**Организация по безопасности и сотрудничеству в Европе**
**Бюро Представителя по вопросам свободы СМИ**

# Искусственный интеллект и дезинформация как вызов многосторонней политике

**Краткий справочный документ для экспертного совещания, организованного Бюро Представителя ОБСЕ по вопросам свободы СМИ 7 декабря 2021 г.**

**Составитель: Дениз Вагнер,**
**советник Представителя ОБСЕ по вопросам свободы СМИ**
**(перевод с англ. яз. В. Гаспаровой)**

г. Вена, ноябрь 2021 года

# СОДЕРЖАНИЕ

# ВВЕДЕНИЕ

Системные проблемы, в принципе, возникают не только в результате дезинформации как таковой – ключевую роль здесь играет именно распространение дезинформации с помощью искусственного интеллекта (ИИ), которое настолько увеличивает масштаб отдельного проблематичного контента, что это приводит к возникновению или усугублению системных последствий, ставящих под угрозу мир и безопасность.

Как отмечалось в первом кратком справочном документе «Международное право и политика в отношении дезинформации в контексте свободы СМИ», международная проблема борьбы с распространением ложной информации и массовой информации, несущей угрозу миру, безопасности и сотрудничеству, возникла за последнее столетие. Существует свод международного права, регулирующего распространение дезинформации, особенно в контексте того ущерба, который она наносит международным отношениям. Сегодня вместе с ростом влияния СМИ, чему способствует усиление роли социальных медиа в информировании общественности, растет и стремление найти решение этой проблемы[1].

Искусственный интеллект играет центральную роль на онлайн-платформах и постепенно становится, если еще не стал, ключевым инструментом формирования и определения информационных пространств в интернете. Благодаря разработке и внедрению искусственного интеллекта, онлайн-платформы способны напрямую влиять на мнения и формы их выражения, что в крупных масштабах также приводит к возникновению системных и структурных угроз всеобщей безопасности. Главной проблемой является безрассудное и повсеместное распространение дезинформации, усиливаемое с помощью ИИ.

В основе этой проблемы лежит информационное насыщение, вызывающее необходимость в структурировании и приоритизации информации, что уже невозможно осуществлять вручную. Без технической поддержки сложно рационализировать, осознавать и осмысленно интерпретировать огромный объем информации, непомерное количество нарративов и контрнарративов, а также темпы создания новостного цикла. Сегодня мы наблюдаем новую тактику ограничения свободы выражения мнения. В то время как цензура направлена на подавление свободы слова, новая тактика делает прямо противоположное, наводняя пространство интернета обилием нарратива, в том числе огромным количеством ложной, неточной и вводящей в заблуждение информации. Это *оружие массового помутнения*[2] оказалось невероятно эффективным в плане порождения хаоса и недоверия к институтам.

[1] Рихтер, А. Г., Краткий справочный документ «Международное право и политика в отношении дезинформации в контексте свободы СМИ» (2021 г.); https://www.osce.org/files/f/documents/8/a/485606.pdf

[2] Christina Nemr and William Gangware, Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age (2019); https://www.state.gov/wp-content/uploads/2019/05/Weapons-of-Mass-Distraction-Foreign-State-Sponsored-Disinformation-in-the-Digital-Age.pdf

# Искусственный интеллект

«Аналитическая машина не претендует на создание чего-то своего, — писала Ада Лавлейс в 1842 году, — она может выполнить любую команду, которую мы сумеем задать. Она может провести анализ, но не способна предугадать какие-либо аналитические соотношения или закономерности[3].

Во многих отношениях это описание справедливо и сегодня. В большинстве международных документов ИИ определяются как машинные системы, так или иначе управляемые при помощи данных и с разной степенью автономности выполняющие цели прогнозирования, предоставления рекомендаций или решений для заданного набора задач, формулируемых человеком, с тем чтобы в конечном итоге оказать влияние на виртуальные и реальные обстоятельства[4].

Таким образом, искусственный интеллект – это интеллект не «машинный», а человеческий, встроенный в структуру данных, которые хранятся и обрабатываются машиной в целях предоставления информации или выполнения процессов с такой скоростью и в таком масштабе, как это делает машина Руба Голдберга[5]. Однако вместо выполнения простых задач сложными способами ИИ предназначен решать сложные задачи простыми способами.

Термин «искусственный интеллект» охватывает множество различных концепций автоматизированных процессов, специфическим компонентом каждого из которых является «алгоритм» – то есть последовательность команд в виде компьютерного кода, который выполняет этот набор команд и в четко определенном формате генерирует выходные данные на основе заданных исходных данных[6].

Системы ИИ часто используются в целях крупномасштабной обработки пользовательских данных и профилирования, что создает риск нарушения прав на неприкосновенность частной жизни и свободу выражения мнения.

В конечном итоге ИИ используется в интернете для поддержки распространения информации среди аудитории («курирование контента»), а также для фильтрации информации в целях выявления и удаления или же понижения популярность противозаконной или нежелательной информации («модерация контента»). Эти процессы влияют на информационную базу, на основании которой люди сегодня взаимодействуют в сети.

---

[3] BBC News, A Point of View: Will machines ever be able to think? (2013); https://www.bbc.com/news/magazine-24565995

[4] Bukovska, OSCE RFOM Strategy Paper to Put a Spotlight on Artificial Intelligence and Freedom of Expression (2019); https://www.osce.org/files/f/documents/9/f/456319_0.pdf; Krivokapic, OSCE Non-Paper on the Impact of Artificial Intelligence on Freedom of Expression (2019); https://www.osce.org/representative-on-freedom-of-media/447829; and OECD Recommendation on Artificial Intelligence (2020); https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449#backgroundInformation

[5] Машина Руба Голдберга представляет собой хитроумное устройство, в котором для решения задач применяется цепная реакция. Таким образом, такая машина применяет сложный метод для выполнения простых задач.

[6] Bukovska, OSCE RFOM Strategy Paper to Put a Spotlight on Artificial Intelligence and Freedom of Expression (2019); https://www.osce.org/files/f/documents/9/f/456319_0.pdf.

## Курирование контента приводит к усилению дезинформации

В интернете хранится колоссальный объем информации. Каждую минуту на YouTube загружается более 500 часов видео, каждый час в Facebook размещается почти 9 миллионов фотографий, а в Instagram публикуется более 500 миллионов историй в день[7].

Аудитории требуется помощь, чтобы ориентироваться в существующем в интернете изобилии контента. В этом отношении полезную функцию выполняют так называемые «рекомендаторы контента» – это управляемые ИИ системы, которые осуществляют поиск в безбрежном объеме информации и предоставляют персонализированные рекомендации относительно выбора контента, предположительно актуального для пользователя[8]. Рекомендательные системы выполняют центральную роль на наиболее популярных сайтах и онлайн-платформах[9], однако такая рекомендация осуществляется отнюдь не беспристрастно. Функционирование таких систем определяется исходным дизайном и коммерческими интересами большинства поисковых систем и платформ социальных сетей. При этом они используют данные о поведении пользователей, чтобы манипулировать их вниманием с конечной целью увеличить доходы от рекламы[10]. Таким образом, эти системы целевой рекламы часто запрограммированы так, чтобы рекомендовать скорее коммерческий, а не информационный контент[11].

Рекомендательные системы используются платформами преимущественно с тем, чтобы предлагать пользователям контент, который, по прогнозам алгоритмов, будет способствовать повышению спроса, дохода и позиции на рынке[12], практически без учета реального содержания распространяемого материала[13]. Это создает финансовые стимулы для разработки и продвижения таблоидного, неоднозначного или иного вызывающего эмоциональный отклик контента, включая недостоверную информацию и дезинформацию.

Исследования показали, что ложные сведения распространяются значительно быстрее, имеют более глубокий и широкий охват, чем правдивые, во всех категориях информации, доступной в интернете. Согласно результатам исследования, проведенного Массачусетским технологическим институтом, истории, основанные на ложной или вводящей в заблуждение информации, ретвитятся на 70 процентов чаще, чем правдивые истории; а на то, чтобы ознакомить 1500 человек с историями, основанными на достоверной информации, нужно почти в шесть раз больше времени, чем на ознакомление такого же количества человек с ложной информацией[14].

---

[7] Domo, Data Never Sleeps 8.0, (2020); https://www.domo.com/learn/infographic/data-never-sleeps-8
[8] Llansó, Hoboken, Leerssen, Harambam, Artificial Intelligence, Content Moderation, and Freedom of Expression (2020); https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf
[9] Cobbe and Singh, Regulating Recommending: Motivations, Consideration, and Principles (2019); https://ejlt.org/index.php/ejlt/article/view/686/979, Table 1
[10] Ricci, Rokach, Shapira, Recommender Systems Handbook (2015); Springer
[11] Bukovska, OSCE RFOM Policy Paper on Freedom of the Media and Artificial Intelligence (2020); https://www.osce.org/files/f/documents/4/5/472488.pdf
[12] Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (2019); Profile Books
[13] Cobbe and Singh, Regulating Recommending: Motivations, Consideration, and Principles (2019); https://ejlt.org/index.php/ejlt/article/view/686/979
[14] Dizikes, Study: On Twitter, false news travels faster than true stories, MIT News (2018); https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308

Несколько лет назад социолог Зейнеп Туфекчи выдвинула тезис о том, что YouTube является так называемым «двигателем радикализации»[15]. По мнению Туфекчи, алгоритмы Google основаны на предположении о том, что людей привлекает информация более экстремальная, чем первый фрагмент контента, изначально просмотренного ими на платформе. Социолог называет это явление «вычислительной эксплуатацией естественного желания человека заглянуть «за кулисы», чтобы еще глубже познать то, что его привлекает». «С каждым новым «кликом» мы испытываем все большее возбуждение от возможности раскрывать новые тайны и постигать все более глубокие истины. YouTube заводит зрителей в кроличью нору экстремизма, а Google увеличивает продажи рекламы»[16].

YouTube является вторым по посещаемости веб-сайтом в мире, при этом 70 процентов его пользовательской активности является результатом просмотра видео, рекомендуемых самой платформой (иными словами, не материалов, преднамеренно найденных пользователями при помощи функции поиска, а тех видео, которые они просматривают по рекомендации самой платформы YouTube или которые автоматически включаются после окончания просматриваемого пользователем видео). Таким образом, распространение дезинформации на этой платформе может принимать огромные масштабы.

Недавнее исследование видеоинформации в интернете, и в частности, материалов на платформе YouTube, показало, что пользователи из неанглоязычных стран наиболее подвержены влиянию контента, который считается тревожным или опасным[17]. Между тем ИИ, предназначенный для модерации контента на платформе Facebook, не способен читать информацию на многих языках тех стран и регионов, где работает эта платформа[18]. Такие «слепые зоны» делают онлайн-платформы особенно уязвимыми, позволяя злоумышленникам публиковать на них вредоносный контент, в том числе дезинформацию[19].

Другой существенный аспект этой проблемы связан с алгоритмами, лежащими в основе рекомендательных систем. Исследования подтверждают способность алгоритмического курирования контента оказывать значительное влияние[20] на общество. Поскольку рекомендательные системы действуют как посредники для распространения дезинформации в цифровой сфере, дезинформация оказывает ощутимое воздействие на общество.

---

[15] Tufekci, YouTube, the Great Radicalizer (2018); https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

[16] Friedersdorf, YouTube Extremism and the Long Tail, The Atlantic (2018); https://www.theatlantic.com/politics/archive/2018/03/youtube-extremism-and-the-long-tail/555350/

[17] Mozilla, YouTube Regrets: A crowdsourced investigations into YouTube's recommendation algorithm (2021); https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf

[18] Canales, Facebook's AI moderation reportedly can't interpret many languages, leaving users in some countries more susceptible to harmful posts, INSIDER (2021); https://www.businessinsider.com/facebook-content-moderation-ai-cant-speak-all-languages-2021-9

[19] В п. 74 отчета независимой международной миссии по установлению фактов по Мьянме, представленном на 39-й сессии Совета ООН по правам человека в 2018 г., упоминается важная роль социальных сетей и выражается сожаление о том, что социальная сеть Facebook не может предоставить данные о распространенности ненавистнических высказываний на своей платформе по отдельным странам, что необходимо для оценки адекватности ее реагирования на них. См. https://undocs.org/en/A/HRC/39/64

[20] Например, в 2014 году социальная сеть Facebook опубликовала исследование, согласно которому она способна активно влиять на эмоциональное состояние пользователей, настраивая свой алгоритм. См. https://www.theguardian.com/technology/2014/jun/29/facebook-users-emotions-news-feeds

Об использовании инструментов ИИ для манипулирования определенными лицами с целью оказать опасное влияние на процесс выборов, демократию и социальную сплоченность, стало известно в результате ряда разоблачений, включая скандал с Cambridge Analytica в 2018 году и недавний скандал с файлами соцсети Facebook, обнародованными Фрэнсис Хауген[21].

## Поляризация

Основываясь на результатах описанной выше оценки предпочтений пользователей, ИИ отправляет отдельным лицам специально подобранный для них контент. Такое генерирование специального контента вызывает озабоченность с точки зрения фрагментации информационных пространств и различных методов поляризации аудитории, в том числе путем сужения выбора видимого пользователю контента, что приводит к частичной «информационной слепоте» (создание так называемых «информационных пузырей»), или же путем рекомендации пользователям контента, подкрепляющего уже имеющиеся у них представления (создание так называемых «эхо-камер»). Вполне естественно, что пользователям комфортно знакомиться исключительно с тем контентом, который поддерживает их собственные взгляды, однако в конечном итоге это может привести к племинизму[22] и серьезно исказить представление об окружающем мире.

Конечно, некоторая степень поляризации неизбежна, и в ней вряд ли можно винить исключительно искусственный интеллект. Многие традиционные СМИ исторически работали на аудиторию единомышленников, то есть представляли собой телеканалы или газеты, имеющие явную политическую ориентацию. Тем не менее, аудиовизуальные СМИ строго регулируются четко прописанными законами и независимыми регулирующими органами, выполняющими контрольные функции и гарантирующими, что новостные программы будут сообщать реальные факты. Печатная пресса также подчиняется собственным кодексам профессиональной этики. В конечном итоге, в большинстве случаев у традиционных СМИ нет никакого карт-бланша на измышление фактов и распространение дезинформации.

Однако онлайн-платформы не связаны журналистской этикой, поэтому достоверность публикуемых на них фактов нельзя считать само собой разумеющейся. Важно, чтобы это осознавало растущее число людей, полагающихся в первую очередь (если не исключительно) на онлайн-платформы как на источники новостей[23].

Технологическая обработка нашего информационного пространства искусственным интеллектом кардинальным образом влияет на то, с какими идеями и информацией мы знакомимся в интернете, тем самым создавая угрозу информационному плюрализму.

---

[21] The Wall Street Journal, The Facebook Files (2021); https://www.wsj.com/articles/the-facebook-files-11631713039 and The Guardian, The Cambridge Analytica Files (2018); https://www.theguardian.com/news/series/cambridge-analytica-files

[22] https://www.theguardian.com/science/blog/2017/dec/04/echo-chambers-are-dangerous-we-must-try-to-break-free-of-our-online-bubbles

[23] Согласно оценкам, 61% «миллениалов» получают новости в основном из социальных сетей. См. https://www.pewresearch.org/fact-tank/2015/06/01/political-news-habits-by-generation/

### Неаутентичное, или манипулятивное, поведение

По данным Европейской комиссии, скоординированное использование фальшивых учетных записей или иных форм неаутентичного поведения с целью искусственного продвижения информации в интернете явно указывает на намерение использовать ложную или вводящую в заблуждение информацию с целью причинения вреда[24].

В начале этого года социальная сеть Facebook опубликовала отчет об угрозах, связанных с операциями по оказанию влияния на свою платформу в период с 2017 по 2020 год[25]. Эти операции определены в отчете как «скоординированные усилия по манипулированию общественными дебатами или их искажению для достижения стратегической цели». Согласно отчету, с 2017 года было выявлено более 150 скоординированных кампаний по оказанию такого влияния в более чем 50 странах.

Инструменты ИИ (например, «армии ботов»), используемые в целях совершения неаутентичных действий, способствуют распространению дезинформации и усиливают ее влияние. В последние годы наблюдается волна активности ботов и троллей, пытающихся манипулировать общественным дискурсом по таким важным вопросам, как выборы или пандемия коронавируса.

Деятельность ботов на основе ИИ, которую называют «оружием массового помутнения», «наводнением пространства» или «открытием всех шлюзов», используется для подавления информации в интернете с целью понизить видимость контента, представляющего общественный интерес.

Это подрывает процесс подлинных общественных дебатов, приводя к разобщенности людей и подпитывая недоверие к демократическим институтам.

Более того, инструменты ИИ также могут использоваться злоумышленниками в попытке заглушить голоса конкретных несогласных в интернете. Примерами этого являются скоординированные кампании по преследованию журналистов, имитирующие общественные движения и использующие управляемые искусственным интеллектом системы распространения информации, которые повышают вирусность таких атак интернете[26].

## Борьба с дезинформацией путем модерации контента

От онлайн-платформ все чаще требуют оказания правительствам более активной поддержки в борьбе с дезинформацией. За последние несколько лет некоторые страны усилили давление на онлайн-платформы, призывая их к автоматизации процессов модерации[27] и, в частности, к

---

[24] European Commission, Tackling COVID-19 disinformation - Getting the facts right (2020); https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020JC0008

[25] Facebook, Threat Report: The State of Influence Operations 2017-2020 (2021); https://about.fb.com/wp-content/uploads/2021/05/IO-Threat-Report-May-20-2021.pdf

[26] Haas, OSCE RFOM Policy paper on freedom of the media and artificial intelligence (2020); https://www.osce.org/files/f/documents/4/5/472488.pdf

[27] См. краткий перечень регуляторных инициатив в области ИИ: Агентство Европейского Союза по основным правам, Инициативы в области политики в отношении ИИ (2016-2019), на сайте: https://fra.europa.eu/en/project/2018/artificial-intelligence-big-data-and-fundamental-rights/ai-policy-initiatives

максимально оперативному удалению контента. Закон Германии о защите Сети (Netzwerkdurchsetzungsgesetz или - NetzDG)[28], принятый в 2017 году, обязывает онлайн-платформы удалять «явно незаконный» контент в течение 24 часов после получения жалобы пользователя[29], а Кодекс поведения ЕС 2016 года призывает платформы удалять или делать недоступными противозаконные человеконенавистнические высказывания в течение максимум 24 часов[30].

Для того чтобы полностью охватить весь требующий модерации онлайн-контент, не хватит никакого количества модераторов, не говоря уже о дополнительном бремени, которое такая работа налагает на тех, кто ее выполняет[31]. Все это обусловливает необходимость в использовании ИИ, по крайней мере, для содействия в модерации контента.

Инструменты ИИ используются онлайн-платформами для масштабного контроля контента с целью выявления целого ряда проблем, включая дезинформацию и манипулятивное поведение. Методы алгоритмической модерации контента используются в целях обнаружения потенциально проблемного контента, а также принятия и исполнения решений о маркировке, обозначении и выделении определенного контента, демонетизизации, понижении или повышении популярности контента с учетом его законности и / или потенциального вреда. Анализ текста и анализ изображений – это два наиболее часто используемых метода модерации для борьбы с распространением потенциально незаконного или опасного контента в интернете. Однако возможности такого автоматизированного анализа контента ограничены, а при его использовании возникает множество проблем.

Во-первых, существует распространенное, но неверное суждение о нейтральности технологий. Проблемы могут возникнуть с самого первого момента разработки алгоритма машинного обучения. Модель машинного обучения разрабатывается на основе набора обучающих данных, подбираемых человеком. Такая модель изучает, воспроизводит и использует данные и навыки, привитые ей разработчиком. Таким образом, дизайн ИИ отражает решения, принятые его создателями. Человек склонен к предвзятости, которая зачастую является врожденной и неотъемлемой частью человеческой природы и меняется под воздействием окружающей среды и опыта[32]. Предубеждения, имеющиеся у разработчиков систем ИИ и заложенные в предоставленных ими данных, могут воспроизводиться системой ИИ на протяжении всего ее жизненного цикла и с ее помощью усиливаться и масштабироваться.

---

[28] Представитель ОБСЕ по вопросам свободы СМИ рассмотрел этот закон и предупредил о его потенциально несоразмерном влиянии на свободу выражения мнения. См. https://www.osce.org/fom/347651.

[29] См. https://www.bmjv.de/DE/Themen/FokusThemen/NetzDG/NetzDG_node.html

[30] См.https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en#theeucodeofconduct

[31] Онлайн-платформы нанимают (и привлекают по внешнему подряду) тысячи модераторов, которые проверяют потенциально опасный и незаконный контент. Эта работа является психологически травмирующей и часто осуществляется на основе нестабильных трудовых договоренностей. Эти человеческие затраты являются веским аргументом в пользу автоматизации. Документальный фильм Блока и Ризевика «Чистильщики» (Block and Riesewick, 'The Cleaners') ярко отображает эти негативные аспекты. См. также: Chotiner, The Underworld of Online Content Moderation, The NewYorker (2019); https://www.newyorker.com/news/q-and-a/the-underworld-of-online-content-moderation

[32] Lebowitz and Lee, 20 cognitive biases that screw up your decisions, INSIDER (2015); https://www.businessinsider.com/cognitive-biases-that-affect-decisions-2015-8

Это явление, широко известное как «алгоритмическая предвзятость», включает в себя расовую, гендерную, классовую и региональную дискриминацию. Джой Буоламвини изобрела термин «кодированный уклон», обозначающий алгоритмическую предвзятость как «встроенные представления, которые распространяются лицами, обладающими полномочиями на кодирование систем»[33]. Этот аспект также особенно сложно измерить, тем более что технология искусственного интеллекта существует в корпоративном «черном ящике» – в большинстве случаев неизвестно, как была разработана та ли иная система ИИ, на каких данных она обучалась, и как функционирует. Именно благодаря результативности систем ИИ нам зачастую становится известно о многих предубеждениях, особенно расовых и гендерных. Примером таких предубеждений является программное обеспечение Google для распознавания лиц на фотографиях, помечавшее чернокожих людей как «горилл»[34]; система автоматической пометки Flickr, помечавшая концентрационные лагеря на карте как «спортивные лагеря» или «спортивные базы в джунглях»[35]; а также программное обеспечение камеры Nikon, ошибочно помечавшее жителей Восточной Азии на фотографиях как «моргнувшие в момент съемки»[36].

Еще более усложняет ситуацию тот факт, что во многих случаях мы даже не знаем, использовался ли ИИ вообще. Отсутствие прозрачности при разработке и применении ИИ особенно затрудняет выявление и устранение алгоритмической предвзятости.

С этой проблемой связано отсутствие разнообразия в составе групп самих разработчиков систем искусственного интеллекта. Как показало исследование, проведенное в 2018 году в 177 технологических компаниях «Кремниевой долины» (США), в десяти крупных компаниях не было ни одной чернокожей женщины, в трех не было ни одного чернокожего сотрудника, а в шести – ни одной женщины-руководителя. Разработкой и приобретением инструментов ИИ, обслуживающих и модерирующих информационное пространство в интернете, занимаются компании и команды, состоящие преимущественно из белых трудоспособных мужчин, и такое недостаточное или искаженное представительство меньшинств или маргинализированных сообществ в процессе разработки систем ИИ приводит к тому, что разработанные системы не отвечают потребностям этих сообществ. В этом отношении было бы несправедливо винить исключительно технологии, отрицая влияние политических и социальных систем, в которых технологии разрабатываются и функционируют[37].

Искусственный интеллект неспособен эффективно понимать или интерпретировать контекст и намерения пользователя, разместившего указанный контент, а в некоторых случаях, и лингвистический, социологический или политический контекст рассматриваемого сообщения. Таким образом, без обеспечения проверки человеком такая модерация контента почти наверняка приведет к незаконным ограничениям.

---

[33] Buolamwini, Fighting the “coded gaze”, Ford Foundation (2018); https://www.fordfoundation.org/just-matters/just-matters/posts/fighting-the-coded-gaze/

[34] Vinent, Google ‘fixed’ its racist algorithm by removing gorillas from its image-labeling tech, THE VERGE (2018); https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

[35] Dewey, Google Maps’ White House glitch, Flickr auto-tag, and the case of the racist algorithm, The Washington Post (2015); https://www.washingtonpost.com/news/the-intersect/wp/2015/05/20/google-maps-white-house-glitch-flickr-auto-tag-and-the-case-of-the-racist-algorithm/

[36] Rose, Are Face-Detection Cameras Racist?, TIME (2010); http://content.time.com/time/business/article/0,8599,1954643,00.html

[37] Digital Freedom Fund, Decolonising Digital Rights; Decolonising Digital Rights: Why It Matters and Where Do We Start? – Digital Freedom Fund

Кроме того, инструменты ИИ не всегда способны определить, является ли тот или иной контент действительно незаконным или опасным. То, насколько конкретный контент нарушает условия обслуживания интернет-платформы (или, в некоторых случаях, противоречит законодательству), зависит от контекста, а именно его технология ИИ не способна учитывать в своей оценке.

Более того, эффективность технологии ИИ зависит от качества наборов данных, используемых для ее обучения. Если эти наборы данных не включают примеры речи на разных языках, которые используют различные сообщества, в частности, маргинальные группы, то разработанная технология не сможет модерировать создаваемый этими группами контент, что может усилить и углубить существующую предвзятость в отношении «недопредставленных» групп.

С учетом этого можно заключить, что для инструментов ИИ всегда будут характерны «ложные срабатывания», когда ИИ будет ошибочно классифицировать контент как нежелательный, а также «ошибочные несрабатывания», когда контент, который должен был быть классифицирован как нежелательный, будет успешно проходить модерацию. Следовательно, для многих умеренность может обернуться несправедливостью.

С точки зрения свободы выражения мнения такие «ложные срабатывания» приводят к цензурированию законного контента, а «ошибочные несрабатывания» лишают возможности устранить вред от дезинформации, тем самым оказывая сдерживающее воздействие на способность отдельных лиц или сообществ взаимодействовать в интернете[38].

Надлежащий надзор и соответствующие процедуры не обеспечиваются и после задействования систем ИИ. Отсутствие соответствующих механизмов подачи и рассмотрения жалоб может привести к тому, что действия, предпринятые в результате алгоритмического принятия решений, будут нарушать право на свободу выражения мнения[39].

Учитывая озабоченность по поводу значимости, масштаба и степени воздействия систем ИИ, приводящих к возникновению ряда проблем, и особенно по поводу их роли в массовом распространении дезинформации, существует необходимость в обеспечении «алгоритмической подотчетности». Подотчетность жизненно важна для создания средств правовой защиты и тем самым для защиты прав и достоинства человека[40].

---

[38] Bukovska, OSCE RFOM Strategy Paper to Put a Spotlight on Artificial Intelligence and Freedom of Expression (2019); https://www.osce.org/files/f/documents/9/f/456319_0.pdf; Krivokapic, OSCE Non-Paper on the Impact of Artificial Intelligence on Freedom of Expression (2019); https://www.osce.org/representative-on-freedom-of-media/447829

[39] ООН, ОБСЕ, ОАГ, АфХПЧН, Совместная декларация о свободе выражения мнения, а также «фейковых новостях», дезинформации и пропаганде (2017 г.); FOM.GAL/3/17 (osce.org)

[40] Исчерпывающий обзор проблем представлен в Стратегическом документе ПССМИ ОБСЕ «В центре внимания: искусственный интеллект и свобода слова» (2019 г.); https://www.osce.org/files/f/documents/9/f/456319_0.pdf

## Концептуальный вызов дезинформации

Ключевой проблемой в борьбе с дезинформацией является невозможность провести четкую грань между фактами и вымыслом и установить явное намерение причинить вред. Непреднамеренные ошибки или определенные формы выражения мнения или убеждений, а также сатиру и пародию нелегко охватить в рамках бинарного анализа фактов или вымысла. Более того, дезинформация, предназначенная для причинения вреда, порой распространяется в интернете третьими сторонами, не имеющими такого намерения (что классифицируется как неумышленное распространение ложной информации).

Главный компонент, который следует учитывать, это *намерение* дезинформировать. В своем недавнем докладе «Дезинформация и свобода мнений и их выражения» Специальный докладчик ООН по вопросам поощрения и защиты права на свободу мнений и их свободное выражение подчеркнула, что некоторые формы дезинформации равносильны подстрекательству к ненависти, дискриминации и насилию, что запрещено международным правом[41].

Другие подробные ссылки на используемые на уровне региона определения дезинформации приведены в более ранних аналитических документах Бюро Представителя ОБСЕ по вопросам свободы СМИ[42].

В конечном итоге важно отметить, что право на свободу выражения мнения является всеобъемлющим и не ограничивается «корректными» заявлениями. Это право также относится и к таким формам «выражения мнения», которые могут рассматриваться как «глубоко оскорбительные»[43]. При этом идеи, информация и мнения, «которые оскорбляют, шокируют или беспокоят государство или любую часть населения»[44], также защищены правом на свободу выражения мнения. В то же время это не оправдывает распространение официальными должностными лицами или представителями государства заведомо лживых или опрометчивых ошибочных утверждений[45].

## Последствия в плане всеобъемлющей концепции безопасности ОБСЕ

Права человека лежат в самой основе ОБСЕ как института. С момента основания СБСЕ / ОБСЕ государства-участники стремились к концептуальному обновлению понятия «безопасность»[46]. Отсутствие конфликта представляет собой одну из составляющих безопасности, при этом другими, не менее важными ее составляющими являются уважение прав человека и основных свобод, а также экономическая и экологическая безопасность. В

---

[41] UNSR Irene Khan, Report on Disinformation and Freedom of Opinion and Expression (2021); A/HRC/47/25 - E - A/HRC/47/25 -Desktop (undocs.org)
[42] Круглые столы с участием на тему дезинформации, ОБСЕ (Expert roundtables on Disinformation | OSCE).
[43] Комитет по правам человека, Замечание общего порядка № 34, CCPR / C / GC / 34, 12 сентября 2011 г., п. 11.4.
[44] «Хэндисайд (Handyside) против Великобритании», жалоба № 5493/72, решение от 7 декабря 1976 г., п. 49.
[45] ООН, ОБСЕ, ОАГ, АфХПЧН, Совместная декларация о свободе выражения мнения, а также «фейковых новостях», дезинформации и пропаганде (2017 г.); FOM.GAL/3/17 (osce.org)
[46] Zannier, Human Rights and OSCE's comprehensive security concept (2017); https://www.osce.org/files/f/documents/b/b/103964.pdf

конечном итоге всеобъемлющее понятие безопасности заключается в обеспечении *безопасности и свободы* людей во всем регионе ОБСЕ.

Обязательства ОБСЕ создают прочную основу для всеобъемлющей безопасности. Хотя ОБСЕ только выиграет от принятия новых обязательств, направленных на решение возникающих проблем в цифровом контексте, существующие обязательства выдержали испытание временем и оказались настолько гибкими, что не потеряли свою актуальность и сегодня.

Когда речь заходит о системных социальных проблемах, таких как дезинформация, контент как таковой не является проблемой. Проблемы возникают тогда, когда информация достигает большой аудитории, и особенно когда она сочетается с другой информацией, усиливающей первоначальный контент[47]. Инструменты искусственного интеллекта, связанные с адресной рекламой, являются одним из ключевых средств широкомасштабного распространения дезинформации в интернете. Основное внимание следует уделять регулированию методов распространения и адресации информации, а не регулированию контента, которое, как объяснялось выше, часто имеет свои недостатки и может еще больше усугубить проблему. Это также во многом соответствует обязательствам ОБСЕ в отношении свободы выражения мнения и свободы СМИ[48].

Демократия требует предоставления гражданам возможности придерживаться противоположных точек зрения, а также общего признания ценности фактов и опоры на них. Алгоритмическое регулирование информации зачастую ограничивает возможность высказывания противоположных мнений, тем самым препятствуя вовлеченности пользователей, но в то же время продвигает сенсационный и вводящий в заблуждение контент для повышения такой вовлеченности. Между тем дезинформация стирает представления об истине и фактах. Опасное сочетание целевой рекламы и дезинформации в интернете не только ослабляет осуществление и реализацию личных прав человека, но может подорвать основы демократии, мира, безопасности и процветания общества.

[47] Cobbe and Singh, Regulating Recommending: Motivations, Consideration, and Principles (2019); https://ejlt.org/index.php/ejlt/article/view/686/979

[48] Еще в 1989 году государства-участники ОБСЕ обязались «обеспечить людям возможность свободно выбирать источники информации» и «с этой целью снять любые ограничения, несовместимые с вышеупомянутыми обязанностями и обязательствами». См https://www.osce.org/files/f/documents/4/f/99565_0.pdf.