Рекомендательные технологии как социотехнический инструмент персонализации информационных потоков и деструктивно-манипулятивного воздействия: вызовы правовому обеспечению информационной безопасности

Никишин Владимир Дмитриевич
кандидат юридических наук, директор Института информационной и медиабезопасности, доцент кафедры информационного права и цифровых технологий Университета имени О.Е. Кутафина (МГЮА), VDNIKISHIN@msal.ru

В российском законодательстве рекомендательные технологии (далее также РТ) получили минимальное нормирование, которое носит во многом декларативный характер, не позволяет обеспечить прозрачность применения РТ и реализацию интернет-пользователями своей свободы воли. В условиях алгоритмически опосредованной коммуникации вред формируется не только и не столько за счет единичных информационных единиц, сколько вследствие системной конфигурации среды. В фокусе анализа находятся конфигурации алгоритмической среды, при которых возникает предсказуемый риск устойчивого вреда, т.е. критерием деструктивности становится не только оценка содержания конкретных публикаций, но и характеристика процедуры как риск-генерирующей. Обосновано, что описанные в работе риски применения РТ противоречат принципу идеологического многообразия и создают предпосылки для нарушения когнитивного иммунитета личности. В работе выделены четыре контура алгоритмического воздействия: спецификация цели; представление данных; оптимизация; архитектура выбора. Синергия этих контуров образует триаду механизмов воздействия: селекция; интенсификация; нормализация. Для аналитики информационной безопасности полезно различать два модуса деструкции: пассивно-кумулятивный и активно-целевой модус. С точки зрения построения модели правового регулирования применения РТ проанализированы аспекты обеспечения процедурной автономии, прозрачности спецификации цели и подотчетности за предсказуемый риск. Особое внимание уделено разграничению посредничества и редакционной ответственности в условиях алгоритмического ранжирования. Типологию вреда, причиняемого РТ, предложено классифицировать по двум критериям: модальности и темпоральности.

Ключевые слова: рекомендательные технологии, рекомендательные системы, ректех, информационная безопасность, семантические капсулы, коллаборативная фильтрация, алгоритмы, ИИ, нейросети, автономия воли, таргетирование, информационное право, цифровое право, средства массовой коммуникации.

Исследование выполнено в рамках программы стратегического академического лидерства «Приоритет-2030».

Рекомендательные технологии можно определить как программный комплекс алгоритмических методов и моделей обработки данных, предназначенный для прогнозирования предпочтений или степени интереса пользователя к определенным элементам (контенту, товарам, услугам и пр.) с их последующей ранжированной выдачей с целью оптимизации взаимодействия пользователя с платформой.

В российском законодательстве рекомендательные технологии (далее также РТ) получили минимальное нормирование после принятия Федерального закона от 31.07.2023 № 408-ФЗ «О внесении изменения в Федеральный закон «Об информации, информационных технологиях и о защите информации»», закрепившего обязанности владельцев информационных ресурсов, применяющих рекомендательные технологии, по информированию пользователей и опубликованию правил их применения, а также предусмотревшего надзорные полномочия по требованию прекращения использования рекомендаций при нарушении закона. Дальнейшая подзаконная детализация реализована приказами Роскомнадзора: приказ № 149 от 06.10.2023 г. установил минимальные требования к содержанию и размещению информации о применении рекомендаций (маркировка и ссылка на отдельную страницу с правилами), приказ № 150 от 06.10.2023 г. – порядок взаимодействия надзорного органа с владельцами ресурсов при выявлении нарушений.

Указанные нормативные правовые акты выступают лишь первым шагом в регулировании применения РТ и защите пользователей от их деструктивного воздействия. К сожалению, текущее регулирование носит во многом декларативный характер и не отвечает реальным вызовам информационной безопасности: требование информирования пользователей о применении РТ сводится к обязанности разместить фраза «На информационном ресурсе при применении информационных технологий предоставления информации осуществляется сбор, систематизация и анализ сведений, относящихся к предпочтениям пользователей сети «Интернет», находящихся на территории Российской Федерации»; а требования к содержанию документа, устанавливающего правила применения РТ, не конкретизированы на уровне подзаконного регулирования и фактически «отданы на откуп» самим операторам РТ.

Такая регуляторика не позволяет обеспечить прозрачность применения РТ и реализацию интернет-пользователем своей свободы воли ввиду динамичной природы рекомендательных технологий как социотехнического инструмента персонализации информационных потоков и деструктивно-манипулятивного воздействия, основные аспекты которого раскрыты далее в настоящей статье.

Деструктивное воздействие рекомендательных технологий ошибочно рассматривать исключительно в аспекте содержания отдельных сообщений и наличия «вредного» контента как такового. В условиях алгоритмически опосредованной коммуникации вред формируется не только и не столько за счет единичных информационных единиц, сколько вследствие системной конфигурации среды: способов отбора, ранжирования, усиления и повторяемости контента, а также особенностей архитектуры выбора. Тем самым деструктивность приобретает процедурный и кумулятивный характер, что требует теоретического анализа механизмов, а не только эмпирической фиксации негативных эффектов.

При таком подходе центральным объектом анализа становится не «вредность» конкретного информационного материала, а контур алгоритмического усиления, который задает вероятность встречи пользователя с тем или иным материалом, частоту повторных экспозиций и направление «следующего шага» в цепочке рекомендаций. Именно этот контур делает воздействие системным: он превращает отдельные информационные единицы в последовательность, обладающую собственной динамикой и внутренними закономерностями. Поэтому для целей информационной безопасности важно различать два уровня: уровень содержания (что именно сообщается) и уровень процедуры (как, в каком порядке и с какой интенсивностью это предъявляется). Во многих ситуациях юридически и социально значимым оказывается именно второй уровень, поскольку он формирует предсказуемые траектории внимания и поведения.

Важным следствием является и то, что в цифровой медиасреде «нейтральность» показа не может предполагаться по умолчанию. Любая система ранжирования неизбежно выражает набор «предпочтений», заложенных в целевую функцию и в правила отбора сигналов. Даже если эти «предпочтения» не связаны с идеологическими установками, они, как минимум, отражают приоритеты оптимизации: удержание внимания, рост взаимодействий, повышение конверсии. В результате различие между «простым доступом к информации» и «организованным распределением внимания» становится принципиальным: второе фактически формирует медийную реальность пользователя, а значит, нуждается в теоретическом описании и в нормативной оценке.

Под деструктивным воздействием рекомендательных технологий понимается их негативное влияние на психическое состояние, поведение и убеждения людей, приводящее к ущербу для личности или общества. Такое воздействие может выражаться в причинении психологического вреда (например, рост тревожности или депрессии от постоянного потребления токсичного контента), нарушении прав и свобод личности (манипуляция мнением, нарушение приватности частной жизни), подрыве социальных связей и институтов (поляризация социальных групп, дезинформация и дискредитация). Критериями установления деструктивного влияния служат: а) вред личности – психологический (стресс, зависимость, искажение самооценки), а иногда и физический (побуждение к опасным действиям); б) вред обществу – разжигание вражды, разрушение доверия, угроза общественной безопасности; в) вред институциональным процессам – манипуляция электоральным поведением, угрозы национальной безопасности. Проще говоря, деструктивным считается воздействие рекомендательных технологий, которое подрывает информационную и медиабезопасность, нарушает нормальное функционирование личности или социума.

Деструктивность в контексте РТ не тождественна ни «неприятности» информации, ни ее политической спорности, ни субъективному несогласию с позицией автора. В фокусе находятся такие конфигурации алгоритмической среды, при которых возникает предсказуемый риск устойчивого вреда – вследствие повторяемости показов, усиления аффективно нагруженных стимулов, безальтернативности маршрутов потребления и специфической уязвимости аудитории. Иначе говоря, критерием становится не оценка содержания как «плохого», а характеристика процедуры как риск-генерирующей.

Отсюда вытекают два дополнительных признака, позволяющих более строго идентифицировать алгоритмически обусловленную деструкцию. Во-первых, признак воспроизводимости: эффект должен проявляться не как единичная случайность, а как повторяемый паттерн, выявляемый на массиве пользователей или в типичных сценариях потребления. Во-вторых, признак управляемости: если риск связан с параметрами выдачи (силой персонализации, автозапуском, глубиной «цепочек», скрытыми настройками), он принципиально поддается снижению посредством процедурных мер. Эти признаки важны и для права: они позволяют переводить обсуждение в плоскость обязанностей по разумным мерам профилактики, а не в плоскость спорных моральных оценок.

Дальнейшая типология деструктивных эффектов требует систематизации через наиболее характерные классы контента и практик, однако следует подчеркнуть: речь идет не о закрытом перечне «запрещенных» тем, а о тех областях, где алгоритмическое усиление чаще всего проявляет себя как самостоятельный фактор риска. Общей для них является одна и та же логика: система ранжирования повышает видимость материала, ориентируясь на сигналы вовлечения, затем закрепляет результат повторением и формирует «коридор» дальнейших рекомендаций, постепенно сужающий выбор. Поэтому ниже перечисляются не просто социально нежелательные явления, а те зоны, в которых связь между механизмом рекомендаций и вредом оказывается наиболее наглядной и теоретически описываемой.

Рекомендательные алгоритмы способны усиливать распространение различных типов деструктивного контента.

Экстремизм и радикализация

Контент террористического или экстремистского характера, пропаганда идеологий ненависти и насилия. Алгоритмы могут рекомендовать все более радикальные материалы пользователям с интересом к подобным темам, «вовлекая» их в экстремистские сообщества.

Дезинформация и фейки

Ложные новости, конспирологические теории, пропаганда, вводящая в заблуждение – все это быстро распространяется благодаря рекомендательным технологиям, ценящим привлекательность контента выше достоверности. Вирусные фейки алгоритмам продвигать легче, чем сложные проверенные новости.

Пропаганда девиантного поведения

Алгоритмы могут рекомендовать уязвимым пользователям контент, провоцирующий саморазрушительные или противоправные действия. Например, распространены случаи, когда подросткам, ищущим безобидные советы по диете, сервисы начинали рекомендовать сообщества анорексии и самоистязания. Исследования показали, что Instagram «затягивал» молодых девушек в поток материалов о нездоровой худобе («skinny binge») и этим ухудшал их восприятие собственного тела.

Формирование зависимостей

Рекомендательные технологии нередко создают эффект затягивания, побуждая пользователей все дольше оставаться онлайн. Это может приводить к цифровой зависимости, особенно у подростков. Персонализированная подача контента повышает риск, что человек втянется в такую зависимость. Так, например, «пузырь фильтров» концептуализируется как следствие гиперперсонализации контента, где алгоритмы, минимизируя когнитивный диссонанс, формируют информационный поток, который больше всего удовлетворяет пользователя платформы.

Манипуляция

Использование скрытых психологических приемов для влияния на мнения и решения людей – еще одна грань деструктивного воздействия на личность пользователя. РТ могут эксплуатировать когнитивные искажения пользователя, подкрепляя существующие убеждения и «слепые зоны». Опасна и целевая манипуляция уязвимых групп: с помощью микротаргетинга алгоритмы позволяют определять пользователей, находящихся в определенном эмоциональном состоянии или социальном положении и направлять на них специально таргетированную пропаганду. Например, утечка документов Facebook показала, что платформа имела техническую возможность выявлять подростков, чувствующих себя «неуверенными» или «ничтожными», и рекламодатели теоретически могли таргетировать на них соответствующие сообщения.

В правовой плоскости это противоречит принципу идеологического многообразия и создает предпосылки для нарушения когнитивного иммунитета личности. Эмпирические исследования при этом подтверждают корреляцию между алгоритмической сегрегацией контента и снижением толерантности к альтернативным точкам зрения.

Переходя к психологическим механизмам, целесообразно оговорить границы интерпретации. В гуманитарно-правовом анализе недостаточно ссылаться на отдельные нейрофизиологические корреляции как на «доказательство» причинности; существенно важнее показать устойчивую психологическую закономерность: эмоционально интенсивные стимулы чаще захватывают внимание и легче закрепляются повторением, тогда как критическая оценка источников и контекста требует ресурсов когнитивного контроля и времени. Рекомендательные системы, оптимизированные под удержание и частоту взаимодействий, структурно смещают выдачу в сторону тех стимулов, которые быстрее вызывают реакцию и сильнее «подкрепляют» поведение. Следовательно, речь идет не о медицинской модели воздействия, а о социально и юридически значимой асимметрии: условия потребления информации проектируются так, что эмоциональная реакция оказывается функционально выгоднее рациональной проверки.

Алгоритмическое усиление, лежащее в основе современных рекомендательных технологий, представляет собой не просто технологический инструмент для персонализации контента в цифровой среде, но и мощный механизм формирования поведения, основанный на принципах, нетождественных принципам информационной безопасности общества. В этой парадигме пользовательская вовлеченность становится ключевой метрикой и «валютой», что закономерно приводит к доминированию контента с высокой эмоциональной валентностью – прежде всего, вызывающего гнев, возмущение, страх или негодование. Нейрофизиологические исследования подтверждают, что подобный контент инициирует мгновенную активацию миндалевидного тела (амигдалы) и других структур лимбической системы, ответственной за эмоциональные реакции, при этом подавляя активность префронтальной коры головного мозга, которая обеспечивает сложные когнитивные процессы, включая критическое осмысление, рациональную оценку и принятие взвешенных решений. С теоретической точки зрения данная технология эксплуатации биологически обусловленных уязвимостей человека интерпретируется как форма скрытой и масштабированной манипуляции, что находится в системном противоречии с задачами обеспечения информационной безопасности личности. Пользователь, лишенный подлинной автономии воли и понимания алгоритмических механизмов, воздействующих на его подсознание, не может считаться свободно и полноценно давшим согласие на обработку своих данных в таких условиях, что ставит под сомнение легитимность всей полученной таким образом поведенческой биографии.

Эксплуатация когнитивных искажений в архитектуре цифровых платформ является прямой операционализацией теории «подталкивания» (nudge theory), разработанной Р. Талером и К.Санстейном. Если в их классической концепции архитектура выбора мягко направляет индивида к социально-полезному поведению (например, к выбору здорового питания или согласию на донорство органов по умолчанию), то в коммерчески ориентированных РТ она трансформируется в инструмент скрытого давления, нацеленного на максимизацию времени пребывания в сервисе и объемов собираемых данных. Это реализуется через целый арсенал интерфейсных решений: бесконечная лента, автовоспроизведение следующего видеоролика, навязчивые уведомления, – которые создают эффект «потерянного времени» и ослабляют пользовательский самоконтроль, целенаправленно эксплуатируя склонность к немедленному вознаграждению дофаминовой системы.

Правовой вызов здесь заключается в отсутствии адекватного нормативного регулирования этих систем, которые по своей сути являются манипулятивными, но не подпадают под традиционные запреты.

Для корректной постановки правовой задачи необходимо уточнить объект регулирования. В случае рекомендательных технологий объектом выступает не только обработка персональных данных и не только содержание распространяемой информации, но прежде всего процедуры алгоритмического воздействия: цели оптимизации, параметры ранжирования, правила усиления и повторяемости, интерфейсные решения, управляющие «трением» выбора, а также режимы работы с уязвимыми группами. Иначе говоря, регулирование должно быть обращено к тому, что можно проверить и воспроизвести: к документируемым решениям и практикам владельца РТ, а не к гипотетическим намерениям или моральным оценкам.

В этой логике наиболее продуктивной становится модель процедурных стандартов: прозрачность ключевых принципов рекомендаций, реальная управляемость персонализации, минимальные требования к работе с жалобами и к снижению предсказуемых рисков, а также подотчетность за повторяемость вредных сценариев. Такой подход позволяет избежать крайностей – либо нормативной «слепоты», когда право фиксирует общие цели без механизма их достижения, либо чрезмерно жестких мер, которые неизбежно вступают в конфликт со свободой доступа к информации и свободой выражения мнений. Процедурные стандарты, напротив, нацелены не на запрет идей, а на изменение условий их распределения и усиления – то есть на саму архитектуру медиапространства.

Наиболее социально опасным проявлением деструктивного воздействия РТ является таргетирование уязвимых групп, которое находит свое теоретическое осмысление в концепции «алгоритмической дискриминации». Как убедительно демонстрируют С. Барокас и А. Зельбст, системы, обученные на ретроспективных данных, не только отражают существующие в обществе предрассудки и структурные неравенства, но и закрепляют, а также масштабируют их. Это происходит двумя основными путями. Во-первых, через использование искаженных датасетов, где субъективная предвзятость «окаменевает» в алгоритме и начинает воспроизводиться в будущем. Во-вторых, через моделирование (анализ прошлых и текущих взаимодействий), а также активную эксплуатацию поведенческих уязвимостей определенных групп: подростков с неустойчивой психикой, лиц, склонных к аддикциям, людей в состоянии финансовой или личностной неустроенности. Алгоритмы выявляют таких пользователей по их поведенческой истории и целенаправленно предлагают им чувствительный контент: пропаганду расстройств пищевого поведения, экстремистские материалы, азартные игры или финансовые схемы «быстрого обогащения». Ярким примером является скандал вокруг социальной сети Facebook, где рекламные платформы позволяли таргетировать объявления на пользователей с интересом к «антисемитским конспирологическим теориям» или с самооценкой «неуверенный».

Следует отметить, что деструктивный потенциал рекомендательных технологий проявляется на множестве уровней: от масштабирования индивидуальных психологических проблем (тревожность, депрессия, цифровая зависимость) до генерации массовых когнитивных искажений, искажений общественного мнения, эрозии социального доверия и подрыва политических институтов через распространение дезинформации и усиление политической поляризации. Механизмом этого воздействия является не просто пассивное отражение предпочтений пользователя, а активное формирование и усиление этих предпочтений через создание «фильтрующего пузыря» и применение инструментов увеличения вовлеченности. Алгоритмы не просто показывают то, что пользователь хочет увидеть; они постепенно и системно смещают его предпочтения в сторону более чувствительного, эмоционально заряженного и, как следствие, более вовлекающего контента, создавая петлю обратной связи, из которой практически невозможно вырваться самостоятельно.

Этот феномен «радикализации через рекомендации» был замечен на примере платформы YouTube, где путь от относительно нейтральных видео к конспирологическим и экстремистским часто прокладывается именно последовательностью рекомендаций алгоритма.

Для противодействия этим системным угрозам требуется комплексный межотраслевой подход, выходящий за рамки традиционного правового регулирования защиты персональных данных. Необходима разработка и имплементация специализированного законодательства, прямо нацеленного на обеспечение прозрачности, подотчетности и безопасности рекомендательных технологий.

Представляется, что ключевым элементом будущего регулирования должно стать внедрение принципа «предустановленного выбора», когда платформы обязаны предлагать пользователям не манипулятивный, а просвещающий и развивающий интерфейс, включая алгоритмические рекомендации, основанные на достоверности информации и ее разнообразии, а не исключительно на метриках вовлеченности. Полагаем, подобные правила целесообразно закрепить и на законодательном уровне.

При этом, следует отметить, что информационная безопасность в контексте рекомендательных технологий выступает как свойство информационной среды сохранять когнитивную и эмоциональную устойчивость индивида и институциональную устойчивость социума при сохранении свободы доступа к информации. В этой рамке «деструктивность», как уже отмечалось, не тождественна неприятности или спорности контента: релевантным является не «ценностный ярлык», а механизм причинения вреда и его рисковый профиль. Последний определяется (1) интенсивностью воздействия (частотой и длительностью экспозиции), (2) персистентностью (устойчивостью эффекта во времени), (3) безальтернативностью (отсутствием реалистичных опций выхода) и (4) уязвимостью аудитории (возраст, психическое состояние, кризисные жизненные обстоятельства и т.п.). Эти четыре параметра позволяют отличать просто «неудобную» информацию от такого воздействия, которое подрывает информационную безопасность.

С теоретической точки зрения деструктивное влияние РТ следует объяснять не свойствами «опасных» сообщений самих по себе, а совокупностью четырех контуров алгоритмического воздействия:

1. Спецификация цели. Большинство промышленных РТ оптимизируют метрики (время удержания, клики, глубина сессии и т.п.), которые коррелируют с экономическими целями, но не являются показателями благополучия пользователя. Неполнота и асимметрия целевой функции порождают систематическую предвзятость: модели «научаются» предпочитать контент с высокой аффективной валентностью и низкой проверяемостью, поскольку такой контент усиливает вовлечение.

2. Представление данных. Обучение на прошлых взаимодействиях с платформой содержит структурные искажения: эффект самоселекции (отвержение невписывающееся в картину мира информации), обратная связь популярности (популярное становится еще популярнее), дискриминация меньшинств за счет редких появлений данных о них. Эти феномены переводят социальные предубеждения и случайные всплески внимания в устойчивые параметры модели.

3. Оптимизация. В реальном времени РТ решают задачу компромисса «исследование–эксплуатация», усиливая краткосрочно выгодные маршруты внимания. Возникают замкнутые петли обратной связи: предпочтение → показ → вовлечение → обновление модели → усиленный показ. При наличии аффективно насыщенного контента такая петля смещает «базовую линию» восприятия и формирует тренд к эскалации, чтобы поддерживать прежний уровень возбуждения, системе приходится предлагать более острый материал.

4. Архитектура выбора. Интерфейсные решения (автовоспроизведение, бесконечная прокрутка, подталкивающие уведомления, скрытие опций отказа) уменьшают «трение» для поведения, выгодного целевой функции, и увеличивают «трение» для поведения, поддерживающего автономию (поиск альтернатив, выключение персонализации). В результате внешняя легитимация выбора сохраняется, тогда как фактическая свобода уменьшается.

Синергия этих контуров образует триаду механизмов воздействия: (а) селекция – системное завышение вероятности появления в ленте аффективно нагруженной и поляризующей информации; (б) интенсификация – наращивание частоты и плотности однотипных нарративов, приводящее к сдвигу ориентиров пользователя и формированию устойчивых паттернов; (в) нормализация – закладывание новых критериев «обычного» через повторяемость и кажущуюся общепринятость, что снижает пороги критического восприятия. Именно триада «селекция–интенсификация–нормализация», а не абстрактная «вредность» отдельных сообщений, объясняет переход от индивидуальной уязвимости к масштабным социальным эффектам.

Для аналитики информационной безопасности полезно различать два модуса деструкции. *Пассивно-кумулятивный модус* возникает, когда агрегирующее воздействие множества «безобидных» на первый взгляд единиц контента вызывает перегрузку, выученную беспомощность, ухудшение настроения, навязчивые паттерны потребления. *Активно-целевой модус* реализуется там, где система, идентифицируя маркеры уязвимости, систематически подбирает специфические стимулы (например, усиленно подает материалы, катализирующие самоповреждающее поведение). Оба модуса опираются на одну и ту же инфраструктуру предсказаний, но различаются по интенсивности и направленности «оптимизации» контента.

С позиций когнитивной психологии ядром деструктивности является конфликт между аффективной активацией и когнитивным контролем. Аффективно заряженный контент усиливает немедленные реакции ориентировочной и лимбической систем, тогда как ресурсы префронтальной регуляции – ограничены и истощаются под длительной нагрузкой. РТ, максимизируя краткосрочное вовлечение, «прикрепляют» выбор к аффективному контуру, создавая поведенческую инерцию и снижая вероятность обращения к критической оценке источников и аргументов. Это не тождественно «запрету эмоций» в публичной сфере; речь идет о системной асимметрии, когда РТ поддерживаются именно те эмоциональные траектории, которые оптимальны для удержания, но субоптимальны для способности пользователя выдерживать сложные, противоречивые сведения.

Правовая перспектива требует перевести описанные механизмы в язык критериев и обязательств. Во-первых, процедурная автономия. Согласие пользователя на обработку данных и применение персонализации не может считаться осознанным, если архитектура выбора скрывает реальные альтернативы, а описание воздействия ограничено общими формулами. Минимальным стандартом является явность и доступность отказа, информирование о ключевых факторах влияния и о последствиях выбора режимов ленты. Во-вторых, прозрачная спецификация цели. Оператор РТ обязан документировать, какие метрики оптимизируются и каким образом предотвращается их конфликт с охраняемыми ценностями (психическое здоровье, защита детей, недискриминация). В-третьих, подотчетность за предсказуемый риск. При наличии данных о том, что определенные контуры рекомендаций ведут к устойчивому вреду у уязвимых групп, бремя доказательства адекватности мер снижения риска не может перекладываться на пользователя.

Доктрина эпистемической безопасности служит мостом между психологией и правом. Ее операционализация в рамках РТ предполагает, как минимум, три измерения: (1) диверсификация источников, (2) контекстуализация достоверности (пометки о статусе источника, наличии проверки, уровне неопределенности), (3) контакт с несогласными точками зрения (контролируемое включение корректных контраргументов, особенно в чувствительных темах). Важно подчеркнуть: эти меры

не конкурируют со свободой выражения; они меняют процедуру доставки и весовых коэффициентов внутри алгоритма, не навешивая содержательных цензурных запретов на сами идеи.

Для целей надзора и внутреннего комплаенса целесообразно ввести критерии «пороговой деструктивности». Такая конструкция отделяет неизбежный «коммуникационный риск» от предсказуемо усиленного алгоритмом риска, подлежащего управлению.

Особого внимания заслуживает разграничение посредничества и редакционной ответственности в условиях алгоритмического ранжирования. В классической модели нейтрального посредника платформа не влияет на вероятность встречи пользователя с определенной информацией за пределами простых форматов индексации. В модели РТ платформа формирует индивидуализированный порядок и интенсивность показа, опираясь на скрытую оценочную функцию. С правовой точки зрения это означает, что нейтральность ограничена: по мере увеличения роли оптимизации растет и объем ожидаемой должной осмотрительности в отношении предсказуемых вредных эффектов для уязвимых групп. Речь не о всеобщем переносе ответственности за речь третьих лиц на оператора, а о специфической ответственности за конфигурацию и эксплуатацию механизма усиления.

Обеспечение информационной безопасности также требует учета контекста применения. Один и тот же алгоритмический контур может быть допустим в развлекательном сервисе для совершеннолетних и недопустим в образовательной среде или детских сервисах. Следовательно, *принцип пропорциональности* в этой сфере должен применяться как тест из четырех шагов: легитимность цели (охраняемая ценность), пригодность меры (способна ли мера снижать конкретный риск), необходимость (есть ли менее ограничительные альтернативы для сопоставимого эффекта), соразмерность в узком смысле (баланс выгод и потерь для затронутых прав). Перевод этого теста в технологическую процедуру (описание альтернативных конфигураций, их прогнозной оценки и причин отказа от более мягких) обеспечивает проверяемость решений оператора РТ.

Наконец, отметим ограничения эпистемологии в оценке деструктивности. Корреляционные наблюдения в социальных сетях часто искажены неучтенной гетерогенностью пользователей и событийными шоками. Отсюда – риск гипердиагностики (объявление негативной тенденцией любых непонравившихся паттернов) и риск гиподиагностики (игнорирование кумулятивного слабого сигнала). Методологическая добросовестность предполагает явное разграничение уровней уверенности: гипотеза риска (обоснованное предположение), сигнал риска (стабильный статистический паттерн), подтвержденный риск (реплицируемый эффект). Только на последнем уровне оправданы строгие организационные вмешательства, тогда как на первых двух – превентивные и мягкие корректировки (увеличение разнообразия, добавление пояснений, ослабление автоплей-контуров).

Дальнейшее теоретико-правовое уточнение требует развести собственно «опасный контент» и «опасную конфигурацию среды», в которой даже формально не противоправные данные при определенной частоте, длительности и отсутствии альтернативных маршрутов восприятия приобретают вредоносный профиль. В этой связи деструктивность предстает как свойство не единичного сообщения, а институционально заданной процедуры его доставки и усиления, что переводит анализ с уровня содержания на уровень архитектуры медиапространства. Именно поэтому информационную безопасность следует понимать как публичное благо: отдельный пользователь не способен самостоятельно модифицировать условия, в которых формируется его информационный «рацион», а потому нуждается в институциональных гарантиях предсказуемости и плюрализма, минимизирующих кумулятивный вред.

Соответственно, ключевым гуманитарно-правовым ориентиром выступает когнитивный иммунитет личности – способность сохранять автономию суждения в среде, где выбор опосредован алгоритмическими фильтрами. В отличие от приватности, охраняющей «периметр данных», когнитивный иммунитет охраняет условия формирования воли и убеждений. Юридически он конкретизируется через требования к процедурной автономии пользователя: информированность о характере персонализации, реальная альтернативность режимов потребления, обратимость профилирования. Эти требования не навязывают «правильного» содержания контента, но задают пределы допустимого воздействия: там, где алгоритмическое усиление предсказуемо снижает способность к самостоятельному выбору и критической оценке, возникает зона, в которой вмешательство оправдано тестом необходимости и соразмерности.

Для аналитики и последующего правоприменения важно развернуть типологию вреда, причиняемого РТ, по двум осям: модальности и темпоральности. По *модальности* различим *прямой вред* (индуцирование противоправного или саморазрушительного поведения), *косвенный вред* (ухудшение психического благополучия, выученная беспомощность, цинизм и т.п.), *диффузный социальный вред* (эрозия доверия, поляризация, нормализация стигматизации и т.п.) и *институциональный вред* (подрыв процедур публичной коммуникации, снижение качества эпистемических практик и т.п.). По *темпоральности* отличим *острые эпизоды* (скачкообразные всплески агрессии или паники) и *хронические эффекты* (формирование привычек потребления, устойчивые когнитивные смещения), причем именно хронические эффекты чаще связаны с конфигурацией алгоритмической среды, а не с единичными сообщениями. Такая рамка позволяет корректно ставить вопрос о причинной релевантности: объектом юридической оценки становится не «пост», а «контур доставки» контента, воспроизводящийся во времени.

Сказанное позволяет сделать методологически важный вывод: в оценке рекомендательных технологий для обеспечения информационной безопасности первичны не моральные ярлыки и не спорность, потенциальная противоправность отдельных сообщений, а параметры риск-профиля алгоритмической процедуры. В дальнейшем это позволяет перевести теоретический анализ в конструктивную плоскость: какие именно элементы системы подлежат раскрытию, какие настройки должны быть доступны пользователю, какие меры профилактики являются разумными, а какие – формальными, и какие артефакты (документы, регламенты, журналы изменений, статистика жалоб) должны обеспечивать проверяемость утверждений владельца РТ. Тем самым теория деструктивного воздействия становится не самостоятельным описанием «опасностей цифровой среды», а основанием для проектирования нормативного и институционального контура медиабезопасности.

Таким образом, теоретические основы деструктивного воздействия РТ в логике информационной безопасности сводятся к следующему. Во-первых, вред возникает не из «плохого контента» как такового, а из несоответствия цели оптимизации охраняемым ценностям и из институционально закрепленной асимметрии между аффективной активацией и когнитивным контролем. Во-вторых, решающую роль играют замкнутые петли обратной связи, которые переводят случайные всплески вовлечения в долговременные сдвиги предпочтений и норм. В-третьих, юридическая релевантность придается тем ситуациям, где риск предсказуем, уязвимость идентифицируема, а альтернативы для пользователя практически отсутствуют. В-четвертых, корректный ответ требует не цензуры содержания, а переопределения процедур: ясного выбора режимов, прозрачной спецификации цели, подотчетности за изменение моделей и верифицируемых методов оценки риска. Такая постановка вопроса позволяет удержать баланс между свободой коммуникации и охраной когнитивного суверенитета, переводя дискуссию из сферы абстрактных опасений в сферу проверяемых инженерно-правовых решений.

Литература

1. Ricci F., Rokach L., Shapira B. Recommender Systems Handbook. Springer, 2015. 1020 p.
2. Bhardwaj A., Bharany S., Kim S Fake social media news and distorted campaign detection framework using sentiment analysis & machine learning // Heliyon. 2024. Vol. 10. Iss. 16. No. e36049.
3. Sunstein C.R. Republic: Divided Democracy in the Age of Social Media. Princeton: Princeton University Press, 2017. 328 p.
4. Heller A.S., Lapate R.C., Mayer K.E., Davidson R.J The face of negative affect: trial-by-trial corrugator responses to negative pictures are positively associated with amygdala and negatively associated with ventromedial prefrontal cortex activity // Journal of cognitive neuroscience. 2014. Vol. 26. No. 9. Pp. 2102–2110.
5. Zuboff S. The Age of Surveillance Capitalism. New York: PublicAffairs, 2019. Pp. 132-140.
6. Thaler R.H., Sunstein C.R. Nudge: improving decisions about health, wealth, and happiness. Yale University Press, 2008. 312 p.
7. Barocas S., Selbst A.D. Big Data's Disparate Impact // California Law Review. 2016. Vol. 104. Pp. 671-732.

Recommendation technologies as a sociotechnical tool for personalizing information flows and for destructive and manipulative influence: challenges to legal support for information security

Nikishin V.D.

Kutafin Moscow State Law University (MSAL)

In Russian legislation, recommendation technologies (hereinafter referred to as RT) have received minimal regulation, which is largely declarative in nature and does not ensure transparency in the use of RT or the exercise of freedom of choice by internet users. In the context of algorithmically mediated communication, harm is not only and not so much caused by individual information units, but rather by the systemic configuration of the environment. The focus of analysis is on the configurations of the algorithmic environment that pose a predictable risk of sustained harm, where the criterion of destructiveness is not only the assessment of the content of specific publications, but also the characterization of the procedure as risk-generating. It is substantiated that the risks of using RT described in the work contradict the principle of ideological diversity and create prerequisites for violating an individual's cognitive immunity. The paper identifies four algorithmic impact loops: target specification, data representation, optimization, and selection architecture. The synergy of these loops forms a triad of impact mechanisms: selection, intensification, and normalization. For information security analytics, it is useful to distinguish between two modes of destruction: passive-cumulative and active-targeted. From the perspective of building a legal framework for the use of RT, the paper analyzes the aspects of ensuring procedural autonomy, transparency in target specification, and accountability for predictable risks. Special attention is paid to the delineation of mediation and editorial responsibility in the context of algorithmic ranking. The typology of harm caused by RT is proposed to be classified according to two criteria: modality and temporality.

Keywords: recommendation technologies, recommendation systems, rektech, information security, semantic capsules, collaborative filtering, algorithms, AI, neural networks, autonomy of will, targeting, information law, digital law, and mass media.

Affilations. The study was carried out as part of the Priority-2030 strategic academic leadership program.

References

1. Ricci F., Rokach L., Shapira B. Recommender Systems Handbook. Springer, 2015. 1020 p.
2. Bhardwaj A., Bharany S., Kim S Fake social media news and distorted campaign detection framework using sentiment analysis & machine learning // Heliyon. 2024. Vol. 10. Iss. 16. No. e36049.
3. Sunstein C.R. Republic: Divided Democracy in the Age of Social Media. Princeton: Princeton University Press, 2017. 328 p.
4. Heller A.S., Lapate R.C., Mayer K.E., Davidson R.J The face of negative affect: trial-by-trial corrugator responses to negative pictures are positively associated with amygdala and negatively associated with ventromedial prefrontal cortex activity // Journal of cognitive neuroscience. 2014. Vol. 26. No. 9. Pp. 2102–2110.
5. Zuboff S. The Age of Surveillance Capitalism. New York: PublicAffairs, 2019. Pp. 132-140.
6. Thaler R.H., Sunstein C.R. Nudge: improving decisions about health, wealth, and happiness. Yale University Press, 2008. 312 p.
7. Barocas S., Selbst A.D. Big Data's Disparate Impact // California Law Review. 2016. Vol. 104. Pp. 671-732.