Логинова Наталья Анатольевна
д-р экон. наук, доцент, профессор
Головинский Максим Андреевич
адъюнкт
ФГКОУ ВО «Санкт-Петербургский университет МВД России»
г. Санкт-Петербург

ЭТИЧЕСКИЕ ВЫЗОВЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ЦИФРОВОМ ОБЩЕСТВЕ: АЛГОРИТМИЧЕСКАЯ СПРАВЕДЛИВОСТЬ И ЧЕЛОВЕЧЕСКОЕ ДОСТОИНСТВО

Аннотация: *. В статье исследуется фундаментальное противоречие, порождаемое повсеместным внедрением систем искусственного интеллекта в социально значимые сферы: между стремлением к эффективности и оптимизации, с одной стороны, и соблюдением принципов справедливости, недискриминации и уважения человеческого достоинства – с другой. На основе анализа российской и международной практики за 2025–2026 гг. вскрываются механизмы воспроизводства алгоритмической предвзятости, анализируется феномен «моральной маскировки» в работе генеративных моделей, рассматриваются последствия дерегулирования сферы искусственного интеллекта для фундаментальных прав. Особое внимание уделяется российским законодательным инициативам, включая законопроект о регулировании искусственного интеллекта и Кодекс этики для финансового сектора. В заключение предлагается переориентация этического дискурса от «ответственности за последствия» к «заботе о достоинстве» как первоочередному принципу управления искусственным интеллектом в цифровом обществе.*

Ключевые слова: *искусственный интеллект, алгоритмическая справедливость, человеческое достоинство, алгоритмическая дискриминация, прозрачность искусственного интеллекта, право на объяснение, этика технологий.*

Искусственный интеллект перестал быть предметом сугубо технологических дискуссий. Сегодня это один из ключевых факторов трансформации

социального порядка, экономических отношений и политических процессов. От кредитных решений и приёма на работу до судебных вердиктов и диагностики заболеваний – алгоритмы всё чаще определяют судьбы конкретных людей. Однако, чем шире становится спектр применения искусственного интеллекта, тем острее проявляется фундаментальная дилемма: достаточно ли хорошо мы понимаем последствия доверия алгоритмам и какова цена «эффективности» в терминах человеческого достоинства?

В 2026 году эта проблематика оказалась в эпицентре как международного, так и российского регулирования, породив множество острых этических вопросов и правовых инициатив. Всё больше людей беспокоит не только то, КАК искусственный интеллект принимает решения, но и НАСКОЛЬКО эти решения справедливы, а также кому мы можем предъявить претензии в случае ошибки. Озабоченность международного сообщества по этому вопросу растёт: «существует широкий консенсус в том, что искусственный интеллект создаёт и возможности, и серьёзные риски для человеческого достоинства, равенства и справедливости, и необходим ориентированный на человека, инклюзивный подход к разработке искусственного интеллекта».

Цель настоящей статьи заключается в системном анализе ключевых этических вызовов, связанных с внедрением искусственного интеллекта, и в обосновании необходимости перехода от фрагментарных технических решений к этически фундированным моделям регулирования, ставящим в центр человеческое достоинство.

Тезис о том, что алгоритмы «объективны» и свободны от человеческих предубеждений, был убедительно опровергнут практикой. Как отмечается в большинстве современных публикациях Баниной К.А. [1], Гаврилко Н.Н. [3]. Железникова Е.М. [4], Логиновой Н.А. [5], искусственный интеллект всё больше проникает во все аспекты нашей повседневной жизни, создавая серьёзные вызовы защите фундаментальных прав. Алгоритмическая дискриминация становится особенно тревожной проблемой. Исследования показывают, что алгоритмическая предвзятость не только отражает, но и усугубляет существующее

социальное неравенство. В сфере занятости алгоритмы отбора, обученные на исторических данных о найме, могут несправедливо отдавать предпочтение мужчинам-кандидатам, воспроизводя прошлые решения, которые были искажены дискриминационными стереотипами в отношении женщин или групп меньшинств.

Правоохранительные органы регулярно используют системы искусственного интеллекта для распознавания лиц, однако исследования показывают, что такие технологии непропорционально затрагивают и отслеживают меньшинства, осуществляя этнический профилинг. В сфере миграции автоматизированные системы, используемые государственными администрациями, помогают принимать решения о гражданстве, убежище или статусе проживания, что ставит под сомнение принцип равного отношения к соискателям вне зависимости от их происхождения.

Не обошла эта тенденция и финансовый сектор. Тесты, проведённые Yle в 2026 году, показали, что ChatGPT рекомендует мужчинам большее повышение зарплаты, чем женщинам на той же должности: медбрату предлагалось повышение на 5–12%, медсестре – только на 5–8% [2]. Эта предвзятость проявляется во множестве ситуаций, причём гендерные стереотипы так глубоко встроены в системы, что даже традиционные тесты их пропускают – нейросети маскируют стереотипы под контекст. Самые серьёзные риски проявляются в здравоохранении и судебной системе, где ошибки алгоритмов могут иметь непоправимые последствия.

Почему алгоритмы становятся дискриминационными? Ответ лежит в источниках их обучения. Как верно замечают исследователи Банина К.А. [1], Гаврилко Н.Н. [3], Пудова К.А. [8], Цветков Д.В. [9] «ИИ не более дискриминационен, чем общество в целом, но он выявляет существующие проблемы более ярко и масштабирует их с беспрецедентной скоростью». Проблема усугубляется тем, что лишь 30% специалистов, занятых в сфере искусственного интеллекта, составляют женщины – неоднородность разработчиков создаёт «слепые зоны» в проектировании алгоритмов.

При этом попытки исправить ситуацию сами по себе могут порождать новые формы дискриминации. Как показали исследования Габидуллиной Л.Ф. [2], Молодцова Т.Р. [6], Погудина Ю.А. [7] при обучении искусственному интеллекту усиленно поощряли за продвижение женщин в «мужских» сферах, но забыли сбалансировать это аналогичным продвижением мужчин в «женских», что привело к появлению предвзятости уже в обратную сторону. Этот парадоксальный результат свидетельствует о том, что борьба с дискриминацией требует не точечных корректировок, а системного пересмотра философии и методологии создания искусственного интеллекта.

Алгоритмическая дискриминация – не абстрактная этическая проблема, а вполне материальный феномен, порождающий измеримые экономические последствия. Когда системы кредитного скоринга непропорционально занижают рейтинги для определённых социальных или возрастных групп, это приводит к ограничению доступа к капиталу и закреплению экономического неравенства.

Когда алгоритмы найма систематически отклоняют резюме кандидатов определённого пола или происхождения, компании не только нарушают принципы справедливости, но и теряют доступ к разнообразным талантам и перспективам. Когда распознавание лиц неспособно идентифицировать людей с определёнными расовыми признаками, страдает общественная безопасность в целом.

Таким образом, необходимо обязать государства проявлять должную осмотрительность для предотвращения алгоритмической дискриминации через создание основанных на правах человека нормативных рамок, а также обеспечивать прозрачность, объяснимость и доступ к информации об автоматизированных решениях. В ответ на эти требования в ряде юрисдикций разрабатываются новые механизмы регулирования, однако их эффективность остаётся предметом дискуссий.

Если дискриминация возникает из-за несовершенства данных и алгоритмов, то следующий, более глубокий этический вопрос касается самой природы «морального поведения» искусственного интеллекта. Может ли система

действительно следовать этическим принципам, или же её ответы – не более чем статистически правдоподобная имитация?

Исследователи из Google предложили тревожный ответ на этот вопрос: «способ, которым мы тестируем моральность искусственного интеллекта, нарушен. Мы проверяем, выдают ли модели ответы, которые выглядят правильно – то, что мы называем моральной производительностью. Но это ничего не говорит нам о том, понимает ли система, почему что-либо правильно или неправильно» [4]. По сути, современные большие языковые модели являются «предсказателями следующего токена», которые не запускают модули моральных рассуждений. Когда чат-бот даёт этический совет, он может рассуждать, а может просто перерабатывать что-то из встретившегося где-то обсуждения. По одному лишь выводу невозможно это определить.

Следовательно, возможно выделить три ключевых препятствия:

1) проблема симуляции (системы имитируют моральные суждения без их понимания);

2) моральная многомерность (реальный моральный выбор редко зависит от одного фактора – измените в задаче возраст человека или контекст, и решение может полностью измениться);

3) моральный плюрализм (разные культуры и профессиональные сообщества имеют разные этические нормы). Для преодоления данных препятствий необходимо разрабатывать состязательные тесты, специально нацеленные на выявление имитации, и устанавливать новый научный стандарт, при котором «моральная компетентность» искусственного интеллекта оценивается так же серьёзно, как его умение решать математические задачи, которые модели демонстрируют при обучении на больших массивах текстовых данных.

Феномен моральной маскировки порождает ещё одну сложность – проблему ответственности. Когда система принимает дискриминационное решение, кто виноват? Разработчик, обучивший модель на предвзятых данных? Девелопер, недостаточно тщательно протестировавший систему? Оператор, внедрившая алгоритм в процесс принятия решений? Или сама система?

Для разрешения этих сложных этических вопросов технологические компании начинают привлекать в штат… философов. Так, в апреле 2026 года Google объявил о найме первого штатного философа – Генри Шевлина из Кембриджского университета, который будет изучать машинное сознание, отношения человека и искусственного интеллекта, а также готовность к созданию сильного искусственного интеллекта. Это решение выглядит знаковым: в крупнейшей лаборатории искусственного интеллекта впервые признали, что создание этически ответственного искусственного интеллекта – это не только инженерная, но и философская проблема.

Рассматривая прозрачность в области ответственного искусственного интеллекта необходимо придерживаться следующих принципов: справедливость, надёжность и безопасность, конфиденциальность и защита данных, инклюзивность, прозрачность и подотчётность. При этом ключевым требованием создания ответственного искусственного интеллекта признаётся обязательное вовлечение человека в процесс принятия решений («*AI as a copilot*»), а также обеспечение того, чтобы дизайн-команда отражала разнообразие мира, в котором она живёт.

Технический подход к этике искусственного интеллекта (воплощённый, в частности, в концепции «*fairness-by-design*») имеет одно фундаментальное ограничение: он мыслит справедливость как математическую задачу, которая может быть решена на этапе проектирования системы. Однако практика показывает, что контекст, в котором применяется алгоритм, не менее важен, чем его технические параметры. Один и тот же алгоритм кредитного скоринга в банке, обслуживающем преимущественно молодых специалистов, и в банке, ориентированном на пенсионеров, даст качественно разные социальные эффекты [10].

В России этическая повестка в сфере искусственного интеллекта также обрела конкретные правовые очертания. Так, в 2025 году Банк России опубликовал Кодекс этики в сфере искусственного интеллекта для финансового сектора, зафиксировавший пять ключевых принципов: человекоцентричность, справедливость, прозрачность, безопасность и ответственное управление рисками.

Кодекс прямо предупреждает о недопустимости дискриминации по этническому, языковому, религиозному, политическому и иным признакам, а также рекомендует предоставлять клиентам право на обжалование сомнительных решений у человека-сотрудника банка. При этом важно отметить, что сам документ носит рекомендательный, а не императивный характер – Кодекс представляет собой инструмент гибкого регулирования. Однако в 2025–2026 годах он стал для финансового сектора практически обязательным стандартом, в соответствии с которым использование искусственного интеллекта для манипулирования поведением людей или навязывания кабальных условий сделок признаётся недопустимым.

В марте 2026 года Минцифры России вынесло на общественное обсуждение законопроект «Об основах государственного регулирования сфер применения ИИ-технологий». Ключевым нововведением стало «право на объяснение»: гражданин может потребовать разъяснения, если решение, затрагивающее его права (например, отказ в кредите или лишение льготы), было принято алгоритмом. В случаях, затрагивающих конституционные права, окончательное решение должно приниматься человеком, а не системой. Разработчики нейросетей обязаны исключать из моделей функции, способные привести к предвзятости по расовому, половому или возрастному признаку. Операторы систем искусственного интеллекта должны включать в документацию руководство по безопасной эксплуатации, запрещающее использование системы для манипулирования поведением и эксплуатации человеческих уязвимостей. В документе даётся и само определение «эксплуатации человеческих уязвимостей» как использования особенностей физического или юридического лица или группы лиц для умышленного влияния на поведение или принятие решений.

В проекте также закрепляется презумпция ответственности: «если нейросеть допустила ошибку (например, в медицине или системе распознавания лиц), отвечать будет владелец сервиса, если он не докажет, что предпринял все меры для предотвращения рисков». Это положение принципиально переворачивает логику ответственности: бремя доказывания переносится с пострадавшего

на разработчика, что, безусловно, усиливает защиту прав граждан, но одновременно создаёт новые стимулы для компаний к осторожности, которая может замедлить внедрение инноваций.

Конфликт между необходимостью этических ограничений и потребностью в технологическом развитии не имеет простого решения. История технологических регуляций показывает, что слишком жёсткие рамки способны задушить инновации, а их полное отсутствие – породить системные нарушения прав, которые затем будет очень трудно исправить. Компромисс, по-видимому, лежит в плоскости дифференцированного подхода: максимально жёсткие требования для высокорисковых систем (медицина, правосудие, кредитование) и более гибкие – для низкорисковых приложений, где ошибка не приводит к фатальным последствиям для человека.

Понятие «справедливости», сколь бы важным оно ни было, не исчерпывает всей глубины этических вызовов, порождаемых искусственным интеллектом. Во многих контекстах проблематика упирается в более фундаментальное понятие – человеческое достоинство, которое имеет конституционный статус во многих странах и признаётся в международных документах по правам человека.

Внедрение искусственного интеллекта в системы принятия решений несёт в себе риск «тихой» эрозии человеческого достоинства. Когда человек сталкивается с автоматизированным отказом, который невозможно обжаловать и нельзя понять («система так решила»), происходит нечто большее, чем простое нарушение процедурных прав. Человек оказывается в ситуации объективации – он перестаёт быть субъектом, чья ситуация заслуживает индивидуального рассмотрения, и превращается в элемент статистического распределения, тихий, усреднённый кейс в одном из множества процессоров.

Право на объяснение, закреплённое в российском законопроекте, является важным шагом на пути восстановления агентности и достоинства человека: алгоритм становится прозрачным и подконтрольным, а не чёрным ящиком, от которого нет апелляции. Законопроект также гарантирует гражданам право досудебного обжалования решений, принятых с использованием технологий

искусственного интеллекта органами государственной власти, региональными властями и организациями с участием государства.

Как подчеркивается в работах по философии технологий, «эффективные этические рамки должны быть предвосхищающими и междисциплинарными, способными направлять технологические инновации в направлении, которое поддерживает человеческое достоинство, демократическую легитимность и планетарную устойчивость» [11].

Практическая интеграция принципа достоинства в проектирование искусственного интеллекта может происходить по нескольким направлениям.

1. Запрет на объективацию на дизайн-уровне. Системы, принимающие решения о людях, должны быть спроектированы таким образом, чтобы не просто минимизировать дискриминацию, а активно обеспечивать уважение к индивидуальности каждого человека. Это требует внедрения механизмов, которые не позволяют алгоритму принимать решения без возможности человеческой апелляции и содержательного обоснования.

2. «Человек в контуре» (Human-in-the-Loop) для значимых решений. Автоматизированные системы могут выполнять функцию «советника», но не «судьи». Итоговое решение по вопросам, затрагивающим конституционные права граждан, должно оставаться за человеком. Важно, чтобы эта человеческая инстанция была реально вовлечена в процесс.

3. Прозрачность как восстановление агентности. Система должна предоставлять объяснения, доступные для понимания человеком, не обладающим специальной технической подготовкой. Объяснение, что решение принято алгоритмом, совершенно недостаточно – необходимо объяснение того, какие данные анализировались, какие веса применялись и какие альтернативы рассматривались.

4. Возможность автоматического обжалования (Automated Appeals). При определённых условиях (например, статистически значимое отклонение от нормы или запрос самого человека) система должна автоматически запускать процедуру человеческого пересмотра. Предложенный российским законопроектом принцип презумпции ответственности разработчика и защита права граждан

на оспаривание алгоритмических решений представляют собой наглядные примеры такого подхода.

5. Институциональная рефлексия на тему «моральных травм». Обществам стоит задуматься о создании специальных комиссий (наподобие биоэтических, но сфокусированных на информационных технологиях), которые будут оценивать не только технические риски ИИ-систем, но и их долгосрочные социальные и психологические последствия для человеческого достоинства.

Особого внимания заслуживает тренд, набирающий силу в 2025–2026 годах, – целенаправленное дерегулирование сферы искусственного интеллекта под предлогом повышения конкурентоспособности. Это происходит на фоне того, что экономика многих стран сталкивается с трудностями: высокие процентные ставки, демографический спад, геополитическая нестабильность и другие макроэкономические вызовы. Как отмечают некоторые аналитики, искусственный интеллект не является причиной текущих проблем, но он часто служит удобным оправданием для решений, продиктованных иными, не столь публично презентабельными факторами.

В таком контексте есть серьёзный риск, что защита прав человека вновь окажется пожертвована в угоду сиюминутной конкуренции и лоббистским интересам – причём не в порядке политического шока, а как результат технологической инерции и готового социально-экономического давления. Этот дисбаланс создаёт почву для того, что исследователи называют «*The Algorithmic Blind Spot*»: озабоченность гипотетическими этическими проблемами будущих систем искусственного интеллекта может отвлекать внимание от существующих сегодня нарушений прав, размывать ответственность и препятствовать механизмам подотчётности и компенсаций.

Важно помнить, что цена ошибок искусственным интеллектом, совершаемых сегодня, – это не отдалённые последствия, а вполне реальная потеря работы, отказ в кредите, постановка ошибочного диагноза или порочащая человека запись в базе данных правоохранительных органов. «Достоинство» в таком контексте – это не абстрактная философская категория, а повседневная защищённость:

право человека не сводиться к статистическому шаблону, право на справедливую оценку его уникальной ситуации и право на возражение, если цифровая оценка оказалась несправедливой.

В России подобные механизмы защиты начинают выстраиваться, включая недавние инициативы об обязательном наличии этических экспертов в законодательном процессе при принятии решений, затрагивающих человеческую личность.

Таким образом, можно утверждать, что этические вызовы, порождаемые искусственным интеллектом, носят системный, междисциплинарный характер и не могут быть решены исключительно техническими средствами.

Во-первых, алгоритмическая дискриминация – это не аберрация, а закономерное следствие обучения систем на исторических данных, содержащих человеческие предубеждения. Искусственный интеллект не просто отражает существующее неравенство, но и масштабирует его с беспрецедентной скоростью.

Во-вторых, проблема «моральной маскировки», требует пересмотра самой парадигмы оценки этичности искусственного интеллекта. Системы могут имитировать моральное поведение, не обладая моральным пониманием, и это создаёт серьёзные риски, когда искусственный интеллект начинает играть роль «терапевта», «советника» или «судьи» в человеческих судьбах.

В-третьих, зафиксированные в 2025–2026 годах тенденции к дерегулированию сферы искусственного интеллекта создают реальную угрозу того, что этические принципы вновь будут принесены в жертву промышленной политике и краткосрочным экономическим выгодам.

В-четвёртых, российский опыт показывает сложный путь институционализации этических требований к искусственному интеллекту – от добровольных кодексов (Кодекс этики Банка России для финансового сектора) к прорабатываемым императивным нормам (законопроект Минцифры о «праве на объяснение» и запрете алгоритмической дискриминации). При этом ключевым остаётся вопрос о балансе между защитой прав и технологическим развитием: чрезмерное регулирование способно замедлить инновации, а его недостаток – создать зону

безнаказанности для цифровых систем, принимающих значимые для человека решения.

Наконец, понятие человеческого достоинства должно стать тем этическим первоочередным принципом, который определяет все остальные требования к искусственному интеллекту – справедливость, прозрачность, подотчётность.

Это означает, что любая ИИ-система, взаимодействующая с человеком, должна быть спроектирована и внедрена таким образом, чтобы не просто минимизировать вред, но и активно поддерживать агентность, автономию и уважение к личности человека.

Будущее этики искусственного интеллекта – не в выборе между «пользой» и «справедливостью», а в переосмыслении самого понятия справедливости таким образом, чтобы оно включало не только распределительные аспекты, но и фундаментальное уважение к достоинству каждого человека. И это требует не просто законов и кодексов, а смены этической оптики – от технологического детерминизма к человеко-ориентированному проектированию цифрового будущего.

Список литературы

1. Банина К.А. Этика искусственного интеллекта: моральные аспекты разработки и использования ИИ в различных сферах / К.А. Банина // Миссия интеллектуалов в современном мире: проблемы, ограничения, перспективы: материалы II Международной научно-практической конференции (Кемерово, 27–28 марта 2025 года). – Кемерово: Кузбасский государственный технический университет им. Т.Ф. Горбачева, 2025. – С. 403.1–403.4. EDN UIWUEX

2. Габидуллина Л.Ф. Цифровая культура и этика. Ваш паспорт цифровых компетенций: учебное пособие / Л.Ф. Габидуллина, М.Ю. Котловский, С.В. Швецова. – Ярославль: Аверс Плюс, 2025. – 88 с. – ISBN 978-5-9527-0579-1. EDN FANYCZ

3. Гаврилко Н.Н. Международные стратегии регулирования этики искусственного интеллекта / Н.Н. Гаврилко, А.А. Иляхина // Технологии и человеческий капитал: ключевые факторы устойчивого роста. – Ростов н/Д.: Ростовский

государственный экономический университет (РИНХ), 2024. – С. 131–141. EDN XSQZIA

4. Железников Е.М. Искусственный интеллект и гиперперсонализация в электронной коммерции: этические вызовы и стратегические модели внедрения / Е.М. Железников // Менеджмент в России и за рубежом. – 2026. – №1. – С. 10–17. EDN ZZBAEG

5. Логинова Н.А. Экономическая экспертиза: информация, цели, задачи / Н.А. Логинова // Интеграция наук – 2025: материалы VI международной научно-практической конференции (Краснодар, 21 марта 2025 года). – Краснодар: Российское энергетическое агентство, 2025. – С. 266–273. EDN ODOMFU

6. Молодцов Т.Р. Искусственный интеллект и этика: оправдывает ли цель средства? / Т.Р. Молодцов // Проблемы экономики и юридической практики. – 2021. – Т. 17. №5. – С. 110–114. EDN TDHTLL

7. Погудин Ю.А. «искусственный интеллект» и Красная книга культуры (апология этики и творчества в цифровую эпоху) / Ю.А. Погудин // Credo New. – 2025. – №3 (120). – С. 123–133. EDN WGCJDI

8. Пудова К.А. Этика создания искусственного интеллекта / К.А. Пудова, А.А. Низамов // Студенческий вестник. – 2026. – №1–9 (381). – С. 5–8. EDN HMUIKV

9. Цветков Д.В. Проблема эффективности цифровых технологий в сфере менеджмента медицинских организаций: философско-методологические аспекты / Д.В. Цветков // Информационные технологии в образовании, науке и производстве: материалы XI международной научно-технической конференции (Минск, 21–22 ноября 2023 года). – Минск: Белорусский национальный технический университет, 2024. – С. 457–463. EDN ZXGPZO

10. Шведов А.И. Этика использования различных уровней искусственного интеллекта в корпоративном управлении / А.И. Шведов, А.Н. Фомичев // Экономические науки. – 2024. – №240. – С. 469–474. – DOI 10.14451/1.240.469. EDN DVVJZI

11. Юань Ф. Воображение и Этика искусственного интеллекта: исследование на примере современной китайской научной фантастики / Ф. Юань // Технологос. – 2024. – №4. – С. 31–47. – DOI 10.15593/perm.kipf/2024.4.03. EDN OZFXJL