УДК 338.2

# ИНТЕГРИРОВАННАЯ МЕТОДИКА ОЦЕНКИ РИСКОВ ПРИМЕНЕНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ФУНКЦИОНАЛЬНЫХ СФЕРАХ УПРАВЛЕНИЯ ОРГАНИЗАЦИЕЙ

***Л.О. Бочкарева***
Государственное автономное образовательное учреждение высшего образования «Московский городской университет управления Правительства Москвы имени Ю.М. Лужкова», Москва, email: Lusik_kru@mail.ru

***Аннотация.*** *Цель исследования заключается в разработке методики оценки рисков применения искусственного интеллекта (ИИ), специфичной для функциональной сферы управления. В процессе проведения исследования применялась теория риск-менеджмента, теория сложных систем, многокритериальное принятие решений, расчетный и аналитический методы, функционально-специфический подход. В результате проведения исследования предложена многоуровневая некомпенсаторная свёртка с переоценкой по жизненному циклу ИИ. Результаты применимы риск-менеджерами и регуляторами. Научная новизна работы определяется пересечением функциональной специфичности, формализованной многокритериальной свёртки и адаптации к российскому нормативному контексту. Методика обеспечивает измеримый инструментарий управления рисками ИИ и может служить основой как для корпоративных регламентов, так и для развития отраслевой стандартизации.*

***Ключевые слова:*** *искусственный интеллект, оценка рисков, риск-менеджмент, функциональная сфера управления, многокритериальное принятие решений, методика, управление персоналом, финансы, маркетинг.*

# INTEGRATED METHOD FOR ASSESSING THE RISKS OF USING ARTIFICIAL INTELLIGENCE IN FUNCTIONAL AREAS OF ORGANIZATION MANAGEMENT

***L.O. Bochkareva***
Moscow Metropolitan Governance Yury Luzhkov University, Moscow, email: Lusik_kru@mail.ru

***Abstract.*** *The purpose of the study is to develop a methodology for assessing the risks of using artificial intelligence (AI) that is specific to the functional area of management. The study used risk management theory, complex systems theory, multi-criteria decision-making, computational and analytical methods, and a functional-specific approach. The study proposed a multi-level non-compensatory convolution with reevaluation based on the AI lifecycle. The results can be used by risk managers and regulators. The scientific novelty of the work is determined by the intersection of functional specificity, formalized multi-criteria convolution, and adaptation to the Russian regulatory context. The methodology provides a measurable toolkit for managing AI risks and can serve as a basis for both corporate regulations and the development of industry-specific standardization.*

***Keywords:*** *artificial intelligence, risk assessment, risk management, functional management, multi-criteria decision-making, methodology, personnel management, finance, marketing.*

Дата поступления статьи в редакцию: 25.05.2026
Дата принятия статьи в печать: 08.07.2026

### *Введение*

Внедрение систем искусственного интеллекта в управление организацией перешло из стадии экспериментов в стадию массовой эксплуатации: по оценке Банка России, к 2025 году ИИ применяет каждая пятая организация финансового рынка, ещё треть планирует внедрение в трёхлетнем горизонте [6]. Сопутствующее расширение рисков — дискриминационных, модельных, приватностных, репутационных — поставило перед менеджментом задачу не только классификации этих рисков, но и их измеримой оценки.

Существующие подходы к оценке рисков ИИ носят преимущественно универсальный характер. Международные рамки управления рисками — NIST AI Risk Management Framework [10] и Регламент Европейского союза 2024/1689 [13] — задают процесс и категории, но сознательно не предписывают конкретных метрик, оставляя их выбор организации. В результате универсальная оценка не учитывает, что один и тот же тип риска проявляется и измеряется по-разному в различных функциональных сферах управления. Дискриминационный риск в подборе персо-

нала, в кредитном скоринге и в маркетинговом таргетинге требует разных индикаторов, разных пороговых значений и разной значимости в общей оценке. Универсальная методика даёт либо чрезмерно общие, либо неприменимые в конкретной функции результаты.

Настоящая статья продолжает исследование, начатое в работе [3], где предложена многоуровневая классификация рисков применения ИИ в российских организациях. Классификация отвечает на вопрос «какие риски существуют»; настоящая работа отвечает на вопрос «насколько велик риск и как его измерить» в привязке к конкретной функции управления.

***Материал и методы исследования***

Существующие методы оценки рисков ИИ распадаются на три класса. Количественные методы опираются на статистические и вероятностные модели, оценку «вероятность Ч ущерб», байесовские и Монте-Карло-подходы. Качественные методы используют экспертные оценки, сценарный анализ и чек-листы. Гибридные методики сочетают оба подхода, однако в подавляющем большинстве останавливаются на процессных чек-листах без формализованной агрегации разнородных оценок.

Систематический обзор показывает, что количественные методики оценки рисков ИИ существуют, но являются сферо-нейтральными: они применяются «по ИИ-системе», а не «по функциональной области управления». Универсальные рамки [10; 13] фиксируют процесс, но не вид свёртки. Репозиторий рисков ИИ [15], консолидирующий 74 фреймворка и более 1700 рисков, носит таксономический характер и не содержит измерительной агрегации. Анализ международных стандартов подтверждает: ни NIST AI RMF, ни ISO/IEC 23894, ни ISO/IEC 42001, ни Регламент ЕС 2024/1689 не предписывают функционально-специфических количественных метрик. Единственное прямое нормативное упоминание измеримости — статья 9(8) Регламента 2024/1689, требующая наличия «заранее определённых метрик и вероятностных порогов», содержание которых отдаётся разработчику [13].

В русскоязычном сегменте также не обнаружено методики, сочетающей функциональную специфичность с формализованной количественно-качественной свёрткой, имеющиеся работы дают либо общие количественные модели цифровых рисков, либо обзоры рисков ИИ в финансовом секторе без формализации [2], либо нейро-нечёткие модели для автоматизированных систем безотносительно функции управления [1].

Таким образом, научная новизна разрабатываемой методики определяется пересечением трёх свойств, по которому, согласно проведённому анализу, не обнаружено опубликованных работ: функциональная специфичность по сферам управления — управление персоналом, финансы, маркетинг; формализованная свёртка количественных и качественных оценок методами многокритериального принятия решений; адаптация к российскому нормативно-правовому контексту. Методика не дублирует существующие стандарты, а реализует их «методологический крючок» — требование измеримости, оставленное стандартами на усмотрение организации.

Методика опирается на четыре теоретические опоры. Теория риск-менеджмента — стандарт ISO 31000:2018 (принципы, рамка, процесс «идентификация → анализ → оценивание → обработка риска») и концепция COSO ERM, интегрирующая риск со стратегией организации. Функциональный подход к управлению обосновывает разделение оценки по сферам: каждая функция имеет собственный контур задач, данных и регуляторных требований. Теория сложных адаптивных систем обосновывает многоуровневость и нелинейность оценки, риск на уровне организации не сводится к сумме рисков отдельных ИИ-компонентов в силу эмерджентности, а малые возмущения способны вызывать каскадные эффекты. Отсюда следует требование иерархической оценки и недопустимость полной компенсации опасных рисков благополучными. Методы многокритериального принятия решений (МКПР) дают аппарат свёртки разнородных оценок.

Выбор аппарата МКПР обоснован сопоставлением методов. Метод анализа иерархий (AHP) и его нечёткая версия пригодны для определения весов критериев по экспертным суждениям, но реализуют полностью компенсаторную аддитивную свёртку. Метод DEMATEL структурирует причинно-следственные связи между типами риска. Метод VIKOR при малом значении параметра компромисса обеспечивает частично-некомпенсаторное ранжирование, при котором опасный частный риск не «растворяется» в средней оценке [12]. Чисто аддитивная взвешенная сумма для риска ИИ непригодна, она позволяет благополучным показателям компенсировать катастрофические. В качестве рабочего аппарата принята трёхуровневая связка: DEMATEL для

структурирования взаимовлияний типов риска, нечёткий AHP для определения сферо-зависимых весов, нечёткий VIKOR (либо мультипликативная некомпенсаторная свёртка) для итогового агрегирования.

*Результаты исследования*

Методика строится на четырёх принципах. Многоуровневая структура оценки: четыре уровня агрегации — частный показатель, тип риска, функциональная сфера, организация в целом. Адаптивность к функциональной сфере: общий каркас инстанцируется для конкретной сферы заданием профиля рисков, весового вектора и комплаенс-контура; методика является не единым инструментом, а порождающей схемой субметодик. Интеграция количественных и качественных показателей: каждый значимый тип риска измеряется и количественными метриками, и качественными индикаторами, сводимыми к единой шкале. Динамический характер: оценка привязана к этапу жизненного цикла ИИ-системы и подлежит переоценке по календарным и событийным правилам.

Формально оценка задаётся на пяти координатных осях: функциональная сфера F (управление персоналом, финансы, маркетинг); тип риска R (девять типов — качество данных, предвзятость, интерпретируемость, безопасность, приватность, операционная устойчивость, регуляторный, репутационный, финансовый риск); природа показателя N (количественная, качественная); уровень оценки L (показатель, тип риска, сфера, организация); этап жизненного цикла C (концепция, данные и разработка, верификация и валидация, развёртывание, эксплуатация и мониторинг, переоценка).

В сфере управления персоналом доминируют риски предвзятости и регуляторный риск. Ядро количественных показателей: коэффициент неблагоприятного воздействия (adverse impact ratio) при найме с порогом 0,80, восходящим к «правилу четырёх пятых»; разность долей положительных решений (statistical parity difference) и разность равных возможностей (equal opportunity difference) с практическим порогом 0,10 [9]; точность прогнозирования увольнений по площади под ROC-кривой с порогом приёмки 0,70; скорректированный регрессионным методом разрыв в оплате труда. Качественные показатели — прозрачность HR-процессов, принятие ИИ-решений сотрудниками, восприятие справедливости алгоритма — операционализируются валидированными опросными шкалами с проверкой внутренней согласованности (коэффициент Кронбаха не ниже 0,70).

Существенное ограничение — теоремы несовместимости критериев справедливости: демографический паритет, выравнивание шансов и предиктивный паритет не могут выполняться одновременно при различных базовых ставках групп [8]. Поэтому субметодика для каждой HR-задачи фиксирует единственную каноническую метрику справедливости: для подбора — коэффициент неблагоприятного воздействия, для прогноза увольнений — разность равных возможностей, для оценки персонала — предиктивный паритет. В российском контексте прямой расчёт метрик справедливости ограничен Федеральным законом «О персональных данных», относящим расовую, национальную принадлежность и состояние здоровья к специальным категориям, для этих признаков применяется косвенный аудит.

В финансовой функции доминируют риски качества данных, финансовый и операционный. Количественные показатели: показатели дискриминирующей способности скоринговых моделей — коэффициент Джини в типичном диапазоне 0,4–0,6 и статистика Колмогорова — Смирнова; индекс стабильности популяции (PSI) с пороговой шкалой 0,10 / 0,25, где превышение 0,25 требует перекалибровки модели; стоимостно-чувствительные метрики обнаружения мошенничества, учитывающие несимметричную стоимость ошибок I и II рода; показатели рыночного риска (Value at Risk) с бэктестингом по «светофорной» схеме. Принципиально разводятся два значения термина «Value at Risk для ИИ-моделей»: VaR, рассчитываемый ИИ-моделью (мера рыночного риска портфеля), и VaR модельного риска (квантиль распределения убытков от ошибок самой модели).

Качественная часть субметодики опирается на практику управления модельным риском. Международным ориентиром служит надзорное руководство SR 11-7; в российском регулировании ему соответствует совокупность нормативных актов Банка России — Указание № 3624-У (определение модельного риска), Положение № 716-П (модельный риск как вид операционного, реестр моделей) [6], Положение № 483-П (независимая валидация). Качество управле-

ния модельным риском оценивается по трём блокам — разработка, валидация, корпоративное управление — с агрегированием в сводный индикатор.

В маркетинговой функции доминируют риски приватности, репутационный и регуляторный. Количественные показатели объединены в четыре группы: точность таргетинга и сегментации (точность, полнота, коэффициент силуэта); эффективность персонализации (приростный отклик, метрики ранжирования рекомендаций); защита персональных данных (k-анонимность с практическим порогом не менее 5, бюджет дифференциальной приватности с интерпретацией значения не выше 1 как сильной гарантии, риск переидентификации); коэффициенты мультиканальной атрибуции на основе значения Шепли — единственного способа распределения, удовлетворяющего аксиомам эффективности, симметрии, нулевого игрока и аддитивности.

Особенность субметодики — явное разделение показателей с абсолютными порогами (приватность, силуэт), универсальными и не зависящими от домена, и показателей с относительными порогами (точность таргетинга, метрики ранжирования), требующими сопоставительной «бенчмарк-карточки» с указанием набора данных, отрасли, периода и протокола оценки. Качественные индикаторы — этичность использования клиентских данных, прозрачность рекомендательных алгоритмов, влияние на бренд. Прозрачность рекомендательных технологий в российском регулировании имеет прямое нормативное основание — статья 10.2-2 Федерального закона «Об информации...», что позволяет построить бинарный индикатор соответствия.

Инструментарий реализует конвейер из пяти операций, преобразующих наблюдения об ИИ-системе в интегральную оценку риска.

Кодирование качественных оценок. Лингвистические экспертные оценки по пятиуровневой шкале (очень низкий — очень высокий риск) представляются триангулярными нечёткими числами вида (l, m, u). Оценки экспертов агрегируются (нижняя граница — минимум, мода — среднее, верхняя — максимум) и дефаззифицируются по формуле центра тяжести $x^* = (l + 2m + u) / 4$.

Нормализация. Все показатели приводятся к единой шкале риска [0; 1] методом минимакса с инверсией для показателей «благо-типа», где большее значение означает меньший риск.

Взвешивание. Структура взаимовлияний типов риска выявляется методом DEMATEL. Веса типов риска определяются для каждой сферы отдельной сессией нечёткого AHP с контролем согласованности (отношение согласованности не выше 0,10); запрет на единый набор весов для всех сфер является аксиомой методики. Веса отдельных показателей при наличии данных по нескольким системам уточняются объективным методом CRITIC.

Свёртка. Каскад агрегации по уровням: внутри типа риска показатели сворачиваются аддитивно; типы риска в композитный индекс сферы — мультипликативной некомпенсаторной свёрткой вида $I = 1 - \prod(1 - \rho)^w$, при которой катастрофическое значение любого частного риска доминирует в итоговой оценке; сферы в оценку организации — геометрической свёрткой. Альтернативой для задач ранжирования служит нечёткий VIKOR с параметром компромисса не выше 0,3. До агрегирования проверяются «ko-критерии» — показатели с абсолютными порогами (коэффициент неблагоприятного воздействия, индекс соответствия прозрачности, бюджет приватности), нарушение которых блокирует эксплуатацию системы независимо от прочих оценок.

Калибровка и интерпретация. Композитный индекс переводится в трёхзонную шкалу («зелёный / жёлтый / красный»); при отсутствии данных о реализованных потерях пороги задаются тертилями эмпирического распределения индекса по сопоставимым организациям с экспертной верификацией. Единая нормализация обеспечивает сопоставимость оценок риска разных функциональных сфер.

Динамический контур реализуется правилом переоценки, объединяющим календарную составляющую (период зависит от сферы — от ежедневной для финансов до еженедельной для управления персоналом и маркетинга) и событийную (триггеры по метрикам дрейфа: индекс стабильности популяции не ниже 0,25, падение ключевой метрики не менее чем на 10 %).

Поскольку доступ к корпоративным данным организаций ограничен, апробация носит иллюстративный характер и проводится на публичных данных. Предлагается комбинированный протокол, принятый в методических исследованиях и совместимый с публичными источниками.

Экспертная валидация проводится методом Дельфи (два-три анонимных раунда, панель 7–12 экспертов, порог консенсуса не ниже 75 %); согласованность экспертных рангов оценивается коэффициентом конкордации Кендалла с порогом 0,5 и проверкой значимости по критерию

хи-квадрат. Содержательная валидность показателей проверяется индексами контент-валидности; внутренняя согласованность качественных шкал — коэффициентом Кронбаха.

Иллюстративный расчёт выполняется на одном-двух открытых кейсах для каждой сферы — публично описанных инцидентах применения ИИ и годовой отчётности публичных компаний. Цель расчёта — продемонстрировать дифференциацию интегральной оценки по сферам и выход за пределы качественной оценки в количественную плоскость.

Анализ чувствительности обязателен для доверия к композитному индексу: применяется глобальный дисперсионный анализ с расчётом индексов Соболя первого и полного порядка на выборке не менее 1000 модельных прогонов [14]. Критерий устойчивости — ни один отдельный весовой параметр не должен объяснять более 30 % дисперсии итоговой оценки; в противном случае весовая схема доминирует над данными и подлежит пересмотру. Робастность ранжирования проверяется сопоставлением рангов при базовом и альтернативных наборах весов.

Внедрение методики целесообразно осуществлять по следующему алгоритму. На первом шаге ИИ-сценарий относится к уровню риска по ступенчатой регуляторной классификации, что определяет глубину оценки. Далее формируется набор показателей выбранной сферы, проверяются «ko-критерии», проводятся кодирование, нормализация, взвешивание и каскадная свёртка, а полученный индекс калибруется в управленческую категорию. Завершается цикл установлением профиля переоценки.

Методика встраивается в существующую систему управления рисками организации как специализированный модуль оценки рисков ИИ, согласованный с процессом ISO 31000 и системой менеджмента ИИ по стандарту ISO/IEC 42001. Требования к компетенциям команды оценки носят междисциплинарный характер: необходимо сочетание экспертизы в риск-менеджменте, машинном обучении и в соответствующей функциональной сфере. Адаптация к типу и размеру организации обеспечивается параметрически — составом оцениваемых типов риска, весовыми векторами и периодичностью переоценки. Качественный комплаенс-контур субметодик опирается на действующие нормативные акты Российской Федерации, что позволяет не конструировать индикаторы соответствия заново, а операционализировать существующие нормы.

### *Заключение*

Таким образом, разработана интегрированная методика оценки рисков применения искусственного интеллекта в функциональных сферах управления организацией. Её преимущество перед универсальными подходами состоит в функциональной специфичности — настройке состава показателей, пороговых значений и весов под конкретную функцию управления — при сохранении единого формального каркаса. Методика объединяет количественные и качественные показатели в едином измерительном инструменте, использует некомпенсаторную свёртку, не допускающую сокрытия опасных частных рисков благополучными показателями, и обладает динамическим характером, привязывая оценку к этапу жизненного цикла ИИ-системы и допуская переоценку по календарным и событийным триггерам.

Научная новизна работы определяется пересечением трёх свойств — функциональной специфичности по сферам управления, формализованной многокритериальной свёртки разнородных показателей и адаптации к российскому нормативному контексту. Методика обеспечивает измеримый инструментарий управления рисками ИИ и может служить основой как для корпоративных регламентов оценки и допуска ИИ-систем к эксплуатации, так и для развития отраслевой стандартизации в области ответственного применения ИИ.

Ограничения исследования связаны с иллюстративным характером апробации на публичных данных и отсутствием эмпирически откалиброванных на российских данных пороговых значений; универсальные пороговые значения, заимствованные из международной практики, требуют валидации в конкретных российских отраслевых контекстах. Направления дальнейших исследований включают: психометрическую валидацию качественных опросных шкал (Colquitt OJS, UTAUT, AIM-FAIR) на российской выборке; накопление статистики реализованных инцидентов применения ИИ для перехода от квантильной калибровки порогов к калибровке по реальным потерям с использованием индекса Юдена; расширение методики на иные функциональные сферы управления (операционная и производственная функции, логистика, корпоративные коммуникации); разработку отраслевых профилей методики для банковской сферы, ритейла, телекоммуникаций и государственного сектора.

## Литература

1. Айдинян А.Р., Цветкова О.Л. Нечёткая модель оценки рисков информационной безопасности автоматизированных систем // Вестник Донского государственного технического университета. 2023. Т. 50, № 2. С. 178-186.

2. Гринько Е.Л., Гарагуц М.А., Вырезуб Д.В. Риски применения искусственного интеллекта в финансовом секторе Российской Федерации // Финансовая экономика. 2024. № 6. С. 122-126. EDN: AVTTRE.

3. Бочкарева Л.О. Многоуровневая система рисков применения искусственного интеллекта в российских организациях: интеграция международного опыта и национальных особенностей // Финансовая экономика. 2025. № 6. С. 14-17. EDN: DIRYOF.

4. Кодекс этики в сфере искусственного интеллекта (Российская Федерация, принят 26.10.2021). М., 2021. [Электронный ресурс]. URL: https://digital.avo.ru/-/kodeks-etiki-v-sfere-iskusstvennogo-intellekta (дата обращения 11.05.2026).

5. Орлов А.И. Организационно-экономическое моделирование. Часть 2. Экспертные оценки: учебник. М.: Издательство МГТУ им. Н.Э. Баумана, 2011. 486 с.

6. Положение Банка России от 08.04.2020 № 716-П «О требованиях к системе управления операционным риском в кредитной организации и банковской группе». М., 2020. [Электронный ресурс]. URL: https://www.garant.ru/products/ipo/prime/doc/74179372/ (дата обращения 11.05.2026).

7. Chang D.-Y. Applications of the extent analysis method on fuzzy AHP // European Journal of Operational Research. 1996. Vol. 95. № 3. P. 649-655. DOI: 10.1016/0377-2217(95)00300-2 EDN: ANQNGV.

8. Chouldechova A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments // Big Data. 2017. Vol. 5. № 2. P. 153-163. DOI: 10.1089/big.2016.0047.

9. Hardt M., Price E., Srebro N. Equality of opportunity in supervised learning // Advances in Neural Information Processing Systems. 2016. Vol. 29. P. 3315-3323.

10. NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg: National Institute of Standards and Technology, 2023.

11. OECD, Joint Research Centre. Handbook on Constructing Composite Indicators: Methodology and User Guide. Paris: OECD Publishing, 2008. 162 p.

12. Opricovic S., Tzeng G.-H. Compromise solution by MCDM methods: a comparative analysis of VIKOR and TOPSIS // European Journal of Operational Research. 2004. Vol. 156. № 2. P. 445-455. DOI: 10.1016/S0377-2217(03)00020-1.

13. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, 2024.

14. Saisana M., Saltelli A., Tarantola S. Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators // Journal of the Royal Statistical Society: Series A. 2005. Vol. 168. № 2. P. 307-323. DOI: 10.1111/j.1467-985X.2005.00350.x EDN: XTGGWH.

15. Slattery P., Saeri A.K., Grundy E.A.C. et al. The AI Risk Repository: a comprehensive meta-review, database, and taxonomy of risks from artificial intelligence // Patterns. 2026. Vol. 7. № 5. DOI: 10.1016/j.patter.2026.101517 EDN: RIDRAE.