УПРАВЛЕНИЕ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИМИ СИСТЕМАМИ

Обзорная статья

УДК 004.89+519.816
DOI 10.37468/2307-1400-2026-2-5-17

Этические риски применения систем поддержки принятия решений на основе искусственного интеллекта в организациях

Матвеев Александр Владимирович[1]

[1] *Санкт-Петербургский университет Государственной противопожарной службы МЧС России имени Героя Российской Федерации генерала армии Е.Н. Зиничева, Санкт-Петербург, Россия*

Аннотация

В статье анализируются этические риски, возникающие при использовании систем поддержки принятия решений на основе технологий искусственного интеллекта (ИИ) в организациях. Такие системы повышают скорость и обоснованность решений за счет обработки больших массивов данных, но одновременно порождают проблемы алгоритмической предвзятости, непрозрачности, размывания ответственности и защиты персональных данных. На основе обзора литературы и анализа реальных практических случаев применения технологий ИИ выделены четыре группы проблем, создающих этические риски (проблемы предвзятости и справедливости, проблемы прозрачности и объяснимости, проблемы подотчетности и ответственности в принятии решений, проблемы конфиденциальности и защиты персональных данных). Показано, что снижение этих рисков требует не только соблюдения регуляторных требований, но и организационных мер (этического проектирования систем, алгоритмического аудита, создания комитетов по этике ИИ и постоянного мониторинга моделей после развертывания). Сочетание справедливости, объяснимости и подотчетности позволяет организациям использовать преимущества ИИ в принятии решений, не отступая от этических стандартов. Научная новизна исследования состоит в систематизации этических рисков систем поддержки принятия решений на основе ИИ применительно к организациям. Практическая значимость исследования определяется возможностью использования предложенных стратегий этичного внедрения ИИ при проектировании и вводе в эксплуатацию систем поддержки принятия решений в организациях.

Ключевые слова: искусственный интеллект, системы поддержки принятия решений, этические риски, прозрачность, предвзятость, конфиденциальность, черный ящик.

Введение

Системы поддержки принятия решений (СППР), использующие технологии искусственного интеллекта (ИИ) постепенно становятся привычным инструментом управления в различных организациях [1, 2]. В их основе лежат методы машинного обучения, анализ больших данных и автоматизации рутинных операций, что приводит к повышению оперативности подготовки решений и сокращает объем ручной работы [3]. Одновременно с ростом отдачи от таких систем становится очевидной и обратная сторона их применения. Справедливость, предвзятость, прозрачность и подотчетность превращаются из абстрактных требований в практические ограничения [4]. Этические аспекты имеют принципиальное

значение, поскольку решения, подготовленные алгоритмом, могут затрагивать конкретных людей (сотрудников, клиентов и других заинтересованных лиц), поэтому требуют не только технической, но и нормативной оценки [5].

Первая из этических проблем СППР на основе ИИ состоит в предвзятости алгоритмов. Модель, обученная на входных данных, которые могут отражать определенное социальное неравенство, будет воспроизводить это неравенство в своих решениях [6, 7]. Немаловажной является и проблема «черного ящика», состоящая в том, что логика получения результата на основе ИИ зачастую является непрозрачной. Лица, принимающие решения (ЛПР) или организация в целом не могут ни объяснить решение, ни защитить его перед регулятором или пострадавшей стороной [8]. Отсутствие обоснования затрудняет установление ответственности, особенно в случаях, когда результаты принятого решения нанесли ущерб конкретному субъекту [9]. Следующая этическая проблема связана с обработкой конфиденциальной информации. Чем больше персональных данных интегрируется в системе, тем выше риск их некорректного использования, утечки или несоблюдения регуляторных требований [10]. Если применение СППР не регулируется и не проверяется по соответствию этическим критериям, то последствия могут нести как заинтересованные лица, так и сами организации в виде потенциальных юридических и финансовых рисков [11].

С учетом этих проблем в статье предлагается критический обзор этических рисков СППР на основе ИИ в организациях. Методом исследования выступает обзор литературы, дополненный анализом реальных практических случаев применения ИИ в организациях.

Обзор исследований

Применение ИИ в СППР способно повысить скорость обработки данных и обоснованности управленческих решений [1, 3]. Вместе с тем предвзятость, непрозрачность, размытая ответственность и уязвимость персональных данных являются предметом научного интереса многих исследователей [12]. В ряде исследований предложены этические принципы, механизмы управления и технические средства, направленные на справедливое и ответственное использование ИИ [4, 13]. В этом разделе рассматриваются основные этические риски использования СППР на основе ИИ и подходы к их управлению.

СППР прошли путь от экспертных систем, основанных на явно заданных правилах, до сложных аналитических платформ, использующих большие данные, машинное обучение и нейронные сети [1]. Ранние системы предлагали решения на основе заранее формализованных правил. СППР на основе ИИ включают в себя предиктивную аналитику, обработку естественного языка и глубокое обучение, работая со структурированными и неструктурированными данными в режиме, близком к реальному времени [3]. Нельзя не отметить чрезвычайную значимость их использования в сфере финансов, здравоохранения, маркетинга и управления персоналом. Данные системы снижают вероятность отдельных типов человеческих ошибок и повышают операционную эффективность [2].

Однако внедрение таких систем сопровождается этическими сложностями, которые необходимо учитывать при их применении. Использование ИИ в принятии решений актуализирует вопросы беспристрастности, ответственности и раскрытия информации. Причем, чем ниже степень участия ЛПР в контуре управления, тем острее данная проблема [2, 14].

1. Проблемы предвзятости и справедливости в СППР на основе ИИ

Предвзятость остается наиболее типичной и обсуждаемой этической проблемой СППР. Алгоритмы способны проявлять социальное неравенство в таких областях, как трудоустройство, кредитование и правоохранительная деятельность [6, 7, 15, 16]. Чтобы реализовать возможности использовать такие СППР и одновременно соблюдать этические ограничения, организациям приходится применять формализованные принципы управления, проводить аудит моделей и сохранять участие человека в принятии решений [17].

Источники алгоритмической предвзятости могут быть различны. Это и смещения в исходных данных, нерепрезентативная выборка, наследование предубеждений от предыдущих версий модели [6]. Так, инструменты подбора персонала на основе ИИ воспроизводили гендерные и расовые различия, поскольку обучались на исторических решениях о найме [15]. Аналогичным образом системы кредитного скоринга связывают с ухудшением доступа к заемным средствам для отдельных категорий заявителей [16, 18].

Устранение подобного рода предвзятости является многофакторной задачей, т.к. она требует применения методов машинного обучения с ограничениями справедливости, использования разнообразных и репрезентативных наборов данных и сохранения человеческого контроля на всех этапах формирования и принятия решений [7, 19]. Также для решения данной задачи может использоваться алгоритмический аудит и инструменты обнаружения смещений [20]. Кроме того, некоторые организации интегрируют процедуры управления решениями ИИ так, чтобы заинтересованные стороны могли проверить и оспорить результат, который считают несправедливым [9, 20].

Риски предвзятости проникают в систему управления на разных этапах ее жизненного цикла, начиная со сбора данных [5]. Если обучающая выборка содержит результаты прошлых дискриминационных решений, модель будет их учитывать в будущих решениях [6]. Системы отбора персонала, обученные на прежних критериях фильтрации резюме, наследуют эти критерии [15]. В финансовом секторе смещенные обучающие данные ведут к систематически худшим условиям для отдельных групп заемщиков [16]. Второй тип смещения возникает на этапе обучения модели, когда используемые алгоритмы оптимизации настроены на точность, а не на справедливость [7]. Если критерий справедливости не задан явно, модель с высокой вероятностью усилит существующее неравенство. В медицинской диагностике недостаточная представленность отдельных групп в обучающих данных приводит к росту диагностических ошибок именно для этих групп [21]. Кроме того, модели глубокого обучения являются непрозрачными, они не объясняют, почему один исход предпочтен другому. Это затрудняет одновременно и проверку справедливости, и установление ответственности [8, 19]. Аспекты предвзятости могут проявляться и на этапе эксплуатации системы и ее дообучения на реальных данных. Чем дольше модель работает, тем сильнее она закрепляет сложившиеся закономерности, в том числе нежелательные [5]. Например, у чат-бота клиентской поддержки смещение может формироваться из самих диалогов с пользователями и влиять на последующие ответы. Без контроля такие эффекты фиксируются в поведении системы и с трудом устраняются из-за петель обратной связи между рекомендациями модели и данными, которые она затем получает [5, 20].

Для снижения этих рисков необходимо интегрировать инструменты обнаружения смещений, задавать ограничения справедливости в моделях и целенаправленно формировать разнообразные обучающие выборки [7]. Алгоритмический аудит и методы устранения смещений позволяют выявлять проблемы на ранних стадиях [20, 22]. Методы объяснимого ИИ (XAI), интегрируемые в СППР, способны снизить данные риски, делая решения доступными для содержательной проверки [19, 23].

Таким образом, на основе анализа исследований можно сделать общий вывод. Если работа с предвзятостью не встроена в проектируемую СППР на основе ИИ с самого начала, то система способна закреплять негативные практики, а доверие к процессу принятия решений в целом снижается [12, 20].

2. Проблемы прозрачности и объяснимости

Как отмечено выше, в большинстве приложений ИИ система работает как «черный ящик», т.е. она выдает рекомендации, не раскрывая, как именно к ним пришла [8]. В большей степени это проявляется при использовании моделей глубокого обучения, т.к. их внутренняя структура и логика не поддается интерпретации, что мешает пользователям сформировать обоснованный уровень доверия [19]. Особенно значима эта проблема там, где рекомендации системы могут

оказывать существенное влияние на жизнь и правовое положение людей (в организациях в сфере здравоохранении, финансов и юриспруденции) [21, 24].

Решением этой проблемы стало развитие методов объяснимого ИИ (XAI), направленных на повышение интерпретируемости моделей [19, 23]. К ним относятся оценка важности признаков, апостериорные объяснения отдельных предсказаний и контрфактические объяснения, показывающие, при каком изменении входных данных решение было бы иным [9, 25, 26].

Организациям, использующим СППР, объяснимость нужна не сама по себе, а как условие анализа генерируемых решений и возможности проверить систему, когда результат оспаривается [9, 22]. В ряде научных исследований отмечается, что в задачах с высокой ценой ошибки предпочтительнее изначально интерпретируемые модели, а не апостериорные объяснения в непрозрачных моделях [8].

Основные проблемы прозрачности и объяснимости обобщены в Таблице 1.

Таблица 1 – Проблемы прозрачности и объяснимости в принятии решений ИИ

Проблема	Влияние	Решение
ИИ как «черный ящик»	Решение невозможно объяснить, а значит обоснованно оспорить или подтвердить	Применение методов объяснимого ИИ (XAI) для повышения прозрачности
Недостаток прозрачности в моделях глубокого обучения	Сложность архитектуры делает логику вывода недоступной для пользователя	Оценка важности признаков, апостериорные и контрфактические объяснения
Решения с высокой ценой ошибки (здравоохранение, финансы, право)	Затрагиваются жизнь и права людей, но обоснование решения недоступно	Требование объяснимости как условие анализа решений и общественного доверия

3. Проблемы подотчетности и ответственности в принятии решений ИИ

Принятие решений на основе ИИ так или иначе затрагивает несколько различных субъектов (разработчиков моделей, ЛПР, организации, использующие данные технологии, регуляторов) [28]. Когда система ИИ, встроенная в СППР, приводит к неблагоприятным последствиям (например, отказу какому-то кандидату по дискриминирующему признаку, или ошибке автономной системы), то установление ответственного лица за данное решение становится самостоятельной проблемой [27, 29]. Анализ практического опыта в случаях ошибок в инструментах найма и ложных задержаний из-за сбоев систем распознавания лиц показывают, что существующих механизмов недостаточно [15, 22].

В ответ регулирующие органы и сами организации развивают инструменты обеспечения подотчетности, такие как нормативное регулирование ИИ, корпоративные кодексы этики и процедуры оценки соответствия [29, 30]. Таким образом, большее внимание к подотчетности при применении ИИ в принятии решений предполагает включение инструментов обеспечения объяснимости решений, регулярного аудита моделей и механизмов рассмотрения жалоб на несправедливость или вред, причиненный СППР на основе ИИ [22, 27]. Отдельного внимания заслуживает необходимость закрепления прав и обязанностей всех участников жизненного цикла системы, использующей технологии ИИ, [28, 31].

4. Проблемы конфиденциальности и защиты персональных данных

СППР на основе ИИ требуют агрегирования больших объемов данных, что обостряет вопросы их безопасности, правовых оснований обработки и допустимости использования [10]. Организациям приходится соблюдать требования законодательства о персональных данных, что является необходимым условием правомерности решений, подготовленных ИИ [27, 29]. При этом баланс между потребностью моделей в данных и защитой частной жизни остается

нетривиальной задачей [10]. Для ее решения используются анонимизация данных, дифференциальная конфиденциальность и федеративное обучение, при котором модель обучается, не выгружая персональные данные с устройств или из локальных хранилищ, данные остаются «на месте», а на сервер передаются только обновления параметров модели (градиенты/веса), которые затем агрегируются в глобальную модель [32]. Сочетание таких методов с усилением мер кибербезопасности позволяет реализовывать СППР, обеспечивающих безопасность субъектов данных и сохраняющих доверие пользователей к решениям СППР [10, 32].

Управление ИИ и этические принципы

Для снижения этических рисков организации и регуляторы внедряют механизмы управления ИИ и этического контроля [28]. Организация экономического сотрудничества и развития (ОЭСР), Европейский союз и др. сформулировали принципы развития технологий ИИ, среди которых прозрачность, непричинение вреда, уважение частной жизни и объяснимость [4, 13]. Сравнительный анализ таких документов показывает содержательное сближение вокруг небольшого числа принципов при заметном расхождении в способах их практической реализации [4]. Типовые рекомендации включают создание комитетов по этике ИИ, регулярный аудит и сохранение человека в контуре принятия решений [17, 22].

Поэтому закрепление этических требований на этапе проектирования системы становится необходимым условием того, что ценности организации не будут утрачены при автоматизации, а сама система не будет использована таким образом, чтобы приносить вред пользователям или другим заинтересованным лицам [13, 28, 30, 33]. Вместе с тем ряд авторов указывает на ограниченность декларативного подхода, т.к. набор принципов сам по себе не обеспечивает этичность практики, если он не подкреплен процедурами, распределением ответственности и санкциями [12, 34].

Управление развитием технологий ИИ и этичная реализация СППР на основе ИИ должно опираться на ряд принципов:

1) Принцип этичного проектирования систем, учитывающего справедливость, подотчетность и прозрачность при построении моделей, включая отбор разнообразных наборов данных, регулярную отчетность о показателях справедливости и проверку того, что модель не ухудшает положение отдельных лиц или групп [7, 13]. Без таких мер решения, подготовленные системой, остаются уязвимыми для обоснованных возражений.

2) Принцип управления рисками и соблюдения нормативных требований. Системы ИИ подпадают под действие законодательства о персональных данных и специализированных норм, разрабатываемых применительно к ИИ [27, 29]. Эти нормы требуют прозрачности автоматизированных решений, защиты персональных данных и явного распределения ответственности за результат [27]. Для их выполнения организации формируют внутренние комитеты по этике ИИ, привлекают внешних аудиторов и создают свои подразделения, выполняющие данные функции [22, 29]. В отсутствие таких механизмов организация принимает на себя юридические риски, которыми не управляет.

3) Принцип постоянной проверки и обратной связи. Модели и технологии ИИ требуют обязательной проверки, мониторинга и оценки качества результатов в процессе эксплуатации [22]. Это гарантирует, что системы ИИ в организациях являются более этичными и нейтральными и соответствуют организационным стандартам. [20]. Использование объяснимого ИИ (XAI) в этом случае дает заинтересованным сторонам возможность понять основания решения и потребовать его пересмотра в случае необходимости [19, 23].

Практика ряда ведущих компаний (Google, IBM, Microsoft и др.) показывает, как эти принципы соединяются в единую систему управления, когда создаются комитеты по этике, инструменты оценки предвзятости, процедуры алгоритмического аудита [22, 30]. Такие примеры позволяют другим организациям отобрать релевантные для своей отрасли механизмы и

выстроить собственную модель управления, в которой ответственное использование ИИ закреплено не декларацией, а реальными процедурами [28, 34].

В Российской Федерации (РФ) регулирование ИИ и этический контроль реализуется на разных уровнях, от стратегических документов до отраслевых рекомендаций и технических стандартов. Ключевым документом в сфере этического регулирования является Кодекс этики в сфере искусственного интеллекта (принят в 2021 году). Он носит рекомендательный характер («мягкое право»). К нему присоединяются участники рынка: технологические компании (Сбер, Яндекс, VK), вузы, органы власти. В кодексе закреплены такие принципы как человекоцентричность, безопасность, прозрачность, конфиденциальность данных, сохранение человеческого контроля и чёткое распределение ответственности. Существуют также отраслевые кодексы этики. Например, Кодекс этики применения искусственного интеллекта в сфере охраны здоровья (утвержден Межведомственной рабочей группой при Минздраве России в феврале 2025 года). Он детализирует этические нормы в сфере медицины, подчёркивает ограниченную автономность систем ИИ, необходимость объяснимости решений («проблема чёрного ящика»), важность конфиденциальности и недопустимость дискриминации. Также в РФ утверждено более 100 национальных стандартов и предварительных национальных стандартов (ПНСТ). Они охватывают разные аспекты: функциональную безопасность (ПНСТ 836-2023), обеспечение доверия к ИИ (ГОСТ 59276), оценку качества систем (ГОСТ 59898), управление рисками (ПНСТ 776-2022), работу с большими данными (ГОСТ 70889, 70466, 59999), а также этические и общественные аспекты (ПНСТ 840-2023).

Направления дальнейших исследований

По мере развития сфер применения СППР на основе технологий ИИ исследовательский интерес смещается к этическим аспектам их применения, что подтверждают многочисленные исследования в данной предметной области. Расширение использования ИИ в разных сферах требует последовательного уточнения этических стандартов, механизмов управления и правовых норм [4, 29]. Дальнейшее развитие целесообразно проводить по трем направлениям: развитие объяснимых систем, междисциплинарное взаимодействие и выработка рекомендаций, снижающих этические риски автоматизированных решений.

1. Развитие объяснимых систем

По мере развития применения СППР в организациях требуется контролировать не только качество работы технологий ИИ, но и этическую приемлемость ее результатов [28]. Особое внимание уделяется объяснимому ИИ (XAI) как подходу, который повышает интерпретируе= мость и позволяет заинтересованным сторонам оценивать рекомендации системы и связанные с ними риски [19, 23].

Технологии повышения конфиденциальности (федеративное обучение и дифференциальная конфиденциальность) предлагают способы усилить защиту данных без существенной потери качества модели [32]. Федеративное обучение исключает централизацию исходных данных и обучает модель в распределенной сети, что снижает последствия возможных утечек и несанкционированного доступа [32]. Условия и границы применимости таких методов в СППР требуют дальнейшего изучения.

Еще одно направление связано с обучением с подкреплением на основе обратной связи от человека, которое позволяет согласовывать поведение системы с человеческими предпочтениями [35]. Такой подход дает возможность настраивать модель по оценкам людей, а не только по формальному целевому показателю.

2. Необходимость междисциплинарного взаимодействия

Этические затруднения, возникающие при использовании СППР, не решаются силами одних только разработчиков. Требуется участие исследователей ИИ, специалистов по этике,

юристов и представителей бизнеса [13, 28]. Междисциплинарное взаимодействие сокращает разрыв между темпом технологического развития и правовыми и этическими ограничениями, регулирующими применение данных технологий [4, 29]. Дальнейшие исследования должны рассмотреть механизмы взаимодействия разработчиков технологий, ЛПР и других пользователей, а также регуляторов, поскольку нормативная база, охватывающая все аспекты использования ИИ, все еще формируется, а работа по унификации требований к ответственности, подотчетности и прозрачности продолжается [27, 29].

Конечные пользователи и другие заинтересованные стороны также должны участвовать в формировании политик управления ИИ [28, 31]. Понимание того, как СППР влияют на реальную практику позволяет разработчикам учитывать разные точки зрения при проектировании систем [21, 24]. Отдельной задачей остается оценка результативности деятельности консультативных советов и корпоративных комитетов по управлению ИИ [22, 34].

3. Нормативное регулирование развития ИИ

Регулирование ИИ находится еще в стадии становления, поэтому последующие исследования должны рассмотреть новые регуляторные инструменты, отвечающие растущим этическим рискам [29]. Действующие модели управления не всегда успевают за изменением технологии, что повышает значимость государственных мер, стимулирующих ответственное использование ИИ [4]. Одна из ключевых задач состоит в правилах распределения ответственности, определяющих, кто отвечает за последствия решения, подготовленного системой [27, 29]. Практически важно также описать, как использование СППР влияет на результативность бизнеса и на объем юридических обязательств при неблагоприятных исходах [30]. Для заполнения пробелов в знаниях исследования должны учитывать отраслевую специфику внедрения этических принципов ИИ [28, 36].

Не менее важно международное взаимодействие в выработке политик и кодексов этики использования ИИ [4, 37]. Регулирование ИИ по своей природе выходит за национальные границы, а отсутствие согласованных международных норм ведет к расхождению этических требований. Роль международных организаций в стандартизации ИИ поэтому заслуживает отдельного анализа [37].

Наконец, требуется усилить аудит систем ИИ и контроль соответствия применимым стандартам [22, 33]. Инструменты этической оценки и независимый аудит третьей стороной снижают риск предвзятых решений и повышают доверие к организации [22]. Отдельного исследования заслуживает вопрос о том, насколько целесообразно сделать этический аудит обязательным для всех сфер, использующих системы ИИ [33].

Стратегии этичного внедрения ИИ

Этичное внедрение ИИ в СППР предполагает системный процесс, включающий участие человека и периодическую проверку моделей на предвзятость и иные нежелательные эффекты [17, 22]. Для этого необходимы ограничения, которые сдерживают применение таких СППР в границах заявленных целей, этических норм и требований законодательства [28, 29]. Можно выделить три подхода этичного внедрения ИИ в СППР в организациях (Таблица 2):

1) *Гибридное взаимодействие ЛПР и ИИ*, при котором контроль остается за ЛПР, а система выполняет вспомогательную функцию [2, 17]. Такой подход показал свою применимость в сферах здравоохранения и финансов, где решения, подготовленные ИИ, требуют профессионального подтверждения [21, 24]. Схема «ЛПР в контуре управления» снижает риск закрепления случайных ошибок и смещений, а также сохраняет персональную ответственность за результат [17].

2) *Предварительная оценка рисков и последствий внедрения*. До ввода СППР в эксплуатацию следует оценить смещения в данных, справедливость решений и ситуации, требующие вмешательства человека [22, 33]. Такая оценка проводится на этапе проектирова-

ния, чтобы предотвратить этические проблемы, которые впоследствии могут нанести потенциальный вред пользователям или заинтересованным лицам[13].

3) *Непрерывный аудит и мониторинг качества*, позволяющие отслеживать этические риски системы во времени [20, 22]. Необходимость постоянного контроля связана с тем, что модели дообучаются на новых данных и могут приобретать новые смещения [5]. На практике это означает регулярный аудит справедливости, проверку на предвзятость и отчетность о применяемых методах.

Таблица 2. Стратегии этичного внедрения ИИ

Стратегия	Описание
Гибридное взаимодействие «ЛПР – ИИ»	Контроль над решением остается за человеком, система выполняет вспомогательную функцию; применяется прежде всего в здравоохранении и финансах
Предварительная оценка рисков и последствий	До развертывания оцениваются смещения в данных, справедливость решений и потребность в человеческом вмешательстве
Непрерывный аудит и мониторинг качества	Регулярные аудиты справедливости, проверка на предвзятость и отчетность как элементы постоянного контроля этических характеристик системы

Заключение

Научная новизна исследования состоит в систематизации этических рисков систем поддержки принятия решений на основе ИИ применительно к организационным системам.

Применение технологий ИИ в СППР сопряжено с рядом этических проблем: алгоритмической предвзятостью, недостаточной интерпретируемостью, рисками для персональных данных и размыванием ответственности за принимаемые решения [12, 28]. Без принятия соответствующих мер данные проблемы препятствуют формированию доверия к системам ИИ, ведут к несправедливым решениям и создают для организациий юридические и репутационные риски [20, 29].

Снижение этих рисков требует комплекса мер:

- развитие этических стандартов, регулярная проверка моделей на предвзятость и обеспечение объяснимости решений [4, 13, 19];
- сохранение ЛПР в контуре принятия значимых решений, что позволяет контролировать и корректировать работу системы [17];
- обучение персонала, формирующее понимание возможностей и ограничений ИИ, а также применимых этических требований [34].

Законодательство о персональных данных и формирующиеся нормы регулирования ИИ создают правовую основу ответственного применения технологий ИИ [27, 29]. Однако соблюдения нормативных требований в целом еще недостаточно. Организациям необходимо последовательно реализовывать принципы справедливости, подотчетности, прозрачности и защиты конфиденциальности данных в процессах принятия решений [4, 13]. Выработка отраслевых стандартов использования ИИ предполагает участие регуляторов, исследователей, компаний и специалистов по этике [28, 38].

Практическая значимость исследования определяется возможностью использования предложенных стратегий этичного внедрения ИИ (гибридное взаимодействие ЛПР и ИИ, предварительная оценка рисков и последствий внедрения, непрерывный аудит и мониторинг качества) при проектировании и вводе в эксплуатацию СППР в организациях. Результаты работы применимы при разработке внутренних регламентов и программ алгоритмического аудита, при формировании комитетов по этике ИИ и требований к документированию принимаемых решений.

Список литературы

1. *Duan Y., Edwards J.S., Dwivedi Y.K.* Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda // International Journal of Information Management. – 2019. – Vol. 48. – P. 63–71. – DOI 10.1016/j.ijinfomgt.2019.01.021.
2. *Shrestha Y. R., Ben-Menahem S. M., von Krogh G.* Organizational decision-making structures in the age of artificial intelligence // California Management Review. – 2019. – Vol. 61, No. 4. – P. 66–83. – DOI 10.1177/0008125619862257.
3. *Матвеев А.В.* Концептуальные основы внедрения искусственного интеллекта в процессы принятия решений в организационных системах // Информатизация и связь. – 2026. – № 2. – С. 28–39. – DOI 10.34219/2078-8320-2026-16-2-28-39. – EDN OHCQAJ.
4. *Jobin A., Ienca M., Vayena E.* The global landscape of AI ethics guidelines // Nature Machine Intelligence. – 2019. – Vol. 1, No. 9. – P. 389–399. – DOI 10.1038/s42256-019-0088-2.
5. *Marabelli M., Newell S., Handunge V.* The lifecycle of algorithmic decision-making systems: Organizational choices and ethical challenges // The Journal of Strategic Information Systems. – 2021. – Vol. 30, No. 3. – Art. 101683. – DOI 10.1016/j.jsis.2021.101683.
6. *Mehrabi N., Morstatter F., Saxena N., Lerman K., Galstyan A.* A survey on bias and fairness in machine learning // ACM Computing Surveys. – 2021. – Vol. 54, No. 6. – Art. 115. – P. 1–35. – DOI 10.1145/3457607.
7. *Смирнова А.И.* Предвзятость как проблема алгоритмов ИИ: этические аспекты // Философия и общество. – 2023. – № 3(108). – С. 118–126. – DOI 10.30884/jfio/2023.03.07. – EDN VNZKTC.
8. *Rudin C.* Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead // Nature Machine Intelligence. – 2019. – Vol. 1, No. 5. – P. 206–215. – DOI 10.1038/s42256-019-0048-x.
9. *Wachter S., Mittelstadt B., Russell C.* Counterfactual explanations without opening the black box: Automated decisions and the GDPR // Harvard Journal of Law & Technology. – 2017. – Vol. 31, No. 2. – P. 841–887.
10. *Čartolovni A., Tomičić A., Mosler E.L.* Ethical, legal, and social considerations of AI-based medical decision-support tools: a scoping review // International Journal of Medical Informatics. – 2022. – Vol. 161. – Art. 104738. – DOI 10.1016/j.ijmedinf.2022.104738.
11. *Песова В.Л.* Этика и риски использования ИИ в финансовом управлении: вызовы для малого и среднего бизнеса // Экономика строительства. – 2025. – № 3. – С. 453–456. – EDN MXTBOG.
12. *Mittelstadt B.* Principles alone cannot guarantee ethical AI // Nature Machine Intelligence. – 2019. – Vol. 1, No. 11. – P. 501–507. – DOI 10.1038/s42256-019-0114-4.
13. *Floridi L., Cowls J., Beltrametti M. [et al.]* AI4People – An ethical framework for a good AI society: opportunities, risks, principles, and recommendations // Minds and Machines. – 2018. – Vol. 28, No. 4. – P. 689–707. – DOI 10.1007/s11023-018-9482-5.
14. *Klaus M.* Transcending weapon systems: the ethical challenges of AI in military decision support systems // Humanitarian Law & Policy. – 2024. – Vol. 24.
15. *Raghavan M., Barocas S., Kleinberg J., Levy K.* Mitigating bias in algorithmic hiring: Evaluating claims and practices // Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* ’20). – New York: ACM, 2020. – P. 469–481. – DOI 10.1145/3351095.3372828.
16. *Obermeyer Z., Powers B., Vogeli C., Mullainathan S.* Dissecting racial bias in an algorithm used to manage the health of populations // Science. – 2019. – Vol. 366, No. 6464. – P. 447–453. – DOI 10.1126/science.aax2342.
17. *Stefan R., Carutasu G.* How to approach ethics in intelligent decision support systems // International Symposium in Management Innovation for Sustainable Management and Entrepreneurship. – Cham: Springer International Publishing, 2019. – P. 25–40.
18. *Волобуев А.В.* Этика искусственного интеллекта, дискриминация и неравенство // Век глобализации. – 2023. – № 3. – С. 48–62. – DOI 10.30884/glob/2023.03.04.
19. *Barredo Arrieta A., Díaz-Rodríguez N., Del Ser J. [et al.]* Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI // Information Fusion. – 2020. – Vol. 58. – P. 82–115. – DOI 10.1016/j.inffus.2019.12.012.
20. *Selbst A. D., Boyd D., Friedler S. A., Venkatasubramanian S., Vertesi J.* Fairness and abstraction in sociotechnical systems // Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* ’19). – New York: ACM, 2019. – P. 59–68. – DOI 10.1145/3287560.3287598.

21. *Amann J., Blasimme A., Vayena E., Frey D., Madai V.I.* Explainability for artificial intelligence in healthcare: a multidisciplinary perspective // BMC Medical Informatics and Decision Making. – 2020. – Vol. 20. – Art. 310. – DOI 10.1186/s12911-020-01332-6.

22. *Raji I. D., Smart A., White R. N. [et al.]* Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing // Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). – New York: ACM, 2020. – P. 33–44. – DOI 10.1145/3351095.3372873.

23. *Adadi A., Berrada M.* Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI) // IEEE Access. – 2018. – Vol. 6. – P. 52138–52160. – DOI 10.1109/ACCESS.2018.2870052.

24. *Матвеев А.В.* Сравнительный анализ методов интерпретируемости моделей искусственного интеллекта в процессах принятия решений // Современные наукоемкие технологии. – 2026. – № 5. – С. 131–138. – DOI 10.17513/snt.40785. – EDN KKJORY.

25. *Ribeiro M. T., Singh S., Guestrin C.* "Why should I trust you?": Explaining the predictions of any classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – New York: ACM, 2016. – P. 1135–1144. – DOI 10.1145/2939672.2939778.

26. *Lundberg S. M., Lee S.-I.* A unified approach to interpreting model predictions // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 4765–4774.

27. *Апостолова Н.Н.* Ответственность за вред, причиненный искусственным интеллектом // Северо-Кавказский юридический вестник. – 2021. – № 1. – С. 112–119. – DOI 10.22394/2074-7306-2021-1-1-112-119.

28. *Rezaei M., Pironti M., Quaglia R.* AI in knowledge sharing, which ethical challenges are raised in decision-making processes for organisations? // Management Decision. – 2025. – Vol. 63, No. 10. – P. 3369–3388.

29. *Давыдов Е.В., Сноговская Ю.О.* Этические и правовые аспекты применения ИИ в экспертной деятельности // Право и управление. – 2026. – № 2. – С. 462–465. – DOI 10.24412/2224-9133-2026-2-462-465. – EDN YDVYFY.

30. *Mikalef P., Conboy K., Lundström J. E., Popovič A.* Thinking responsibly about responsible AI and 'the dark side' of AI // European Journal of Information Systems. – 2022. – Vol. 31, No. 3. – P. 257–268. – DOI 10.1080/0960085X.2022.2026621.

31. *Biondi G. [et al.]* Ethical design of artificial intelligence-based systems for decision making // Frontiers in Artificial Intelligence. – 2023. – Vol. 6. – Art. 1250209. – DOI 10.3389/frai.2023.1250209.

32. *McMahan H. B., Moore E., Ramage D., Hampson S., Agüera y Arcas B.* Communication-efficient learning of deep networks from decentralized data // Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). – PMLR, 2017. – Vol. 54. – P. 1273–1282.

33. *Mökander J., Axente M., Casolari F., Floridi L.* Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation // Minds and Machines. – 2022. – Vol. 32, No. 2. – P. 241–268. – DOI 10.1007/s11023-021-09577-4.

34. *Hagendorff T.* The ethics of AI ethics: An evaluation of guidelines // Minds and Machines. – 2020. – Vol. 30, No. 1. – P. 99–120. – DOI 10.1007/s11023-020-09517-8.

35. *Christiano P. F., Leike J., Brown T. B., Martic M., Legg S., Amodei D.* Deep reinforcement learning from human preferences // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 4299–4307.

36. *Берман Н.Д.* Преодоление вызовов внедрения ИИ-агентов в образовательной среде: фокус на этические аспекты и риски // ЦИТИСЭ. – 2025. – № 3(45). – С. 166–177. – EDN MFAQIU.

37. *Алборова М.Б.* § 3. Этические аспекты ИИ: Рекомендация ЮНЕСКО и иных организаций // Международная безопасность в эпоху искусственного интеллекта: Учебник для вузов. – 2-е изд., испр. и доп. – Москва: Издательство «Аспект Пресс», 2025. – С. 439–457. – EDN ZEYFVO.

38. OECD. Recommendation of the Council on Artificial Intelligence. – OECD/LEGAL/0449. – Paris: OECD, 2019.

Статья поступила в редакцию 25 февраля 2026 г.
Принята к публикации 15 июня 2026 г.

Ссылка для цитирования: Матвеев А.В. Этические риски применения систем поддержки принятия решений на основе искусственного интеллекта в организациях // Национальная безопасность и стратегическое планирование. 2026. № 2(54). С. 5-17. DOI: https://doi.org/10.37468/2307-1400-2026-2-5-17

Review article

Ethical risks of artificial intelligence-based decision support systems in organizations

Matveev Alexander V. [1]

[1]*Saint Petersburg University of the State Fire Service of the Ministry of Emergency Situations of Russia named after Hero of the Russian Federation General of the Army E.N. Zinichev, Saint Petersburg, Russia*

Abstract

This article analyzes the ethical risks arising from the use of artificial intelligence (AI)-based decision support systems in organizations. While these systems improve the speed and validity of decisions by processing large amounts of data, they also raise issues of algorithmic bias, opacity, erosion of responsibility, and privacy protection. Based on a literature review and an analysis of real-world AI case studies, four groups of issues that create ethical risks are identified: bias and fairness issues, transparency and explainability issues, accountability and responsibility in decision-making, and privacy and privacy issues. It is shown that mitigating these risks requires not only compliance with regulatory requirements but also organizational measures (ethical system design, algorithmic audit, the creation of AI ethics committees, and continuous monitoring of models after deployment). The combination of fairness, explainability, and accountability allows organizations to leverage the benefits of AI in decision-making without deviating from ethical standards. The scientific novelty of this study lies in its systematization of the ethical risks of AI-based decision support systems as applied to organizations. The practical significance of this study lies in the potential use of the proposed strategies for the ethical implementation of AI in the design and implementation of decision support systems in organizations.

Keywords: artificial intelligence, decision support systems, ethical risks, transparency, bias, confidentiality, black box.

References

1. *Duan Y., Edwards J. S., Dwivedi Y. K.* Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda // International Journal of Information Management. – 2019. – Vol. 48. – P. 63–71. – DOI 10.1016/j.ijinfomgt.2019.01.021.

2. *Shrestha Y. R., Ben-Menahem S. M., von Krogh G.* Organizational decision-making structures in the age of artificial intelligence // California Management Review. – 2019. – Vol. 61, No. 4. – P. 66–83. – DOI 10.1177/0008125619862257.

3. *Matveev A.V.* Conceptual Foundations of Implementing Artificial Intelligence in Decision-Making Processes in Organizational Systems // Informatization and Communication. - 2026. - No. 2. - Pp. 28–39. - DOI 10.34219/2078-8320-2026-16-2-28-39. - EDN OHCQAJ.

4. *Jobin A., Ienca M., Vayena E.* The global landscape of AI ethics guidelines // Nature Machine Intelligence. – 2019. – Vol. 1, No. 9. – P. 389–399. – DOI 10.1038/s42256-019-0088-2.

5. *Marabelli M., Newell S., Handunge V.* The lifecycle of algorithmic decision-making systems: Organizational choices and ethical challenges // The Journal of Strategic Information Systems. – 2021. – Vol. 30, No. 3. – Art. 101683. – DOI 10.1016/j.jsis.2021.101683.

6. *Mehrabi N., Morstatter F., Saxena N., Lerman K., Galstyan A.* A survey on bias and fairness in machine learning // ACM Computing Surveys. – 2021. – Vol. 54, No. 6. – Art. 115. – P. 1–35. – DOI 10.1145/3457607.

7. *Smirnova A.I.* Bias as a problem in AI algorithms: ethical aspects // Philosophy and Society. – 2023. – No. 3(108). – P. 118–126. – DOI 10.30884/jfio/2023.03.07. – EDN VNZKTC.

8. *Rudin C.* Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead // Nature Machine Intelligence. – 2019. – Vol. 1, No. 5. – P. 206–215. – DOI 10.1038/s42256-019-0048-x.

9. *Wachter S., Mittelstadt B., Russell C.* Counterfactual explanations without opening the black box: Automated decisions and the GDPR // Harvard Journal of Law & Technology. – 2017. – Vol. 31, No. 2. – P. 841–887.

10. *Čartolovni A., Tomičić A., Mosler E. L.* Ethical, legal, and social considerations of AI-based medical decision-support tools: a scoping review // International Journal of Medical Informatics. – 2022. – Vol. 161. – Art. 104738. – DOI 10.1016/j.ijmedinf.2022.104738.

11. *Pesova V.L.* Ethics and risks of using AI in financial management: Challenges for small and medium businesses // Construction Economics. - 2025. - No. 3. - Pp. 453-456. - EDN MXTBOG.

12. *Mittelstadt B.* Principles alone cannot guarantee ethical AI // Nature Machine Intelligence. – 2019. – Vol. 1, No. 11. – P. 501–507. – DOI 10.1038/s42256-019-0114-4.

13. *Floridi L., Cowls J., Beltrametti M. [et al.]* AI4People – An ethical framework for a good AI society: opportunities, risks, principles, and recommendations // Minds and Machines. – 2018. – Vol. 28, No. 4. – P. 689–707. – DOI 10.1007/s11023-018-9482-5.

14. *Klaus M.* Transcending weapon systems: the ethical challenges of AI in military decision support systems // Humanitarian Law & Policy. – 2024. – Vol. 24.

15. *Raghavan M., Barocas S., Kleinberg J., Levy K.* Mitigating bias in algorithmic hiring: Evaluating claims and practices // Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* ’20). – New York: ACM, 2020. – P. 469–481. – DOI 10.1145/3351095.3372828.

16. *Obermeyer Z., Powers B., Vogeli C., Mullainathan S.* Dissecting racial bias in an algorithm used to manage the health of populations // Science. – 2019. – Vol. 366, No. 6464. – P. 447–453. – DOI 10.1126/science.aax2342.

17. *Stefan R., Carutasu G.* How to approach ethics in intelligent decision support systems // International Symposium in Management Innovation for Sustainable Management and Entrepreneurship. – Cham: Springer International Publishing, 2019. – P. 25–40.

18. *Volobuev A.V.* Ethics of artificial intelligence, discrimination and inequality // Century of globalization. - 2023. - No. 3. - P. 48–62. – DOI 10.30884/glob/2023.03.04.

19. *Barredo Arrieta A., Díaz-Rodríguez N., Del Ser J. [et al.]* Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI // Information Fusion. – 2020. – Vol. 58. – P. 82–115. – DOI 10.1016/j.inffus.2019.12.012.

20. *Selbst A. D., Boyd D., Friedler S. A., Venkatasubramanian S., Vertesi J.* Fairness and abstraction in sociotechnical systems // Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* ’19). – New York: ACM, 2019. – P. 59–68. – DOI 10.1145/3287560.3287598.

21. *Amann J., Blasimme A., Vayena E., Frey D., Madai V.I.* Explainability for artificial intelligence in healthcare: a multidisciplinary perspective // BMC Medical Informatics and Decision Making. – 2020. – Vol. 20. – Art. 310. – DOI 10.1186/s12911-020-01332-6.

22. *Raji I.D., Smart A., White R.N. [et al.]* Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing // Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* ’20). – New York: ACM, 2020. – P. 33–44. – DOI 10.1145/3351095.3372873.

23. *Adadi A., Berrada M*. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI) // IEEE Access. – 2018. – Vol. 6. – P. 52138–52160. – DOI 10.1109/ACCESS.2018.2870052.

24. *Matveev A.V.* Comparative analysis of methods for interpretability of artificial intelligence models in decision-making processes // Modern science-intensive technologies. - 2026. - No. 5. - P. 131–138. - DOI 10.17513/snt.40785. - EDN KKJORY.

25. *Ribeiro M.T., Singh S., Guestrin C.* “Why should I trust you?”: Explaining the predictions of any classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – New York: ACM, 2016. – P. 1135–1144. – DOI 10.1145/2939672.2939778.

26. *Lundberg S.M., Lee S.-I.* A unified approach to interpreting model predictions // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 4765–4774.

27. *Apostolova N.N.* Liability for harm caused by artificial intelligence // North Caucasian Legal Bulletin. – 2021. – No. 1. – P. 112–119. – DOI 10.22394/2074-7306-2021-1-1-112-119.

28. *Rezaei M., Pironti M., Quaglia R.* AI in knowledge sharing, which ethical challenges are raised in decision-making processes for organisations? // Management Decision. – 2025. – Vol. 63, No. 10. – P. 3369–3388.

29. *Davydov E.V., Snogovskaya Yu.O*. Ethical and legal aspects of using AI in expert activities // Law and Management. – 2026. – No. 2. – P. 462–465. – DOI 10.24412/2224-9133-2026-2-462-465. – EDN YDVYFY.

30. *Mikalef P., Conboy K., Lundström J. E., Popovič A*. Thinking responsibly about responsible AI and ‘the dark side’ of AI // European Journal of Information Systems. – 2022. – Vol. 31, No. 3. – P. 257–268. – DOI 10.1080/0960085X.2022.2026621.

31. *Biondi G. [et al.]* Ethical design of artificial intelligence-based systems for decision making // Frontiers in Artificial Intelligence. – 2023. – Vol. 6. – Art. 1250209. – DOI 10.3389/frai.2023.1250209.

32. *McMahan H. B., Moore E., Ramage D., Hampson S., Agüera y Arcas B.* Communication-efficient learning of deep networks from decentralized data // Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). – PMLR, 2017. – Vol. 54. – P. 1273–1282.

33. *Mökander J., Axente M., Casolari F., Floridi L.* Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation // Minds and Machines. – 2022. – Vol. 32, No. 2. – P. 241–268. – DOI 10.1007/s11023-021-09577-4.

34. *Hagendorff T.* The ethics of AI ethics: An evaluation of guidelines // Minds and Machines. – 2020. – Vol. 30, No. 1. – P. 99–120. – DOI 10.1007/s11023-020-09517-8.

35. *Christiano P. F., Leike J., Brown T. B., Martic M., Legg S., Amodei D.* Deep reinforcement learning from human preferences // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 4299–4307.

36. *Berman N.D.* Overcoming the Challenges of Implementing AI Agents in the Educational Environment: Focus on Ethical Aspects and Risks // CITISE. - 2025. – No. 3 (45). – Pp. 166-177. – EDN MFAQIU.

37. *Alborova, M.B*. § 3. Ethical Aspects of AI: Recommendation of UNESCO and Other Organizations // International Security in the Age of Artificial Intelligence: Textbook for Universities. – 2nd ed., corrected and enlarged. – Moscow: Aspect Press Publishing House, 2025. – Pp. 439-457. – EDN ZEYFVO.

38. OECD. Recommendation of the Council on Artificial Intelligence. – OECD/LEGAL/0449. – Paris: OECD, 2019.

For citation: Matveev A.V. Ethical risks of artificial intelligence-based decision support systems in organizations // National security and strategic planning. 2026. № 2(54). pp. 5-17. DOI: https://doi.org/10.37468/2307-1400-2026-2-5-17

Информация об авторах:

Матвеев Александр Владимирович – кандидат технических наук, доцент, заведующий кафедрой прикладной математики и информационных технологий, Санкт-Петербургский университет Государственной противопожарной службы МЧС России имени Героя Российской Федерации генерала армии Е.Н. Зиничева, г. Санкт-Петербург, Россия
ORCID: https://orcid.org/ 0000-0002-0778-3218
SPIN-код: 5778-8832
e-mail: fcvega_10@mail.ru

Information about authors:

Matveev Alexander V. – Candidate of Technical Sciences, Associate Professor, Head of the department of applied mathematics and information technologies, St. Petersburg University of the State Fire Service of the Ministry of Emergency Situations of Russia named after Hero of the Russian Federation, Army General E.N. Zinichev, St. Petersburg, Russian Federation
ORCID: https://orcid.org/ 0000-0002-0778-3218
SPIN: 5778-8832
e-mail: fcvega_10@mail.ru