

Противодействие языку ненависти в социальных сетях: актуальные проблемы современности

Рабочий документ

01

Данный документ был подготовлен исследователями Оксфордского института изучения интернета при поддержке ЮНЕСКО в качестве вклада в реализацию Стратегии и плана действий Организации Объединенных Наций по борьбе с языком ненависти, а также в рамках финансируемого Европейским союзом проекта «#CoronavirusFacts: борьба с «дезинфодемией» вокруг темы COVID-19 в условиях конфликтоопасной социальной среды». Это один из совместных проектов Оксфордского института изучения интернета и ЮНЕСКО, осуществляемых в рамках подготовки пакета методологических инструментов, в котором представлен обзор существующих подходов, информационных ресурсов и исследовательских проектов, касающихся мониторинга, анализа распространенности и долгосрочных последствий языка ненависти в интернете, а также оценки потенциальных возможностей и практических методов противодействия этому явлению. В связи с этим будут приветствоваться любые замечания и комментарии в отношении данного рабочего документа, которые могли бы помочь в проведении более развернутого исследования по этой теме.

Пресечение и искоренение языка ненависти представляет собой деятельность на нескольких уровнях, направленную на устранение первопричин и побудительных мотивов этого явления, предотвращение его перехода в насильственную форму и решение связанных с ним серьезных социальных проблем. Для выработки эффективных мер противодействия языку ненависти, в том числе посредством образовательной и просветительской работы, крайне важно более внимательно отслеживать и анализировать этот феномен, опираясь на достоверные и объективные данные. В эпоху цифровых технологий это означает также необходимость более глубокого понимания ситуации с точки зрения распространенности, вирулентности и воздействия языка ненависти в интернете.

Выявление ненавистнической риторики в сети интернет представляет для исследователя целый ряд сложностей методологического характера, включая дефиниции, используемые для определения этой проблемы, социальный и исторический аспекты, лингвистические тонкости, разнообразие представленных в сети онлайн-сообществ и видов ксенофобского контента, встречающегося в интернете (форма подачи, используемые образы и т.п.). С технической точки зрения язык ненависти в интернете сложен для исследования ввиду нестабильности функционирования систем, используемых для его выявления, непрозрачности характера проприетарных алгоритмов, отсутствия доступа к корпоративным данным и т.д. Ясное видение возможных путей решения этих сложных задач является необходимым условием для более глубокого понимания причин появления и способов распространения ненавистнических высказываний в интернете, и следовательно для выработки эффективных мер противодействия.

В Стратегии и плане действий ООН по борьбе с языком ненависти обозначен ряд приоритетных направлений работы, касающихся отслеживания и анализа ненавистнических высказываний. В указанных документах, в частности, предусматривается, что соответствующие организации системы ООН должны располагать необходимыми возможностями для «распознавания, отслеживания, сбора данных и анализа тенденций, связанных с пропагандой ненависти». Что касается исследования проблемы распространения языка ненависти в интернете, то организациям системы ООН рекомендуется поощрять «большой акцент на исследовании взаимосвязи между неправомерным использованием возможностей интернета и социальных сетей для пропаганды ненависти и факторами, подталкивающими людей к насилию», а также поощрять «картирование новых потенциальных угроз и возможностей, возникающих в контексте развития технологий и интернет-платформ, применительно к проблеме распространения языка ненависти», и «определение протоколов действий, учитывающих новые виды ненавистнического контента в цифровой среде».

За истекший год ситуация с пандемией COVID-19 дополнительно высветила актуальность Стратегии и плана действий ООН по борьбе с языком ненависти в условиях, когда распространившаяся по всему миру волна человеконенавистнических высказываний еще больше обострила проблему нетерпимости и дискриминации в отношении определенных групп населения и дестабилизировала общества и политические системы.

Цель настоящего рабочего документа – дать краткий обзор наиболее важных аспектов, которые необходимо принимать во внимание в рамках противодействия распространению языка ненависти в социальных сетях, будь то в форме введения компаниями-владельцами социальных сетей конкретных требований, принятия карательных правовых норм или использования мер профилактического и просветительского характера. Документ включает три раздела: первая его часть посвящена определению термина «язык ненависти» и обзору связанных с этим проблемным аспектом рамочных нормативных документов; в части 2 анализируются и рассматриваются инструменты и методы отслеживания языка ненависти в интернете, а также обсуждаются способы оценки его распространенности в глобальной сети; в третьей части исследуются некоторые негативные последствия ненавистнической риторики и рассматриваются возможные меры противодействия и профилактики.

1

ОПРЕДЕЛЕНИЕ ПОНЯТИЯ «ЯЗЫК НЕНАВИСТИ»

Сложность борьбы с пропагандой ненависти и законодательного регулирования этой проблемы начинается с определения самого понятия. На сегодняшний день не существует согласованного на международном уровне определения термина «язык ненависти». Между тем это понятие затрагивает сразу несколько аспектов правового характера, таких как право на свободу мнений и их свободное выражение, дискриминация, пропаганда дискриминации, вражды или насилия и подстрекательство к ним.

Основательница проекта «Опасная речь»¹ Сьюзан Бенеш видит две основные сложности в определении термина «язык ненависти». Во-первых, «ненависть» представляет собой размытый с точки зрения смысла термин и может проявляться с разной степенью интенсивности, а следовательно – иметь разные последствия: «означает ли “ненависть” применительно к ненавистнической риторике, что автор высказывания ненавидит сам или подстрекает к то-му, чтобы ненавидели другие, или, может быть, пытается заставить людей испытывать чувство, что их ненавидят»?² Во-вторых, «язык ненависти» по сути означает, что определенный человек или группа лиц являются объектом ненавистнических высказываний из-за своих отличительных особенностей или принадлежности к какой-либо группе. В связи с этим правовая норма или определение должны уточнять, распространяется ли их действие на все типы идентичности и социальные группы, и если нет, то о каких именно группах идет речь. Автор проекта «Опасная речь» утверждает, что чрезмерно широко толкуемые нормы закона могут неправомерно использоваться против социально незащищенных групп населения или политической/гражданской оппозиции, что в определенных ситуациях может пагубно сказываться на положении именно тех групп, которых должен защищать закон о противодействии языку ненависти. С другой стороны,

можно утверждать, что слишком узкая нацеленность определения на конкретные группы или типы социально-культурной идентичности, может являться причиной неравенства в правах или недостаточности правовых инструментов для решения самой проблемы.

Рамки настоящего документа не позволяют рассмотреть вышеперечисленные аспекты во всех деталях. Тем не менее, обзор международных норм и национального законодательства разных стран мира указывает на многосложный характер этой проблемы и различные толкования, которые могут применяться к понятию «язык ненависти».

На глобальном уровне наряду с не имеющей обязательной юридической силы Всеобщей декларацией прав человека существует Международный пакт о гражданских и политических правах (МПГПП), в котором закреплено право на свободу выражения мнений (статья 19) и запрет любой пропаганды ненависти, представляющей собой подстрекательство к дискриминации, вражде или насилию (статья 20). Кроме того, в статьях 19 и 20 предусмотрены допустимые меры ограничения свободы выражения мнений, причем «только такие ограничения, **которые предусмотрены законом и являются необходимыми: (а) для уважения прав и репутации других лиц; (б) для охраны государственной безопасности, общественного порядка, здоровья или нравственности населения.**

В дополнение к указанным принципам в «Рабатском плане действий по запрещению пропаганды национальной, расовой или религиозной ненависти, представляющей собой подстрекательство к дискриминации, вражде или насилию», было предложено **«шесть пороговых**

¹ The Dangerous Speech Project. 2021. <https://dangerousspeech.org/>

² Benesch, S. 2021. Dangerous Speech: A Practical Guide.

The Dangerous Speech Project, 2021, p. 7. <https://dangerousspeech.org/>

критериев» для обоснования ограничений свободы слова, учи-тывающих социально-политический контекст, статус говорящего, намерение вызвать враждебность публики к определенной группе, содержание высказывания, охват потенциальной аудитории и вероятность негативных последствий.

Огромное значение в плане противодействия языку ненависти имеет Международная конвенция о ликвидации всех форм расовой дискриминации, содержащая более строгое чем в статье 20 МПГПП положение, не требующее наличия намерения или «пропаганды ненависти» и включающее «распространение» в перечень наказуемых деяний. Кроме того, отношение к рассматриваемой теме имеют Конвенция о предупреждении преступления геноцида и наказании за него и Конвенция о ликвидации всех форм дискриминации в отношении женщин (КЛДЖ). На основе дискуссий с официальными лицами ООН, экспертами научно-образовательного сообщества и представителями общественных организаций международная правозащитная организация «Статья 19» разработала «Камденские принципы свободы выражения мнений и равенства». Указанные принципы служат методологической основой для толкования положений МПГПП, а содержащиеся в них разъяснения в отношении терминов «подстрекательство», «дискриминация», «враждебность» и «насилие» призваны удерживать заинтересованные стороны от неправомерного применения статьи 20.

В процессе преобразования международных норм и принципов в положения национального законодательства каждая страна применяет свой собственный подход к определению термина «язык ненависти» с точки зрения формы выражения, потенциального объекта и характера последствий ненавистнических высказываний для их признания в качестве языка ненависти. Отсутствие единого для всех определения является одним из основных препятствий на пути противодействия распространению языка ненависти в интернете, где подобный дискурс не обязательно ограничивается пределами национальной территории.

Наличие определений для языка ненависти имеет важное значение и с точки зрения исследовательской и правозащитной деятельности, в частности, для оценки потенциальных социальных последствий. Негативные последствия ненавистнической риторики могут проявляться как на уровне отдельных людей (в виде психологической травмы), так и на уровне социальных групп, сообществ и общества в целом (в виде ущемления в правах и лишения общественных благ). Учитывая, что язык ненависти используется в отношении групп, имеющих общие характерные особенности, чрезвычайно важное значение имеет анализ ситуации на уровне конкретных сообществ. Последствия пропаганды ненависти затрагивают различные группы населения в разной степени: больше всего при этом страдают социально отчужденные группы населения. Более того, эти последствия накапливаются теми, кто страдает от них, причем ключевой переменной в оценке тяжести этих последствий является предыдущий опыт объекта языка ненависти.³

ЯЗЫК НЕНАВИСТИ В ОНЛАЙНОВОЙ СРЕДЕ

Ненавистническая риторика в интернете принципиально не отличается от языка ненависти вне сети. Тем не менее, существуют отличия в характере информационного взаимодействия в среде, где имеет место/происходит распространение языка ненависти, а также в использовании определенных слов, обвинений и конспирологических теорий, которые могут очень быстро меняться, достигать наивысшей точки и бесследно исчезать. Ненавистнические высказывания могут распространяться молниеносно – за несколько часов или даже минут.

В опубликованном ЮНЕСКО в 2015 году докладе под названием «Противодействие пропаганде ненависти в интернете» отмечалось, что размещаемые в сети посты ксенофобского содержания могут создаваться и распространяться с минимальными издержками, не предусматривают в отличие от других письменных материалов этапа редактирования, имеют су-

³ Gelber, K. and McNamara, L. 2016. Evidencing the harms of hate speech. *Social Identities*, Vol. 22, No. 3, pp. 324-341.

щественно разнящуюся по охвату аудиторию в зависимости от популярности сообщения и могут быть опубликованы в разных странах, поскольку серверы и головные офисы интернет-платформ не обязательно находятся в той же стране, что и пользователь и его целевая аудитория. Помимо этого размещенные в интернете материалы ксенофобского содержания могут оставаться доступными более длительное время, периодически вызывая повышение к себе интереса, проникать в другие сети коллективного доступа или возникать вновь, а также распространяться анонимно. В связи с этим широко обсуждается вопрос о том, кто является модератором интернет-пространства, а также следует ли, и если да, то в какой именно момент, удалять соответствующий контент из сети.

Иллюстрацией такой дискуссии стало обсуждение в Германии закона «О мерах в отношении социальных сетей» (NetzDG), который был представлен в 2017 году и в котором содержалось требование к платформам социального общения, имеющим более двух миллионов подписчиков, ввести прозрачную процедуру модерации материалов противоправного характера (в том числе ненавистнических высказываний) и удалять в течение 24 часов контент, признанный незаконным, а также регулярно отчитываться о принятых мерах. Указанный закон подвергся серьезной критике на том основании, что подталкивал интернет-платформы к выполнению функций «негосударственной цензуры» с сохранением за судами полномочий по принятию решений. В связи с этим было высказано опасение в том, что предложенные сроки исполнения и предусмотренные штрафы могут привести к ситуации, когда во избежание вышесказанных штрафов платформы социальных сетей начнут «излишне усердно» удалять контент. В 2020 году закон был пересмотрен и обязал сайты социальных сетей передавать выявленные материалы противоправного характера в Федеральное управление уголовной полиции. Следующий, почти одновременный с вышеупомянутым пересмотр расширил права пользователей, обязав платформы упростить и сделать более удобной для пользователей процедуру подачи жалоб о размещении незаконного контента, а также обеспечить возможность обжалования решений, касающихся его удаления или неудаления.

Таким образом, разработка правовых норм по борьбе с языком ненависти в интернете и вне сети нередко сопряжена со сложной процедурой экспертной оценки, что обусловлено одно-временным отсутствием определений и требованием соблюдения свободы выражения мнений в рамках действующего законодательства. С учетом перечисленных сложностей борьба с языком ненависти должна включать не только методы правового воздействия.

2

ИНСТРУМЕНТЫ И МЕТОДЫ ОЦЕНКИ И МОНИТОРИНГА ЯЗЫКА НЕНАВИСТИ

Правила и инструменты, применяемые для выявления, мониторинга и модерации ненавистнического контента в интернете, отличаются в зависимости от местной специфики, участвующих сторон и конкретных интернет-платформ.

Методы выявления можно условно разделить на две категории: методы более комплексного характера, изначально предполагающие фильтрацию по ключевым словам с задействованием механизмов краудсорсинга, и методы, основывающиеся на использовании людей-модераторов контента, которые оценивают материалы, помеченные пользователями как «язык ненависти», и определяют, являются ли они таковыми. Несмотря на то, что ручной метод имеет явное преимущество, позволяя принимать во внимание контекст высказывания и быстро реагировать на меняющуюся ситуацию, этот подход является более трудоемким, затратным по времени и дорогостоящим, ограничивая возможность его применения в широких масштабах, а также оперативность принятия решений. С учетом этих сложностей и в свете растущих объемов производимого социальными сетями контента, а также развития технологий машинного обучения и методов обработки естественного языка, интернет-платформы и специалисты-исследователи смогли разработать и во все большей степени полагаются сегодня на технологии автоматического выявления. Во многих проектах последнего времени используется широкое разнообразие методов в различной их комбинации. Ниже приводятся основные используемые при этом термины:

- **Машинное обучение:** методы, предполагающие использование компьютерных алгоритмических систем, способных к самообучению на основе полученного опыта и использования баз данных.
- **Автоматическая обработка естественного языка:** методы, позволяющие обрабатывать и анализировать большие объемы текстов на естественном языке.
- **Методы обработки текста по ключевым словам:** методы, использующие онтологическую модель или словари, позволяющие выявлять текст, содержащий ключевые слова потенциально ненавистнического характера.
- **Дистрибутивная семантика:** методы количественной оценки и категоризации семантического сходства между словами, фразами и предложениями на основе их распределения (дистрибуции) в больших выборках данных.
- **Анализ тональности (текста):** методы, позволяющие дать оценку конкретному тексту с точки зрения эмоционального отношения к объекту высказывания.
- **Метаданные об источнике:** некоторые методы предполагают использование для создания модели объекта высказывания метаинформации об имеющихся данных, например сведения о пользователях, ассоциируемых с определенными сообщениями, в том числе статистику сетевого ресурса, например число подписчиков.
- **Глубокое обучение:** класс алгоритмов машинного обучения, использующих многоуровневый подход для последовательного извлечения из первичных данных информации более высокого уровня.

Компании-владелец социальных сетей в значительной степени перешли от практики реагирования на сообщения, помеченные пользователями как ненавистнический контент, к превентивному выявлению и удалению такого контента с помощью своих автоматизированных систем еще до того, как его увидят пользователи. Перечисленные методы, безусловно, необходимы для обеспечения масштабного противодействия языку ненависти. Тем не менее, их применение создает определенные сложности: автоматическое выявление ненавистнического контента неизбежно сопряжено с ошибками и случаями удаления материалов, не содержащих ненавистнических высказываний. Кроме того, чрезмерное удаление контента может иметь демотивирующий эффект и подрывать свободу слова.

В целях повышения эффективности отслеживания языка ненависти ведется постоянная работа по совершенствованию инструментов выявления соответствующего контента. Так, например, **интерфейс прикладного программирования Perspective (Perspective API)**⁴ является инструментом с открытым исходным кодом, разработанным компанией Jigsaw (бизнес-инкубатор Google) и командой Google по методам противодействия нарушениям и широко используемым новостными агентствами и сервисами Google. В нем применена технология машинного обучения для оценки фраз на предмет потенциальной токсичности их использования в коммуникации. Инструмент «Perspective API» реализован на семи языках (английский, французский, немецкий, итальянский, португальский, испанский и русский). По заявлению **Facebook**, последняя версия его инструмента мониторинга и выявления ненавистнического контента на его платформах обеспечивает улучшенный семантический анализ и понимание содержания, включая анализ изображений, комментариев и других элементов текста⁵.

Над созданием инструментов для выявления ненавистнического контента работали также специалисты-исследователи и организации, представляющие гражданское общество. Ниже приводятся некоторые примеры результатов такой работы:

- Кенийская интернет-платформа **Umati** стала одной из первых, обеспечивших ручной мониторинг в сети постов, написанных на распространенных в Кении языках.
- В 2017 году Т. Дэвидсон вместе с коллегами разработал инструмент **Hate Sonar**, в котором использована модель логистической регрессии, обучаемая на данных веб-форумов и социальной сети Twitter.
- В рамках проектов ADL и D-Lab, осуществляемых на базе Калифорнийского университета в Беркли, был разработан **индекс онлайн-ненависти (ИОН)**, предназначенный для преобразования человеческого понимания языка ненависти при помощи самообучающейся интеллектуальной системы в гибко настраиваемый инструмент, который может быть использован интернет-платформами для выявления масштабов и распространенности языка ненависти в сети.
- Команда проекта «Measures & Counter-measures» («Меры и контрмеры») из Института Алана Тьюринга (Великобритания) разработала инструмент на основе технологии глубокого обучения **для выявления в социальных сетях предубеждений в отношении представителей восточноазиатских стран**.
- В 2020 году К. Мун совместно с коллегами разработал **инструмент для выявления ненавистнического контента на корейском языке**. Этот инструмент был обучен распознавать не только ненависть, но и предвзятое отношение.
- **«Hate Meter»** – это самообучающаяся интеллектуальная система, позволяющая путем обработки естественного языка отслеживать антимусульманские ненавистнические высказывания в интернете. Данная платформа функционирует на английском, итальянском и французском языках.
- Система **COSMOS** собирает и анализирует в режиме реального времени данные из Twitter, обеспечивая поиск по ключевым словам, анализ тональности текста и обработку естественного языка.
- Инструмент **MANDOLA** обнаруживает ненавистнический контент с помощью комбинации технологий, а именно анализа тональности, обработки естественного языка, компьютерного осмысления и глубокого обучения.

⁴ Более подробную информацию об API Perspective можно найти на сайте <https://www.perspectiveapi.com/>

⁵ Источник: <https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech/>

Важно отметить, что эффективность мониторинга языка ненависти в интернете зависит от доступности данных, в первую очередь данных с платформ социального общения. Следует также учесть, что многие из разработанных на сегодня инструментов являются одноязычными, чаще всего англоязычными, в связи с чем необходимы дополнительные исследования в отношении эффективности многоязычных методов выявления языка ненависти. При этом подавляющее большинство научных изысканий и усилий по мониторингу языка ненависти на платформах социальных сетей ориентированы на США и страны Европы, что объясняет отставание других регионов не только с точки зрения доступности соответствующих инструментов и данных, но и в плане понимания масштабов распространенности языка ненависти и основных тенденций в этих регионах. Важность устранения этих диспропорций обусловлена в числе прочего зависимостью языка ненависти от конкретного контекста.

РАСПРОСТРАНЕННОСТЬ ЯЗЫКА НЕНАВИСТИ НА ПЛАТФОРМАХ СОЦИАЛЬНОГО ОБЩЕНИЯ

Используя автоматизированные средства обнаружения, разработанные на основе существующих на сегодня методов, Twitter, Facebook, Instagram и YouTube все чаще сообщают о помеченном и/или удаленном на их платформах контенте. В период с января по март 2021 года видеохостинг YouTube удалил 85 247 материалов, нарушающих политику запрета использования языка ненависти. В двух его предыдущих отчетах приводились сравнимые цифры. За тот же период Facebook сообщил о выявлении 25,2 млн. единиц контента, в отношении которых были приняты меры, а Instagram информировал о 6,3 млн. выявленных единиц подобного контента. Согласно последнему опубликованному платформой Twitter в порядке транспарентности отчету, в период с июля по декабрь 2020 года компания удалила 1 628 281 единиц контента, признанного нарушающим политику запрета использования языка ненависти.

В социальных сетях распространенность языка ненависти определяется на основе выборки контента, просматриваемого пользователями. Такая оценка дает представление (причем приблизительное) лишь в отношении ненавистнического контента, размещенного на платформе, и не учитывает материалы, которые компания уже обнаружила и удалила в инициативном порядке. На сегодняшний день Facebook является единственной платформой, предоставляющей информацию о показателях распространенности. По информации компании, в период с января по март 2021 года показатель распространенности языка ненависти колебался в пределах от 0,05% до 0,06%, что говорит о некотором снижении этого показателя по сравнению с двумя предыдущими отчетами. В ряде исследований указывается, что показатель распространенности языка ненависти на основных интернет-платформах, таких как Twitter и Wikipedia, составляет менее 1% от общего объема контента, в то время как на более нишевых альтернативных платформах, таких как Gab и 4chan, 5-8% размещенного контента может носить оскорбительный характер.⁶ Данные о распространенности языка ненависти на платформах социального общения остаются неполными, что отчасти объясняется недостаточной прозрачностью политики интернет-платформ, в том числе в отношении доступа к их данным.

⁶ Zannettou et al. 2018; Mathew et al. 2018; Hine et al. 2017; Vidgen et al. 2019.

3

ПРОТИВОДЕЙСТВИЕ ЯЗЫКУ НЕНАВИСТИ В ИНТЕРНЕТЕ

Необходимо сразу же подчеркнуть, что принятие мер по борьбе с языком ненависти и, соответственно, выбор подходящих инструментов и практических подходов, а также мер профилактического характера осложняется несколькими факторами. В зависимости от конкретных условий ответы, которые различные заинтересованные стороны дают на ключевые вопросы, говорят об отсутствии консенсуса по данной теме. В чем проявляются негативные последствия использования языка ненависти, и насколько серьезными должны быть эти последствия, чтобы оправдать введение мер, регулирующих свободу слова? Если говорить более детально, то какого рода последствия и в результате каких видов ненавистнического речевого поведения, требуют введения мер регулирования с соблюдением международных норм в области прав человека и стандартов, касающихся свободы выражения мнений?

Дополнительные сложности в деле борьбы с языком ненависти создает и сама структура управления интернет-пространством. В числе таких сложностей – долговременный, мобильный, анонимный и межюрисдикционный характер онлайн-контента, широкий спектр организационных структур управления интернет-платформами и разнородная, объединяющая широкий круг заинтересованных сторон система управления интернетом.

Несмотря на это, многие группы и отдельные люди тем или иным образом активно вовлечены в борьбу с ненавистническими высказываниями в интернете, участвуя в профилактических усилиях по формированию у пользователей глобальной сети неприятия к ксенофобскому дискурсу.

ГОСУДАРСТВЕННЫЕ СРЕДСТВА ПРАВОВОЙ ЗАЩИТЫ

Важным инструментом борьбы с языком ненависти являются судебные инстанции. Несмотря на различие существующих в разных регионах мира позиций и точек зрения в отношении языка ненависти, в том числе в онлайн-среде, и их постепенное изменение по мере углубления понимания этой проблемы, существует ряд признанных на международном уровне принципов, региональных соглашений, национальных норм и примеров судебной практики, соответствующих международным стандартам в области прав человека и содержащих положения, которые касаются языка ненависти в интернете и вне сети и о которых рассказывалось в начале данной публикации.

При этом мы можем наблюдать неожиданно возникающие проблемы, связанные с сугубо правовым реагированием на распространение языка ненависти в интернете. Эти проблемные аспекты касаются баланса прав, злоупотребления со стороны влиятельных заинтересованных сторон возможностью ограничения прав граждан, широкого использования т.н. порога криминализации подстрекательства к насилию, а также недостаточной изученности взаимосвязи между ненавистнической риторикой и насилием вне цифровой среды. Не менее важным с точки зрения противодействия языку ненависти в интернете является ключевой для обеспечения правовой защиты вопрос об ограничении полномочий отдельных государств в отношении контроля интернет-пространства. Для эффективного противодействия языку ненависти в

интернете невозможно полагаться исключительно на национальные средства правовой защиты.⁷

В 2016 году рядом крупных технологических компаний был согласован с Европейской комиссией «Кодекс поведения, направленный на пресечение противоправного распространения языка ненависти в интернете», который требует от ИТ-компаний проведения проверки по заявлениям о ненавистнических высказываниях в течение одного дня после получения со-ответствующей информации. Реализация этого подхода сопряжена с рядом сложностей, обусловленных большим разнообразием служебных и рабочих определений языка ненависти. Тем не менее, он представляет собой важный шаг в направлении развития сотрудничества в этой области и согласования как правовых, так и неюридических методов борьбы с распространением ненавистнической риторики в интернете.

МЕРЫ, ПРИНИМАЕМЫЕ КОМПАНИЯМИ СЕКТОРА ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

В 2021 году YouTube и Facebook сообщили об увеличении объемов выявленного и помеченного ими ненавистнического контента, а также доли контента, помеченного ими в сравнении с долей контента, помеченного пользователями. Это свидетельствует о расширении применения автоматизированных систем выявления языка ненависти. При этом данная тенденция сопровождается увеличением числа единиц восстановленного контента в сравнении с предыдущими отчетными периодами. Так, в период с января по март 2021 года Facebook восстановил 408 700 единиц контента, а Instagram – 43 700 единиц. Несмотря на то, что отчетная информация свидетельствует о более частом чем раньше принятии интернет-платформами мер в отношении ненавистнических высказываний, мы не знаем, обусловлено ли это отмечаемым параллельно ростом

числа таких нарушений, ужесточением проводимой платформами политики или увеличением числа ложных заявлений.

Компании-владельцы социальных сетей подпадают под действие национальных правовых систем, в связи с чем их деятельность напрямую регулируется национальным законодательством, что, как правило, положительно сказывается на оперативности их реагирования на просьбы о сдерживании ненавистнической риторики. При этом сами платформы социального общения не имеют территориальной привязки, и по этой причине от них ожидается лишь соблюдение их собственных условий предоставления услуг, которые могут быть более или менее строгими в сравнении со стандартами, установленными международными соглашениями, о которых говорилось в предыдущем разделе. К практическим мерам, принимаемым веб-сайтами социальных сетей, относятся удаление материалов, признанных разжигающими ненависть, а также направление предупреждений пользователям, размещающим материалы ксенофобского содержания, ограничение или запрет их деятельности на платформе. Стандарты сообщества в этом вопросе постоянно эволюционируют, особенно в том, что касается соотношения между автоматизированными и осуществляемыми человеком способами модерации контента.

В свете перечисленных выше сложностей в последние годы получило развитие движение на основе широкого участия заинтересованных сторон, выступающее за повышение транспарентности интернет-компаний в качестве средства обеспечения их большей подотчетности. Речь идет о предложенном принятии по меньшей мере в 30 странах и регионах мер нормативно-правового регулирования, в том числе в рамках разрабатываемого в настоящее время в Европе закона о цифровых услугах. Наряду с этим ИТ-компаниями были предприняты шаги по повышению прозрачности своей деятельности. В 2021 году международная неправительственная организация «Access Now» внесла в свой индекс более 70 компаний, регулярно предоставляющих

⁷ С 2013 года в рамках своей инициативы в области профессиональной подготовки судей ЮНЕСКО содействует повышению уровня осведомленности судебных работников в различных регионах мира в отношении международных и региональных стандартов в области свободы выражения мнений, доступа к информации и безопасности журналистов, сознавая, что для многих работников судебной системы вопросы, связанные с рассмотрением дел, касающихся языка ненависти, являются крайне актуальными. В общей сложности более 23 000 работников судебных органов прошли соответствующую подготовку в том числе в рамках массовых открытых онлайн-курсов 'MOOC' обучения по месту работы и участия в семинарах-тренингах, а также благодаря публикации ряда методических пособий и руководств.

отчетность в порядке транспарентности⁸, тем не менее, таких компаний должно быть гораздо больше.

В кратком тематическом обзоре ЮНЕСКО под названием «Впусти в себя солнце: транспарентность и подотчетность в цифровую эпоху» повышение транспарентности представлено в качестве некоего компромиссного направления усилий между чрезмерным государственным регулированием онлайн-контента, приведшим к несоразмерному ограничению прав человека, и подходом, предполагающим полную свободу действий, который не позволил эффективно решить проблему контента, будь то язык ненависти или дезинформация. В указанном обзоре представлен набор из 26 глобально значимых принципов транспарентности, охватывающих проблематику контента и процедур, меры надлежащей осматрительности, вопросы правовой защиты и полноценной реализации прав, а также коммерческие аспекты, вопросы сбора и использования персональных данных и доступа к информации.

НЕЮРИДИЧЕСКИЕ МЕРЫ РЕАГИРОВАНИЯ И ПРОФИЛАКТИЧЕСКИЕ МЕРОПРИЯТИЯ

Другие неюридические меры реагирования касаются проведения исследований и информационно-разъяснительных кампаний с участием гражданского общества. Они могут также носить профилактический характер с целью формирования у интернет-пользователей неприятия к ненавистнической риторике. К числу таких мер относятся, в частности, инициативы, направленные на осознание истинных причин и возможных последствий пропаганды ненависти в интернете, в том числе с помощью образования, а также усилия по внедрению более эффективных мер правового и технического характера.

Основное внимание при этом уделяется образовательным и просветительским инициативам, предполагающим постоянную профилактическую работу. Просветительские мероприятия помогают повысить осведомленность широкой общественности о пагубных последствиях языка ненависти, бороться с глубинными причинами этого явления и эффективно распознавать технологии манипулирования и риторику, используемую для распространения языка ненависти в интернете и вне сети. Так, по всему миру были разработаны и осуществляются программы медийной и информационной грамотности, направленные на формирование у интернет-аудитории навыков критического анализа опубликованной в сети информации, а также распознавания вызывающего тревогу и ненавистнические настроения контента и дезинформации. Аналогичным образом предпринимаются усилия по выработке контраргументов, направленных на противодействие языку ненависти с помощью позитивного нарратива, как, например, проведенная в 2017 году в Германии, Великобритании и Франции по инициативе и под руководством Facebook кампания «Инициатива поощрения гражданского мужества в сети Интернет». Существует ряд других инициатив гражданского общества, направленных на активное поощрение изменений на уровне самих интернет-платформ. Так, например, в июле 2020 года в результате проведения информационной кампании «Хватит зарабатывать на ненависти!» более 1200 компаний со всего мира объединились в коалицию и призвали к рекламному бойкоту крупных интернет-платформ, а также потребовали модерации ненавистнических высказываний и приостановки рекламы на аккаунтах, поощряющих дискриминацию определенных социальных групп. Указанная информационная кампания стала еще одним голосом в поддержку все более многочисленных призывов бороться с языком ненависти в интернете, особенно на фоне усилившихся нападков на социально отчужденные группы в период пандемии COVID-19. Это побудило несколько компаний-владельцев социальных сетей внести изменения в правила, установленные ими для своих пользовательских сообществ.

⁸ <https://www.accessnow.org/transparency-reporting-index/>

РЕКОМЕНДАЦИИ

Критически важное значение для понимания, мониторинга, а также сбора данных и анализа тенденций, связанных с распространением языка ненависти в интересах принятия практических мер противодействия, имеет выработка научно обоснованной политики, направленной на борьбу с разжиганием ненависти в интернете и на предотвращение перерастания ненавистнической риторики в насилие. При этом необходимо обеспечить соблюдение права на свободу выражения мнений. Сформулированные ниже рекомендации призваны помочь в определении перечня основных практических мер с целью преодоления новых сложных проблем, которые возникли в связи с широким распространением языка ненависти, и в частности, на минимизацию его пагубных последствий для мира, стабильности и полноценной реализации каждым человеком своих основных прав.

1. Поощрять выработку всеобъемлющих определений языка ненависти при одно-временном соблюдении права на свободу выражения мнений

- Обеспечить соответствие определений международным стандартам, в частности, предусмотренным в МПГПП и Рабатском плане действий.

2. Создавать коалиции с участием широкого круга заинтересованных сторон

- Поощрять обмен данными и опытом между правозащитными организациями, провайдерами интернет-услуг и широкой общественностью.
- Содействовать расширению возможностей заинтересованных сторон, в частности местных сообществ, в вопросах выявления и мониторинга ненавистнической риторики в социальных сетях с учетом конкретных условий и местных языков.
- Организовать с привлечением широкого круга заинтересованных сторон обсуждение тенденций, связанных с пропагандой ненависти, методами распространения ненавистнического дискурса и способами борьбы с ним.
- Выступать во взаимодействии с группами специалистов и широкой общественностью в поддержку создания платформ для выработки определений и рабочих процедур, которые должны касаться не только государств Северной Америки и Западной Европы, но и охватывать все большее число стран по всему миру.

3. Собирать данные и поощрять политику открытого доступа там, где данные уже собраны, с соблюдением требований о защите персональных данных

- Собирать качественные данные о лицах, ставших объектом ненавистнических высказываний, с тем чтобы улучшить понимание масштабов проблемы и характер причиненного вреда.
- Разъяснять на уровне компаний-владельцев интернет-платформ важность совершенствования методов обеспечения транспарентности, в том числе публикации открытых данных о жалобах по поводу ненавистнического контента и принятых практических мерах, а также о корректности работы используемых ими систем модерации контента, в частности, в исследовательских целях.
- Поддерживать разработку недорогих, доступных и удобных в использовании инструментов и методик, которые могут быть применены с целью мониторинга и выявления языка ненависти в многоязычной и мультикультурной среде, причем в сроки, позволяющие принимать меры противодействия.

4. Поощрять интернет-платформы к тому, чтобы они предлагали продуманные варианты правовой защиты лицам, чей контент был удален

- Содействовать сотрудничеству между компаниями-владельцами социальных сетей и группами гражданского общества, занимающимися вопросами цифровых прав, с тем чтобы обеспечить

соответствие процедур модерации и удаления контента потребностям сообщества.

5. Содействовать повышению грамотности в вопросах, касающихся языка ненависти, а также медийно-информационной грамотности и развитию навыков использования цифровых технологий посредством осуществления образовательных программ

- Предоставить финансирование и ресурсы, необходимые для разработки образовательных программ, способствующих формированию неприятия языка ненависти, с учетом современных тенденций распространения ненавистнической риторики и форм реагирования на соответствующие вызовы. Это потребует налаживания тесного взаимодействия между компаниями-владельцами социальных сетей, исследовательскими организациями и заинтересованными сторонами сектора образования.
- Отдавать предпочтение методам разъяснительной профилактической работы, направленной на информирование о пагубных последствиях использования языка ненависти в интернете и способствующих повышению медийно-информационной грамотности, перед мерами противодействия и минимизации последствий.
- Содействовать налаживанию и поддержанию партнерских отношений между образовательными учреждениями и компаниями – владельцами социальных сетей с целью расширения доступа к информации и ресурсам по противодействию языку ненависти на платформах социального общения путем проведения целенаправленных кампаний по распространению соответствующей информации или перенаправления пользователей на внешние ресурсы.

6. Оказывать поддержку организациям, активно вовлеченным в борьбу с распространением языка ненависти в интернете

- Оказывать содействие и обеспечить адекватную ресурсную поддержку специализированным организациям, занимающимся мониторингом и пресечением распространения языка ненависти, в особенности тем, которые лучше других способны учесть местную специфику.

Настоящее исследование является частью сборника дискуссионных документов, подготовленных по заказу ЮНЕСКО и Офиса Специального советника ООН по предупреждению геноцида. Указанные документы являются прямым вкладом в реализацию Стратегии и Плана действий ООН по борьбе с языком ненависти и публикуются в контексте Многостороннего форума заинтересованных сторон и Конференции на уровне министров по противодействию языку ненависти в сфере и при помощи образования, состоявшихся в сентябре и октябре 2021 года.

Вспышка коронавирусной инфекции высветила актуальность Стратегии и плана действий Организации Объединенных Наций по борьбе с языком ненависти в условиях, когда распространившаяся по всему миру волна человеконенавистнических высказываний еще больше обострила проблему нетерпимости и дискриминации в отношении определенных групп населения и дестабилизировала общества и политические системы. В дискуссионных документах предпринята попытка раскрыть основные проблемные аспекты, связанные с этим глобальным вызовом, и предложить возможные меры реагирования и рекомендации.

Данное исследование было инициировано секцией ЮНЕСКО по вопросам свободы выражения мнений и обеспечения безопасности журналистов в рамках финансируемого Европейским союзом проекта «#CoronavirusFacts: борьба с «дезинформацией» вокруг темы COVID-19 в условиях конфликтной социальной среды». Документ был подготовлен исследователями Оксфордского института изучения интернета Джонатаном Брайтом, Антонеллой Пери-ни, Энн Плойн и Реей Уисс.

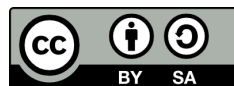
Данный документ является частью сборника дискуссионных документов, подготовленных по заказу ЮНЕСКО и Офиса Специального советника ООН по предупреждению геноцида (OSAPG). Документы являются непосредственным вкладом в Стратегию и План действий ООН и публикуются в контексте Многостороннего форума заинтересованных сторон и Ми-нистерской конференции по борьбе с языком вражды через образование в сентябре и октябре 2021 года.

Вспышка пандемии COVID-19 подчеркнула актуальность Стратегии и Плана действий ООН, породив волну языка вражды по всему миру, что еще больше усугубляет нетерпимость и дискриминацию в отношении определенных групп и дестабилизирует общества и политические системы. В этих дискуссионных документах делается попытка раскрыть ключевые вопросы, связанные с этим глобальным вызовом, и предложить возможные ответы и рекомендации.

Этот документ был подготовлен по заказу Секции ЮНЕСКО по свободе выражения мнений и безопасности журналистов в рамках проекта “#CoronavirusFacts, Addressing the ‘Disinfodemic’ on COVID-19 in conflict-prone environments”, финансируемого Европейским Союзом. Он был подготовлен Джонатаном Брайтом, Антонеллой Перини, Анной Плуин и Реей Уиссом из Окс-фордского института Интернета при Оксфордском университете.

Опубликовано в 2022 г. Организацией Объединенных Наций по вопросам образования, науки и культуры
7, Place de Fontenoy, 75352 Paris 07 SP, France

© ЮНЕСКО 2022



Данный документ предлагается в открытом доступе под лицензией Attribution-ShareAlike 3.0 IGO (CC-BY-SA 3.0 IGO) (<http://creativecommons.org/licenses/by-sa/3.0/igo/>). Используя содержание этого документа, пользователи соглашаются с правилами пользования Репозитория открытого доступа ЮНЕСКО (www.unesco.org/open-access/terms-use-ccbysa-rus).

Название оригинала: *Addressing hate speech on social media: contemporary challenges*

Опубликовано в 2021 г. Организацией Объединенных Наций по вопросам образования науки и культуры.

Использованные названия и представление материалов в настоящем документе не являются выражением со стороны ЮНЕСКО какого-либо мнения относительно правового статуса какой-либо страны, территории, города или района или их соответствующих органов управления, равно как и линий разграничения или границ.

Ответственность за взгляды и мнения, высказанные в данном документе, несут авторы. Их точка зрения может не совпадать с официальной позицией ЮНЕСКО и не накладывает на Организацию никаких обязательств.

Настоящий документ был подготовлен при финансовой поддержке Европейского Союза. Ответственность за его содержание несут исключительно авторы, и оно не обязательно отражает точку зрения Европейского Союза.

Графический дизайн: Дин Дорат

CI/FEJ/2021/DP/01 /REV