



Генеральная Ассамблея

**Доклад эксперта на тему:
«Определение границ допустимого контента в цифровой среде в эпоху
искусственного интеллекта: роль государств, частных платформ и
международных организаций»**

Аннотация

Развитие цифровых платформ и генеративного искусственного интеллекта усложнило традиционную проблему границ свободы выражения мнений. Вопрос сегодня касается не только удаления незаконного материала, но и алгоритмического ранжирования, демонетизации, возрастных ограничений, маркировки синтетического контента, автоматизированной модерации и процедур обжалования.

Международное право прав человека предоставляет исходную нормативную рамку. Статья 19 Международного пакта о гражданских и политических правах защищает право искать, получать и распространять информацию и идеи независимо от государственных границ и выбранного средства коммуникации. Ограничения свободы выражения допустимы лишь при соблюдении требований законности и необходимости для достижения предусмотренной Пактом легитимной цели; Комитет ООН по правам человека дополнительно требует строгого соблюдения соразмерности. [2, 3]

Статья 20(2) МПГПП требует запрещать законом пропаганду национальной, расовой или религиозной ненависти, представляющую собой подстрекательство к дискриминации, вражде или насилию. Поскольку общего международно-правового определения «языка ненависти» не существует, Рабатский план действий предлагает оценивать, достигает ли высказывание соответствующего порога, с учётом контекста, положения автора, умысла, содержания и формы, масштаба распространения и вероятности причинения вреда. [2, 4]

Настоящий доклад предлагает распределённую четырёхфункциональную модель цифрового управления. Международные стандарты прав человека задают пределы допустимого вмешательства; государства определяют незаконность контента и обеспечивают надлежащую правовую процедуру; цифровые платформы и разработчики ИИ осуществляют операционное управление своими сервисами при условии прозрачности и подотчётности; ООН и иные международные механизмы обеспечивают координацию, научную оценку, совместимость подходов и развитие потенциала.

Главный вывод заключается в том, что ни государство, ни частная компания, ни международная организация не должны становиться единственным субъектом, определяющим допустимость глобальной публичной речи.

Содержание

1. Введение	4
2. Незаконный контент, вредный контент и международно-правовой порог	5
3. Генеративный ИИ, синтетический контент и происхождение информации.....	6
4. Алгоритмическое распространение и власть рекомендаций	8
5. Автоматизированная модерация и риск масштабируемой ошибки	9
6. Права детей.....	10
7. Международно-правовые и правозащитные основы	11
8. Государства.....	12
9. Частные цифровые платформы.....	13
10. Международные организации.....	14
11. Современная практика: Digital Services Act	15
12. Современная практика: ЮНЕСКО	16
13. Современная практика: Глобальный цифровой договор	17
14. Современная архитектура ООН по искусственному интеллекту	17
15. Рамочная конвенция Совета Европы об ИИ.....	18
16. Основные системные риски	19
17. Предлагаемая четырёхфункциональная модель	21
18. Рекомендации государствам	22
19. Рекомендации цифровым платформам и разработчикам ИИ	23
20. Иерархия мер вместо автоматического удаления	24
21. Роль Организации Объединённых Наций	25
22. Заключение	26
23. Библиография	27

1. Введение

Интернет и цифровые платформы превратились в ключевую инфраструктуру общественного общения. Однако власть современной платформы заключается не только в возможности удалить публикацию. Она способна повысить или снизить её алгоритмическую видимость, ограничить монетизацию, установить возрастной барьер, добавить предупреждение, изменить положение материала в поиске или рекомендовать его многомиллионной аудитории.

Генеративный искусственный интеллект добавляет к этой системе новый уровень. Он позволяет быстро создавать тексты, изображения, аудио и видео, а одновременно используется платформами для выявления и классификации огромных объёмов пользовательского контента. Следовательно, управление современной цифровой средой включает по меньшей мере пять самостоятельных вопросов: кто определяет юридическую незаконность; кто применяет правила конкретного сервиса; кто управляет алгоритмическим охватом; каким образом устанавливается происхождение синтетического материала; кто и как рассматривает жалобу на ошибочное решение.

Глобальный цифровой договор, принятый в составе Пакта во имя будущего, закрепляет цель создания инклюзивного, открытого, безопасного и защищённого цифрового пространства, основанного на международном праве и правах человека. Он отдельно призывает технологические компании и социальные платформы повышать прозрачность условий использования, модерации, рекомендательных алгоритмов и обращения с пользовательскими данными. [8]

В отношении контента, создаваемого ИИ, Глобальный цифровой договор предусматривает развитие инструментов идентификации, подтверждения происхождения, маркировки, водяных знаков и иных технических средств. Однако эти меры рассматриваются в контексте снижения потенциального вреда; из документа не следует, что вся информация, созданная с использованием ИИ, должна считаться недопустимой. [8]

Таким образом, главная задача международного сообщества заключается не в формировании единого каталога разрешённых и запрещённых мнений, а в создании легитимной процедуры принятия решений о цифровом контенте.

2. Незаконный контент, вредный контент и международно-правовой порог

Для разработки устойчивой системы регулирования необходимо прежде всего разделять юридическую незаконность и более широкое понятие вреда.

Незаконность контента определяется применимым правом с учётом международных обязательств государства. При этом сама по себе ложность, спорность, оскорбительность либо потенциальная общественная вредность высказывания не означает автоматически, что государство вправе запретить его распространение.

Статья 19 МПГПП устанавливает высокий порог для вмешательства. Ограничение должно быть предусмотрено законом, преследовать одну из разрешённых целей и соответствовать строгим критериям необходимости и соразмерности. Комитет ООН по правам человека подчёркивает, что чрезмерно широкие ограничения

несовместимы с этим стандартом и что государству необходимо обосновать конкретную необходимость вмешательства. [2, 3]

Особенно осторожно следует обращаться с понятием «язык ненависти». Универсального международно-правового определения hate speech нет. Статья 20(2) МПГПП использует более конкретную конструкцию: пропаганда национальной, расовой или религиозной ненависти должна достигать уровня подстрекательства к дискриминации, вражде или насилию. Рабатский план действий помогает оценивать этот порог индивидуально и контекстуально. [2, 4]

Это имеет непосредственное значение для автоматизированной модерации. Система способна хорошо распознавать отдельные слова, визуальные признаки или статистические закономерности, но правовая оценка подстрекательства может зависеть от исторического контекста, намерения, статуса говорящего, вероятности вреда, сатиры, цитирования, журналистского документирования или академического использования.

Следовательно, эффективная цифровая политика должна стремиться не к максимальному объёму удалённой информации, а к максимально точному соотношению между характером вреда и характером вмешательства.

3. Генеративный ИИ, синтетический контент и происхождение информации

Генеративный ИИ существенно снизил стоимость массового производства убедительных цифровых материалов. Эта технология имеет многочисленные общественно полезные применения, включая образование,

науку, перевод, творчество и повышение доступности информации. Одновременно она может использоваться для мошеннического олицетворения, манипуляции, производства вводящих в заблуждение материалов и иных форм вредоносной деятельности.

Поэтому само происхождение материала из генеративной системы не должно автоматически определять его правовой статус.

Более пропорциональный подход состоит в оценке риска конкретного использования. Синтетическое учебное изображение, художественный материал, политическая сатира, мошенническая имитация личности и сфальсифицированный кризисный видеоматериал создают принципиально разные риски и не должны регулироваться одинаково.

Механизмы provenance, authenticity, labeling и watermarking способны повышать прозрачность и помогают пользователю оценить происхождение информации. Глобальный цифровой договор прямо призывает технологические компании и разработчиков совершенствовать такие инструменты. [8]

При этом маркировка должна рассматриваться как один элемент системы управления риском, а не как замена анализу контекста. Она не решает самостоятельно проблему достоверности: человеческий материал также может быть ложным, а синтетический — достоверным или общественно полезным.

4. Алгоритмическое распространение и власть рекомендаций

Современное управление информационной средой нельзя оценивать исключительно через число удалённых публикаций.

Платформа может оставить материал доступным, но исключить его из рекомендаций; ограничить монетизацию; понизить его позицию в поиске; ограничить распространение для несовершеннолетних; добавить предупреждение; либо, напротив, активно рекомендовать материал огромной аудитории.

Следовательно, существенная часть цифровой власти реализуется через распределение внимания, а не формальный запрет.

Глобальный цифровой договор поэтому отдельно говорит о прозрачности рекомендательных алгоритмов и призывает платформы предоставлять исследователям доступ к релевантным данным при соответствующих гарантиях конфиденциальности. [8]

Глобальные принципы ООН по обеспечению целостности информации дополняют эту рамку, выделяя общественное доверие и устойчивость, здоровые стимулы, расширение возможностей пользователей, независимые и плюралистические СМИ, прозрачность и исследования. Документ также обращает внимание на экономические стимулы технологических компаний и взаимодействие бизнес-моделей с информационными рисками. [9]

Поэтому будущая политика должна охватывать не только удаление контента, но и системные механизмы, определяющие его видимость и распространение.

5. Автоматизированная модерация и риск масштабируемой ошибки

Масштаб крупнейших платформ делает автоматизированную модерацию практически неизбежным элементом цифровой инфраструктуры. Автоматизация способна быстро выявлять повторяющийся и очевидно противоправный или нарушающий правила материал, но одновременно масштабирует ошибочные классификации.

Эмпирические исследования показывают, что проблема не является чисто гипотетической. Исследование Fasching и Lelkes, опубликованное в Findings of ACL 2025, сравнило семь ведущих модерационных систем на более чем 1,3 млн синтетических предложений и обнаружило заметные расхождения в том, как разные системы классифицировали идентичный контент; различия были особенно выражены применительно к отдельным демографическим группам. [17]

Исследование AfriHate, опубликованное NAACL 2025, сформировало наборы данных по hate speech и abusive language для 15 африканских языков с участием носителей соответствующих языков и культур. Авторы показывают, что качество классификации существенно зависит от языка и что нехватка локальных данных и контекстуальной экспертизы способна приводить как к недостаточной модерации, так и к цензурированию на основе примитивного keyword spotting. [18]

Следовательно, автоматическая система не должна рассматриваться как безошибочный финальный источник оценки контекста.

Вместо универсального требования ручной проверки каждой публикации более реалистичны четыре гарантии:

- предварительное тестирование систем;

- регулярный анализ ошибок и различий между языками и группами;
- понятное объяснение значимых решений;
- доступная возможность содержательного пересмотра, включая человеческий пересмотр там, где автоматическая классификация способна существенно затронуть права пользователя.

6. Права детей

Защита несовершеннолетних является самостоятельной частью цифрового регулирования, но не отменяет их права на выражение мнений и доступ к информации.

Замечание общего порядка № 25 Комитета ООН по правам ребёнка прямо распространяет анализ прав ребёнка на ИИ, алгоритмы, автоматизированные системы и цифровую среду. Комитет указывает, что фильтры и другие ограничения выражения детей должны быть законными, необходимыми и соразмерными, а основания ограничений должны быть прозрачными и сообщаться в доступной возрасту форме. [6]

Комитет также предупреждает, что автоматизированное фильтрование, профилирование и принятие решений не должны манипулировать или подменять способность ребёнка формировать и выражать собственные взгляды. [6]

Поэтому «безопасность детей» не должна использоваться как основание для неопределённой массовой фильтрации. Предпочтительна система, сочетающая возрастные гарантии, *privacy-by-design*, цифровую

грамотность, прозрачность, возможность жалобы и оценку конкретных рисков.

7. Международно-правовые и правозащитные основы

В центре регулирования цифрового контента должны находиться международные права человека.

Статья 19 Всеобщей декларации прав человека и статья 19 МПГПП закрепляют свободу мнений, выражения и получения информации. Статья 20 МПГПП устанавливает отдельные обязательства применительно к пропаганде войны и определённым формам подстрекательства к дискриминации, вражде или насилию. [1, 2]

При этом необходимо различать обязанности государств и ответственность предприятий.

Система ООН исходит из того, что государства обязаны защищать людей от нарушений прав человека третьими лицами, включая предприятия; бизнес, в свою очередь, несёт ответственность уважать права человека, не допускать негативного воздействия и реагировать на такое воздействие. UN Guiding Principles являются ключевым международным стандартом этой корпоративной ответственности. [5]

Для технологических компаний human rights due diligence особенно значима потому, что дизайн платформы, рекомендательных систем, рекламы, модерации и процедуры апелляции способен воздействовать на права одновременно большого числа пользователей.

ЮНЕСКО развивает эту логику в Guidelines for the Governance of Digital Platforms 2023 года. Документ использует многосторонний подход и распределяет роли

между государствами, платформами, межправительственными организациями, гражданским обществом, СМИ, академическим и техническим сообществом. Его разработке предшествовал глобальный процесс консультаций, собравший более 10 000 комментариев из 134 стран. [7]

Следовательно, уже существующий международный подход направлен скорее на регулирование процедур, ответственности и системных рисков, чем на создание мирового списка разрешённых убеждений.

8. Государства

Государства сохраняют центральную роль в определении юридической незаконности и применении юридической ответственности.

Именно национальное законодательство в пределах применимой юрисдикции определяет многие конкретные составы противоправного поведения. Государственные институты обладают также полномочиями по расследованию, судебному рассмотрению и назначению предусмотренных правом последствий.

Однако государственное регулирование цифрового контента связано с особым риском чрезмерного ограничения законной речи.

Если такие понятия, как «ложная информация», «экстремизм», «оскорбительный материал» или «угроза общественному порядку», формулируются неопределённо, правоприменение может выйти далеко за пределы конкретного вреда. Международный стандарт статьи 19 МПГПП поэтому требует ясного правового основания, легитимной цели, необходимости и соразмерности. [2, 3]

Оптимальная государственная модель должна включать:

- ясность правовой нормы;
- независимый судебный или иной эффективный контроль;
- уведомление о причинах ограничения;
- возможность обжалования;
- защиту журналистской, академической, художественной и общественно значимой деятельности;
- недискриминацию;
- прозрачность государственных запросов на ограничение контента в той мере, в какой это совместимо с законными целями расследования и безопасности.

Государству следует прежде всего устанавливать правовые рамки и гарантии, а не пытаться централизованно определять допустимость каждого отдельного высказывания.

9. Частные цифровые платформы

Большая часть ежедневных решений о цифровом контенте принимается не судами и не международными организациями, а частными платформами.

Пользовательские условия и community standards способны быть строже минимального порога юридической незаконности. Это само по себе не означает, что каждая корпоративная норма должна быть превращена в государственный запрет.

Профессиональная сеть, детский сервис, образовательная платформа и видеохостинг могут иметь различные правила использования.

Однако масштаб крупнейших платформ создаёт проблему частной инфраструктурной власти. Компания может влиять на публичную дискуссию посредством

удаления, рекомендаций, поиска, рекламной инфраструктуры, монетизации и дизайна пользовательского интерфейса.

Практика DSA демонстрирует значимость различия между законом и правилами платформы. По данным Европейской комиссии, в первом полугодии 2025 года 99% соответствующих решений о модерации принимались в целях соблюдения собственных условий платформ, а не вследствие сообщений о незаконности по праву ЕС или государств-членов. [11]

Это означает, что прозрачность должна отвечать как минимум на вопрос: почему именно ограничен материал — потому что он незаконен или потому что нарушает частные правила сервиса?

Платформам целесообразно обеспечивать ясность правил, предсказуемое их применение, уведомление пользователя, доступную систему апелляций, оценку воздействия на права человека, отчётность по автоматизированной модерации и достаточную языковую и культурную компетентность систем и персонала. [5, 7]

10. Международные организации

Глобальная природа цифровой среды создаёт очевидную потребность в международной координации.

Одна публикация может создаваться в одной стране, размещаться инфраструктурой в другой, распространяться компанией из третьей и одновременно достигать пользователей в десятках или сотнях юрисдикций.

Однако из трансграничности интернета не следует необходимость создания единого мирового органа, который определял бы допустимость конкретных мнений.

Более реалистичная и институционально совместимая с ООН роль международных организаций заключается в формировании принципов, развитии доказательной базы, координации, совместимости подходов, содействии исследовательскому сотрудничеству и помощи государствам, обладающим ограниченными экспертными ресурсами.

Генеральная Ассамблея в рамках своей компетенции обсуждает вопросы повестки и выносит государствам рекомендации. Официальные материалы UN Model United Nations подчёркивают рекомендательный характер резолюций Генеральной Ассамблеи по вопросам повестки. [19]

Следовательно, модельная резолюция по данной теме должна прежде всего согласовывать стандарты и рекомендуемые меры, а не претендовать на создание всемирной исполнительной службы модерации.

11. Современная практика: Digital Services Act

Digital Services Act Европейского союза представляет интерес прежде всего как пример процедурного и системного регулирования платформ. [10]

Его ценность состоит не в создании универсального перечня недопустимых политических или общественных высказываний, а в развитии процедур прозрачности, объяснения решений, жалоб и оценки рисков.

Европейская комиссия сообщила в феврале 2026 года, что платформы отменили почти 50 млн решений, затрагивавших пользовательский контент или аккаунты. Из

165 млн решений, обжалованных пользователями через внутренние механизмы, около 30% были отменены. [11]

Такой масштаб пересмотра показывает, что механизм апелляции не является формальностью: даже крупные системы модерации систематически принимают решения, которые затем изменяются.

DSA также развивает statements of reasons, посредством которых пользователь должен получать конкретную информацию о причинах удаления или иного ограничения его материала. [12]

Для международного сообщества этот опыт важен не как обязательная глобальная модель, а как подтверждение принципа: регулировать необходимо не только конечный результат модерации, но и процедуру его принятия и пересмотра.

12. Современная практика: ЮНЕСКО

Руководящие принципы ЮНЕСКО по управлению цифровыми платформами предлагают многостороннюю модель, в которой различные субъекты имеют разные роли и обязанности. [7]

Правозащитный подход ЮНЕСКО сочетает human rights due diligence, прозрачность и подотчётность платформ, empowerment пользователей, независимый надзор и участие заинтересованных сторон. [7]

Особая ценность этой модели заключается в отказе от предположения, что государство или технологическая компания обладают достаточной информацией и легитимностью для единоличного управления всей цифровой средой.

13. Современная практика: Глобальный цифровой договор

Глобальный цифровой договор сформировал универсально согласованную межправительственную рамку, связывающую цифровую безопасность, права человека, информационную целостность, управление данными и международное управление ИИ. Его следует отличать от международного договора с самостоятельным механизмом исполнения. [8]

Он призывает цифровые компании повышать прозрачность модерации и рекомендаций, поддерживает исследовательский доступ к данным при соответствующих гарантиях и предусматривает развитие технических механизмов идентификации синтетического контента. [8]

Особенно важно, что документ одновременно призывает противодействовать дезинформации, hate speech и другим формам цифрового вреда и подчёркивает необходимость соответствия международному праву и защиты свободы выражения.

Тем самым Глобальный цифровой договор поддерживает не максималистскую модель удаления информации, а модель управления рисками в рамках прав человека.

14. Современная архитектура ООН по искусственному интеллекту

26 августа 2025 года Генеральная Ассамблея консенсусом приняла резолюцию A/RES/79/325, учредив Независимую международную научную группу по искусственному интеллекту и Глобальный диалог по вопросам управления ИИ. [13]

Научная группа состоит из 40 экспертов. Её мандат включает подготовку ежегодных научно обоснованных оценок возможностей, рисков и последствий ИИ в немилитарной сфере, а также тематических материалов. Члены группы были назначены Генеральной Ассамблеей в феврале 2026 года. [14]

Первый предварительный доклад группы был опубликован 1 июля 2026 года и использовался как доказательная база для первой ежегодной сессии Глобального диалога. [14]

Первая сессия Глобального диалога прошла в Женеве 6–7 июля 2026 года. ООН сообщает о более чем 4 200 зарегистрированных участниках почти из 170 государств и более чем 1 500 письменных представлениях, поступивших на этапе подготовки. Вторая сессия запланирована на 3–4 мая 2027 года в Нью-Йорке. [15]

При этом Глобальный диалог не является регуляторным, правоприменительным или переговорным органом по выработке обязательного договора. Его функция заключается в обмене практиками и формировании общего понимания вопросов управления ИИ. [15]

Это подтверждает более общий вывод: современная архитектура ООН движется в направлении научного обеспечения, координации и совместимости подходов, а не создания глобального органа цензуры.

15. Рамочная конвенция Совета Европы об ИИ

Рамочная конвенция Совета Европы об искусственном интеллекте, правах человека, демократии и верховенстве права CETS No. 225 была открыта для подписания 5

сентября 2024 года и является первым международным договорным инструментом об ИИ такого рода. [16]

Европейский союз ратифицировал Конвенцию 15 мая 2026 года. Однако по состоянию на 30 августа 2026 года условие её вступления в силу — пять ратификаций, включая как минимум три государства — члена Совета Европы — ещё не выполнено. [16]

Следовательно, значение Конвенции заключается прежде всего в демонстрации возможного перехода от этических рекомендаций и *soft law* к договорным обязательствам. Она не должна представляться как уже действующее универсальное регулирование цифровой модерации.

16. Основные системные риски

1. Государственная цензура. Чем менее определены понятия запрещённой информации, тем легче режим противодействия конкретному вреду превращается в механизм подавления журналистики, общественной критики или политической дискуссии. Международный стандарт статьи 19 требует поэтому точности и соразмерности ограничений. [2, 3]

2. Корпоративная концентрация информационной власти. Отсутствие государственного вмешательства само по себе не обеспечивает нейтрального информационного пространства, поскольку платформы продолжают определять правила сервиса, рекомендации, рекламу, поиск и монетизацию.

3. Масштабируемая алгоритмическая ошибка. Автоматизированная система способна повторить ошибочную классификацию не один раз, а в отношении

огромного числа пользователей; академические исследования подтверждают значительные различия между модерационными моделями и между языковыми контекстами. [17, 18]

4. Цифровая фрагментация. Несовместимые национальные режимы способны увеличивать транзакционные издержки, создавать конфликт юрисдикций и затруднять деятельность небольших государств, которые не располагают собственными ресурсами для сложного технологического надзора.

5. Чрезмерное удаление законной информации. Угроза особенно высока там, где платформа несёт очевидные издержки за сохранение проблемного материала, но почти не сталкивается с последствиями за ошибочное удаление. Масштаб отмен решений в рамках DSA демонстрирует практическую ценность систем апелляции. [11]

6. Неравенство языков и культур. Технологическая безопасность не является глобальной, если качественная модерация существует только для небольшой группы доминирующих языков. AfriHate показывает, насколько важны локальные данные и участие носителей соответствующего культурного контекста. [18]

17. Предлагаемая четырёхфункциональная модель

Вместо выбора одного «владельца» границ допустимого контента предлагается распределить полномочия и ответственность между четырьмя функциями.

Функция	Основной субъект	Основная задача	Ключевые гарантии
Нормативная	Международные правозащитные стандарты	Задают минимальные границы государственного вмешательства: свобода выражения, недискриминация, частная жизнь, права ребёнка и эффективные средства правовой защиты.	Законность, легитимная цель, необходимость, соразмерность.
Публично-правовая	Государства	Определяют незаконность поведения в национальном праве в соответствии с международными обязательствами; создают процессуальные гарантии и судебную защиту.	Ясность норм, независимый контроль, уведомление и обжалование.
Операционная	Платформы и разработчики ИИ	Применяют правила сервисов, управляют рекомендациями и модерацией, реагируют на новые формы цифрового вреда.	Прозрачность, human rights due diligence, объяснимость, апелляции, анализ ошибок.
Доказательно-координационная	ООН, ЮНЕСКО и международные научные механизмы	Формируют сопоставимые стандарты прозрачности, доказательную базу, обмен практиками, совместимость подходов и экспертные возможности государств.	Инклюзивность, научная независимость, развитие потенциала, международный диалог.

Первая — нормативная функция международных правозащитных стандартов. Международное право должно задавать минимальные границы государственного вмешательства: свобода выражения мнений, недискриминация, неприкосновенность частной жизни, права ребёнка и эффективные средства правовой защиты. Международная система не должна превращаться в перечень разрешённых и запрещённых политических мнений.

Вторая — публично-правовая функция государства. Государства определяют незаконность поведения в своём законодательстве в соответствии с международными обязательствами, создают процессуальные

гарантии, обеспечивают судебную защиту и устанавливают правила деятельности платформ там, где это необходимо для защиты прав и иных законных интересов.

Третья — операционная функция платформ и разработчиков. Частные сервисы применяют собственные правила, управляют рекомендациями, осуществляют модерацию и реагируют на новые формы цифрового вреда. Их операционная гибкость должна сопровождаться прозрачностью, human rights due diligence, объяснением решений, механизмами жалоб и систематическим анализом ошибок.

Четвёртая — доказательно-координационная функция международных организаций. ООН, ЮНЕСКО, научные механизмы и другие международные институты должны помогать формировать сопоставимые стандарты прозрачности, научную доказательную базу, обмен практиками, совместимость подходов и экспертные возможности государств.

18. Рекомендации государствам

Генеральной Ассамблее целесообразно рекомендовать государствам:

- обеспечивать соответствие любых ограничений свободы выражения требованиям законности, легитимной цели, необходимости и соразмерности;
- избегать неопределённых общих запретов на «ложную» или «вредную» информацию;
- различать незаконность и более широкие системные риски;
- обеспечивать уведомление, обжалование и независимый контроль при существенном вмешательстве;

- учитывать права ребёнка при проектировании цифровых мер безопасности;
- повышать прозрачность государственных запросов на ограничение информации в пределах, совместимых с законными целями;
- развивать экспертизу для языков и культурных контекстов, недостаточно представленных в глобальных технологиях.

Международно-правовую основу такого подхода образуют прежде всего статьи 19 и 20 МПГПП и связанные с ними правозащитные стандарты. [2–4, 6]

19. Рекомендации цифровым платформам и разработчикам ИИ

Генеральной Ассамблее целесообразно призвать платформы и разработчиков ИИ:

- публиковать понятные правила модерации;
- различать решения, основанные на законе, и решения, основанные на terms of service;
- объяснять существенные ограничения пользователю;
- предоставлять доступные механизмы обжалования;
- проводить оценку воздействия на права человека;
- анализировать различия качества модерации между языками и группами;
- обеспечивать содержательный пересмотр значимых автоматизированных решений;
- создавать безопасные механизмы доступа исследователей к данным;
- повышать прозрачность рекомендательных систем;
- совершенствовать provenance и labeling для синтетических материалов высокого риска;

- публиковать сопоставимую статистику по модерации, апелляциям, отменам решений и степени автоматизации.

Такой подход совместим с направлениями Глобального цифрового договора, UN Guiding Principles и Руководящих принципов ЮНЕСКО. [5, 7, 8]

20. Иерархия мер вместо автоматического удаления

Если контент не достигает порога незаконности, удаление не должно автоматически становиться первым и единственным средством воздействия.

В зависимости от характера риска могут использоваться менее ограничительные инструменты:

- предупреждение;
- дополнительный контекст;
- маркировка происхождения;
- снижение алгоритмического усиления;
- ограничение монетизации;
- возрастная мера;
- пользовательские настройки рекомендаций.

Выбор меры должен зависеть от серьезности потенциального вреда, вероятности его наступления, масштаба распространения, характера аудитории и воздействия меры на свободу выражения.

Для незаконного материала государство вправе предусматривать и более строгие последствия, но соответствующие нормы и процедуры должны отвечать применимым международным обязательствам. [2, 3]

21. Роль Организации Объединённых Наций

ООН находится в уникальном положении, поскольку способна обеспечить универсальное участие государств, но при этом не является мировым наднациональным правительством.

Наиболее перспективная роль ООН состоит не в определении допустимости отдельных публикаций, а в пяти взаимосвязанных направлениях.

1. Развитие общих правозащитных ориентиров. На базе МПГПП, Глобального цифрового договора, UN Guiding Principles, документов ЮНЕСКО и иных релевантных международных стандартов. [2, 5, 7, 8]

2. Содействие совместимости национальных подходов. Целью не должна быть полная унификация всех уголовных, административных и гражданско-правовых норм, но государства могут согласовывать минимальные процессуальные стандарты прозрачности, объяснимости, права на жалобу и независимого контроля.

3. Научная оценка. Независимая международная научная группа по ИИ способна предоставлять государствам общую доказательную базу, особенно значимую для стран, которые не располагают собственными масштабными научно-регуляторными возможностями. [13, 14]

4. Постоянный глобальный диалог. Глобальный диалог по управлению ИИ позволяет государствам, бизнесу, гражданскому обществу, академическому и техническому сообществу обсуждать совместимость подходов без превращения этой площадки в глобального регулятора. [15]

5. Развитие потенциала. Глобальные стандарты будут недостаточно легитимны, если их фактически способны

разрабатывать и проверять только наиболее технологически развитые государства и крупнейшие компании. Особым направлением должна стать поддержка многоязычных исследований, локальных экспертных сообществ, независимых регуляторов и исследовательской инфраструктуры в развивающихся странах.

22. Заключение

Вопрос определения границ допустимого контента в цифровой среде не имеет устойчивого решения в форме передачи исключительного контроля одному субъекту.

Концентрация полномочий у государства способна создавать риск политизации и чрезмерного ограничения законной речи.

Полная передача решений платформам оставляет значительную часть глобальной публичной сферы под контролем частных правил, рекомендательных систем и экономических стимулов.

Создание международного органа, ежедневно определяющего допустимость конкретных высказываний для всего мира, в свою очередь, не соответствует нынешней институциональной структуре ООН и многообразию национальных правовых систем.

Наиболее устойчивой является распределённая модель.

Международные стандарты прав человека задают пределы допустимого вмешательства. Государства создают ясные юридические нормы, процедуры и средства правовой защиты. Платформы сохраняют способность оперативно управлять сервисами, но повышают прозрачность и подотчётность. Международные организации обеспечивают

координацию, научную оценку, совместимость подходов и развитие потенциала.

Искусственный интеллект делает такую архитектуру одновременно более необходимой и более сложной, поскольку решения о допустимости и видимости контента могут приниматься в огромном масштабе, автоматически и незаметно для пользователя.

Поэтому легитимность цифрового управления должна определяться не только вопросом «какой контент был ограничен?», но и вопросами «кто принял решение?», «на каком основании?», «почему была выбрана именно эта мера?», «было ли решение объяснено?» и «может ли человек эффективно его оспорить?»

Главный принцип будущей резолюции Генеральной Ассамблеи: безопасность цифровой среды не должна становиться общим оправданием цензуры, а свобода выражения не должна исключать точные, доказательные и соразмерные меры против конкретного вреда.

23. Библиография

Внутритекстовые ссылки приведены в квадратных скобках. Приоритет отдан международным договорам, официальным документам ООН и иных международных организаций, региональным нормативным источникам и первичным рецензируемым исследованиям.

[1] Universal Declaration of Human Rights. United Nations, 1948

<https://www.un.org/en/about-us/universal-declaration-of-human-rights>

[2] International Covenant on Civil and Political Rights. United Nations, 1966; статьи 19–20

<https://2covenants.ohchr.org/About-ICCPR.html>

[3] Human Rights Committee. General Comment No. 34. Article 19: Freedoms of opinion and expression. CCPR/C/GC/34, 2011

- [4] Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence. Office of the United Nations High Commissioner for Human Rights, 2012
- [5] Guiding Principles on Business and Human Rights: Implementing the United Nations “Protect, Respect and Remedy” Framework. United Nations, 2011
<https://europe.ohchr.org/areas-of-work/business-and-human-rights>
- [6] Committee on the Rights of the Child. General Comment No. 25 on children’s rights in relation to the digital environment. CRC/C/GC/25, 2021
- [7] Guidelines for the Governance of Digital Platforms: Safeguarding freedom of expression and access to information through a multistakeholder approach. UNESCO, 2023
<https://www.unesco.org/en/articles/guidelines-governance-digital-platforms>
- [8] Pact for the Future. Annex I: Global Digital Compact. A/RES/79/1, United Nations General Assembly, 2024
<https://www.un.org/pact-for-the-future/en/annex-i-global-digital-compact>
- [9] United Nations Global Principles for Information Integrity. United Nations, 2024
<https://www.un.org/en/node/218792>
- [10] Digital Services Act — Regulation (EU) 2022/2065. European Union, 2022
- [11] Two years of Digital Services Act allows 50 million content moderation decisions by platforms to be reversed. European Commission, 17 February 2026
<https://digital-strategy.ec.europa.eu/en/news/two-years-digital-services-act-allows-50-million-content-moderation-decisions-platforms-be-reversed>
- [12] The Digital Services Act brings transparency to content moderation. European Commission
<https://digital-strategy.ec.europa.eu/en/policies/dsa-brings-transparency>
- [13] Terms of reference and modalities for the establishment and functioning of the Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance. A/RES/79/325, United Nations General Assembly, 2025
- [14] Independent International Scientific Panel on AI — FAQ and Preliminary Report. United Nations, 2026
<https://www.un.org/independent-international-scientific-panel-ai/en/faq>
- [15] Global Dialogue on Artificial Intelligence Governance — FAQ, documents and outcomes of the First Annual Dialogue. United Nations, 2026
<https://www.un.org/global-dialogue-ai-governance/en/faq>
- [16] Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, CETS No. 225. Council of Europe, opened for signature 5 September 2024; status checked 30 August 2026
<https://www.coe.int/en/web/conventions/full-list?module=signatures-by-treaty&treatynum=225>

- [17] Fasching, N.; Lelkes, Y. Comparative analysis of content moderation systems. Findings of ACL 2025
<https://aclanthology.org/2025.findings-acl.1144/>
- [18] AfriHate: A Multilingual Collection of Hate Speech and Abusive Language Datasets for African Languages. NAACL 2025
<https://aclanthology.org/2025.naacl-long.92/>
- [19] How decisions are made at the United Nations. United Nations Model United Nations guidance
<https://www.un.org/en/model-united-nations/how-decisions-are-made-un>