопоставления хэш-значений, применяемые для блокирования загружаемых материалов, а также инструменты для удаления или понижения рейтинга контента задним числом, часто с трансграничным эффектом. Заметные проблемы

возникают тогда, когда технологии ИИ используются для мониторинга национального законодательства или даже для предоставления правоохранительным органам возможности мониторинга электронных сообщений пользователей под предлогом обеспечения безопасности и общественного порядка. В результате этого под особым давлением может оказаться индивидуальная и групповая анонимность, что может привести к охлаждающему воздействию на свободу выражения мнения и свободу медиа, а также на безопасность журналистов. Хотя влияние модерации контента с использованием ИИ на противоправное поведение по-прежнему неясно, технологии ИИ не зависят от контекста и склонны к чрезмерно широкому применению правил, которые они призваны установить. Это означает, что они регулярно генерируют так называемые ложноположительные и ложноотрицательные результаты при выявлении предположительно незаконного контента в сети, что может привести к произвольным ограничениям законного выражения или невозможности ограничить незаконное выражение мнения.

Рабочая группа указала на потенциальное негативное воздействие использования инструментов на основе ИИ для модерации контента на свободу выражения мнения отдельных лиц, а также на более широкие общественные риски для свободы медиа, демократии и верховенства права. Политические рекомендации, выдвинутые рабочей группой по изучению модерации контента и вопросов незаконного контента, позволяют государствам-участникам ОБСЕ выявлять, анализировать и оценивать значительные системные риски, связанные с инструментами модерации контента, включая случаи их использования для предотвращения быстрого распространения незаконного контента в сети. Эти рекомендации объединены с рекомендациями рабочей группы по легальному, но деструктивному контенту, включая ненавистническую риторику, и содержат указания в отношении гарантий свободы слова при модерации контента с использованием ИИ, а также руководство по обеспечению прозрачности, доступа к данным, независимого надзора, средств правовой защиты и рамок должной осмотрительности в вопросах прав человека.

Работа данной экспертной группы и подготовка этой части доклада проходила под руководством председателя **профессора Мартина**

Шейнина, при поддержке докладчиков **профессоров Маттиаса Кеттеманна** и **Марлены Висняк**.

Модерация контента – язык ненависти

Инструменты на основе ИИ, используемые для обнаружения и оценки потенциально деструктивного, но легального контента, с особым акцентом на ненавистническую риторику в интернете и алгоритмическую дискриминационную предвзятость

Вторая рабочая группа по модерации контента рассмотрела фактическое и прогнозируемое негативное воздействие автоматизированных и основанных на ИИ инструментов для выявления и оценки ненавистнической риторики в интернете на права человека отдельных лиц, с акцентом на право маргинализированных групп на свободу выражения и свободу мнения. Влияние дискриминационной предвзятости может проявляться в виде «предвзятой цензуры» в отношении контента, размещенного представителями конкретных общественных групп, которые часто становятся объектом ненавистнических высказываний и оскорблений в сети. Хотя язык ненависти сам по себе сильно зависит от контекста и его трудно обнаружить и удалить автоматически, группы, которые могут стать мишенью для оскорблений в интернете, могут быть вынуждены молчать, поскольку их собственные сообщения подвергаются цензуре. Для обучения автоматизированных инструментов выявлению и распознаванию различных категорий контента используются наборы данных. Если эти наборы не включают образцы сообщений на различных языках разных сообществ, или если в обучающих данных не представлены определенные группы, это может привести к ошибочной классификации, которая оказывает несоразмерное воздействие на маргинализированные сообщества. Автоматизированные инструменты могут либо пропустить потенциально ненавистнический контент (ложноотрицательные результаты), либо ошибочно пометить законные высказывания как ненавистническую риторику (ложноположительные результаты).

Совместные рекомендации по гарантиям свободы слова при использовании ИИ для модерации контента содержат указания

относительно прозрачности, доступа к данным, независимого надзора, средств правовой защиты и рамок должной осмотрительности в отношении прав человека. Конкретные рекомендации рабочей группы по «языку ненависти» направлены на создание условий для выявления и устранения системных рисков, особенно для маргинализированных групп, возникающих в результате использования систем модерации контента на основе ИИ, применяемых для выявления потенциально опасного контента, такого как ненавистнические высказывания. Эти рекомендации содержат руководство по созданию ориентированных на права человека автоматизированных инструментов модерации контента и по расширению цифрового участия маргинализированных групп в публичном дискурсе.

Работа этой экспертной группы и разработка данной части доклада осуществлялась под руководством председателя **профессора Лорны Вудс**, при поддержке докладчиков **Эмми Бевенси** и **Кэти Пентни**.

Использование ИИ в курировании контента

Часть А: Курирование контента – плюрализм медиа

Инструменты на основе ИИ, предназначенные для курирования и персонализации интернет-контента с акцентом на системы рекомендации контента, и их влияние на плюрализм медиа

В первой части раздела о курировании контента анализируется негативное воздействие алгоритмических систем рекомендации контента на права человека с акцентом на абсолютное право на свободу мнения, а также плюрализм и свободу медиа. В нем рассматриваются: распространение потенциально деструктивного контента, такого как вводящий в заблуждение, вызывающий поляризацию или разжигающий ненависть контент; влияние рекомендательных систем на разнообразие мнений и идей; влияние алгоритмического курирования на право на формирование мнения и плюрализм медиа; риск поляризации общества. Алгоритмический отбор контента основан на политике интернет-посредников, которые следуют своим внутренним экономическим интересам и интересам рекламодателей, а не ориентируются на достоверность, разнообразие или общественный интерес (например,

ценность новостей). Такой подход влияет на взаимодействие в социальных сетях и свободный обмен информацией, а также оказывает давление на профессиональную журналистику, направляя доходы от рекламы посредникам. Более того, пользователи обращаются к новостям реже, чем к предлагаемому в индивидуальном порядке пакету информационных сообщений, поэтому каждому отдельному сообщению приходиться бороться за внимание пользователей в новостной ленте, что стимулирует использование кликбейта для их привлечения. Хотя эта модель содействует рекламе и приносит прибыль посредникам, она представляет собой проблему с точки зрения плюрализма медиа.

После описания проблем в этой части доклада выдвигается ряд политических рекомендаций для государств-участников ОБСЕ по обеспечению реальной прозрачности интернет-посредников, повышению самостоятельности и контроля со стороны отдельных пользователей, а также рекомендации по продвижению разнообразия мнений, информации, представляющей общественный интерес, и плюрализма медиа.

Работой экспертной группы и подготовкой доклада руководила председатель **профессор Криштина Розгони** при поддержке докладчиков **Люсьена Хайца** и **Бояны Костич**.

Часть B: Курирование контента – слежка

Инструменты на базе ИИ, используемые в рекламе на основе слежки, и их связь с курированием контента посредством профилирования частных лиц и прогнозирования их поведения

Вторая часть раздела о курировании контента посвящена связи между курированием контента и рекламой. Использование ИИ в целевой рекламе связано с направлением конкретных рекламных объявлений отдельным пользователям путем использования автоматизированных статистических данных – например, машинного обучения, обработки естественного языка, распознавания речи и изображений. Различные формы использования данных, включая психологическое профилирование и нанотаргетинг, возможны благодаря обработке данных, извлечению сигналов и автоматизированному анализу широкого спектра различных

типов данных – таких как пользовательский контент, данные о местоположении, поведенческие модели, психографика, информация о расе, экономическом статусе, поле, возрасте, поколении, уровне образования, уровне дохода и занятости пользователя. Краткосрочное и долгосрочное, а также прямое и косвенное влияние основанной на слежке рекламы на поведение человека, его благосостояние и общество в целом пока неизвестно, но системы на базе ИИ неоднократно выдавали необъективные и ошибочные результаты.

В этой части доклада обсуждается далеко идущее воздействие автоматизированных процессов и процессов на базе ИИ, используемых в рекламе на основе слежки, на личное взаимодействие, общение и участие людей в демократических дебатах. От нарушения неприкосновенности частной жизни до фрагментации информационного пространства, реклама на основе слежки может нанести серьезный ущерб праву свободно формировать и придерживаться мнения, а также искать, получать и передавать информацию. Рабочая группа, подготовившая эту часть доклада, рассматривает такие вопросы, как отсутствие объяснимости и прозрачности алгоритмических систем, использующих личные и поведенческие данные людей; манипулятивные маркетинговые техники, использующие определенные характеристики и слабые стороны пользователей для повышения убедительности сообщения; дискриминация, вызванная оптимизирующими рекламу алгоритмами; распространение потенциально деструктивного контента для повышения вовлеченности пользователей с целью увеличения прибыли.

Основанные на этом анализе рекомендации включают меры, направленные на повышение прозрачности, предотвращение и смягчение рисков для прав человека, обусловленных такой практикой, как навязчивое таргетирование и персонализация контента. В рекомендациях также подчеркивается необходимость борьбы с основанными на слежке бизнес-моделями нескольких доминирующих интернет-посредников. Рекомендации в области политики, адресованные государствам-участникам ОБСЕ, включают в себя расширение прав и возможностей отдельных пользователей по осуществлению контроля над своими

данными, а также получаемой и передаваемой ими информацией, а также более эффективную защиту свободы мнения в цифровой экосистеме.

Работой экспертной группы и подготовкой доклада руководил председатель **профессор Владан Йолер** при поддержке докладчиков **Холли Сарджеант** и **Юлии Хаас**.

ИСПОЛЬЗОВАНИЕ ИИ В МОДЕРАЦИИ КОНТЕНТА

Использование ИИ в модерации контента в контексте реагирования на угрозы безопасности и язык ненависти

Эта часть доклада посвящена использованию ИИ в модерации контента, а также последствиям в плане прав человека, вытекающим из использования инструментов ИИ для целевого воздействия на конкретные категории пользовательского контента. В ней подчеркиваются недостатки использования ИИ при модерации контента в контексте как явно незаконного контента, такого как материалы террористического или экстремистского характера, так и потенциально деструктивного, но законного контента, такого как ненавистническая риторика, в частности, с точки зрения маргинализированных групп. В заключении приводятся оперативные и технические рекомендации для государств-участников ОБСЕ, ориентированные на обеспечение прав человека. Эти рекомендации направлены на устранение существующего негативного воздействия используемых при модерации контента инструментов ИИ на право искать, получать и распространять любого рода информацию и идеи.

1. Определение масштабов модерации контента

Два экспертных семинара были посвящены использованию инструментов ИИ для модерации контента, в первую очередь двум категориям пользовательского интернет-контента: незаконному контенту и потенциально деструктивному, но законному контенту, с особым акцентом на ненавистническую риторику. В следующих разделах описывается масштаб работы экспертных групп в каждой из этих областей.

1.1 Угрозы безопасности и незаконный контент в интернете

Автоматизированные средства обнаружения потенциально незаконного контента в интернете – также называемые проактивными мерами – находятся в центре научных и политических дебатов. Частные субъекты и разработчики политики часто представляют ИИ в качестве панацеи, которая в конечном итоге сможет решить очень сложные вопросы, связанные с распространением незаконного контента, включая пропаганду терроризма. Однако такой взгляд на технологию,

представленный в качестве обоснования для ускорения «внедрения ИИ во всех отраслях экономики, как в частном, так и в государственном секторе»[5] не учитывает системные риски, связанные с проактивной идентификацией, обнаружением и удалением пользовательского контента. Хотя устранение угроз безопасности является законным и необходимым шагом, ответные меры не должны осуществляться в ущерб правам человека. Риски связаны с автоматизированными системами принятия решений, которые используются онлайн-платформами, часто по требованию государства, либо напрямую, через юридически обязательные законодательные рамки, либо косвенно, через усиление давления на платформы, чтобы заставить их «делать больше».

Независимо от используемого технологического метода, такие автоматизированные инструменты могут накладывать предварительные ограничения на право на свободу выражения мнения и свободу информации. На практике это означает, что они могут априори исключать определенные лица, группы, идеи или средства выражения из общественного дискурса. В международной системе прав человека и в различных конституционных законах существуют строгие требования к обоснованию предварительных ограничений на свободу выражения мнения. Эти требования вытекают из опасений по поводу чрезмерного ограничения свободного обмена информацией. В этом отношении инструменты ИИ вызывают особую озабоченность, поскольку они ограждены от любого общественного контроля, не учитывают контекст и действуют крайне непрозрачно, что исключает любую возможность применения эффективного средства правовой защиты и возмещения ущерба. Хотя предварительный отбор контента для ограничения распространения вредоносных программ и материалов, связанных с сексуализированным насилием над детьми, широко признается в качестве положительного использования автоматизации, необходимо сохранять осторожность при применении той же логики к другим типам высказываний, что относится к более широкой сфере управления контентом.[6]

5 Европейская комиссия, Приложение к сообщению Комиссии Европейскому парламенту, Европейскому совету, Совету, Европейскому экономическому и социальному комитету и Комитету по делам регионов, «Скоординированный план реализации европейской стратегии в области ИИ» (7 декабря 2018 года, COM(2018) 795 final, стр. 4, <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56017>.
6 Emma Llanso, "No amount of 'AI' in content moderation will solve filtering's prior-restraint problem" Big Data & Society 7(1), p. 1-2, <https://journals.sagepub.com/doi/pdf/10.1177/2053951720920686>.

Работа экспертной группы имеет особое значение с учетом того, что в регионе ОБСЕ за последнее время появилось большое количество законодательных инициатив по регулированию потенциально незаконного контента в интернете. В итоговом отчете группы содержатся рекомендации по улучшению регулирования инструментов ИИ для модерации контента с учетом прав человека. Они призваны способствовать определению ориентированных на права человека регулирующих мер, принимаемых в ответ на распространение и тиражирование незаконного контента в глобальной сети.

Хотя этот итоговый доклад не содержит определение потенциально незаконного контента, представленные в нем рекомендации по безопасности ориентированы на проактивные методы обнаружения и оценки:

- **Контента, который является незаконным независимо от контекста** типичным примером такого контента является сексуализированное насилие над детьми, которое запрещено рядом международных правовых документов, таких как Будапештская конвенция Совета Европы, Лансаротская конвенция, Конвенция №182 Международной организации труда, Конвенция ООН о правах ребенка и другие. Однако даже в отношении этой категории контента национальные законы не предусматривают единообразных мер.

- **Контента, являющегося частью более крупного преступления** случае со ставшими вирусными видео с обезглавливанием, как минимум одно насильственное преступление имело место в «реальной жизни». Любая инициатива по модерации контента без учета элементов преступления, имевшего место в реальной жизни, может лишить жертв возможности для возмещения ущерба. Кроме того, публикация, равно как и удаление такого контента, может повлиять (в качестве улики) на расследование и документирование нарушений прав человека.

- **Законного контента, который является незаконным в определенном контексте**
  это относится к контенту, который не является незаконным сам по себе, однако способ его публикации в интернете может быть приравнен к уголовному преступлению. Типичным примером такого

контента является публикация изображений наготы без согласия соответствующего лица или несанкционированная публикация личной информации.

- **Контента, который является незаконным в основном в силу декларируемых намерений и последствий**
  к этой категории относится подстрекательство к насилию или терроризму. Обычно преступлением является не сам контент, а скорее (субъективное) намерение, стоящее за его публикацией, в сочетании с (объективным) риском того, что он может побудить некоторых пользователей к насилию. К этой категории также относятся, например, ксенофобия, подстрекательство к дискриминации и разжигание ненависти.

1.2 Язык ненависти в интернете

Благодаря большому количеству интернет-посредников всех мастей и размеров был сформирован глобальный рынок идей, позволяющий людям по всему миру обмениваться и получать информацию и идеи. В то же время, это способствовало распространению и усилению ненавистнической риторики.[7] Государствам приходится отстаивать противоречивые интересы, одновременно защищая свободу слова отдельных лиц и права и свободы объектов и получателей ненавистнических высказываний, а также общества в целом. В частности, осуществление прав человека может быть ограничено для маргинализированных групп, которые подвергаются дискриминационным предубеждениям и зачастую вынуждены молчать из-за таких общественных явлений, как язык ненависти. Проявление ненависти не является уникальным для интернет-контекста. Напротив, она существовала в «реальном мире» во всех обществах и на протяжении всей истории и была направлена против отдельных лиц и групп на основе идентифицируемых характеристик, таких как раса, пол/гендер, религия и сексуальная ориентация. Однако сетевое измерение ставит новые задачи с точки зрения объема, охвата и воздействия языка ненависти. Например, только за последний квартал 2020 года Facebook удалил более 20 миллионов

7 See, e.g., European Commission, Countering illegal hate speech online: 5th evaluation of the Code of Conduct (June 2020) at <https://ec.europa.eu/info/sites/default/files/codeofconduct_2020_factsheet_12.pdf>.

материалов,[8] а Google за тот же период удалил с YouTube около 100 тысяч видеороликов, пропагандирующих ненависть.[9] Отличительной чертой языка ненависти, распространяемого в интернете, является то, что тем, на кого он рассчитан (или широкой общественности) труднее избежать его или отгородиться от него, поскольку он может проникать в традиционно безопасные пространства, включая жилища людей, зачастую анонимно, а иногда при помощи скоординированных клеветнических кампаний.[10]

Усилия по борьбе с языком ненависти значительно активировались в последние несколько лет в связи с тем, что это явление набирает силу, а его воздействие на общество вызывает все большую озабоченность. В фокусе оказались вопросы о том, как эффективнее бороться с языком ненависти в сети, о разграничении ролей и обязанностей государств и частных субъектов в модерировании языка ненависти, а также о роли, которую должны играть автоматизированные системы принятия решений в обнаружении и удалении такого контента.

На международном уровне не существует общепринятого определения языка ненависти, что позволяет судам и трибуналам определять границы допустимости и недопустимости высказываний. Под понятие «язык ненависти» может подпадать широкий спектр высказываний: от незаконной ненавистнической риторики, такой как подстрекательство к насилию и геноциду, в наиболее экстремальной части спектра, до потенциально незаконных ненавистнических высказываний, таких как угрозы насилия и преследования, и до высказываний, которые, не будучи незаконными, являются деструктивными и оскорбительными.[11] С распространением онлайн-платформ и пользовательского контента

8 Центр прозрачности Facebook, Отчет об обеспечении соблюдения норм сообщества: <https://transparency.facebook.com/community-standards-enforcement#hate-speech>.
9 Отчет о прозрачности деятельности Google, «Политика в области противодействия языку ненависти» (октябрь 2020 - декабрь 2020): <https://transparencyreport.google.com/youtube-policy/featured-policies/hate-speech?hl=en>.
10 M. Williams and M. de Reya, "Hatred Behind the Screens: A Report on the Rise of Online Hate Speech" (2019) p. 18 at <https://hatelab.net/wp-content/uploads/2019/11/Hatred-Behind-the-Screens.pdf>.
11 Стратегия и план действий ООН по борьбе с ненавистнической риторикой (2020), Таблица 1, стр. 16, <https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf>.

задача по определению и регулированию языка ненависти все чаще возлагается на частные компании, однако государства по-прежнему несут основную ответственность за защиту прав человека, включая свободу выражения мнения, недискриминацию и доступ к соответствующим средствам правовой защиты. Крайне важно обеспечить наличие соответствующих рекомендаций в отношении деятельности частных корпораций на цифровом рынке идей и надзора за их выполнением.

Свобода выражения относится не только к информации или идеям, которые воспринимаются положительно, но также к информации и идеям, которые задевают, шокируют или вызывают беспокойство.[12] Любое ограничение этой свободы должно быть законным, соразмерным и соответствовать международному праву. В то время как с удалением незаконного контента все довольно очевидно, в отношении контента, который не является незаконным, но может быть деструктивным и нарушать права других людей, дело обстоит сложнее. Определить эту вторую категорию контента и выработать соответствующие меры реагирования на нее довольно сложно. Объектами ненавистнических высказываний зачастую становятся традиционно маргинализированные группы общества, чьи голоса не слышны и которые не представлены в кабинетах власти. Поэтому очень важно обеспечить более широкое участие таких групп и их представителей в принятии решений по этим фундаментальным вопросам, включая обсуждение эффективных мер борьбы с языком ненависти. На практике отсутствие широкого участия и представленности в процессе принятии решений привело к чрезмерно обобщенным и недостаточно инклюзивным подходам к проблеме языка ненависти, особенно в интернете.[13]

Непонимание и отсутствие инклюзивности может привести к ситуациям, когда свободное выражение мнения маргинализированными группами населения неправомерно квалифицируется как ненавистническая риторика, что позволяет автоматизированным системам принятия

12 «Хэндисайд против Соединенного Королевства», жалоба № 5493/72 (ЕСПЧ, 7 декабря 1976 года) [49]; Стратегия и план действий ООН, стр. 14.

13 See, e.g., M. K. Land and R. J. Hamilton, "Beyond Takedown: Expanding the Toolkit for Responding to Online Hate" in Predrag Dojcinovic (ed.) Propaganda, War Crimes Trials and International Law: From Cognition to Criminality 143 (Routledge, 2020), p. 2 at <https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3514234_code858831.pdf?abstractid=3514234&mirid=1>.

решений эффективно блокировать публикации таких групп и их отдельных представителей. Это может быть следствием непонимания контекста, включая динамику отношений в группе и за ее пределами. Например, термины «квир» и «гей» могут использоваться как гомофобные или трансфобные оскорбления, определяемые и регулируемые как ненавистническая риторика; однако они также могут быть частью лексики членов сообщества ЛГБТК+ или использоваться для «просоциальных функций», таких как создание сообществ и групп по интересам и оказание людям помощи в том, чтобы лучше подготовиться к возможной враждебности со стороны других людей.[14] Аналогичное проиллюстрированному выше непонимание контекста и намерений приводит к необоснованному удалению контента, публикуемого в интернете представителями расовых меньшинств.[15] Регулирование языка ненависти должно зависеть от контекста – от намерений автора сообщения, его вероятных последствий, а также конкретного значения слов или изображений в данном социально-политическом контексте. Исследования показали, что системы автоматизированного принятия решений просто не в состоянии справиться с этой контекстуальной задачей. Поэтому обобщенные или чрезмерно инклюзивные подходы могут привести к цензуре в отношении представителей маргинализированных групп, нарушая их свободу выражения.[16] Этот эффект подавления свободы выражения должен стать предметом первоочередной озабоченности как государств, так и интернет-посредников.

Если политика в отношении языка ненависти является недостаточно инклюзивной – то есть, не учитывает законные, но деструктивные высказывания – интернет-пространство может стать небезопасной или неблагоприятной средой для представителей маргинализированных групп, фактически вытесняя их глобальной сети. Это особенно проблематично в свете той важной роли, которую это пространство играет на нашем новом

14 Thiago Dias Oliva, "Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online" (2021) Sexuality & Culture 25, p. 705-7, <https://www.researchgate.net/publication/345501707_Fighting_Hate_Speech_Silencing_Drag_Queens_Artificial_Intelligence_in_Content_Moderation_and_Risks_to_LGBTQ_Voices_Online>.

15 Thomas Davidson, Debasmita Bhattacharya and Ingmar Weber, "Racial Bias in Hate Speech and Abusive Language Detection Datasets" (2019), <https://www.aclweb.org/anthology/W19-3504.pdf>; Maarten Sap et al, "The Risk of Racial Bias in Hate Speech Detection" (2019), <https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf>.

16 Там же.

(цифровом) рынке идей, куда люди все чаще обращаются для обмена информацией, ознакомления с новостями и участия в общественных дебатах. В результате может возникнуть «дефицит демократии», когда представители маргинализированных групп – женщины и небинарные люди, расовые и этнические меньшинства, члены ЛГБТК+ сообщества и т.д. – не могут или не хотят полноценно участвовать в демократическом дискурсе.[17] Более того, политика может быть недостаточно инклюзивной, поскольку не учитывает межсекторальность – то есть ненавистническую риторику, направленную против отдельных лиц или групп на основе двух или более идентифицирующих факторов.[18]

Что касается рекомендаций, ориентированных на борьбу с языком ненависти, необходимо иметь в виду следующие замечания, касающиеся сферы их применения:

- Хотя данный доклад не содержит определение языка ненависти, основное внимание в нем уделяется законной, но деструктивной ненавистнической риторике в интернете. Рекомендации направлены на регулирование и модерацию такой риторики в соответствии с правами человека.

- В свете несоразмерного воздействия модерации языка ненависти на маргинализированные группы, в докладе представлены рекомендации для государств-участников ОБСЕ по обеспечению защиты таких групп и их представителей, включая их право на свободу выражения мнения, недискриминацию и доступ к адекватным средствам правовой защиты.

- Хотя модерация контента происходит на разных уровнях – как более подробно рассмотрено ниже –настоящий доклад в первую очередь фокусируется на крупномасштабной (англ. industrial) модерации языка ненависти, чтобы отразить масштаб и степень ее воздействия на свободу выражения.

17 Nani Jansen Reventlow, "The power of social media platforms: who gets to have their say online?" Lilith (February 4, 2021), <https://www.lilithmag.nl/blog/2021/2/3/the-power-of-social-media-platforms-who-gets-to-have-their-say-online>.
18 Стратегия и план действий ООН, стр. 28.

2. Инструкция по модерации контента

Обнаружение и модерация незаконного или потенциально деструктивного, но законного контента является сложной задачей. Хотя привлечение интернет-посредников к ответственности имеет крайне важное значение, понимание ограничений самой технологии (а также задействованных бизнес-моделей) помогает государственным органам принимать более действенные меры в отношении корпораций, осуществляющих свою деятельность в сфере социальных сетей. В данном разделе представлен обзор инструментов и методов алгоритмической модерации контента, а также определены несколько слабых мест широкомасштабной модерации.

Инструменты ИИ зачастую используются для модерации контента без адекватного обоснования и аргументации решений. Это означает, что пользователи часто не знают, почему было принято то или иное автоматизированное решение и какая именно информация подтолкнула систему к принятию этого решения.

Типы модерации контента

Согласно определению Робин Каплан, существует три основные модели модерации контента:

- **Модерация силами компании** (англ. **artisanal moderation**): модерация контента небольшими группами штатных модераторов.

- **Модерация силами сообщества** (англ. **community-reliant moderation**): модерация контента, преимущественно осуществляемая на добровольной основе представителями сообщества из различных подразделений интернет-посредника; такой моделью пользуется, к примеру, Wikipedia или Reddit.

- **Крупномасштабная модерация** (англ. **industrial moderation**): модерация контента с привлечением большого количества

сторонних специалистов в сочетании с использованием собственных автоматических систем машинного обучения.[19]

В данном докладе основное внимание уделяется воздействию и аспектам третьего типа – крупномасштабной модерации.

Средства модерации контента можно разделить на следующие категории:

- **Обнаружение**
  нахождение и идентификация контента, который может нарушать политику интернет-посредника.

- **Принятие решения**
  определение того, действительно ли обнаруженный контент нарушает политику посредника.

- **Приведение в исполнение**
  принятие мер в отношении контента в соответствии с требованиями, изложенными в политике посредника.

- **Обжалование (Апелляция)**
  возврат к этапу вынесения решения, если пользователь оспаривает или обжалует решение посредника.

- **Политика**
  набор принципов, правил или инструкций, определяющих, какой контент является приемлемым на платформе посредника. На практике эти принципы пересматриваются и обновляются на основе других компонентов процесса модерации контента.[20]

Выявление незаконного контента или контента, нарушающего условия предоставления услуг посредника – или прогнозирование того, что выявленный контент попадет в одну из этих категорий – может привести к ряду последствий. Наиболее распространенными из них

19 Robyn Caplan, Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches (2018), <https://datasociety.net/library/content-or-context-moderation/>.
20 Meedan, Content Moderation Toolkit: Toolkit for Civil Society and Moderation Inventory, at <https://meedan.com/reports/toolkit-for-civil-society-and-moderation-inventory/>.

являются пометка или удаление контента. В случае удаления контент немедленно изымается и в некоторых случаях больше не загружается. Помимо удаления, существует ряд инструментов для решения вопроса «проблемного» контента задним числом. К таким инструментам можно отнести следующие:

- **Демонетизация контента**
  на таких платформах, как YouTube и Twitch, где авторы могут получать прибыль от популярности своего контента, условия предоставления услуг могут быть предусматривать возможность лишения пользователей прибыли от определенных типов контента. Хотя такая демонетизация может иметь свои преимущества, она зачастую непропорционально применяется в отношении представителей маргинализированных групп, либо из-за вышеупомянутых проблем с алгоритмом, либо по причинам намеренного приглушения голосов таких групп.

- **Понижение приоритетности и ранжирования контента**
  тот контент, который пользователь видит на онлайн-платформах, обычно контролируется рядом частных алгоритмов, призванных повысить заинтересованность. Интернет-посредники могут понижать ранжирование или удалять из списка рекомендаций оскорбительные или деструктивные аккаунты. Хотя это может быть полезно с точки зрения борьбы с распространением деструктивного контента, такого как ненавистническая риторика или дезинформация, это часто приводит к росту популярности и без того популярного контента, такого как основные новости, потенциально за счет приглушения голосов маргинализированных групп населения.

- **Приостановка или ограничение функций аккаунта**
  временная приостановка аккаунта служит сдерживающим фактором для пользователей, нарушающих правила сообщества, без применения постоянного запрета в их отношении. Хотя такие приостановки могут предотвратить новые проявления ненавистнической риторики, они также могут быть использованы против маргинализированных групп, которые пытаются создать для себя безопасное пространство в интернете.

- **Удаление аккаунта**
  полное удаление аккаунта может лишить пользователя возможности сохранить обширную базу подписчиков, следовательно, может быть особенно действенной мерой в отношении злостных нарушителей. Однако, как показал недавний опыт удаления аккаунтов в Telegram, экстремистские сообщества быстро приспосабливаются к ситуации, воссоздавая свои каналы и возвращая свою аудиторию после удаления аккаунтов.

- **Блокировка/отключение уведомлений/отписка**
  эти опции обеспечивают форму субъективной модерации, которая позволяет пользователям решать, какой контент они не хотят видеть в своей личной ленте. На платформах социальных сетей на базе протокола Secure-Scuttlebutt блокировки являются прозрачными, что превращает выражение доверия и недоверия в инструмент для ограничения распространения нежелательных сообщений.[21]

Крупномасштабная модерация контента с использованием алгоритмов

Алгоритмическая модерация контента включает в себя ряд методов, применяемых в статистике и информатике и различающихся по степени сложности и эффективности. Все эти методы предназначены для идентификации, сопоставления, прогнозирования или классификации пользовательского контента на основе его точных свойств или общих характеристик. Автоматизированные инструменты используются интернет-посредниками для масштабного регулирования контента по целому ряду тем, включая терроризм, сцены насилия, «токсичные высказывания», несанкционированное изображение наготы, жестокое обращение с детьми и обнаружение спама. Два типа алгоритмической модерации контента, а именно, анализ текста и изображений, используются преимущественно, хотя и не исключительно, для борьбы с потенциально незаконным контентом в интернете. Когда определенный фрагмент контента отмечается инструментами ИИ как потенциально незаконный, он обычно включается в очередь или список приоритетов для проверки «модератором-экспертом». Затем он может быть удален

21 Более подробную информацию можно найти на децентрализованной платформе Scuttlebutt, <https://scuttlebutt.nz/>.

или рассмотрен с помощью одного из вышеупомянутых инструментов задним числом.

При анализе текста, системы машинного обучения регулярно используют инструменты обработки естественного языка (NLP). Такие инструменты подвергают текст комплексному синтаксическому разбору, пытаясь приблизить анализ к пониманию текста человеком. Инструменты NLP учатся определять то, какие эмоции передает конкретный текст – положительные или отрицательные (так называемый анализ настроения), и, следовательно, классифицировать его принадлежность или непринадлежность к определенной категории пользовательского контента. Инструменты NLP предназначены для прогнозирования результатов на основе помеченных образцов, например, «оскорбительный» или «неоскорбительный» контент. Наиболее известным примером инструмента NLP является API Perspective от Google/Jigsaw – набор инструментов с открытым исходным кодом, позволяющий операторам веб-сайтов, исследователям и другим лицам использовать свои модели машинного обучения для оценки «токсичности» сообщения или комментария.

С другой стороны, автоматическое обнаружение и идентификация изображений и видео часто предполагают обнаружение контента, который ранее был идентифицирован как незаконный, а также обнаружение нового контента, который может быть добавлен к категории незаконного. Технологии обнаружения и идентификации изображений используют так называемые хэш-значения. Хэш – это уникальное числовое значение, также называемое «цифровым отпечатком пальца», которое генерируется определенным алгоритмом, работающим с файлом изображения. Простая технология хэширования оценивает размерность изображения или цветовые значения пикселей. Простое изменение пикселя изображения полностью меняет хэш файла, что позволяет легко обойти инструмент. Более сложные инструменты используют перцептивное хэширование, которое включает в себя отпечатки пальцев изображений и видео, смешанные с другими характеристиками контента, например, такими как измеряемая в герцах частота повторений за промежуток времени в аудио. Перцептивные хэши более надежны и могут идентифицировать изображения и видео даже после их изменения. Типичным примером перцептивного хеширования является PhotoDNA,

разработанная компанией Microsoft и используемая для борьбы с насилием в отношении детей в интернете.

После произошедшего в 2019 году теракта в Крайстчёрче в Новой Зеландии, Facebook, Google, Twitter и Microsoft создали Глобальный интернет-форум по борьбе с терроризмом (GIFCT) в рамках своих обязательств по содействию добровольному соблюдению компаниями Кодекса поведения ЕС по противодействию незаконному разжиганию ненависти в интернете. В рамках GIFCT четыре компании обмениваются передовым опытом разработки инструментов алгоритмической модерации контента. Они также используют сверхсекретную и непрозрачную хэш-базу данных террористического контента, в которой делятся друг с другом цифровыми отпечатками «незаконного контента», включая изображения, видео, аудио и текстовые сообщения. Например, в течение нескольких часов после теракта в Крайстчёрче Facebook загрузил хэши примерно 800 различных версий видеоролика с изображениями стрелка. Теоретически, каждое видео, загруженное пользователями Facebook, YouTube и Twitter, теперь может быть хэшировано и проверено по базе данных. Контент, совпадающий с записью в базе данных, будет немедленно заблокирован. Эта база данных находится в ведении исключительно частных лиц и не подлежит общественному контролю, что создает серьезные проблемы для журналистского и художественного контента.

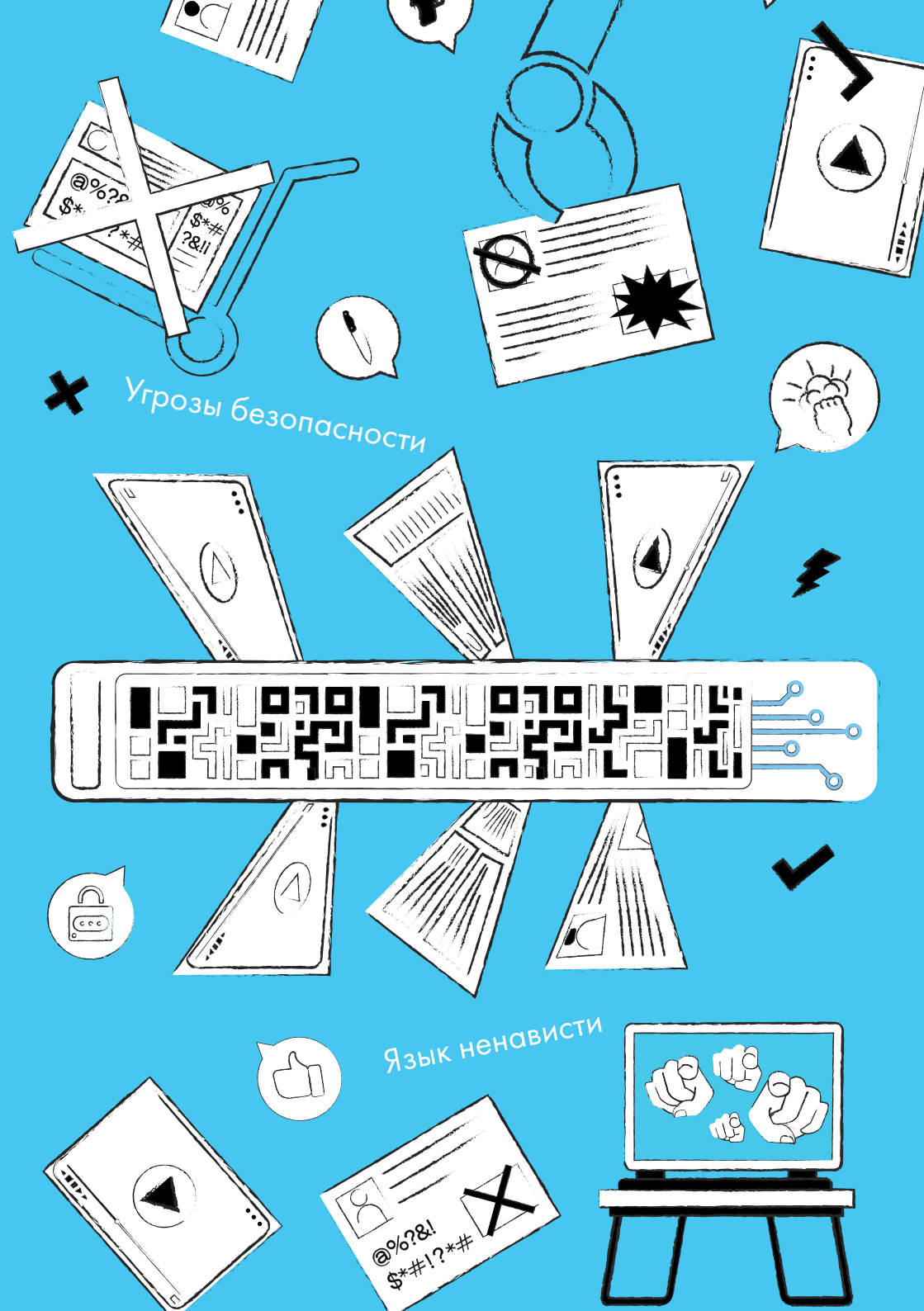
Недостатки алгоритмической модерации контента

Существует несколько уязвимых мест в проектировании и разработке алгоритмических инструментов модерации контента, а также в бизнес-моделях интернет-посредников, которые используют эти инструменты, и об этом должны помнить разработчики политики. Одно из основных уязвимых мест при разработке алгоритма машинного обучения возникает, когда команда людей определяет правила аннотирования обучающих данных, которые будут использоваться в модели машинного обучения. Этот шаг очень важен, поскольку ИИ, по сути, является копировальной машиной. Системы ИИ учатся тому, чему их учат люди, и даже в этом случае возможны некоторые отклонения. Предвзятость людей и предвзятость, заложенная в самих данных, будет воспроизводиться на протяжении всего жизненного цикла системы ИИ.

Например, если система, пытающаяся выявить преступность в интернете, опирается на структурно расистские данные, это приведет к более глубокому закреплению расистских результатов. Дополнительные проблемы возникают из-за тонкостей речи или их использования в группе и вне ее, что может привести к систематическому неверному обозначению терминов, нанося вред тем, кого система призвана защищать. ИИ не предусматривает единого простого решения для борьбы с ненавистью в интернете в силу того, что каждый тип ненависти на основе идентичности и каждый контекст отличаются друг от друга, а также в силу постоянного изменения ландшафта по мере адаптации противников. Среда, в которой создаются и внедряются системы машинного обучения, особенно для такой деликатной цели, как обнаружение языка ненависти, глубоко динамична и контекстуальна.

Хотя прозрачность и участие в процессах решения этих проблем могли бы значительно снизить связанные с ними риски, интернет-посредники имеют частные интересы, которые противоречат требуемой прозрачности. Например, алгоритм машинного обучения для выявления языка ненависти сам по себе является товаром, который можно продать. Поэтому посредники, скорее всего, придадут ему статус коммерческой технологии. Более того, многие посредники утверждают, с той или иной степенью достоверности, что совместное использование таких алгоритмов позволит противникам злоупотреблять ими. Кроме того, трудности могут возникнуть при объяснении конкретных типов решений, для которых используется инструмент ИИ, или гарантии того, что совместное использование инструмента ИИ направлено на решение новых форм именно той проблемы, для решения которой он был создан.

Надежная система модерации контента должна обеспечивать соразмерность и точность реакции на незаконный или потенциально деструктивный контент, одновременно стремясь решить технические и социально-политические вопросы, связанные с самой модерацией. Хотя государства обычно имеют косвенный контроль над практикой модерации, применяемой корпорациями, полная осведомленность о существующей политике и практике, а также об альтернативных возможностях помогает информировать и направлять процесс разработки политики.


Угрозы безопасности
Язык ненависти
@%?&!
$*#!?*#

3. Рекомендации по использованию ИИ в модерации контента с учетом прав человека

3.1 Рекомендации по поводу обеспечения прозрачности

Рекомендации по обеспечению прозрачности алгоритмов

- **Государства должны обязать интернет-посредников предоставлять документацию об инструментах искусственного интеллекта, используемых ими для модерации контента.** Вся обнародованная информация должна быть понятной и доступной для всех пользователей. Платформы обязаны обнародовать информацию о целях проведения анализа защищенных личных данных пользователей (т.е. данных об их возрасте, расовой и половой принадлежности, ограниченных возможностях и др.) или групп пользователей, данных о членстве в сообществах и о доверенных лицах. Платформы обязаны раскрывать следующую информацию:

    - Обучающие данные, включая содержание и происхождение наборов данных, используемых для обучения алгоритмов; методы обучения моделей ИИ; переменные/функции/характеристики, влияющие на алгоритмическое курирование контента; рекомендательные системы и/или системы ранжирования пользователей (например, по возрасту, полу и т.д.) и информацию о том, в какой степени пользователи могут контролировать эти переменные; процессы управления обучающими данными (например, их сбор, хранение, предварительная и последующая обработка, передача, хранение).

    - Информацию об услугах по обогащению данных, таких как предварительная подготовка и очистка данных (аннотирование/маркировка данных, анализ настроений, распознавание изображений, проверка преобразования речи в текст и т.д.) и выполнение задач с присутствием человека в контуре управления

(модерация контента человеком, разработка непрерывного цикла обратной связи, проверка алгоритмических результатов и моделей и т.д.), включая документирование данных о человеке или группе лиц, занимающихся обогащением данных, а также сведения об уровне их подготовки.

- Процессы и результаты тестирования, оценки и валидации этих моделей, включая оценку качества и точности.

- **Государства должны обязать интернет-посредников документировать контент-ориентированные модели.** Необходимо законодательно обязать посредников обнародовать критерии, параметры и функции, используемые в моделях машинного обучения, предназначенных для курирования или модерации контента, либо иного метода анализа данных или распознавания шаблонов. Это включает в себя дезагрегирование данных для моделей машинного обучения, предназначенных для изъятия и удаления пользовательского контента, а также для моделей, предназначенных для усиления или ослабления «теневой блокировки» и «дерэнкинга» контента. Любая раскрытая информация должна быть понятной и доступной для всех пользователей, при этом обеспечивая им конфиденциальность и защиту их собственных данных.

- **Государства должны обеспечить разработку множества наборов данных, основанных на различных атрибутах, поскольку только измеряемые и регистрируемые атрибуты могут быть использованы в целях обучения или оценки алгоритма.** Многие широко доступные наборы данных основаны на неизменяемых характеристиках (таких как этнические группы) или категориях, которые регистрируются и регулируются государством (например, юридический пол, финансовый доход или профессия). При этом часто не отслеживаются такие характеристики, как сексуальная ориентация и гендерная идентичность. Это значительно ограничивает возможности противодействия межсекторной дискриминационной предвзятости, присущей некоторым алгоритмическим системам.

- **Государства должны обеспечить прозрачность и защиту прав человека при использовании систем ИИ в государственном секторе, в том числе для модерации контента.** Государствам следует усилить контроль и ввести строгие требования к прозрачности информации в случаях использования систем ИИ государственными ведомствами, например, при применении технологий распознавания лиц или мониторинга контента, распространяемого на онлайн-платформах.

- **Государства должны обязать интернет-посредников уведомлять пользователей о применении в отношении них автоматизированных процессов или об использовании автоматизированных систем для модерации контента третьих лиц, а также обязать платформы разъяснять принципы работы таких механизмов.** Платформы обязаны подробно информировать пользователей об основаниях для удаления контента, указывая, какое именно правило было нарушено, и уведомляя о возможности запросить проведение проверки с участием человека.

- **Государства обязаны обнародовать все направленные интернет-посредникам запросы и полученные на них ответы**, а также требовать от платформ раскрытия информации о корректировке или изменении моделей машинного обучения, используемых для модерации потенциально противоправного контента, в результате запроса со стороны государства.

- **Государства должны обязать интернет-посредников предоставлять исследователям и организациям гражданского общества доступ к наборам данных и моделям** с целью их оценки и проведения исследований в общественных интересах. При необходимости можно создавать институциональные наблюдательные советы и внедрить независимый процесс аккредитации.

- **Государства должны требовать доказательств целесообразности используемых инструментов мониторинга.** Таким доказательством может стать, например, подробное описание случаев, в которых использование автоматизированных инструментов позволило верно идентифицировать противоправный

контент, чего невозможно было бы достичь при помощи неавтоматизированных средств. Доказательства целесообразности важны для решения вопроса о необходимости применения того или иного инструмента. В конечном итоге, только доказанная целесообразность является тем критерием, который можно сопоставить с ущербом, нанесенным в результате ограничения прав человека, и таким образом оценить соразмерность примененного инструмента.

Рекомендации по обеспечению прозрачности, ориентированной на пользователя

- **Государства должны обязать интернет-посредников надлежащим образом раскрывать информацию о том, что в отношении пользователя применяется или будет применен процесс алгоритмического принятия решений, включая модерацию контента, и предоставлять пользователям возможность как минимум отклонить такое принятие решений.** Пользователям следует предоставить возможность контролировать инструменты обнаружения модерации контента. В идеале такая возможность по умолчанию обеспечивается механизмом предоставления соответствующего согласия (англ. «opt-in»). Надлежащее уведомление позволяет отдельным пользователям при желании отклонить автоматизированное принятие решений в отношении себя. Интернет-посредники должны разработать процедуры предоставления согласия и обеспечения конфиденциальности, обеспечивающие пользователям возможность информированного выбора в соответствии с законами о защите данных.

- **Государства должны обеспечить пользователям доступ к имеющимся у интернет-посредников данным профилирования**[22] **включая любые выводы в отношении пользователей.** Эти данные должны по запросу предоставляться

22 Согласно Общему регламенту по защите данных (GDPR), «профилирование» — это любая форма автоматической обработки персональных данных для анализа или предугадывания личных аспектов физического лица; это означает, что «простая оценка или классификация лиц на основе характеристик» может считаться профилированием вне зависимости от того, является ли ее целью предугадывание.

пользователям в понятном и доступном для них формате. Пользователи также должны иметь возможность редактировать и удалять свои профили. Хотя в Европейском союзе это право в значительной степени обеспечивается Общим регламентом по защите данных (GDPR), существует необходимость создания эффективных и доступных процедур или интерфейсов, предоставляющих пользователям доступ к такой информации. Поэтому минимальные стандарты обязательств по обеспечению прозрачности, ориентированной на пользователя, изложенные в пункте «f» части второй статьи 13 и пункте «g» части второй статьи 14 Общего регламента, должны быть обязательными для государств региона ОБСЕ.

- **Государства должны включить в законодательство требование, обязывающее интернет-посредников предоставлять разъяснительную информацию об используемых моделях, исходных данных, показателях эффективности и тестировании моделей машинного обучения, излагая эту информацию связно, на понятном и соответствующем возрасту пользователя языке.** Такая информация предоставит пользователям возможность оспорить алгоритмическое принятие решений и/или отказаться от использования сайта. Право возражать против использования автоматизированных систем принятия решений должно существовать и в случае участия человека в процессе принятия решений.

- **Государства должны обязать интернет-посредников надлежащим образом разъяснять пользователям алгоритмический процесс принятия решений.** Как минимум, следует разъяснять пользователям принцип принятия конкретных автоматизированных решений в отношении модерации контента, чтобы предоставить им возможность оспаривать такие решения. Разъяснение такого рода должно быть составлено на понятном пользователю языке и включать в себя статистические данные, использованные для принятия решения, а также подробное изложение политики посредника, лежащей в основе принятого решения.

Рекомендации в отношении требований к прозрачности, необходимых для обеспечения эффективного доступа к средствам правовой защиты и возмещения ущерба лицам, ставшим объектами ненавистнической риторики

- **Государства должны обязать интернет-посредников предоставлять обоснование принятых решений, с разъяснением процесса и конкретных действий в отношении контента, помеченного как ненавистнические высказывания.** Обоснованное решение о мерах, принимаемых в отношении подобного контента, должно быть доведено до сведения всех пользователей, которых оно касается, с разъяснением прав каждой заинтересованной стороны и четко сформулированными инструкциями о порядке обжалования принятого решения. Такое же правило должно применяться и в отношении встречных уведомлений, независимо от того, будут ли они отклонены или же решение будет принято в пользу контент-провайдера.

- **Государства должны обязать интернет-посредников сохранять все данные об удалении контента в соответствии со стандартами защиты данных.** Это включает, в частности, информацию о том, какие решения об удалении не прошли проверку человеком, о попытках пользователей обжаловать принятое решение, а также о жалобах, по которым не было принято никаких действий в отношении контента. Кроме того, интернет-посредники должны по возможности включать в свои отчеты о прозрачности статистику и информацию о категориях ненавистнической риторики, в отношении которых были приняты меры (например, о том, какие именно защищенные характеристики были нарушены), о проценте и количестве успешных апелляций, а также о предоставленных средствах правовой защиты.

- **Государства должны обязать интернет-посредников сохранять весь контент, классифицированный как ненавистнический, автоматически заблокированный и удаленный,** включая отдельные сообщения, видео, изображения и целые аккаунты. При условии соблюдения требований к защите

данных и конфиденциальности, этот контент должен по запросу предоставляться исследователям, чтобы обеспечить дополнительный надзор за механизмами возмещения ущерба, справедливым разрешением претензий и эффективностью механизмов обжалования, особенно в отношении маргинализированных групп.

Рекомендации в отношении требований к прозрачности, необходимых для эффективного общественного надзора

- **Государства должны назначить надзорные органы, обладающие опытом в области обеспечения равноправия и недискриминации, и наделить их соответствующими полномочиями, с целью обеспечения мониторинга и недопущения несправедливого или дискриминационного воздействия автоматизированно принимаемых решений на маргинализированные группы.** В качестве таких органов могут выступать национальные институты по правам человека, институты омбудсменов или уполномоченных по вопросам информации и конфиденциальности, которые дополняют деятельность национальной судебной системы. Крайне важно, чтобы государство обеспечило таким органам возможность выполнения надзорных функций, предоставляя им адекватные и значимые законодательные полномочия, а также достаточное и гарантированное количество ресурсов.

- **Органы по обеспечению равноправия должны иметь возможность проводить стратегические судебные процессы для оспаривания дискриминационных результатов автоматизированно принимаемых решений.** Эти органы должны получать достаточную финансовую поддержку и иметь штат сотрудников, занимающихся конкретно этой тематикой и работающих над повышением прозрачности процесса применения автоматизированных мер.

- **Государства должны гарантировать, что обязательные требования к прозрачности в отношении интернет-**

посредников будут уделять особое внимание качеству, а не количеству. Цифры сами по себе служат лишь основной для сравнения, но не содержат ценной информации о том, каким образом интернет-посредники работают с пользовательским контентом. Поэтому посредники должны включать в свои отчеты о прозрачности такие данные, как количество всех полученных уведомлений; категория выдавших их субъектов, включая частных лиц, административные органы или суды; причины проведения оценки законности контента или нарушения интернет-посредником условий предоставления услуг; а также информацию о том, кем был отмечен контент – частными лицами, автоматизированными инструментами или доверенными модераторами контента.

- **Государства должны закрепить в законодательстве требование о представлении отчетов о прозрачности, дающих четкое представление о применяемых методах модерации контента,** включая его удаление, демонетизацию или деприоритизацию, приостановку работы аккаунта, удаление аккаунта или любые другие действия в отношении отмеченного контента или аккаунтов пользователей.

- **Государства должны установить минимальные требования к отчетам о прозрачности,** в том числе указать конкретные сроки уведомления поставщика контента о применении каких-либо мер; конкретные сроки подачи встречного уведомления; точный срок до момента ограничения контента; сроки процедуры апелляции; количество полученных апелляций и способы их разрешения.

- **Конкретно в отношении использования языка ненависти (ненавистнической риторики), государства должны обязать посредников публиковать количество получаемых в год сообщений об оскорбительном или токсичном поведении в интернете.** Сюда необходимо включать сведения о том, какое количество таких сообщений содержат ненавистническую риторику, направленную против таких защищенных характеристик, как расовая, этническая, религиозная или половая принадлежность.

Особое внимание необходимо уделить комплексному анализу того, каким образом в различных формах дискриминационного обращения могут быть сгруппированы такие индивидуальные характеристики, как раса, класс, пол и др.

- **Государства должны обязать интернет-посредников публиковать агрегированные данные о количестве модераторов контента, имеющихся у них в каждом регионе, а также о том, на каком языке работают эти модераторы.** Они должны предоставлять конкретную информацию о процессе обучения модераторов методам выявления потенциально опасного контента в плане гендерной и иной идентичности, а также их ознакомления с международными стандартами в области прав человека.

- **Государства должны обнародовать полноценные отчеты о мерах, предпринимаемых ими в ответ на распространение законного, но потенциально деструктивного контента.** Государственные органы обязаны регулярно предоставлять общественности полноценную информацию о количестве, характере и правовых основаниях всех запросов на ограничение контента, направленных интернет-посредникам; о действиях, предпринятых на основании этих запросов; и о случаях ограничения контента на основании договоров о взаимной правовой помощи.

Рекомендации по структурам предоставления доступа к данным независимым заинтересованным сторонам, обладающим соответствующим опытом

- **Государства должны ввести требование в адрес интернет-посредников о составлении внешних отчетов**, доступ к которым должен предоставляться всем соответствующим независимым заинтересованным сторонам и государственным органам, включая исследователей и организации гражданского общества. Необходимо обязать интернет-посредников предоставлять возможность для проведения независимого внешнего аудита

любой автоматизированной модели при условии сохранения коммерческой тайны и конфиденциальности/сохранности данных.

- **Государства должны ввести требование в адрес онлайн-платформ о предоставлении доступа к данным и составлении внешних отчетов**, доступ к которым должен предоставляться всем соответствующим независимым заинтересованным сторонам и государственным органам, включая исследователей, организации гражданского общества и пострадавших пользователей. Платформы обязаны предоставлять возможность для проведения независимого внешнего аудита своих алгоритмических моделей при условии сохранения коммерческой тайны и конфиденциальности/сохранности данных. Государствам следует ввести критерии для обеспечения независимости и компетентности аудиторов.

- **Любое законодательство или политика в области управления контентом, инициированные государствами, должны основываться на фактах и исследованиях.** Органам государственной власти следует предоставить полноправный доступ к данным, хранящимся у интернет-посредников, в соответствии с адекватными требованиями к защите данных, с тем чтобы они могли разрабатывать политику на основании фактических данных и обеспечивать адекватный независимый общественный надзор. С этой целью государства должны установить требования в отношении предоставления доступа к данным третьим сторонам, четко сформулировав, кто именно может получить доступ к данным, и к каким именно, кто и каким образом должен собирать и проверять данные перед предоставлением доступа к ним.

- **Государства должны установить критерии независимости и компетентности аудиторов.** Интернет-посредники должны добровольно проходить регулярную комплексную и эффективную аудиторскую оценку со стороны независимых аудиторских органов, обладающих необходимой компетенцией, которым необходимо предоставить описание потенциальных правовых или иных последствий системы. Тем не менее, такая оценка рисков

всегда осуществляется в качестве вторичной меры, в то время как первоочередной мерой должна быть предварительная оценка потенциального воздействия на права человека, проводимая под общественным надзором.

- **Организациям гражданского общества, научным деятелям, проводящим исследования в интересах общества, и журналистам следует предоставить возможность осуществлять содержательный мониторинг и аудит автоматизированных систем принятия решений.** Независимые третьи стороны, осуществляющие аудит, должны иметь доступ ко всей необходимой им информации, такой как исходный код, критерии данных и показатели эффективности, для проведения эффективного надзора за саморегулированием интернет-посредников. Полученная информация должна предоставлять третьим сторонам возможность проводить аудит и составлять отчеты о функционировании, эффективности или ошибочности конкретных автоматизированных решений о том, какой контент должен быть удален, а какой может остаться у интернет-посредника.

3.2 Рекомендации по соблюдению прав человека при управлении контентом

- Государства должны разработать политику в области прав человека, уделяя особое внимание таким важнейшим правам человека, как право на свободу выражения мнения и свободу средств массовой информации, право на неприкосновенность частной жизни, право на недискриминацию и право на жизнь, свободу и безопасность. Государства несут ответственность в рамках международного права в области прав человека и должны выполнять позитивное обязательство по защите прав человека от вмешательства других лиц, включая частные организации или физических лиц. Таким образом, государства должны придерживаться юридически обязательных требований международного права в сфере прав человека, и обеспечить полное отражение этих требований в национальном

законодательстве, регулирующем функционирование платформ и управление контентом.

- **Государствам следует воздержаться от включения в законодательство требования к онлайн-платформам о применении автоматизированных инструментов для обнаружения и идентификации потенциально незаконного или опасного контента**, что в некоторых юрисдикциях называется «упреждающей мерой».

- **Государства должны предоставить четкие указания о том, какой контент классифицируется как незаконный в соответствии с применимым законодательством.** Независимые судебные органы должны дать подробную оценку того, что представляет собой незаконный контент, а также разграничить различные типы/категории такого контента. Государства должны требовать от платформ раскрытия информации о том, какие автоматизированные инструменты используются в отношении тех или иных категорий незаконного контента, и каким образом они функционируют, а также каковы предполагаемые цели их использования (обнаружение, идентификация, стирание/удаление, управление доступом к трафику, усиление/ослабление, «теневой запрет» и др.).

    - **Государства должны гарантировать и закрепить в законодательстве надлежащие меры по защите прав человека при алгоритмической модерации контента** путем создания механизмов смягчения негативных последствий использования систем ИИ для модерации и курирования пользовательского контента, включая ненавистническую и противоправную риторику. С этой целью можно обязать интернет-посредников проводить правозащитную экспертизу систем ИИ для обнаружения, идентификации и устранения потенциально деструктивного контента. Посредники обязаны оценивать точность и частоту ошибок в системе ИИ, а также потенциальный вред от так называемых ложноотрицательных и ложноположительных

результатов работы этих систем, при этом предпринимая усилия для предотвращения и смягчения дискриминационных последствий использования систем ИИ в целом и уделяя особое внимание защите свободы выражения мнения и свободы медиа. Большое значение имеют разнообразные наборы данных, а также знание и понимание местного контекста, лингвистических нюансов и закодированного языка.

- **Государствам следует включить в законодательство требование о проведении должной проверки соблюдения прав человека в бизнес-моделях процессов сбора данных.** Бизнес-модели процессов сбора данных могут усилить негативное воздействие на права человека, поощряя потенциально законный, но деструктивный контент в интернете. Интернет-посредники, чьи бизнес-модели зависят от целевой рекламы и массового сбора и анализа пользовательских данных, должны запрашивать у пользователей предварительное согласие на сбор данных и персонализированную модерацию и курирование контента. Как минимум, такие интернет-посредники должны обеспечивать пользователям возможность выразить отказ от сбора данных и/или алгоритмической модерации контента, а также предоставлять альтернативные средства обеспечения безопасности пользователей в сети.

    - **Целесообразно будет предоставить механизм «согласия по умолчанию» для алгоритмических систем модерации контента,** поскольку такой механизм обеспечивает большую защиту тем пользователям, которые недостаточно хорошо осведомлены о принципах работы этих систем. Интернет-посредники должны разрабатывать политику «предоставления согласия» и политику конфиденциальности, которые предоставят пользователям возможность информированного выбора и обеспечат соблюдение законов о защите данных. Механизм «предоставления согласия» должен позволять пользователям хотя бы в минимальной степени осуществлять контроль над рекомендательными системами.

Рекомендации по обязательной оценке влияния систем ИИ на права человека

- **Государства обязаны требовать проведения прозрачной, независимой и всесторонней предварительной оценки воздействия на права человека (ОВПЧ) в рамках четко сформулированной нормативной базы и под надзором регулирующего органа или независимых сторон, обладающих соответствующей квалификацией.** Оценка должна включать анализ продуктов, услуг и систем посредников и их воздействия на права человека, уделяя особое внимание праву пользователей на свободу выражения мнения, а также проблемам, связанным с плюрализмом медиа. Предварительная оценка воздействия на права человека должна проводиться максимально открыто и прозрачно, при активном участии отдельных лиц и групп, пострадавших от ненавистнической или противозаконной риторики. В оценке также должны участвовать представители затронутых сообществ и групп заинтересованных сторон, включая гражданское общество и маргинализированные группы. Результаты оценки воздействия на права человека должны быть обнародованы и изложены на доступном и понятном языке.

- **Государства должны обязать компании проводить регулярную оценку воздействия моделей алгоритмической модерации контента на права человека** на своих платформах в течение всего жизненного цикла систем искусственного интеллекта. Компании должны осуществлять значимое взаимодействие с внешними заинтересованными сторонами, обладающими соответствующим опытом в области прав человека, а также проектирования, разработки и развертывания систем модерации незаконного контента в интернете. Особое внимание в рамках такого взаимодействия следует уделять применению инклюзивных и интерактивных подходов в отношении маргинализированных и уязвимых групп. Планирование, методология и результаты оценки воздействия на права человека должны соответствовать

признанной передовой практике и быть общедоступными. При проведении оценки необходимо также учитывать вопросы соразмерности, что, в свою очередь, требует оценки целесообразности/необходимости того или иного вмешательства и оценки наносимого им ущерба правам человека.

- **Государства должны требовать от интернет-посредников разработки внутренних процессов, позволяющих выявлять и предотвращать риски ущемления прав человека.** Все интернет-посредники должны создавать соответствующие внутренние механизмы, включая внутренний аудит, хотя структура и масштаб таких механизмов зависят от размера компании-посредника. В случаях применения ненавистнической риторики особенно необходимы критерии оценки риска, помогающие определить, оказывает ли подобная риторика несоразмерное воздействие на отдельных представителей или группы из числа маргинализированных сообществ, и если да, то как именно. Особое внимание следует уделить межсекторальному анализу того, каким образом комбинации таких атрибутов, как раса, класс, пол и другие индивидуальные характеристики приводят к различным формам дискриминационного обращения.

3.3 Рекомендации по обеспечению эффективных средств правовой защиты и возмещению ущерба

- **Государства должны требовать от интернет-посредников создания оперативных механизмов рассмотрения жалоб.** Во-первых, пострадавший пользователь должен иметь возможность запросить дополнительную информацию о работе алгоритмического инструмента модерации контента, особенно если в результате его работы контент был удален. Во-вторых, пользователю необходимо предоставить возможность запрашивать проверку с участием человека. В-третьих, пользователи должны иметь доступ ко всей информации, необходимой для обжалования решения, в том числе в судебные инстанции. Помимо прочего,

это включает в себя информацию о цели использования алгоритмического инструмента модерации контента, условиях его применения, метриках оценки (ложноположительные / ложноотрицательные результаты) и т.д.

- **С целью обеспечить доступ пользователей к эффективным средствам правовой защиты, государства должны требовать предоставления конкретного обоснования решений по управлению контентом**, независимо от того, принимались ли эти решения на основании проверки человеком или автоматизированной проверки. Пользователей необходимо уведомлять о касающихся их решениях по модерации контента, включая его удаление, а также демонетизацию, приостановление действия или удаление аккаунта.

- **Государства должны обязать интернет-посредников предоставлять пользователям реальные возможности для обжалования принятого решения.** Это особенно важно в случае удаления контента и приостановления действия аккаунта. Процесс обжалования (апелляции) должен быть доступным и своевременным и предусматривать эффективные средства правовой защиты, в том числе восстановление удаленного контента или отмену решения о приостановлении действия аккаунта. Если первоначальное решение в отношении контента было принято автоматизированными средствами, процесс обжалования должен включать проверку, осуществляемую человеком. В случае отклонения поданной пользователем апелляции, пользователю должны быть предоставлены четкие аргументы в обоснование принятого решения. При обжаловании решения об удалении своего контента или приостановления действия своего аккаунта, пользователь должен иметь возможность представить дополнительные доказательства. Процедура обжалования на уровне интернет-посредника может предусматривать такие средства правовой защиты, как отмена решения, принесение извинений, предоставление развернутого ответа, предоставление объяснений, внесение исправлений, восстановление аккаунта или

же совокупность нескольких средств правовой защиты. Однако эти средства правовой защиты не должны заменять собой эффективные судебные средства правовой защиты и восстановление справедливости в судебном порядке. В целом, онлайн-платформы должны обеспечить дополнительную проверку с обязательным участием человека".

- **Государства должны поощрять политику и исследовательские инициативы, направленные на изучение влияния дизайна интерфейса на поведение пользователей, а также на решение таких проблем, как обманчивые интерфейсы, известные как «темные паттерны» (англ. «dark patterns»). Помимо автоматизированных средств обнаружения,** интернет-посредники по-прежнему полагаются на сообщения пользователей о преследованиях и травле на своих сайтах. Пользователи имеют право на соответствующее возмещение ущерба, нанесенного направленной против них ненавистнической риторикой.

- **Интернет-посредники должны совершенствовать дизайн интерфейса существующих механизмов информирования о злоупотреблениях, обеспечивая их доступность и эффективность, с учетом возраста и интересов пользователей.** Интернет-посредники должны регулярно собирать отзывы и интересоваться мнением пользователей и организаций гражданского общества, особенно представляющих исторически маргинализированные и подверженные риску группы, с целью повышения эффективности и доступности своих механизмов информирования. Кроме того, пользователям, пометившим какой-либо контент как неподобающий, следует предоставлять информацию о принятом решении и мерах в отношении контента, о котором они сообщили.

- **Государства должны требовать четкого формулирования и доступности стандартов сообщества и условий предоставления услуг, на которых основываются решения**

о модерации контента. Необходимы понятные правила и инструкции о допустимом и недопустимом использовании услуг интернет-посредника, а также о последствиях нарушения условий предоставления услуг. Такая прозрачность необходима как для отдельных пользователей, так и в целях обеспечения полноценного общественного и государственного надзора. Интернет-посредники должны регулярно предоставлять всем пользователям исчерпывающую и изложенную понятным языком информацию об изменениях в условиях предоставления услуг.

- **Государства должны создавать благоприятные условия для проведения общественных дискуссий, включая свободу медиа.** Государства должны предпринимать упреждающие и проактивные меры по борьбе со структурными и институционализированными формами ненависти и распространением языка ненависти в интернете. К таким мерам относятся инициирование и поддержка информационных кампаний для информирования общественности – особенно пользователей платформ социальных сетей – о том вреде, который наносят людям направленные против них преследования и оскорбления в интернете, а также об их влиянии на общество и негативном воздействии на маргинализированные группы. Такие проактивные меры должны также включать в себя инвестиции в исследования о потенциальном позитивном использовании ИИ для создания более безопасных, управляемых сообществами онлайн-пространств; инициативы по пресечению применения ненавистнической риторики, выходящие за рамки удаления или приостановки действия аккаунтов; создание возможностей и форумов для диалога между интернет-посредниками, гражданским обществом и маргинализированными группами для улучшения практики выявления и модерации высказываний в сети.

3.4 Рекомендации по позитивному использованию ИИ для создания безопасных и управляемых сообществом пространств для маргинализированных групп

- **Государства должны призвать интернет-посредников к предоставлению маргинализированным сообществам права участвовать в принятии решений в процессе разработки и внедрения новых продуктов ИИ.** Многие из них обладают знаниями и опытом, позволяющими осуществить весь процесс, от этапа подготовки данных до стадии развертывания системы ИИ, таким образом, чтобы максимально увеличить положительное воздействие и снизить до минимума внешние факторы, негативно влияющие на исторически маргинализированные и подверженные риску сообщества.

- **Государства должны поддерживать существующие идеи в области ИИ, инициируемые самими маргинализированными сообществами или проводимые в их интересах, а также содействовать усилиям по углублению знаний и расширению возможностей таких сообществ для использования ими потенциально полезного ИИ.** Некоторые группы уже разрабатывают рекомендации для сообществ и проводят семинары о методах использования ИИ для расширения возможностей сообществ, а не для наблюдения за ними или их дальнейшей маргинализации.[23]

- **Государства должны поддерживать не отдельные «универсальные решения», а широкое разнообразие подходов.** Примером может служить дополнение для браузера, цель которого – выявлять и удалять женоненавистнический контент, подобно принципу работы блокировщика рекламы. Некоторые специалисты в сфере одноранговых (P2P) технологий выступают за использование методов субъективной модерации.[24] В этом

23 Например, см. <https://alliedmedia.org/wp-content/uploads/2020/09/peoples-guide-ai.pdf>.

24 Emmi Bevensee, The Decentralised Web of Hate: White supremacist are starting to use Peer-

случае может существовать целый ряд алгоритмов модерации ИИ, а отдельный пользователь может активировать одну или несколько систем одновременно. Такие инструменты, основанные на алгоритмах, не цензурируют все содержание сообщения, а просто меняют предлагаемый пользователю контент в индивидуальном порядке. Широкий выбор стратегий позволяет находить надежные решения путем прикладного экспериментирования.

- **Государства должны по возможности способствовать использованию открытого исходного кода в существующих частных моделях и обеспечивать обратную связь с сообществом при их реализации.** При том что модели с закрытым исходным кодом бывают более прибыльными, такие проекты, как Hugging Face, показывают, что мир ИИ может быть надежным, прибыльным и в то же время открытым.[25]

3.5 Рекомендации по формализации сотрудничества с правоохранительными органами

- **Государства должны адекватно применять гарантии, запрещающие обязательную передачу данных**, особенно правоохранительным органам, и в этом плане принимать конкретные меры по защите маргинализированных и уязвимых групп.

- **При мониторинге и/или отслеживании контента в интернете государства должны придерживаться международного права в области прав человека, включая проверку на соответствие трем критериям допустимости ограничений свободы выражения мнения.** Отдавая распоряжение о мониторинге и отслеживании контента или давая платформам указание об удалении контента, идентичного или схожего с контентом, ранее признанным незаконным, государства должны убедиться в том, что эти меры предусмотрены законом, преследуют легитимную цель, являются необходимыми и задействуют наименее интрузивные средства для эффективного достижения цели. В частности, должна

to-Peer technologies. Are we prepared? At <https://rebelliousdata.com/p2p/>.

25 Более подробную информацию можно найти на сайте <https://huggingface.co/>.

быть четко определена легитимная цель любой государственной меры, приводящей к использованию инструментов ИИ для управления контентом, а преимущества такой меры должны быть четко изложены, чтобы продемонстрировать, насколько эти доказанные преимущества соразмерны ущербу, наносимому правам человека.

4. Заключение

Распространение потенциально незаконного и деструктивного контента в интернете, а также последствия алгоритмического принятия решений по-прежнему представляют собой сложный вопрос со множеством нюансов. Спустя почти пять десятилетий после подписания Хельсинкского Заключительного акта сегодня как никогда необходимо сотрудничество между государствами-участниками ОБСЕ для решения новых проблем, связанных с модерацией интернет-контента и распространением незаконной информации, а также законной, но деструктивной ненавистнической риторики в глобальной сети. Это особенно актуально с увеличением количества мощных интернет-посредников, которые выступают в роли привратников и модераторов мнений на этом новом цифровом рынке идей, важность которого растет с каждым днем.

Целью данного раздела доклада является формулирование принципиального подхода к регулированию незаконного контента и законной, но деструктивной ненавистнической риторики в интернете, уделяя особое внимание воздействию подобной риторики и алгоритмического принятия решений на маргинализированные группы. Государства несут основную ответственность за соблюдение, продвижение и реализацию прав человека, включая право на свободу выражения мнения, право на свободу медиа и право на защиту от дискриминации. Эта ответственность предполагает эффективное регулирование деятельности интернет-посредников на всех этапах процесса, от дизайна и разработки алгоритмических моделей до средств правовой защиты, доступ к которым необходимо обеспечить пострадавшим лицам и группам.

В данном разделе изложен ряд упреждающих, превентивных и ответных рекомендаций, которые призваны помочь государствам-участникам

ОБСЕ в решении этой задачи. Рекомендации касаются различных соответствующих аспектов, таких как обеспечение прозрачности алгоритмов; проведение надлежащей проверки на соблюдение прав человека; обеспечение доступа к эффективным средствам правовой защиты и возмещения ущерба; эффективное вовлечение гражданского общества и пострадавших сообществ на всех этапах жизненного цикла алгоритмического инструмента; и содействие позитивному использованию ИИ с целью создания безопасных и управляемых сообществом пространств для маргинализированных групп.

ИСПОЛЬЗОВАНИЕ ИИ В КУРИРОВАНИИ КОНТЕНТА

Курирование контента и плюрализм медиа

Эта часть посвящена **использованию ИИ в целях курирования контента** и содержит анализ влияния систем рекомендации контента на основе данных на разнообразие и плюрализм медиа. Эта и следующая глава, посвященная недостаткам курирования контента на основе ИИ и целевой рекламы, содержат рекомендации по предотвращению негативного воздействия инструментов ИИ, предназначенных для курирования контента, на право человека на свободу мнений и их выражения.

1. Определение масштабов влияния процессов курирования контента на плюрализм медиа

1.1 Влияние алгоритмического курирования контента и рекомендательных систем на основе анализа данных на плюрализм и разнообразие медиа

Разнообразие и плюрализм медиа представляют собой основные демократические принципы, качество реализации которых зависит от увеличения числа доминирующих интернет-посредников и их влияния на общественный дискурс. Интернет-посредники, в частности, платформы социальных сетей, стали важным источником, точками доступа и ключевыми распространителями информации, включая новостной контент. Распространение и растущая агрегация информации происходит в основном посредством систем алгоритмического курирования контента[26] и рекомендательных систем. Используя методы оптимизации и анализа связанных и несвязанных с человеком аспектов, эти системы «поставляют» пользователю персонализированный контент, адаптированный к его индивидуальному профилю, поэтому каждый пользователь получает различные категории и объем контента. Системы

26 Курирование контента можно понимать как набор алгоритмических и управляемых человеком процессов, которые поддерживают распространение контента среди аудитории, например, ранжирование контента или анализ редакционных данных. См.: Б. Буковска и др., «В фокусе – искусственный интеллект и свобода слова» #SAIFE (2020), стр.19.

рекомендации контента, которые ранжируют контент с целью определить, что именно следует рекомендовать тому или иному пользователю, тем самым ограничивают свободу пользователя искать и передавать информацию, а также влияют на общий информационный ландшафт и свободу медиа. Дизайн рекомендательных систем существенно влияет на то, что пользователь может увидеть в интернете, а что остается скрытым – и для кого. Процесс алгоритмического курирования контента также определяется[27] ценностями и целями создателя алгоритма,[28] социально-техническими факторами, саморегулированием (например, условиями предоставления услуг) и государственным регулированием. Учитывая, насколько вездесущим стал интернет-контент и какую огромную роль он играет в формировании мнений и принятии решений, возникает ключевой вопрос: на ком лежит ответственность за определение и реализацию политики, направленной на приоритизацию и кодификацию в контексте плюрализма и разнообразия медиа[29] в эпоху цифровой информации?

В этой части доклада приводится концептуальный обзор ключевых процессов алгоритмического курирования контента и их преобразующего воздействия на плюрализм медиа. Кроме того, в ней содержится ряд рекомендаций в адрес государств-участников ОБСЕ по ориентированному на права человека подходу к алгоритмическому курированию контента. По существу, эта часть посвящена влиянию алгоритмического курирования контента и основанных на анализе данных рекомендательных систем на плюрализм и разнообразие медиа в демократическом обществе, а также роли государства как основного гаранта права человека на свободу выражения мнения, который несет ответственность за создание благоприятных условий для реализации этого права.

27 K. Klonick, The New Governors: The People, Rules, and Processes Governing Online Speech, The Harvard Law Review, p. 1664.
28 Radsch, Courtney. "Digital Information Access." In A New Global Agenda: Priorities, Practices, and Pathways of the International Community, edited by D. Ayton-Shenker, 72–83. Rowman & Littlefield Publishers, 2018, <https://books.google.com/books?id=tyJJDwAAQBAJ>.
29 Philip M Napoli, "Rethinking Program Diversity Assessment: An Audience-Centered Approach" (1997) 10 Journal of Media Economics 59-74.; N. Helberger & M. Wojcieszak (2018). Exposure Diversity. In P. M. Napoli (Ed.), Mediated Communication (pp. 535-560). (Handbooks of Communication Science; том 7). De Gruyter Mouton, <https://doi.org/10.1515/9783110481129-029>.

1.2 Противоречия между алгоритмическим курированием контента и свободой самовыражения

Возможность фильтровать, приоритизировать и регулировать интернет-контент с учетом личных предпочтений и интересов пользователя часто противоречит праву человека искать, получать и распространять разнообразную информацию.[30] В качестве базового принципа интернет-посредники обычно приоритизируют и показывают контент пользователю, основываясь на спрогнозированном системой потенциальном сценарии взаимодействия пользователя с этим контентом. Подобно системам персонализированной и поведенческой рекламы, системы рекомендации контента собирают данные о пользователях и иных лицах для создания цифровых профилей, оценки сходства между ними и формирования выводов на основе анализа этих данных.

Модель многих онлайн-платформ,[31] для которых вовлечение пользователей и получение прибыли являются более приоритетной задачей, чем применение подходов с учетом прав человека, может привести и приводит к эксплуатации и навязчивому использованию данных, распространению дезинформации и созданию алгоритмических контуров обратной связи.[32] Доказано, что такая модель оказывает негативное влияние на плюрализм контента, особенно созданного маргинализированными сообществами или для них. Эта модель закрепляет пробелы в информации[33] и создает препятствия для защиты интересов пользователей, тем самым отображая и закрепляя структурное социальное неравенство. Существуют также свидетельства того, что процесс модерации контента выгоден тем группам, которые уже и так доминируют в онлайн-пространстве и нарративах, а не

30 P. Leersen, The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems, European Journal of Law and Technology (2020), p.12.
31 Ranking Digital Rights, It's the Business Model: How Big Tech's Profit Machine is Distorting the Public Sphere and Threatening Democracy (2020).
32 Bodó, B., Helberger, N., Eskens, S., & Möller, J, Interested in diversity: The role of user attitudes, algorithmic feedback loops, and policy in news personalization. Digital Journalism (2019), p. 219.
33 A. Causevic and A. Sengupta, Whose Knowledge Is Online? Practices of Epistemic Justice for a Digital New Deal, IT for Change (2020), <https://itforchange.net/digital-new-deal/2020/10/30/whose-knowledge-is-online-practices-of-epistemic-justice-for-a-digital-new-deal/>.

маргинализированным группам, информации и нарративам.[34] Более того, установлено, что системы алгоритмического обнаружения контента (например, поисковые системы) обостряют расизм, предлагая дискриминационные поисковые фразы и разграничивая изображения членов маргинализированных сообществ по расовому, языковому и гендерному признаку.[35]

В большинстве своем алгоритмические системы курирования и рекомендации контента основаны на собственных (внутренних) правилах, на интересах и предположениях посредников, а не на демократических или общественных ценностях.[36] Рекомендация контента имеет огромное значение для обеспечения развития и господства крупных интернет-посредников и лежит в основе их бизнес-моделей. Поскольку рекомендательные системы представляют собой «ключевую логику, управляющую потоками информации, от которых мы зависим»,[37] они позволяют интернет-посредникам контролировать потоки информации и знаний. Это оказывает более масштабное влияние на общественные интересы, представительство и равенство или неравенство влияния как в сети, так и за ее пределами.[38] Рекомендательные системы интернет-посредников значительно изменили логику общественной коммуникации, включая доступ к новостям, критической информации и общему контенту в интересах общества. Эти системы существенно ограничивают возможности равного доступа к информации, публикуемой журналистами и медиа, при этом ограничивая доступ самих журналистов к информации и оказывая давление на журналистику в результате того, что деньги за рекламу уходят к посредникам. Результаты последних исследований в области

34 B. Marshall, Algorithmic misogynoir in content moderation practice, Heinrich-Böll-Stiftung (2021), p.7,11. См. также: M. E, Mazzoli and D. Tambini, Prioritisation uncovered: The Discoverability of Public Interest Content Online. Council of Europe (2020), p. 44.

35 Safiya Umoja Noble, Algorithms of Oppression How Search Engines Reinforce Racism, NYU Press (2018).

36 C. Radsch. "Digital Information Access." In A New Global Agenda: Priorities, Practices, and Pathways of the International Community, edited by D. Ayton-Shenker, 72–83. Rowman & Littlefield Publishers, 2018. <https://books.google.com/books?id=tyJJDwAAQBAJ>.

37 T. Gillespie (2018). Custodians of the internet. См. по ссылке <https://www.researchgate.net/publication/327186182_Custodians_of_the_internet_Platforms_content_moderation_and_the_hidden_decisions_that_shape_social_media>.

38 P. Leerssen, The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems. См. по ссылке <file:///Users/eliskapirkova/Downloads/Leerssen%20EJLT_corr.pdf>.

алгоритмической приоритизации, определяемой как «ряд дизайнерских и алгоритмических решений, приводящих к повышению значимости и открываемости контента»[39] продемонстрировали потенциальный риск поляризации мнений и взглядов в интернете. Например, важным фактором в процессе приоритизации контента являются индивидуальные политические предпочтения и/или принадлежность. Поэтому приоритизацию может усилить и закрепить поляризацию мнений и взглядов в интернете, особенно среди пользователей, находящихся у противоположных полюсов политического спектра, которые, вероятно, уже потребляют преимущественно аффилированный контент.[40] Исследования также показали, что «некоторые группы общества легче поддаются избирательному воздействию, чем другие».[41] В то время как активная персонализация, основанная на предоставляемых пользователем данных, обычно приводит к предоставлению более разнообразной информации, пассивная персонализация, основанная на алгоритмическом отборе контента, имеет тенденцию усугублять так называемый эффект нахождения в «информационном пузыре».[42]

Предвзятость и дискриминация, в том числе гендерная, могут возникать в основанном на этапе анализа данных в процессе алгоритмического принятия решений по нескольким причинам и на разных уровнях систем курирования контента, и их, как правило, довольно трудно выявить и смягчить. Существует мнение, что исключение сенситивной информации, зависящей от идентификации пользователя, представляет собой достаточную меру для защиты от дискриминации. Однако дискриминация может происходить и происходит, несмотря на такие «меры защиты», учитывая масштабы и разнообразие информации, содержащейся в наборах данных, которые формируются алгоритмами. Предвзятость алгоритмов может быть следствием их разработки и реализации, включая

39 M.E. Mazzoli and D. Tambini. Prioritisation uncovered: The Discoverability of Public Interest Content Online. Council of Europe (2020), p.12.
40 B. Stark, D. Stegmann, Are Algorithms a Threat to Democracy? The Rise of Intermediaries: A Challenge for Public Discourse. Retrieved from <https://algorithmwatch.org/wp-content/uploads/2020/05/Governing-Platforms-communications-study-Stark-May-2020 -Algorithm-Watch.pdf>.
41 B. Bodó, N. Helberger, S. Eskens & J. Möller, Interested in diversity: The role of user attitudes, algorithmic feedback loops, and policy in news personalization. Digital Journalism (2019), p.15.
42 D. Wagner, Artificial Intelligence and Disinformation as a Multilateral Policy Challenge, <https://www.osce.org/files/f/documents/d/0/506702.pdf>.

использование нерепрезентативных или неполных обучающих данных, или лежащих в их основе индивидуальных, опытных или ценностно-ориентированных данных, отражающих историческое/структурное неравенство. Предвзятость алгоритмов может оказывать коллективное неравное воздействие на сообщества, особенно на маргинализированные группы, даже при отсутствии намеренной дискриминации. Поэтому необходимо исследовать как преднамеренные, так и непреднамеренные последствия алгоритмов. Существующей государственной политики может оказаться недостаточно для выявления, смягчения и исправления последствий применения алгоритмов в отношении отдельных лиц или общества в целом. Помимо преднамеренных усилий по формированию индивидуального внимания (прямое манипулирование), существует также опасность внесения в алгоритм нежелательных и косвенных предубеждений путем включения больших данных на различных уровнях курирования контента. Как прямая, так и косвенная дискриминация, которая вызвана алгоритмами с использованием больших данных, является одной из наиболее актуальных угроз в процессе курирования контента, управляемого алгоритмами.

Совокупный эффект фильтрации и персонализации контента практически создает этапы (слои) ограничений в плане открываемости, а значит, и доступности разнообразного медиаконтента. Вышеупомянутые проблемы серьезно ограничивают плюрализм медиа, понимаемый как разнообразие источников информации (внешний плюрализм) и контента (внутренний плюрализм).[43] В контексте данного документа под плюрализмом медиа также понимается распределение коммуникативной власти (или «влияния») в обществе. Необходимой предпосылкой для справедливого распределения коммуникативного влияния является деконцентрация власти и децентрализации ресурсов в информационной экосистеме,[44] а также поддержка альтернативных моделей, предлагающих разнообразие нарративов и контента. Очевидно, что алгоритмизированные процессы курирования контента изменяют понятия плюрализма и разнообразия средств массовой информации, которые

43 P. M. Napoli, "Rethinking Program Diversity Assessment: An Audience-Centered Approach" (1997) 10 Journal of Media Economics 59-74.; Helberger, N., & Wojcieszak, M. (2018). Exposure Diversity. In P. M. Napoli (Ed.), Mediated Communication (стр. 535-560). (Handbooks of Communication Science; том 7). De Gruyter Mouton. <https://doi.org/10.1515/9783110481129-029>.
44 M. Moore and D. Tambini (eds) (2018) Digital Dominance: The Power of Google, Amazon, Facebook, and Apple. New York: Oxford University Press.

необходимы для обеспечения демократических публичных дебатов и создания инклюзивного общества.

Именно на этом фоне государственные и негосударственные субъекты, в первую очередь интернет-посредники и медиа-организации, а также международные и региональные организации, представители гражданского общества и научных кругов должны разработать политику, способствующую созданию благоприятной среды для плюрализма медиа. Это означает создание условий для доступности, наличия, открываемости и потребления различных категорий контента или медиаконтента через различные средства массовой информации и многочисленные каналы.

2. Алгоритмическое курирование контента и основанныенаанализеданныхрекомендательные системы: влияние на плюрализм медиа

2.1 Типология

Очевидным источником влияния, а значит и коммуникативной власти интернет-посредников и компаний-владельцев социальных сетей являются их системы рекомендации контента, которые также «придают вес их роли в демократической культуре».[45] По сути, «рекомендательная система» включает в себя различные технологии, которые фильтруют, извлекают и упорядочивают информацию для отдельных пользователей. К факторам ранжирования информации относится уровень заинтересованности в конкретном контенте, тип контента, время его первого опубликования или предыдущий опыт взаимодействия пользователей с подобным контентом. Ранжируя контент, рекомендательные системы могут формировать и изменять способность людей составлять собственное мнение.

Основная цель рекомендательных систем заключается в фильтрации больших объемов информации в интернете. Этот алгоритмически управляемый процесс работает по-разному:

- **Фильтрация, основанная на контенте:** пользователи получают рекомендации по поводу контента, основанные на их заявленных или

45 K. Klonick, The New Governors: The People, Rules, and Processes Governing Online Speech, The Harvard Law Review, p. 1663.

подразумеваемых предпочтениях. Например, если пользователю нравится классическая музыка или же он хочет получать новости о любимой спортивной команде, то рекомендательная система будет отдавать предпочтение тем элементам, которые соответствуют его интересам, и вероятно обеспечит вовлечение этого пользователя.

- **Коллаборативная фильтрация:** пользователи получают рекомендации по поводу контента, на основе информации о людях, с которыми они тесно связаны или с которыми у них есть общие характеристики (демографическая категория, предпочтения в плане контента и др.). Например, при чтении новостей система рекомендует пользователю те статьи, с которыми ознакомился или которыми поделился его/ее друг, или же при совершении покупок в интернете система рекомендует товары, которые приобрели люди с похожей историей покупок.

- **Гибридная фильтрация:** комбинация вышеупомянутых методов фильтрации и курирования контента. Например, система рекомендует пользователю понравившуюся его/ее другу новостную статью, но только если она касается определенной темы, представляющей интерес для пользователя, и объединяет эту информацию с широким спектром различных метаданных, таких как местоположение пользователя, история использования и т.д.

Все эти процессы основаны на анализе данных пользователей, их профилях и взаимодействии с данной платформой, а также на информации, полученной из лежащей в их основе архитектуры рекламных технологий. Алгоритм определяет точный способ генерации рекомендаций в отношении контента, используя методы фильтрации, основанной на контенте, и коллаборативной фильтрации. На основе пользовательских рекомендаций система создает рекомендательную стратегию, определяющую принципы объединения данных для расчета потенциальной вовлеченности с целью удовлетворения критериев оптимизации. Проще говоря, алгоритмическое курирование контента – это стратегия, используемая рекомендательной системой для определения того, как можно наилучшим образом использовать собранные данные для достижения заранее определенных целей оптимизации.[46]

46 Для достижения цели оптимизации алгоритм может подчеркнуть различные способы приоритизации полученных данных. Например, алгоритм может отдать приоритет

Все крупные интернет-посредники, в частности, платформы социальных сетей, используют так называемые «открытые системы рекомендации контента».[47] Эти системы по умолчанию используют пользовательский контент в общем списке источников рекомендателя, при этом исключая определенные элементы контента, например, на основании нарушения условий предоставления услуг. Чтобы оптимизировать вовлечение, эти системы «персонализируют» онлайн-опыт, отдавая предпочтение контенту, который предположительно является интересным для пользователя, создается с учетом его/ее предыдущей вовлеченности и поведения. Соответственно, видео, результаты поиска, новостные статьи или иная информация, которую видит пользователь, уникальна для каждого отдельного пользователя и отличается от того, что видят другие. По этой и другим причинам, алгоритмические системы рекомендации контента потенциально подрывают и нарушают демократические процессы.[48] Они способны ограничивать возможности и доступ человека к различным точкам зрения, ценностям и нарративам, тем самым создавая угрозу плюрализму и разнообразию. Для предотвращения такого воздействия может потребоваться вмешательство и контроль со стороны государства.

Алгоритмическая фильтрация и адаптация онлайн-контента с учетом предполагаемых личных предпочтений и интересов пользователя уменьшает доступ пользователей к разносторонней информации, что может негативно повлиять на разнообразие и общественный дискурс, а также на конфиденциальность. Таким образом, системы курирования контента способны серьезно ограничить источники информации,

актуальности новостной статьи. Другая стратегия может заключаться в рассмотрении популярности статей в качестве критерия ранжирования для выбора окончательных рекомендаций, рассчитанных на конкретного пользователя. Еще одним часто используемым способом курирования контента является точность. Оптимизированный по точности подход направлен на то, чтобы как можно точнее смоделировать предпочтения пользователя. В этом случае вырабатываются рекомендации, которые соответствуют существующим предпочтениям пользователя. В зависимости от собранных данных и характера рассматриваемого элемента, существует множество способов совершенствования каждой из вышеприведенных стратегий.

47 В отличие от закрытой рекомендательной системы, которая предлагает пользователям товары из ограниченного списка вариантов. Эти списки курируются владельцем платформы.

48 N. Helberger (2019) On the Democratic Role of News Recommenders. Digital Journalism 7(8). Routledge: 993–1012. DOI: 10.1080/21670811.2019.1623700.

используемые пользователем для формирования собственного информированного мнения, и тем самым повлиять на мыслительный процесс человека. Исследователи пока не дали окончательного заключения, однако такой процесс потенциально ограничивает способность человека формировать свое мнение и делает его уязвимым для манипулятивного вмешательства. Поскольку системы основаны на методах интрузивного сбора данных и архитектуре массового убеждения, неизбежная персонализация контента может оказать значительное влияние на когнитивную автономию человека и нарушить его право на формирование собственного мнения.

2.2 Курирование и приоритизация контента, представляющего общественный интерес

Методы, с помощью которых онлайн-платформы курируют контент при помощи рекомендательных систем, недостаточно прозрачны и чрезвычайно редко подвергаются общественному и/или государственному контролю. Если интернет-посредники и предусматривают критерий разнообразия в рекомендательных системах, то это обычно делается разработчиками для привлечения пользователей и увеличения прибыли. Курируемое платформами разнообразие в основном используется не для поддержки демократических дебатов, а для оптимизации финансовой прибыли путем обеспечения долгосрочного вовлечения пользователей для достижения так называемой «цели оптимизации» – то есть увеличения доходов от рекламы или повышения стоимости платформы/услуги за счет увеличения трафика.[49] Проще говоря, ориентированные на бизнес критерии курирования контента в основном были разработаны для оптимизации экономической выгоды и вовлечения пользователей в корпоративных интересах, а не для отражения и обеспечения подлинно разнообразного контента.[50] Рекомендательные системы также могут иметь непредвиденные последствия для масштабных общественных целей и способны негативно влиять на абсолютное право на свободу мысли и мнения.

49 По словам представителя Facebook: «Facebook приносит прибыль только потому, что, когда вы складываете множество крошечных взаимодействий, которые ничего не стоят сами по себе, их стоимость внезапно достигает миллиардов долларов», K. Klonick, The New Governors, p. 1627.
50 K. Klonick, The New Governors, p. 1664.

Кроме того, процессы рекомендательных систем интернет-посредников, как правило, исключают возможность выбора и контроля и автономность отдельных пользователей, которые являются необходимыми условиями для обеспечения автономии личности в процессе поиска и передачи разнообразной информации и идей. После обнародования информации о ряде уязвимых аспектов рекомендательных систем[51] общество и государства ужесточили требования к более эффективной и осмысленной приоритизации «диверсифицированного» воздействия медиа в рекомендательных процессах. В частности, были затронуты вопросы законности и охвата политических высказываний, а также распространения и нормализации определенных систем ценностей как риторики, защищенной законом о свободе слова, даже если такой контент нарушает условия предоставления услуг или международные стандарты в области прав человека. Учитывая полное отсутствие информации о том, каким образом платформы регулируют и приоритизируют риторику в сети, очевидно, что рекомендательные системы и коррелирующая логика оптимизации могут подорвать принцип доступной всем «справедливой возможности участия»[52] В то же время следует признать, что основанные на правах человека рекомендательные системы могут и способствовать плюрализму, например, в условиях авторитаризма и контроля над медиа.

Важно определить, что представляет собой общественный интерес применительно к алгоритмическому курированию контента, и как он влияет на приоритетность различных типов контента. Концепция «контента, представляющего общественный интерес», и ее определение вызывают столько же споров, сколько и определение «диверсифицированного воздействия». В принципе, контент, представляющий общественный интерес – это информация, «в получении которой заинтересована общественность».[53] Другой подход к контенту, представляющему общественный интерес, заключается в том, чтобы рассматривать его как контент, имеющий отношение к благополучию граждан, жизни сообщества или местного населения. Очевидные примеры включают информацию, связанную с пандемией КОВИД-19 или с процессами демократических выборов. Объем зачастую недостоверного контента по

51 Наиболее известным и цитируемым был случай с «напалмовой девочкой» – сделаной репортером фотографией, которая была удалена Facebook на основании политики в отношении наготы.

52 K. Klonick, The New Governors, p. 1664.

53 M.E. Mazzoli and D. Tambini, Prioritisation uncovered: The Discoverability of Public Interest Content Online. Совет Европы (2020), стр. 13.

этим вопросам побудил посредников публично заявить о приоритетном значении достоверности информации. За сравнительно короткий срок несколько платформ продемонстрировали свою способность перенастроить алгоритмы рекомендательных систем, с тем чтобы отфильтровывать или помечать ложную информацию и отдавать предпочтение контенту, предоставляемому заслуживающими доверие органами здравоохранения. Однако причины эффективности этих модификаций или лежащие в их основе логические мотивы остаются предметом жарких споров,[54] а недостаточная прозрачность в отношении базовых данных и осуществляемого платформами выбора при модерации контента по-прежнему вызывают вопросы. Столкнувшись с растущими призывами со стороны общественности и государств возложить на них ответственность за информирование пользователей по вопросам, касающимся здоровья, интернет-посредники и платформы социальных сетей, невзирая на существующие разногласия, продемонстрировали свою способность к рефлексии и изменению принципов приоритизации и ранжирования контента.[55]

Это указывает на необходимость обеспечить повышенное общественное внимание и оказывать политическое давление на платформы, с тем чтобы они обеспечили прозрачность своих рекомендательных процессов и перестроили рекомендательные системы и цели оптимизации для решения структурных проблем современной медиасреды. Эта проблема выходит за рамки управления контентом и затрагивает антимонопольное законодательство, вопросы собственности на средства массовой информации и правила концентрации медиа.[56] Это также подчеркивает настоятельную необходимость приоритизации целей политики в области плюрализма и разнообразия медиа и вмешательств в целях создания более благоприятного цифрового пространства.

54 M. Cinelli, The COVID-19 social media infodemic, The Nature (2020), стр.10; см. также: Global Disinformation Index, Why is tech not defunding COVID-19 disinfo sites? (2020), <https://disinformationindex.org/2020/05/why-is-tech-not-defunding-covid-19-disinfo-sites>.

55 Европейская комиссия, Совместное сообщение для Европейского парламента, Европейского совета, Совета, Европейского экономического и социального комитета и Комитета по делам регионов, «Борьба с дезинформацией по поводу COVID-19 – исправление фактов» (Tackling COVID-19 disinformation - Getting the facts right), JOIN(2020) 8 final, 10 июня 2020 г., раздел 5.

56 M. E. Mazzoli and D. Tambini, Prioritisation uncovered: The Discoverability of Public Interest Content Online. Council of Europe (2020), p. 23.

2.3 Агрегация новостей и плюрализм медиа

Агрегаторы новостей функционируют как центральный узел распространения информации в интернете, направляя читателей к новостным статьям и иному контенту, который агрегатор считает новостным. Этот процесс осуществляется преимущественно алгоритмами, поэтому агрегаторов новостей иногда называют «алгоритмическими привратниками».[57] В работе агрегаторов новостей часто возникает противоречие между «алгоритмической» и «редакционной» логикой.[58] «Алгоритмическая логика» существенно влияет на разнообразие, а также на политический дискурс, например, отдавая предпочтение новизне по сравнению с другими критериями информативности (например, общественной значимостью, разнообразием и др.). Исследование процессов курирования контента в AppleNews, который использует как модерацию человеком (в Top Stories), так и алгоритмическое курирование контента (в Trending Stories), показало, что модерируемый человеком контент отличается «более разнообразным и более справедливым распределением источников, чем истории, отобранные алгоритмом».[59] Согласно тому же исследованию, Trending Stories («популярные новости») включали почти исключительно «мягкие новости» (например, истории о знаменитостях), а Top Stories («главные новости») предназначались для «жестких новостей» (например, политического контента).[60] Было установлено, что такая практика серьезно влияет на плюрализм источников и распространение новостей, а значит, и на плюрализм контента. Это влияние осуществляется двумя способами: во-первых, агрегаторы создают так называемый «эффект расширения рынка», поскольку предоставляют людям возможность ознакомиться с новостными изданиями, чей бренд менее известен или популярен. Во-вторых, внедрение агрегаторов побудило некоторых пользователей ограничить или прекратить прямое использование новостных изданий, что привело к так называемому «эффекту замещения». Поскольку реакция пользователей во многом определяется первым впечатлением, для привлечения внимания и вовлечения пользователей в новостной ленте обычно используются кликбейты, что способствует

57 Napoli 2014.
58 T. Gillespie, PJ Boczkowski, KA Foot, Media technologies: Essays on communication, materiality, and society, MIT Press (2014).
59 J. Bandy and N. Diakopoulos, Auditing News Curation Systems: A Case Study Examining Algorithmic and Editorial Logic in Apple News, Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020), p. 43.
60 Там же.

размещению приносящей прибыль рекламы. Такой подход ставит под сомнение стабильность средств массовой информации, а, следовательно, их независимость и плюрализм, усугубляя общее давление и финансовые трудности, с которыми традиционно сталкиваются медиа вследствие использования интернет-посредниками моделей концентрированной рекламы и эксплуатации данных.

Наблюдается растущий дисбаланс между информационным и коммуникационным воздействием традиционных медиа и онлайн-платформ, а также между созданием контента и его курированием. Учитывая снижающуюся популярность традиционной бизнес-модели, основанной на подписке, традиционные медиа ведут борьбу за выживание. Все больше людей получают новости исключительно из других источников, статьи в которых чаще всего доступны «бесплатно». Готовность пользователей платить за качественные новости снизилась, в то время как использование «бесплатных» сайтов-агрегаторов новостей и платформ социальных сетей возросло. В результате у многих новостных онлайн-изданий не осталось другого выбора, кроме как искать новые источники дохода. Они вынуждены перенимать множество приемов, используемых крупными платформами, которые оказывают значительное влияние на логику развития индустрии цифровых рекламных технологий (например, использование целевой рекламы, публикация спонсируемого контента или сбор и продажа данных пользователей). Эта тенденция имеет глубоко негативные последствия для свободы медиа во всем мире, порождая ситуацию, в которой традиционные медиа-организации вынуждены конкурировать с социальными медиакомпаниями и посредниками за одни и те же источники дохода, в то же время подстраиваясь под рекомендательные системы посредников и их политику курирования контента. Такая ситуация способствует подрыву доверия к медиа, снижает ответственность за создание и распространение дезинформации и другого проблемного контента.

Некоторые традиционные медиа-организации и сами используют алгоритмические инструменты, при этом персонализация и оптимизация контента играют неотъемлемую роль в процессах медиапроизводства. Между «логикой персонализации новостей» и «логикой персонализации

платформ» сохраняются серьезные различия"[61]: новостные медиа сталкиваются с такими проблемами, как системный недостаток технологических и финансовых ресурсов, обесценивание традиционной редакционной и профессиональной этики, превалирование новых частных интересов и экономических стимулов. Алгоритмические модели курирования контента и рекомендательные стратегии, разработанные и внедренные независимыми традиционными медиа и особенно общественными вещателями, могли бы обеспечивать альтернативные и более оптимальные методы ознакомления аудитории с разнообразными материалами и даже предлагать отдельным пользователям модели с заложенным в них «разнообразием».[62]

Факты показывают, что журналисты ценят «редакционную логику», включая «прозрачность, разнообразие, редакционную автономию, широкое предоставление информации, актуальность для конкретного человека, удобство использования и эффект неожиданности», а не алгоритмическую, ориентированную на бизнес логику рекомендательных систем.[63]

Процессы и практика курирования и рекомендации контента, основанные на алгоритмах, создают угрозы плюрализму и разнообразию медиа и вызывают обеспокоенность в плане осуществления права на свободу выражения мнения. Ниже перечислены основные факторы, способствующие возникновению таких угроз:

- Финансовая нестабильность и фискальное давление на традиционные медиа: онлайн-платформы получили огромную экономическую власть, в первую очередь за счет доходов от рекламы, и используют этот рычаг, чтобы диктовать условия для курирования всего онлайн-

61 B. Bodó, Selling News to Audiences – A Qualitative Inquiry into the Emerging Logics of Algorithmic News Personalization in European Quality News Media, Digital Journalism (2019), p.17-18.

62 Подробнее об этом можно узнать в докладе N. Helberger, Diversity by design — Diversity of content in the digital age, The Government of Canada (2020), p. 8; Natali Helberger, Kari Karppinen & Lucia D'Acunto (2018) Exposure diversity as a design principle for recommender systems, Information, Communication & Society, 21:2, 191-207, DOI: 10.1080/1369118X.2016.1271900.

63 Т В этом исследовании приняли участие редакции новостей из Нидерландов и Швейцарии; M. Bastian, N.Helberger & M. Makhortykh, Safeguarding the Journalistic DNA: Attitudes towards the Role of Professional Values in Algorithmic News Recommender Designs, Digital Journalism (2021), p.21.

контента, включая новостной и редакционный медиаконтент. Этот дисбаланс сил включает в себя дисбаланс «силы мнения»[64] и «воздействия на процессы формирования индивидуального и общественного мнения», что, в свою очередь, позволяет «этим платформам изменять саму структуру и баланс рынка медиа и тем самым оказывать прямое и постоянное воздействие на плюралистическую общественную сферу».[65]

- Традиционные медиа, вынужденные перенимать аналогичные бизнес-модели и «логику» социальных сетей, упускают возможность участвовать в изменении «правил новых коммуникационных режимов»,[66] и в создании более диверсифицированного медиаландшафта. Тем не менее, существуют альтернативные модели алгоритмической обработки, в центре которых находится контент, представляющий общественный интерес, и профессиональная журналистская практика. Как правило, такие модели внедряются традиционными и общественными медиа-организациями и служат для смягчения потенциальных проблем, вызванных отсутствием приоритизации контента, представляющего общественный интерес.[67]

- Совершенствование алгоритмического курирования контента с целью увеличения разнообразия медиаисточников сопряжено с рядом проблем:

    - Даже если интернет-посредники и компании социальных сетей «обучают и программируют» алгоритмы «во имя благих целей» – с тем чтобы предоставлять неоднородной аудитории неоднородный контент – такой практике недостает реальной прозрачности, и люди не имеют возможности влиять на дизайн и логику, которые управляют этими системами. Это представляет собой значительный системный риск для реализации свободы выражения мнения.

64 N. Helberger, The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power. Digital Journalism, 8(6), 842-854 (2020).
65 Там же, стр. 846.
66 Для углубленного обсуждения этой проблемы см. N. Helberger, The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power. Digital Journalism, 8(6), 842-854 (2020)
67 M.E, Mazzoli and D.Tambini, Prioritisation uncovered: The Discoverability of Public Interest Content Online. Council of Europe (2020).

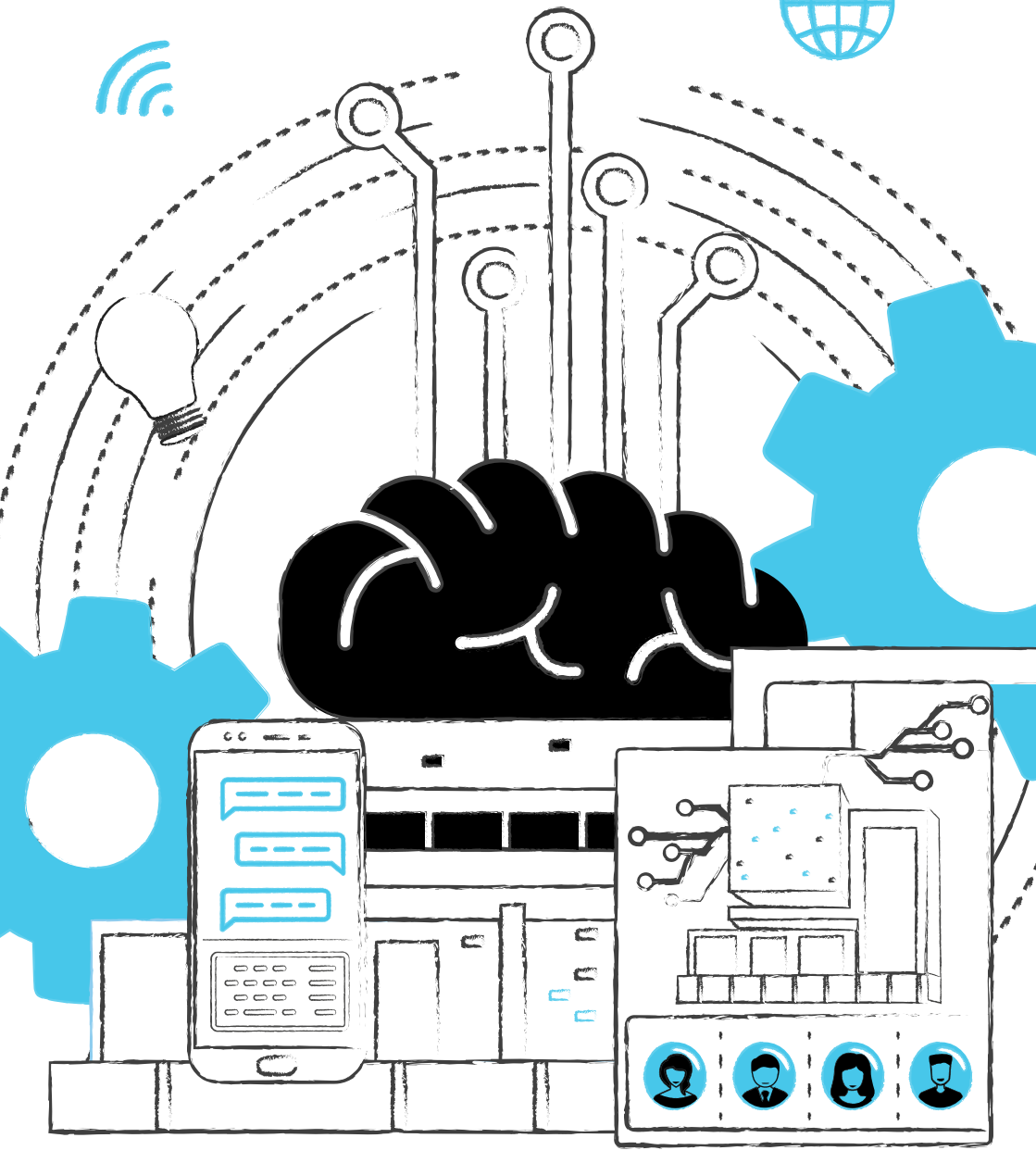
- Хотя персонализированный контент и процессы оптимизации могут способствовать удовлетворению различных индивидуальных, групповых и общественных потребностей и создавать потенциал для разнообразия, такой потенциал определяется целями государственной политики и соответствующими мерами.

- Практически нет информации о том, как циркулирует в интернете контент, создаваемый маргинальными сообществами и для этих сообществ, и как рекомендательные системы относятся к такому контенту. Исследования показывают[68] что определенный контент и риторика воспринимаются этими системами несколько иначе, что вызывает опасения по поводу неравной доступности контента и отсутствия механизмов предотвращения дискриминационных результатов работы алгоритмов, а также создает необходимость обеспечения справедливого и равного участия общественности и общественных дискуссий.

- Последствия алгоритмического курирования контента могут привести к еще большим нарушениям прав человека и принципа верховенства права, если их усиливает определенный национальный контекст, в частности сочетание с системным захватом медиа со стороны государства или частных компаний и с монополизацией контроля над общественным диалогом. В таких обстоятельствах дополнительные алгоритмические ограничения плюрализма и разнообразия медиа приводят к еще большему суммарному ограничению права человека на свободу выражения мнения.

68 См., например: A. Chinmayi, Facebook's Faces, Forthcoming Harvard Law Review Forum Volume 135 (2021) and K. Klonick, The New Governors: The People, Rules, and Processes Governing Online Speech, The Harvard Law Review (2018); C. O'Neil, Facebook's VIP "Whitelist" Reveals Two Big Problems, Bloomberg Opinion (2021), Retrieved from: <https://www.bloomberg.com/opinion/articles/2021-09-15/facebook-s-xcheck-vip-whitelist-reveals-two-big-problems>.

Плюрализм медиа



3. Рекомендации по использованию ИИ в курировании контента с учетом прав человека

В рамках международной системы защиты прав человека государства являются основными гарантами плюрализма медиа. Они в конечном итоге гарантируют реализацию прав человека и несут ответственность за создание благоприятных условий для реализации прав на свободу выражения мнения и свободу медиа. Приведенные ниже рекомендации в адрес государств-участников ОБСЕ, выработанные в ходе семинара, касаются таких вопросов, как укрепление плюралистического медиаландшафта и плюрализма мнений (3.1); содействие созданию благоприятной среды для разнообразия медиаконтента и индивидуального ознакомления с различными медиа (3.2); и создание условий для индивидуальной свободы выбора и возможности контролировать контент (3.3)

3.1 Рекомендации по укреплению плюралистического медиаландшафта и плюрализма мнений

Цель этой части доклада заключается в том, чтобы предложить государствам-участникам нормативную повестку, способствующую плюрализму в медиасреде и «сосуществованию разнообразных и конкурирующих интересов, что является основой для демократического равновесия»[69]. Сегодня реализацию этой повестки затрудняют такие факторы, как ограничение возможностей для создания демократического медиапространства, несбалансированное доминирование цифровых платформ и чрезмерная концентрация рынка. Государства-участники ОБСЕ должны создать условия для инновационного развития, обеспечения независимости и устойчивости средств массовой информации, особенно тех, которые ориентированы на общественные интересы, а также внедрить модели разработки, курирования и распространения контента, способствующие созданию таких условий.

- **С помощью регуляторных инициатив государства должны обеспечить равные условия для всех медиа, устраняя препятствия на пути к созданию справедливых и эффективных**

69 A. Roksa-Zubcevic et al, Media Regulatory Authorities and media pluralism, Regional Publication. Council of Europe (2021), pp.12-14..

рыночных условий. Рыночные условия должны обеспечить медиа возможность получать и использовать новые технологии и разрабатывать альтернативные бизнес-модели, в том числе модели алгоритмического курирования контента, которые будут способствовать диверсификации медиаландшафта и распространению контента, представляющего общественный интерес.

- **Государствам следует проанализировать, как вопрос о контенте, представляющем общественный интерес, отражен в уже существующей и будущей политике в области плюрализма медиа,** особенно с учетом роста значимости онлайн-платформ в распространении информации об общественном здравоохранении во время пандемии КОВИД-19.

- **Публично-частные партнерства между государством и соцмедиа компаниями и другими посредниками должны быть абсолютно прозрачными** и подлежать гражданскому надзору и общественному контролю. Это должно включать в себя нормативно-правовую базу для обеспечения плюрализма медиа.

- **Государства должны использовать политику и законодательство для преодоления существующих в настоящее время дисбаланса и монополизации рынка,** особенно в отношении интернет-посредников и контролируемого государством распространения контента. Любое вмешательство государства должно обеспечивать продемократический режим, который будет подлинно независимым и предложит структурные решения для укрепления плюрализма.

- **Государства должны способствовать плюрализму, технологическим и медийным инновациям** путем финансирования комплексных независимых исследований, которые помогут субъектам медиа, институтам общественного контроля и научному сообществу понять текущее распределение влияния – особенно в плане последствий работы рекомендательных систем, анализа логики рекомендаций и их конечного воздействия на плюрализм и разнообразие медиа.

3.2 Рекомендации по созданию благоприятных условий для разнообразия медиаконтента и индивидуального ознакомления с плюралистической информацией

В этой части доклада разнообразие рассматривается как нормативная концепция с целью ответить на следующие вопросы: должны ли интернет-посредники обеспечивать равный доступ к общественному пространству и возможность участия в нем, и если да, то каким образом

- **Нормативно-правовые и политические меры со стороны государства должны поддерживать и развивать интернет как пространство для участия в демократических процессах и демократического представительства.** Любое государственное регулирование цифрового пространства должно иметь четко определенные рамки, необходимые и соразмерные для достижения прозрачных целей, в полном соответствии с международной системой прав человека.

- **Государства должны принимать участие и поддерживать межотраслевой диалог для сбора наиболее актуальных и значимых данных о влиянии алгоритмического курирования контента,** таком как поляризация, пробелы в информации и др. Независимый надзор и прозрачный мониторинг над соблюдением принципа разнообразия требует **междисциплинарного подхода под руководством академических институтов или организаций гражданского общества при поддержке государства.** Межотраслевой контроль за соблюдением принципа разнообразия должен использоваться для выявления контента и аудиторий, которые рискуют быть лишенными возможности участия в общественной жизни и/или представительства или исторически уже сталкивались с такой угрозой.

- **Государства должны использовать инклюзивный подход и обеспечивать участие различных заинтересованных сторон в алгоритмическом курировании контента.** Демократические режимы не являются самодостаточными – для процветания демократии граждане должны иметь возможность принимать

информированные решения. Обеспечивая открытый диалог и межотраслевое взаимодействие с интернет-посредниками, государства совместно с организациями гражданского общества, маргинализированными сообществами, организациями медиа, журналистами и их представителями могут способствовать устойчивому межотраслевому сотрудничеству, в том числе между государственными и негосударственными субъектами. Более того, государства должны добиваться разнообразия в составе команд разработчиков алгоритмических систем курирования контента, с тем чтобы при разработке и внедрении алгоритмов были представлены различные интересы и перспективы.

- **Государства должны оказывать поддержку и предоставлять ресурсы существующим независимым органам регулирования медиа,** которые привлекают всех национальных участников и экспертов медиа к созданию и обеспечению экономической, правовой и политической среды, в которой разнообразие культивируется как основная демократическая цель.

- **Государствам следует разработать основанную на фактах и исследованиях законодательную базу для обеспечения подотчетности интернет-посредников, в том числе путем отражения в законе требования о надлежащей проверке соблюдения прав человека.** Для обеспечения общественного контроля, частью любой стратегии снижения рисков или любого внешнего аудита должна быть оценка воздействия на права человека.

- **Государства должны укреплять независимые органы регулирования медиа и другие компетентные учреждения и привлекать их к обеспечению общественного надзора и проведению исследований.** В частности, эти органы должны участвовать в оценке воздействия на права человека с целью устранения порождаемых интернет-посредниками рисков в отношении разнообразия и плюрализма, включая риски, связанные с маргинализированными сообществами. Наряду с проведением таких оценок следует обеспечить наличие механизмов

подотчетности, а также обнародование и публикацию результатов оценок, аудитов и т.п.

- **Государства должны увеличить государственное финансирование независимой, качественной журналистики и/или предоставлять финансовые ресурсы независимым заинтересованным сторонам,** обладающим соответствующим опытом и надежной репутацией в области прав человека. Эти независимые организации способны предложить альтернативы существующим бизнес-моделям, которые ориентированы на получение дохода и основаны на анализе данных, тем самым способствуя развитию децентрализованных технологических систем алгоритмического курирования контента, продвигающих такие общественные ценности, как разнообразие медиа, инклюзивность и толерантность.

- **Государства должны гарантировать, что любое потенциальное вмешательство в эту область не будет ограничивать позитивные функции персонализации или независимость медиа,** но будет предусматривать поддержку и меры для привлечения внимания разработчиков контента к обеспечению его разнообразия и отражению в нем общественных интересов. Персонализация может быть полезна человеку, если используется для детализации поиска и ускоряет получение информации.

- **Государства должны создавать и гарантировать адекватные механизмы доступа к данным, позволяющие опытным исследователям, организациям гражданского общества и другим независимым заинтересованным сторонам, таким как медиа, получать доступ к данным, хранящимся у интернет-посредников.** В то же время необходимо предотвратить злоупотребления такими механизмами путем использования принципов этики или создания независимого органа, выполняющего функцию надзора.

3.3 Рекомендации по поводу создания условий для индивидуальной свободы выбора и возможности контролировать контент

Алгоритмический дизайн должен предусматривать расширение прав и возможностей человека, а дизайн, ориентированный на общественные интересы и права человека, должен быть первоочередной задачей при внедрении алгоритмов.

- **Государства должны поддерживать инициативы саморегулирования и совместного регулирования контента и создавать условия, способствующие индивидуальному контролю со стороны пользователя над контентом, который он или она видят в интернете.** С этой целью можно предусмотреть в законодательстве такие требования, как необходимое согласие на использование по умолчанию систем рекомендации контента, а также простая идентификация и возможность выбора редакционной и нередакционной персонализации.

- **Государства обязаны обеспечить прозрачность и обоснованность персонализации предварительно отобранных новостей и обработки данных.** Это включает прозрачность критериев, принципов и различных механизмов, на основании которых принимаются решения о приоритетности контента (с тем чтобы укрепить общественное доверие и предоставить общественности возможность понять, учитываются ли при этом коммерческие или общественные интересы). Государства также должны требовать от посредников введения надлежащих процессуальных мер. Например, вводя ограничения на новостные ленты, посредники должны информировать пользователей о своей политике и предоставлять эффективные механизмы возмещения ущерба. Аналогичным образом, **государства должны поддерживать инициативы по саморегулированию и совместному регулированию, которые обеспечивают прозрачность процессов, используемых посредниками** для принятия решений о приоритетности контента.

- **Государства обязаны создавать устойчивые программы медиа- и цифровой грамотности для всех групп населения.** Люди часто не знают и/или не осознают последствий алгоритмического курирования контента в плане осуществления ими своих прав и основных свобод.

- **Государства обязаны уделять особое внимание защите права на доступ, поиск и распространение любых мнений и идей среди всех возрастных групп. В частности, государства должны расширять возможности для формирования индивидуального мнения,** в том числе для молодых людей, лишенных надлежащего доступа к традиционному медиаконтенту.

4. Заключение

Курирование контента, по крайней мере в контексте плюрализма и разнообразия медиа, в значительной мере отошло на задний план более широких дискуссий по поводу управления контентом. Это упущение, усугубленное непониманием важности медийного и информационного разнообразия для неоднородной аудитории, имеет негативные последствия, которые не менее опасны, чем риски, связанные с незаконным контентом и дезинформацией. Все управляемые алгоритмами процессы курирования и модерации контента тесно взаимосвязаны и рассматривать их следует комплексно.[70] Эти процессы чрезвычайно важны, поскольку именно алгоритмы определяют, какую именно информацию видят пользователи, какой контент приоритизируется, а какой исключается. Онлайн-привратники все чаще полагаются на рекомендательные системы, которые систематически анализируют модели поведения человека и составляют профили конкретных пользователей, помогающие определить, какая информация с большей вероятностью их заинтересует. Иными словами, привратники собирают данные, чтобы определить, какой персонализированный контент следует предлагать отдельным пользователям, чтобы стимулировать их вовлеченность и генерировать еще больше данных о них – даже если принимаемые решения противоречат принципам демократического дискурса, информационного разнообразия, плюрализма медиа, а также

70 Иными словами, чрезмерное удаление «законного» контента на самом деле также представляет угрозу для разнообразия медиа.

праву человека на неприкосновенность частной жизни. По этой причине, как четко сформулировано в рекомендациях, междисциплинарные исследования и прозрачность политики и практики курирования контента интернет-посредниками являются важнейшими предпосылками для обеспечения плюрализма и разнообразия медиа при разработке и внедрении алгоритмов.

В настоящем итоговом докладе подчеркиваются проблемные аспекты персонализированных систем рекомендации контента, используемых интернет-посредниками и, в частности, социальными медиаплатформами. В нем описывается негативное влияние этих систем на сплоченность общества, разнообразие, качество информации в общественном дискурсе, а также конфиденциальность. На сетевой опыт индивидуальных пользователей стратегически влияют решения, которые принимаются в корыстных целях и реализуются с помощью алгоритмов, без уведомления пользователей и без контроля со стороны государственных органов, и которые основаны на интрузивном сборе и анализе данных в обход законов о конфиденциальности и защите данных, что в значительной степени негативно сказывается на разнообразии информации, плюрализме медиа и праве на неприкосновенность частной жизни.

Существует множество доказательств того, что сила убеждения онлайн-платформ способна направлять и усиливать определенные общественные нарративы и типы дискурса по сравнению с другими. Для стран с нестабильными или деспотичными политическими системами эта сила убеждения, усиленная алгоритмами, может иметь катастрофические последствия в плане реализации индивидуальных прав человека. Ранжируя и дифференцируя контент и рекомендации, интернет-посредники изменяют конфигурацию общественных дебатов, расширяя возможности тех, кто уже находится в привилегированном положении. Платой за это является сокращение разнообразия в целом и, в частности, неблагоприятное воздействие на исторически маргинализированные группы, которые продолжают оставаться на задворках общественного дискурса в ходе этого процесса, который тиражирует и усиливает неравенство и несправедливость. В то время как алгоритмическое курирование контента способно внести раскол в общество и ограничить активность пользователей и распространение информации,

разнообразие медиа способствует социальной сплоченности, толерантности и распределению коммуникационных возможностей. Именно государства, в первую очередь, а также негосударственные субъекты, в частности, посредники и медиа-организации, обязаны гарантировать, что плюрализм медиа, равноправный доступ к информации и полноценное осуществление прав человека будут лежать в основе правил, влияющих на информационное онлайн-пространство. А именно эти правила являются кирпичиками в строительстве подлинно демократического цифрового общества.

Курирование контента и бизнес-модели на основе слежки

Эта глава посвящена использованию ИИ в курировании контента с целью продемонстрировать связь между капитализмом слежки и целевой рекламой, а также их влияние на свободу выражения мнения. В ней подчеркиваются недостатки курирования контента на основе ИИ и целевой рекламы и приводятся ориентированные на права человека рекомендации в адрес государств-участников ОБСЕ по устранению негативного воздействия инструментов ИИ, используемых в процессе курирования контента, на право на свободу выражения мнения.

1. Определение масштабов воздействия бизнес-моделей на основе слежки при их использовании для курирования контента

1.1 Влияние автоматизированного принятия решений на право на свободу мнения

Международная система прав человека различает внутреннее и внешнее измерение права на свободу мнения. В то время как внешнее измерение этого права может подлежать законным, соразмерным и недискриминационным ограничениям, необходимым в демократическом обществе, внутреннее измерение свободы мнения, так называемый «forum internum», является абсолютным и не допускает отступлений.[71] Статья 19 Всеобщей декларации прав человека, а также Международный пакт о гражданских и политических правах обеспечивают защиту этого абсолютного права от любых ограничений или вмешательства. По словам Специального докладчика ООН по вопросам свободы выражения мнения, «недобровольное раскрытие мнения запрещается, а автономность мышления должна утверждаться».[72]

71 Управление и Верховный комиссар по правам человека, Замечание общего порядка № 22: Статья 188 МПГПП (Свобода мысли, совести и религии), см. на сайте <https://www.refworld.org/docid/453883fb22.html>, 1993.

72 Доклад Специального докладчика по вопросу о поощрении и защите права на свободу мнений и их свободное выражение Айрин Хан, «Дезинформация и свобода мнений и их свободное выражение», см. на сайте: <https://undocs.org/A/HRC/47/25>, 2021.

Основанные на сборе данных бизнес-модели крупных онлайн-платформ позволяют рекламной индустрии разрабатывать и применять стратегии таргетинга, опирающиеся на анализ данных. Благодаря такому подходу компании выявляют и используют поведенческие модели и характеристики отдельного человека или сообщества. Зонтичным термином, охватывающим все эти манипулятивные методы, является «реклама на основе слежки», под которым подразумевается цифровая реклама, направленная на отдельных лиц или группы лиц, обычно посредством отслеживания и профилирования на основе персональных данных. Контекст размещения конкретной рекламы может быть случайным, поскольку такая реклама нацелена на отдельного человека и может следовать за ним в различных контекстах.[73] В большинстве случаев реклама на основе слежки является частью автоматизированного процесса, в ходе которого каждое отдельное рекламное объявление выбирается и размещается за считанные миллисекунды. Это означает, что ни «издатель рекламы» (например, владелец сайта или приложения), ни рекламодатель (например, владелец продвигаемого бренда) не выбирают, кому и где будет показана их реклама – это решение автоматически принимают технологические системы, которые часто контролируются сторонними посредниками (так называемыми рекламно-технологическими или AdTech-компаниями).[74]

Реклама на основе слежки в значительной степени способствовала использованию характеристик человека для повышения убедительности сообщения, тем самым необоснованно посягнув на абсолютную свободу человека формировать собственное мнение и его возможность независимо мыслить. Пользователями услуг медиа-платформ манипулируют, заставляя их думать или принимать решения, которые в противном случае они могли бы и не принять. Реклама на основе слежки эксплуатирует уязвимые аспекты человека, даже если не выявляет их напрямую. С помощью аудиторий «двойников» (т.н. lookalike-аудиторий) рекламодатели способны дублировать группы людей по определенным параметрам, чтобы охватить новых пользователей, обладающих аналогичными характеристиками.

73 Norwegian Consumer Council, Time to ban surveillance-based advertising: The case against commercial surveillance online, см. на сайте: <https://www.forbrukerradet.no/wp-content/uploads/2021/06/20210622-final-report-time-to-ban-surveillance-based-advertising.pdf>, 2021.

74 Norwegian Consumer Council, Out of control: How consumers are exploited by the online advertising industry, см. на сайте: <https://fil.forbrukerradet.no/wp-content/uploads/2020/01/2020-01-14-out-of-control-final-version.pdf>, 2020.

Автоматизированные инструменты и доминирующее положение нескольких интернет-платформ расширяют их возможности манипулирования, поскольку каждый отдельный пользователь их услуг постоянно и в любое время может стать мишенью для рекламы.

Возникает необходимость запрета методов, которые оказывают негативное влияние на абсолютное право человека на свободу мнения и свободу мысли, в частности, в результате масштабного манипулирования мыслями и мнением людей без их ведома или согласия. Это явление, в частности, включает в себя целенаправленное наблюдение за поведением пользователей и индивидуальное отслеживание их присутствия на разных сайтах и с различных устройств. Подобная непрерывная слежка со стороны компаний создает риск систематического манипулирования людьми, выходящего за рамки традиционных форм рекламного воздействия. Реклама на основе слежки выбирает пользователей в качестве мишеней, используя скрытые методы[75] и эксплуатируя их слабости, что открывает новые возможности для манипулирования. В частности, в сочетании с алгоритмами курирования контента, позволяющими получить максимальный доход, реклама на основе слежки способна влиять на слова и поступки отдельных людей, что негативно сказывается на разнообразии информации, взглядов и мнений.

Несмотря на заявления онлайн-платформ о том, что от рекламы на основе слежки уже никуда не деться, интернет строился не на бизнес-модели «жутковатой рекламы», а на прямо противоположных подходах. На самом деле, все совсем наоборот. Государства не должны напрямую или косвенно защищать бизнес-модели, которые базируются на рекламе на основе слежки и нарушают международное право в области прав человека. Искоренение неправомерных моделей также означает создание альтернативных возможностей, не нарушающих права человека, включая инновационные формы контекстной рекламы, основанные на минимальной персонализации и не задействующие индивидуальный таргетинг.[76] Это также обеспечит появление на цифровом рынке новых игроков.

75 Усилия гражданского общества по достижению большей прозрачности наталкиваются на препятствия: <https://algorithmwatch.org/en/defend-public-interest-research-on-platforms/>.

76 Natasha Lomas, Data from Dutch public broadcaster shows the value of ditching creepy ads, см. на сайте: <https://techcrunch.com/2020/07/24/data-from-dutch-public-broadcaster-shows-the-value-of-ditching-creepy-ads/?guccounter=1>, 2020.

Реклама на основе слежки имеет далеко идущие последствия в плане межперсонального общения, принимаемых человеком решений и его участия в демократических дебатах. Меры, направленные на повышение прозрачности, помогают лучше понять масштаб проблем, однако их недостаточно для предотвращения и пресечения нарушений прав человека. Индивидуальный и общественный вред, причиняемый навязчивым таргетингом и персонализацией, требует систематических ответных мер. Любое инвазивное отслеживание пользователей, от вторжения в частную жизнь до курирования контента, существенно ущемляет право на свободу выражения мнения. Государства обязаны защищать это абсолютное право от подобных вмешательств путем создания адекватной нормативно-правовой базы, устанавливающей и обеспечивающей надежные гарантии прав человека.

1.2 Руководство по онлайн-таргетингу

В то время как многие посредники выбирают пользователей (а также непользователей) в качестве своей мишени, профилируя и отслеживая их поведение на сайтах, лидирующие позиции в индустрии рекламных технологий занимают онлайн-привратники, имеющие беспрецедентный доступ к большим объемам пользовательских данных. На практике реклама на основе слежки начинается с «издателей рекламы» – публикующих рекламу компаний, которые управляют веб-сайтом или мобильным приложением, предоставляющим услуги или контент. «Издатели» предоставляют место для размещения рекламы на своих платформах и/или доступ к данным о своих пользователях. Их торговые партнеры – это «маркетологи», то есть компании, которые стремятся продать свою продукцию наиболее ценным клиентам. Однако между этими сторонами находятся другие продавцы и онлайн-биржи рекламы, которые остаются в тени, не вступают в прямые отношения с пользователями, принимают решения о том, какие объявления будут размещены на тех или иных сайтах, и получают за это процент от сделки. Эта сложная рекламная сеть собирает, анализирует и объединяет огромное количество персональных данных без ведома пользователей. Ни «издатели», ни «маркетологи» не в состоянии полностью или даже частично контролировать этот процесс.

Крупные интернет-посредники занимают доминирующее положение в рекламной экосистеме, выполняя одновременно все три роли – издателей

рекламы, маркетологов и сторонних поставщиков. Это привилегированное положение усиливается благодаря практически неограниченному доступу к данным, включая данные, поступающие от собственных сервисов посредника и от третьих лиц. Это создает огромный дисбаланс сил, который подпитывает недобросовестную конкуренцию на цифровом рынке и создает риск системных нарушений прав человека.

Основное внимание в этой части доклада уделяется интернет-посредникам, чьи бизнес-модели в значительной степени зависят от онлайн-таргетинга. Хотя многие интернет-посредники направляют рекламу пользователям, анализируя данные об их поведении и отслеживая, какие вебсайты они посещают, что само по себе является нарушением прав человека, именно интернет-посредники, имеющие беспрецедентный доступ к большим объемам данных пользователей, являются лидерами индустрии рекламных технологий. Например, крупные платформы социальных сетей, такие как Facebook, разработали детальные системы для своего рекламного интерфейса, благодаря которым они контролируют основные доходы от рекламы во всем мире. Организация по защите цифровых прав Panoptykon («Паноптикон») составила карту и подробно описала экосистему рекламных технологий, разработанную привратником Facebook, а также ее влияние на права человека. «Паноптикон» отмечает, что Facebook является не просто пассивным посредником между рекламодателем и пользователями.[77] а позволяет рекламодателям выбирать критерии, которые затем интерпретируются алгоритмом Facebook для достижения желаемых целей рекламодателя.

В данном докладе Facebook приводится в качестве практического примера того, как реклама на основе слежки работает на практике. Во-первых, рекламодатели могут выбирать свою целевую аудиторию на основе критериев таргетинга, которые определяет посредник. Существует ряд критериев, которые рекламодатели могут использовать с этой целью. Среди прочих,[78] они могут выбрать целевую аудиторию или так называемых «двойников» (lookalike-аудитория). Оба критерия были введены Facebook в последние годы и часто описываются следующим образом:

77 Panoptykon Foundation, Who (really) targets you? Facebook in Polish election campaigns, 2020.
78 Там же.

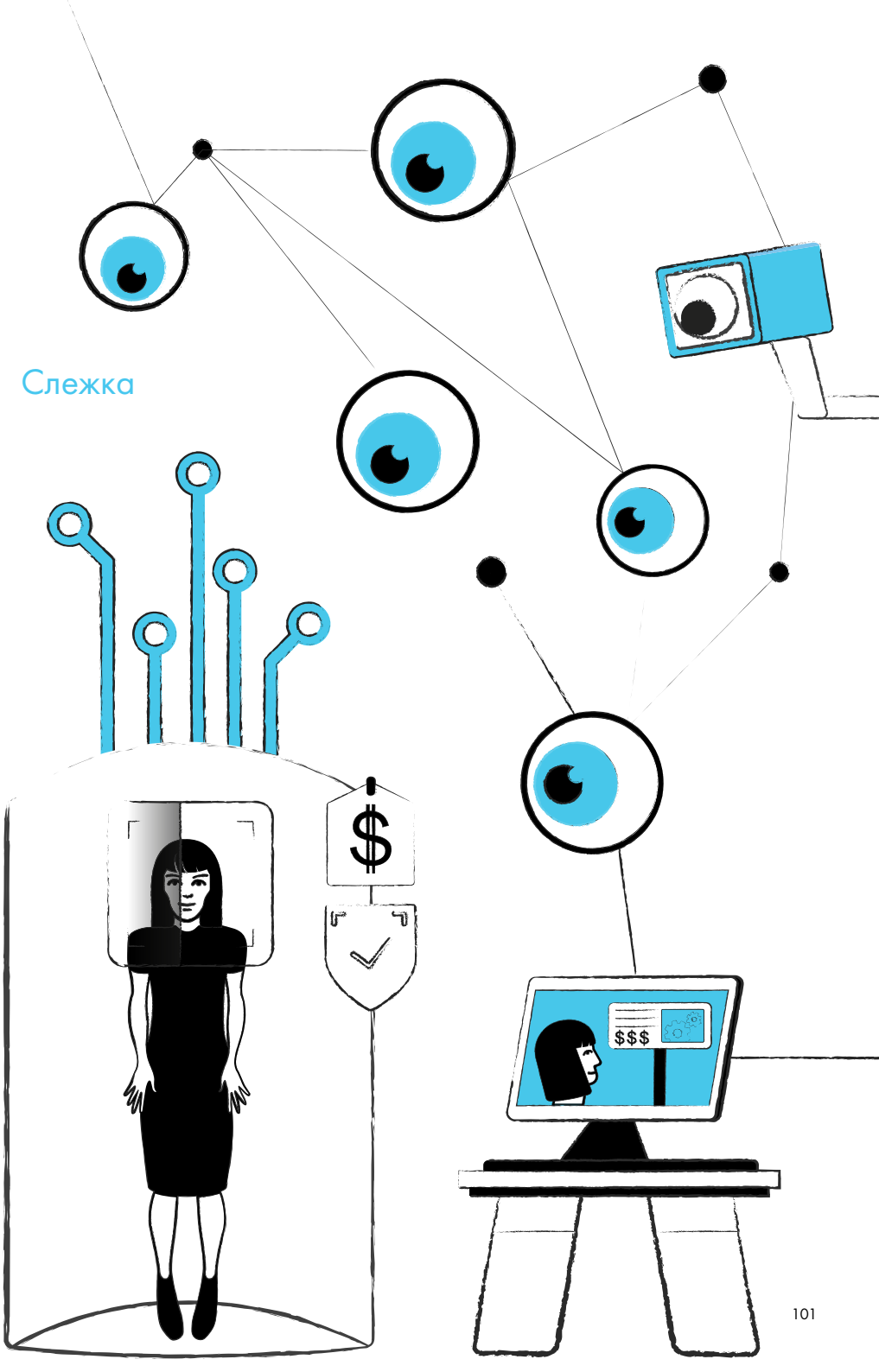
- **Критерий пользовательской аудитории** основан на собственной имеющейся у рекламодателя информации о пользователях, которую он может передать посреднику. Соответственно, алгоритм посредника сопоставляет эту информацию со своими собственными данными о пользователях, без раскрытия профилей пользователей рекламодателям.

- **Критерий lookalike-аудитории** позволяет рекламодателям выбирать в качестве мишени группу пользователей, похожих на искомую. На практике посредник прогнозирует, какая аудитория имеет общие характеристики с первоначальной целевой группой – так называемая «посевная аудитория».[79] Аудитории «двойников» определяются алгоритмом сопоставления посредника.

Онлайн-платформы способны с высокой точностью выбирать людей в качестве мишеней для рекламы, поскольку обладают данными и информацией как об отдельных пользователях, так и о людях, не являющихся пользователями их услуг. Анализ больших данных позволяет им предсказывать поведение людей, используя данные, предоставленные непосредственно пользователями или полученные путем наблюдения за онлайн-активностью и моделями поведения пользователей и других людей. Высокочувствительные алгоритмы создают профили на основе поведенческих данных – привычек, предпочтений, антипатий и взаимодействия с пользователями. Эти профили могут даже включать выводы, сделанные на основе часов наибольшей активности пользователя в интернете. Как создание, так и последующее использование профилей нарушает неприкосновенность частной жизни, предполагает допущения и может привести к дискриминации. Алгоритмы также способны извлекать дополнительную информацию о человеке, которую он не имел намерения раскрывать. В основе этого метода лежит идея о том, что чем больше компании знают о своих пользователях, тем больше вероятность того, что они смогут успешно прогнозировать их поведение и потенциально манипулировать ими. Затем эта информация используется для предоставления конкретного контента и рекламы «в нужное время и в нужном контексте», чтобы побудить пользователей покупать определенную продукцию или услуги или смотреть определенные виде.[80]

79 Panoptykon Foundation, Who (really) targets you? Facebook in Polish election campaigns, 2020. См. также: Norwegian Council for Consumer Protection, Out of control: How consumers are exploited by the online advertising industry, 2020.
80 Vladan Joler, The Human Fabric of the Facebook Pyramid, SHARE Lab Foundation, 2017.

Слежка



2. Рекомендации по регулированию рекламы на основе слежки с учетом прав человека

2.1 Рекомендации по расширению прав и возможностей пользователей и их личной свободы выбора в онлайн-экосистеме

- **Государствам следует применять подход, ориентированный на человека и пользователя, для расширения прав и возможностей пользователей, свободы выбора каждого человека и возможностей контроля над собственными данными.** Существует значительный риск, связанный с отсутствием у нас возможности узнать, проводилось ли в отношении нас профилирование или идентификация, и если да, то каким образом и с помощью какого алгоритма. Однако только прозрачности информации недостаточно для того, чтобы люди могли контролировать использование своих данных. Прозрачность должна сочетаться с надежным, осуществимым правом на отказ от применения в отношении себя подобных методов.

- **Государствам следует гарантировать, что укрепление прав и возможностей пользователей и свободы личного выбора, а также создание дополнительных систем внешнего надзора и расследования не являются взаимоисключающими факторами.** В нынешней социально-политической обстановке важно уделять первостепенное внимание свободе выбора и возможностям контроля со стороны пользователей.

- **Государства должны инвестировать в исследования с** целью разработки эмпирической базы для выявления и изучения влияния рекламы на основе слежки на независимость и свободный выбор пользователей. Без дальнейших эмпирических исследований может произойти чрезмерное упрощение опыта пользователей, основанного на скудных данных и исследованиях, направленных на конкретные онлайн-сообщества.

- **Государства должны способствовать созданию нормативно-правовой базы для улучшения качества информации, распространяемой среди пользователей,** чтобы предоставить им возможность свободного выбора рекламы, которую они просматривают и на которую реагируют. Нормативно-правовая база также должна гарантировать, что пользователи будут лучше информированы о том, какие данные о них собираются и как они используются (включая причины, по которым конкретный пользователь становится мишенью для конкретной рекламы).

- **Государствам следует четко сформулировать, в каких случаях существующее регулирование медиа и контента применимо к виртуальному контенту.** В случае выявления пробелов государствам следует пересмотреть и разработать политику и рекомендации по модерации интернет-контента в контексте т.н. «черного ящика» – недоступности онлайн-экосистем и платформ.

- **Государства должны продвигать практику ведения бизнеса, обеспечивающую альтернативы нынешней рекламе на основе слежки.** Текущая бизнес-практика интернет-посредников создает проблему концентрации власти, которая ограничивает персональную независимость и свободный выбор пользователей.

- **Государства должны обязать частных субъектов действовать в соответствии с Руководящими принципами предпринимательской деятельности в аспекте прав человека ООН,** с тем чтобы корпоративные ценности и структуры управления не были нацелены на максимизацию прибыли в ущерб правам человека и демократическим ценностям. Общественность все чаще требует от предприятий действовать не в коммерческом вакууме, а отражать демократические ценности и приоритеты.

- **Государства должны поощрять частный сектор к использованию неюридических методов для обеспечения большей прозрачности и подотчетности.** Частные инициативы по разработке кодексов этики играют важнейшую роль в обеспечении корпоративной социальной ответственности. Однако сами по

себе такие подходы к саморегулированию не могут обеспечить эффективную защиту от потенциального нарушения рекламой на основе слежки абсолютного права на свободу мнения.

Рекомендации по работе с населением и повышению осведомленности широкой общественности о рекламе на основе слежки

- **Государства должны способствовать повышению осведомленности и цифровой грамотности,** чтобы пользователи понимали, как управлять собственным медиапотреблением и использованием интернет-посредников. Пользователи должны иметь четкое представление о том, по какой причине им демонстрируют адресный контент, а также о принципах обработки своих персональных данных и предоставления доступа к ним. Пользователи должны иметь представление о том, какой объем их персональных данных подвергается обработке, к каким категориям информации предоставляется доступ, кто имеет право на такой доступ, и как конкретная информация связана с защищенными характеристиками. Повышение уровня цифровой грамотности важно для расширения прав и возможностей пользователей и усиления их устойчивости к постоянно адаптирующейся отрасли.

- **Государства должны понимать механизмы влияния основанной на слежке рекламы на права человека на равенство и недискриминацию в сочетании с правом на свободу мнения и его свободное выражение.** Реклама на основе слежки создает различный и потенциально дискриминационный опыт как внутри групп людей, обладающих определенными характеристиками, так и между этими группами. Это еще более осложняет ситуацию, поскольку некоторые группы не обладают достаточной цифровой грамотностью, что может усилить негативное влияние моделей на основе слежки.

- **Государства должны рекомендовать частным субъектам рассмотреть концепцию «социальной лицензии»,** которая направлена на то, чтобы частные и государственные поставщики услуг действовали ответственно и этично в наилучших интересах общества.

- **Государствам следует инвестировать в исследования** для создания прочной эмпирической базы, которая обеспечит реализацию инициатив по повышению осведомленности и информированности, направленных на решение практических вопросов, связанных с реакцией людей на манипуляции в интернете и стратегии целевой рекламы.

Рекомендации по отражению в законодательстве требований к прозрачности: различные уровни прозрачности

- **Государства должны обеспечить значительную прозрачность информации в рекламе на основе слежки.** Персонализация рекламы на основе слежки означает, что разные люди видят разные рекламные объявления, в зависимости от целого ряда факторов, включая время суток, контекст, демографические данные, персональные характеристики и поведенческие модели. При этом алгоритмические системы, в которые поступают данные пользователей, глубоко непрозрачны, и их часто сравнивают с «черным ящиком». Таким образом, пользователям (или регулирующим органам) практически непонятны решения, на которые опирается реклама на основе слежки, в результате чего пользователи не в состоянии понять, почему им показывают конкретную рекламу в конкретный момент времени, и каким образом в этом процесс используются их личные данные.

- **Назначенные надзорные органы, обладающие опытом в области обеспечения равенства и недискриминации, необходимо наделить полномочиями по мониторингу и устранению неравного или дискриминационного воздействия рекламы на основе слежки на маргинализированные группы. Государства должны рассмотреть различные подходы к ответственности за деструктивную рекламу на основе слежки,** а также проанализировать преимущества моделей саморегулирования и совместного регулирования, корпоративной подотчетности и управления, механизмов судебного разбирательства или альтернативных электронных судов для установления ответственности за негативные последствия такой рекламы.

- **Государства должны наделить органы по защите равенства полномочиями для проведения стратегических судебных процессов** по оспариванию дискриминационных результатов применения автоматизированных мер.

- **Государства в сотрудничестве с научными кругами, гражданским обществом и независимыми заинтересованными сторонами должны** переориентировать усилия по обеспечению прозрачности на получение большего доступа к крупномасштабным дезагрегированным данным, что позволит провести исследования и понять суть профилирования и рекламы на основе анализа данных. Прозрачность необходима для того, чтобы государства и общественность были в курсе того, как развертывается реклама на основе слежки. Это позволит проводить эффективные исследования и оспаривать деструктивные процессы. Доступ к данным для проведения эффективных исследований в интересах общества должен быть предусмотрен законодательством.

Рекомендации по решению вопроса о взаимосвязи между индивидуальной и групповой конфиденциальностью

- Международное право в области прав человека дает определение индивидуальных прав, в то время как у онлайн-профилирования есть коллективные аспекты и последствия. Цифровые профили основаны на заключениях и предположениях о сложной системе данных и сетей. Алгоритмическое профилирование позволяет соотнести разные характеристики и связи для составления профиля представителей маргинализированных групп. Поэтому **государства должны обеспечить соблюдение интернет-посредниками права на свободу мнения и осознание важного пересечения этого права с правом на свободу объединений и правом на свободу слова.**

- **Государства должны учитывать имеющиеся у существующих правовых механизмов ограничения в плане гарантий соблюдения коллективных прав в контексте рекламы на основе слежки.** Существующие правовые системы отражают индивидуальные права и распространяются на коллективные права

только в том случае, если человек принадлежит к определенной группе. Даже если человек делает осознанный выбор и отказывается от передачи своих персональных данных, он все равно может стать объектом профилирования как часть более широкой группы, выбранной системами искусственного интеллекта в качестве мишени или отнесенной ими к целевой категории. Государства должны обеспечить защиту и регулирование процесса использования персональных данных, включая метаданные или демографически идентифицируемые данные, которые считаются чрезвычайно актуальными и ценными, когда речь идет о методике рекламы.

2.2 Рекомендации по разработке национальных и международных нормативных инициатив по эффективному устранению негативного воздействия рекламы на основе слежки на права человека

Рекомендации по обеспечению абсолютного права на свободу мнения

- **В соответствии с международными стандартами в области прав человека, государства должны уважать и продвигать абсолютное право на свободу мнения,** что включает в себя право на конфиденциальность собственных мыслей и мнения, право на защиту от манипулирования ими и право не подвергаться наказанию за их выражение.

- **Государства должны акцентировать права каждого человека беспрепятственно придерживаться своего мнения; искать, получать и распространять информацию и идеи с помощью любых средств массовой информации, независимо от физических границ; и не подвергаться незаконному или произвольному вмешательству в личную жизнь.** Государства обязаны гарантировать людям возможность формировать собственное мнение и предоставить защиту от манипулирования этим мнением путем тайного профилирования, которое помогает выявить, когда пользователь наиболее восприимчив к попыткам повлиять

на его поведение, и использовать его слабости. Неправомерное влияние может стать результатом применения таких методов, как непрозрачная или не поддающаяся проверке массовая целевая реклама; наблюдение и слежка за поведением или сбивающие с толку особенности дизайна («темные паттерны»); либо использование дисбаланса сил для влияния на мышление (скорость, масштаб, недосягаемость, эффект «черного ящика» и систематическое скрытое влияние). Такие методы слежения и таргетинга могут привести к возникновению самоцензуры и приспособленчества и их дальнейшему распространению среди определенных слоев населения. Государства должны четко сформулировать политику и установить границы между законным влиянием и незаконным манипулированием с применением алгоритмических технологий, и рассмотреть возможность введения моратория или запрета в отношении последнего.

- **Государства должны предоставить пользователям возможность опротестовывать в юридических инстанциях** меры, примененные в отношении них интернет-посредниками в результате несовершенства системы или же ошибочной классификации пользователей, которая влияет на опыт человека в интернете независимо от точности такой классификации.

- **Государства должны законодательно обеспечить анонимность и шифрование данных,** в том числе гарантировать, что мнение человека не будет разглашено против его воли.

- **Государства должны инвестировать средства в кампании по повышению цифровой и медийной грамотности и просвещению, информируя общественность о том, как целевая реклама на основе профилирования и слежки влияет на деятельность человека в интернете и угрожает свободе мнения.** Постоянная слежка нарушает права человека и ограничивает его право на свободу мнения и его выражения. Государства должны обеспечить каждому достаточные инструменты и разнообразие информации для свободного формирования собственного мнения и использования положительных аспектов свободы выражения мнения.

- **Государства должны рассматривать рекламу на основе слежки в социотехническом контексте модерации и курирования контента,** а также определить пути решения проблемы централизации власти, в том числе в результате объединения нескольких услуг.

Рекомендации по обеспечению полноценной прозрачности и регуляторных мер в отношении онлайн-таргетинга

- **Государства должны разработать политику в области прав человека, уделяя внимание таким важнейшим правам,** как право на свободу выражения мнения, свободу средств массовой информации, неприкосновенность частной жизни и право на свободу от дискриминации. Государства имеют обязательства в рамках международного права в области прав человека, в том числе позитивное обязательство по защите прав человека от вмешательства других сторон, включая частных субъектов или физических лиц. Соответственно, государства должны гарантировать полное соответствие национального законодательства и политики, регулирующих деятельность интернет-посредников и рекламной индустрии, требованиям международного законодательства в области прав человека.

- **Государства должны гарантировать, что частные субъекты будут действовать в соответствии с надлежащей правовой процедурой и стандартами законности, легитимности и допустимости надзора со стороны независимого и беспристрастного судебного органа, в соответствии с Руководящими принципами предпринимательской деятельности в аспекте прав человека ООН.** Государства должны создать нормативную базу, при помощи которой компании смогут продемонстрировать неукоснительное выполнение своих обязанностей в соответствии с Руководящими принципами.

- **Государства должны эффективно применять существующие законы о защите данных и конфиденциальности.** В этом контексте они обязаны предусмотреть такие принципы, как минимизация данных и целевое ограничение. Государства также

должны эффективно внедрять конкурентное и антимонопольное законодательство, а также другие нормы, направленные на укрепление автономии человека.

- **Государства должны ввести требования в отношении корпоративной слежки – в том числе в отношении целевой рекламы, использующей отслеживание и профилирование. В частности, необходимо ввести требование о надлежащей проверке соблюдения прав человека** и о сборе информации о соблюдении компанией Руководящих принципов предпринимательской деятельности в аспекте прав человека ООН.

- **Государства должны обязать интернет-посредников предоставлять документацию о методах отслеживания и профилирования на основе искусственного интеллекта, используемых ими в рекламных целях.** Они должны требовать от интернет-посредников предоставления разъяснений относительно используемых моделей, категорий собираемых данных и целей их сбора, а также показателей эффективности и результатов тестирования используемых моделей. Государства должны обязать интернет-посредников надлежащим образом разъяснять принципы работы своих рекламных и бизнес-моделей, процессы алгоритмического принятия решений, а также принятия автоматизированными системами решений, влияющих на пользователя. Любая разглашенная информация должны быть изложена понятным и доступным для пользователей языком. Необходимо включать информацию о сборе защищённых персональных характеристик или их аналогов. Другая дополнительная информация должна предоставляться исследователям и регулирующим органам в порядке, обеспечивающем конфиденциальность.

- **Государства должны требовать проведения предварительной оценки воздействия на права человека любых бизнес-моделей, основанных на сборе данных и рекламе.** Оценка должна проводится в рамках четкой нормативной базы и быть прозрачной, независимой и инклюзивной (с проведением содержательных консультаций с потенциально затронутыми группами и другими

заинтересованными сторонами). Этот процесс должен включать в себя надзор со стороны регулирующего органа или независимых заинтересованных сторон, обладающих соответствующей квалификацией, с тем чтобы смягчить негативное воздействие рекламных моделей, предотвратить дискриминацию и обеспечить свободу мнения и его выражения.

- **Государства должны принять новые или усилить существующие ограничения в отношении категорий данных, подлежащих сбору, и методов их использования, а также в отношении категорий данных, которые могут быть разглашены рекламодателям, брокерам данных или третьим сторонам.**

- **Государства обязаны четко определить, каким образом рекламные методы наносят «ущерб»** (как индивидуально, так и коллективно) демократическим процессам, основываясь на принципе предосторожности, чтобы определить рамки запрета деструктивных методов рекламы на основе слежки. К таким запретам относится, например, запрет на использование сложных методов воздействия, основанных на психологических моделях, учитывающих психологические слабости и подверженность манипулированию. В отношении методов сбора данных для целевой рекламы, не выходящих за установленные рамки, государства должны гарантировать строгие требования к прозрачности (например, в отношении размещения скрытой рекламы) и проверки на соблюдение прав человека, ставя превыше всего интересы отдельного человека.

- **Государства должны запретить методы неизбирательного массового сбора и анализа данных пользователей для целевой рекламы, которые наносят пользователям индивидуальный или коллективных ущерб, или нарушают их право на свободу мнения.** Это включает, например, целевую рекламу, основанную на масштабном отслеживании уязвимых аспектов пользователей или таких защищенных персональных характеристик, как этническая или половая принадлежность, религиозные убеждения или сексуальная ориентация. Запреты и ограничения на рекламу на основе слежки

могут применяться по модели запрета обманчивой и сублиминальной рекламы или по модели ограничений на рекламу алкоголя, табака, азартных игр или экологически опасных материалов. Необходимо предусмотреть особые меры для защиты уязвимых/восприимчивых групп, таких как дети и молодежь.

- **Государства должны гарантировать, чтобы персонализированная реклама, использующая методы извлечения персональных данных, применялась только с информированного согласия и на основании сознательного выбора.** Государства должны обеспечить пользователям возможность выбирать, какие из их данных можно использовать и в каких целях, а также каким образом они хотят участвовать в онлайн-дебатах и получать целевую рекламу (включая просмотр персонализированной рекламы и в целом согласие быть объектом слежки ради получения рекламы). Для менее навязчивых рекламных моделей необходимо как минимум предусмотреть возможность отказать в сборе данных, а также гарантировать альтернативные средства обеспечения безопасности пользователей в интернете. Согласие должно быть однозначным, предоставленным без принуждения и основанным на осознанном выборе, а также должно соответствовать законам о защите данных, и учитывать тот факт, что рекламные модели могут влиять на права человека не только путем сбора и анализа персональных данных, но и путем использования другой информации и метаданных. Пользователи должны иметь контроль над тем, какие их данные подлежат сбору, сохранению или удалению, и как именно они будут использованы для рекламы. Государства должны содействовать отражению требований конфиденциальности в разработке систем и их применению по умолчанию.

- **Государства должны обязать интернет-посредников предоставлять информацию о своей модели доходов и обеспечивать прозрачность сети.**

- **Государства должны обязать интернет-посредников каждый раз уведомлять пользователей при применении любой формы**

отслеживания и профилирования, сообщать пользователям о том, как работают такие механизмы, и предоставлять возможность для выражения согласия или отказа в простой и удобной для пользователя форме. Государства должны обязать посредников раскрывать информацию о том, что послужило основанием для демонстрирования пользователю того или иного рекламного контента – его собственная история, информация о местоположении, данные о его активности в социальных сетях, демографические характеристики или иные сведения (включая прокси-аудитории или аудитории двойников («lookalike-аудитории»), в которых группируются пользователи, обладающие определенными характеристиками). Посредников также следует обязать предоставлять информацию о параметрах таргетинга и о категориях аудитории (как на основе поведения, так и на основе контента), а также о руководящих принципах оценки категорий аудитории и о том, проверяются ли алгоритмически созданные категории человеком перед их использованием.

- **Государства должны обеспечить пользователям доступ к данным их профилирования, которые хранятся у интернет-посредников,** а также к любым сделанным в отношении них заключениям (включая метаданные, такие как присвоенные категории, и список рекламодателей, пытающихся оказать на них влияние). Эти данные должны предоставляться пользователям по запросу в понятном и доступном формате. Пользователи должны обладать возможностью редактировать и удалять собственный профиль.

- Путем введения определенных запретов и обязательного требования о предоставлении информации о том, какие именно данные подлежат сбору, хранению и анализу, и для принятия каких рекламных решений они используются, **государства должны решить проблему скрытой рекламы на основе слежки, которая может повлиять на способность человека использовать услуги интернет-посредников в качестве форумов для свободного выражения мнения, доступа к информации и участия в общественной жизни.**

- **Государства должны обязать посредников предоставлять регулярные отчеты о прозрачности,** устанавливая минимальные требования к собираемым данным, используемым категориям и автоматизации, а также к тому, как они влияют на контент и предоставляемую рекламу. Посредники также обязаны предоставлять обязательные, функциональные рекламные базы данных/библиотеки.

- **Государствам следует создать структуру, в рамках которой интернет-посредники будут предоставлять информацию о проведенных ими оценках воздействия на права человека и обеспечивать внешнюю независимую экспертизу.** Интернет-посредники должны проводить оценку того, какие риски ограничения права на свободу выражения и информации могут быть связаны с их политикой и практикой целевой рекламы, а также оценку рисков дискриминации.

- Для обеспечения независимого внешнего аудита рекламных моделей **государства должны требовать от интернет-посредников ведения отчетности в соответствии с принципами конфиденциальности и защиты данных и предоставления доступа к отчетам всем соответствующим государственным органам и независимым заинтересованным сторонам,** включая исследователей и организации гражданского общества.

- **Государства должны требовать от интернет-посредников предоставления исследователям и организациям гражданского общества доступа к своим рекламным данным,** позволяющим им оценить рекламную практику, ее индивидуальное и коллективное воздействие, а также использовать эти данные для проведения исследований, ориентированных на общественные интересы.

- **Государства должны гарантировать демократическое управление,** а также назначить и наделить полномочиями надзорные органы, обладающие опытом в области обеспечения равенства и недискриминации, с целью обеспечения мониторинга

и недопущения неравного или дискриминационного воздействия рекламы на маргинализированные группы.

- **Государства должны укрепить независимость органов по защите данных** и обеспечить их адекватной политической поддержкой, финансовыми ресурсами и полномочиями.

- **Государства должны поощрять координацию действий различных заинтересованных сторон,** наращивание потенциала и изучение влияния дизайна интерфейса на поведение пользователей, а также изучение таких вопросов, как «темные паттерны». Государства также должны способствовать проведению исследований по вопросам маргинализации и гендерной природы цифрового наблюдения и влияния рекламы, а также исследований негативных внешних воздействий бизнес-моделей, основанных на сборе личной информации в глобальном масштабе, что позволяет осуществлять микро-таргетинг отдельных лиц с учетом их конкретных атрибутов, черт и предпочтений. Кроме того, необходимы исследования таких вопросов как потенциальное деструктивное влияние целевой рекламы на поведение человека; связь между бизнес-моделями, основанными на слежке, и заинтересованностью посредников в распространении деструктивного контента; дискриминационный ущерб в результате алгоритмического принятия решений; и связанные с этим серьезные последствия.

- **Государства должны воздерживаться от получения произвольного доступа к данным, собираемым интернет-посредниками.** Запросы на предоставление данных должны основываться на принципах легитимности, законности, необходимости и соразмерности, и предусматривать судебный надзор. Государства должны адекватно применять гарантии безопасности, запрещающие обязательную передачу данных, особенно правоохранительным органам, и в этом плане принимать конкретные меры по защите маргинализированных и уязвимых групп.

- **Государства должны решить проблему сосредоточения власти,** которая подразумевает сбор данных как источника влияния на рынке и еще больше усиливает влияние ряда доминирующих посредников в ущерб потенциальным конкурентам и издателям новостей. Меры могут включать, например, обязательное обеспечение совместимости, переносимости данных (права владения данными) с помощью безопасных механизмов и/или децентрализации власти.

- **Государства должны рассмотреть вопрос о влиянии концентрации рынка цифровой рекламы на существующие медиа и доступность информации, представляющей общественный интерес.** Они должны инвестировать в сильные общественные медиа и независимую журналистику.

- **Государства должны инвестировать в изучение альтернативных источников дохода, которые не зависят от коммерциализации поведения отдельного человека, не влияют на новые модели поведения и не формируют их.** Примеры подобных альтернатив включают контекстную рекламу или таргетинг по простым критериям, на которые соглашается пользователь; прямые связи между поставщиками контента и рекламодателями без посредничества рекламного сектора, который монетизирует контент, отредактированный другими (включая медиа); а также усилия по поощрению инноваций, учитывающих права человека.

- **Государствам следует рассмотреть вариант платформ публичных служб,** которые служат обществу и полностью подотчетны ему, основываясь на демократическом управлении.

2.3 Основные принципы предотвращения попыток государства злоупотребить бизнес-моделями на основе слежки

Бизнес-модели сбора данных, связанные с рекламой на основе слежки, могут стать предметом злоупотребления и со стороны государств. Государственные органы все больше полагаются на получение данных частными компаниями, которые служат «резервуарами потребительских данных». Правительства регулярно получают доступ к различным данным,

предоставляемым частным сектором. В последние годы группами по защите гражданских прав был зафиксирован ряд случаев, когда государственные органы заключали соглашения с брокерами данных с целью получить доступ к личным данным пользователей. Например, Фонд электронных рубежей (EFF) описал случай, когда Погранично-таможенная служба США купила данные[81] системы автоматического распознавания номерных знаков (ALPR) у компании Vigilant с целью поиска лиц, подлежащих депортации.[82] Такие неформальные соглашения между государством и частными организациями представляют серьезную угрозу в плане защиты прав человека. Необязательная и несоразмерная слежка может нарушить безопасность в интернете и затруднить доступ к информации и идеям.[83] Слежка может негативно повлиять на самовыражение людей в интернете, особенно журналистов и членов гражданского общества, и привести к самоцензуре из страха перед постоянным наблюдением. Более того, слежка оказывает несоразмерно сильное воздействие на свободу слова маргинализированных групп, включая расовые, религиозные, этнические, гендерные и сексуальные меньшинства, а также журналистов и правозащитников.[84] Это в равной степени относится к слежке как со стороны государственных органов, так и со стороны частных компаний.

В частности, передовые разработки ИИ открыли новые возможности для массовой слежки со стороны государства, которая опирается на архитектуру бизнес-моделей посредников. Серьезные последствия для защиты прав человека в интернете могут иметь различные инструменты мониторинга контента, используемые государствами для установления взаимосвязей между целевыми пользователями или для присвоения значения или отношения их сообщениям в социальных сетях при помощи обработки естественного языка и анализа эмоциональной окраски сообщений. Когда этот процесс дополняется машинным обучением, государства получают возможность обнаруживать связи и взаимоотношения, которые в противном случае было бы невозможно выявить. В результате в авторитарных государствах правозащитники, политические активисты и маргинализированные группы

81 <https://www.techdirt.com/articles/20190321/09165441842/vigilant-customers-are-lying-about-ices-access-to-plate-records.shtml>.
82 <https://www.aclunc.org/blog/documents-reveal-ice-using-driver-location-data-local-police-deportations>.
83 A/HRC/23/40, <https://undocs.org/en/A/HRC/23/40>.
84 A/HRC/29/32, <https://undocs.org/en/A/HRC/29/32>.

могут подвергаться преследованию за свои убеждения и взгляды, а также несоразмерным и суровым наказаниям.

Данная часть доклада содержит общие принципы, которым необходимо следовать государствам в целях предотвращения массовых нарушений прав человека:

1. **Государственные ведомства и особенно правоохранительные органы должны иметь весьма ограниченный и узконаправленный доступ к данным, не выходящий за рамки конкретных идентификаторов или категорий.**

2. **Сбор данных правоохранительными органами должен быть обоснован конкретными подозрениями.** Правоохранительные органы могут получать доступ только к конкретным записям и контенту. Массовый мониторинг, включая распознавание лиц, недопустим, поскольку может привести к массовой слежке.

3. **Данные, собранные в рамках специальных полномочий в области национальной безопасности, не должны использоваться для каких-либо иных государственных целей, включая правоохранительную деятельность.** Они должны храниться в течение ограниченного срока, после чего ненужные данные должны быть удалены.

4. **Метаданные, например, содержащие информацию о том, где, когда и с кем общался человек, могут содержать чрезвычайно откровенные сведения о частной жизни человека, поэтому им следует обеспечить высокий уровень правовой защиты.**

5. **Необходимо предусмотреть уголовную ответственность за незаконную слежку и обеспечить эффективные средства правовой защиты.** Незаконно собранные данные должны быть неприемлемы в качестве доказательств, а лицам, изобличившим незаконное поведение, должна быть предоставлена защита.

3. Заключение

В данной части доклада подчеркивается влияние бизнес-моделей целевой рекламы и сбора данных и слежки с использованием ИИ на курирование контента, плюрализм информации и способность людей формировать собственное мнение, придерживаться и выражать его, а также получать свободный доступ к информации.

В докладе анализируется связь целевой рекламы с ростом числа могущественных интернет-посредников, которые одновременно выступают в роли привратников в плане выражения мнения и обнародования информации на цифровом рынке идей. В докладе также рассматривается вопрос о растущей зависимости между ценностью и, следовательно, узнаваемостью онлайн-контента и его вкладом в получение рекламной прибыли посредниками. Здесь говорится о том, как данные, характеристики и уязвимые аспекты отдельных лиц используются для целевой рекламы, и описывается влияние коммерческих соображений на управление контентом и информационным пространством в интернете. В докладе также рассматривается вопрос о том, как нынешняя цифровая экосистема может помешать осуществлению абсолютного права на свободу мнения и права искать, получать и распространять информацию любого рода, независимо от физических границ.

Более того, в докладе подчеркивается связь бизнес-моделей интернет-посредников, занимающихся сбором данных, со слежкой со стороны государства. Слежка, как со стороны государственных органов, так и со стороны частных организаций, отрицательно влияет на способность отдельных лиц, в частности журналистов и представителей гражданского общества, к самовыражению, оказывая несоразмерно сильное влияние на маргинализированных лиц и группы. В докладе изложен ряд рекомендаций в адрес государств-участников Организации по безопасности и сотрудничеству в Европе (ОБСЕ) в отношении принятия проактивных, упреждающих и ответных мер. Данные рекомендации, ориентированные на защиту прав человека, сосредоточены на гарантиях абсолютной свободы мнения, обеспечении значимой прозрачности, осуществлении мер регулирования таргетинга в интернете, а также на общих принципах

предотвращения использования государствами бизнес-моделей на основе слежки. Несмотря на то, что некоторые проблемы, такие как недостаточная аргументируемость, прозрачность и подотчетность систем на базе ИИ, связанных с рекламой и управлением контентом, требуют безотлагательного решения, в докладе также отмечается необходимость решения проблем более масштабной экосистемы, основанной на слежке, в целях реальной защиты и продвижения свободы мнения и его выражения в цифровую эпоху.

Данная публикация подготовлена благодаря финансовой поддержке Австрии, Болгарии, Чешской Республики, Финляндии, Франции, Нидерландов, Швеции, Швейцарии и США.

ОБСЕ
Представитель по
вопросам свободы СМИ