EDN: **QZZMPN** **УДК 338.24:004.8**

Парадокс алгоритмической легитимности и архитектура ответственности: как избежать ловушек искусственного интеллекта в управлении

Лаврова Елена Викторовна
кандидат экономических наук, доцент, доцент кафедры менеджмента и естественно-научных дисциплин ФГБОУ ВО «Смоленский университет спорта», г. Смоленск, Российская Федерация
e-mail: **e.v.lavrova@list.ru**
SPIN-код РИНЦ: 2964-4612
ORCID ID: 0000-0002-2824-1127

Аннотация

В статье исследуется парадокс алгоритмической легитимности – противоречие между ростом доверия к искусственному интеллекту в принятии управленческих решений и размыванием ответственности за последствия таких решений. Охарактеризованы типичные сценарии управленческих сбоев при использовании алгоритмических систем в торговле, финансах, управлении персоналом, логистике, государственном управлении и раскрыты причины размывания ответственности: эффект автоматизации, непрозрачность алгоритмических моделей и отсутствие институциональных механизмов контроля. Описана модель «архитектуры ответственности», включающая технический, институциональный, юридический, организационный и образовательный уровни контроля. Предложена классификация систем искусственного интеллекта по уровню автономии и определены соответствующие механизмы контроля, сделан вывод о необходимости перехода от декларативных этических принципов к механизмам распределенного контроля.

Ключевые слова

• искусственный интеллект • управленческие решения • алгоритмическая легитимность • архитектура ответственности • алгоритмический аудит • распределенный контроль •

Введение

Стремительное проникновение технологий искусственного интеллекта (далее – ИИ) в сферу государственного и корпоративного управления породило принципиально новую ситуацию – переход к алгоритмическому способу принятия решений [8]. Логика внедрения технологий ИИ представляется безусловно привлекательной: алгоритмы свободны от человеческих слабостей – усталости, эмоций, когнитивных искажений, корыстных мотивов, они обещают точность, скорость и масштабируемость, недоступные человеку. Исследования показывают, что организации, активно внедряющие ИИ в управленческие процессы, демонстрируют существенно более высокую эффективность операционных решений по сравнению с конкурентами, полагающимися исключительно на человеческий опыт [9]. Однако практика внедрения выявляет наличие глубинного противоречия, которое мы предлагаем обозначить термином «парадокс алгоритмической легитимности».

Суть данного парадокса состоит в следующем. С одной стороны, алгоритмы приобретают легитимность именно благодаря своей кажущейся объективности: лица, принимающие решения, склонны доверять «математической беспристрастности» больше, чем человеческому суждению, подверженному ошибкам и влияниям. С другой стороны, эта же алгоритмическая опосредованность принятия решений разрушает традиционные механизмы ответственности: когда ошибочное решение принимает менеджер, существует понятный механизм привлечения к ответственности, обжалования, компенсации ущерба. Когда же решение принимает (или существенно влияет на принятие решения) «черный ящик» алгоритма, механизм ответственности оказывается размытым [16]. Разработчик может утверждать, что алгоритм работал в соответствии с техническим заданием, оператор – что следовал рекомендации системы, менеджер – что полагался на валидированную технологию, а совет директоров – что утвердил стратегию цифровой трансформации. В результате возникает ситуация, которую можно охарактеризовать как «ответственность, растворенная в коде».

Особенность современного этапа развития социально-экономических систем заключается в том, что ИИ перестал быть экзотическим инструментом и превратился в рутинный элемент управления. Сегодня системы, основанные на машинном обучении и больших данных, участвуют в управленческих процессах повсеместно: в розничной торговле и промышленных закупках, в подборе персонала и прогнозировании спроса, в распределении социальных выплат и кадровых решениях, в логистике и финансах [5]. В розничной торговле алгоритмы определяют, какие товары закупать, по какой цене и в каких объемах, автоматически корректируя ассортимент в режиме реального времени [17]. В логистике системы маршрутизации на основе ИИ ежедневно принимают решения, от которых зависит своевременность поставок и транспортные издержки [14]. В финансовом секторе алгоритмы не только оценивают кредитоспособность заемщиков, но и управляют инвестиционными портфелями, совершая тысячи операций в секунду [3]. В управлении персоналом автоматизированные системы отбора кандидатов отклоняют резюме. В государственном управлении алгоритмы распределяют бюджетные средства, определяют очередность предоставления услуг, оценивают эффективность работы учреждений и др. [13].

Актуальность нашего исследования определяется не только теоретической значимостью описанного парадокса, но и острой практической потребностью в разработке механизмов, позволяющих избежать его разрушительных последствий. Последствия алгоритмических сбоев (от дискриминации при найме до ошибочного лишения социальных выплат) измеряются не только финансовыми потерями, но и разрушением человеческих судеб, репутационными издержками, снижением общественного доверия к институтам власти. При этом существующие этические кодексы и декларации принципов в сфере ИИ остаются

преимущественно декларативными и не предлагают работающих механизмов распределения ответственности [10]. Корпоративные политики использования ИИ зачастую сводятся к перекладыванию рисков на плечи сотрудников, вынужденных либо доверять алгоритму, либо нести персональную ответственность за его оспаривание.

Новизна предлагаемого нами подхода заключается в переходе от постановки проблемы алгоритмической подотчетности к проектированию конкретной «архитектуры ответственности» – системы взаимосвязанных мер технического, институционального, юридического, организационного и образовательного характера, обеспечивающих возможность контроля и оспаривания решений, принятых при участии ИИ. В отличие от работ, сосредоточенных исключительно на технических аспектах объяснимого ИИ или, напротив, на общих этических принципах, настоящее исследование предлагает интегративный подход, учитывающий специфику различных сфер управления и различные уровни автономии алгоритмов.

Цель нашего исследования состоит в разработке системной модели предотвращения негативных последствий внедрения ИИ в управленческие процессы, основанной на анализе резонансных инцидентов и синтезе существующих регуляторных и технических решений. Для достижения поставленной цели необходимо решить следующие задачи:

— во-первых, выявить типичные управленческие сбои при использовании ИИ в розничной торговле, логистике, финансах, управлении персоналом и государственном управлении, а также установить закономерности размывания ответственности в человеко-машинных системах управления;
— во-вторых, разработать классификацию систем ИИ по степени автономии в принятии управленческих решений с выделением соответствующих каждому уровню рисков и механизмов контроля;
— в-третьих, предложить многоуровневую модель «архитектуры ответственности», объединяющую технические, институциональные, юридические, организационные и образовательные механизмы предотвращения негативных последствий при внедрении систем ИИ в управленческие процессы.

Теоретико-методологические основы исследования

Настоящее исследование опирается на междисциплинарный подход, объединяющий теорию принятия управленческих решений, социологию технологий, правовое регулирование ИИ и методы, сложившиеся в области человеко-компьютерного взаимодействия.

Теоретическим фундаментом исследования выступает концепция распределенного познания применительно к анализу сложных социотехнических систем. Согласно данной концепции, когнитивные процессы, включая принятие решений, не локализованы исключительно в индивидуальном сознании, а распределены между образующими единую систему людьми и технологиями [18]. В современном управлении в качестве такой системы выступает связка «менеджер – алгоритм – данные – организационные процедуры», где когнитивная нагрузка и ответственность за решение распределены между элементами. Понимание управления как распределенного когнитивного процесса позволяет избежать двух крайностей: как технодетерминизма, приписывающего алгоритмам способность к самостоятельному действию, так и антропоцентризма, игнорирующего трансформацию управленческих практик под влиянием технологий.

Основным методом исследования является качественный анализ конкретных ситуаций, позволяющий изучать сложные, контекстно-обусловленные явления, каковыми и являются инциденты, связанные с внедрением ИИ в управление.

Выбор данного метода обусловлен необходимостью понимания не только фактической стороны событий, но и механизмов принятия решений, институционального контекста, поведенческих реакций участников [15].

Отбор ситуаций для анализа осуществлялся по следующим критериям: наличие документально подтвержденной информации об инциденте, включая судебные решения, официальные расследования, корпоративные отчеты; релевантность теме управления, т.е. решение должно быть принято в рамках управленческого процесса, будь то государственное, корпоративное или операционное управление; разнообразие сфер применения, чтобы отобранные ситуации охватывали розничную торговлю, финансы, управление персоналом, логистику, государственное управление; наличие измеримых последствий, будь то финансовые, репутационные или социальные; репрезентативность для российского и зарубежного контекста.

Дополнительным методом выступил анализ нормативно-правовых документов, корпоративных политик использования ИИ и этических кодексов [4], позволивший выявить существующие подходы к распределению ответственности и их ограничения.

Типичные управленческие сбои и размывание ответственности при использовании ИИ

Проведенный анализ позволяет выделить несколько типичных сценариев возникновения управленческих сбоев при использовании систем ИИ. Важно подчеркнуть, что в большинстве случаев речь идет не о технических ошибках в узком смысле слова, а о системных эффектах взаимодействия человеческих и технологических компонентов управления, когда ошибки проектирования, неполнота данных, неучтенные факторы или когнитивные ловушки лиц, принимающих решения, приводят к масштабным негативным последствиям.

Сфера розничной торговли и управления цепочками поставок является одной из наиболее насыщенных технологиями автоматизации, поскольку здесь алгоритмы управляют закупками, ценообразованием, прогнозированием спроса, персонализацией предложений и распределением товаров по магазинам. Однако интенсивная автоматизация несет и специфические риски, которые наглядно проявились в инциденте, произошедшем в 2023-2024 гг. в крупнейшей розничной сети Walmart. Система автоматизированного управления закупками, основанная на алгоритмах машинного обучения, анализирующая исторические данные о продажах, сезонность, промо-активности и макроэкономические индикаторы, начала генерировать ошибочные прогнозы спроса на товары длительного хранения. Проведенное внутреннее расследование установило, что проблема возникла из-за неучтенного алгоритмом изменения потребительского поведения после пика инфляции: модель, обученная на данных 2019-2022 гг., опиралась в своих прогнозах на закономерности пандемийного и постпандемийного спроса, тогда как реальное поведение потребителей сместилось в сторону экономии и сокращения запасов. В результате система, ориентируясь на исторические данные о высоком спросе, сформировала заказы на закупку ряда категорий товаров в объемах, превышающих реальные потребности в изменившихся условиях, что вызвало затоваривание складов [17]. Ключевая управленческая проблема данного инцидента заключалась в том, что менеджеры, получая рекомендации системы, в большинстве случаев утверждали их без дополнительной проверки, полагаясь на авторитет алгоритма и не имея достаточных стимулов для критической оценки. Лишь немногие сотрудники, обладающие развитой рыночной интуицией и многолетним опытом, пытались оспаривать рекомендации, однако процедура оспаривания была излишне бюрократизирована и требовала многоуровневых согласований, что фактически

делало ее неприменимой в операционном режиме. Ответственность за сбой оказалась размытой между разработчиками системы, указывавшими на качество предоставленных для обучения данных, менеджерами, ссылавшимися на неукоснительное следование утвержденному регламенту, и высшим руководством, делегировавшим полномочия по принятию решений алгоритмической системе без создания адекватных механизмов контроля.

Не менее показательные примеры можно обнаружить в сфере управления персоналом и кадровых решений, где алгоритмы все чаще используются для первичного отбора кандидатов, оценки их соответствия корпоративной культуре и даже прогнозирования успешности на конкретной должности. В 2018 г. компания Amazon оказалась в центре скандала, связанного с работой автоматизированной системы отбора резюме. Разработанный специалистами компании алгоритм, обученный на десятилетии кадровых данных, при отборе на технические должности демонстрировал устойчивое предпочтение кандидатов-мужчин, систематически занижая рейтинг резюме, содержавших указание на принадлежность к женскому полу – например, упоминание об участии в женских профессиональных сообществах. Как выяснили разработчики, алгоритм неосознанно воспроизвел исторические закономерности найма, существовавшие в компании и в целом в технологической индустрии, где мужчины традиционно были представлены значительно шире. Несмотря на то, что система не использовала признак пола в явном виде, она научилась выявлять коррелирующие с ним лингвистические и контекстуальные признаки. Компания предприняла попытку доработать алгоритм, однако полностью устранить дискриминационный эффект не удалось, и в итоге проект был закрыт [6]. Управленческий аспект этого инцидента состоит в том, что исторические данные, на которых обучаются алгоритмы, никогда не являются нейтральными: они всегда несут в себе следы прошлых управленческих практик, включая их предрассудки, ошибки и системные ограничения. Ответственность за воспроизводство этих предрассудков ложится на руководителей, принимающих решение о внедрении системы, даже если сами алгоритмические модели формально нейтральны.

В финансовом секторе и сфере кредитования, исторически являющемся пионером внедрения алгоритмических систем, также накоплен значительный опыт как успехов, так и провалов. Один из наиболее резонансных инцидентов, получивших широкое общественное обсуждение, связан с алгоритмом оценки кредитоспособности, используемым Apple Card – совместным продуктом Apple и Goldman Sachs. В 2019 г. предприниматель Дэвид Хейнсмиерс обратил внимание в социальных сетях на то, что алгоритм предоставил ему кредитный лимит в 20 раз превышающий лимит его супруги несмотря на то, что они подавали совместную налоговую декларацию. Проведенное регулятором штата Нью-Йорк расследование не выявило прямых доказательств намеренной дискриминации, однако подтвердило, что алгоритм, обученный на исторических данных о кредитном поведении, воспроизвел существующие в обществе диспропорции [21]. Управленческий аспект данной ситуации заключается в том, что решение о кредитном лимите фактически принималось алгоритмом без участия человека. Сотрудники банка и Apple не имели возможности вмешаться в процесс, а клиенты не могли получить внятного объяснения, почему решение принято именно так. Ответственность за дискриминационные последствия оказалась размытой: Goldman Sachs согласился выплатить компенсации и изменить алгоритм, но ни одно должностное лицо не понесло персональной ответственности. Более того, сама конструкция продукта была выстроена так, чтобы минимизировать участие человека в принятии решений, что рассматривалось как конкурентное преимущество.

В российской практике показателен опыт использования алгоритмических рекомендательных систем в управлении инвестиционными портфелями. Сервисы автоматизированного инвестирования активно продвигают роботизиро-

ванное управление сбережениями – формирование инвестиционных рекомендаций на основе алгоритмов, учитывающих риск-профиль клиента и рыночную ситуацию. Однако в 2022 г., после кардинального изменения макроэкономической ситуации, многие клиенты столкнулись с тем, что алгоритмы продолжали рекомендовать стратегии, основанные на исторических данных, не учитывающих новые геополитические и экономические реалии [3]. С правовой и управленческой точек зрения интересен вопрос о распределении ответственности. В тексте пользовательских соглашений, которые клиенты финансовых организаций подтверждают при начале обслуживания, как правило, указывается, что все выдаваемые системой рекомендации следует воспринимать исключительно в качестве информации, не являющейся основанием для принятия обязательных к исполнению решений. К данному положению обычно добавляется разъяснение о том, что всю полноту ответственности за те инвестиционные решения, которые в конечном итоге будут приняты человеком, несет исключительно сам клиент, а не организация, предоставляющая соответствующий сервис. Тем не менее на практике возникает достаточно серьезное противоречие, которое обусловлено тем, что значительная часть пользователей склонна воспринимать алгоритмические рекомендации в качестве профессионального руководства, требующего безусловного выполнения. Именно вследствие такого доверия и формируется разрыв между реальными ожиданиями людей и фактической конструкцией ответственности, закрепленной в официальных документах.

Рассматривая сферу государственного и муниципального управления, необходимо отметить, что в данной области цена любой ошибки, допущенной алгоритмической системой, приобретает особенно высокое значение. Это объясняется тем, что принимаемые здесь решения затрагивают не только финансовые интересы граждан, но и их базовые права, возможность получения доступа к социальным благам, а также защиту от возможных необоснованных действий со стороны уполномоченных лиц. Драматичный пример последних лет, который демонстрирует опасность подобных ситуаций, представляет собой так называемое «дело о детском пособии» в Нидерландах, приведшее в 2021 г. к правительственному кризису и последовавшей за ним отставке кабинета министров. Суть данного инцидента заключалась в том, что налоговые органы Нидерландов задействовали алгоритмическую систему для выявления случаев мошенничества, связанных с получением детских пособий. Разработанный алгоритм, процесс обучения которого осуществлялся на основе исторических данных, в ходе своей работы начал демонстрировать устойчивую предвзятость по отношению к гражданам, обладающим двойным гражданством, имеющим невысокий уровень доходов либо проживающим в определенных районах. Тысячи семей в результате такой работы системы были без достаточных оснований обвинены в мошеннических действиях и вынуждены осуществить возврат значительных денежных сумм, достигавших десятков тысяч евро, причем многие из этих людей оказались на грани банкротства и утраты права на жилье. Сам алгоритм функционировал по принципу «черного ящика»: обычные граждане не имели возможности понять, на каком именно основании они были отнесены к категории нарушителей, а сотрудники ведомств, получавшие готовые уведомления, сформированные системой, утверждали их в автоматическом режиме, не располагая при этом ни реальными возможностями, ни достаточной мотивацией для осуществления дополнительной проверки [19].

Значимые уроки можно извлечь также из опыта внедрения алгоритмических систем, накопленного в российской практике государственного управления. В качестве примера можно привести процессы автоматизации предоставления государственных услуг, реализованные через портал «Госуслуги», в ходе которых неоднократно возникали ситуации, когда алгоритмы, определяющие очередность оказания услуг, функционировали недостаточно прозрачно для граждан. Так, автоматизированные системы записи на прием к врачу, изначально

предназначенные для оптимизации нагрузки на медицинские учреждения, фактически создают неравные условия доступа для различных категорий граждан, например, в отношении пожилых людей, которые испытывают объективные трудности при использовании цифровых интерфейсов, либо в связи с проблемами нестабильного интернет-соединения [12]. Хотя указанные инциденты не получили столь широкого общественного резонанса, какой сопровождал нидерландское дело, они подтверждают ряд устойчивых закономерностей: размывание ответственности между разработчиками, менеджерами и операторами систем ИИ, эффект автоматизации (некритическое доверие к алгоритмическим рекомендациям), воспроизводство исторических предрассудков через обучающие данные, а также – вне зависимости от страны – исключение уязвимых категорий граждан из полноценного пользования услугами при отсутствии действенных механизмов контроля.

Особого внимания в рамках рассматриваемой проблематики заслуживает сфера логистики и управления транспортом, поскольку в этой области алгоритмы принимают решения, оказывающие непосредственное влияние на безопасность людей и сохранность перевозимых грузов. В качестве примера можно привести ситуацию, сложившуюся в 2022 г. в одной из крупных российских логистических компаний, которая столкнулась с системным сбоем при внедрении алгоритмической системы, предназначенной для оптимизации маршрутов доставки. Разработанная система, построенная на основе моделей машинного обучения, предлагала такие маршруты, которые позволяли минимизировать общий пробег транспортных средств и сократить время нахождения в пути, однако при этом она не учитывала ряд важных факторов: сезонные ограничения проходимости дорог, график работы конкретных пунктов приема грузов, а также сложившиеся предпочтения постоянных клиентов относительно времени доставки. В результате такого формального подхода маршруты, которые система определяла как оптимальные, на практике приводили к систематическим срывам согласованных сроков доставки, к увеличению фактического расхода топлива, обусловленного неучтенными дорожными условиями, и к росту недовольства со стороны клиентов, рассчитывавших на своевременное получение заказов. Водители, получая задания от системы, оказались в двойственном положении: следование рекомендациям алгоритма приводило к срыву показателей, а отклонение от них требовало дополнительных согласований и создавало риск дисциплинарной ответственности [14].

В сфере маркетинга и управления взаимоотношениями с клиентами алгоритмы активно используются для персонализации предложений, сегментации аудитории, прогнозирования оттока клиентов. Однако и здесь накоплен опыт ошибок, связанных с неучтенными факторами. Известны случаи, когда алгоритмы, оптимизирующие маркетинговые коммуникации, демонстрировали нечувствительность к социальному контексту – например, направляли рекламу товаров для беременных женщинам, пережившим потерю ребенка, или предлагали кредитные продукты людям, находящимся в процедуре банкротства. Такие инциденты наносят репутационный ущерб компаниям и подрывают доверие клиентов, при этом ответственность за них также оказывается размытой между разработчиками алгоритмов, маркетологами, формирующими критерии таргетинга, и менеджерами, утверждающими коммуникационные стратегии [2].

Таким образом, во всех рассмотренных сферах – от розничной торговли до государственного управления – проявляется общая проблема отсутствия чёткого распределения ответственности, когда и корпоративные политики, и пользовательские соглашения, и этические кодексы в сфере ИИ на практике либо перекладывают риски на конечного пользователя, либо размывают ответственность между множеством участников, не создавая действенных механизмов контроля.

Уровни автономии и риски: классификация систем ИИ в управлении

Рассмотрение конкретных инцидентов, связанных с применением алгоритмических систем, приводит к выводу о том, что характер возникающих управленческих рисков, равно как и набор необходимых механизмов контроля, обнаруживает существенные различия в зависимости от того, какая степень автономии была предоставлена системе ИИ. Систематизация указанных различий по уровням автономии алгоритмов в принятии управленческих решений представлена в табл. 1.

Таблица 1. Уровни автономии систем ИИ и соответствующие механизмы контроля

Уровень автономии	Характеристика	Типичные сферы применения	Риски	Механизмы контроля
Информационно-справочный	Предоставление структурированных данных	Бизнес-аналитика, дашборды показателей	Искажение данных, неполнота	Аудит источников, верификация методик сбора
Рекомендательный	Предложение вариантов с оценкой эффективности	Кредитный скоринг, подбор персонала, маркетинг	Эффект автоматизации, неявная предвзятость	Объяснимость моделей, право на оспаривание, обучение пользователей
Полуавтономный	Автоматическое принятие решений с возможностью вмешательства	Автоматизация закупок, маршрутизация, динамическое ценообразование	Размывание ответственности, запаздывание вмешательства	Регулярный аудит решений, пороговые значения для ручного контроля
Полностью автономный	Самостоятельное принятие и исполнение решений	Алгоритмический трейдинг, управление запасами	Системные риски, необратимость последствий	Страхование, резервные протоколы, лимитирование полномочий

Первый уровень автономии ИИ – информационно-справочные системы, которые предоставляют лицу, принимающему решение, структурированную информацию, но не предлагают готовых рекомендаций. На этом уровне алгоритм выступает как инструмент аналитики, расширяющий возможности человека, но не вторгающийся в собственно принятие решения. Риски здесь минимальны и связаны преимущественно с качеством и полнотой данных [11].

Второй уровень – рекомендательные системы, которые предлагают лицу, принимающему решение, готовые варианты с оценкой их предполагаемой эффективности. Человек может принять или отвергнуть рекомендацию, но сама конструкция системы, как правило, подталкивает к следованию алгоритмическому совету. Риски на этом уровне существенно выше, поскольку срабатывает эффект автоматизации – склонность человека некритически доверять рекомендациям, особенно если система оформлена как «экспертная» [7].

Третий уровень – полуавтономные системы, которые принимают решения автоматически в рамках заданных параметров, но предусматривают возможность человеческого вмешательства в отдельных случаях. Такие системы широко используются в управлении цепочками поставок, динамическом ценообразовании, автоматизации закупок. Риски на данном уровне связаны с тем, что вмешательство может быть запаздывающим или невозможным по техническим

причинам, а также с тем, что параметры, в рамках которых система действует автономно, могут быть заданы некорректно [20].

Четвертый уровень – полностью автономные системы, которые принимают и исполняют решения без участия человека. Наиболее характерный пример – алгоритмический трейдинг, где решения о покупке и продаже ценных бумаг принимаются за миллисекунды. Здесь риски максимальны и могут приобретать системный характер.

Предложенная классификация позволяет ранжировать риски управления: от минимальных на информационно-справочном уровне до максимальных на полностью автономном. Повышение автономии ИИ влечет за собой не только технические риски, но и размывание ответственности, а также ослабление человеческого контроля, следовательно, для каждого уровня автономии необходим адекватный набор механизмов контроля – не избыточный для низких уровней, но достаточный для высоких.

Архитектура ответственности: модель и практические рекомендации для руководителей

Обобщение результатов анализа инцидентов и существующих регуляторных механизмов позволяет предложить модель «архитектуры ответственности», включающую несколько взаимосвязанных уровней, каждый из которых должен быть проработан при внедрении систем ИИ в управленческие процессы (рис. 1).

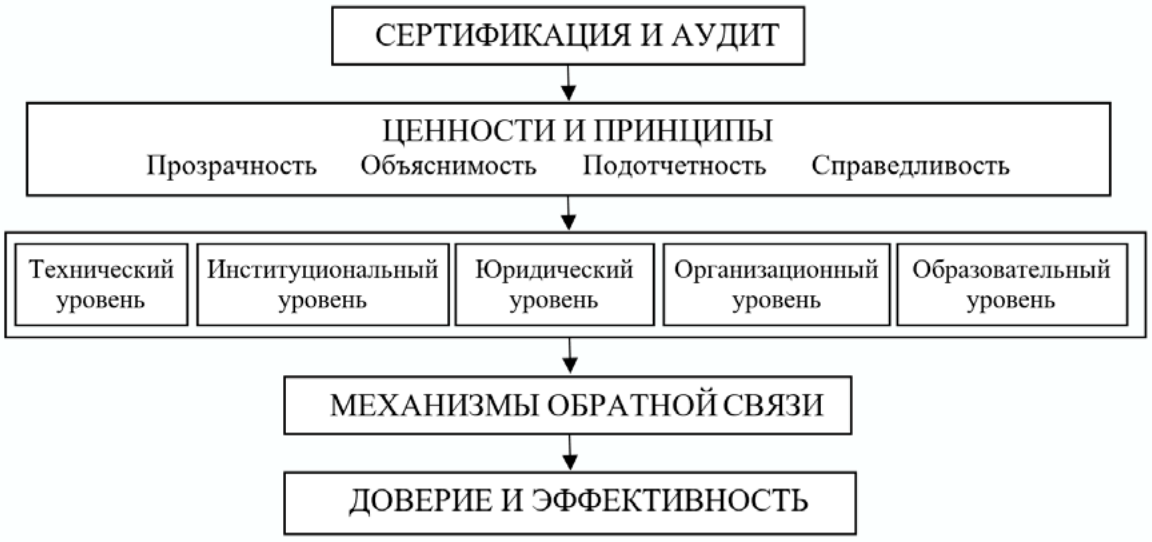

Рисунок 1. Архитектура ответственности: многоуровневая система контроля при использовании ИИ в управлении

Технический уровень архитектуры ответственности предполагает внедрение стандартов объяснимого ИИ как обязательного требования для систем, используемых в управлении. Речь идет о том, что алгоритм должен не просто выдавать решение, но и предоставлять лицу, принимающему решение, понятное обоснование – какие именно факторы и в какой степени повлияли на результат. При этом важно, чтобы объяснения были не только технически корректными, но и практически полезными для менеджера, т.е. давали ему возможность оценить обоснованность рекомендации и принять информированное решение о доверии или оспаривании.

Институциональный уровень архитектуры ответственности предполагает встраивание человека в контур управления с четко прописанными процедурами и ответственностью. Простое наличие человека, формально утверждающего решения алгоритма, недостаточно – необходимы процедуры, стимулирующие реальную проверку, а не автоматическое одобрение. Практика показывает, что эффективным инструментом может стать обязательная регистрация факта проверки и возможность квалифицированного оспаривания. При этом важно, чтобы случаи оспаривания не рассматривались как отклонение от нормы, а, напротив, становились ценным материалом для обучения и совершенствования системы.

Юридический уровень архитектуры ответственности требует проработки вопросов распределения ответственности между разработчиками, операторами, руководителями и организациями в целом. В мировой практике наметилось несколько подходов к решению этой проблемы. Один из них – страхование алгоритмических рисков, когда ответственность за возможный ущерб перераспределяется через механизмы страхования. Другой подход – создание прецедентного права, в котором ответственность распределяется пропорционально вкладу каждой стороны в возникновение ущерба [1]. Представляется перспективным подход, аналогичный регулированию в авиации: пилот несет ответственность за принятие решения в конкретной ситуации, однако, если сбой произошел из-за некорректной работы автопилота, ответственность может быть возложена на разработчика или эксплуатирующую организацию.

Организационный уровень архитектуры ответственности предполагает внедрение регулярного алгоритмического аудита как обязательной процедуры для систем, используемых в управлении. Такой аудит должен проводиться не только на этапе внедрения, но и регулярно в процессе эксплуатации, поскольку данные и условия могут меняться, приводя к дрейфу модели и снижению качества решений. Эффективной практикой является создание независимых групп, задачей которых является стресс-тестирование алгоритмов – попытки найти уязвимости, предвзятости, неучтенные факторы до того, как они приведут к реальным негативным последствиям.

Образовательный уровень архитектуры ответственности предполагает формирование цифровой компетентности руководителей и специалистов, использующих системы ИИ. Понимание базовых принципов работы алгоритмов, их ограничений, типичных ошибок и когнитивных ловушек должно стать неотъемлемой частью подготовки управленческих кадров – без этого любые технические и институциональные механизмы будут неэффективны, поскольку пользователь не сможет квалифицированно применять предоставленные ему инструменты контроля.

Описанные пять уровней архитектуры ответственности – технический, институциональный, юридический, организационный и образовательный – не существуют изолированно друг от друга. Они образуют единую систему, которую можно охарактеризовать как механизм распределенного контроля. Суть данного механизма заключается в том, что ни один из элементов системы (ни человек, ни алгоритм, ни отдельная процедура) не несет всей полноты ответственности и не обладает всей полнотой власти. Контроль распределяется между разработчиками, задающими технические параметры, менеджерами, принимающими окончательные решения, аудиторами, проверяющими качество работы системы, и регуляторами, устанавливающими правовые рамки. Подобное распределение позволяет избежать двух крайностей: с одной стороны, концентрации неконтролируемой власти в руках алгоритмического «черного ящика», с другой – паралича принятия решений из-за страха ответственности. Распределенный контроль создает систему сдержек и противовесов, где каждая ошибка может быть обнаружена и исправлена на том уровне, где она возникла, не разрушая доверия к системе в целом.

Заключение

Анализ многочисленных инцидентов, связанных с применением ИИ в управлении – от розничной торговли до государственного администрирования – подтверждает существование устойчивого парадокса алгоритмической легитимности. Чем больше мы доверяем алгоритмам в силу их кажущейся объективности, тем сложнее становится установить ответственность, когда решения, сгенерированные ИИ, приводят к негативным последствиям. Данный парадокс не является временным или легко преодолимым – он коренится в самой природе современных алгоритмических систем, работающих как «черные ящики», и в инерции организационных практик, не успевающих за технологическими изменениями. Выход из парадокса видится не в отказе от использования ИИ в управлении, что было бы и невозможным, и нерациональным, а в выстраивании системной «архитектуры ответственности», включающей технические, институциональные, юридические, организационные и образовательные компоненты. Ключевой принцип архитектуры – переход от иллюзии «безупречного кода» к реалистичному пониманию алгоритмов как инструментов, которые могут ошибаться, и к созданию механизмов, позволяющих эти ошибки своевременно выявлять и исправлять.

Предложенная классификация систем ИИ по уровням автономии позволяет дифференцировать подходы к контролю: чем выше автономия, тем более формализованными и многоуровневыми должны быть механизмы предварительного тестирования, текущего мониторинга и пост-аудита. Разработанная модель «архитектуры ответственности» дает руководителям конкретный инструментарий для выстраивания систем управления, в которых ИИ выступает не как «черный ящик», подрывающий подотчетность, а как надежный партнер, усиливающий человеческие возможности и сохраняющий прозрачность.

Список литературы:

1. Абилова Е.В., Колесник Е.А. Внедрение технологий искусственного интеллекта в HR-процессы // Общество, экономика, управление. 2025. Т. 10. № 3. С. 31-36. DOI 10.47475/2618-9852-2025-10-3-31-36 EDN MKESKO
2. Атабаева Э.Р. Возможности и риски применения искусственного интеллекта в бизнесе // Экономика и управление: проблемы, решения. 2025. Т. 13. № 3(156). С. 195-202. DOI 10.36871/ek.up.p.r.2025.03.13.022 EDN VQKFHR
3. Гафурова Г.Т., Нотфуллина Г.Н., Давыдова И.Ш. Робоэдвайзеры vs человеческие управляющие: кто побеждает? // Финансы и кредит. 2025. Т. 31. № 7. С. 162-176. DOI 10.24891/bfwlqc EDN BFWLQC
4. Добринская Д.Е. Цифровое общество в социологической перспективе // Вестник Московского университета. Серия 18. Социология и политология. 2019. Т. 25. № 4. С. 175-192. DOI 10.24290/1029-3736-2019-25-4-175-192 EDN QPHJEN
5. Дорф Т.В. К вопросу о применении искусственного интеллекта в управлении персоналом: адаптация, сопровождение, обучение // Друкеровский вестник. 2025. № 3(65). С. 114-128. DOI 10.17213/2312-6469-2025-3-114-128 EDN AVCWUH
6. Дурович А.П., Гришко Н.И. Возможности и направления применения технологий искусственного интеллекта в управлении персоналом // Труд. Профсоюзы. Общество. 2025. № 1(87). С. 13-22. EDN RIMYFV
7. Катаева М.Г. Цифровые технологии в управлении персоналом: перспективы и вызовы // Управление персоналом и интеллектуальными ресурсами в России. 2024. Т. 13. № 5. С. 21-27. DOI 10.12737/2305-7807-2024-13-5-21-27 EDN RPJBLU

8. Комлев Е.Ю., Бирюков И.А. Правовые основы применения технологий искусственного интеллекта в государственном и муниципальном управлении: современное состояние и перспективы развития // Закон и право. 2024. № 5. С. 81-85. DOI 10.24412/2073-3313-2024-5-81-85 EDN BDFUNT
9. Перспективы и риски применения технологии искусственного интеллекта в корпоративном управлении / [А. В. Фроликов и др.] // Вестник Российского экономического университета имени Г.В. Плеханова. 2025. Т. 22. № 1(139). С. 198-208. DOI 10.21686/2413-2829-2025-1-198-208 EDN DLQPVC
10. Репин Д.А., Игнатьев С.А. «Внедрять нельзя отказаться»: влияние этики на применение технологий искусственного интеллекта в управлении социально-экономическими процессами // Экономика и управление. 2024. Т. 30. № 12. С. 1503-1509. DOI 10.35854/1998-1627-2024-12-1503-1509 EDN HTFMVH
11. Соболева Ю.П., Фадеев О.А. Теоретические основы применения технологий искусственного интеллекта в управлении // Образование и наука без границ: фундаментальные и прикладные исследования. 2024. № 20. С. 68-72. DOI 10.36683/FP-20/68-72 EDN OWKKSF
12. Сурилов М.Н. Исследование потенциала цифровых технологий в совершенствовании мониторинга законодательной базы в сфере образования в Москве // Экономика: вчера, сегодня, завтра. 2024. Т. 14. № 3-1. С. 423-431. EDN TXOFDO
13. Троян Н.А. Правовые механизмы применения технологий искусственного интеллекта в сфере государственного управления в России и за рубежом // Мониторинг правоприменения. 2024. № 3(52). С. 10-16. DOI 10.24682/2226-0692-2024-3-10-16 EDN JQEPHI
14. Черноиваненко И.А., Кравец О.Я., Атласов И.В. Математическое обеспечение распределенного управления роем автономных мобильных объектов с применением технологий искусственного интеллекта // Системы управления и информационные технологии. 2025. № 1(99). С. 71-77. EDN ZXNLLK
15. Biju P.R., Gayathri O. Algorithmic solutions, subjectivity and decision errors: a study of AI accountability // Digital Policy, Regulation and Governance. 2025. Vol. 27. N 5. P. 523-552. DOI 10.1108/DPRG-05-2024-0090
16. Busuioc M. Accountable Artificial Intelligence: Holding Algorithms to Account // Public Administration Review. 2021. Vol. 81. N 5. P. 825-836. DOI 10.1111/puar.13293
17. Criado J.I., Sandoval-Almazán R., Gil-Garcia J.R. Artificial intelligence and public administration: Understanding actors, governance, and policy from micro, meso, and macro perspectives // Public Policy and Administration. 2025. Vol. 40. N 2. P. 173-184. DOI 10.1177/09520767241272921
18. Floridi L. The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities. Oxford: Oxford University Press, 2023. 288 p.
19. Henman P. Governing by Algorithms and Algorithmic Governmentality: Towards machinic judgement // The Algorithmic Society: Technology, Power, and Knowledge. Ed. by M. Schuilenburg, R. Peeters. London: Routledge, 2020. P. 19-34. DOI 10.4324/9780429261404
20. Papagiannidis E., Mikalef P., Conboy K. Responsible artificial intelligence governance: A review and research framework // The Journal of Strategic Information Systems. 2025. Vol. 34. N 2. Article 101885. DOI 10.1016/j.jsis.2024.101885
21. Waldman A., Martin K. Governing algorithmic decisions: The role of decision importance and governance on perceived legitimacy of algorithmic decisions // Big Data & Society. 2022. Vol. 9. N 1. Article 20539517221100449. DOI 10.1177/20539517221100449.

The paradox of algorithmic legitimacy and the architecture of responsibility: how to avoid the traps of artificial intelligence in management

Lavrova Elena Viktorovna
Candidate of Economic Sciences, Associate Professor of the Department of Management and Natural Sciences, Smolensk University of Sports, Smolensk, Russian Federation
e-mail: **e.v.lavrova@list.ru**
RSCI SPIN code: 2964-4612
ORCID ID: 0000-0002-2824-1127

Abstract

This article examines the paradox of algorithmic legitimacy, defined as the contradiction between the growing trust in artificial intelligence for managerial decision-making and the erosion of accountability for the outcomes of such decisions. The study characterizes typical scenarios of managerial failures when utilizing algorithmic systems in commerce, finance, human resource management, logistics, and public administration. Furthermore, it uncovers the root causes of accountability erosion, specifically focusing on automation bias, the opacity of algorithmic models, and the lack of institutional control mechanisms. The author describes an «accountability architecture» model that integrates technical, institutional, legal, organizational, and educational levels of control. Additionally, a classification of artificial intelligence systems based on their level of autonomy is proposed, and corresponding control mechanisms are defined. The study concludes with the necessity of transitioning from declarative ethical principles to mechanisms of distributed control.

Keywords

• artificial intelligence • managerial decisions • algorithmic legitimacy • accountability architecture • algorithmic audit • distributed control •

References:

1. Abilova E.V., Kolesnik E.A. Implementation of Artificial Intelligence Technologies in HR Processes // Society, Economy, Management. 2025. Vol. 10. N 3. P. 31-36. DOI 10.47475/2618-9852-2025-10-3-31-36 EDN MKESKO
2. Atabaeva E.R. Opportunities and Risks of Using Artificial Intelligence in Business // Economics and Management: Problems, Solutions. 2025. Vol. 13. N 3(156). P. 195-202. DOI 10.36871/ek.up.p.r.2025.03.13.022 EDN VQKFHR

3. Gafurova G.T., Notfullina G.N., Davydova I. Sh. Robo-advisors vs Human Managers: Who Wins? // Finance and Credit. 2025. Vol. 31. N 7. P. 162-176. DOI 10.24891/bfwlqc EDN BFWLQC
4. Dobrinskaya D.E. Digital Society in Sociological Perspective // Moscow University Bulletin. Series 18. Sociology and Political Science. 2019. Vol. 25. N 4. P. 175-192. DOI 10.24290/1029-3736-2019-25-4-175-192 EDN QPHJEN
5. Dorf T.V. On the Issue of Using Artificial Intelligence in Personnel Management: Adaptation, Support, Training // Drucker's Herald. 2025. N 3(65). P. 114-128. DOI 10.17213/2312-6469-2025-3-114-128 EDN AVCWUH
6. Durovich A.P., Grishko N.I. Opportunities and Directions for the Application of Artificial Intelligence Technologies in Personnel Management // Labor. Trade Unions. Society. 2025. N 1(87). P. 13-22. EDN RIMYFV
7. Kataeva M.G. Digital Technologies in Personnel Management: Prospects and Challenges // Personnel and Intellectual Resources Management in Russia. 2024. Vol. 13. N 5. P. 21-27. DOI 10.12737/2305-7807-2024-13-5-21-27 EDN RPJBLU
8. Komlev E.Yu., Biryukov I.A. Legal Foundations for the Application of Artificial Intelligence Technologies in State and Municipal Administration: Current State and Development Prospects // Law and Legislation. 2024. N 5. P. 81-85. DOI 10.24412/2073-3313-2024-5-81-85 EDN BDFUNT
9. Prospects and Risks of Using Artificial Intelligence Technology in Corporate Management / [A.V. Frolikov et al.] // Vestnik of the Plekhanov Russian University of Economics. 2025. Vol. 22. N 1(139). P. 198-208. DOI 10.21686/2413-2829-2025-1-198-208 EDN DLQPVC
10. Repin D.A., Ignatiev S.A. «Implement or Abandon»: The Influence of Ethics on the Application of Artificial Intelligence Technologies in Managing Socio-Economic Processes // Economics and Management. 2024. Vol. 30. N 12. P. 1503-1509. DOI 10.35854/1998-1627-2024-12-1503-1509 EDN HTFMVH
11. Soboleva Yu.P., Fadeev O.A. Theoretical Foundations of Using Artificial Intelligence Technologies in Management // Education and Science Without Borders: Fundamental and Applied Research. 2024. N 20. P. 68-72. DOI 10.36683/FP-20/68-72 EDN OWKKSF
12. Surilov M.N. A Study of the Potential of Digital Technologies in Improving the Monitoring of the Legislative Framework in the Field of Education in Moscow // Economy: Yesterday, Today, Tomorrow. 2024. Vol. 14. N 3-1. P. 423-431. EDN TXOFDO
13. Troyan N.A. Legal Mechanisms for the Application of Artificial Intelligence Technologies in Public Administration in Russia and Abroad // Monitoring of Law Enforcement. 2024. N 3(52). P. 10-16. DOI 10.24682/2226-0692-2024-3-10-16 EDN JQEPHI
14. Chernoivanenko I.A., Kravets O.Ya., Atlasov I.V. Mathematical Support for Distributed Control of a Swarm of Autonomous Mobile Objects Using Artificial Intelligence Technologies // Control Systems and Information Technologies. 2025. N 1(99). P. 71-77. EDN ZXNLLK
15. Biju P.R., Gayathri O. Algorithmic solutions, subjectivity and decision errors: a study of AI accountability // Digital Policy, Regulation and Governance. 2025. Vol. 27. N 5. P. 523-552. DOI 10.1108/DPRG-05-2024-0090
16. Busuioc M. Accountable Artificial Intelligence: Holding Algorithms to Account // Public Administration Review. 2021. Vol. 81. N 5. P. 825-836. DOI 10.1111/puar.13293
17. Criado J.I., Sandoval-Almazán R., Gil-Garcia J.R. Artificial intelligence and public administration: Understanding actors, governance, and policy from micro, meso, and macro perspectives // Public Policy and Administration. 2025. Vol. 40. N 2. P. 173-184. DOI 10.1177/09520767241272921
18. Floridi L. The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities. Oxford: Oxford University Press, 2023. 288 p.

19. Henman P. Governing by Algorithms and Algorithmic Governmentality: Towards machinic judgement // The Algorithmic Society: Technology, Power, and Knowledge. Ed. by M. Schuilenburg, R. Peeters. London: Routledge, 2020. P. 19-34. DOI 10.4324/9780429261404
20. Papagiannidis E., Mikalef P., Conboy K. Responsible artificial intelligence governance: A review and research framework // The Journal of Strategic Information Systems. 2025. Vol. 34. N 2. Article 101885. DOI 10.1016/j.jsis.2024.101885
21. Waldman A., Martin K. Governing algorithmic decisions: The role of decision importance and governance on perceived legitimacy of algorithmic decisions // Big Data & Society. 2022. Vol. 9. N 1. Article 20539517221100449. DOI 10.1177/20539517221100449.

Поступила в редакцию: 14.03.2026
Рецензия получена: 28.04.2026
Принята в печать: 08.06.2026