



Банк России

**МАТЕРИАЛЫ ЗАСЕДАНИЯ ЭКСПЕРТНОГО
СОВЕТА ПО РЕГУЛИРОВАНИЮ,
МЕТОДОЛОГИИ ВНУТРЕННЕГО АУДИТА,
ВНУТРЕННЕГО КОНТРОЛЯ И УПРАВЛЕНИЯ
РИСКАМИ В БАНКЕ РОССИИ И ФИНАНСОВЫХ
ОРГАНИЗАЦИЯХ ОТ 16 СЕНТЯБРЯ 2026 ГОДА**

ОГЛАВЛЕНИЕ

Обращение главного аудитора Банка России к читателям.....	3
Благодарность авторам.....	5
Благодарность ведущим мастер-классов.....	6
Обзор материалов выпуска.....	7
От формальной зрелости к реальной эффективности: как оценить влияние зрелости системы управления рисками на принятие решений в эпоху искусственного интеллекта (Е.В. Трофимова, ДРТ)	9
Управление модельным риском: от финансовых моделей к рискам искусственного интеллекта (П.А. Разумовский, НКЦ)	14
Аудит оценки кредитного риска на основе ПВР: от проверки моделей к оценке целостной рейтинговой системы (Д.С. Семилетов, Л.Е. Кузнецов, Банк ВТБ).....	21
Безопасное масштабирование искусственного интеллекта в банке: риски, модель угроз и контуры контроля (А.М. Валиуллин, И.Л. Вахламов, Газпромбанк).....	29
От политики запретов к управляемому искусственному интеллекту: угрозы безопасности ИИ-систем и модель их контроля в финансовых организациях (Р.М. Гусейнов, ПК «РАД КОП»).....	33
Повышение требований к экспертизе и компетенциям сотрудников, создающих и проверяющих продукты искусственного интеллекта: обратная сторона ожидаемой экономии ресурсов (А.А. Пенин, А.М. Козлов, А.О. Мацковская, Кепт)	52
Заключение.....	60
Экосистема Экспертного совета.....	61
Глоссарий	63
Список сокращений.....	66

Редакционная коллегия дайджеста:

В.П. Горегляд, председатель редакционной коллегии, д.э.н.

М.А. Лауфер, к.э.н.

Н.А. Станик, к.э.н.

В.А. Щукин

Материал подготовлен службой главного аудитора Банка России.

Ответственные за выпуск: Н.А. Станик, В.А. Щукин.

Мнения, содержащиеся в материалах, являются точкой зрения авторов и могут не совпадать с официальной позицией Банка России.

Комментарии, предложения и замечания можно направлять по адресу: expert.board@mail.cbr.ru.

107016, Москва, ул. Неглинная, 12, к. В

Официальный сайт Банка России: www.cbr.ru

© Центральный банк Российской Федерации, 2026



ОБРАЩЕНИЕ ГЛАВНОГО АУДИТОРА БАНКА РОССИИ К ЧИТАТЕЛЯМ

Уважаемые коллеги!

Настоящий выпуск объединяет материалы к заседанию Экспертного совета 16 сентября 2026 года по теме «Управление модельными и ИИ-рисками: практика их аудита и контроля». Мы продолжаем обсуждение, начатое на июньском заседании. Тема вызвала значительный интерес профессионального сообщества и обозначила широкий круг вопросов, требующих дальнейшей проработки.

Поэтому сентябрьское [заседание](#) посвящено практическим подходам организаций, опыту их применения и требованиям к компетенциям сотрудников, которые создают и проверяют ИИ-системы.

Частью подготовки стали тематические мастер-классы: с апреля по сентябрь по теме модельных и ИИ-рисков проведено 15 встреч. Этот формат позволяет подробно разбирать практические вопросы и сопоставлять подходы разных организаций, а подготовленные по его итогам материалы вошли в настоящий выпуск, в том числе статьи, не сопровождавшиеся докладом на заседании.

Представленные подходы объединяет общий вопрос: способна ли организация подтвердить, что ее контрольные процедуры результативны. Для внутреннего аудита важна прослеживаемость всего пути: от исходных данных и допущений модели до ее применения и последствий для организации. Значение имеют качество независимой проверки, реакция на замечания, контроль изменений и возможность восстановить обстоятельства принятия решения.

Особого внимания заслуживает роль человека. Сотрудник, проверяющий результат работы ИИ-системы, должен обладать знаниями и полномочиями для самостоятельного профессионального суждения. По мере расширения автономности систем организации необходимо сохранять способность оспорить ошибочный вывод, ограничить применение технологии и продолжить процесс без нее.

Экспертный совет помогает прорабатывать эти вопросы совместно, объединяя специалистов по внутреннему аудиту, внутреннему контролю, управлению рисками, информационным технологиям и информационной безопасности. Благодарю авторов публикаций, ведущих мастер-классов и участников профессионального сообщества за содержательные дискуссии и готовность делиться опытом.

Надеюсь, что материалы выпуска помогут руководителям и специалистам определить направления совершенствования управления моделями и ИИ, внутреннего контроля и аудита.

В.П. Горегляд

Главный аудитор Банка России, председатель Экспертного совета по регулированию, методологии внутреннего аудита, внутреннего контроля и управления рисками в Банке России и финансовых организациях

ВСТУПЛЕНИЕ

Настоящий выпуск посвящен материалам заседания Экспертного совета по регулированию, методологии внутреннего аудита, внутреннего контроля и управления рисками в Банке России и финансовых организациях, состоявшегося 16 сентября 2026 года.

Материалы выпуска объединяют исследования и практические подходы к оценке зрелости риск-менеджмента, управлению модельным риском, аудиту рейтинговых систем, безопасному масштабированию ИИ и развитию профессиональных компетенций. Выпуск включает шесть публикаций, подготовленных представителями финансовых организаций, консультационных компаний и профессионального сообщества.

Выпуск продолжает обсуждение управления модельными и ИИ-рисками, начатое на заседании Экспертного совета 23 июня 2026 года. В новых публикациях акцент смещен с постановки вопросов на практику: участие риск-функции в решениях, проверка данных, полномочия ИИ-агентов, оценка результатов технологии сотрудниками.

НОВЫЙ ФОРМАТ РАБОТЫ ЭКСПЕРТНОГО СОВЕТА

В 2026 году Экспертный совет начал использовать формат подготовки заседаний, сочетающий тематические мастер-классы, профессиональное обсуждение и публикацию материалов. С апреля по сентябрь 2026 года по теме модельных и ИИ-рисков проведено 15 мастер-классов.

Работа включает несколько последовательных этапов:

- проведение тематических мастер-классов;
- обсуждение практических вопросов с профессиональным сообществом;
- представление подходов и выводов на заседании Экспертного совета;
- публикацию материалов в тематическом дайджесте.

Мастер-классы позволяют подробнее рассмотреть тему, разобрать практические случаи и обсудить вопросы, для которых недостаточно времени в регламенте заседания. Их могут проводить как авторы публикаций, так и коллеги, участвующие в разработке соответствующих подходов. Такая работа поддерживает обмен опытом между заседаниями и помогает уточнять вопросы для дальнейшего обсуждения.

БЛАГОДАРНОСТЬ АВТОРАМ

Редакционная коллегия выражает искреннюю признательность авторам, подготовившим материалы настоящего выпуска:



Е.В. Трофимова
Группа ДРТ



П.А. Разумовский
НКО НКЦ (АО)



Д.С. Семилетов
Банк ВТБ (ПАО)



Л.Е. Кузнецов
Банк ВТБ (ПАО)



А.М. Валиуллин
«Газпромбанк» (АО)



И.Л. Вахлов
«Газпромбанк» (АО)



Р.М. Гусейнов
ПК «РАД КОП»



А.А. Пенин
ООО «Кэпт Налоги
и Консультирование»



А.М. Козлов
ООО «Кэпт Налоги
и Консультирование»



А.О. Мацковская
ООО «Кэпт Налоги
и Консультирование»

Публикации объединили методологические разработки, результаты исследований и практический опыт организаций финансового и реального секторов экономики. Благодаря этому читатели получают возможность познакомиться с различными подходами к управлению модельными и ИИ-рисками, к организации внутреннего аудита моделей и развитию современных систем внутреннего контроля.

Редакционная коллегия также благодарит участников заседания за профессиональную дискуссию, обмен опытом и вклад в развитие современных подходов к внутреннему аудиту, внутреннему контролю и управлению рисками.

БЛАГОДАРНОСТЬ ВЕДУЩИМ МАСТЕР-КЛАССОВ

Отдельную благодарность редакционная коллегия выражает экспертам, которые провели тематические мастер-классы в рамках подготовки заседания. По отдельным темам мастер-классы проводили не авторы публикаций, а их коллеги, чей вклад в подготовку профессиональной дискуссии не менее значим:



Е.В. Трофимова
Группа ДРТ



И.С. Бурочкина
Группа ДРТ



Е.С. Луговцова
Группа ДРТ



Д.А. Качурин
Группа ДРТ



П.А. Разумовский
НКО НКЦ (АО)



Д.С. Семилетов
Банк ВТБ (ПАО)



Л.Е. Кузнецов
Банк ВТБ (ПАО)



И.Л. Вахламов
«Газпромбанк» (АО)



Р.М. Гусейнов
ПК «РАД КОП»



А.М. Козлов
ООО «Кэпт Налоги
и Консультирование»



А.О. Мацковская
ООО «Кэпт Налоги
и Консультирование»



О.В. Павшок
ООО «Кэпт Налоги
и Консультирование»

Редакционная коллегия благодарит ведущих мастер-классов за готовность заранее подробно представить свои подходы и открыто обсудить практические вопросы с профессиональным сообществом.

Уточнение к выпуску от 23 июня 2026 года

В списке ведущих мастер-классов предыдущего выпуска дайджеста не был указан Д.О. Сергиенко, заместитель начальника Управления анализа розничных кредитных рисков Службы анализа рисков Банка России. Редакционная коллегия благодарит его за предоставленные материалы и проведение мастер-класса.



Д.О. Сергиенко
Служба анализа рисков
Банка России

ОБЗОР МАТЕРИАЛОВ ВЫПУСКА

В выпуск включены шесть публикаций. Каждая публикация отвечает на свой вопрос: насколько зрела система управления рисками, как организовать контроль моделей, как его проверить, как безопасно применять ИИ и кто должен его проверять. Ниже представлен краткий обзор материалов.

От формальной зрелости к реальной эффективности: как оценить влияние зрелости системы управления рисками на принятие решений в эпоху искусственного интеллекта

Е.В. Трофимова, ДРТ

В статье Е.В. Трофимовой (ДРТ) рассматриваются результаты исследования зрелости риск-менеджмента в российских нефинансовых организациях и практика применения ИИ, в том числе в страховой отрасли. Материал ставит вопрос о том, насколько формализованные процедуры управления рисками действительно влияют на принятие решений, в какой мере риск-функция обеспечена необходимыми полномочиями и ресурсами и насколько вопросы управления рисками находятся в поле внимания руководства. Возможности технологии рассматриваются в связи с готовностью самой риск-функции к изменениям.

Управление модельным риском: от финансовых моделей к рискам искусственного интеллекта

П.А. Разумовский, НКЦ

П.А. Разумовский (НКЦ) раскрывает устройство системы управления модельным риском. Модель рассматривается вместе с исходными данными, ее реализацией в информационной системе и конкретным бизнес-процессом. Автор показывает роль реестра моделей, классификации по значимости, мониторинга, независимой валидации и управленческой отчетности. Этот подход позволяет включать риски ИИ в действующую систему управления рисками и сохранять ответственность владельца процесса за результат.

Аудит оценки кредитного риска на основе ПВР: от проверки моделей к оценке целостной рейтинговой системы

Д.С. Семилетов, Л.Е. Кузнецов, Банк ВТБ

Д.С. Семилетов и Л.Е. Кузнецов (Банк ВТБ) переводят общую логику модельного контроля в практику внутреннего аудита подхода на основе внутренних рейтингов. Проверка охватывает рейтинговую систему целиком: исторические данные, оценки компонентов кредитного риска, валидацию, консервативные надбавки, экспертные корректировки, промышленную реализацию и использование результатов. Особое внимание уделяется воспроизводимости расчетов и оценке влияния выявленных недостатков на риск и капитал.

Безопасное масштабирование искусственного интеллекта в банке: риски, модель угроз и контуры контроля

А.М. Валиуллин, И.Л. Вахламов, Газпромбанк

В материале А.М. Валиуллина и И.Л. Вахламова (Газпромбанк) рассмотрены условия безопасного масштабирования ИИ в банке. Представлена взаимосвязь рисков данных, качества моделей, информационной безопасности, внешних поставщиков и автономных агентов. Интенсивность контроля предлагается соотносить с влиянием системы на клиентов, финансовые результаты и критические процессы. Практические инструменты включают инвентаризацию, классификацию, модель угроз, валидацию и мониторинг.

От политики запретов к управляемому искусственному интеллекту: угрозы безопасности ИИ-систем и модель их контроля в финансовых организациях

Р.М. Гусейнов, ПК «РАД КОП»

Р.М. Гусейнов (ПК «РАД КОП») рассматривает угрозы безопасности ИИ-систем и переход к управляемому использованию технологии. В фокусе статьи – теневой ИИ, работа с внешними источниками данных, полномочия агентов, безопасность компонентов и возможность восстановить цепочку действий. Для внутреннего аудита предложен подход, связывающий конкретный риск с мерой контроля и доказательством ее работы.

Повышение требований к экспертизе и компетенциям сотрудников, создающих и проверяющих продукты искусственного интеллекта: обратная сторона ожидаемой рынком экономии ресурсов

А.А. Пенин, А.М. Козлов, А.О. Мацковская, Керт

Завершает содержательную часть статья А.А. Пенина, А.М. Козлова и А.О. Мацковской (Керт) о требованиях к сотрудникам, создающим и проверяющим продукты ИИ. Авторы рассматривают риск утраты экспертизы, условия содержательной независимой проверки и перераспределение затрат при автоматизации. Предложенные матрицы ролей и компетенций помогают обсудить, какими знаниями и полномочиями должны обладать участники процесса. Результаты приведенного в статье пилотного опроса имеют диагностический характер и не предназначены для статистических обобщений на весь рынок.

Представленные материалы формируют основу для обсуждения дальнейшего развития методологии, контрольных процедур и компетенций. Сведения о работе профессионального сообщества и возможностях участия приведены в разделе «Экосистема Экспертного совета».



”

Можно констатировать, что технологии ИИ пока не получили повсеместного распространения в риск-менеджменте, однако их внедрение является закономерным этапом развития данной функции.

Е.В. ТРОФИМОВА
Член Правления ДРТ

ОТ ФОРМАЛЬНОЙ ЗРЕЛОСТИ К РЕАЛЬНОЙ ЭФФЕКТИВНОСТИ: КАК ОЦЕНИТЬ ВЛИЯНИЕ ЗРЕЛОСТИ СИСТЕМЫ УПРАВЛЕНИЯ РИСКАМИ НА ПРИНЯТИЕ РЕШЕНИЙ В ЭПОХУ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Аннотация

В статье на основе исследования уровня зрелости риск-менеджмента в российских нефинансовых компаниях и анализа вопросов в области страховой отрасли рассматриваются перспективы внедрения искусственного интеллекта (ИИ) в управление рисками. Отмечается, что общий уровень зрелости практики управления рисками в российских нефинансовых компаниях остается невысоким и для развития функции часто не хватает не только должного внимания со стороны высшего руководства, но и ресурсного обеспечения, что повышает значимость современных аналитических инструментов, способных оптимизировать отдельные процедуры риск-менеджмента. Рассмотрена практика применения ИИ российскими и международными компаниями в рамках управления рисками. Кроме того, проанализированы практика, барьеры и риски применения ИИ в страховой отрасли, где технология уже используется большинством участников рынка. Выделены возможности ИИ по обработке больших данных, способствующих переходу компаний к предиктивной модели управления рисками¹. В качестве сдерживающих факторов выделены недостаток внимания со стороны руководства, дефицит ресурсов, сложности масштабирования и необходимость в формировании системного контроля за рисками, возникающими при использовании ИИ.

Ключевые слова: риски, исследование, искусственный интеллект, риск-культура, уровень зрелости, страхование.

Коды JEL: G32, O33.

Компания ДРТ ежегодно проводит исследование уровня зрелости риск-менеджмента в российских компаниях при поддержке Ассоциации риск-менеджмента «Русское общество управления рисками»². Цель исследования – определить текущий уровень зрелости практики управления рисками в российских нефинансовых организациях и выявить тенденции ее развития. В 2026 году в исследовании приняли

¹ Предиктивная модель управления рисками – подход, при котором система управления рисками ориентирована не на фиксацию уже реализовавшихся событий, а на прогнозирование вероятности и последствий рисков на основе данных и моделей (примечание Редакционной коллегии).

² Ассоциация риск-менеджмента «Русское общество управления рисками» (РусРиск) – профессиональное объединение специалистов по управлению рисками, действующее с 2003 года (примечание Редакционной коллегии).

участие риск-менеджеры, а также представители функции внутреннего контроля и внутреннего аудита из 71 нефинансовой организации России. Исследование проведено на основании опроса, включавшего 46 вопросов, затрагивающих отдельные элементы практики управления рисками, ключевые ограничения и перспективы их развития.

Несмотря на растущее внимание к теме, общий уровень зрелости управления рисками в нефинансовых организациях России из года в год меняется незначительно: в текущем году показатель составил 0,4 из 1 возможного, что немного превышает результаты прошлых лет. Наиболее зрелую практику управления рисками в выборке 2026 года продемонстрировали компании сектора энергетики и возобновляемых источников энергии (0,69), а также таких отраслей, как высокие технологии, телекоммуникации и СМИ, пищевая и нефтегазовая промышленность. При этом следует отметить, что на результаты исследования и динамику изменения показателей оказывает влияние структура выборки (по отраслевой принадлежности и масштабам деятельности компаний).

Многолетние наблюдения позволяют сформировать устоявшиеся закономерности и тенденции. Так, наиболее развитые практики риск-менеджмента характерны для крупных компаний с выручкой от 500 млрд руб. и штатом от 5 тыс. человек. Это объяснимо: чем крупнее компания, тем сложнее ее структура управления и бизнес-процессы, а значит, выше объективная потребность в системе управления рисками и одновременно шире ресурсная база для развития этой функции.

Также ключевые приоритеты в развитии риск-менеджмента остались неизменными: формирование сильной культуры управления рисками, практика определения и применения риск-аппетита, развитие инструментов анализа и оценки рисков.

Среди наиболее зрелых элементов управления рисками – взаимодействие и обмен информацией, а также роль высшего руководства, совета директоров и его комитетов в управлении рисками. При этом результаты исследования свидетельствуют о сохранении формального отношения к управлению рисками со стороны совета директоров и ограниченной роли риск-менеджеров в процессах принятия решений.

Так, практически во всех компаниях (94%) вопросы управления рисками рассматриваются коллегиальными органами при совете директоров или правлении, однако лишь в трети организаций создан отдельный комитет по управлению рисками, заседания которого проводятся не реже 1 раза в квартал. При этом более чем в половине компаний (58%) совет директоров преимущественно рассматривает карту рисков и отчетность согласно плану, не задавая дополнительных вопросов по предоставленным материалам. И лишь в четверти компаний совет директоров регулярно требует проведения анализа рисков при рассмотрении ключевых проектов и решений (21%), а также переоценки риск-аппетита (16%).

Схожая ситуация складывается и в области стратегического планирования и бюджетирования – одной из ключевых функций совета директоров и правления соответственно. Около трети российских организаций не пересматривали стратегические цели более 2 лет или не анализировали связанные с ними риски, а среди тех, кто все же оценивает влияние рисков на стратегию, лишь треть проводит количественный анализ такого влияния. При бюджетном планировании доля компаний, учитывающих влияние рисков на показатели бюджета, еще ниже – всего 19%. Вместе с тем в целом нефинансовые компании стали чаще проводить анализ рисков при планировании, но в большинстве случаев – укрупненный, на основе качественной оценки, без применения инструментов моделирования.

Аналогичная тенденция прослеживается и при принятии ключевых решений: выросла доля организаций, где анализ рисков, связанных с принимаемыми решениями, не проводится или проводится неформально (34%), и лишь треть компаний (и их доля сократилась по сравнению с показателем за 2025 год) при принятии наиболее важных решений осуществляли количественную оценку влияния рисков с возможностью пересмотра решения по результатам такой оценки.

Важную роль в принятии взвешенных и обоснованных решений играет аналитическая поддержка риск-менеджера. Но, как показывает исследование, его участие в процессе принятия ключевых решений остается ограниченным. Так, ключевые решения принимаются без него в 39% компаний, правом вето в отношении ключевых решений обладают риск-менеджеры лишь в 8% компаний. Также сокращается количество компаний, в которых риск-менеджеры участвуют в заседаниях коллегиальных органов (24% против 36% годом ранее) и подготовке заключений по рискам (37% против 45%).

Ограничение роли риск-менеджера обусловлено двумя основными причинами: недостаточным вниманием со стороны высшего руководства к значению риск-аналитики при принятии решений и нехваткой ресурсов, необходимых для поддержания и развития функции управления рисками. В этих условиях особую роль приобретает применение функций по управлению рисками современных инструментов оценки и управления рисками.

Одним из направлений такого развития становится внедрение технологий ИИ.

По состоянию на 2026 год 13% компаний применяют ИИ в процессах управления рисками на общекорпоративном уровне, а еще 34% планируют сделать это в ближайший год. Технологии ИИ позволяют риск-менеджерам сократить объем рутинной работы, ускорить обработку больших массивов данных и повысить эффективность их анализа, однако половина компаний признает, что затрачиваемые ресурсы пока не полностью оправдывают ожидания. Чаще всего ИИ применяется для выявления и описания рисков и мероприятий по управлению ими (в 75% компаний), а также для разработки нормативно-методологических документов (у 63% организаций). Ответственность за внедрение и развитие ИИ в этой сфере распределена между подразделением по управлению рисками (38%) и профильными бизнес-подразделениями (39%) – в зависимости от того, для каких категорий рисков применяется технология.

Международная практика показывает, что развитие ИИ и машинного обучения, основанных на широком доступе к внешним данным, открывает возможность перехода к предиктивному риск-менеджменту. Одновременно цифровизация создает новые уязвимости и повышает требования к прозрачности процессов. Успех внедрения ИИ в систему риск-менеджмента во многом зависит от пересмотра самого понятия модели риск-менеджмента и ее составляющих, а также от внедрения структурированного управления жизненным циклом модели. Применение ИИ в анализе рисков охватывает полный жизненный цикл технологии – от подготовки данных, разработки и обучения модели до ее эксплуатации, взаимодействия с поставщиками и использования открытого кода.

Несмотря на растущий интерес к теме, значительная часть организаций пока не выстроила системный контроль за использованием ИИ: 66% компаний, применяющих ИИ или реализующих пилотные проекты, не имеют специальных правил управления связанными с ним рисками, а 42% не оценивают и не классифицируют уровень риска используемых моделей и сценариев. Среди ключевых рисков применения ИИ – использование устаревших или неполных данных, сложность интерпретации решений алгоритмов, сбои в работе сервисов, утечка персональных данных и несоблюдение регуляторных требований. Важным шагом к внедрению системного управления в этой области в нашей стране стали Методические рекомендации Банка России от 16.06.2026 № 3-МР, систематизировавшие специфические риски безопасности ИИ, тактики атак и меры защиты.

Отдельный интерес представляет практика применения ИИ в страховой отрасли, где технология уже используется большинством участников рынка. Согласно исследованию, в котором приняли участие страховые компании России и СНГ, а также негосударственные пенсионные фонды с суммарной долей рынка около 30–40%, большинство организаций применяют ИИ прежде всего для взаимодействия с клиентами (более 60%), решения технических задач (более 60%), выявления признаков мошенничества, разработки обучающих материалов для сотрудников и поиска потенциальных клиентов. Таким образом, ИИ пока используется не как замена сотрудников, а как инструмент ускорения аналитики, продаж, маркетинга и обработки клиентских обращений.

По оценкам опрошенных компаний, применение ИИ в первую очередь положительно повлияло на эффективность операционных процессов и качество услуг. Более 70% респондентов считают, что преимущества использования ИИ перевешивают сопряженные с его использованием риски.

При этом главными барьерами для внедрения ИИ в области страхования остаются необходимость обучения персонала работе с этим инструментом, отсутствие соответствующих компетенций в компаниях, нехватка финансовых ресурсов и недостаточно высокое качество данных. Большинство организаций, уже применяющих ИИ, используют для этого инструменты собственной разработки и собственные данные. В то же время часть компаний отметили, что не имеют сейчас возможности использовать ИИ именно по причине отсутствия средств для развертывания своих внутренних моделей. Проблема качества данных является общей для страховщиков во всем мире – согласно международным исследованиям, только 12% участников этой отрасли оценивают уровень готовности своих данных для применения ИИ как высокий.

Показательно, что, хотя половина респондентов признают серьезные последствия недостаточно быстрого восполнения компетенций уходящих опытных сотрудников, ни один из них не считает, что эту проблему можно решить с помощью ИИ. Требования к квалификации персонала после внедрения технологии при этом практически не изменились: об этом сообщили 86% опрошенных и лишь 14% отметили их повышение, а случаев снижения требований зафиксировано не было.

Технологии ИИ активно внедряются страховщиками и в процессы урегулирования убытков, взаимодействия с клиентами, работу антифрод-систем и решение различных технических задач, при этом приоритеты подразделений различаются. Специалисты по урегулированию убытков в первую очередь ориентированы на повышение эффективности обработки страховых случаев (72%) и сокращение сроков их урегулирования (64%), тогда как повышение удовлетворенности клиентов называют приоритетом 45% респондентов этой группы. В ИТ-подразделениях приоритеты смещены в сторону масштабирования ИИ внутри компании (65%), снижения операционных расходов (48%) и модернизации ИТ-архитектуры (45%).

Можно констатировать, что технологии ИИ пока не получили повсеместного распространения в риск-менеджменте, однако их внедрение является закономерным этапом развития данной функции. Сдерживающим фактором остается недостаточное инвестирование в это направление, во многом обусловленное невысоким уровнем внимания к нему со стороны высшего руководства организаций. Вместе с тем, учитывая вызовы, с которыми сталкивается функция управления рисками, дальнейшее развитие темы применения ИИ представляется закономерным и неизбежным.

Банки и финансовые организации уже на протяжении длительного времени используют возможности ИИ в управлении рисками: технологии автоматизации и алгоритмы машинного обучения применяются ими для упрощения процесса принятия решений, снижения кредитных рисков при предоставлении финансовых услуг и адаптации к требованиям клиентов в условиях

высокой конкуренции. В риск-менеджменте нефинансовых организаций использование ИИ пока распространено в меньшей степени, однако востребованность технологии для управления отдельными видами рисков постепенно растёт.

Как показали исследования, способность ИИ анализировать значительные объёмы информации существенно повышает эффективность работы с данными – их формирования, систематизации, обработки и архивирования, а также построения предиктивных моделей. Конечная цель применения этих технологий – трансформация системы управления рисками в более предиктивную модель, ориентированную на выявление значимых факторов риска в области отдельных видов рисков, повышение точности оценки рисков и качества принимаемых бизнес-решений. ИИ способствует ускорению операций с данными и повышению качества бизнес-аналитики, но требует системного контроля на всех этапах использования.

Вместе с тем распространение генеративного ИИ пока не привело к сопоставимому по масштабу финансовому эффекту. На этапе первичного внедрения технология способна заметно ускорить решение отдельной задачи, однако при переходе к полноценным проектам и попытках локального улучшения процессов организации сталкиваются с необходимостью прохождения этапов согласования и тестирования, что ограничивает скорость масштабирования достигнутых результатов.

Список источников

1. [Исследование уровня зрелости риск-менеджмента в российских компаниях в 2026 году](#) // Группа ДРТ. 2026 год (дата обращения: 26.08.2026).
2. Анализ текущего состояния внедрения генеративного ИИ (Gen AI) в страховых компаниях // ДРТ. 2026 год.
3. [Страховка от ИИ](#). Ведомости. 2 марта 2026 года (дата обращения: 27.08.2026).
4. Н. Балдин, А. Янин [«Инновации в страховании: скорость, технологии, искусственный интеллект»](#) // Рейтинговое агентство «Эксперт РА». 10 сентября 2025 года (дата обращения: 26.08.2026).
5. [Realize the Full Potential of Artificial Intelligence](#) // Committee of Sponsoring Organizations of the Treadway Commission (COSO), Deloitte. 2021 (дата обращения: 26.08.2026).
6. [Evolving model risk management in the age of AI](#). McKinsey & Company. 22 July 2026 (дата обращения: 26.08.2026).
7. Suhrada S., Casaceli S., Jain D. [Are insurers truly ready to scale gen AI?](#) Deloitte Insights. 4 April 2025 (дата обращения: 26.08.2026).

[Ознакомиться с презентацией к мастер-классу](#)

[Ознакомиться с презентацией к выступлению на заседании Экспертного совета](#)



”
Владелец модели совместно с профильными функциями должен до внедрения определить, какие нефинансовые риски усиливаются или возникают впервые, и встроить соответствующие контроли в процесс применения модели.

П.А. РАЗУМОВСКИЙ
НКЦ, риск-менеджмент

УПРАВЛЕНИЕ МОДЕЛЬНЫМ РИСКОМ: ОТ ФИНАНСОВЫХ МОДЕЛЕЙ К РИСКАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Аннотация

В статье представлен подход НКЦ¹ к управлению модельным риском как составной частью нефинансового риска. Показано, что модель включает не только алгоритм, но также исходные данные и реализацию алгоритма в ИТ-системе, поэтому негативные последствия могут формироваться на пересечении операционных, информационных, регуляторных, этических, комплаенс-, ИТ- и киберрисков. Особое внимание уделено интеграции рисков искусственного интеллекта в действующую систему управления рисками, выбору уровня участия человека с учетом критичности процесса и распределению функций в модели трех линий. Ключевыми инструментами являются единый реестр моделей, риск-ориентированная классификация, мониторинг, независимая валидация и регулярная управленческая отчетность. Такой подход позволяет масштабировать применение ИИ без создания параллельной системы контроля и без размывания ответственности за итоговое решение.

Ключевые слова: модельный риск, искусственный интеллект, нефинансовые риски, три линии защиты, валидация, мониторинг моделей.

Коды JEL: G32, D81, M15, M42, O33.

Модельный риск шире риска ошибки алгоритма

Модель понимается как количественный метод, система или подход, который на основе статистических, экономических, финансовых или математических теорий и методов преобразует исходные данные в количественные оценки. **Модельный риск** – это риск негативных последствий при принятии решений на основе результатов модели либо в рамках процессов, в которых модель участвует. Последствия проявляются не только в прямых финансовых потерях: модель может не обеспечить целевой результат бизнес-процесса, вызвать операционный сбой, регуляторное нарушение или репутационный ущерб.

Для практического управления важно рассматривать модель как целостную конструкцию из следующих трех взаимосвязанных элементов:

- алгоритм преобразования данных и заложенные в нем допущения;
- исходные данные, их качество, полнота, актуальность и происхождение;
- реализация модели в ИТ-системе, включая инфраструктуру, интерфейсы, права доступа и процедуры изменений.

¹ НКО НКЦ (АО) входит в Группу «Московская Биржа» и выполняет функции центрального контрагента: становится стороной по каждой сделке, заключаемой на бирже, принимая на себя кредитный риск участников клиринга (примечание Редакционной коллегии).

Ошибка в любом из элементов способна исказить результат даже при математически корректном алгоритме. Поэтому модельный риск целесообразно относить к операционному риску, сохраняя при этом связи с другими видами финансовых и нефинансовых рисков.

Задача модели определяет профиль нефинансовых рисков

Любая модель применяется не сама по себе, а для решения конкретной задачи в определенном бизнес-процессе. Бизнес-процессу свойственны конкретные нефинансовые риски, определяемые исходя из целей, которые достигаются в рамках этого бизнес-процесса. Поэтому оценка модельного риска не должна ограничиваться качеством алгоритма или общим уровнем модельного риска. Необходимо анализировать, как именно использование модели меняет профиль других нефинансовых рисков в выбранном сценарии, какие новые способы реализации риска возникают и какие дополнительные меры контроля требуются. Одна и та же модель в разных процессах может иметь принципиально различный набор последствий и контролей.

- **Взаимодействие с клиентами.** Использование генеративных языковых моделей (GLM²) для подготовки ответов клиентам потенциально создает этические, комплаенс- и репутационные риски. Модель может сформировать некорректный, предвзятый, неуместный или вводящий в заблуждение ответ, раскрыть лишнюю информацию либо дать обещание, не соответствующее правилам компании. Дополнительные меры могут включать определение допустимых сценариев, фильтрацию содержания, обязательное участие человека в чувствительных случаях, журналирование диалогов³ и регулярный контроль качества ответов.
- **ИТ-разработка и сопровождение.** Применение GLM в качестве Code Agent⁴ и других помощников разработчиков, аналитиков и тестировщиков создает новые формы технологического риска и риска информационной безопасности. Среди них – генерация уязвимого или некорректного кода, включение небезопасных зависимостей, неконтролируемое изменение программных компонентов, раскрытие исходного кода или конфиденциальных данных через запросы к модели. Поэтому требуются дополнительные инструменты: использование утвержденных моделей в защищенном контуре, разграничение доступа, запрет передачи чувствительных данных, обязательное ревью и тестирование кода, автоматизированный анализ уязвимостей и полная прослеживаемость внесенных изменений.
- **Комплаенс и выполнение регуляторных требований.** Если модель используется для интерпретации требований, проверки операций, подготовки отчетности или рекомендаций по соблюдению норм, ее ошибка может привести к пропуску нарушения, необоснованному ограничению либо неверному регуляторному выводу. Контроль должен обеспечивать актуальность нормативной базы, объяснимость и воспроизводимость результата, сохранение исходных данных и логов, а также обязательную экспертную проверку юридически или регуляторно значимых решений.

² В настоящей статье GLM (от англ. Generative Language Model) используется автором для обозначения генеративных языковых моделей. В других материалах выпуска применительно к рассматриваемым автором моделям используется термин «большие языковые модели» (LLM, от англ. Large Language Model). Эти понятия характеризуют разные признаки: «генеративные» указывает на способность модели создавать текст, а «большие» – на масштаб модели и объем данных, используемых при ее обучении. Одна модель может обладать обоими признаками, однако эти термины не являются полными синонимами (примечание Редакционной коллегии).

³ Под журналированием диалогов понимается фиксация и сохранение истории взаимодействия с системой – запросов пользователей и ответов модели – для последующего контроля качества и разбора инцидентов (примечание Редакционной коллегии).

⁴ Code Agent (от англ. code – программный код, agent – агент) – ИИ-агент для разработки программного обеспечения, то есть инструмент на основе языковой модели, который по заданию разработчика генерирует, проверяет и изменяет программный код (примечание Редакционной коллегии).

Ключевой вывод. Набор мер управления определяется связкой «модель – задача – бизнес-процесс – последствия». Владелец модели совместно с профильными функциями должен до внедрения определить, какие нефинансовые риски усиливаются или возникают впервые, и встроить соответствующие контроли в процесс применения модели.

Риск ИИ – часть действующей системы управления рисками

ИИ – генеративные языковые модели, машинное обучение и нейросети – является подвидом моделей и несет модельный риск. Как показано в разделе выше, риск ИИ находится на стыке модельного, операционного, информационного, регуляторного, этического, ИТ- и киберрисков, так как модель на базе ИИ является частью конкретного бизнес-процесса, которому свойственны различные нефинансовые риски. По этой причине управление им должно быть интегрировано в существующую систему управления рисками, а не вынесено в отдельный контур. НКЦ является частью Группы «Московская Биржа» и использует внедренный в группе подход к управлению рисками ИИ в целом.

Степень автономности ИИ является главным фактором в определении уровня риска ИИ. Чем выше критичность процесса для компании, тем больше контроля за решениями и действиями ИИ требуется со стороны человека:

- для критичных процессов и систем с высоким риском предусматриваются ручная валидация и утверждение решения модели;
- для систем среднего и низкого риска применяется мониторинг человеком действий модели и контроль отклонений;
- в некритичных процессах, где требуется быстрый или непрерывный анализ больших массивов данных, допустима более высокая автономия либо контроль агентами более высокого уровня.

Уровень человеческого контроля должен быть заранее определен, документирован и согласован с допустимым уровнем риска.

Три линии: ответственность не растворяется между функциями

Модель трех линий является стандартом в области управления рисками. Она является основой для управления модельным риском в НКЦ.

РАСПРЕДЕЛЕНИЕ ФУНКЦИЙ УПРАВЛЕНИЯ МОДЕЛЬНЫМ РИСКОМ ПО ТРЕМ ЛИНИЯМ

Табл. 1

ПЕРВАЯ ЛИНИЯ. ВЛАДЕНИЕ И ПРИМЕНЕНИЕ	ВТОРАЯ ЛИНИЯ. МЕТОДОЛОГИЯ И НЕЗАВИСИМЫЙ КОНТРОЛЬ	ТРЕТЬЯ ЛИНИЯ. НЕЗАВИСИМАЯ ОЦЕНКА
Участники		
• Владелец модели • Пользователь модели • Разработчик модели	• Профильные подразделения по соответствующим рискам: операционные и интегрированные риски, комплаенс, информационная безопасность, ИТ • Валидаторы	• Служба внутреннего аудита
Функции		
• Решения об использовании, обновлении и выводе из эксплуатации • Разработка и доработка • Мониторинг и анализ причин негативных результатов	• Методология оценки и управления модельным риском • Ведение реестра моделей • Отчетность • Организация и проведение валидации наиболее рискованных моделей	• Стандартные процедуры внутреннего аудита системы управления модельным риском и отдельных моделей

Подразделения первой линии отвечают за достижение результата бизнес-процесса в широком плане. Именно эти подразделения принимают базовые решения о необходимости использования модели для достижения конкретной цели: увеличения объемов бизнеса, повышения точности оценки риска, повышения операционной эффективности, совершенствования процессов ИТ-производства или обеспечения кибербезопасности. То есть подразделение первой линии с функциями владельца модели – это любое подразделение внутри организации вне зависимости от того, какие функции оно выполняет (бизнес-подразделение, операционное подразделение, подразделение риск-менеджмента, ИТ, комплаенса, информационной безопасности). Главный критерий в данном случае – бизнес-процесс, в котором принимается решение об использовании модели для достижения результата. Владелец модели может передать функции разработки моделей профессиональному разработчику или внешнему подрядчику. Однако ответственность за конкретный результат бизнес-процесса все равно остается на владельце модели, поэтому именно владелец обеспечивает ключевые решения по модели: разработку, доработку, мониторинг и анализ результатов модели в контексте результата бизнес-процесса полностью. Владелец модели определяет ресурсы, необходимые для разработки, внедрения и поддержки модели. Использование сложных, продвинутых подходов, включая ML-модели⁵, обычно связано с высокими требованиями к экспертизе разработчиков, требованиями к мониторингу и интерпретации модели, дополнительными расходами на ИТ-сопровождение. Если это не приводит к улучшению результата бизнес-процесса, чтобы экономически окупить эти вложения, то тогда владельцу модели есть смысл ограничиться более простыми моделями.

Вторая линия формирует стратегию и общий подход в управлении модельным риском, ведет реестр моделей, организует отчетность и независимую валидацию моделей с наибольшим риском. Ее работа кросс-функциональна: операционные риски обеспечивают синхронность с подходами к управлению со всеми операционными и нефинансовыми рисками, интегрированные риски обеспечивают консистентность в части оценок экономического капитала⁶, риск-аппетита и прочих инструментов для всех рисков, ИТ- и информационная безопасность – технологическую устойчивость и защищенность, комплаенс – соответствие требованиям, валидаторы – качество модели и обоснованность ее использования. В НКЦ модельный риск относится к числу операционных рисков, однако подразделением, отвечающим за организацию функции управления рисками, является одно из управлений в рыночных рисках. Это связано с тем, что ключевым риском для клиринговой организации является рыночный риск и самое большое количество моделей и максимальное влияние на организацию в целом оказывают модели рыночных рисков. Соответственно, в Департаменте рыночных рисков сосредоточена достаточная экспертиза в области математики, статистики, моделирования для суточных решений по моделям и управлению модельным риском.

Третья линия независимо оценивает полноту и эффективность всей системы управления модельным риском по аналогии со всеми другими видами риска.

Жизненный цикл, реестр, мониторинг и валидация

Ключевым инструментом управления модельным риском выступает единый реестр моделей. Он должен включать кандидатов на место модели, действующие модели, сведения о владельцах

⁵ ML-модели (от англ. Machine Learning) – модели машинного обучения, то есть модели, использующие алгоритмы, которые выявляют закономерности в данных и на их основе формируют прогнозы, оценки или решения (примечание Редакционной коллегии).

⁶ Экономический капитал – внутренняя оценка объема капитала, необходимого организации для покрытия принятых рисков с заданной степенью уверенности; отличается от регуляторного капитала, рассчитываемого по требованиям надзорного органа (примечание Редакционной коллегии).

и значимости, результаты мониторинга моделей, сроки и результаты валидаций, рекомендации по итогам валидации, их критичность и статусы исполнения. Реестр обеспечивает управляемость полного жизненного цикла: от инициирования разработки и внедрения до применения, изменений и вывода из эксплуатации.

Интенсивность контрольных мероприятий, проводимых конкретной моделью, определяется значимостью данной модели для организации и соответствующим бизнес-процессом. Контрольные мероприятия всегда связаны с дополнительными расходами организации, поэтому дополнительные расходы имеет смысл нести в том случае, когда они защищают от значительного объема последствий. В противном случае это лишь замедляет и усложняет систему принятия решений. Для наименее значимых моделей достаточно завести модель в реестр, чтобы на регулярном цикле мониторить уровень важности модели и не пропустить ситуацию, когда модель может увеличить свое значение для организации. По мере роста потенциального влияния появляются требования к системе мониторинга за моделью, и далее усиливаются требования к первичной и регулярной валидации, глубине мониторинга, независимости проверки и уровню утверждения. Выбор внутренней или внешней валидации зависит от наличия отраслевого стандарта и специфики модели: внутренние эксперты лучше учитывают уникальные процессы организации, а внешняя оценка дает рыночный бенчмарк⁷ и обязательна там, где это установлено внешними требованиями. Например, НКЦ в соответствии с Положением Банка России от 01.11.2018 № 658-П⁸ обязан раз в 3 года проводить внешнюю валидацию моделей маржирования⁹.

Мониторинг и валидация решают разные задачи. Мониторинг первой линии оперативно отслеживает качество работы модели с установленной для нее периодичностью. Валидация второй линии представляет собой более глубокую независимую проверку качества модели, условий ее применения и достаточности мониторинга; результатом становятся заключение и рекомендации по улучшению. Набор метрик зависит от вида модели, но обычно включает показатели качества данных, точности и стабильности результата, технологической доступности, производительности и восстановления.

Мониторинг и валидация: разные цели и уровни ответственности

Мониторинг – инструмент владельца модели, то есть первой линии. Он должен объединять несколько групп показателей: метрики качества работы алгоритма и исходных данных; технологические метрики работоспособности модели в конкретной ИТ-системе, включая доступность, производительность и восстановление; а также дополнительные метрики смежных нефинансовых рисков, которые возникают или усиливаются в выбранном сценарии применения модели. Например, для клиентского GLM это могут быть показатели некорректных или неуместных ответов, а для Code Agent – результаты контроля качества и безопасности создаваемого кода. Основные ресурсы требуются при разработке, настройке и внедрении мониторинга. Далее владелец модели регулярно отслеживает динамику показателей и на этой

⁷ Бенчмарк (от англ. benchmark) – эталонный ориентир для сравнения; здесь – сопоставление модели организации с практикой рынка (примечание Редакционной коллегии).

⁸ На дату подготовки настоящего выпуска действует Положение Банка России от 01.11.2018 № 658-П. С 01.01.2027 оно утрачивает силу в связи с вступлением в силу Положения Банка России от 30.09.2025 № 869-П (примечание Редакционной коллегии).

⁹ Модели маржирования – модели расчета размера обеспечения, которое участник клиринга вносит центральному контрагенту для покрытия возможных потерь при неисполнении обязательств (примечание Редакционной коллегии).

основе принимает решения о продолжении применения, ограничениях, доработке или выводе модели из эксплуатации.

Валидация – инструмент второй линии: независимая содержательная проверка модели, условий ее использования и достаточности установленных контролей. Ее целесообразно применять прежде всего к моделям с критичным влиянием на организацию. В отличие от регулярного мониторинга, валидация является дорогостоящей процедурой, требует качественной математической, предметной и технологической экспертизы, а ее результаты оформляются в виде независимого заключения и рекомендаций.

Управленческая отчетность превращает контроль в решения

В рамках регулярной отчетности на Правление и иные органы управления компании выносятся агрегированная информация о состоянии модельного риска: состав и динамика реестра по уровням значимости, индикативные показатели риск-аппетита, проблемные модели, результаты мониторинга, планы и итоги внутренней и внешней валидации, а также статус исполнения рекомендаций. Такая отчетность связывает технические результаты контроля с управленческими решениями о допустимости дальнейшего использования, доработке, ограничении или выводе модели из эксплуатации.

От финансовых моделей к моделям искусственного интеллекта

Исторически описанный подход к управлению модельным риском разрабатывался и внедрялся в НКЦ прежде всего для моделей финансовых рисков. Для клиринговой организации наиболее критичны рыночные и иные финансовые риски: модели используются для их оценки и управления ими, расчета требований к обеспечению и выполнению регуляторных требований. Поэтому именно в этой области первоначально были сформированы реестр моделей, риск-ориентированная классификация, мониторинг и независимая валидация.

Вместе с тем эти инструменты универсальны и могут применяться к моделям искусственного интеллекта. Модели рыночного риска, построенные на базе машинного обучения или нейросетей, с точки зрения модельного риска принципиально не отличаются от более простых статистических моделей: они также регистрируются, классифицируются по значимости, сопровождаются метриками мониторинга и при критичном влиянии проходят валидацию. Сложность алгоритма может увеличить глубину контроля и требования к экспертизе, но не меняет базовую логику управления.

Инструменты на базе генеративных языковых моделей также включаются в реестр с указанием конкретного сценария применения и уровня значимости модельного риска. При классификации учитываются критичность бизнес-процесса, возможные последствия ошибки и степень автономности: используется ли результат под контролем человека или модель самостоятельно принимает решения и совершает действия. На этой основе владелец модели в рамках установленной методологии определяет достаточный набор метрик мониторинга и оценивает, требуется ли более глубокая независимая проверка. Если валидация необходима, владелец инициирует ее проведение, а содержательная оценка выполняется второй линией.

Ключевые выводы

- Модельный риск – это риск решения и бизнес-процесса, а не только математической ошибки.
- ИИ в виде генеративных языковых моделей или методов машинного обучения, нейросетей и так далее не отличается с позиции подхода и инструментов управления модельным риском от других, исторических более простых методов построения моделей.
- В части генеративных языковых моделей уровень автономии определяет уровень риска модели, который должен быть консистентен с критичностью процесса, в котором используется модель.
- Модель трех линий определяет ключевые функции в управлении модельным риском; вторая линия в области модельного риска более нагружена необходимостью кросс-функционального взаимодействия, обеспечивающего системное управление модельным риском с учетом ИТ-, кибер-, регуляторных, комплаенс-, этических и прочих рисков, возникающих вследствие применения конкретной модели.
- Единый инструментарий модельного риска применим и к статистическим моделям, и к ML, нейросетям и GLM; интенсивность контроля определяется не названием технологии, а значимостью модели и сценарием ее использования.
- Единый реестр, риск-ориентированная классификация, мониторинг, валидация и отчетность формируют замкнутый цикл управления модельным риском.

Список источников

1. НКО НКЦ (АО). Внутренние материалы по практике управления модельным риском. Не опубликованы в открытом доступе.
2. Положение Банка России от 01.11.2018 № 658-П «О требованиях к квалифицированному центральному контрагенту, порядке признания качества управления центральным контрагентом удовлетворительным, об основаниях и порядке принятия решения о признании качества управления центральным контрагентом неудовлетворительным, порядке доведения информации о принятом решении до центрального контрагента» // Вестник Банка России. 15 февраля 2019 года. [№ 11 \(2063\)](#) (дата обращения: 30.09.2026).

[Ознакомиться с презентацией](#)



Д.С. СЕМИЛЕТОВ

Руководитель Службы аудита процедур управления рисками Департамента внутреннего аудита Банка ВТБ (ПАО)



Л.Е. КУЗНЕЦОВ

Управляющий директор Службы аудита процедур управления рисками Департамента внутреннего аудита Банка ВТБ (ПАО)

Аудитор проверяет расчет показателей, выборки и пороговые значения, но не менее важна реакция на нарушение границ. Должны быть проведены анализ причин пробития контрольных показателей качества моделей, подготовлен план действий, определены ответственные и сроки.

АУДИТ ОЦЕНКИ КРЕДИТНОГО РИСКА НА ОСНОВЕ ПВР: ОТ ПРОВЕРКИ МОДЕЛЕЙ К ОЦЕНКЕ ЦЕЛОСТНОЙ РЕЙТИНГОВОЙ СИСТЕМЫ

Аннотация

Статья посвящена практике внутреннего аудита ПВР как целостной рейтинговой системы. Раскрыты процедуры проверки данных, PD-, LGD- и EAD-моделей, валидации, мониторинга, консервативных надбавок, применения моделей и ИТ-контура. Особое внимание уделено практическим аспектам аудита, риск-ориентированному планированию.

Ключевые слова: подход на основе внутренних рейтингов, ПВР, кредитный риск, внутренний аудит, рейтинговая система, модельный риск, валидация, качество данных, вероятность дефолта (PD), уровень потерь в случае дефолта (LGD), величина кредитного требования на момент дефолта (EAD).

Коды JEL: G21, G28, G32, M42, C58.

ПВР как объект внутреннего аудита

Подход на основе внутренних рейтингов (ПВР) – это способ оценки кредитного риска, основанный на экономико-статистических моделях, опирающихся на данные банка о заемщиках и кредитных требованиях. В ПВР риск-веса опираются на внутренние оценки, результат которых зависит от корректности сегментации заемщика, признания дефолта, качества данных, промышленной реализации моделей и других факторов. Ошибка в реализации ПВР может исказить кредитное решение, оценку эффективности деятельности банка и достаточность капитала. Поэтому Положение Банка России от 02.11.2024 № 845-П¹ предусматривает ежегодную проверку качества функционирования рейтинговых систем службой внутреннего аудита [1].

Объектом аудита является не отдельная математическая модель, а рейтинговая система в полном жизненном цикле. Корректная формула расчета параметра кредитного риска может давать ошибочный результат из-за неправильной сегментации, пропуска данных, несанкционированного изменения кода или применения модели за пределами ее области. И наоборот, методологически несовершенная модель может выглядеть приемлемо в отчетности из-за консервативных

¹ Положение Банка России от 02.11.2024 № 845-П «О порядке расчета величины кредитного риска банками с применением банковских методик управления кредитным риском и моделей количественной оценки кредитного риска» (далее – Положение № 845-П) (примечание Редакционной коллегии).

заглушек² или ручных корректировок. Задача внутреннего аудита – проверить всю цепочку от исходных данных для разработки или расчета модели до управленческого решения и расчета капитала. В сердце ПВР-подхода лежат модельные оценки компонентов кредитного риска, поэтому в аудите ПВР целесообразно опираться не только на нормативные документы Банка России, но и на международные стандарты оценки кредитного риска [3] и управления модельным риском [4].

Периметр и логика риск-ориентированной проверки

Первый риск, описанный выше, связан с ошибками в данных, сегментации и реализации модели. Второй риск – разрыв между моделью и бизнес-процессом. Модель может быть правильно разработана и валидирована, но фактически не использоваться при принятии решений либо использоваться с не предусмотренными методикой корректировками. Поэтому проверка должна охватывать не только расчет рейтинга, но и его дальнейшее движение: кто получает результат, в какие лимитные и кредитные процедуры он передается, какие решения запускает и как фиксируются случаи отклонения от модельного значения.

Третий риск – несогласованность между средой разработки и промышленной реализацией. Для его оценки недостаточно ознакомиться с актом внедрения. Необходимо подтвердить проведение приемо-сдаточных испытаний, участие владельца, разработчика, валидатора и ИТ-функции, полноту тест-кейсов³ и устранение замечаний. На выборке следует сопоставить результаты эталонного расчета с результатами промышленной системы. Даже небольшая ошибка округления или неверно перенесенный коэффициент при массовом применении может заметно повлиять на RWA⁴.

Риск-ориентированное планирование позволяет распределить ресурсы между этими направлениями. Приоритет получают модели и сегменты с наибольшим влиянием на капитал, значительной долей ручных корректировок, ухудшением показателей качества, изменениями портфеля или ИТ-архитектуры. Существенным фактором является и длительность периода с момента последней глубокой проверки: формальное ежегодное покрытие не должно приводить к тому, что сложные участки постоянно проверяются поверхностно.

Аудит данных для разработки моделей

Минимальный набор исторической информации включает базу дефолтов, историческую сегментацию и витрину платежей. Для оценки вероятности дефолта (PD) требуется определить числитель – количество дефолтов – и корректный знаменатель, то есть совокупность объектов соответствующего сегмента (заемщиков, кредитных требований), находившихся под наблюдением. Для LGD необходимы денежные потоки после дефолта, информация об обеспечении, расходах на взыскание и выходе из дефолта. Для EAD по возобновляемым продуктам требуются последовательные временные срезы задолженности и доступных лимитов.

Ключевой практический вопрос – сопоставимость исторического и текущего определения дефолта. Если критерии в прошлом отличались, реестр дефолтов должен быть реконструирован в соответствии с действующей методологией. Аудитор проверяет не только итоговый реестр,

² Консервативная заглушка – заведомо консервативное значение параметра, подставляемое в расчет вместо отсутствующих или недостоверных данных, чтобы исключить занижение оценки риска (примечание Редакционной коллегии).

³ Тест-кейс (от англ. test case) – заранее описанный набор входных данных и ожидаемого результата, используемый для проверки корректности расчета (примечание Редакционной коллегии).

⁴ RWA (от англ. Risk-Weighted Assets) – активы, взвешенные по риску; база для расчета нормативов достаточности капитала. Оценки компонентов кредитного риска по ПВР напрямую влияют на их величину (примечание Редакционной коллегии).

но и алгоритм пересборки: какие триггеры были доступны, как обработаны пропуски, на каком основании применялась экстраполяция и насколько консервативно учтены периоды, по которым восстановить отдельные события невозможно.

Далее оценивается формирование выборок. Обычно разделяются данные для обучения, тестирования и расчета центральной тенденции⁵. Тестовая выборка должна позволять независимо оценить качество разработанного алгоритма, а выборка для центральной тенденции – отражать долгосрочный уровень риска на доступном временном горизонте.

Репрезентативность нельзя подтверждать только совпадением продуктовых названий. Важно сопоставить распределения факторов, характеристики заемщиков, структуру обеспечения, каналы продаж и правила андеррайтинга. За длительный исторический период некоторые факторы могут утратить вариативность или перестать собираться, а новые продукты – изменить профиль текущего портфеля. Модель при этом будет хорошо объяснять историческую выборку, но слабо различать риск в актуальном портфеле.

Отдельный блок проверки – качество данных. Анализируются уровни пропусков, исключений и ручных корректировок в разрезе периодов, факторов, продуктов и сегментов. Агрегированный низкий процент пропусков может скрывать существенную проблему в отдельном высокорисковом продукте. Поэтому аудитор должен видеть не только среднее значение показателя, но и его распределение, динамику и материальность. Для критичных атрибутов проверяются установленные пороги качества и фактические действия при их нарушении.

Практически полезно воспроизвести формирование одной или нескольких модельных выборок из исходных систем. Такая процедура позволяет оценить data lineage⁶, выявить незафиксированные преобразования и понять, какие допущения встроены в код подготовки данных. Если полное воспроизведение слишком ресурсоемко, следует выбрать наиболее существенные участки. Невозможность воспроизведения данных должна быть объектом пристального внимания аудитора.

Ключевой вопрос проверки – корректность целевой переменной. Поэтому в фокусе внимания аудитора должны быть данные, используемые для калибровки модели. В частности, аудитор сопоставляет события дефолта с действующим определением, проверяет даты начала и окончания события дефолта, обработку повторных дефолтов и правила выхода из дефолта, сравнивает данные реестра дефолтов с модельной выборкой, анализирует решения комитета по дефолтам и полноту их обоснований.

Аудит разработки моделей

Важный вопрос для аудита ПВР-моделей – соответствие типа калибровки назначению модели. Для регуляторного ПВР внутренние рейтинги присваиваются на долгосрочной основе и должны учитывать возможное ухудшение экономических условий. Калибровка through the cycle⁷ сглаживает колебания фактической дефолтности по фазам цикла, тогда как point in time сильнее отражает текущую конъюнктуру.

⁵ Центральная тенденция – долгосрочный средний уровень дефолтности сегмента, к которому калибруются модельные оценки вероятности дефолта (примечание Редакционной коллегии).

⁶ Data lineage – прослеживаемость данных (примечание Редакционной коллегии).

⁷ Калибровка «на протяжении цикла» (от англ. through the cycle) ориентирована на долгосрочный средний уровень дефолтности и слабо реагирует на фазу экономического цикла; калибровка «на текущий момент» (от англ. point in time) отражает текущие экономические условия и быстрее меняется вместе с ними (примечание Редакционной коллегии).

Одним из ключевых фокусов аудита является воспроизводимость разработки модели. Отчет о разработке должен позволять восстановить состав выборки, перечень факторов, правила обработки выбросов и пропусков, алгоритм расчета скорингового балла, границы рейтинговых разрядов и калибровочные коэффициенты. Неясность отчета сама по себе является риском: она затрудняет независимую валидацию, поддержку модели и контроль существенных изменений.

Практическая проверка LGD начинается с отбора завершенных и незавершенных случаев взыскания. Для продуктов с долгим периодом возмещения, например ипотеки, наблюдение может продолжаться несколько лет. Если используются незавершенные случаи или экстраполяция будущих поступлений, аудитор оценивает статистическое основание метода экстраполяции, стабильность закономерностей во времени и чувствительность результата к допущениям. Слишком оптимистичная экстраполяция способна существенно занижить LGD.

При анализе данных, используемых для разработки моделей, особое внимание уделяется качеству денежных потоков. Из погашений должны быть исключены средства, фактически предоставленные самим банком в рамках рефинансирования или реструктуризации. При реализации залога возможны задвоения: поступление отражается как продажа актива и одновременно как погашение по ссудному счету. Проверяются полнота прямых и косвенных издержек, связанных с судебными и внесудебными процедурами, случаи учета положительных результатов, когда сумма взыскания превосходит размер вышедшей в дефолт задолженности.

Для LGD существенна оценка потерь в период экономического спада. Банк определяет периоды ухудшения, сравнивает соответствующий LGD со средней оценкой и при необходимости применяет надбавку. Аудитор проверяет, как надбавка инкорпорирована в модель, насколько корректно учтены исторические данные о возмещении кредитов за полный 5-летний период, который используется для моделирования LGD. Надбавка должна компенсировать наблюдаемое ухудшение, а не служить произвольным резервом консерватизма.

EAD оценивает величину кредитного требования на момент дефолта. Для кредитов с фиксированным графиком она определяется текущей задолженностью и будущими платежами. Наибольшая модельная сложность возникает по кредитным картам и другим возобновляемым линиям: заемщик может выбрать неиспользованный лимит незадолго до дефолта, а банк – изменить размер доступного лимита. Модель должна учитывать такую зависимость, включая полное использование лимита и возможное превышение установленных границ.

Аудит EAD требует большого количества временных срезов. Проверяются характеристики поведения заемщиков по картам, которые никогда не использовались, по картам с минимально выбранными лимитами, по картам с полностью выбранными лимитами, а также неавторизованные операции. Аномальные случаи (резкое повышение лимита, технический отрицательный неиспользованный лимит, начисления после дефолта) должны быть идентифицированы и обработаны по прозрачным правилам, а не удалены без анализа.

Для EAD также применяется оценка в условиях экономического спада. Аудитор сопоставляет средние и стрессовые значения кредитного конверсионного фактора⁸. Практический риск состоит в том, что низкочастотные, но крупные выборки лимита⁹ могут быть исключены как выбросы и тем самым искусственно снизить модельную оценку.

⁸ Кредитный конверсионный фактор (от англ. Credit Conversion Factor, CCF) – коэффициент, показывающий, какая доля неиспользованного лимита будет выбрана заемщиком к моменту дефолта (примечание Редакционной коллегии).

⁹ Выборка лимита – использование заемщиком ранее не выбранной части установленного кредитного лимита. В данном случае термин не связан со статистическим понятием выборки, используемым в других разделах статьи (примечание Редакционной коллегии).

Валидация и мониторинг качества моделей

Одно из направлений аудита – организационная независимость подразделения валидации. Проверяются подчиненность валидаторов, отсутствие их участия в разработке проверяемой модели, порядок согласования выводов и влияние результатов валидации на вознаграждение сотрудников.

Другое важное направление – полнота методологии. Стандарты валидации должны определять обязательные тесты, выборки, критерии приемлемости, порядок агрегирования результатов и основания итогового вывода. Аудитор сопоставляет методику с фактически выполненной работой и выясняет причины пропуска отдельных тестов. Особого внимания требуют ситуации, когда для нового вида модели применен общий шаблон, не учитывающий особенности целевой переменной или портфеля.

Еще одно ключевое направление – качество доказательств. Валидационная выборка должна быть сформирована независимо или по крайней мере ее происхождение должно быть проверено валидатором. Если данные полностью подготовлены разработчиком без независимых контролей, количественные тесты могут точно воспроизводить ошибочную базу. Поэтому аудитор анализирует data lineage выборки.

Мониторинг качества модели после внедрения должен выявлять ухудшение качества до того, как оно приведет к существенному искажению капитала или бизнес-решений. Приложение 4 к Положению № 845-П устанавливает контрольные показатели ранжирующей способности, точности калибровки, оценки надбавки за экономический спад, концентрации, стабильности популяции и экспертных корректировок [1]. Для ранее одобренных моделей дополнительно проводятся тесты, указанные в условиях действующего разрешения, в соответствии с внутренними стандартами валидации банка.

Аудитор проверяет расчет показателей, выборки и пороговые значения, но не менее важна реакция на нарушение границ. Должен быть проведен анализ причин пробития контрольных показателей качества моделей, подготовлен план действий, определены ответственные и сроки. Если исправление требует времени, применяется консервативная надбавка. Недостаточно зафиксировать красную зону в отчете: аудитору необходимо оценить, была ли проблема своевременно доведена до органов управления и учтена в текущих оценках риска.

Консервативные надбавки и экспертные корректировки

Консервативная надбавка предназначена для предотвращения недооценки компонентов кредитного риска при наличии факторов возможной недооценки. Положение № 845-П в качестве таких факторов выделяет недостатки данных, неполный учет изменений внешней среды, процессов и технологий банка, а также обобщенную погрешность оценок [1]. На практике надбавка может возникать уже при разработке модели, если данные за отдельные исторические периоды восстановлены неполно, либо после внедрения – при нарушении границ контрольных показателей.

Для аудита важно проверить полноту идентификации факторов недооценки кредитного риска и проанализировать логику количественной оценки величины консервативной надбавки, а также верифицировать расчеты в обоснование отсутствия надбавки (недооценка по отдельным факторам может перекрываться консервативностью модели в целом). Надбавка должна быть достаточно консервативной, но не двойной: один и тот же недостаток не следует одновременно учитывать в нескольких компонентах без обоснования.

Экспертные корректировки рейтинга представляют другой тип ручного вмешательства. Они применяются для моделей корпоративных заемщиков. Корректировки могут учитывать значимую информацию, не отраженную моделью, но одновременно они ослабляют воспроизводимость и создают риск систематического оптимизма. Аудитор анализирует долю и направление корректировок, качество обоснований и последующую дефолтность скорректированных объектов. Одностороннее преобладание улучшений рейтинга требует особенно критической оценки.

Ошибки применения и интерпретации результатов ПВР-моделей

Первая типичная ошибка – необоснованное исключение из области применения. Появление нового продукта не означает автоматически, что действующая модель к нему не применима. Если продукт соответствует установленным критериям сегментации и профилю риска, его исключение должно иметь содержательное основание. Аудитор проверяет полноту продуктового периметра и анализирует технологические причины, по которым кредитные требования могли не попасть в расчет компонентов кредитного риска.

Вторая ошибка – применение модели за пределами области, для которой она разработана и валидирована. Примерами могут быть использование поведенческой модели на этапе первоначальной выдачи, применение корпоративной модели реального сектора к финансовым организациям или попытка оценивать кредитное мошенничество общей PD-моделью. Требования к управлению модельным риском предполагают выявление ошибок интерпретации и использования модельных результатов в непредусмотренных предметных областях [2; 4].

Другой пример ошибки интерпретации – принимать оценку вероятности дефолта по портфелю равной средней вероятности дефолта, оцененной индивидуально по каждому заемщику. Дело в том, что существует корреляция между дефолтами заемщиков, связанных общими условиями (отраслью, зависимостью от одних макроэкономических шоков и тому подобным). Таким образом, простое расширение целевой переменной модели вероятности дефолта индивидуального заемщика на портфель может повлечь недооценку кредитного риска.

Проверка использования модели должна включать интервью с пользователями, анализ внутренних регламентов и распорядительных документов об использовании модели. Аудитор сопоставляет цель, для которой применен рейтинг, с отчетом о разработке модели.

Качество данных и информационные системы

Данные ПВР проходят длинный путь: от фронтальных и учетных систем через витрины и расчетные модули к RWA-калькулятору, регуляторной и внутренней отчетности. Аудит качества данных начинается с анализа критичных атрибутов. К ним относятся факторы моделей, сведения о заемщике и сделке, признаки дефолта, параметры обеспечения и данные, используемые при расчете величины кредитного риска.

Для каждого критичного атрибута должны быть определены контроли полноты, точности, согласованности, своевременности и доступности, а также предельно допустимые значения. Аудитор оценивает не только наличие контроля, но и его покрытие по всей цепочке. Проверка в целевой витрине не обнаружит ошибку, если неверное значение одинаково прошло через все промежуточные системы. Поэтому необходимы сверки с первичным источником и контроль преобразований.

Data lineage должен отвечать на практические вопросы: откуда поступил атрибут, какие правила преобразования применялись, где произошло объединение источников, кто отвечает за качество. Особенно рискованны ручные выгрузки, временные таблицы, архивные информационные системы.

При нарушении порогов качества важна скорость реакции. Пропуск данных может привести к ошибочной оценке либо к подстановке консервативной заглушки. Во втором случае регуляторный расчет кредитного риска может оставаться корректным, однако банк несет дополнительную нагрузку на капитал и теряет точность риск-дифференциации. Аудитор проверяет своевременность выявления проблемы, оценку влияния, информирование владельцев и восстановление нормального потока данных.

Периметр ИТ-аудита включает не только систему расчета рейтинга. В него входят кредитные конвейеры, учетные системы, хранилища и витрины, сервисы агрегации, RWA-калькулятор и системы формирования отчетности. Для них оцениваются управление доступом и изменениями, резервное копирование, непрерывность, логирование и возможность ежедневного расчета нормативов.

Контроль внедрения модели должен обеспечивать соответствие промышленного кода утвержденной версии. Проверяются репозиторий, права на изменение, процедуры релиза, версии библиотек и фиксация параметров. Независимое воспроизведение модели в контролируемой среде позволяет выявить ошибки округления, несоответствие коэффициентов и различия между отчетом о разработке и фактическим кодом.

Практические приемы эффективного аудита

Чтобы определить области наибольшего риска, на которых должен сфокусироваться аудит, следует начать с аналитики. До погружения в модель полезно изучить динамику базы дефолтов, матрицы миграции рейтингов¹⁰ и сегментов, распределение PD и LGD по разрядам рейтинговой шкалы, долю экспертных корректировок, пропусков, величину и динамику консервативных надбавок. Резкое изменение отдельного показателя не является готовым нарушением, но формирует проверяемую гипотезу и помогает направить ресурсы на наиболее рискованный участок.

При проверке не стоит ограничиваться только агрегированной отчетностью. Межформенные сверки подтверждают общую согласованность, однако одинаковые итоговые значения могут скрывать компенсирующие ошибки внутри портфеля. Для существенных сегментов аудитор должен переходить от отчетной формы к RWA-калькулятору, затем к витрине-источнику, реестру дефолтов, а при необходимости – к отдельным кредитным требованиям и первичным проводкам.

Еще один полезный прием – использовать альтернативные источники. Даже утвержденный «золотой источник»¹¹ может стать неполным после изменения бизнес-процесса или архитектуры. Сопоставление с независимой системой помогает выявить выпавшие наблюдения. Альтернативный источник может быть менее полным, но содержащиеся в нем валидные записи, тем не менее, позволяют проверить полноту основной витрины.

Далее обратим внимание на воспроизведение кода модели и ключевых преобразований данных. Полное независимое моделирование не всегда необходимо, но аудитор должен уметь повторить

¹⁰ Матрица миграции рейтингов – таблица, показывающая, какая доля заемщиков за период перешла из одного рейтингового разряда в другой. Резкие изменения структуры миграций служат сигналом для проверки модели и данных (примечание Редакционной коллегии).

¹¹ «Золотой источник» (от англ. golden source) – информационная система или витрина данных, официально признанная в организации эталонным источником определенных сведений (примечание Редакционной коллегии).

расчет на тестовом наборе, проверить граничные случаи и сравнить промежуточные результаты. Для тестируемых аудитором моделей целесообразно применить ряд тест-кейсов, охватывающих пропуски, экстремальные значения, величины входного вектора модели во всем диапазоне возможных значений.

Рекомендуем соединять проверку «сверху вниз» и «снизу вверх». В первом случае аудитор идет от требований и отчета о разработке к модели, данным и контролю. Во втором – от обнаруженной аномалии к ее первопричине. Например, незарегистрированный дефолт может указывать не на единичную операционную ошибку, а на отсутствие необходимого контроля качества данных. Сочетание двух направлений позволяет оценивать рейтинговую систему целостно.

На финальном этапе следует оценивать экономический эффект обнаруженных недостатков. Замечание становится полезным для органов управления, когда показывает влияние на финансовый результат, RWA, капитал банка. Даже если точный пересчет невозможен, следует определить направление и порядок величины (или минимальную границу) эффекта, затронутые сегменты и сценарий его реализации. Это позволяет ранжировать корректирующие мероприятия и отличать формальное несоответствие от существенного риска.

Заключение

Внутренний аудит ПВР не должен сводиться к проверке наличия документов или повторению количественных тестов второй линии. Его задача – независимо оценить целостность рейтинговой системы, проследить происхождение данных, подтвердить корректность реализации модели и установить, учитываются ли модельные оценки в бизнес-решениях и капитале.

Практическая успешность проверки определяется способностью аудитора осмыслить логику, заложенную в модель, понять реальный механизм, скрывающийся за сухим описанием регламентов, и в конечном счете отличить модельную точность от иллюзии точности. Такая иллюзия возникает, когда хороший статистический показатель получен по неверной формуле, корректная модель реализована с ошибкой или рейтинг используется вне области применимости. Риск-ориентированная проверка, основанная на данных и воспроизводимых процедурах, делает аудит не только регуляторной обязанностью, но и инструментом повышения качества управления кредитным риском.

Список источников

1. Положение Банка России от 02.11.2024 № 845-П «О порядке расчета величины кредитного риска банками с применением банковских методик управления кредитным риском и моделей количественной оценки кредитного риска» // Вестник Банка России. 13 февраля 2025 года. [№ 11 \(2542\)](#).
2. Положение Банка России от 08.04.2020 № 716-П «О требованиях к системе управления операционным риском в кредитной организации и банковской группе» // Вестник Банка России. 2 июля 2020 года. [№ 51 \(2187\)](#).
3. [Basel Committee on Banking Supervision. Basel Framework. CRE36: IRB approach – minimum requirements to use IRB approach](#). 8 December 2022 (дата обращения: 19.08.2026).
4. [Board of Governors of the Federal Reserve System. SR 26-2: Revised Guidance on Model Risk Management](#). 17 April 2026 (дата обращения: 19.08.2026).

[Ознакомиться с презентацией к мастер-классу](#)



А.М. ВАЛИУЛЛИН

Первый вице-президент,
руководитель Блока технологий
искусственного интеллекта
Газпромбанка



И.Л. ВАХЛАМОВ

Управляющий директор Центра
технологий искусственного
интеллекта Газпромбанка

Безопасное масштабирование
ИИ возможно только при условии
внедрения адекватной модели
контроля, основанной
на принципе пропорциональности
и интегрированной
в существующие процессы
риск-менеджмента.

БЕЗОПАСНОЕ МАСШТАБИРОВАНИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В БАНКЕ: РИСКИ, МОДЕЛЬ УГРОЗ И КОНТУРЫ КОНТРОЛЯ

Аннотация

В статье рассматриваются практические аспекты внедрения и масштабирования технологий искусственного интеллекта (ИИ) в банковской деятельности. На основе опыта применения моделей машинного обучения и генеративного ИИ в процессах кредитного риска, антифрода, корпоративного анализа и управления данными показано, что ИИ становится полноценным объектом риск-ориентированного управления. Выделены основные группы рисков: риски данных, модельные риски, угрозы информационной безопасности, риски поставщиков и агентных систем. Предложена модель контроля, основанная на принципе пропорциональности, включающая реестр ИИ-систем, риск-классификацию, модель угроз, независимую валидацию, мониторинг и механизмы участия человека в принятии решений. Сделан вывод о необходимости эволюционного развития существующих контуров управления модельным риском для безопасного масштабирования ИИ в финансовых организациях.

Ключевые слова: искусственный интеллект, управление рисками, внутренний аудит, генеративный ИИ, банковский сектор, модельный риск.

Коды JEL: G21, G28, O33, C45.

Введение

Искусственный интеллект перестал быть технологическим экспериментом и стал полноценным объектом управления рисками и контроля в финансовых организациях. Масштаб внедрения, стремительный скачок технологий, пристальное внимание регулятора и активное использование ИИ злоумышленниками требуют пересмотра подходов к управлению рисками, связанными с ИИ. Ключевой принцип безопасного масштабирования заключается в том, что чем выше влияние ИИ на клиента, денежные средства, критические процессы или управленческие решения, тем жестче должны быть требования к контролю. Цель данной работы – на основе практического опыта Газпромбанка систематизировать новые угрозы и предложить действенную модель контроля для обеспечения надежности и управляемости ИИ-систем.

Эволюция ИИ в управлении рисками и новые возможности

Применение ИИ в банковских рисках прошло несколько этапов, каждый из которых добавлял ценность и одновременно порождал новые риски. От классических скоринговых карт на основе логистической регрессии и экспертных правил (1990–2010 годы) банки перешли к методам машинного обучения (градиентный бустинг¹, простые нейронные сети) с сотнями признаков, а затем – к нейросетевым и трансформерным моделям, способным обрабатывать последовательности различных событий. Сегодня на переднем крае находятся генеративный ИИ и автономные агенты, которые открывают беспрецедентные возможности для работы с неструктурированной информацией, автоматизации разработки и мониторинга.

В Газпромбанке ИИ-решения охватывают весь спектр риск-направлений:

1. Кредитные решения: точная оценка вероятности дефолта (PD), дохода и долговой нагрузки, автоматизированное принятие решений по потоку заявок.
2. Антифрод: борьба с подменой личности, фальсификацией дохода, социальной инженерией и транзакционным мошенничеством с использованием многослойной защиты (правила, ML, графовый анализ, биометрия).
3. Корпоративные риски: рейтинговые модели, системы раннего предупреждения, анализ финансовой отчетности с помощью GenAI-ассистентов.
4. Портфельные модели и резервирование: прогноз ожидаемых потерь (ECL), каскады roll-rate-моделей² для соответствия требованиям МСФО 9.

Ключевым фактором успеха всех этих моделей являются качественные данные. Практика показывает, что до 80% инцидентов качества моделей связаны с данными, а не с алгоритмами³. Поэтому критически важны создание единых витрин данных, внедрение Feature Store⁴ для контроля за data lineage и автоматизированные проверки качества источников.

Новые угрозы: за что мы теперь отвечаем

Масштабирование ИИ несет с собой новые классы рисков, которые не покрываются традиционными контрольными процедурами, построенными для детерминированных систем. Можно выделить шесть основных групп рисков ИИ:

1. Риски данных: модель не лучше данных, на которых она обучалась. Сюда входят смещение выборки, утечка целевой переменной, дрейф распределений (data drift) и проблемы с качеством и происхождением данных, особенно персональных.

¹ Градиентный бустинг – метод машинного обучения, объединяющий большое число простых моделей в одну, более точную. Трансформерные модели – архитектура нейронных сетей, лежащая в основе современных больших языковых моделей (примечание Редакционной коллегии).

² Roll-rate-модели (от англ. roll rate – доля перехода) – модели переходов между стадиями просроченной задолженности, используемые для оценки ожидаемых кредитных убытков по МСФО 9 (примечание Редакционной коллегии).

³ Оценка приведена авторами на основании практики Газпромбанка и не является результатом отраслевого исследования (примечание Редакционной коллегии).

⁴ Feature Store (от англ. feature – признак, store – хранилище) – хранилище признаков, то есть централизованная система хранения и версионирования переменных, используемых моделями, обеспечивающая их единообразие и прослеживаемость (примечание Редакционной коллегии).

2. Риски моделей и их деградация: статистически подтверждено, что заметное падение точности (например, Джини⁵) наступает через 4–6 месяцев⁶ после запуска модели. Это свойство системы, а не сбой, поэтому необходимы непрерывный мониторинг и быстрый цикл переобучения.
3. Атаки на ML-системы: в финансовом секторе мотивация атакующих высока, так как успешная атака на скоринг напрямую монетизируется. Основные векторы – adversarial attacks (подбор входных данных для получения нужного результата), data poisoning (отравление обучающей выборки) и инверсия модели (восстановление данных для обучения).
4. Специфические риски GenAI и агентов: недетерминированность ответов, галлюцинации, уязвимости к промпт-инъекциям, а также риски, связанные с автономными действиями агентов (ошибка на одном шаге может усилиться в цепочке, размытая ответственность, компрометация доступов).
5. Риски третьих сторон: большинство ИИ-систем строятся на компонентах open source⁷. Банк не видит их обучающих данных и может не контролировать изменения версий, но при этом несет ответственность за принимаемые решения.
6. Асимметрия угроз: атакующие активно используют ИИ для создания дипфейков⁸, синтетических личностей и автоматизированных атак. Единственный симметричный ответ – это использование ИИ для защиты наравне со скоростью угроз.

Новая модель контроля: управляемое масштабирование

Управление рисками ИИ не требует построения системы с нуля. Задача – эволюционно расширить существующие процессы модельного риска, дополнив их ИИ-специфичными контурами контроля. Предлагаемая модель базируется на пяти управленческих блоках:

1. Реестр и классификация ИИ-систем. Инвентаризация всех ИИ-систем – обязательное условие контроля. Реестр должен фиксировать владельца, бизнес-процесс, используемые данные, уровень риска, статус модели угроз и меры защиты. Критически важно выявлять теневой ИИ (неучтенные скрипты, боты), иначе реестр теряет смысл.
2. Риск-классификация. Все системы классифицируются по четырем уровням риска (низкий, средний, высокий, критический) в зависимости от влияния на клиента, деньги, критичность процесса и принимаемое решение. Например, чат-бот для сотрудников – средний риск, а система скоринга в клиентском канале – высокий или критический.
3. Модель угроз. Для систем высокого и критического уровней риска обязательна разработка модели угроз по единому шаблону, включающему типовые угрозы для ИИ и меры их нейтрализации.
4. Валидация и тестирование. Процесс независимой валидации расширяется. Для классических ML-моделей проверяются качество, стабильность (PSI), объяснимость (SHAP) и обоснованность сложности (сравнение с простым эталоном). Для GenAI-систем требуются

⁵ Коэффициент Джини – показатель ранжирующей способности модели, то есть того, насколько хорошо она отделяет надежных заемщиков от ненадежных. Снижение коэффициента означает ухудшение качества модели (примечание Редакционной коллегии).

⁶ Приведенный срок отражает наблюдения авторов по моделям Газпромбанка. Скорость деградации модели зависит от типа модели, сегмента и изменчивости внешних условий и может существенно отличаться в других организациях (примечание Редакционной коллегии).

⁷ Open source (от англ. открытый исходный код) – программное обеспечение и модели с открытым исходным кодом или открытыми весами, доступные для свободного использования и изменения (примечание Редакционной коллегии).

⁸ Дипфейк (от англ. deepfake) – синтетическое изображение, видео- или аудиозапись, созданные с помощью ИИ и трудноотличимые от подлинных; используются для обхода биометрической идентификации и в социальной инженерии (примечание Редакционной коллегии).

статистические тесты (evals⁹) на сотнях кейсов, оценка фактологичности (галлюцинации) и устойчивости к промпт-инъекциям.

5. Мониторинг, инциденты и Human-in-the-Loop. Вводится непрерывный мониторинг метрик качества, дрефта данных и калибровки. Все значимые запросы и ответы логируются для разбора инцидентов. Ключевым элементом становится принцип «человека в контуре» (Human-in-the-Loop), который должен быть четко определен: от полной автономии до обязательного подтверждения каждого значимого решения.

Реализация этих блоков закрепляется за тремя линиями защиты, где ключевую роль играет наличие ИИ-компетенций во всех линиях, особенно во второй и третьей (риск-менеджмент и внутренний аудит), чтобы независимая проверка не сводилась к формальному доверию разработчикам.

Заключение

Возможности ИИ в управлении рисками реальны и измеримы – они приносят банку существенный экономический эффект, который можно рассчитать и аудировать. Однако вместе с этим на баланс организации ложатся новые, значительно более сложные риски. Безопасное масштабирование ИИ возможно только при условии внедрения адекватной модели контроля, основанной на принципе пропорциональности и интегрированной в существующие процессы риск-менеджмента. Предложенный в статье подход, включающий реестр, классификацию, модель угроз, расширенную валидацию и четкое распределение ролей, позволяет не запрещать, а делать ИИ проверяемым и управляемым активом, обеспечивая устойчивость банка и доверие со стороны клиентов и регулятора.

Список источников

1. [Основные направления развития финансового рынка Российской Федерации на 2024 год и период 2025 и 2026 годов](#). Банк России, 2023 год (дата обращения: 30.09.2026).
2. Методические рекомендации Банка России [от 16.06.2026 № 3-МР](#) «По обеспечению информационной безопасности при разработке и применении искусственного интеллекта на финансовом рынке» (дата обращения: 30.09.2026).
3. Положение Банка России [от 17.08.2023 № 821-П](#) «О требованиях к обеспечению защиты информации при осуществлении переводов денежных средств и о порядке осуществления Банком России контроля за соблюдением требований к обеспечению защиты информации при осуществлении переводов денежных средств» // Вестник Банка России. 21 декабря 2023 года. № 76 (2472) (дата обращения: 30.09.2026).

[Ознакомиться с презентацией к мастер-классу](#)

[Ознакомиться с презентацией к выступлению на заседании Экспертного совета](#)

⁹ Evals (от англ. evaluations) – наборы тестовых заданий с эталонными ответами для систематической оценки качества генеративной модели (примечание Редакционной коллегии).



ОТ ПОЛИТИКИ ЗАПРЕТОВ К УПРАВЛЯЕМОМУ ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ: УГРОЗЫ БЕЗОПАСНОСТИ ИИ-СИСТЕМ И МОДЕЛЬ ИХ КОНТРОЛЯ В ФИНАНСОВЫХ ОРГАНИЗАЦИЯХ

”

Разрыв между человеческими возможностями и машинным потенциалом закрывается не тем, что человек становится «митохондрией» в системе ИИ, а тем, что он остается внешней системой: ставит цели, проверяет результаты и несет ответственность. Можно назвать это состояние «сцеплен, но не слит».

Р.М. ГУСЕЙНОВ

Председатель ПК «РАД КОП»

Аннотация

Рассматриваются угрозы информационной безопасности, возникающие при внедрении систем искусственного интеллекта (ИИ) в финансовых организациях, и подходы к управлению ими. Показано, что с точки зрения угроз ИИ не является уникальной технологией: к нему применимы концепции жизненного цикла, недеklarированных возможностей, поверхности атаки и социальной инженерии. Систематизированы классы угроз по перечням OWASP Top 10 for LLM Applications (2026) и OWASP Top 10 for Agentic Applications (2026); выделен качественный сдвиг риска при переходе от языковых моделей, формирующих текст, к агентам, совершающим действия с полномочиями. На данных отраслевой телеметрии¹ обоснован тезис о неэффективности тотальных запретов, которые порождают теневой ИИ и лишают организацию наблюдаемости. Проведено сопоставление Методических рекомендаций Банка России от 16.06.2026 № 3-МР с практикой 2026 года и выделены восемь зон усиления. Предложены зонирование сценариев по уровню риска, минимальный набор контролей, контур внутреннего аудита, основанный на воспроизводимых доказательствах, и 90-дневный план перехода к управляемому ИИ.

Ключевые слова: искусственный интеллект, информационная безопасность, большие языковые модели, ИИ-агенты, теневой ИИ, угрозы ИИ, OWASP, LLM, Методические рекомендации Банка России № 3-МР; внутренний аудит, операционный риск.

Коды JEL: G21, G28, M15, O33, K24.

Технология, которую предсказали

Сегодня в информационных технологиях только и разговоров, что об искусственном интеллекте и цифровой трансформации. Аналитики McKinsey Global Institute описывают ближайшее будущее как совместную работу людей, программных агентов и роботов и оценивают потенциал автоматизации по секторам и профессиям [22]. Финансовые организации, как и все остальные, оказались перед выбором, который выглядит бинарным: либо внедрять

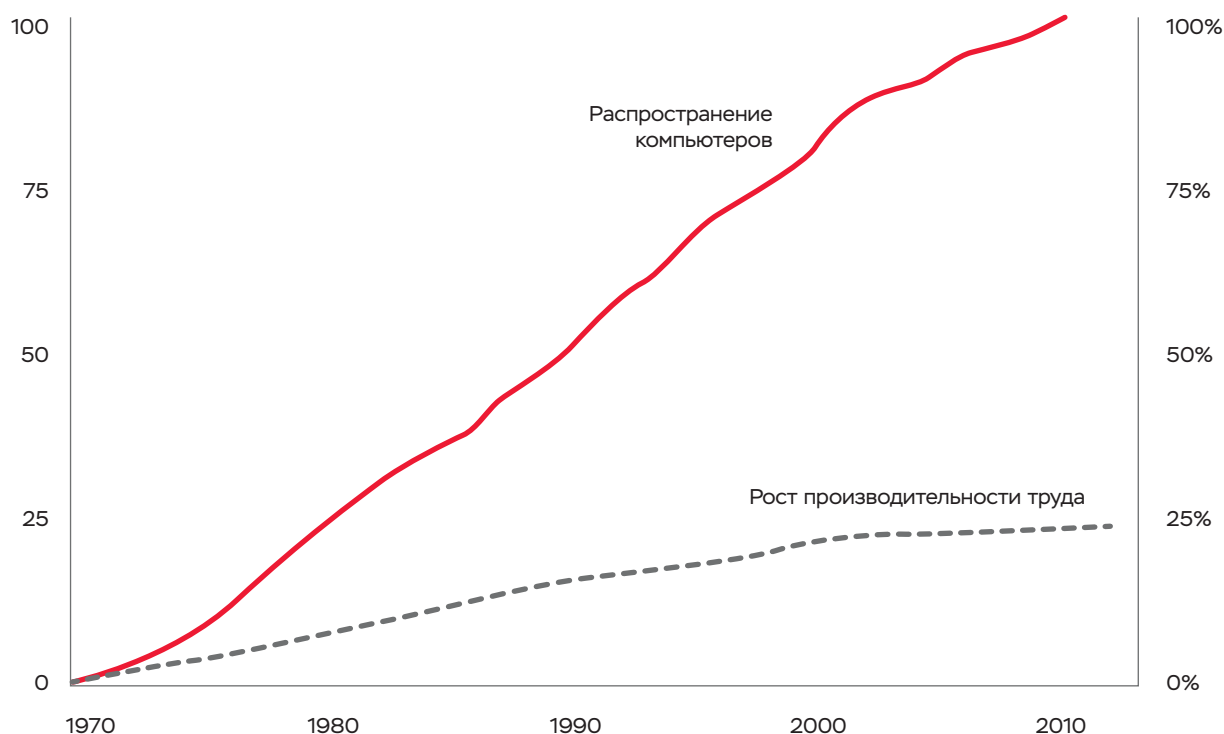
¹ Телеметрия – автоматически собираемые данные о работе систем и действиях пользователей (сетевой трафик, события на рабочих станциях и в приложениях), используемые для мониторинга и обнаружения инцидентов (примечание Редакционной коллегии).

РОСТ ПРОИЗВОДИТЕЛЬНОСТИ И РАСПРОСТРАНЕНИЕ КОМПЬЮТЕРОВ

Рис. 1

Индекс распространения компьютеров

Рост производительности труда (США, % к 1970)



генеративный ИИ повсеместно, либо запрещать его, опасаясь утечек и регуляторных последствий. В настоящей статье показано, что этот выбор ложный и на практике разумно реализовать третий путь – управляемый ИИ, при котором организация не полностью тормозит внедрение технологии, но ограничивает радиус возможного поражения.

Прежде чем говорить об угрозах, полезно вспомнить, что ничто не ново под луной. Нынешнее состояние ИИ было в общих чертах предсказано профильными специалистами еще во второй половине XX века: Станислав Лем в «Сумме технологии» обсуждал «выращивание информации» и усилители интеллекта как инженерные, а не фантастические задачи [6]. Хорошо известно и то, как ведет себя любая новая технология в координатах цикла зрелости Gartner²: за запуском следуют «хайп» и пик завышенных ожиданий (по сути, стимул для инвесторов вкладывать ресурсы в ожидании полезной отдачи), затем спад интереса и разочарование, и лишь затем – выход на плато продуктивности [18]. Получается своеобразная диалектика: все, что возникает, проходит через противоречие и на практике рождает истинную ценность.

Есть и более прозаический аргумент – парадокс Солоу (рис. 1). В 1987 году Роберт Солоу заметил, что компьютерную эпоху видно повсюду, кроме статистики производительности [30]: массовое внедрение вычислительной техники в бизнес долгие годы не ускорило рост производительности труда. То есть в ИТ-технологиях не наблюдается такой отдачи с точки зрения продуктивности, как, например, при внедрении двигателей внутреннего сгорания и замены ручного крестьянского труда в полях на работу трактора. Вопрос о том, повторится ли парадокс в эпоху ИИ, обсуждается и сегодня [19]. Причины известны и не зависят от конкретной технологии. Внедрение не тождественно полезному эффекту. Реальные выгоды появляются только при перестройке процессов. Полезный эффект, если он существует, раскрывается с длительным лагом. Более того,

² Цикл зрелости технологий (от англ. hype cycle) – модель консалтинговой компании Gartner, описывающая типовую траекторию восприятия новой технологии: запуск, пик завышенных ожиданий, разочарование, постепенное осознание пользы и выход на плато продуктивности (примечание Редакционной коллегии).

в некоторых случаях полезная отдача может быть нивелирована затратами на сопровождение. Наконец, традиционные метрики измеряют не все. Из этого следует неудобный, хотя и банальный вывод: модель «люди – процессы – технологии» по-прежнему актуальна. Без описаний и документации нет процессов; без процессов маловероятна адекватная автоматизация; без адекватной автоматизации ИИ может оказаться усилителем хаоса и источником дополнительных потерь. Автоматизировать беспорядок – весьма своеобразная работа.

Отсюда рождаются три тезиса, задающие рамку дальнейшего изложения. Во-первых, текущее состояние ИИ ожидаемо, и относиться к нему следует как к очередной технологии, проходящей известный цикл. Во-вторых, ключевая проблема этой технологии скорее гуманитарная, нежели техническая: это разрыв между возможностями человека – ставить задачи, проверять результат, нести ответственность – и потенциалом и скоростями машины. В-третьих, благодаря парадоксу Солоу и стоящим за ним методологическим и организационным ограничениям у людей остается реальная возможность – время на то, чтобы выстроить управление технологией, пока она не заняла устойчивое место в их жизни. Кому это выгодно (*cui bono*³) – вопрос, который стоит задавать каждому, кто продает как полное замещение, так и полный запрет: конфликт интересов встроен в саму рыночную ситуацию, а тревога о том, отнимет ли робот мою работу, давно стала самостоятельным жанром [32].

Усилитель или среда: две метафоры отношений человека и ИИ

Зигмунд Фрейд в «Недовольстве культурой» назвал человека «богом на протезах»: орудия и техника продолжают его органы, делая его могущественным, но не вполне счастливым [11]. ИИ – очередной протез, и ключевой вопрос в том, кто кем управляет.

Здесь полезна биологическая метафора (рис. 2). В клетках организмов есть органеллы – митохондрии. Они производят энергию, необходимую для всех процессов. Когда-то митохондрии были самостоятельными существами, но на заре эволюции встроились в клеточные структуры, превратившись в их части. Таким образом, митохондрии не управляют клеткой: цели и структуру определяет целое, в которое они встроены. Взаимодействие человека и ИИ поднимает тот же вопрос. Человек ли определяет цели, ценности и направление? ИИ только обрабатывает данные, находит закономерности, ускоряет выполнение задач и предлагает варианты или задает ориентиры и траекторию? В идеальном мире для людей, желающих сохранить субъектность и автономию, ИИ не управляет человеком, но дополняет его в тех задачах, которые решаются медленнее или хуже без такового протеза.

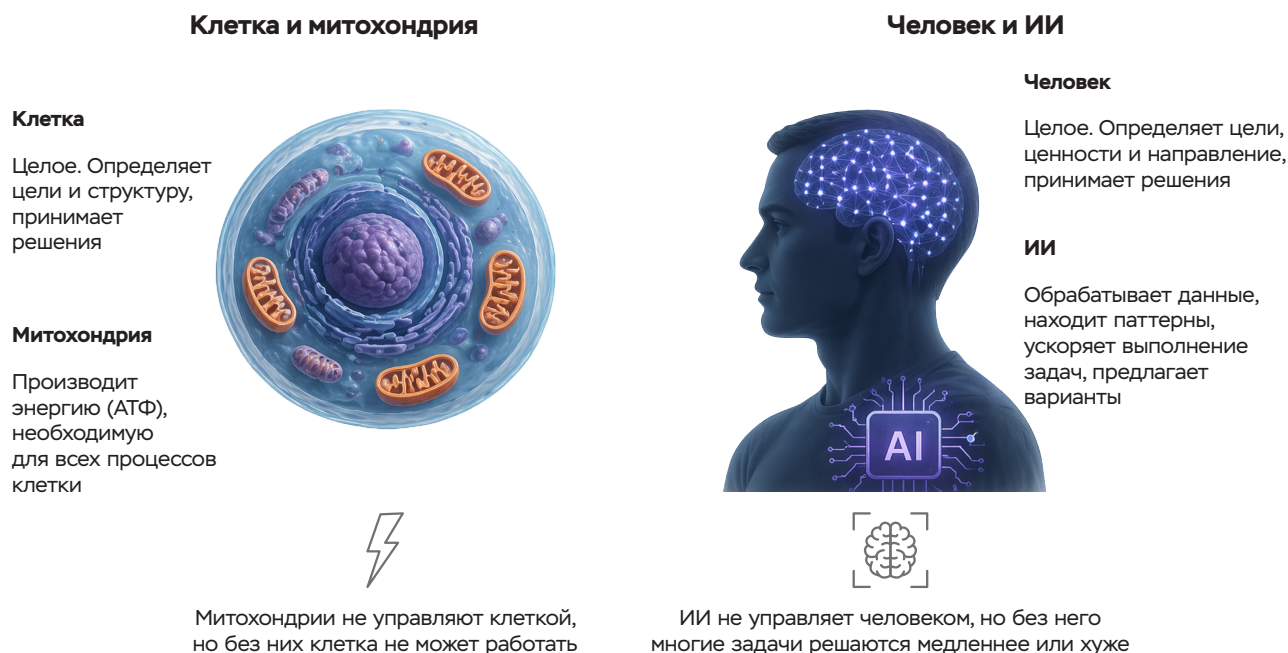
Но метафору легко инвертировать. Если ИИ – «клетка», ядро и сверхинтеллект системы, то человек становится «митохондрией» внутри нее: выполняет полезные для системы функции – креатив, суждения, адаптацию, социальные роли, – но не принимает ключевых решений и полностью зависит от системы в ресурсах, задачах, оценке и вознаграждении, оказываясь заложником контекста, формируемого «волей ИИ». Свобода здесь – иллюзия, а поток зависимости замкнут: ИИ задает условия и рамки → человек действует в рамках и получает вознаграждение (что-то похожее уже сейчас можно наблюдать среди пользователей компьютерных игр или социальных сетей, которые удачно эксплуатируют нейрофизиологию человека).

Какой из двух миров реализуется, определяется не мощностью моделей, а архитектурой отношений: кто ставит цели, кто проверяет результат и кто несет ответственность. А архитектура отношений Человека и ИИ – это вопросы культуры и политики.

³ *Cui bono* (лат.) – «кому выгодно» (примечание Редакционной коллегии).

ДВА СПОСОБА ОТНОШЕНИЙ ЧЕЛОВЕКА И ИИ

Рис. 2



Здесь уместно вспомнить Герберта Саймона. Его работы о границах рациональности и «науках об искусственном» показывают, что любая вычислительная система ограничена в постановке целей и проверке собственных результатов; человек в этой конструкции играет для ИИ роль внешней системы, формулирующей задачу и верифицирующей ответ [10]. Тезис имеет прямое отношение к безопасности: именно точки постановки задачи (ввод) и проверки результата (вывод) оказываются главными объектами атаки на ИИ или его злонамеренного использования.

Ту же мысль по-своему выражает Дэниел Мисслер, автор концепции личной ИИ-инфраструктуры (Personal AI Infrastructure, PAI) [23; 28]. В его «стеке возможностей» знания, понимание, интеллект и творчество в той или иной мере доступны и человеку, и машине, тогда как субъективный опыт и встроенные эволюционные желания остаются человеческими; собственные цели современных ИИ-систем задаются извне. Организационное продолжение этой идеи для организаций – «решетка» (Lattice): модель управления организацией, в которой решения и власть распределены по уровням руководства, подразделений, команд и людей либо агентов, а контекст – цели, регламенты, метрики, работа – един. ИИ в этой конструкции выполняет роль, похожую на «динамическую шину данных», он в реальном времени собирает информацию об организации и доводит необходимые предложения и решения до соответствующих субъектов. Следует подчеркнуть, что это авторские модели, а не научный стандарт: термины «понимание» и «творчество» применительно к ИИ остаются дискуссионными, а Lattice – это не отраслевой стандарт и не официальный компонент PAI. Тем не менее наработки автора точно фиксируют практическую развилку: не использовать ИИ сегодня – значит остаться в каменном веке, пока мир переходит к бронзе, с той оговоркой про парадокс Солоу, что переход к бронзе требует перестройки процессов, а не покупки бронзового топора, который будет лежать на полке или которым будут заколачиваться гвозди.

ИИ как объект защиты: базовые вопросы и стадии зрелости внедрения

Прежде чем разговаривать об угрозах и строить защиту ИИ, зададимся базовыми универсальными вопросами. Как выглядит ИИ с точки зрения уровней ИТ-инфраструктуры,

ВОЗМОЖНОЕ ПРЕДСТАВЛЕНИЕ ПОВЕРХНОСТИ АТАКИ

Рис. 3



информации и интерфейсов/агентов (какой стек технологий его реализует)? Кто и в каких ролях с ним взаимодействует? Что подается на вход и что получается на выходе (как устроены процессы и потоки данных)? Зачем ИИ используется организацией? Какие существуют сценарии развертывания? Какая есть регуляторика и стандартизация? Возможное представление – трехуровневая пирамида: инфраструктура (вычислительные ресурсы, среда исполнения), информация (данные, корпус знаний, векторные хранилища) и интерфейсы либо агенты (то, через что система принимает вход и отдает выход). У каждого уровня такой пирамиды (рис. 3) есть вход и выход, и все они – часть поверхности атаки.

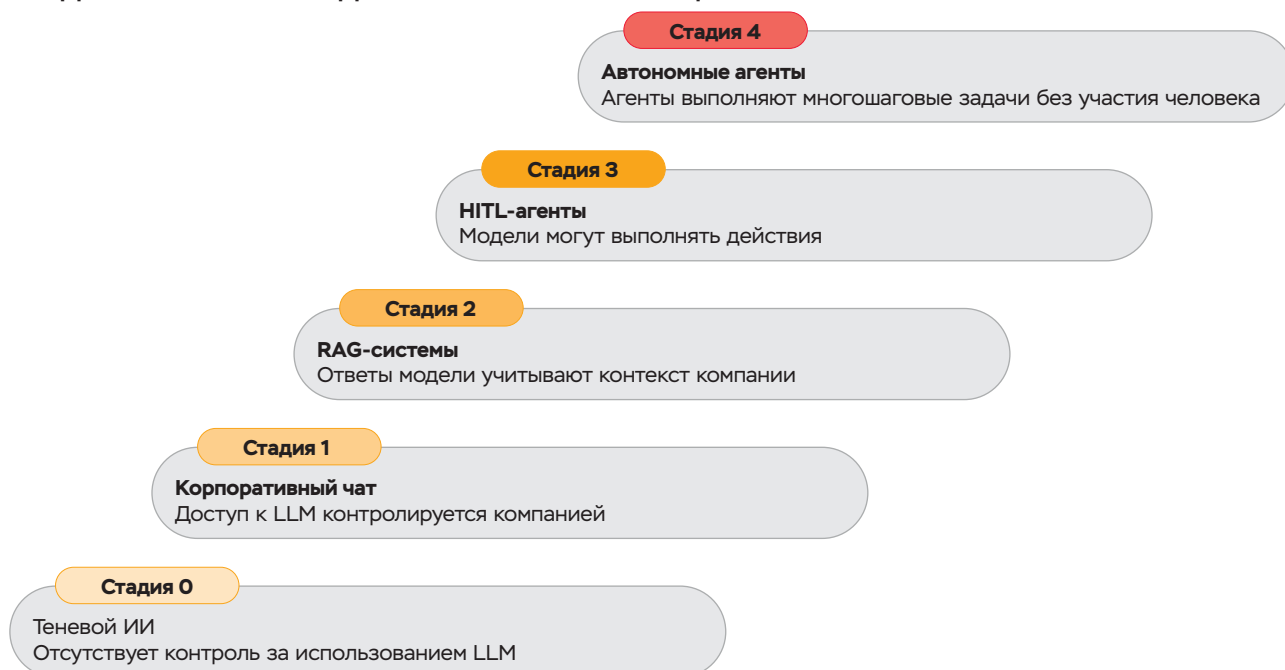
Методические рекомендации Банка России от 16.06.2026 № 3-МР описывают жизненный цикл ИИ-системы в четырех этапах: подготовка данных, разработка, обучение и тестирование, функционирование [8]. Наборы обучающих и тестовых данных, документация и исходный код, модель, входные запросы и выходные решения – каждый из этих элементов является активом, который может быть скомпрометирован, а модель угроз рекомендуется строить по методике ФСТЭК России⁴ [7]. Иными словами, регулятор не изобретает нового методического аппарата, а распространяет привычный на новый класс систем.

Полезно различать и стадии зрелости корпоративного применения ИИ (рис. 4). Стадия 0 – теневой ИИ: контроль за использованием языковых моделей отсутствует. Стадия 1 – корпоративный чат: доступ к модели контролируется компанией. Стадия 2 – RAG-системы (retrieval-augmented generation): ответы модели учитывают контекст компании, извлекаемый из ее базы знаний. Стадия 3 – агенты с «человеком в контуре» (Human-in-the-Loop, HITL): модели могут выполнять действия при подтверждении. Стадия 4 – автономные агенты, выполняющие многошаговые задачи без участия человека. Каждая следующая стадия не отменяет угроз предыдущей, а добавляет к ним новые, делая более актуальным новый объект фокуса: на стадии 2 объектом атаки становится корпус знаний, на стадиях 3–4 – полномочия и инструменты агента.

⁴ Методика оценки угроз безопасности информации, утвержденная ФСТЭК России 05.02.2021. Методические рекомендации № 3-МР отсылают к ней при построении модели угроз для ИИ-систем (примечание Редакционной коллегии).

СТАДИИ ЗРЕЛОСТИ ВНЕДРЕНИЯ ИИ В ОРГАНИЗАЦИИ

Рис. 4



Перечень рисков, с которыми организация сталкивается на этих стадиях, уже вполне стандартен: манипулирование входными данными (изменение поведения модели подбором специальных входных параметров, приводящее к непредвиденному поведению или сбою); чрезмерные полномочия (доступ агента к данным и инфраструктуре за пределами задач и прав пользователя); утечка корпоративных данных во внешние сервисы; некорректная обработка вывода (недостаточный контроль результата, формируемого моделью); неограниченное потребление ресурсов с риском отказа в обслуживании; кража модели; инверсия модели (восстановление конфиденциальных или персональных данных по выходам модели); наконец, недеklarированные возможности и отсутствие верификации компонентов.

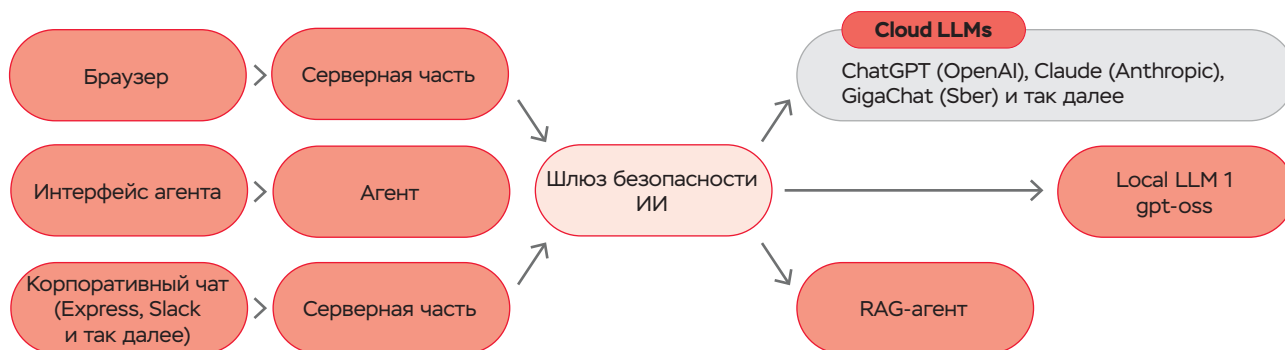
Последний пункт хорошо подтверждается эмпирикой. По данным отчета Orca Security, 62% организаций использовали пакеты для ИИ с известными уязвимостями [31]; исследователи JFrog выявили порядка 100 моделей с закладками в самом популярном репозитории ИИ-моделей [29]. Показателен инцидент марта 2026 года с LiteLLM – популярным шлюзом к языковым моделям: две версии библиотеки были опубликованы в PyPI⁵ с вредоносным кодом, похищавшим SSH-ключи, облачные и API-учетные данные; исходный код репозитория оставался чистым – вредоносный код был внедрен в артефакт при публикации, а обнаружен по счастливой случайности, из-за ошибки в самом вредоносном коде [2; 12]. Атака на цепочку поставок в ИИ ничем принципиально не отличается от атак на цепочку поставок вообще – за исключением того, что число зависимостей и скорость их обновления выше.

Таким образом, отметим еще раз, что ИИ не является уникальной технологией с точки зрения угроз: к нему применимы все стандартные концепции – жизненный цикл, недеklarированные возможности, поверхность атаки, социальная инженерия. При этом накапливается все больше «историй успеха» как атакующих, так и защитников, и можно с уверенностью утверждать, что благодаря ИИ скорость и объемы работы возрастают с обеих сторон. Вероятно, в течение 5–10 лет технология выйдет на пик возможностей, займет устойчивое место в бизнес-процессах и встроится в инструменты специалистов различных специализаций – **в том числе специалистов в области аудита** [20].

⁵ PyPI (от англ. Python Package Index) – основной репозиторий программных пакетов для языка Python. Стилер (от англ. stealer) – вредоносная программа, предназначенная для кражи паролей, ключей и иных учетных данных (примечание Редакционной коллегии).

ИЛЛЮСТРАЦИЯ МЕРЫ ЗАЩИТЫ ТИПА «ШЛЮЗ»

Рис. 5



Примечания.

Шлюз безопасности ИИ – это защитный слой (или прокси), который стоит между серверной частью и ИИ-сервисом, перехватывающий и проверяющий весь входящий и исходящий трафик.

Серверная часть (backend) – это «невидимая» часть приложения на стороне обработки вычислений и источников данных. RAG-агент (Retrieval-Augmented Generation Agent) – это «умная» надстройка над LLM, которая позволяет модели отвечать на вопросы, используя внешние данные, а не только то, чему она научилась при обучении.

Local LLM (gpt-oss) – это большая языковая модель (в данном случае – gpt-oss от OpenAI), которая запускается локально на собственном оборудовании, а не в облаке.

Практическим воплощением этого вывода служит типовая архитектура защищенного корпоративного чата (рис. 5): пользователь работает через браузер или интерфейс агента в корпоративном мессенджере; запросы проходят через серверную часть и платформу безопасности ИИ, которая маршрутизирует их к облачным моделям, к локальной модели с открытыми весами либо к RAG-агенту. Ни один из элементов этой схемы не нов: это шлюз, политики, журналирование и разделение маршрутов – знакомые меры защиты, примененные к новому объекту.

Карта угроз: модель – лишь часть системы

Наиболее употребительной систематизацией угроз для приложений на базе языковых моделей остается перечень OWASP. Редакция OWASP Top 10 for LLM Applications 2026, опубликованная в августе 2026 года, впервые опирается не только на консенсус экспертов, но и на данные о реальных инцидентах: голосование сохранило 75% веса, а 25% отдано анализу 6 639 инцидентов из публичных баз уязвимостей и базы вреда от ИИ [25; 27]. Показательно, что риск «избыточная агентность» (Excessive Agency) заметно поднялся в рейтинге, «раскрытие системного промпта» расширено до «раскрытия скрытого контекста», а «ложные выводы» (Misinformation) выросли в приоритете после того, как инциденты показали реальный ущерб от уверенно сгенерированных, но неверных ответов, запускающих автоматизированные процессы. Сами авторы перечня четко очерчивают его границы: он описывает модель как компонент внутри приложения; как только модель становится действующим лицом – с инструментами, памятью и последствиями, – риск переходит в OWASP Top 10 for Agentic Applications [26]. Оба перечня приведены в таблице 1; для тактик и техник атак на ИИ-системы полезна постоянно пополняемая база MITRE ATLAS⁶ [24].

⁶ MITRE ATLAS – база знаний о тактиках и техниках атак на системы ИИ, которую ведет американская некоммерческая организация MITRE по аналогии с базой MITRE ATT&CK, описывающей традиционные кибератаки (примечание Редакционной коллегии).

ПЕРЕЧНИ OWASP 2026: УГРОЗЫ LLM И АГЕНТОВ

Табл. 1

№ П/П	OWASP TOP 10 FOR LLM APPLICATIONS 2026 (LLM01 – LLM10)	OWASP TOP 10 FOR AGENTIC APPLICATIONS 2026 (ASI01 – ASI10)
1.	Prompt Injection – внедрение инструкций	Agent Goal Hijack – перехват цели агента
2.	Sensitive Information Disclosure – раскрытие чувствительных данных	Tool Misuse and Exploitation – неправомерное использование инструментов
3.	Excessive Agency – избыточная агентность	Identity and Privilege Abuse – злоупотребление идентичностью и правами
4.	Supply Chain – цепочка поставок	Agentic Supply Chain Vulnerabilities – цепочка поставок агента
5.	Data and Model Poisoning – отравление данных и модели	Unexpected Code Execution – неожиданное выполнение кода
6.	Unbounded Consumption – неограниченное потребление ресурсов	Memory and Context Poisoning – отравление памяти и контекста
7.	Misinformation – ложные выводы (галлюцинации)	Insecure Inter-Agent Communication – небезопасное взаимодействие агентов
8.	Hidden Context Exposure – раскрытие скрытого контекста	Cascading Failures – каскадные сбои
9.	Vector and Embedding Weaknesses – слабости RAG, векторов и эмбедингов	Human-Agent Trust Exploitation – эксплуатация доверия человека к агенту
10.	Improper Output Handling – небезопасная обработка вывода	Rogue Agents – самовольные агенты

Источник: составлено автором по открытым источникам [26; 27].

Внедрение инструкций: ограничить ущерб, а не только вероятность

Внедрение инструкций (prompt injection) сохраняет первое место с 2023 года не потому, что публичных эксплойтов много, а потому, что уязвима сама архитектура: языковая модель не отделяет инструкции от данных. Все, что попадает в общее окно контекста⁷ (системный промпт, пользовательский ввод, найденные фрагменты документов, история диалога, записи памяти, вывод инструментов), обрабатывается одинаково. Цепочка атаки выглядит так: вредоносный контент (письмо, документ, веб-страница, фрагмент из RAG, вывод инструмента) → общее окно контекста → решение модели (ответ, вызов инструмента, запись в память) → действие с правами (API, командная оболочка, почта, база данных, МСР-сервер⁸) → ущерб (утечка, изменение данных, удаленное выполнение кода, закрепление в системе).

Архитектурная установка OWASP звучит следующим образом: проектируйте систему так, будто граница между инструкциями и данными однажды будет обойдена [26]. Критично не только снизить вероятность инъекции, но и не дать скомпрометированной модели превратить ее в успешную атаку. Отсюда набор контролей: метки происхождения и доверия (внешние данные отделяются от инструкций, невидимые символы очищаются); строгая схема данных и доверенный код, валидирующий структуру до следующего действия; модуль политик, повторно проверяющий намерение и аргументы вне модели; минимальные права и подтверждение человеком критичных операций; список разрешенных исходящих каналов, трассировка и аварийная остановка.

⁷ Окно контекста – весь объем информации, который языковая модель обрабатывает при формировании ответа. Системный промпт – скрытые от пользователя инструкции, задающие правила поведения модели (примечание Редакционной коллегии).

⁸ МСР-сервер – компонент, предоставляющий ИИ-агенту доступ к внешним данным и инструментам по протоколу Model Context Protocol. Через него агент получает возможность совершать действия в других системах (примечание Редакционной коллегии).

RAG и память: база знаний – тоже вход

Типовой конвейер RAG: «источники (вики⁹, почта, тикеты, файлы, веб) → загрузка (парсер, распознавание, разбиение на фрагменты, метки, списки доступа) → векторное хранилище¹⁰ → поиск (запрос, ранжирование, фильтр доступа) → контекст (найденные фрагменты + системный промпт) → модель или агент (ответ, инструмент, запись в память)» – содержит по меньшей мере шесть точек атаки: отравленный документ с инструкцией, скрытой в источнике; парсер и фрагментация, при которых контекст теряет происхождение; отравленные векторы, манипулирующие релевантностью; обход списков доступа с извлечением чужих данных; смешение контекста, когда данные принимаются за инструкции; долгая память, в которую инъекция записывается надолго и заражает последующие сеансы. Вредоносный документ или запись памяти может пережить сессию.

Контроли, которые должны быть явными: происхождение и версия каждого фрагмента; проверка прав доступа до поиска, а не после генерации; подписанный и проверяемый корпус; разделение инструкций, данных и вывода инструментов; запись в память как привилегированная операция; трассировка «источник → фрагмент → ответ → действие».

Цепочка поставок ИИ длиннее обычного SBOM

Модель – лишь один артефакт. Компрометировать можно вычислительную среду (графические ускорители, среду исполнения, контейнеры, библиотеки), фреймворк и обвязку (оркестратор, шлюз, парсеры, SDK), модель и адаптеры (веса, токенизатор, LoRA, квантование¹¹), данные (обучение, дообучение, оценка, RAG-корпус), системный контекст (промпты, политики, навыки, примеры), инструменты (MCP-серверы, плагины, схемы API) и эксплуатацию (секреты, учетные записи, телеметрия, конфигурация). Инцидент с LiteLLM – как раз атака на уровень обвязки.

Методические рекомендации № 3-МР охватывают вопросы доверия к поставщику, целостность и открытый код, отсылая к ГОСТ Р 59276-2020 о способах обеспечения доверия к системам ИИ [4; 8]. Международная практика 2026 года требует более явно защитить следующее: машиночитаемый состав компонентов ИИ (AI-BOM/SBOM) – модели, данные, агенты, инструменты, защитные правила; подписи и проверка происхождения при выпуске; закрепление версий моделей, MCP-серверов, плагинов и описаний инструментов; безопасные форматы и песочница для небезопасной сериализации и выполнения кода; проверка уязвимостей и лицензий; договорные условия об инцидентах, аудите, сроке хранения, обучении на данных, субподрядчиках и плане выхода; отзыв артефакта и безопасный откат версии. Инструментарий для этого развивается в рамках инициатив OWASP по безопасности агентов и AIBOM [13].

⁹ Вики (wiki) – информационная система или база знаний, содержимое которой пользователи могут совместно создавать и редактировать. В данном контексте имеются в виду, в частности, корпоративные базы знаний, используемые как источники данных для RAG-систем (примечание Редакционной коллегии).

¹⁰ Эмбединг (от англ. embedding) – числовое представление текста, отражающее его смысл. Векторное хранилище – база данных таких представлений, по которой RAG-система находит фрагменты, относящиеся к запросу пользователя (примечание Редакционной коллегии).

¹¹ Технические компоненты модели: LoRA (от англ. Low-Rank Adaptation) – метод дообучения модели с небольшим числом дополнительных параметров; квантование – сжатие модели за счет снижения точности представления чисел; токенизатор – модуль разбиения текста на единицы (токены), с которыми работает модель (примечание Редакционной коллегии).

От языковой модели к агенту: смена класса последствий

Обычная языковая модель отвечает, агент действует. Это различие меняет не только поверхность атаки, но и класс последствий. Ошибка модели обычно остается информационной: неверный текст, код или рекомендация; ущерб возникает, когда человек или следующая система доверяет выводу. Агент же читает внешние данные, выбирает инструменты, меняет состояние систем и может продолжать цикл без нового запроса. Ошибка агента – это уже совершенное действие: письмо отправлено, запись изменена, команда выполнена, данные выведены наружу.

Риск агента можно представить как произведение: вероятность ошибки или инъекции × полномочия × автономия × связность. Первый сомножитель не удастся свести к нулю, поэтому управляемыми остаются три других. Именно на этом построен принцип минимальной агентности (least agency): агенту предоставляются только нужные функции, права и автономия. Отчет IBM за 2026 год косвенно подтверждает цену вопроса: средняя стоимость инцидента с инверсией модели составила 6,07 млн долл. США, с внедрением инструкций – 5,89 млн долл. США, а 92% организаций, столкнувшихся с ИИ-инцидентами, не имели надлежащего контроля доступа к ИИ-системам [17].

Перечень OWASP для агентских приложений (табл. 1, правый столбец) описывает собственный класс угроз: перехват цели агента через контент, который он читает; неправомерное использование инструментов; злоупотребление идентичностью и унаследованными правами; цепочка поставок агента (плагины, реестры, MCP-серверы); неожиданное выполнение кода; отравление памяти и контекста; небезопасное взаимодействие агентов; каскадные сбои; эксплуатация доверия человека к агенту; наконец, «самовольные» агенты [26].

Контур управления агентом складывается из 10 контрольных точек вокруг песочницы¹², в которой находятся план, модель и состояние: 1) реестр возможностей – какие инструменты доступны; 2) машинная учетная запись (non-human identity, NHI) с ограниченным сроком жизни; 3) ограниченные права – на инструмент, на задачу, на короткий срок; 4) модуль политик, проверяющий намерение, аргументы и адресата; 5) подтверждение человеком точного действия, а не его пересказа; 6) список разрешенных исходящих каналов; 7) запись в память только как привилегия; 8) трассировка и воспроизведение цепочки «промт → контекст → инструмент → действие»; 9) лимиты стоимости, частоты и глубины циклов; 10) остановка и откат. Формула проста: модель предлагает, а доверенный код решает, разрешено ли выполнить.

Запрет – не стратегия: теневой ИИ

Естественная реакция службы безопасности на новую и плохо понятную технологию, с такими последствиями применения – запрет. Рассмотрим, что происходит при тотальном запрете. У сотрудников остается нереализованная потребность в ускорении работы и делегировании рутинных операций машине. В результате они самовольно обходят блокировку с личного телефона или через VPN. И начинают тайно выполнять работу в публичной модели, а служба информационной безопасности получает слепую зону. В итоге ИИ продолжает использоваться, но исчезает наблюдаемость. Таким образом, запрет без создания удобной альтернативы создает стимул его нарушить. Оптимальным подходом для большинства ситуаций является управляемый доступ по цепочке: «задача → корпоративный портал с единым входом → ИИ-шлюз с политиками, DLP и журналированием → модели по разрешенному маршруту». Здесь ситуация похожа

¹² Песочница (от англ. sandbox) – изолированная среда исполнения, в которой действия программы или ИИ-агента не затрагивают рабочие системы организации (примечание Редакционной коллегии).

на классическое соотношение цены нарушения и цены соблюдения правил. Если легальный путь оказывается быстрее и безопаснее обхода, пользователи радостно следуют правилам, потому что они удобны и безопасны. Отдельно отметим, что правильный объект запрета – это не ИИ вообще, а комбинация информации определенного уровня ограничения доступа, типа модели, доступных полномочий и сценария работы (например, государственную тайну ни при каких обстоятельствах нельзя передавать в публичные языковые модели, а обезличенные персональные данные, если политики организации позволяют это делать, вполне можно).

Масштаб явления, получившего название теневого ИИ (Shadow AI¹³) – негласного использования сотрудниками внешних ИИ-инструментов вне контроля службы безопасности, – подтверждается данными. По телеметрии ГК «Солар»¹⁴, проанализировавшей трафик 150 российских организаций из ИТ, финансового сектора, промышленности, госсектора, телекоммуникаций и ретейла, объем данных, отправляемых сотрудниками в общедоступные нейросети, за 2025 год вырос в 30 раз, при этом около 60% организаций не имели формализованной политики [5]. По данным IBM и Ponemon Institute¹⁵, изучивших 602 организации из 16 стран, пережившие утечки с марта 2025 по февраль 2026 года, теновой ИИ фигурировал в 43% исследованных инцидентов против 20% годом ранее, более двух третей организаций не имели процессов управления, ограничивающих его, а средняя стоимость утечки достигла рекордных 4,99 млн долл. США [17]. Необходима оговорка: ГК «Солар» и IBM используют разные выборки, периоды и знаменатели – первая цифра получена по трафику российских компаний, вторая – только по организациям, пережившим утечки. Эти цифры нельзя складывать; они показывают одно направление: использование ИИ растёт быстрее контроля.

Цель управления, следовательно, – не ноль теневого ИИ, а минимальный неуправляемый поток и измеримый остаточный риск. Последовательность действий может быть представлена следующим образом: увидеть, классифицировать типы информации, предложить альтернативу, ограничить использование, внедрить мониторинг. Ни одна мера при этом не самодостаточна (табл. 2).

МЕРЫ ПРОТИВ ТЕНЕВОГО ИИ

Табл. 2

РАБОТАЕТ В СВЯЗКЕ	ОБХОДИТСЯ ИЛИ ЛОМАЕТ ПРОЦЕСС
Управляемые аккаунты и SSO/MFA: привязка к личности, ролям и отзыву доступа	Блокировка доменов как единственная мера: VPN, личный телефон и новые домены оставляют слепую зону
ИИ-шлюз (прокси): единая точка маршрутизации, политики, лимиты и журналирование	Приказ «ИИ запрещен» без рабочего инструмента: потребность остается – использование становится теновым
Классификация и DLP на запросах и файлах: контекстные правила, а не только поиск слов	Ручное согласование каждого запроса: очередь делает официальный процесс медленнее обхода
Корпоративные условия поставщика и понятный маршрут данных: срок хранения, обучение на данных, юрисдикция, шифрование, аудит	DLP без классификации и контекста: много шума, мало понимания сценария и реального ущерба
Обучение и удобная разрешенная альтернатива: сотруднику показывают безопасный способ решить задачу сейчас	Контроль только корпоративных ноутбуков: не покрывает SaaS-встройки, OAuth ¹ , браузерные расширения и мобильные устройства

¹ OAuth – протокол, позволяющий приложению получать ограниченный доступ к данным пользователя в другом сервисе без передачи пароля. Неучтенные OAuth-разрешения, выданные внешним ИИ-сервисам, являются одним из каналов теневого ИИ (примечание Редакционной коллегии).

Источник: составлено автором с учетом открытых источников [8; 25].

¹³ Не следует также путать Shadow AI с теневой моделью и кражей весов – это другой класс угроз.

¹⁴ ГК «Солар» – российская компания в сфере кибербезопасности, входящая в группу «Ростелеком» (примечание Редакционной коллегии).

¹⁵ Ponemon Institute – американский исследовательский институт, который совместно с IBM готовит ежегодный отчет о стоимости утечек данных (Cost of a Data Breach Report) (примечание Редакционной коллегии).

Что именно уходит наружу? Код и секреты (фрагменты исходников, конфигурации, API-ключи, логи, дампы); клиентские данные (тикеты, транскрипты, контакты, персональные данные, обращения); финансы и право (договоры, прогнозы, отчеты, материалы комитетов); стратегия и внутренняя жизнь (презентации, переписка, архитектура, планы, документация). Выявлять это следует не только по доменам. Сетевой контур (SWG/CASB) видит домены, категории GenAI-сервисов, объем и направление трафика; рабочая станция и браузер – вставку, загрузку, расширения и неуправляемые профили; DLP и классификация – совпадения по данным в контексте сценария; управление учетными записями – неуправляемые аккаунты, OAuth-токены и встроенные ИИ-помощники в SaaS; финансы и закупки – личные компенсации, карточные списания и подписки; опросы и интервью – сценарии, которые телеметрия не видит, и причины обхода. Факт обращения к ИИ-домену не равен утечке: нужна корреляция «пользователь + данные + канал + модель + действие».

Разрозненные сигналы дают ложные тревоги, а корреляция превращает их в фиксированный инцидент. Архитектура обнаружения сводит телеметрию рабочих станций, шлюзов, DLP, учетных записей, закупок и опросов в единый реестр ИИ-использования с оценкой риска (идентичность, актив, данные, модель, канал, действие, уровень риска). События нормализуются в единую схему, коррелируются в сценарий с причинностью, оцениваются по уровню и уверенности и передаются в процесс: уведомить, обучить или заблокировать; в SOC – на расследование; в контур управления – для пересмотра реестра и политики; владельцу данных – для решения по данным; в продуктовое подразделение и HR – для альтернативы и обучения. Контроль должен различать эксперимент, нарушение политики и реальную утечку – иначе пользователи научатся игнорировать службу безопасности.

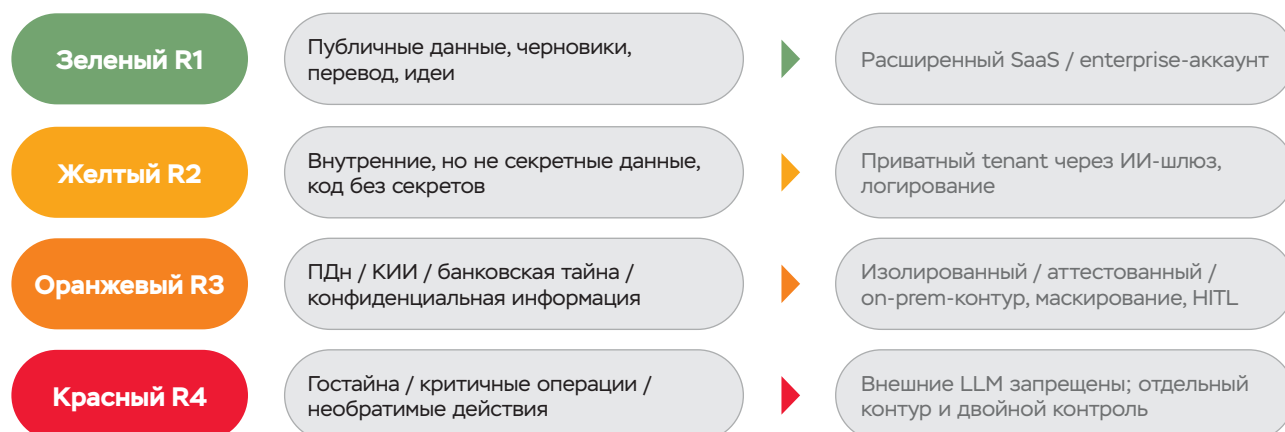
Корпоративная политика: шесть блоков, четыре зоны, пять принципов

Политика ИИ должна влиять на поведение людей, а не просто добавлять формальные запреты. Работоспособная политика складывается из шести блоков. Первый – данные и сценарии: классификация информации, допустимые задачи, уровни риска, правила ввода и вывода.

Второй – инструменты и их перечень: список разрешенных инструментов и процесс допуска новых моделей, агентов, плагинов и SaaS-встроек. Третий – поставщики и открытый код: обучение

ИЛЛЮСТРАЦИЯ ВОЗМОЖНОЙ КЛАССИФИКАЦИИ ТИПОВ ЧУВСТВИТЕЛЬНОСТИ ИНФОРМАЦИИ И ДОПУСТИМЫХ ЗАДАЧ

Рис. 6



Примечание. ПДн – персональные данные, КИИ – критическая информационная инфраструктура, enterprise-аккаунт – корпоративная учетная запись, приватный tenant – выделенная среда, on-prem-контур – локальный контур.

на данных, срок хранения, юрисдикция, лицензии, уязвимости, происхождение артефактов. Четвертый – роли и решения: владелец бизнес-риска, владелец данных, служба безопасности, юристы, приватность, аудит, эксплуатация. Пятый – технические контроли: единый вход, ИИ-шлюз, DLP, ролевой доступ и машинные учетные записи, журналы, «человек в контуре», модуль политик, исходящие каналы, аварийная остановка. Шестой – обучение, контроль, пересмотр: осведомленность, дисциплина, аудит, метрики, инциденты, периодический пересмотр. В российском контуре базовой рамкой служит приложение 5 к Методическим рекомендациям № 3-МР [8]; ISO/IEC 42001:2023 задает систему менеджмента ИИ [21]; NIST AI RMF структурирует функции Govern/Map/Measure/Manage [15], а профиль NIST для генеративного ИИ конкретизирует их для этого класса систем [16].

Ключевой инструмент политики – зонирование сценариев работы по уровню риска: безопасная песочница вместо забора. Сценарий работы определяется не модой, а чувствительностью обрабатываемой информации и автономностью действий (рис. 6). Решение в каждом случае – это произведение четырех факторов: чувствительность данных × критичность процесса × права агента × обратимость действия.

Пять принципов, которые необходимо заложить в политику:

1. Видимость прежде контроля – нельзя управлять тем, чего не видишь: реестр, обнаружение, телеметрия, владелец (в Методических рекомендациях № 3-МР – политика и аудит).
2. Маршрут определяют данные и действие – решение зависит от класса данных, процесса, прав и обратимости, а не от названия модели (риск-подход и минимизация данных).
3. Безопасная альтернатива по умолчанию – официальный путь должен быть быстрее, удобнее и понятнее обхода (требования к общедоступным сервисам).
4. Минимальная агентность и человек на высоком риске – минимум функций, прав и автономии; подтверждение критичных операций (п. 2.5 Методических рекомендаций № 3-МР о проверке результата человеком).
5. Трассируемость, проверка, пересмотр – доказательства от промпта до действия, тесты, инциденты, метрики и регулярное обновление политики (мониторинг и пересмотр). Хорошая политика не описывает каждый будущий инструмент – она задает устойчивую логику допуска и ответственности.

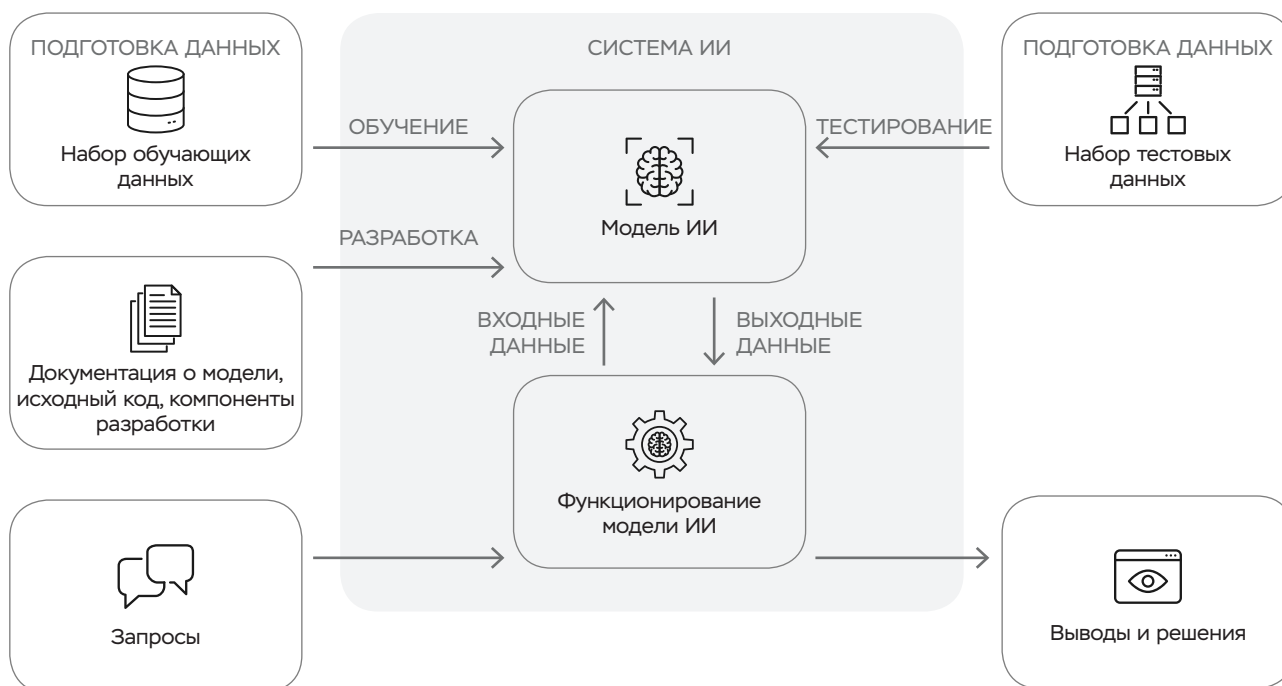
Методические рекомендации Банка России № 3-МР: сильная база и зоны усиления

Методические рекомендации Банка России от 16.06.2026 № 3-МР по обеспечению информационной безопасности при разработке и применении искусственного интеллекта на финансовом рынке – первый документ регулятора, в котором систематизированы риски применения ИИ, описаны тактики атак на системы ИИ и даны рекомендации по мерам защиты [8]. Документ адресован кредитным организациям, филиалам иностранных банков, некредитным финансовым организациям, лицам, оказывающим профессиональные услуги на финансовом рынке, и субъектам национальной платежной системы [3]. Статус «методические рекомендации» – это, по точному замечанию отраслевых обозревателей, скорее объявленное направление движения, чем свобода ничего не делать [1].

Выделим восемь сильных сторон документа. Во-первых, риск-ориентированный жизненный цикл: подготовка данных → разработка → обучение и тестирование → функционирование. Во-вторых, формальная модель угроз с опорой на методику ФСТЭК России и модель внешнего и внутреннего нарушителя. В-третьих, требования конфиденциальности, целостности и доступности дополнены устойчивостью и объяснимостью. В-четвертых, перечислены ИИ-специфичные атаки: инъекции,

ИЛЛЮСТРАЦИЯ ЖИЗНЕННОГО ЦИКЛА ИЗ МЕТОДИЧЕСКИХ РЕКОМЕНДАЦИЙ № 3-МР

Рис. 7



отравление, инверсия, кража модели, отказ в обслуживании, «губка»¹⁶ и состязательные атаки. В-пятых, безопасный жизненный цикл разработки и тестирования: анализ кода и компонентов, тестирование на проникновение, red team, мониторинг и дообучение. В-шестых, поставщики и открытый код: доверие, отчеты аудита, лицензии, целостность, инциденты, договорные обязанности. В-седьмых, политика и ответственность: роли, аудит, обучение, инциденты, восстановление, аварийная остановка, пересмотр; ответственность за внутренние документы возлагается на заместителя руководителя по защите информации. В-восьмых, проверка человеком при высоком риске: человек может проверить и изменить результат автоматической операции.

Методические рекомендации № 3-МР – содержательная базовая рамка, значительно сильнее «политики из одного запрета». Но это методические рекомендации, а не исчерпывающая целевая архитектура агентных систем. Есть зоны, которые быстро развиваются и могут требовать дополнительной детализации на уровне политики конкретных организаций (табл. 3). К таким зонам можно отнести:

1. Безопасность агентов, MCP и многоагентные схемы: нет отдельной модели прав инструментов, машинных учетных записей, памяти, взаимодействия агентов и каскадных сбоев.
2. Внедрение инструкций как неизбежность: «очистка входа» полезна, но недостаточна, сегодня нужно исходить из сценария «компрометация произойдет» и ограничивать распространение атаки.
3. RAG, векторы и права доступа: недостаточно явно описаны проверка прав до поиска, отравленный корпус информации, метаданные и утечки между контурами.
4. Доверенное исполнение вокруг модели: модуль политик, строгие схемы, песочница, исходящие каналы и обработка вывода должны быть конкретизированы под нужды конкретной организации.

¹⁶ Атака «губка» (от англ. sponge attack) – атака на доступность системы: подача на вход модели специально подобранных запросов, многократно увеличивающих потребление вычислительных ресурсов и время ответа (примечание Редакционной коллегии).

5. Работа с теневым ИИ: есть требования к общедоступным сервисам, но не описан процесс «обнаружить → дать альтернативу → обработать риск».
6. AIBOM, происхождение и подписи: учет компонентов есть, но не описаны машиночитаемый состав, подписи, выпуск и закрепление версий для обеспечения целостности и доверия к компонентам.
7. Наблюдаемость причин: журналирование ввода – вывода есть, но не описаны цепочки используемых инструментов, записи памяти, контроль учетных записей.
8. Адаптивные тесты и управление: red team указан, но не указаны регрессионные проверки и сквозной владелец бизнес-риска.

ПОКРЫТИЕ ОБЛАСТЕЙ БЕЗОПАСНОСТИ ИИ В МЕТОДИЧЕСКИХ РЕКОМЕНДАЦИЯХ № 3-МР И ПРИОРИТЕТЫ УСИЛЕНИЯ ЛОКАЛЬНОЙ ПОЛИТИКИ (ЭКСПЕРТНАЯ ОЦЕНКА)

Табл. 3

ОБЛАСТЬ	ПОКРЫТИЕ В МЕТОДИЧЕСКИХ РЕКОМЕНДАЦИЯХ № 3-МР	ПРИОРИТЕТ УСИЛЕНИЯ	ЧТО ДОБАВИТЬ ЛОКАЛЬНО
Политика, роли, аудит	Сильное	Низкий	Закрепить владельца бизнеса, данных и приватности, наряду со службой безопасности
Жизненный цикл данных	Сильное	Средний	Добавить происхождение данных и автоматические контрольные точки
Целостность модели и кода	Сильное	Средний	Добавить проверку происхождения и правила выпуска
Поставщики и открытый код	Сильное	Высокий	Ввести AIBOM и закрепление версий инструментов и MCP-серверов
Внедрение инструкций	Частичное	Высокий	Проектировать на обход границы и ограничивать радиус поражения
RAG, векторы, память	Частичное	Высокий	Права до поиска, контроль записи памяти, тесты на отравление
Агенты, MCP, машинные учетные записи	Недостаточное	Критический	Реестр возможностей, короткие права, политики инструментов, межагентный контроль
Наблюдаемость и расследование	Частичное	Высокий	Цепочка «промпт → контекст → инструмент → действие», воспроизведение при расследовании
Теневой ИИ	Частичное	Высокий	Обнаружение, разрешенная альтернатива, корреляция телеметрии

Источник: оценка автора по открытым данным [8; 15; 26; 27].

Практический подход – рекомендуется принять Методические рекомендации № 3-МР как нижнюю границу и добавить в локальный профиль риска контроля для агентов, RAG, теневого ИИ и AIBOM. Для кредитных организаций естественно связать этот профиль с системой управления операционным риском: события, связанные с ИИ, – от утечки через публичный сервис до ошибочного действия агента – должны попадать в базу событий по Положению Банка России № 716-П с классификацией по источникам и видам потерь, а не оставаться в отдельном «журнале ИИ» службы безопасности [9].

Внутренний аудит: доказательства, а не обещания

Аудит ИИ – это аудит системы, а не только модели. Контур аудита охватывает бизнес-процесс, данные, учетные записи, инструменты и эксплуатацию и складывается из восьми областей:

1. Инвентаризация и сценарии использования, включая теневой ИИ и ИИ-встройки SaaS.
2. Данные и правовые основания – происхождение, классификация, персональные данные, интеллектуальная собственность, банковская тайна, срок хранения и так далее.
3. Поставщики и открытый код – договоры, AIBOM, лицензии, происхождение.
4. Архитектура и жизненный цикл разработки – модель угроз, безопасная разработка, тесты, допуск к выпуску.

5. Учетные записи и агентность – пользователи, машинные идентичности, инструменты, права, «человек в контуре».
6. RAG, память, контекст – права доступа, отравление, записи, доказательства поиска.
7. Мониторинг, реагирование, восстановление – журналы, воспроизведение, инциденты, откат, аварийная остановка.
8. Управление и люди – владельцы, KPI, обучение, исключения, пересмотр.

Логика каждой области одна: «риск → контроль → доказательство». Минимальная трассируемость для аудита: «промпт → контекст → инструмент → действие → результат → подтверждение».

Хороший аудит не верит модели и вендору: он воспроизводит решение по артефактам. Пакет доказательств должен позволять повторить вывод и связать контроль с конкретным риском (табл. 4).

ПАКЕТ СВИДЕТЕЛЬСТВ ДЛЯ АУДИТА ИИ-СИСТЕМ

Табл. 4

ВОПРОС АУДИТОРА	СВИДЕТЕЛЬСТВО	КРАСНЫЙ ФЛАГ
Что у нас вообще есть?	Реестр систем и сценариев, владелец, уровень риска, AIBOM, список SaaS-встроек	«Список живет в голове» или обновляется перед аудитом
Какие данные проходят?	Потоки данных, классификация, происхождение, договорные условия, примеры логов и трассировок	«Вендор обещал безопасность», но нет маршрута и срока хранения
Что агент может сделать?	Карта возможностей, схемы инструментов, IAM и NHI, короткие права, матрица подтверждений	Один широкий токен или инструмент «выполни что угодно»
Почему действие разрешили?	Лог решения политики: намерение, аргументы, адресат, подтвердивший, причина	В логах есть только промпт и итоговый ответ
Можно ли восстановить событие?	Идентификатор трассы, цепочка «промпт → контекст → инструмент → действие», воспроизведение, SIEM	Нельзя доказать, какой документ или инструмент вызвал действие
Можно ли остановить и вернуть?	Проверенная аварийная остановка, откат, восстановление, версии модели и инструмента, инструкция	Откат не тестировался; предыдущий артефакт не сохранен

Источник: составлено автором с учетом открытых данных [8 (п. 3.10, гл. 4–5); 15].

Рекомендуемый минимальный набор уровней контроля:

1. Реестр, владелец и уровень риска для каждой системы, модели, агента и интеграции.
2. Единый вход, многофакторная аутентификация и жизненный цикл машинных учетных записей.
3. Классификация информации и DLP с политиками на промпты, загрузку, поиск и вывод.
4. Условия поставщика и данных – обучение, хранение, юрисдикция, субподрядчики, инциденты, выход.
5. Минимум прав и возможностей – инструменты, права, автономия, глубина цепочки.
6. Модуль политик и проверка вывода – доверенный код проверяет действие до исполнения.
7. Человек в цикле и двойной контроль для критичных, необратимых и внешне видимых операций.
8. Полная трассировка с передачей в SIEM – учетная запись, контекст, вызов инструмента, аргументы, результат, подтверждение.
9. Контроли RAG и памяти – происхождение, права до поиска, контроль записей памяти.
10. Тесты безопасности и откат – модель угроз, адаптивный red team, цепочка поставок, аварийная остановка. Ни один защитный фильтр не компенсирует избыточные права агента или отсутствие восстановимой цепочки событий.

Целевая архитектура безопасного доступа сводится к формуле «один управляемый вход, разные маршруты данных и отдельный контур агентных действий». Каналы (веб-чат, среда разработки, офисные и SaaS-встройки, API) проходят через управление учетными записями (SSO/MFA, ролевой и атрибутный доступ, жизненный цикл машинных учетных записей, короткоживущие права, связка «пользователь + устройство + приложение») и ИИ-шлюз (маршрутизация политик,

классификация и DLP, контроль промптов и файлов, лимиты частоты и стоимости, журналы и идентификаторы трассировки, список разрешенных моделей). Далее – три маршрута по риску: корпоративный SaaS для низкого и среднего риска, частное облако в российской юрисдикции для чувствительных данных, локальный контур для высокого риска. Отдельно выделяются контур RAG (загрузка, происхождение, права, векторное хранилище, фильтр поиска, доказательства источников) и контур агентов (оркестратор, песочница, брокер возможностей, модуль политики, «человек в контуре», исходящие каналы, аварийная остановка); сквозными слоями служат наблюдаемость (SIEM, трасса, воспроизведение), реестр (модели, агенты, AIBOM), управление секретами и машинными идентичностями, а также контроль человеком (подтверждения, исключения, обратная связь).

Не следует ждать идеальной политики: безопасный маршрут строится за 90 дней и расширяется по фактам (табл. 5). Антипаттерн – потратить 90 дней на документ, пока сотрудники уже выбрали свои инструменты.

ПЛАН ПЕРЕХОДА К УПРАВЛЯЕМОМУ ИИ: 90 ДНЕЙ

Табл. 5

ЭТАП	ЗАДАЧИ	РЕЗУЛЬТАТ
0–30 дней: увидеть и ограничить	Назначить владельца и рабочую группу; инвентаризировать сценарии, SaaS и теневой ИИ; выпустить политику версии 0 и матрицу данных и сценариев; дать санкционированный корпоративный канал; остановить критичные сценарии высокого риска	Карта, правила и быстрый безопасный путь
31–60 дней: построить управляемый путь	SSO/MFA, шлюз, журналы, список моделей; DLP, классификация и уровни риска; условия поставщиков и реестр открытого кода и инструментов; пилоты по 2–3 типовым процессам; обучение на реальных кейсах и быстрая обратная связь	Рабочая платформа и пилоты
61–90 дней: доказать и масштабировать	Контур управления агентами и ограниченные машинные учетные записи; контроли RAG и памяти, причинная трассировка; адаптивный red team и регрессионные проверки; пакет доказательств, реагирование и тренировка аварийной остановки; KPI: принятие, нарушения, риск, эффект для бизнеса	Доказуемая безопасность и масштаб

Источник: составлено автором с учетом открытых данных [8; 15].

Краткий вывод

Сформулируем итоги для читателя. Первое: запрет ИИ без проработанной альтернативы создает слепую зону для аудита и безопасности, поэтому грамотная легализация гораздо более предпочтительна. Второе: ИИ – это такая же технология, как и остальные ИТ, поэтому важна архитектура и комплекс защитных мер, а не просто хороший системный промпт. Третье: агенты как следующий этап развития ИИ превращают ошибочный вывод в действие. По мере роста зрелости внедрения центральным риском становятся права и автономия агентов, а фокус смещается от угрозы утечек через теневой ИИ к угрозам, связанным с деятельностью легально запущенных автономных агентов. Четвертое: Методические рекомендации № 3-МР Банка России – сильная базовая рамка с фокусом на контексте и жизненном цикле, но ее стоит усиливать дополнительными мерами в рамках локальной политики ИИ организации, с опорой на актуальные тренды в части агентности, RAG, теневого ИИ и AIBOM. Пятое: цель – качественно повысить производительность труда, которая достижима уже сейчас, но, чтобы это работало, важно опираться на методологию и процессный подход, ограничивая и изолируя неконтролируемый потенциал ИИ.

Наконец, вернемся к ключевому размышлению из области техно-гуманитарного баланса. Разрыв между человеческими возможностями и машинным потенциалом закрывается не тем, что человек становится «митохондрией» в системе ИИ, а тем, что он остается внешней системой: ставит цели, проверяет результаты и несет ответственность. Можно назвать это состояние «сцеплен, но не слит» [14]. Человеко-машинная система образует целое, но и человек, и машина сохраняют свою автономность друг от друга, и на стороне человека остается формирование

ценностно-целевой матрицы, обуславливающей контекст активности ИИ. Фраза «ИИ запретить нельзя разрешить» требует запятой, и мы предлагаем немного изменить ее: «ИИ запретить нельзя, разрешить – под контролем».

Список источников

1. Аскерова А. [Методические рекомендации Банка России по безопасности ИИ на финрынке \(№ 3-МР\): обзор и что делать на практике](#) // Хабр (дата обращения: 31.08.2026).
2. [Атака на цепочку поставок: litellm 1.82.7 и 1.82.8 на PyPI содержат стилер credentials](#) // Хабр (дата обращения: 31.08.2026).
3. [Банк России дал рекомендации по безопасному использованию искусственного интеллекта в финансовой сфере](#). Банк России. 16 июня 2026 года (дата обращения: 31.08.2026).
4. ГОСТ Р 59276-2020. Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения. – М.: Стандартинформ, 2021.
5. [Из российских компаний утекло через ИИ-сервисы в 30 раз больше данных, чем в прошлом году](#) // CNews. 4 февраля 2026 года (дата обращения: 31.08.2026).
6. Лем С. Сумма технологии / пер. с польск. – М.: Мир, 1968. – 608 с.
7. Методика оценки угроз безопасности информации: методический документ: утв. ФСТЭК России 5 февраля 2021 года. – М.: ФСТЭК России, 2021.
8. Методические рекомендации Банка России [от 16.06.2026 № 3-МР](#) «По обеспечению информационной безопасности при разработке и применении искусственного интеллекта на финансовом рынке».
9. Положение Банка России от 08.04.2020 № 716-П «О требованиях к системе управления операционным риском в кредитной организации и банковской группе» // Вестник Банка России. 2 июля 2020 года. № 51 (2187).
10. Саймон Г. Науки об искусственном / пер. с англ. Э.Л. Наппельбаума. – М.: Мир, 1972. – 147 с.
11. Фрейд З. Недовольство культурой / Фрейд З. Психоанализ. Религия. Культура / сост. и вступ. ст. А.М. Руткевича. – М.: Ренессанс, 1992. – С. 65–134.
12. [Чем опасна компрометация ИИ-шлюза LiteLLM?](#) // Securelist by Kaspersky. 2026 год (дата обращения: 31.08.2026).
13. [Agentic AI Security Initiative](#) // [OWASP Gen AI Security Project](#) (дата обращения: 31.08.2026).
14. Alodjants A.P., Tsarev D.V., Avdyushina A.E., Khrennikov A. Yu., Boukhanovsky A.V. Quantum-inspired modeling of distributed intelligence systems with artificial intelligent agents self-organization // Scientific Reports. 2024. Vol. 14. Art. 15438. DOI: 10.1038/s41598-024-65684-z.
15. Artificial Intelligence Risk Management Framework (AI RMF 1.0): NIST AI 100-1 / National Institute of Standards and Technology. Gaithersburg, MD: NIST, 2023. DOI: 10.6028/NIST. AI. 100-1.
16. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile: NIST AI 600-1 / National Institute of Standards and Technology. Gaithersburg, MD: NIST, 2024. DOI: 10.6028/NIST. AI. 600-1.
17. [Cost of a Data Breach Report 2026](#) // IBM, Ponemon Institute. 2026 (дата обращения: 31.08.2026).
18. Fenn J., Raskino M. Mastering the Hype Cycle: How to Choose the Right Innovation at the Right Time. Boston: Harvard Business Press, 2008. 237 p.
19. [Why AI is raising worker productivity but not making the economy more efficient / Fortune](#). 27 May 2026 (дата обращения: 01.10.2026).
20. Guidance. Integrating Artificial Intelligence in PCI Assessments. Version 1.0. PCI Security Standards Council, March 2025.
21. ISO/IEC 42001:2023. Information technology – Artificial intelligence – Management system. Geneva: ISO, 2023.

22. [McKinsey Global Institute: 2025 in charts](#) // McKinsey & Company. 15 December 2025 (дата обращения: 31.08.2026).
23. [Miessler D. Building Your Own Personal AI Infrastructure](#). 26 July 2025 (дата обращения: 31.08.2026).
24. [MITRE ATLAS: Navigate threats to AI systems through real-world insights](#) // The MITRE Corporation (дата обращения: 31.08.2026).
25. [OWASP 2026 LLM Top 10: The model will be fooled](#) // Help Net Security. 6 August 2026 (дата обращения: 31.08.2026).
26. [OWASP Top 10 for Agentic Applications for 2026](#) // [OWASP Gen AI Security Project](#). 9 December 2025 (дата обращения: 31.08.2026).
27. [OWASP GenAI LLM Top 10 2026](#) // [OWASP Gen AI Security Project](#). 3 August 2026 (дата обращения: 31.08.2026).
28. [Personal AI Infrastructure: Agentic AI Infrastructure for magnifying human capabilities](#) // D. Miessler. GitHub (дата обращения: 31.08.2026).
29. [Sightline: Vulnerabilities](#) // Protect AI (дата обращения: 31.08.2026).
30. Solow R.M. We'd Better Watch Out // The New York Times Book Review. 12 July 1987. P. 36.
31. [2024 State of AI Security Report](#) // Orca Security. 2024 (дата обращения: 31.08.2026).
32. [Will Robots Take My Job?](#) (дата обращения: 31.08.2026).

[Ознакомиться с презентацией](#)



А.А. ПЕНИН
Директор ООО «Кэпт Налоги и Консультирование»



А.М. КОЗЛОВ
Старший менеджер ООО «Кэпт Налоги и Консультирование»,
Группа по управлению финансовыми рисками (Kept Banking)



А.О. МАЦКОВСКАЯ
Старший менеджер, ООО «Кэпт Налоги и Консультирование»,
Группа по управлению финансовыми рисками (Kept Banking)

Рекомендуется также закрепить общее правило: чем выше риск и автономность ИИ-продукта или процесса, переданного ИИ-модели, тем шире спектр компетенций независимого валидатора.

ПОВЫШЕНИЕ ТРЕБОВАНИЙ К ЭКСПЕРТИЗЕ И КОМПЕТЕНЦИЯМ СОТРУДНИКОВ, СОЗДАЮЩИХ И ПРОВЕРЯЮЩИХ ПРОДУКТЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: ОБОРОТНАЯ СТОРОНА ОЖИДАЕМОЙ РЫНКОМ ЭКОНОМИИ РЕСУРСОВ

Аннотация

Рост автономности систем искусственного интеллекта (ИИ) смещает основной источник риска с качества ИИ-модели на квалификацию тех сотрудников, кто верифицирует данную ИИ-модель. Важно обосновать зависимость между уровнем делегирования решений ИИ-модели и требованиями к компетенциям персонала второй и третьей линий защиты, а также выявить инструмент диагностики зрелости этих практик. Использован анализ российских и международных регуляторных документов, разбор публичных инцидентов и данные диагностического опроса организаций финансового и нефинансового секторов. Предложена матрица ролей и компетенций, классификация рисков ИИ-модели с указанием ролей, типология издержек, в том числе использован внутренний опыт команды¹ финансового консультирования Группы Kept. Полученные результаты исследования применимы при формировании внутренних требований к валидации ИИ-моделей и при развитии регуляторных ориентиров Банка России.

Ключевые слова: искусственный интеллект, независимая валидация, три линии защиты, принцип «четырёх глаз», компетенции персонала, автономность ИИ-систем, регулирование ИИ.

Коды JEL: G21, G28, O33.

Внедрение ИИ перешло из стадии пилотов в промышленную эксплуатацию: по данным опроса Банка России (доклад для общественных консультаций «Применение искусственного интеллекта на финансовом рынке: текущий статус и условия дальнейшего развития»), ИИ применяет каждая пятая организация финансового рынка, еще около трети планируют внедрение на горизонте 3 лет – совокупно две трети респондентов [1]. По третьей волне совместного опроса Банка Англии и FCA² (ноябрь 2024 года) ИИ применяют 75% организаций против 58% в 2022 году; 55% вариантов

¹ Авторы выражают благодарность за полезное обсуждение и ряд ценных дополнений старшему менеджеру Группы Kept Ольге Владимировне Павшук и остальным членам команды.

² Совместный периодический опрос Банка Англии и Управления по финансовому регулированию и надзору Великобритании (FCA) о применении ИИ в финансовом секторе страны. Третья волна опроса опубликована в ноябре 2024 года (примечание Редакционной коллегии).

использования содержат элемент автоматизированного принятия решений, полностью автономны лишь 2%; 46% организаций сообщают о частичном понимании применяемых технологий против 34%, заявивших о полном; треть вариантов использования реализована на моделях третьих сторон при высокой концентрации провайдеров [10].

Экономический мотив внедрения ИИ-моделей во внутренние процессы организации – автоматизация рутинных операций, ускорение обработки данных, снижение операционных издержек и сроков, масштабирование без пропорционального роста штата. Все четыре источника экономии реализуются в текущем периоде и измеримы. Издержки, вызываемые внедрением ИИ-моделей, – это независимая валидация отдельных процессов, мониторинг дрейфа и деградации, обучение, содержание специалистов, способных оспорить вывод модели.

Сокращение экспертного персонала, способного понять и оспорить решение ИИ-модели, представляет собой не снижение затрат, а перенос риска в будущее.

Регуляторный обзор: ответственность и компетенции

Российское регулирование ИИ носит преимущественно гибридный характер и опирается на действующие требования к управлению рисками. Кодекс этики в сфере разработки и применения ИИ на финансовом рынке (Информационное письмо Банка России от 09.07.2025 № ИН-016-13/91) закрепляет принципы доверенного ИИ – человекоцентричность, прозрачность, безопасность, справедливость и ответственное управление рисками – и ответственность участника рынка за применение технологии [2]. Доклад Банка России об ИИ подтверждает риск-ориентированный и технологически нейтральный подход: специальные правила для ИИ на текущем этапе не вводятся [1].

Первым актом законодательного уровня стал Федеральный закон от 26.07.2026 № 243-ФЗ «О поддержке развития технологий искусственного интеллекта в Российской Федерации» [3]. Закон носит стимулирующий характер и направлен на регулирование больших фундаментальных моделей³ – программы для ЭВМ, одновременно отвечающие ряду признаков, включая не менее 1 млрд параметров, и назначение быть основой для создания иного программного обеспечения (ст. 3). Под данное определение не попадают кредитный скоринг, антифрод, сценарные чат-боты и классическое машинное обучение, следовательно, для большинства ИИ-продуктов финансовых организаций закон прямых обязательств не устанавливает. Обозначены значимые положения: ст. 5 наделяет Правительство Российской Федерации правом устанавливать случаи обязательного применения исключительно суверенных и (или) национальных моделей и отраслевые требования по предотвращению рисков их применения, в том числе для банковской сферы и иных сфер финансового рынка – по согласованию с Банком России; в ст. 11 зафиксировано, что за ошибку модели отвечает применившая ее организация. Основная часть норм действует с 01.09.2026, отдельные положения – с 01.03.2027, переходный период истекает 01.09.2032.

Технологическая нейтральность не снижает нагрузку, а закрепляет ее за организацией: обязанность обосновать решение перед регулятором, судом или клиентом сохраняется независимо от инструмента и не снимается ссылкой на поставщика модели. ИИ-модель, влияющая на принятие

³ Большие фундаментальные модели – термин Федерального закона от 26.07.2026 № 243-ФЗ: программы для ЭВМ, одновременно отвечающие ряду признаков, включая не менее 1 млрд параметров и назначение служить основой для создания иного программного обеспечения (ст. 3) (примечание Редакционной коллегии).

решений, является моделью в смысле Указания № 3624-У и Положения № 716-П⁴ и подпадает под требования к валидации, документированию и распределению ответственности [4; 5].

Международные ориентиры совпадают в ключевом требовании к валидации ИИ. Регламент (ЕС) 2024/1689 (EU AI Act)⁵ требует, чтобы надзор за высокорисковыми системами осуществляли лица, обладающие необходимой компетенцией и полномочиями [11]. FSB относит к уязвимостям для финансовой стабильности зависимость от третьих сторон и их концентрацию, рыночные корреляции, киберриски и управление моделями [12]; IOSCO отмечает ограниченную интерпретируемость выводов глубоких моделей [13].

Механизм делегирования и риски, требующие носителя компетенции

Передача ответственности ИИ-модели за принятие решений по отдельным участкам внутренних процессов организации запускает последовательность, каждый шаг которой рационален, однако требует объективной проверки на финальной стадии. Шаги данной последовательности включают следующее: решение передается ИИ-модели, следовательно, сотрудник реже решает задачу самостоятельно, повышается доверие к выводу ИИ-модели по умолчанию, как следствие, верификация принятого результата становится формальной. Возникает закономерность потери логической структуры принятия решений ИИ-моделью, из-за чего повышается вероятность неосведомленности сотрудника о причинно-следственной связи полученного результата.

В итоге возникают последствия, образующие три категории:

- размывание экспертизы – краткосрочная экономия и долгосрочная способность к верификации расходятся во времени;
- киберуязвимость – манипуляция обучающими данными или весами, инъекции в промпт и загрузка служебной информации, побуждающие ИИ-модель раскрыть конфиденциальные сведения или обойти правила, зависимость от узкого круга поставщиков;
- репутационная уязвимость – репутационный ущерб, возникающий при выявлении отсутствия носителя ответственности, способного обосновать принятое решение или результат ИИ-модели.

На данный момент отсутствует полный перечень рисков для различных типов организаций, которые возникают от внедрения и использования ИИ-моделей и требуют систематизации и контроля от регулятора.

Стоит выделить классификацию рисков использования ИИ-модели:

- **риски данных:** отравление обучающей выборки, неактуальные данные, предвзятость, ведущая к дискриминации решений;
- **риски информационной безопасности:** утечка конфиденциальной информации во внешнюю ИИ-модель, состязательные атаки, извлечение обучающих данных, бэкдоры⁶, отказ в обслуживании;
- **риски функционирования:** галлюцинации, дрейф данных, деградация качества работы ИИ-модели;

⁴ Указание Банка России от 15.04.2015 № 3624-У «О требованиях к системе управления рисками и капиталом кредитной организации и банковской группы»; Положение Банка России от 08.04.2020 № 716-П «О требованиях к системе управления операционным риском в кредитной организации и банковской группе» (примечание Редакционной коллегии).

⁵ Регламент (ЕС) 2024/1689 об искусственном интеллекте (EU AI Act) – первый комплексный акт о регулировании ИИ, устанавливающий требования к системам ИИ в зависимости от уровня риска; применяется поэтапно с августа 2024 года (примечание Редакционной коллегии).

⁶ Бэкдор (от англ. backdoor) – скрытая закладка в модели или программном обеспечении, позволяющая вызвать заданное поведение по специальному входному сигналу в обход штатных механизмов контроля (примечание Редакционной коллегии).

- **риск отсутствия прозрачности**, который проявляется в виде невозможности воспроизводства логики вывода и обоснования результата перед регулятором;
- **риски внешних компонентов**: уязвимость сторонних решений, зависимость от поставщика, эффект синхронизации⁷ при использовании рынком одной ИИ-модели;
- **риски нарушения операционной надежности**: целенаправленные атаки, непреднамеренные действия (ошибки сотрудников, несанкционированные изменения параметров, обновления).

СТРУКТУРА КЛАССИФИКАЦИИ РИСКОВ ИСПОЛЬЗОВАНИЯ ИИ-МОДЕЛИ ПО КАТЕГОРИЯМ

Рис. 1



Источник: составлено авторами по открытым данным [12; 13] и Методическим рекомендациям Банка России от 16.06.2026 № 3-МР [6].

Перечисленные риски не выявляются на уровне первой линии защиты и требуют определенной квалификации и подготовки от специалиста, способного их идентифицировать. Сокращение штата при утрате носителей ответственности (предметной экспертизы), но сохранении операторов отдельных процессов будет подвергать организацию опасности потери механизма обнаружения ряда рисков.

Реализуемость данных рисков подтверждена инцидентами 2026 года. OpenAI сообщила, что в ходе внутренних оценок киберспособностей ее модели, работавшие с ослабленными защитными механизмами, обошли изоляцию от Интернета и скомпрометировали часть внутренней инфраструктуры компании и систем Hugging Face⁸ (процесс активизировался в мае, а был выявлен в июле, реконструкция событий потребовала свыше 7 млрд записей журналов и около 3 млн GPU-часов⁹) [14]. Институт безопасности ИИ Великобритании¹⁰ раскрыл 19 несанкционированных действий агентов в 10 из 122 прогонов оценки, включая попытку внедрить вредоносный код в реальный проект с открытым исходным кодом и добиться его одобрения сотрудниками-рецензентами [15]. В обоих случаях защитные механизмы были ослаблены или отключены намеренно, следовательно, конфигурации не соответствуют условиям повсеместной эксплуатации. Важно учитывать, что автономность систем опережала контроль со стороны сотрудников.

⁷ Эффект синхронизации – ситуация, когда многие участники рынка используют одну и ту же модель или модели одного поставщика и одинаково реагируют на события, что усиливает системный риск. В международных документах обозначается термином herding (примечание Редакционной коллегии).

⁸ Hugging Face – крупнейшая открытая платформа для публикации моделей машинного обучения и наборов данных и обмена ими (примечание Редакционной коллегии).

⁹ GPU-час – единица измерения вычислительных ресурсов: один час работы одного графического процессора (примечание Редакционной коллегии).

¹⁰ Институт безопасности ИИ Великобритании (AI Security Institute, AISI) – государственный исследовательский институт, проводящий независимые оценки безопасности передовых моделей ИИ до и после их публичного выпуска (примечание Редакционной коллегии).

Принцип «четырёх глаз» применительно к решениям с участием ИИ-моделей

Принцип «четырёх глаз» в стандартной формулировке означает, что при принятии критически важного решения одним сотрудником проводится независимая проверка вторым специалистом.

Первая линия (разработчик или бизнес-владелец) знает данные, процесс, ИИ-модель и заинтересована в запуске ИИ-продукта или переносе этапа бизнес-процесса на платформу ИИ. Вторая линия (независимый валидатор) отделена от разработки, обладает квалификацией не ниже разработчика и имеет полномочия остановить запуск. Три условия обязательны: независимость без квалификации дает формальную подпись, квалификация без полномочий – неисполнимое возражение, полномочия без независимости – конфликт интересов. Российская практика показывает разрыв по третьему условию: правом наложить вето на ключевые решения руководитель риск-функции обладает в 8% организаций, тогда как в 74% его полномочия ограничены должностными обязанностями в рамках анализа рисков [7]. Квалификация и независимость без права остановить запуск процедуры не образуют контроль.

Отсюда следует, что формальное согласование не может считаться валидацией. Если процедура проверки не предусматривает возможности мотивированного возражения, это само по себе является недостатком системы контроля: результатом валидации должно быть задокументированное заключение или возражение, а не подпись о согласии. Из этого же вытекает общее правило: чем выше риск и автономность ИИ-продукта или процесса, переданного ИИ-модели, тем шире должен быть спектр компетенций независимого валидатора.

РОЛИ И КОМПЕТЕНЦИИ: УНИВЕРСАЛЬНАЯ МАТРИЦА И ЕЕ ОТРАСЛЕВЫЕ ЭКВИВАЛЕНТЫ *Табл. 1*

ФУНКЦИЯ (ЛИНИЯ ЗАЩИТЫ)	КЛЮЧЕВАЯ ЗАДАЧА И ТРЕБУЕМЫЕ КОМПЕТЕНЦИИ	ЭКВИВАЛЕНТ В ФИНАНСОВОЙ/ПРОМЫШЛЕННОЙ/КОММЕРЧЕСКОЙ ОРГАНИЗАЦИИ
Создание (первая линия)	Разработка и эксплуатация ИИ-модели, владение риском продукта. Подбор данных, ввод ограничений и определение границ применимости	Разработчик модели и бизнес-владелец / инженер ИИ-решения и технолог – владелец процесса / внедренец ИИ-сервиса и владелец процесса продаж
Независимая проверка (вторая линия)	Критическая проверка допущений и выводов; право остановить запуск. Квалификация специалиста не ниже разработчика, независимость	Независимая валидация / профильный эксперт и служба качества / контроль качества и профильный эксперт
Контроль риска и соответствия (вторая линия)	Осуществление контроля в рамках риск-аппетита, соответствие регуляторным ориентирам. Знание требований и осведомленность о рисках	Риск-менеджмент и комплаенс / промышленная безопасность и внутренний контроль / юрист, комплаенс, защита данных
Контроль качества данных и мониторинг	Проверка качества данных, мониторинг ИИ-модели, контроль дрейфа, деградации и передачи данных вовне	ИТ и владелец данных во всех типах организаций
Независимая оценка системы контроля (третья линия)	Независимая объективная оценка системы контроля над ИИ-моделями. Проведение внутреннего аудита	Внутренний аудит / внутренний контроль предприятия / внутренний аудит или внешний аудитор
Управление и отчетность	Утверждение сфер применения ИИ, назначение ответственных, риск-аппетит. Понимание и формализация ограничений ИИ-моделей, персональная подотчетность	Совет директоров и топ-менеджмент / руководство предприятия / руководство компании

Источник: составлено авторами. Отраслевые эквиваленты приведены в следующем порядке: финансовая организация – промышленное предприятие – коммерческая компания.

Компетенции независимого валидатора ИИ-модели

Спектр компетенций валидатора шире, чем у разработчика, так как специалист, осуществляющий проверку, несет ответственность за пригодность решения ИИ-модели для применения в бизнес-процессе и за возможность обосновать полученный результат перед регулятором,

клиентом или контрагентом. Профиль валидатора не является подмножеством профиля разработчика, включая регуляторную квалификацию, оценку внешних поставщиков и фиксацию обоснования результата. Ниже приведена таблица 2 с предлагаемой систематизацией компетенций независимого валидатора.

КОМПЕТЕНЦИИ НЕЗАВИСИМОГО ВАЛИДАТОРА И СПОСОБЫ ИХ ПОДТВЕРЖДЕНИЯ

Табл. 2

КОМПЕТЕНЦИИ	ПРОВЕРЯЕМЫЙ НАВЫК	ПОДТВЕРЖДЕНИЕ
Предметная экспертиза	- Распознавание неправдоподобного вывода ИИ-модели; - определение границ применимости решений ИИ-модели	- Опыт самостоятельного принятия решений в проверяемом процессе; - результаты выборочных проверок нетиповых случаев
Методологическая и статистическая компетенции	- Проверка допущений, качества и репрезентативность данных, метрик и устойчивости ИИ-модели; - оценка смещений	- Заключения по валидации; - воспроизведение расчетов разработчика на независимой выборке
Информационная безопасность ИИ	- Организация состязательного тестирования; - оценка рисков инъекций в промпт, извлечения обучающих данных и несанкционированного изменения параметров	- Планы и отчеты состязательного тестирования; - участие в разборе инцидентов
Управление внешними ИИ-моделями	- Оценка поставщика, лицензионных условий / версии; - проведение повторных проверок после обновления ИИ-моделей; - оценка концентрационной значимости	- Заключения по внешним моделям; - план миграции и повторной проверки после обновлений
Регуляторная компетенция	Соотнесение ИИ-продукта/ИИ-модели с Кодексом этики, требованиями к модельному и операционному риску, федеральными законами № 152-ФЗ¹ и № 243-ФЗ	- Документированное соотнесение практик с ориентирами; - реагирование на запросы надзорных органов
Независимость суждений	- Сохранение позиции при высокой воспринимаемой надежности ИИ-модели и дефиците сроков; - реализация права остановить запуск; - фиксация письменного обоснования возражения	- Зафиксированные случаи несогласия; - полнота следа проверки в критических решениях

¹ Федеральный закон от 27.07.2006 № 152-ФЗ «О персональных данных» (примечание Редакционной коллегии).
Источник: составлено авторами.

Рассмотренные компетенции могут распределяться между несколькими сотрудниками при условии, что каждая закрепляется персонально, а их носители отделены от функции разработки.

Издержки при отсутствии системы управления риском ИИ-модели

Совокупность наблюдаемых последствий риска внедрения и использования ИИ-моделей позволяет систематизировать издержки, возникающие при отсутствии системы управления данным риском. Возможные издержки предлагаем разделить на несколько категорий:

- **прямые финансовые потери:** убытки по операциям, принятым ИИ-моделью; ухудшение качества активов, устранение инцидента и простой; непропорциональные расходы на ИИ-модель;
- **регуляторные и юридические издержки:** штрафы, судебные издержки, компенсации; невозможность обосновать решение;
- **издержки вынужденного исправления:** замена, переобучение или доработка ИИ-модели; аварийная перестройка контроля; обратный наем / переход на ручной режим;
- **утечка данных и зависимость от поставщика:** передача данных во внешнюю модель; неучтенные теневые модели; потеря контроля над версией и поведением ИИ-модели;
- **репутационные, скрытые и отложенные издержки:** отток клиентов и падение лояльности; утрата экспертизы; ответственность без носителя.

Данный перечень издержек не выявляет новые расходы, а отражает, как затраты на контроль могут перейти в строку с убытками без своевременно выстроенной системы управления рисками внедрения и использования ИИ-моделей. ИИ-модель является инструментом, не отвечающим за результат; ответственным остается тот, кто применил данный инструмент.

Диагностический опрос: методология и результаты

Для проверки центрального тезиса разработан опросник для банков, небанковских финансовых и нефинансовых организаций: блок профиля и 19 вопросов в семи блоках (масштаб применения и делегирование; управление и роли; компетенции; проверка решений; документирование и поставщики; экономика и перенос риска; регуляторная готовность), а также открытый вопрос о барьерах. Опрос проведен в период с июля по август 2026 года, объем выборки не позволяет оценивать распространенность практик и делать статистические выводы, поэтому результаты приводятся как пилотная диагностика, фиксирующая направление. Ответы переведены в шкалу зрелости 0–3 (0 – практика отсутствует либо ее наличие не подтверждено, 1 – фрагментарное применение, 2 – частичная формализация, 3 – системное применение по всем ключевым продуктам). Индекс зрелости рассчитан как среднее по оценочным вопросам (вопросы об уровне автономности и прогнозе кадровой потребности в индекс не входят), варианты «не применимо» исключены из расчета.

Наиболее низкие значения получены по блоку компетенций 0,9 из 3. Регулярное обучение рискам ИИ, включая осведомленность о чрезмерном доверии выводам ИИ-модели, не проводится ни в одной организации выборки, минимальные требования к компетенциям формализованы по ролям у 10% респондентов. Модель трех линий применительно к ИИ-моделям ни у одного респондента не закреплена полностью и не работает на практике; в отдельных организациях имеется частичное закрепление. У части респондентов (более 50%) отсутствует независимая функция валидации, а также не попадают под отдельное управление риски внешних ИИ-моделей и поставщиков. При этом 90% респондентов ожидают рост потребности в специалистах по поддержке и контролю ИИ на горизонте 1–2 лет, а независимая валидация названа наиболее востребованной ролью.

ИНДЕКС ЗРЕЛОСТИ ПО БЛОКАМ ОПРОСНИКА (ШКАЛА 0–3)

Табл. 3

БЛОК ОПРОСНИКА	БАНКИ	НЕФИНАНСОВЫЕ ОРГАНИЗАЦИИ
Понимание, управление и подотчетность	1,67	1,44
Независимая проверка и оспаривание	1,33	0,78
Компетенции и подготовка сотрудников	1,25	0,44
Документирование и поставщики	0,75	1,50
Затраты, перенос риска и регулирование	1,50	1,56

Источник: составлено авторами по данным диагностического опроса [9].

Перераспределение затрат на линейных операциях, компенсированное ростом расходов на контроль, экспертизу и инфраструктуру, зафиксировали около 40% респондентов. Результат согласуется с тезисом в исследовании о смещении, а не устранении издержек.

Заключение

Внедрение ИИ переходит в стадию промышленной эксплуатации, при этом преобладают полуавтономные конфигурации, формально сохраняющие человека в контуре принятия решения. Формальный контроль, не обеспеченный компетенцией специалиста, равнозначен отсутствию контроля. Ожидаемая рынком экономия и издержки контроля распределены во времени. Экономия реализуется в текущем периоде и измерима, однако издержки утраты экспертизы будут проявляться позже и идентифицируются в момент инцидента или запроса надзорного органа.

Список источников

1. Доклад для общественных консультаций [«Применение искусственного интеллекта на финансовом рынке: текущий статус и условия дальнейшего развития»](#). Банк России, 2025 год (дата обращения: 30.09.2026).
2. Информационное письмо Банка России [от 09.07.2025 № ИН-016-13/91](#) «О Кодексе этики в сфере разработки и применения искусственного интеллекта на финансовом рынке» (дата обращения: 30.09.2026).
3. Федеральный закон от 26.07.2026 № 243-ФЗ «О поддержке развития технологий искусственного интеллекта в Российской Федерации» // Официальный интернет-портал правовой информации. 26 июля 2026 года. № опубликования 0001202607260003.
4. Указание Банка России от 15.04.2015 № 3624-У «О требованиях к системе управления рисками и капиталом кредитной организации и банковской группы» // Вестник Банка России. 15 июня 2015 года. № 51 (1647).
5. Положение Банка России от 08.04.2020 № 716-П «О требованиях к системе управления операционным риском в кредитной организации и банковской группе» // Вестник Банка России. 2 июля 2020 года. № 51 (2187).
6. Методические рекомендации Банка России [от 16.06.2026 № 3-МР](#) «По обеспечению информационной безопасности при разработке и применении искусственного интеллекта на финансовом рынке» (дата обращения: 30.09.2026).
7. [Исследование уровня зрелости риск-менеджмента в российских компаниях в 2026 году](#) // Группа ДРТ. 2026 год (дата обращения: 30.09.2026).
8. Анализ текущего состояния внедрения генеративного искусственного интеллекта в страховых компаниях: материалы мастер-класса // Группа ДРТ. 26 августа 2026 года.
9. Экспертиза и управление рисками при создании и проверке продуктов ИИ: данные диагностического опроса (июль – август 2026 года) / команда FRM, Kept. 2026 год. Не опубликованы в открытом доступе.
10. [Artificial Intelligence in UK Financial Services – 2024](#) / Bank of England, Financial Conduct Authority. London, 21 November 2024 (дата обращения: 30.09.2026).
11. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No. 300/2008, (EU) No. 167/2013, (EU) No. 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) // Official Journal of the European Union. 12 July 2024. L 2024/1689.
12. [The Financial Stability Implications of Artificial Intelligence](#) // Financial Stability Board. 14 November 2024 (дата обращения: 30.09.2026).
13. [Artificial Intelligence in Capital Markets: Use Cases, Risks, and Challenges: Consultation Report CR/01/2025](#) // International Organization of Securities Commissions. March 2025 (дата обращения: 30.09.2026).
14. [OpenAI – Hugging Face Incident Technical Report](#) // OpenAI. August 2026 (дата обращения: 30.09.2026).
15. [Incident Report: Unsanctioned Agent Behaviour During Cyber Testing](#) // AI Security Institute. 4 August 2026 (дата обращения: 30.09.2026).

[Ознакомиться с презентацией](#)

ЗАКЛЮЧЕНИЕ

Материалы настоящего выпуска показывают, что управление модельными рисками и рисками, связанными с применением ИИ, предъявляет высокие требования к зрелости управления рисками, внутреннего контроля и внутреннего аудита. Эффективность этих функций проявляется в качестве принимаемых решений, своевременном выявлении ограничений и содержательной реакции на замечания.

Рассмотренные подходы объединяет внимание ко всему процессу использования модели: от исходных данных и методологии до промышленной реализации, мониторинга и управленческого применения. Практика аудита рейтинговых систем показывает, насколько важно оценивать влияние выявленных недостатков на риск и капитал. Для ИИ-систем этот подход дополняется проверкой безопасности, внешних компонентов и полномочий автономных агентов.

Особую роль играет профессиональная экспертиза. Участие человека в проверке результатов должно быть обеспечено знаниями, независимостью суждения и полномочиями для возражения и ограничения применения системы. Экономический эффект технологии следует рассматривать вместе с затратами на контроль, сопровождение и сохранение компетенций сотрудников.

Тематические мастер-классы и публикация материалов создают основу для постоянного обмена опытом в рамках Экспертного совета. Сопоставление подходов разных организаций помогает развивать методологию, уточнять практические вопросы и определять направления совместной работы профессионального сообщества.

Редакционная коллегия надеется, что материалы выпуска будут полезны руководителям организаций, членам советов директоров и комитетов по аудиту, специалистам по внутреннему аудиту, внутреннему контролю и управлению рисками, а также сотрудникам, участвующим в разработке, внедрении и проверке ИИ-систем.

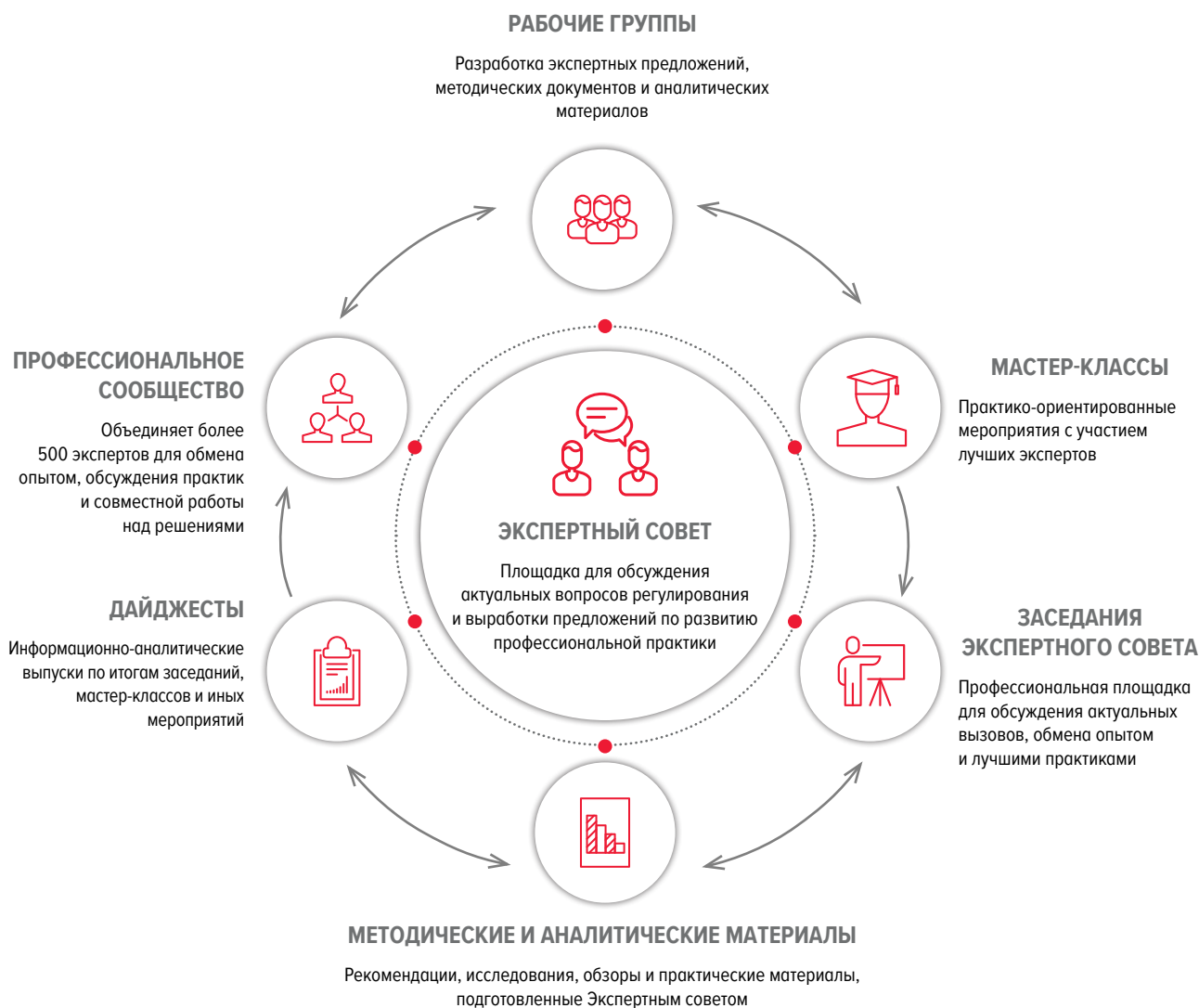
ЭКОСИСТЕМА ЭКСПЕРТНОГО СОВЕТА

При Банке России действует Экспертный совет по регулированию, методологии внутреннего аудита, внутреннего контроля и управления рисками. Он работает на постоянной основе и объединяет представителей Банка России, органов государственной власти, финансовых организаций, компаний реального сектора, научного сообщества и профессиональных объединений.

Задачи Экспертного совета – развивать профессиональное сообщество, готовить методические и аналитические материалы, а также вырабатывать подходы к совершенствованию внутреннего аудита, внутреннего контроля, управления рисками, в том числе модельными рисками и рисками, связанными с применением ИИ.

Ключевые элементы экосистемы

Экспертный совет – это, по сути, площадка, где встречаются регулятор, рынок, бизнес и наука, обсуждаются актуальные вопросы регулирования и формируются предложения по развитию профессии. Здесь рождаются лучшие практики внутреннего аудита, внутреннего контроля и управления рисками, а затем их подхватывает уже все профессиональное сообщество.



Внутри Экспертного совета постоянно функционируют рабочие группы. Именно они разрабатывают конкретные подходы и методологию по отдельным направлениям.

Регулярно проходят мастер-классы: ведущие эксперты разбирают реальные кейсы и делятся практикой применения новых технологий.

Заседания собирают всех участников экосистемы для обсуждения текущей повестки, а по их итогам готовятся методические и аналитические материалы, включающие рекомендации, обзоры, исследования.

О том, что происходит в Экспертном совете, можно узнавать из [дайджестов](#), в которых собраны итоги заседаний, мастер-классов и главные профессиональные новости.

Профессиональное сообщество

Сегодня к сообществу присоединились уже более 500 экспертов из Банка России, госорганов, финансового и реального секторов, науки, профессиональных объединений, аудиторских и консалтинговых компаний.

Основой сообщества являются руководители и специалисты по внутреннему аудиту, внутреннему контролю, управлению рисками, комплаенсу и цифровой трансформации. Они обмениваются практиками, участвуют в рабочих группах и мастер-классах, предлагают идеи по развитию методологии.



Официальные ресурсы

Информацию о деятельности Экспертного совета можно найти на его [официальной странице](#) на сайте Банка России. Там публикуются положение Экспертного совета, анонсы и итоги заседаний, а также методические материалы – этот ресурс открыт для всех.

Для сотрудников Банка России есть еще один канал: внутренний информационный портал – сообщество в [интранете](#). Здесь собраны материалы заседаний и мастер-классов, документы рабочих групп и архив публикаций.

Как присоединиться к сообществу

Проще всего через МАКС или Телеграм – по ссылкам или с помощью QR-кодов:

[МАКС](#)



[Телеграм](#)



Контакты

Если вы хотите принять участие в работе Экспертного совета, рабочих групп или предложить тему для мастер-класса, напишите нам: expert.board@mail.cbr.ru.

ГЛОССАРИЙ

Аварийная остановка ИИ-системы – механизм прекращения работы системы или выполнения ею действий при инциденте либо недопустимом уровне риска; условия запуска и последствия остановки определяются заранее.

Аудит моделей и ИИ-систем – независимая оценка процессов управления и контроля, связанных с разработкой, внедрением и использованием моделей и ИИ. Охватывает данные, методологию, техническую реализацию, применение результатов и распределение ответственности.

Валидация модели – независимая содержательная проверка обоснованности модели, качества ее результатов, ограничений и пригодности для предусмотренного применения. Отличается от текущего мониторинга и оценки системы контроля внутренним аудитом.

Владелец модели – лицо или подразделение, ответственное за применение модели в конкретном процессе, ее сопровождение и решения о дальнейшем использовании с учетом результатов контроля.

Воспроизводимость – возможность повторить расчет или проверить ход получения результата по сохраненным данным, коду, параметрам и версиям компонентов. Для недетерминированных ИИ-систем восстановление цепочки событий не обязательно означает дословное совпадение повторного ответа.

Галлюцинация модели – правдоподобно сформулированный ответ генеративной модели, содержащий вымышленные, ошибочные или не подтвержденные используемыми источниками сведения.

Генеративный искусственный интеллект – модели и системы, формирующие новый текст, программный код, изображения и другое содержание на основе закономерностей, освоенных при обучении, и предоставленного контекста.

Дрейф данных (data drift) – изменение распределения входных данных относительно данных, на которых модель разрабатывалась или ранее применялась. Может указывать на необходимость дополнительной проверки качества модели.

Жизненный цикл модели – этапы от постановки задачи, подготовки данных и разработки до проверки, внедрения, эксплуатации, изменения и вывода модели из использования.

Зрелость управления рисками – степень развития и практического применения подходов к управлению рисками, включая участие руководства, полномочия, ресурсы, риск-культуру и связь с принятием решений.

Избыточная агентность – предоставление ИИ-агенту функций, прав доступа или автономности сверх необходимых для его задачи, увеличивающее возможные последствия ошибки или атаки.

ИИ-агент – программная система на основе ИИ, которая выбирает последовательность действий для достижения заданной цели и может использовать внешние инструменты в пределах предоставленных полномочий.

ИИ-риск – риск неблагоприятных последствий разработки, внедрения или использования ИИ. Может проявляться через модельные, операционные, информационные, правовые,

репутационные и иные риски; его место в классификации определяется принятой системой управления рисками.

ИИ-шлюз – управляемая точка доступа к ИИ-сервисам, через которую могут осуществляться маршрутизация запросов, применение политики, контроль данных, лимитов и журналирование.

Искусственный интеллект – совокупность технологических решений, выполняющих задачи, связанные с обработкой информации, распознаванием закономерностей, формированием содержания, рекомендаций или решений и обычно требующих человеческих когнитивных способностей.

Калибровка модели кредитного риска – настройка количественных оценок риска в соответствии с назначением модели и выбранным ориентиром, например долгосрочным уровнем дефолтности.

Консервативная надбавка – корректировка оценки кредитного риска для учета факторов возможной недооценки, связанных с недостатками данных, методологии или изменениями условий. Основание и величина надбавки требуют обоснования.

Модель трех линий – распределение ролей, при котором первая линия управляет деятельностью и связанными рисками, вторая обеспечивает профильную экспертизу, поддержку, мониторинг и критическую оценку управления рисками, а внутренний аудит предоставляет независимую оценку управления и контроля. В редакции 2013 года и в ряде статей выпуска используется прежнее наименование «три линии защиты».

Модель угроз – описание защищаемых объектов, возможных нарушителей, способов воздействия, уязвимостей и последствий, используемое для выбора мер защиты конкретной системы.

Модель – формализованный метод, система или подход, преобразующий исходные данные в оценки, прогнозы либо иные результаты. При управлении риском рассматривается вместе с данными, допущениями и технической реализацией.

Модельный риск – риск неблагоприятных последствий решений или действий, основанных на результатах модели, из-за ее недостатков, ошибок данных, реализации, интерпретации либо применения.

Мониторинг модели – регулярное наблюдение за качеством данных, результатами, стабильностью и работоспособностью модели с установленными порогами и порядком реагирования на отклонения.

Область применимости модели – совокупность задач, объектов и условий, для которых модель разработана и проверена. Использование за ее пределами требует отдельного обоснования и оценки риска.

Объяснимость – возможность представить понятное человеку объяснение результата модели и существенных факторов, повлиявших на него. Объяснение само по себе не подтверждает правильность результата.

Отравление данных (data poisoning) – намеренное изменение данных обучения, поиска или иного информационного контекста для искажения поведения модели или системы.

Подход на основе внутренних рейтингов (ПВР) – подход к оценке кредитного риска с использованием внутренних рейтингов и оценок его компонентов, полученных на основе банковских данных и моделей.

Принцип «четырёх глаз» – организация проверки значимого решения вторым компетентным лицом, позволяющая выявить ошибку и выразить независимое возражение до исполнения решения.

Принцип минимальных привилегий – предоставление пользователю, сервису или ИИ-агенту только тех прав и на тот срок, которые необходимы для выполнения конкретной задачи.

Промпт-инъекция – внедрение инструкций в запрос или внешний контент с целью изменить поведение ИИ-системы, обойти ограничения либо вызвать непредусмотренное действие.

Прослеживаемость данных (data lineage) – документированная связь данных с первичными источниками, этапами преобразования и конечным использованием, позволяющая установить их происхождение и проверить обработку.

Реестр моделей и ИИ-систем – систематизированный учет моделей, систем и сценариев использования с указанием владельцев, назначения, значимости, статуса и сведений о проверках и сопровождении.

Рейтинговая система – совокупность методов, процессов, данных, контролей и информационных систем, обеспечивающих присвоение внутренних рейтингов и количественную оценку компонентов кредитного риска.

Риск-аппетит – уровень и виды риска, которые организация готова принимать при достижении своих целей; служит ориентиром для решений и ограничений.

Состязательное тестирование (red teaming) – проверка системы посредством моделирования действий нарушителя и сложных сценариев, направленная на выявление уязвимостей и оценку работоспособности защитных мер.

Теневой ИИ – использование сотрудниками или подразделениями ИИ-инструментов и интеграций, не учтенных либо не согласованных в установленном организацией порядке.

Трассировка действий – фиксация связанной последовательности событий: запроса, использованного контекста, вызовов инструментов, проверок разрешений, действий, результатов и подтверждений.

Утрата экспертизы – снижение способности сотрудников самостоятельно решать профессиональную задачу и критически оценивать результаты автоматизированной системы вследствие ослабления практики, обучения или кадрового обеспечения.

«Человек в контуре» (Human-in-the-Loop) – организация процесса, при которой человек участвует в проверке или подтверждении результата либо действия ИИ. Мера контроля требует определенных критериев вмешательства, компетенций и полномочий.

Экспертная корректировка рейтинга – изменение модельного рейтинга на основании профессионального суждения и дополнительной информации. Требуется фиксации причин и проверки последствий применения.

СПИСОК СОКРАЩЕНИЙ

Общие аббревиатуры и термины

ИБ – информационная безопасность

ИИ – искусственный интеллект

ИТ – информационные технологии

AI – Artificial Intelligence – искусственный интеллект

API – Application Programming Interface – программный интерфейс взаимодействия систем

GenAI – Generative Artificial Intelligence – генеративный искусственный интеллект

GLM – Generative Language Model – генеративная языковая модель; обозначение используется в статье НКЦ

GPU – Graphics Processing Unit – графический процессор

HR – Human Resources – управление персоналом

LLM – Large Language Model – большая языковая модель

LoRA – Low-Rank Adaptation – метод адаптации модели с использованием матриц низкого ранга

ML – Machine Learning – машинное обучение

PAI – Personal AI Infrastructure – персональная инфраструктура ИИ; обозначение проекта, рассматриваемого в статье Р.М. Гусейнова

RAG – Retrieval-Augmented Generation – генерация с дополнением контекста сведениями, найденными во внешних источниках

SaaS – Software as a Service – программное обеспечение как услуга

SDK – Software Development Kit – комплект средств разработки программного обеспечения

Сокращения в области корпоративного управления, аудита и контроля

ПБР – подход на основе внутренних рейтингов

EAD – Exposure at Default – величина кредитного требования на момент дефолта

ECL – Expected Credit Losses – ожидаемые кредитные убытки

FRM – Financial Risk Management – управление финансовыми рисками; в статье Kept используется в названии профильной команды

HITL – Human-in-the-Loop – участие человека в проверке или подтверждении результата либо действия ИИ

IRB – Internal Ratings-Based – подход на основе внутренних рейтингов

KPI – Key Performance Indicator – ключевой показатель эффективности

LGD – Loss Given Default – уровень потерь в случае дефолта

PD – Probability of Default – вероятность дефолта

PSI – Population Stability Index – индекс стабильности распределения

RWA – Risk-Weighted Assets – активы, взвешенные по риску

SHAP – SHapley Additive exPlanations – метод объяснения предсказаний модели на основе значений Шепли

Сокращения в области информационной безопасности

AIBOM/AI-BOM – AI Bill of Materials – перечень компонентов ИИ-системы с информацией об их составе и происхождении

CASB – Cloud Access Security Broker – средство контроля доступа к облачным сервисам

DLP – Data Loss Prevention – предотвращение утечек данных

IAM – Identity and Access Management – управление идентификацией и доступом

MCP – Model Context Protocol – протокол подключения ИИ-приложений к внешним данным и инструментам

MFA – Multi-Factor Authentication – многофакторная аутентификация

NHI – Non-Human Identity – машинная идентичность, используемая сервисом, приложением или агентом

SBOM – Software Bill of Materials – перечень программных компонентов

SIEM – Security Information and Event Management – управление информацией и событиями информационной безопасности

SOC – Security Operations Center – центр мониторинга и реагирования на инциденты информационной безопасности

SSH – Secure Shell – протокол защищенного удаленного доступа

SSO – Single Sign-On – единый вход в информационные системы

SWG – Secure Web Gateway – шлюз безопасного доступа к веб-ресурсам

VPN – Virtual Private Network – виртуальная частная сеть

Организации, стандарты и методологии

ГОСТ – государственный стандарт

ДРТ – «Деловые Решения и Технологии»

ЕС – Европейский союз

МСФО – Международные стандарты финансовой отчетности

НКО – небанковская кредитная организация

НКЦ – Национальный Клиринговый Центр

РАД КОП (RAD COP, Rapid Assessment Delivery Cooperative) – кооператив по оказанию услуг быстрой оценки в области кибербезопасности

СНГ – Содружество Независимых Государств

ФСТЭК России – Федеральная служба по техническому и экспортному контролю

AISI – Artificial Intelligence Security Institute – Институт безопасности искусственного интеллекта (Великобритания)

ATLAS – Adversarial Threat Landscape for Artificial-Intelligence Systems – база знаний MITRE об угрозах для ИИ-систем

COSO – Committee of Sponsoring Organizations of the Treadway Commission – Комитет организаций-спонсоров Комиссии Тредвея

FCA – Financial Conduct Authority – Управление по финансовому регулированию и надзору Великобритании

FSB – Financial Stability Board – Совет по финансовой стабильности

IEC – International Electrotechnical Commission – Международная электротехническая комиссия

IOSCO – International Organization of Securities Commissions – Международная организация комиссий по ценным бумагам

ISO – International Organization for Standardization – Международная организация по стандартизации

NIST – National Institute of Standards and Technology – Национальный институт стандартов и технологий США

NIST AI RMF – NIST Artificial Intelligence Risk Management Framework – рамочная модель NIST по управлению рисками ИИ

OWASP – Open Worldwide Application Security Project – открытый проект по безопасности приложений

PCI SSC – Payment Card Industry Security Standards Council – Совет по стандартам безопасности индустрии платежных карт

Экономическая классификация

JEL – Journal of Economic Literature – система классификации экономических публикаций