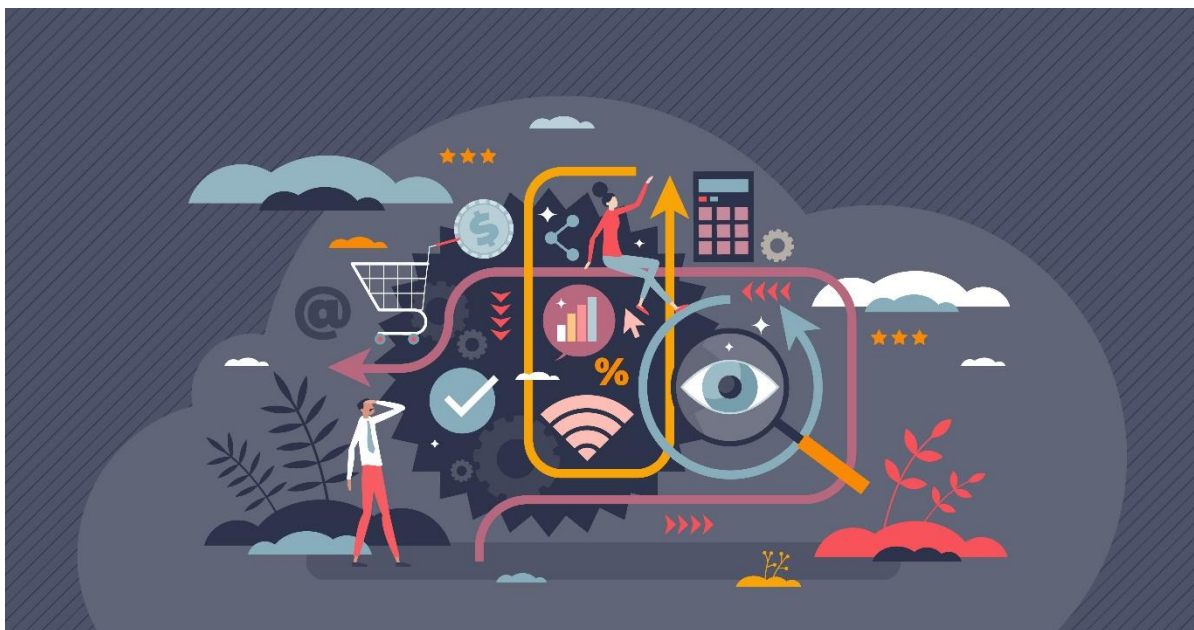




CENTER
FOR GLOBAL
IT-COOPERATION

**ТОКСИЧНЫЙ КОНТЕНТ:
ПРАКТИКИ И ИНСТРУМЕНТЫ САМОРЕГУЛИРОВАНИЯ**

(справка)



Москва, февраль 2022

1. ВВЕДЕНИЕ	3
2. ОПРЕДЕЛЕНИЯ И КАТЕГОРИИ	4
3. МЕЖДУНАРОДНЫЕ ОРГАНИЗАЦИИ	6
4. ПЛАТФОРМЫ	9
5. ОБЩЕСТВЕННЫЕ И ПРАВОЗАЩИТНЫЕ ОРГАНИЗАЦИИ	10
6. ПРАКТИКА СТРАН	12
<i>Великобритания</i>	12
<i>Французская Республика</i>	14
<i>Нидерланды</i>	16
<i>Индия</i>	17
<i>Косово</i>	18
7. ВЫВОДЫ	19

1. ВВЕДЕНИЕ

В справке приведены основные площадки, которые развивают практики и инструменты по защите граждан от вредоносного контента. Материал в том числе дает общее представление о подходе организаций к категорированию контента, который формально не подпадает под законодательные ограничения, однако, несет в себе риски и может оказать деструктивное воздействие и являться опасным, вредоносным или нежелательным для широкой аудитории или отдельных категорий граждан (пользователей).

Приводятся краткие сведения о наиболее распространенных за рубежом подходах, способствующих информационно-просветительской задаче и развитию инструментов саморегулирования применительно к вредоносному контенту в Интернете.

Международная практика последних лет свидетельствует об ужесточении мер регулирования цифрового контента, при этом не всегда обеспечивается баланс безопасности и соблюдения прав человека, что вызывает резонансные общественные обсуждения и противостояние в среде правозащитных организаций.

Среди зарубежных экспертов превалирует мнение, что введение новых запретительных мер, включая блокирование сайтов, должно быть предметом широкой общественной дискуссии с учетом всех ключевых заинтересованных сторон: государства, бизнеса, гражданского общества, академического сообщества, Интернет-пользователей.

Материал по своему наполнению продолжает систематизацию информации, изложенной ранее в Обзоре CGITC "Сравнительно-правовой анализ мер по противодействию распространения противоправного (деструктивного) контента в сети Интернет".

2. ОПРЕДЕЛЕНИЯ И КАТЕГОРИИ

Попытки дать определение и классификацию токсичному контенту делались многими экспертами. Наиболее часто цитируемым исследованием в этой области является исследование Д. Кумара “Разработка классификации токсичного содержимого для различных точек зрения”¹ (Designing Toxic Content Classification for a Diversity of Perspectives), который предполагает, что термин «токсичный контент» используется как зонтик применительно к атакам на основе личных данных, таким как расизм в социальных сетях, травля в онлайн-играх или ответах на посты, троллинг, угрозы насилием, сексуальные домогательства и многое другое. Эти атаки представляют собой подмножество злоупотреблений, происходящих из ненависти и домогательств, более широкой угрозы, которая охватывает любые действия, при которых злоумышленник пытается причинить эмоциональный вред цели (например, преследование, доксинг, сексторция и насилие со стороны интимного партнера). В отличие от проблем классификации спама, фишинга или связанных с ними злоупотреблений, которые могут полагаться на экспертные оценки, токсичный контент по своей сути является субъективной проблемой.

В исследовании К. Курита “На пути к надежной классификации токсического содержания”² дается попытка решения проблем, связанных с обнаружением и фильтрацией токсичного контента в Интернете, который препятствует конструктивному обмену идеями, исключает чувствительных людей из онлайн-диалогов и оказывает воздействие на психическое и физическое здоровье получателей, способствует разжиганию ненависти и ненормативной лексику.

В статье ВНИИ МВД России 2020 года “Токсичный” контент в сети Интернет и его влияние на радикализацию молодежи”³ под токсичным контентом предлагается понимать “различные виды негативной информации, оказывающей деструктивное психологическое воздействие на личность, социальные группы и общество в целом”. Авторы статьи одним из наиболее опасных видов токсичного контента в сети “Интернет” определяют экстремистский контент – информацию, возбуждающую социальную, расовую, национальную или религиозную ненависть и вражду и стимулирующую насилие.

Термин «деструктивный контент» (destructive) в англоязычных материалах применительно к регулированию Интернета используется крайне редко. В материалах Еврокомиссии, например, присутствует термин «Защита от жестокого (агрессивного, оскорбительного) поведения» (Protection against abusive behaviour).

В русскоязычных материалах и документах деструктивный контент в большинстве случаев связан с понятиями «деструктивное поведение» и «деструктивные действия»,

¹ <https://www.usenix.org/system/files/soups2021-kumar.pdf>

² <https://arxiv.org/pdf/1912.06872.pdf>

³ <https://cyberleninka.ru/article/n/toksichnyy-kontent-v-seti-internet-i-ego-vliyanie-na-radikalizatsiyu-molodezhi/viewer>



которые включают в себя намеренное нарушение социальных отношений, включая экстремизм, причинение физического ущерба, моральное унижение людей, жестокость к животным, вандализм и другое.

Термин употребляется чаще всего применительно к защите детей в Интернет-среде. На практике под деструктивным контентом подразумеваются те виды информации, которые перечислены в ФЗ от 29 декабря 2010 г. № 436 «О защите детей от информации, причиняющей вред их здоровью и развитию». При этом, на различных площадках, в том числе в Госдуме ведется дискуссия о необходимости расширения перечня информации, которую следует отнести к деструктивной. В июле 2021 года Общественным советом Роскомнадзора было принято решение о создании двух постоянных комиссий в составе совета: комиссии по защите детей от деструктивного и опасного контента и комиссии по защите персональных данных.

3. МЕЖДУНАРОДНЫЕ ОРГАНИЗАЦИИ

Вопросы определения подходов к вызовам, связанным с распространением токсичного контента входят в повестку ряда международных форумов.

В июне 2021 года на пресс-конференции в Елисейском дворце президент Франции Эммануэль Макрон подтвердил свое внимание к онлайн-регулированию и, в частности, к токсичному контенту. Он призвал к расширению международного сотрудничества, поскольку саммит Большой семерки (G7) проходит в конце этой недели в Великобритании. «Третья большая тема, которая может извлечь выгоду из эффективной многосторонности и которую мы собираемся поднять во время этого саммита G7, - это онлайн-регулирование», - сказал Макрон. «Эта тема, и я уверен, что мы поговорим об этом снова, имеет важное значение для наших демократий».⁴

Подходы Организации экономического сотрудничества и развития (ОЭСР) в отношении одного из наиболее проблемных видов токсичного контента, а именно террористического и экстремистского характера изложены в соответствующем отчете⁵ Организации. В отчете представлен обзор политик и процедур по борьбе с террористическим и экстремистским контентом в 50 ведущих мировых онлайн-сервисах по обмену контентом с акцентом на прозрачность. Результаты исследования показали, что только пять из 50 сервисов выпускают отчеты о прозрачности в отношении террористического и экстремистского контента, и эти сервисы используют разные подходы в своих отчетах, разные определения терроризма и экстремизма, сообщают разные типы информации, используют разные методы измерения и оценки и выпускают отчеты с разной частотой и в разное время. Небольшое количество отчитывающихся компаний и различия в том, что, когда и как они отчитываются, не позволяют получить четкое и полное межотраслевое представление об эффективности мер компаний по борьбе с таким контентом в Интернете и о том, как они могут повлиять на права человека. По мнению ОЭСР, данную ситуацию можно было бы улучшить, если бы больше компаний выпускало отчеты о прозрачности в отношении террористического и экстремистского контента и включили бы более сопоставимую информацию.

В **Евросоюзе** пока в массе своей действуют нормы «мягкого регулирования». В 2016 г. Европейская комиссия согласовала с операторами сервисов Facebook, Microsoft, Twitter и Google (YouTube) «Кодекс поведения ЕС по противодействию разжиганию ненависти и вражды в Интернете»⁶ (в последующие годы к Кодексу присоединились также Instagram, Snapchat, TikTok и LinkedIn). Кодекс предписывает платформам иметь правила и

⁴ <https://techcrunch.com/2021/06/10/macron-says-g7-countries-should-work-together-to-tackle-toxic-online-content/>

⁵ <https://www.oecd.org/sti/current-approaches-to-terrorist-and-violent-extremist-content-among-the-global-top-50-online-content-sharing-services-68058b95-en.htm>

⁶ The EU Code of conduct on countering illegal hate speech online. European Commission // Режим доступа: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en (дата обращения: 01.08.2021).

стандарты, запрещающие разжигание ненависти, создавать системы и группы для проверки контента, предположительно нарушающего такие стандарты. Проверка большей части подозрительных сообщений должна осуществляться в течение 24 часов, при необходимости платформы должны удалять такие сообщения или ограничивать к ним доступ. От платформ также требуется предоставлять отчеты о прозрачности.

В 2017 г. **Еврокомиссия** опубликовала Обращение к другим институциональным структурам ЕС со своим видением борьбы с незаконным контентом в Интернете и призывом к усилению ответственности онлайн-платформ⁷. В документе изложены руководящие принципы для платформ по внедрению передовых методов предотвращения, обнаружения, удаления незаконного контента и ограничения доступа к незаконному контенту. Предполагается, что описанные действия должны осуществляться совместными усилиями онлайн-платформ, национальных властей государств-членов ЕС и других заинтересованных сторон. В документе подчеркивается, что у онлайн-платформ есть реальные и эффективные возможности по предотвращению использования их инфраструктуры для совершения правонарушений, включая технические средства для выявления и удаления противозаконной информации.

В настоящее время в Евросоюзе ведется активное обсуждение проекта Регламента о едином рынке для цифровых сервисов (Digital Services Act)⁸. Целью данного Регламента является гармонизация и унификация требований к обеспечению прозрачности, подотчетности, добросовестности и ответственности провайдеров Интернет-сервисов, в частности, в сфере модерации контента.

Модерация контента определяется в документе как «деятельность, осуществляемая провайдером посреднического сервиса, направленная на обнаружение, выявление и устранение незаконного контента (информации), противоречащего условиям пользовательского соглашения, который размещается (распространяется) пользователем сервиса; принятие мер, влияющих на доступность и видимость контента (информации), таких как понижение приоритета контента, ограничение доступа или удаление контента, а также принятие мер, влияющих на возможность пользователя размещать (распространять) информацию, таких как удаление или приостановление действия аккаунта» (пункт (p) статьи 2).

Одним из принципов регулирования деятельности провайдеров Интернет-сервисов в проекте Регламента назван принцип пропорциональности, предполагающий дифференциацию требований к провайдерам в зависимости от характера и охвата их

⁷ Communication on Tackling Illegal Content Online - Towards an enhanced responsibility of online platforms. European Commission // Режим доступа: <https://digital-strategy.ec.europa.eu/en/library/communication-tackling-illegal-content-online-towards-enhanced-responsibility-online-platforms> (дата обращения: 01.08.2021).

⁸ Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC // Режим доступа: <https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN> (дата обращения: 01.08.2021).



деятельности. Самые крупные провайдеры (ст. 25), играющие системообразующую роль в политическом и экономическом пространстве (оказывающие услуги более чем 45 миллионам пользователей, около 10 процентов от населения ЕС), должны соблюдать максимальное количество требований. Такие провайдеры, помимо прочего, должны будут разрабатывать и применять систему управления рисками, связанными с распространением нелегального контента, нарушением прав граждан и свободы слова, дискриминацией, умышленным манипулированием сервисом и т. п. Алгоритмы используемых систем рекомендаций (recommender systems) и таргетирования рекламы должны быть раскрыты и понятны пользователям, при этом пользователю должна быть предоставлена возможность влиять на параметры этих алгоритмов (ст. 29).

4. ПЛАТФОРМЫ

Многие цифровые платформы используют свои ресурсы и опыт для разработки общих сервисов и технологий, которые могут быть приняты в отрасли. К ним относятся Microsoft PhotoDNA, общая система для обнаружения и реагирования на изображения сексуального насилия над детьми, и Google Perspective API, который использует машинное обучение для обнаружения и маркировки потенциально вредного или «токсичного» контента для модераторов. В ноябре 2018 года Microsoft и другие компании организовали «хакатон» для разработки технологии анти-груминга, которая будет бесплатно лицензирована для небольших компаний по всему миру.

Вопросы выстраивания сбалансированной системы внутрикорпоративной цензуры, фильтрации и удаления токсичного (деструктивного) контента уже несколько лет находятся в поле зрения руководства глобальных цифровых платформ. С одной стороны, приходится учитывать специфику законодательной классификации контента по степени вредности и доступности для различных групп населения в странах операционной деятельности цифровых платформ. С другой – фокус внимания всегда на подходах компаний к работе со своей целевой аудиторией в целях максимального охвата рынка и извлечения прибыли.

Общая схема взаимодействия ИТ-платформ с токсичным (деструктивным) контентом сводится к следующей формуле. В основе системы внутрикорпоративной цензуры лежит профильное законодательство страны происхождения цифровой платформы (или страны с наиболее прибыльным для платформы рынком), а также иные обязательства, добровольно взятые на себя той или иной платформой. По такому принципу работает, например, экосистема Google – там, где это возможно (нет прямых угроз для свободной деятельности по местному законодательству) используется нормы и классификация токсичного (деструктивного) контента с ранжированием доступа по возрасту, предусмотренные нормативно-правовыми актами США. Кроме того, распространена практика прокладки или, как минимум, разделения ответственности цифровых платформ с пользователями за распространение нежелательного контента (по умолчанию предусматривается пользовательскими соглашениями без возможности выбора).

По указанной схеме активно выстраивают работу в странах и регионах производители (дистрибьютеры) многопользовательских онлайн игр, которых в ряде стран мира все чаще начинают причислять к цифровым платформам.

5. ОБЩЕСТВЕННЫЕ И ПРАВООЩИТНЫЕ ОРГАНИЗАЦИИ

UK Safer Internet Centre

Центр⁹ является партнером трех ведущих организаций: Childnet International, Internet Watch Foundation и SWGfL. Основная миссия - сделать интернет безопасным местом для детей и подростков.

В январе 2011 года Европейская комиссия включила Центр в число партнеров по проблемам безопасного интернета в Великобритании.

Три функции центра:

1) Информационно-просветительская. Предоставление консультаций (включая телефон доверия) и поддержки детям и подросткам, родителям и опекунам, персоналу, работающему с детьми. Центр также координирует проведение Дня безопасного Интернета по всей Великобритании.

2) Предоставление поддержки специалистам, работающим с детьми и подростками, по вопросам безопасности в Интернете

3) Горячая линия.

На сайте Центра представлен доклад по вредоносному контенту, к нему привязаны различные интерактивные инструменты для жалоб и получению дополнительной информации по различным видам вредного контента. По каждому виду контента предоставлен набор полезных ссылок для детей и родителей, которые коррелируют с определенной тематикой.

Ниже выдержка из доклада о видах вредоносного контента:

1. Угрозы (Threats). Выделяют 2 типа угроз:

1.1 Гипотетическая угроза. Это может быть выражение несогласия путем несерьезных угроз, вероятность осуществления которых крайне мала. Обычно это не противоречит общественным стандартам социальных сетей, если нет других факторов, которые необходимо учитывать.

1.2 Реальная (правдоподобная, убедительная) угроза. Угроза представляет реальную опасность для жизни, подвергая кого-либо непосредственному риску причинения вреда, например, угроза жизни. Об угрозах такого рода всегда следует сообщать в полицию как о чрезвычайной ситуации. Другими угрозами такого рода может быть "раскрытие" чьего-либо поведения с целью шантажа. Они могут быть использованы для принуждения человека сделать что-то, чего он не хочет, например,

⁹ <https://saferinternet.org.uk/about>

отправить интимное изображение или другое действие, о котором он может впоследствии пожалеть.

2. Самозванство (Impersonation). Присваивание себе личности другого человека, чтобы преследовать его или обманывать. Может включать такие действия, как создание фальшивых учетных записей или захват учетных записей, обычно с целью нападения на человека.

3. Издевательства и притеснения (Bullying & Harassment). Может включать обидные высказывания в адрес отдельного человека или группы людей, троллинг, распространение слухов и исключение людей из онлайн-сообществ. В случае преследования поведение повторяется и направлено на причинение беспокойства или страданий. О повторных домогательствах следует сообщать в полицию.

4. Самоповреждение или самоубийство (Self-harm or Suicide). Большинство платформ не допускают контент, который поощряет, инструктирует или прославляет членовредительство или самоубийство. На некоторых платформах действуют процессы защиты пользователей, которые просматривают или делятся подобным контентом.

5. Жестокое обращение в Интернете (Online Abuse). Широкий термин, который охватывает любую форму оскорбления, совершенного в социальной сети, на сайте, игровой платформе или в приложении. Как правило, это словесное оскорбление, но может также включать оскорбления, основанные на изображениях.

6. Насильственный контент (Violent Content). Может включать графический контент, включая gore-контент, например, видео с обезглавливанием или сцены, прославляющие жестокое обращение с животными. Большинство из них противоречит общественным стандартам различных платформ.

7. Нежелательные сексуальные намерения и действия (Unwanted Sexual Advances). Часто является гендерным насилием и может принимать форму сексуальных высказываний или настойчивых и непрошенных сообщений сексуального характера. Отправителю совершенно безразлично, хочет ли получатель видеть такие предложения.

8. Порнографический контент (Pornographic Content). Контент для взрослых (изображение обнаженного тела или сексуальных сцен), который не является незаконным, но нарушает условия большинства онлайн-платформ.

9. Иной вредоносный контент. Некоторые виды вредоносного (вредного) контента могут явно не подпадать ни под один из перечисленных выше.

6. ПРАКТИКА СТРАН

Великобритания

В декабре 2020 года Правительство Великобритании доработало “Белую книгу” (Online Harms White Paper)¹⁰, в которой изложена программа действий по борьбе с контентом или деятельностью, которые наносят вред отдельным пользователям, особенно детям, или угрожают образу жизни в Великобритании, либо подрывают национальную безопасность, либо подрывают общие права, обязанности и возможности для содействия интеграции. В документе отмечается, что в настоящее время существует ряд нормативных и добровольных инициатив, направленных на решение этих проблем, но они не идут достаточно далеко или быстро, или не являются достаточно последовательными между различными компаниями, чтобы обеспечить безопасность пользователей Великобритании в Интернете. В Белой книге выдвигаются амбициозные планы по созданию новой системы подотчетности и надзора для технологических компаний, выходящей далеко за рамки саморегулирования. Новая нормативная база для обеспечения безопасности в Интернете четко определит обязанности компаний по обеспечению безопасности британских пользователей, особенно детей, в Интернете с помощью самых решительных мер по противодействию незаконному контенту и деятельности. Подчеркивается, что борьба с незаконным и вредным контентом и деятельностью в Интернете является одной из составляющих более широкой миссии Великобритании по разработке правил и норм для Интернета, включая защиту персональных данных, поддержку конкуренции на цифровых рынках и продвижение ответственного цифрового проектирования. В частности, в книге выделены следующие блоки вредоносного контента:

- Сексуальная эксплуатация детей и надругательство над ними в Интернете
- Террористический контент в Интернете
- Контент, незаконно загруженный из тюрем
- Борьба с серьезным насилием в Интернете
- Продажа опиоидов через Интернет
- Борьба с анонимными злоупотреблениями в Интернете
- Кибербуллинг
- Самоповреждение и самоубийство
- Использование сексуальных изображений несовершеннолетними
- Онлайн-дезинформация
- Онлайн-манипуляции
- Оскорбление общественных деятелей в Интернете

В качестве возникающих проблем отмечены:

¹⁰ <https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper>

- **Экранное время**

Время работы с экраном и его влияние на детей - вопрос, вызывающий все большую озабоченность. Исследование Internet Matters показало, что почти половина родителей (47%) обеспокоены тем, сколько времени их ребенок проводит в Интернете, а 88% принимают меры по ограничению использования ребенком устройств.

- **Проектная зависимость**

Некоторые онлайн-продукты были разработаны таким образом, чтобы стимулировать постоянное использование. Они включают, казалось бы, небольшие, но воздействующие на человека функции, которые стимулируют людей продолжать пользоваться приложением или платформой. Одним из распространенных примеров является "бесконечная прокрутка", когда информация загружается непрерывно по мере того, как пользователь прокручивает страницу вниз, побуждая его продолжать прокрутку.

В Белой книге также выделены некоторые виды вредоносного контента, которые “не имеют ясного определения”, к таким видам отнесены, например, пропаганда калечащих операций на женских половых органах (FGM - Female Genital Mutilation); интриги и подстрекательство; различные принуждающие действия.

В документе указано, что в Великобритании сформировался динамичный и инновационный рынок, связанный с безопасностью в Интернете, разрабатывающий инструменты для бизнеса, чтобы защитить своих пользователей от вреда. Например:

- SuperAwesome, один из самых быстрорастущих технологических МСП в Великобритании, предоставляет инструменты и технологии для защиты цифровой конфиденциальности детей;
- Crisp, МСП со сложными инструментами на основе искусственного интеллекта для поддержки модерации и мониторинга контента, помогает сотням компаний по всему миру запускать более безопасные платформы - каждый месяц его системы оценивают миллиарды единиц контента на предмет незаконного или вредного содержания и помогают выявлять рецидивистов, постоянно размещающих неподобающий контент;
- Yoti, поставщик цифровых идентификационных данных, сотрудничает с социальной сетью Yubo, чтобы использовать машинное обучение для определения возраста пользователей веб-сайта, чтобы определить, соответствуют ли пользователи сайта возрастной группе для их платформы - важный шаг в обеспечении безопасности детей в Интернете.

Правительство поддерживает развитие этой зарождающейся экосистемы технологий безопасности в Великобритании. В частности, с помощью программ ускорителей Национального центра кибербезопасности (NCSC) правительство обеспечивает быструю разработку решений проблем кибербезопасности, таких как аутентификация, безопасность мобильных устройств и управление идентификацией, которые направлены на снижение

вреда, наносимого в Интернете, путем повышения безопасности онлайн-среды пользователей, предоставления им большего контроля над своим взаимодействием и затруднения доступа для тех, кто стремится использовать технологии для злоупотреблений.

Французская Республика

В 2019 году Президент Франции Эммануэль Макрон и Премьер-министр Новой Зеландии Джасинда Ардерн после атаки на мечети в новозеландском городе Крайстчерч (ChristChurch), где преступник убил 51 человека, транслируя происходящее в прямом эфире в социальных сетях, организовали в Париже встречу десяти глав государств и высокопоставленных представителей интернет-компаний. Цель встречи заключалась в принятии «Крайстчерчского призыва к удалению из интернета материалов террористического и насильственного экстремистского характера»¹¹.

В июне 2021 года на пресс-конференции в Елисейском дворце президент Франции Эммануэль Макрон подтвердил свое особое внимание к онлайн-регулированию и, в частности, к токсичному контенту. Эта инициатива дала значительные результаты - в сентябре 2019 года на 74-й сессии Генеральной Ассамблеи ООН в Нью-Йорке после нескольких раундов переговоров с интернет-компаниями президент Французской Республики и премьер-министр Новой Зеландии сделали заявления о принятии следующих мер¹²:

1. Реформа структуры Глобального интернет-форума по борьбе с терроризмом (ГИФБТ) и его системы управления для обеспечения большей независимости от компаний-учредителей (Facebook, Microsoft, Twitter и YouTube).

2. Создание в рамках реформированного ГИФБТ рабочих групп, специализирующихся на изучении использования интернета террористами и экстремистами, прибегающими к насильственным действиям, чтобы лучше понять это явление; борьба с ограничениями, накладываемыми алгоритмами, и разработка единой системы обмена данными с соблюдением конфиденциальности и основных прав пользователей.

3. Общий для государств и компаний протокол управления кризисами, создаваемый в увязке с работой, проводимой Европоллом и Европейской комиссией, с целью эффективного и оперативного реагирования в случае террористической атаки и/или вирусного распространения в интернете материалов террористического характера.

В настоящее время эти заявления продолжают претворяться в жизнь:

¹¹ <https://www.christchurchcall.com/supporters.html>

¹² <https://www.diplomatie.gouv.fr/ru/politique-etrangere/cifrovaya-diplomatiya/evnenements/article/l-appel-de-christchurch-quelles-avancees-12-mai-2021>

- ГИФБТ теперь является некоммерческой организацией, юридически не зависящей от четырех компаний – учредителей форума;
- протокол управления кризисами был распространен и пересмотрен совместно с представителями государств, компаний и гражданского общества, поддерживающих «Крайстчерчский призыв», в ходе двухдневного рабочего семинара, организованного Google в Веллингтоне 3-4 декабря 2019 г.;
- был официально создан новый Независимый консультативный комитет в состав которого входят представители государств (Канада, Франция, Новая Зеландия, Великобритания, США, Япония и Гана), Европейской комиссии, Контртеррористического исполнительного директората ООН и 12 организаций гражданского общества (включая организации, специализирующиеся на борьбе с терроризмом и насильственным экстремизмом, защитников цифровых технологий, свободы слова и прав человека, а также представителей академического сообщества), которые составляют большинство в этом объединении;
- ГИФБТ принял на работу своего первого исполнительного директора Николаса Расмуссена.

Помимо этих достижений, «Крайстчерчский призыв» позволил укрепить сотрудничество между Францией, поддержавшими призыв государствами и крупными интернет-компаниями, реализуемое в рамках борьбы с материалами террористического характера в интернете. Он также позволил установить новый диалог с международным гражданским обществом. Сегодня, спустя два года после принятия призыва, его поддерживают 54 государства, Европейская комиссия, Совет Европы, ЮНЕСКО, а также основные поставщики онлайн-услуг (Amazon, Facebook, Google, Microsoft, Dailymotion, Twitter, YouTube, Qwant).

Для того чтобы привлечь гражданское общество к участию в выполнении принятых в Париже обязательств, Франция и Новая Зеландия инициировали создание сети организаций международного гражданского общества, поддерживающих «Крайстчерчский призыв». С этой сетью, состоящей из 47-и организаций, два раза в месяц проводятся консультации. В 2020 г. сеть приняла участие в опросе в ходе широкой консультации со сторонниками «Крайстчерчского призыва», организованной с тем, чтобы лучше понять, как они выполняли обязательства по призыву, и определить меры, которые необходимо принять для дальнейшего прогресса в этом отношении. В этой связи были определены следующие основные темы: совершенствование реагирования во время кризисов, публикация докладов о прозрачности, реализация мер по охвату всех материалов, относящихся к категории материалов террористического и насильственного экстремистского характера, и увеличение сообщества сторонников призыва. Этим темам было отведено центральное место на саммите по случаю второй годовщины «Крайстчерчского призыва», который состоялся 14 мая 2021 г.

Франция продолжает внимательно следить за работой, которая будет проводиться в рамках взятых обязательств. Она является членом Независимого консультативного

комитета реформированного ГИФБТ (наряду с США, Великобританией, Канадой, Японией, Ганой и Новой Зеландией).

Структура реформированного ГИФБТ:

- правление, в состав которого входят четыре компании-учредителя (Facebook, Microsoft, Twitter и YouTube) и могут быть включены другие члены (предприятия малого бизнеса и неправительственные организации). Правление осуществляет набор на должность исполнительного директора;
- независимый консультативный комитет, состоящий из представителей государств (Канада, Франция, Новая Зеландия, Великобритания, США, Япония и Гана), Европейской комиссии, Контртеррористического исполнительного директората ООН и 12 организаций гражданского общества;
- постоянная группа в составе исполнительного директора и генерального секретариата, отвечающая за реализацию трех стратегических направлений деятельности ГИФБТ (предотвращение, реагирование и исследование);
- рабочие группы, созываемые исполнительным директором на основе приоритетов Правления и Независимого консультативного комитета. Первые запланированные рабочие темы касаются правовых и технических аспектов, академических и практических исследований, протокола управления кризисами, воздействия алгоритмов и механизмов деятельности;
- ежегодный многосторонний форум, позволяющий обмениваться информацией и передовым опытом между государствами, гражданским обществом и компаниями. Участники получают полугодовые доклады ГИФБТ и в течение года могут проводить обсуждения посредством телефонных конференций и онлайн-семинаров.

Нидерланды

В исследовании¹³ Glushko & Samuelson Information Law and Policy Lab¹⁴ (август 2021 г.) отмечается, что практика модерации контента провайдерами посреднических услуг, например, платформами социальных сетей, регулируется их собственными «Руководством сообщества» (Community Guidelines – CG) и «Условиями использования» (Terms of Use – ToU). В проекте Еврокомиссии Digital Services Act (DSA)¹⁵ модерация контента определяется как "деятельность, осуществляемая поставщиками посреднических услуг, направленная на обнаружение, идентификацию и устранение незаконного контента или информации, несовместимой с их правилами и условиями". Статья 12 DSA устанавливает

¹³ In between illegal and harmful: a look at the Community Guidelines and Terms of Use of online platforms in the light of the DSA proposal and the fundamental right to freedom of expression (Part 1 of 3)

<https://dsa-observatory.eu/2021/08/02/in-between-illegal-and-harmful-a-look-at-the-community-guidelines-and-terms-of-use-of-online-platforms-in-the-light-of-the-dsa-proposal-and-the-fundamental-right-to-freedom-of-expression-part-1-of-3/>

¹⁴ <https://ilplab.nl>

¹⁵ <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM:2020:825:FIN>

требования к условиям и положениям, указывая, что они должны предоставлять пользователям информацию о процедурах, мерах и инструментах, используемых при модерации контента. Эта информация должна быть не только изложена ясным и недвусмысленным образом, но также должна быть общедоступной и в легкодоступном формате. Более того, платформы обязаны действовать добросовестно, объективно и соразмерно при модерации контента, не забывая об основных правах и интересах заинтересованных сторон. Поскольку DSA не предлагает никаких дополнительных указаний относительно того, как поступать с вредным контентом, у платформ остается значительный простор для самоуправления и свободы действий, что впоследствии сказывается на онлайн-свободе выражения мнений их пользователей. Поэтому очень важно, чтобы «Руководство сообщества» (CG) и «Условия использования» (ToU) были, по крайней мере, достаточно ясными и прозрачными. Таким образом, пользователи платформы могут понимать и прогнозировать, когда и как их свобода выражения может быть ограничена.

Эксперты изучили как определяется вредоносный (harmful) контент в «Руководствах сообщества» и «Условиях использования» 6-и известных крупнейших онлайн-платформ: YouTube, Twitter, Snapchat, Instagram, TikTok и Facebook. Были выбраны 4 категории контента, которые не квалифицируются как «незаконный» (illegal), а указываются как «вредный/вредоносный» контент. К ним относятся: 1) нагота, 2) дезинформация, 3) самоубийства и членовредительство (включая расстройства пищевого поведения) и 4) нападки (язык вражды).

Выяснено, что наиболее четкие и ясные определения имеют термины «нагота» и «самоубийство и членовредительство», остальные категории представлены в разнообразной трактовке (существуют общеупотребимое, научное и образное/творческое восприятие и значение). Соответственно возникают и разные подходы платформ к мерам по реагированию на контент, который можно отнести к данным категориям.

Так, например, из рассмотренных платформ TikTok - единственная, которая предлагает четкое определение «дезинформации». YouTube, напротив, вообще не определяет, что считать таковой.

Индия

В Индии 25 февраля 2021 года Министерством электроники и информационных технологий были приняты Правила информационных технологий (Руководство для посредников и Кодекс этики цифровых медиа) и опубликовано сообщение о целях данного документа¹⁶. Документ предусматривает 3-уровневую систему противодействия распространению незаконного контента в Интернете (саморегулирование платформ;

¹⁶ Government notifies Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules 2021 // Режим доступа: <https://pib.gov.in/Pressreleaseshare.aspx?PRID=1700749> (дата обращения: 01.08.2021)



регулирование на уровне зарегистрированных Министерством саморегулируемых объединений платформ; государственный контроль). Для исполнения предписаний регулятора об удалении запрещенного контента владельцы платформ должны назначать уполномоченное лицо (grievance officer). Кроме того, платформам необходимо обеспечить возможность направления жалоб от пользователей на нарушение требований законодательства.

Косово

Предложения законодательно отрегулировать «токсичный» контент онлайн-СМИ вызвали озабоченность Совета по прессе Косово и Ассоциации журналистов Косово, которые охарактеризовали эту инициативу как «вредную», назвав ее нарушением «международных правил журналистики».¹⁷

¹⁷<https://prishtinainsight.com/press-council-and-ajk-criticise-call-for-state-regulation-of-online-content/>

7. ВЫВОДЫ

1. Вредоносный контент (может быть обозначен как «деструктивный», «оказывающий вредное/негативное воздействие», «токсичный») не имеет пока ясного и единого (универсального для всех) строго научного описания и системы категорирования. Существует разница в определениях многочисленных видов такого контента – соответственно возникают и различия в подходах по внесению материалов в стоп-листы, блокированию или запрещению (на уровне саморегулирования). Кроме того, с развитием технологий возникают новые виды контента (или поведения в сети), которые можно отнести к опасным, вредным или по меньшей мере к негативным/нежелательным.

Следует констатировать, что лидирующие позиции в исследованиях по обозначенным вопросам и установлению норм и правил играют западные экспертные центры и организации. Россия, как правило, вынуждена анализировать уже выпущенные за рубежом рекомендации, руководства и кодексы на английском языке, которые получили широкое распространение и были приняты за определенный «эталон» значительной частью мирового сообщества.

2. Практика крупнейших платформ в отношении вредоносного контента, имеет различия, во многих случаях строго не определена, в том числе, в части исключений к запрещению (exceptions to content restrictions). Это нередко вызывает споры и противодействие со стороны правительств, отдельных социальных групп или индивидуальных пользователей (граждан).

«Руководства сообщества» и «Условия использования» платформ не предоставляют достаточно подробностей и пояснений о том, какой конкретный тип ограничения будет использован в отношении определенного типа контента. Следовательно, такая практика модерации, как, например, удаление контента может восприниматься пользователем как неожиданная и необоснованная¹⁸.

3. Проблема вредоносного контента характерна для таких сфер как литература, кинематограф, телевидение и СМИ, игровая индустрия и др. В этих сферах вопросы саморегулирования, цензуры и самоцензуры в определенной степени уже прошли определенную “проработку”, отдельные полезные подходы могли бы быть взяты на вооружение применительно к контенту в Интернете. Безусловно в сфере цифрового пространства, с учетом обширных технологических возможностей, данная проблема приобрела более сложные формы и требует систематизации на основе объективных исследований и оценки воздействия, включая воздействие на когнитивные функции человека.

¹⁸ <https://dsa-observatory.eu/2021/08/02/in-between-illegal-and-harmful-a-look-at-the-community-guidelines-and-terms-of-use-of-online-platforms-in-the-light-of-the-dsa-proposal-and-the-fundamental-right-to-freedom-of-expression-part-1-of-3/>

4. Особое значение имеет возрастной фактор и маркировка продуктов/контента, их адаптивность для различных возрастных категорий. Так, например, показ вполне объективной (достоверной) картины боевых действий, может оказать негативное воздействие на психику ребенка. Это затрудняет универсализацию подходов к саморегулированию контента по определенным категориям - добавляется фактор возраста аудитории и, соответственно, идентификации пользователя (подтверждение его принадлежности к определенной возрастной группе).

5. Саморегулирование и расширение списка вредоносного контента, включая выявление новых возникающих рисков, прежде всего должно служить основанием для дальнейшего совершенствования законодательной базы.

6. Понимание новых рисков и более понятная формализация категорий существующих и вновь возникающих вредоносных материалов (контента) само по себе не будет работать без обратной связи, то есть наличия доступных и удобных цифровых инструментов и приложений для подачи соответствующих уведомлений или жалоб со стороны пользователей, общественных наблюдателей и правозащитных организаций.

7. Приобретают все большее значение технические возможности и наличие аппаратно-программных средств для системного и точного выявления в сети вредоносного контента, который нередко имеет весьма размытые границы (отличительные характеристики) и может быть формально идентифицирован и воспринят как законный и разрешенный материал. В этом отношении набор машиночитаемых признаков "вредоносности" (негативного влияния, ущерба и т.п.) и соответствующие стандарты и протоколы выявления деструктивного контента требуют постоянной актуализации. Эффективное решение таких задач может быть найдено только на государственном уровне (с учетом недостаточного интереса крупных платформ к таким процессам и их нацеленности на получение доходов).

8. Такие факторы как образование населения, цифровая грамотность, преобладающие в обществе этические ориентиры и ценности имеют первостепенное значение для реализации долгосрочной, плановой политики в сфере борьбы с распространением вредоносного контента в сети. В свою очередь эти факторы во многом обеспечиваются благоприятной экономической ситуацией и благополучием семей.