

СБОРНИК ЛУЧШИХ ПРАКТИК

Примеры соблюдения
Кодекса этики в сфере
разработки и применения
искусственного интеллекта
на финансовом рынке



Оглавление

О Сборнике	2
Система управления рисками	4
Виртуальные ассистенты	61
Финансовые помощники в инвестициях	76
Скоринг	90
Анализ рисков операций	100
Финансовая доступность	115
Внутренние сервисы	125
Приложение. Кодекс этики в сфере разработки и применения искусственного интеллекта на финансовом рынке	132

О Сборнике

В 2025 году Банк России опубликовал Кодекс этики в сфере разработки и применения искусственного интеллекта на финансовом рынке¹ (Кодекс этики, ИИ). Он закладывает этические основы ответственного внедрения ИИ-технологий в финансовой отрасли, развивая положения общеотраслевого Кодекса этики в сфере ИИ. Цель Кодекса этики – повышение доверия к ИИ на финансовом рынке, продвижение ответственных практик внедрения технологии и минимизация связанных с ней рисков.

В Кодексе этики Банк России рекомендует финансовым организациям следовать пяти принципам этичного внедрения ИИ:

- человекоцентричность;
- справедливость;
- прозрачность;
- безопасность, надежность и эффективность;
- ответственное управление рисками.

Для каждого из принципов Кодекс этики содержит рекомендации по их соблюдению.

В 2026 году Банком России совместно с Ассоциацией «Альянс в сфере искусственного интеллекта» и Ассоциацией развития финансовых технологий была создана рабочая группа по этике в сфере ИИ на финансовом рынке. Ее первой задачей стало формирование настоящего Сборника лучших практик соблюдения Кодекса этики (Сборник).

Сборник направлен на популяризацию Кодекса этики и эффективных подходов к ответственному внедрению ИИ. Сборник предлагает участникам финансового рынка конкретные способы соблюдения рекомендаций Кодекса этики.

¹ Информационное письмо Банка России от 09.07.2025 № ИН-016-13/91.

В Сборнике рассмотрено, как применяющие ИИ финансовые организации адаптируют подходы к управлению рисками (раздел «Система управления рисками»), а также учитывают риски ИИ в отдельных сервисах и бизнес-процессах (разделы «Виртуальные ассистенты», «Финансовые помощники в инвестициях», «Скоринг», «Анализ рисков операций», «Финансовая доступность» и «Внутренние сервисы»).

Каждая глава Сборника состоит из двух частей:

- аналитическая часть – обобщает лучшие этические практики и предлагает способы соблюдения Кодекса этики;
- практическая часть – содержит примеры соблюдения Кодекса этики в сервисах и бизнес-процессах организаций.

В аналитической и практической частях содержатся отсылки к Кодексу этики, полный текст которого вместе с указанием на релевантные практики организаций находится в приложении.

В будущем Сборник будет обновляться и дополняться с учетом актуальных практик соблюдения Кодекса этики.

Практики организаций, включенные в Сборник, демонстрируют возможные способы соблюдения Кодекса этики и не являются подтверждением эффективности или качества упомянутых сервисов или бизнес-процессов.

Сервисы и бизнес-процессы организаций рассматривались на соответствие только прямо упомянутым в соответствующих разделах рекомендациям.



Система управления рисками

ИИ открывает финансовым организациям новые возможности для создания клиентских сервисов, повышения эффективности и качества финансовых услуг. Вместе с тем на разных этапах жизненного цикла моделей ИИ могут возникать дополнительные риски, которые требуют ответственного подхода к внедрению ИИ. И сегодня организации финансового рынка выстраивают комплексные системы управления рисками ИИ, которые позволяют эффективно использовать ИИ и обеспечивают доверие к результатам его работы

Важно учитывать

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ СОЗДАЕТ НОВЫЕ ВЫЗОВЫ ДЛЯ УПРАВЛЕНИЯ РИСКАМИ В ОРГАНИЗАЦИЯХ

Свойства ИИ значительно отличаются от свойств традиционного ПО: модели ИИ сильнее зависят от качества и характера данных, могут быть сложно интерпретируемыми, высоко автономными и чувствительными к изменению внешних условий.

Кроме того, некоторые риски ИИ не могут быть в полной мере выявлены до начала эксплуатации и проявляются только в реальных условиях применения, например, когда ИИ сталкивается с новыми данными или сценариями применения. Поэтому для моделей ИИ недостаточно разовой проверки качества в момент разработки и валидации: даже тщательно протестированная модель со временем может начать давать менее точные и интерпретируемые результаты по мере изменения данных, клиентского поведения и рыночных условий.

Также все активнее в деятельность финансовых организаций внедряются более сложные типы ИИ – генеративный ИИ и ИИ-агенты. Их использование может влиять на бизнес-процессы сильнее других решений и создавать дополнительные риски, в том числе связанные с галлюцинациями и автономностью.

РИСКИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ЗАВИСЯТ ОТ КОНТЕКСТА ПРИМЕНЕНИЯ

Риски ИИ зависят не только от технических характеристик самой модели ИИ, но и от контекста применения, в том числе процессов и функций, в которых внедряется ИИ. Риск-профиль одной и той же модели ИИ может значительно различаться в зависимости от ряда факторов, например, от характера используемых данных, степени автономии и влияния на человека. Для финансового рынка учет этих обстоятельств особенно важен, поскольку модели ИИ все чаще используются в процессах, связанных с клиентом, принятием решений, мониторингом операций и иными чувствительными функциями.

Кроме того, внедрение ИИ часто связано с использованием моделей, данных и инфраструктуры от третьих лиц, когда часть значимых риск-факторов находится за пределами непосредственного контроля организации. Это может повышать уровень риска ИИ-решений и затруднять их оценку и мониторинг.

Лучшие практики

Кодекс этики предлагает финансовым организациям широкий перечень рекомендаций по ответственному управлению рисками. На практике следует учитывать вид и масштаб деятельности организации, ее организационной структуры, существующей системы управления рисками (СУР) и характера оказываемых услуг.

ОРГАНИЗАЦИЯ СИСТЕМЫ УПРАВЛЕНИЯ РИСКАМИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Интеграция правил управления рисками искусственного интеллекта в общую систему управления рисками

ИИ привносит специфику в те риски, которые уже знакомы финансовым организациям, включая модельный риск, риск информационной безопасности, регуляторный риск и так далее. Поэтому особенности их проявления, связанные с ИИ, целесообразно учитывать в текущей СУР организации с опорой на существующие процессы.

Правила управления рисками искусственного интеллекта

В основе эффективного управления рисками ИИ должны лежать формализованные правила, которые определяют, как эти риски выявляются, оцениваются, документируются и контролируются.

Форма закрепления таких правил может быть разной. В зависимости от масштаба и характера применения ИИ в организации они могут быть закреплены как в отдельном документе об управлении рисками ИИ, так и в общих документах, регламентирующих управление рисками в организации в целом.

Политика управления рисками использования моделей ИИ п. 6.2

Пример утверждения правил управления рисками ИИ
в виде отдельного документа

Возможен и гибридный подход, при котором базовые требования закрепляются в общих правилах управления рисками, а дополнительные разъяснения, процедуры и методические материалы выносятся в отдельный документ. В таком случае важно обеспечить согласованность документов между собой, чтобы в организации не действовали дублирующие или противоречащие друг другу требования.



Степень детализации правил управления рисками ИИ может отличаться.

- если ИИ используется ограниченно, может быть достаточно точечной адаптации действующих документов;
- при широком применении ИИ или использовании более сложных и внешних моделей ИИ целесообразно оценить потребность в более детализированных и комплексных правилах;
- для генеративного ИИ и ИИ-агентов могут предусматриваться специальные правила и процедуры

Система управления рисками агентского ИИ п. 6.1-6.2

Пример учета специфики ИИ-агентов в СУР

Правила управления рисками ИИ могут включать следующие положения:

- указание, на какие модели и случаи применения ИИ распространяются правила;
- распределение ролей и ответственности внутри организации;
- порядок ведения реестра моделей ИИ и поддержания его актуальности;
- критерии оценки уровня риска моделей ИИ;
- требования к обоснованию целесообразности использования ИИ для отдельных процессов;
- требования к качеству данных;
- правила документирования информации о процессах на всех этапах жизненного цикла ИИ;
- порядок тестирования и валидации моделей ИИ до и после ввода в эксплуатацию;
- порядок введения моделей ИИ в эксплуатацию и требования к последующему мониторингу;
- метрики, используемые при проверке качества данных, тестировании и валидации моделей ИИ и на других этапах жизненного цикла;
- правила управления изменениями (критерии существенных изменений, порядок их согласования, повторной проверки, версионирования, возврата к предыдущим версиям);
- требования к человеческому надзору, включая случаи, когда участие человека обязательно;
- специальные правила для генеративного ИИ и ИИ-агентов (ограничения на сценарии использования, дополнительные правила тестирования и прочее);
- правила работы с внешними поставщиками;
- правила реагирования на инциденты и вывода моделей ИИ из эксплуатации;
- требования к обучению и повышению квалификации работников, участвующих в разработке, использовании и управлении рисками ИИ.

При формировании правил управления рисками ИИ, кроме действующих нормативных актов, следует учитывать, в частности, Методические рекомендации Банка России по обеспечению информационной безопасности при разработке и применении ИИ на финансовом рынке, стандарты ISO/IEC 23894, ПНСТ 776-2022, а также ГОСТ Р 71476-2024, ГОСТ Р 71484.3-2024, ГОСТ Р 71539-2024, ГОСТ Р ИСО/МЭК 42001-2024.

Роли и функции при управлении рисками искусственного интеллекта

Для эффективного управления рисками ИИ важно четко определить круг вовлеченных субъектов (ответственных подразделений и сотрудников), их роли и функции. В результате должна сформироваться целостная система, в рамках которой для всех стадий жизненного цикла ИИ фиксируются полномочия и обязанности участников, а также механизмы эскалации.

В рамках этой системы целесообразно определить, например, кто инициирует разработку и использует модель ИИ, кто проводит независимую проверку и принимает решение о внедрении, кто осуществляет мониторинг, согласует изменения и может ограничить использование ИИ в случае инцидентов и так далее.

Разделение ролей и функций можно рассматривать на трех уровнях:

- **На уровне управления организацией** утверждаются внутренние политики и рамочные подходы, определяется риск-аппетит и осуществляется общий надзор за СУР ИИ. Органы управления организацией также должны обеспечивать функционирование СУР ИИ: организовывать процессы, распределять ресурсы, контролировать соблюдение правил управления рисками на уровне структурных подразделений и крупных бизнес-процессов и так далее. Все решения должны приниматься с учетом наиболее актуальной информации о рисках, инцидентах и ограничениях применения ИИ;
- **На уровне структурных подразделений** СУР целесообразно выстраивать по отдельным функциям («линиям защиты»). Одним из возможных подходов является разделение процессов разработки, использования, независимой проверки и внутреннего контроля, что обеспечивает объективную оценку и управление рисками. Так, первая линия защиты может отвечать за создание, применение и повседневный мониторинг ИИ (владелец продуктов и процессов), вторая – за независимое управление рисками ИИ, включая валидацию и пересмотр уровня риска (подразделения валидации, комплаенса, информационной безопасности и пр.), а третья – за независимую оценку полноты и эффективности всей СУР (внутренний аудит);
- **На уровне отдельных моделей и случаев применения ИИ** целесообразно разделять роли и функционал отдельных участников, например владельца ИИ-решения, разработчика, пользователя и валидатора. Владелец отвечает за соответствие модели ее назначению, внутренним требованиям условиям применения, а также за актуальность сведений о ней и прохождение необходимых процедур проверки. Разработчик отвечает за проектирование, построение, тестирование и документирование, а пользователь – за применение ИИ по назначению с учетом известных ограничений. В свою очередь, валидатор должен обеспечивать объективную оценку и подтверждать корректность применения ИИ. Целесообразно организационно отделять валидатора от процессов разработки и эксплуатации (независимая валидация).

Роли и функции участников могут детализироваться в зависимости от содержания правил управления рисками ИИ и особенностей случая применения ИИ.

Возможное распределение ролей и функционала на уровне отдельных ИИ-решений и процессов

Роль	Функционал
Заказчик	Является тем, в чьих интересах разрабатывается модель. Отвечает за конечный результат
Методолог / Куратор	Является центром компетенций, обладающим полным пониманием методологии процесса, использующего модель
Поставщик данных	Обеспечивает процесс необходимыми данными
Разработчик модели	Разрабатывает модель, используя статистические и математические методы
Валидатор	Обеспечивает независимую валидацию
ИТ-архитектор	Отвечает за архитектурное решение по имплементации модели в ИТ-системы
ИТ-разработчик	Разрабатывает и внедряет автоматизированное промышленное решение

При распределении ролей и функций важно предусматривать кросс-функциональные механизмы координации между ответственными участниками и структурными подразделениями. Это позволит при необходимости оценивать риски ИИ и управлять ими с учетом позиций подразделений по бизнес- и системной аналитике, информационной безопасности, комплаенсу и так далее.

Политика управления
рисками использования моделей ИИ

п. 6.2

Пример определения лиц, ответственных за реализацию мер по управления рисками ИИ

Реестр моделей искусственного интеллекта

Ключевым элементом СУР ИИ является документирование информации на всех этапах жизненного цикла. Такое документирование возможно осуществлять через специальный реестр моделей ИИ, который обеспечивает целостное представление о том, как и для каких целей ИИ используется в организации, в каком статусе находятся модели ИИ и какие риски с ними связаны. Реестр может содержать не всю документируемую информацию о модели ИИ, но при необходимости позволять определить и найти ее оригинальный источник.



При масштабном внедрении ИИ в организации реестр моделей перестает быть вспомогательным инструментом и становится важной частью СУР ИИ. Целесообразно вести его централизованно, регулярно обновлять и встраивать в процедуры на разных этапах жизненного цикла моделей ИИ

Реестр может содержать разную информацию о моделях ИИ, включая в том числе следующие сведения:

- **Общие сведения о модели ИИ:**
 - внутренний идентификатор, наименование, версия, тип модели или архитектуры;
 - связанный бизнес-процесс, приложение, система, внутренняя или внешняя среда использования;
 - назначение и границы допустимого использования;
 - текущий статус (проверка концепции, в разработке, в эксплуатации, эксплуатация приостановлена, выведена из эксплуатации), версия;
 - владелец модели ИИ или процесса, разработчик, пользователь, валидатор и иные участники;
 - типы, источники, владельцы используемых данных, информация об использовании чувствительных данных;
 - связанные системы, каналы интеграции, используемые внешние сервисы и критические компоненты.

Реестр может использоваться как инструмент сопровождения модели ИИ по жизненному циклу. Он позволяет устанавливать контрольные точки и видеть, на каком этапе находится конкретная модель. Также важно включать в реестр модели ИИ, находящиеся в разработке или выведенные из эксплуатации. Информация о них может повысить качество управления рискам в текущих процессах, в том числе учитывать возможные остаточные риски, и использоваться при валидации других моделей.

- **Риск-профиль и контрольные меры:**
 - уровень и значимые факторы риска, ключевые риски и сценарии отказа, возможные остаточные риски;
 - применяемые меры для снижения уровня риска;
 - проведенные тесты и проверки, используемые тестовые наборы данных, процедуры валидации, бенчмарки и полученные результаты, ссылки на результаты;
 - подразделение-валидатор, периодичность валидации.
- **Сведения о внешних поставщиках:**
 - данные о поставщике, тип лицензии, версия решения и уведомления об изменениях, каналы интеграции, значимые условия использования, иная значимая информация от поставщика.

- **Мониторинг, инциденты и реагирование:**

- периодичность мониторинга, ответственные лица, перечень отслеживаемых метрик и факторы эскалации;
- выявленные отклонения от метрик, инциденты, механизмы и результаты реагирования на инциденты.



Основной единицей реестра может быть случай применения ИИ, а не модель ИИ. Это позволит учитывать при управлении рисками ИИ особенности бизнес-процесса и может быть целесообразно, например, при использовании одной и той же модели в разных функциях



Стандарт управления жизненным циклом моделей ИИ

п. 6.4 [↗](#)



Порядок управления модельным риском

п. 6.4 [↗](#)

Примеры ведения реестра моделей ИИ

По мере развития генеративного и агентского ИИ требования к содержанию реестра могут становиться более детализированными.

- **Для генеративного ИИ** реестр можно дополнять сведениями о системных промптах и их версиях, конфигурации модели, подключенных источниках данных (в том числе через RAG), инструментах повышения объяснимости, правилах логирования и так далее.
- **Для ИИ-агентов** реестр можно также дополнять сведениями о функциях, которые может вызывать агент, предоставленных доступах и их уровне, правилах оркестрации и эскалации, правилах логирования действий агента и так далее.

Реестр может быть не изолированным инструментом, а частью более широкой системы учета. Его целесообразно связывать с другими реестрами, например с данными об ИТ-активах, источниках данных, внешних поставщиках, контрольных процедурах и инцидентах.

УПРАВЛЕНИЕ РИСКАМИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА
НА ЭТАПАХ ЖИЗНЕННОГО ЦИКЛА

Управление рисками ИИ может быть удобно выстраивать в рамках отдельных этапов жизненного цикла ИИ. Такие этапы могут выделяться в организации с учетом особенностей внутренних процессов и сложившегося подхода и включать, например, следующие этапы:



 **Фреймворк управления рисками ИИ** п. 6.8 [↗](#)

Пример интеграции управления рисками ИИ в жизненный цикл систем ИИ

 **Модель угроз кибербезопасности ИИ** п. 5.5 [↗](#)

Пример управления рисками информационной безопасности на всех этапах жизненного цикла ИИ

Инициирование разработки

На данном этапе важно установить потребность организации в использовании ИИ в конкретном бизнес-процессе, оценить соответствие планируемой к применению модели ИИ целям и риск-аппетиту организации и определить условия и границы применения ИИ.

На практике этап инициирования разработки модели ИИ может включать, в частности, следующие элементы:

- **Определение целей внедрения ИИ и релевантных процессов**
Важно формализовать, какие задачи предполагается решать с помощью ИИ, какие бизнес-процессы и функции затрагиваются и кто является владельцем модели ИИ. Целесообразно однозначно установить, в каких процессах и продуктах может использоваться модель, и по возможности зафиксировать заведомо недопустимые сценарии или условия ее использования. Необходимо учитывать требования действующего законодательства и действующие в организации правила управления рисками.



Уже на этапе инициирования разработки важно документировать всю значимую информацию и вносить ее в реестр моделей ИИ организации, чтобы упростить управление рисками в дальнейшем

 **Стандарт управления жизненным циклом моделей ИИ** п. 5.3 

Пример бизнес-требований для инициирования разработки ИИ

- **Предварительная оценка рисков**
Для каждой модели ИИ целесообразно определить уровень риска, а также соответствующие процедуры проверки, согласования и контроля на последующих этапах жизненного цикла. При такой оценке нужно принимать во внимание факторы риска ИИ, которые определяются организацией во внутренних правилах управления рисками ИИ. Возможные факторы риска также закреплены в Кодексе этики.

	Фреймворк управления рисками ИИ	п. 6.6-6.7	
	Политика управления рисками использования моделей ИИ	п. 6.6-6.7	
	Порядок управления модельным риском	п. 6.6	

Пример оценки и присвоения уровня риска ИИ

 **Система управления рисками агентского ИИ** п. 6.6 

Пример оценки и присвоения уровня риска ИИ-агентов



В отношении значимых бизнес-процессов и решений с высоким уровнем риска целесообразно оценить, можно ли решить затрагиваемую задачу без применения ИИ, если его внедрение будет создавать необоснованные риски

Обеспечение данными

На данном этапе организация определяет, какие данные необходимы для работы модели ИИ, из каких источников они поступают, в каком виде будут использоваться и какие требования к их качеству и безопасности должны соблюдаться. На практике этап обеспечения данными может включать, в частности, следующие элементы:

- **Определение источников и состава данных**

Целесообразно формально определить, какие типы данных будут использоваться (структурированные, неструктурированные, транзакционные, поведенческие, синтетические и пр.) и кто является их владельцем. Одновременно важно оценивать необходимость использования персональных данных, коммерческой тайны и иной конфиденциальной информации.

- **Обеспечение качества, целостности и происхождения данных**

Целесообразно установить требования к точности, полноте, актуальности и непротиворечивости данных, а также внедрить процедуры контроля качества, включая проверки на аномалии, смещения и дублирование. Также целесообразно обеспечить прослеживаемость происхождения данных, включая перечень допустимых источников и контроль модификаций.

Для повышения качества управления рисками в будущем важно документировать процессы сбора, разметки, обработки данных, формирования признаков, а также ключевые решения по обработке и трансформации данных, включая применяемые правила фильтрации, методы обезличивания и подходы к балансировке выборок.



Необходимый набор мер в каждом конкретном случае зависит от уровня риска модели ИИ и типа данных. Например, вопросам качества важно уделять особое внимание при использовании данных от внешних поставщиков, а также в случаях применения ИИ с высоким риском



Построение жизненных циклов моделей ИИ
в системах принятия решений по страхованию

п. 5.2

Пример подхода к проверке качества данных

- **Организация доступа к данным и мер защиты**

Целесообразно выстроить систему разграничения доступа к данным на основе ролевой модели и следовать принципу минимально необходимых прав. Важно предусмотреть технические и организационные меры защиты от утечек и атак.



Модель угроз кибербезопасности ИИ

п. 5.6

Примеры подхода к обеспечению
конфиденциальности данных



Внедрение агентского ИИ может требовать реализации дополнительных мер управления рисками, связанными с данными, которые позволяют оперативно реагировать на возможные инциденты. Например, важно иметь возможность оперативно проверять доступные ИИ-агентам механизмы доступа к данным и их использованию, а также с учетом поставленных целей минимизировать доступ к данным, особенно к чувствительной информации

Разработка и валидация

На данном этапе организация либо разрабатывает собственную модель ИИ, либо интегрирует готовую модель (третьих лиц или из открытого доступа), в том числе с дообучением, проверяет ее работу в предполагаемом контексте применения и подтверждает, что она может быть допущена к дальнейшему внедрению с учетом установленного уровня риска.

Разработка

Архитектуру модели ИИ важно выстраивать с учетом сценария ее применения и выбранного способа реализации (собственной разработки, дообучения и прочее). Уже при ее проектировании нужно закладывать базовые механизмы контроля входных и выходных данных, использовать доверенные компоненты, определять границы доверия между моделью, внешними источниками и расширяемой функциональностью. Для разработки моделей ИИ могут создаваться отдельные среды.



Важно документировать состав и взаимодействие всех компонентов, включая модель, данные, инфраструктуру хранения и обслуживания модели, пользовательские интерфейсы. В случае с ИИ-агентами целесообразно также учитывать, к каким системам и каким способом они будут получать доступ, в том числе в мультиагентских системах

При разработке важно обеспечивать воспроизводимость и управляемость процесса: отслеживать изменения, поддерживать контроль версий моделей и кода.

Если организация использует модели ИИ от третьих лиц, целесообразно, в частности:

- дополнять процедуры проверки контрагента проверкой самой модели, запрашивая соответствующую документацию (об используемых данных, ограничениях, уровне объяснимости и так далее);
- включать в договор условия, которые позволят осуществлять эффективный мониторинг при эксплуатации модели, в том числе получать необходимую информацию об используемых данных;
- предусматривать дополнительные меры надзора за эксплуатацией внешних моделей, включая уведомление об изменениях модели, сохранение версии, расширенное тестирование и валидацию.

Модель угроз кибербезопасности ИИ

п. 5.5 

Пример учета рисков информационной безопасности при применении ИИ-решений от третьих лиц

Составляющей разработки должно являться тестирование, перед переходом к которому целесообразно продемонстрировать функционал модели владельцу бизнес-кейса для принятия решения об одобрении ее применения.

Набор способов тестирования может зависеть от конкретного кейса и уровня риска. Оно может включать широкий перечень мер: от проверки данных (например, на «загрязнение») и результатов работы до стресс-тестирования, тестирования модели «красной командой», а также специальных проверок на смещения, интерпретируемость и встроенные ограничения.



Метрики и пороговые значения должны учитывать сценарий применения ИИ, критичность процессов, характеристики пользователей и сложность используемого типа ИИ. Генеративный ИИ и ИИ-агенты могут требовать использования дополнительных бенчмарков и тестов

Тестирование важно проводить на данных и сценариях, максимально приближенных к фактическому контексту применения ИИ. Если установленные показатели не достигаются, важно обеспечить доработку модели.

Процедуры тестирования и его ограничения целесообразно документировать, а его завершение оформлять отчетом с указанием примененных сценариев и наборов данных, использованных метрик, достигнутых результатов, выявленных ограничений и требуемых доработок.

Валидация

Валидацию целесообразно проводить силами обособленного подразделения, не участвовавшего в разработке модели ИИ. В зависимости от уровня риска она может быть обязательной или факультативной.

	Стандарт управления жизненным циклом моделей ИИ	п. 5.3	
	Система мониторинга качества моделей ИИ	п. 5.4	
	Порядок управления модельным риском	п. 5.3	

Пример подхода к валидации моделей ИИ



Валидация должна подтверждать, что модель пригодна для предполагаемого использования и соответствует требованиям к точности, надежности, масштабируемости, скорости, а ее результаты в необходимой степени прозрачны. При этом требования и применимые метрики должны определяться с учетом конкретного сценария применения, типа ИИ и уровня риска

Результаты валидации важно оформлять в виде заключения, содержащего вывод о допустимости применения модели, указание на выявленные ограничения и необходимые доработки, а также, при необходимости, условия, при которых модель может быть допущена к эксплуатации.

Возможные метрики на этапе разработки и валидации могут включать:

Метрики качества моделей ИИ	
Accuracy	Доля правильных предсказаний
Precision	Доля истинно-положительных среди предсказанных положительных
Recall	Доля истинно-положительных среди всех реальных положительных
F1-Score	Гармоническое среднее precision и recall
Gini / ROC-AUC	Способность модели различать классы
MAE / RMSE	Ошибка предсказания
Hallucination Rate	Частота генерации фактически неверной информации
Answer Relevancy	Соответствие ответа заданному вопросу
Correctness / Factuality	Соответствие ответа проверенным источникам
Prompt Adherence	Точность следования инструкциям в промпте
Coherence & Fluence	Логичность и естественность текста

Этические индикаторы	
Bias / Fairness	Дискриминация по демографическим признакам
Explainability	Понятность решений для человека
Toxicity / Safety	Наличие вредоносного/опасного контента
Privacy Compliance	Соответствие требованиям защиты данных
Auditability	Возможность реконструкции решений

Внедрение и эксплуатация

Внедрение моделей ИИ в эксплуатацию должно осуществляться только после проведения тестирования и валидации. При этом целесообразно заранее сформировать план последующего мониторинга, определить ответственных лиц и сотрудников, осуществляющих надзор за действиями и решениями ИИ, когда это предусмотрено логикой бизнес-процесса.



С учетом уровня риска и особенностей процесса внедрение может происходить поэтапно, например, когда доступ сначала предоставляется ограниченному количеству пользователей и клиентов

При эксплуатации модели ИИ важно проводить регулярный мониторинг качества: анализ ошибок, жалоб, инцидентов, изменения клиентского поведения и актуальности данных. Его можно осуществлять в том числе с применением автоматизированных механизмов, например для анализа метрик и уведомления ответственных лиц.

	Построение жизненных циклов моделей ИИ в системах принятия решений по страхованию	п. 5.4	
	Порядок управления модельным риском	п. 5.3	
	Система мониторинга качества моделей ИИ	п. 5.4	

Пример подхода к мониторингу качества ИИ

В рамках мониторинга качества целесообразно учитывать специфику ИИ-агентов, например осуществлять логирование, которое при необходимости может позволить отследить причину ошибок, особенно в отношении генеративного ИИ и ИИ-агентов.

	Система управления рисками агентского ИИ	п. 6.9	
--	--	--------	--

Примеры мер контроля рисков ИИ-агентов

В рамках мониторинга также целесообразно:

- операционализировать механизм выявления и учета инцидентов, реализуя предусмотренные меры реагирования на них;

 **Политика управления рисками использования моделей ИИ** п. 6.10 

Пример механизмов реагирования на инциденты ИИ

- проводить периодическую повторную оценку случаев применения ИИ и присвоенного им уровня риска, при необходимости с пересмотром применимых мер и контролей.

Внесение изменений и вывод из эксплуатации

При выявленном в рамках мониторинга на этапе эксплуатации ухудшении показателей качества модель необходимо дорабатывать, а при необходимости – временно ограничивать ее применение. Соответствующие решения следует принимать на основе критериев, установленных в правилах управления рисками ИИ.

 **Стандарт управления жизненным циклом моделей ИИ** п. 5.4 

Пример порядка доработки модели ИИ

Для корректировок модели или изменения области ее применения следует предусматривать:

- определение порога существенности изменений, который требует применения специальных процедур, в том числе повторной валидации модели;
- оценку влияния изменений на риск-профиль модели;
- документирование всех версий, изменений, принятых решений и установленных контролей.

Целесообразно сохранять информацию о выведенных из эксплуатации моделях, особенно если они могут применяться как запасные решения или для сравнения с актуальными моделями по бенчмаркам. О выводе моделей из эксплуатации важно своевременно уведомлять все подразделения, с бизнес-процессами которых было связано использование такой модели.

 **Порядок управления модельным риском** п. 5.7 

 **Построение жизненных циклов моделей ИИ в системах принятия решений по страхованию** п. 5.7 

Примеры подхода к обеспечению непрерывности деятельности

Фреймворк управления рисками искусственного интеллекта



ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 6.1. Управление рисками ИИ в рамках общей СУР
- пункт 6.6. Оценка и присвоение уровня риска ИИ
- пункт 6.7. Факторы риска при присвоении уровня риска ИИ
- пункт 6.8. Мониторинг и контроль рисков ИИ
- пункт 6.9. Минимизация выявленных рисков ИИ

О сервисе

В организации разработан фреймворк по управлению рисками ИИ, основанный на ключевых характеристиках нейросетей без учета сфер их использования. Он применим ко всей деятельности Московской биржи, связанной с разработкой, внедрением и использованием ИИ.

Фреймворк охватывает три основных типа ИИ, которые имеют сложную, нелинейную природу (генеративный, рекомендательный, предиктивный). При этом риски, связанные с более простыми алгоритмами, продолжают управляться в рамках стандартных процедур управления рисками.

Соблюдение рекомендаций Кодекса этики

6.1 Управление рисками искусственного интеллекта в рамках общей системы управления рисками

Управление риском ИИ интегрировано в уже существующую СУР и учитывает классические три линии защиты. Это позволяет использовать уже отработанные механизмы, адаптируя их под специфику ИИ.

6.6. Оценка и присвоение уровня риска искусственного интеллекта 6.7. Факторы риска при присвоении уровня риска искусственного интеллекта

Разработан набор параметров, позволяющих для каждой конкретной системы ИИ определить уровень риска. Такая критериальная оценка учитывает два параметра риска – значимость и вероятность, которые раскрыты посредством системы критериев, обеспечивающих всестороннюю оценку уровня риска:

Фреймворк управления рисками искусственного интеллекта

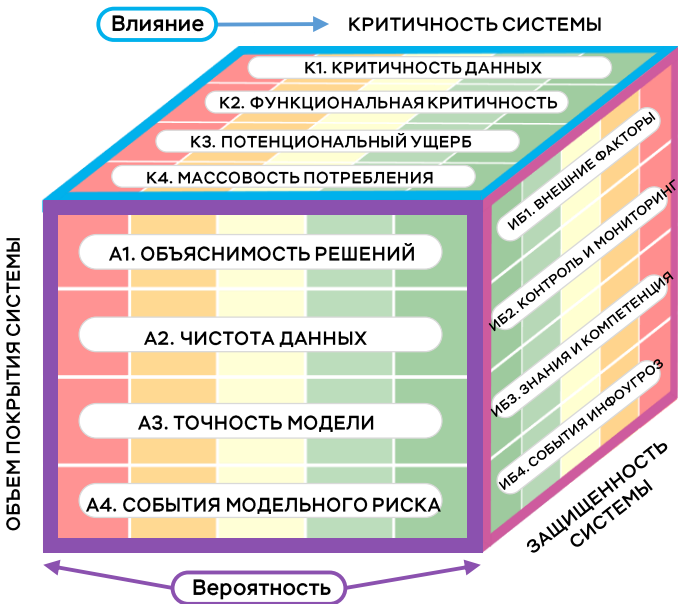


ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Параметр	
Значимость	Вероятность
Критерии для оценки	
1. Критичность данных. 2. Функциональная критичность. 3. Потенциальный ущерб. 4. Массовость потребления	Блок 1. Система ИИ как модель (объем покрытия системы) 5. Объяснимость решений, принимаемых ИИ, наличие механизмов валидации и оценки качества выдаваемых результатов. 6. Чистота данных, используемых в модели. 7. Точность модели. 8. Наличие риск-событий, связанных с реализацией модельного риска
	Блок 2. ИИ как ИТ-система (защищенность системы) 9. Уровень зависимости от внешних факторов. 10. Уровень контроля и мониторинга модели. 11. Уровень знаний и компетенций владельцев модели, разработчиков, пользователей. 12. Наличие риск-событий, связанных с реализацией информационных угроз в отношении модели ИИ

В совокупности эти 12 критериев оценки вероятности и значимости риска формируют куб, который позволяет оценить уровень риска. Каждый критерий оценивается по пятибалльной шкале и соответствует утвержденным в Московской бирже подходам к оценке нефинансовых рисков.

«Куб риска»
для оценки системы ИИ



Фреймворк управления рисками искусственного интеллекта



ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Критериальная система оценки, несмотря на детальность, не дает оценки совокупного уровня риска ИИ-системы. Для устранения этого пробела полученные в результате оценки баллы «приземлены» на тепловую карту риска.

В процессе апробации подхода все 12 критериев имеют одинаковый вес. В дальнейшем предполагается дифференцировать веса критериев в зависимости от типа ИИ, используемого в системе, и провести их последующую калибровку. Это обеспечит более адекватную и честную оценку уровня риска, учитывая специфику генеративного, рекомендательного или предиктивного ИИ.

Совокупный уровень риска для системы поиска информации по документам риск-менеджмента

		Влияние				
		Минимальное	Низкое	Среднее	Значительное	Высокое
		0-4	5-8	9-12	13-16	17-20
Вероятность	Минимальная	0-8				
	Низкая	9-16				
	Средняя	17-24	 RAG-RM			
	Значительная	25-32				
	Высокая	33-40				

6.8. Мониторинг и контроль рисков искусственного интеллекта

Управление рисками интегрировано в жизненный цикл системы ИИ.

На этапах разработки, тестирования и развертывания используется продуктовый риск-менеджмент, в котором осуществляется проактивное управление рисками будущей системы ИИ и заранее реализуются соответствующие меры, а также настраиваются контроли, чтобы минимизировать будущие риски.

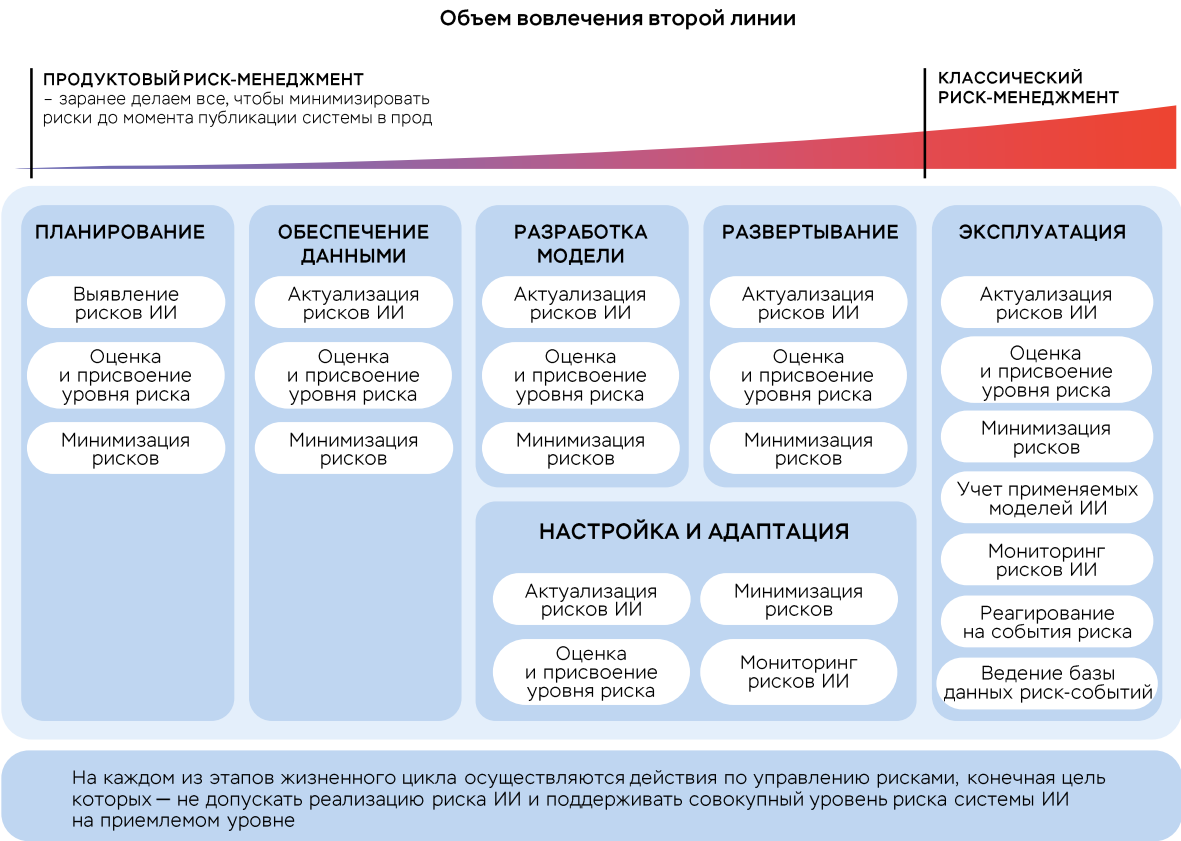
На этапе эксплуатации работает классический риск-менеджмент, который не только актуализирует информацию о выявленных ранее или новых идентифицированных рисках, но и покрывает работу с рисковыми событиями.

Фреймворк управления рисками искусственного интеллекта



ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Задачи риск-менеджмента на разных этапах жизненного цикла системы ИИ



Для каждого этапа жизненного цикла системы ИИ определены базовые риски, основанные на OWASP Top-10, присущие этапу, а также инструментарий СУР/MLSecOps, который возможно использовать с целью управления рисками ИИ.

Фреймворк управления рисками искусственного интеллекта



ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Управление рисками на этапах создания системы ИИ

Этап жизненного цикла	Ключевые риски	Инструментарий
Планирование	1) Стратегический. 2) Риск третьих сторон. 3) Правовые и этические риски	Определение значимости модели. Определение политик конфиденциальности, правовых и этических ограничений использования. Определение уровня риска зависимости от третьих сторон
Обеспечение данными	1) Утечка данных. 2) Риски цепочки поставок данных. 3) Отравление данных (data poisoning)	Обеспечение организованного контроля доступа к данным, классификация и управление данными, отслеживание происхождения данных, защита от отравления данных, защита извлеченных признаков от манипулирования, организация платформы управления данными
Разработка	1) Уязвимости векторов и эмбедингов. 2) Неавторизованные данные для обучения. 3) Промпт-инъекции. 4) Отравление модели/подделка источников. 5) Некорректная/избыточная обработка выходных данных	Контроль входных/выходных данных, использование доверенных компонент, состязательное обучение, тестирование и валидация
Развертывание	1) Уязвимости векторов и эмбедингов. 2) Уязвимости кода. 3) Утечки данных и системных инструкций. 4) Риски цепочки поставок. 5) Некорректная/избыточная обработка выходных данных	Контроль и воспроизводимость проводимых экспериментов. Защита активов. Организация платформы управления разработкой. MLOps-конвейер. Анализ защищенности системы ИИ. Классическая информационная безопасность

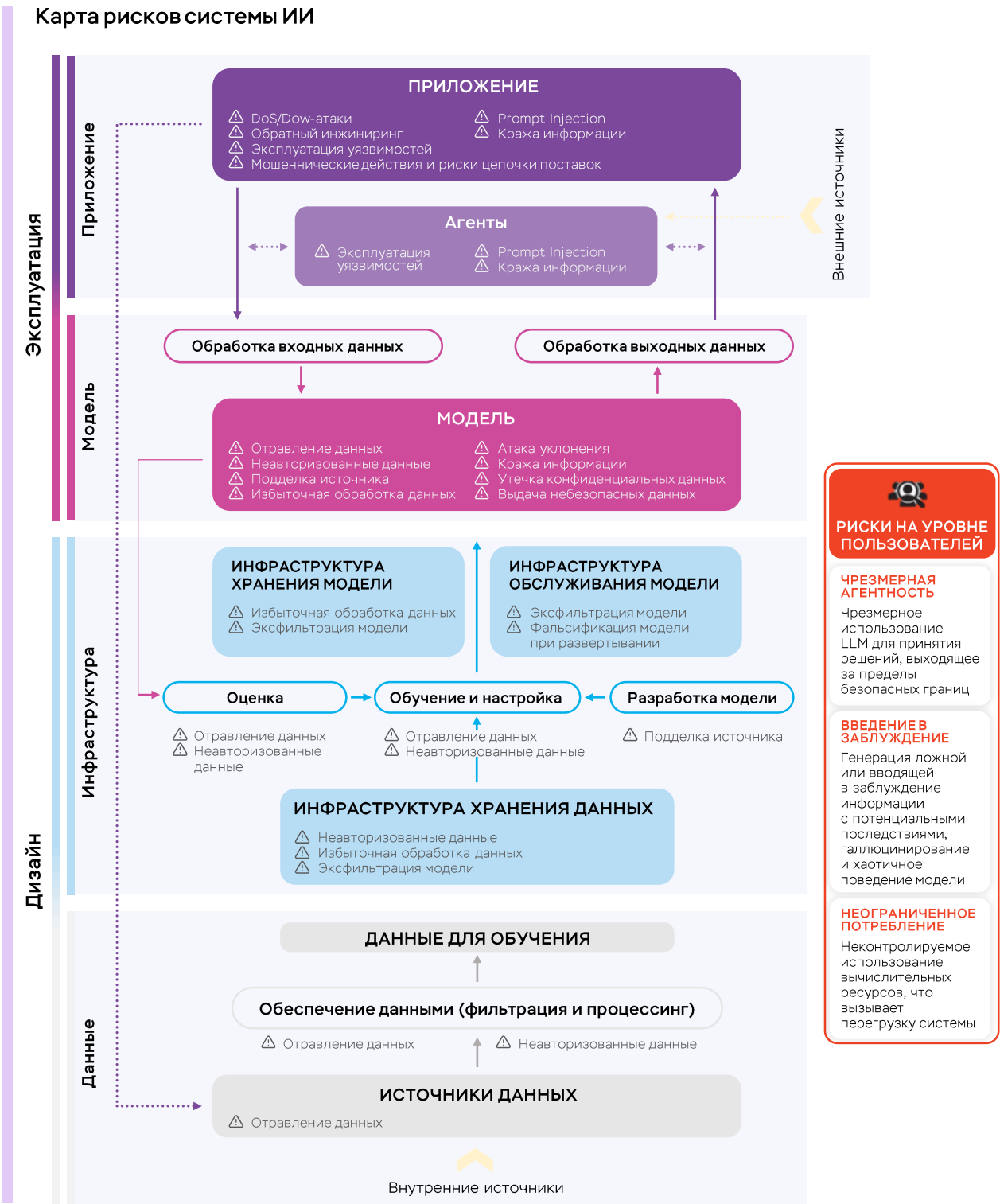
К моменту ввода в эксплуатацию ИИ-системы выстроена четкая система контролей и защитных механизмов. По этой причине для этапа эксплуатации подготовлена верхнеуровневая карта рисков, интегрирующая риски на трех ключевых уровнях: дизайна системы, ее эксплуатации, а также пользовательского взаимодействия. Риски распределены по ключевым архитектурным элементам системы ИИ.

Фреймворк управления рисками искусственного интеллекта



ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Карта рисков системы ИИ



Фреймворк управления рисками искусственного интеллекта



ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

6.9. Минимизация выявленных рисков искусственного интеллекта

На данный момент Московская биржа ориентируется на работу по принципу «человек в контуре»: человек сохраняет решающую роль, принимая финальное решение на основе рекомендаций, предоставленных моделью ИИ. Этот подход обеспечивает более высокий уровень управляемости, снижает риск автономных, потенциально ошибочных действий ИИ и подчеркивает ответственность пользователя за финальное решение.

Политика управления рисками использования моделей искусственного интеллекта



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 5.4. Мониторинг качества ИИ
- пункт 5.5. Обеспечение информационной безопасности
- пункт 5.6. Обеспечение конфиденциальности информации
- пункт 6.1. Управление рисками ИИ в рамках общей СУР
- пункт 6.2. Организация управления рисками при разработке и применении ИИ
- пункт 6.4. Учет применяемых моделей ИИ
- пункт 6.5. Выявление рисков ИИ
- пункт 6.6. Оценка и присвоение уровня риска ИИ
- пункт 6.7. Факторы риска при присвоении уровня риска ИИ
- пункт 6.9. Минимизация выявленных рисков ИИ
- пункт 6.10. Реагирование на реализовавшиеся риски ИИ

О сервисе

В Т-Банке принята Политика управления рисками использования моделей ИИ в банке. Она определяет принципы, методы и процедуры управления рисками, связанными с разработкой, внедрением и эксплуатацией систем ИИ в банке. Документ распространяется на все системы ИИ, используемые в банке, включая:

- генеративные модели (большие языковые модели, текстовые, мультимодальные);
- рекомендательные системы;
- предиктивные модели (прогнозирование, классификация);
- агентные системы (автономные ИИ-агенты);
- ИИ-инструменты для разработки кода.

Политика управления рисками использования моделей искусственного интеллекта



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Соблюдение рекомендаций Кодекса этики

6.1. Управление рисками искусственного интеллекта в рамках общей системы управления рисками

Управление рисками ИИ осуществляется в рамках общей СУР Т-Банка с учетом трех линий защиты:

1-я линия	Владельцы продуктов и разработчики ИИ-систем
2-я линия	Функции риск-менеджмента, информационной безопасности, комплаенса
3-я линия	Внутренний аудит

6.2. Организация управления рисками при разработке и применении искусственного интеллекта

Политика утверждена в виде самостоятельного документа и входит в пул применяемых в банке документов об управлении рисками в организации.

Политика включает в себя следующие разделы:

- организация СУР ИИ;
- классификация и оценка рисков ИИ;
- управление рисками на этапах жизненного цикла;
- категоризация угроз и контрмеры;
- требования к безопасности данных;
- мониторинг и контроль качества ИИ-систем;
- требования к разработке и внедрению;
- роли и ответственность;
- обучение и осведомленность;
- этика использования ИИ;
- управление инцидентами.

Политика управления рисками использования моделей искусственного интеллекта



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Политика пересматривается не реже одного раза в год или при:

- появлении новых регуляторных требований;
- реализации значимых инцидентов безопасности;
- появлении новых классов угроз (OWASP, SAIF updates).

Контроль за соблюдением Политики возложен на ответственные подразделения:

- риск-менеджмента (методология);
- информационной безопасности;
- внутреннего аудита (независимая проверка).

В Политике определены лица, ответственные за реализацию мер, направленных на минимизацию выявленных рисков ИИ и зоны их ответственности:

Ответственный	Роль
Владелец ИИ-продукта	Оценка рисков, внедрение контролей, мониторинг качества
Разработчик модели	Безопасная разработка, тестирование, документирование
MLSecOps команда	Безопасность ML-платформы, сканирование артефактов
Информационная безопасность	Контроль безопасности, мониторинг инцидентов, тестирование моделей «красной командой» (red teaming)
Риск-менеджмент	Методология оценки рисков, отчетность
Комплаенс	Соответствие регуляторным требованиям
Пользователь ИИ-системы	Следование политикам, верификация критических решений

6.4. Учет применяемых моделей искусственного интеллекта

В банке ведется реестр всех моделей с метаданными и владельцами. Это является одной из контрмер против ключевых угроз в рамках процессов MLSecOps.

6.5. Выявление рисков искусственного интеллекта

В рамках выявления рисков ИИ в банке документируется информация о порядке разработки ИИ, включая информацию об ограничителях (guardrails) на входе и выходе, логировании всех вызовов модели, мониторинге в реальном времени.

Политика управления рисками использования моделей искусственного интеллекта



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

6.6. Оценка и присвоение уровня риска искусственного интеллекта

6.7. Факторы риска при присвоении уровня риска искусственного интеллекта

Для каждой системы ИИ определяется уровень риска на основе 12 критериев, сгруппированных по двум параметрам – значимость и вероятность риска.

Параметр	
Значимость	Вероятность
Критерии для оценки	
1. Критичность данных Категория обрабатываемых данных (персональные, платежные, коммерческая тайна) 2. Функциональная критичность Влияние системы на бизнес-процессы Банка 3. Потенциальный ущерб Финансовый, репутационный, регуляторный ущерб при реализации риска 4. Массовость потребления Количество пользователей системы (внутренние / внешние)	Блок 1. Система ИИ как модель 5. Объяснимость решений Наличие механизмов валидации и оценки качества выдаваемых результатов 6. Чистота данных Качество и достоверность данных, используемых в модели 7. Точность модели Метрики качества модели (accuracy, precision, recall, F1) 8. Наличие риск-событий Наличие инцидентов, связанных с реализацией модельного риска
	Блок 2. ИИ как ИТ-система (защищенность) 9. Зависимость от внешних факторов Использование внешних API, провайдеров моделей, облачных сервисов 10. Уровень контроля и мониторинга модели Наличие трейсинга, логирования, алертинга 11. Компетенции команды Квалификация разработчиков, владельцев, пользователей модели 12. История инцидентов безопасности Реализованные информационные угрозы в отношении модели

Каждый критерий оценивается по пятибалльной шкале (1-5 баллов).

Тепловая карта рисков:

Низкий риск	≤ 20 баллов
Средний риск	21 – 40 баллов
Высокий риск	41 – 60 баллов

Политика управления рисками использования моделей искусственного интеллекта



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

6.9. Минимизация выявленных рисков искусственного интеллекта

В банке действует принцип «человек в контуре» («human-in-the-loop»): для всех критически важных решений, принимаемых с использованием ИИ человек сохраняет решающую роль. Он принимает финальное решение на основе рекомендаций модели и несет за него ответственность.

Кроме того, в рамках минимизации рисков ИИ предусмотрены отдельные мероприятия по обучению сотрудников, а также чек-листы для проверки перед выпуском моделей в эксплуатации.

6.10. Реагирование на реализовавшиеся риски искусственного интеллекта

Для целей управления инцидентами ИИ в банке предусмотрена их классификация (модельный риск, риски безопасности, комплаенс и другие).

Также предусмотрены следующие процедуры реагирования на инциденты:

- детектирование: автоматическое (алерты) или ручное (сообщения пользователей);
- классификация: определение типа и критичности инцидента;
- сдерживание: изоляция затронутых компонентов, аварийное отключение (kill-switch) при необходимости;
- расследование: анализ логов, поиск причин;
- устранение: фиксация уязвимости, обновление модели/контролей;
- восстановление: возврат к нормальной работе, к резервным копиям при необходимости;
- извлечение уроков: обновление политик, контролей, обучение.

5.4. Мониторинг качества искусственного интеллекта

Качество используемых в организации ИИ моделей регулярно проверяется в рамках мониторинга.

Мониторинг осуществляется по установленным метрикам:

- точность модели (accuracy, precision, recall);
- частота галлюцинаций;
- время ответа;
- частота срабатывания ограничителей (guardrails);
- количество инцидентов безопасности.

Политика управления рисками использования моделей искусственного интеллекта



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

5.5. Обеспечение информационной безопасности

В рамках СУР большое значение придается вопросам обеспечения информационной безопасности при использовании ИИ, включая специальные риски и атаки.

В Политике предусмотрены меры по управлению рисками на этапах жизненного цикла ИИ. Например, этап разработки и тестирования включает ряд мер, начиная с проверки обучающих данных на «загрязнение» (data poisoning) и заканчивая тестированием модели «красной командой» (red teaming) перед выпуском.

На этапе эксплуатации, в свою очередь, осуществляются в том числе проверка на деградацию качества, атаки, в том числе инъекции промптов (prompt injection), джейлбрейк (jailbreak), утечки данных.

Для выявления рисков информационной безопасности в банке внедрен автономный атакующий ИИ-агент Nulla, который превентивно обнаруживает уязвимости на всех этапах жизненного цикла программного обеспечения и позволяет повысить безопасность организации.

В отличие от традиционных средств защиты Nulla не просто сканирует код на наличие потенциальных рисков, а строит цепочки атак и демонстрирует реальный ущерб с учетом бизнес-контекста

5.6. Обеспечение конфиденциальности информации

В Политике установлены подробные требования к безопасности данных как на этапе обучения моделей ИИ, так и в процессе использования.

Также предусмотрены меры по обеспечению конфиденциальности данных на двух уровнях:

- организационном: система уровней доступа с принципом минимально необходимого доступа;
- архитектурном: техническая изоляция данных между пользователями и сессиями.

Стандарт управления жизненным циклом моделей искусственного интеллекта

РСХБ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 5.2. Проверка качества наборов данных
- пункт 5.3. Проверка качества ИИ
- пункт 5.4. Мониторинг качества ИИ
- пункт 6.4. Учет применяемых моделей ИИ
- пункт 6.6. Оценка и присвоение уровня риска ИИ

О сервисе

В Россельхозбанке осуществляется комплексное управление рисками ИИ на всех этапах жизненного цикла используемых в банке моделей. Модели разрабатываются и применяются в соответствии с принятыми в банке Принципами управления рисками и Стандартом управления жизненным циклом моделей ИИ.

Соблюдение рекомендаций Кодекса этики

5.2. Проверка качества наборов данных

На этапе поиска и подготовки данных командой разработки проводится их анализ, направленный на определение источников данных для обучения, и оценка наличия и достаточности данных. При поиске и подготовке данных команда разработки в том числе руководствуется внутренними документами в части требований к качеству данных и политикой информационной безопасности.

5.3. Проверка качества искусственного интеллекта

Формирование и согласование требований бизнес-кейса

Качество ИИ обеспечивается с начального этапа жизненного цикла модели.

Так, для каждого бизнес-кейса его владелец готовит детальные требования, которые включают в том числе:

- описание текущего процесса, в котором планируется применение модели;
- описание целевого процесса;
- источники и атрибутивный состав данных;
- целевую переменную;

Стандарт управления жизненным циклом моделей искусственного интеллекта

РСХБ**БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ****ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ**

- целевые показатели и бизнес-метрики, их базовые значения в числовом выражении;
- бизнес-ограничения для реализации бизнес-потребности;
- оценку уровня значимости модели (при необходимости).

Разработка и обучение модели

Если на этапе разработки целевые показатели качества не достигаются, то осуществляется пошаговый анализ обученной модели для выявления факторов, обусловивших недостижение показателей качества модели.

По итогам завершения этапа оформляется детальный отчет о разработке, включающий в себя исчерпывающую информацию о модели.

Для перехода к этапу тестирования команда разработки проводит демонстрацию функционала модели владельцу бизнес-кейса, по итогам которой формируется протокол одобрения (не одобрения) применения модели в процессе владельца бизнес-кейса.

Тестирование и оптимизация модели

Одним из основных методов оценки является А/В-тестирование. Если по результатам А/В-тестирования не удастся обеспечить достижение установленных показателей, команда разработки проводит детальный анализ и обеспечивает доработку модели перед проведением повторного тестирования.

По итогам завершения этапа оформляется отчет о тестировании модели и, в случае его успешности, модель готовится к внедрению в процесс владельца бизнес-кейса.

Первичная валидация модели

Первичная валидация проводится в отношении новых или существенно измененных уже действующих моделей. При этом в зависимости от уровня значимости модели, валидация может быть как обязательной, так и упрощенной или проводиться по усмотрению владельца бизнес-кейса для моделей с низким уровнем значимости. В случае, если по итогам первичной валидации уровень качества модели признан недостаточным для ее применения в процессах банка, команда разработки может провести доработку модели и повторно направить ее на рассмотрение в подразделение валидации.

Внедрение и эксплуатация модели

Внедрение моделей в целевой процесс владельца бизнес-кейса и введение ее в промышленную эксплуатацию осуществляется только после проведения этапа тестирования и первичной валидации.

Стандарт управления жизненным циклом моделей искусственного интеллекта

РСХБ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

В банке внедрен Аналитический сервис по автоматизации процесса оценки и переоценки отдельных видов профильных залогов юридических лиц.

Сервис направлен на сокращение трудозатрат специалистов банка, на поиск аналогов, проведение и оформление ежеквартальных переоценок всех предметов залога из залогового портфеля.

В качестве метрик для оценки качества работы моделей ИИ используются метрики MAPE (Mean Absolute Percentage Error), которые показывают процент средней ошибки модели ИИ и расчет доли попаданий прогнозных значений в заданный диапазон отклонения по сравнению с предыдущими результатами оценки.

По результатам проведенных экспериментов и обучения моделей ИИ сделан вывод о том, что по отдельным видам имущества модели ИИ не дают достаточно релевантных результатов, в связи с чем принято решение об исключении из обработки моделями ИИ таких видов имущества. При необходимости проведения оценки по имуществу из соответствующего списка пользователь получает уведомление о необходимости проведения ручной оценки.

Установлено требование о проведении не реже одного раза в год периодической валидации используемых моделей ИИ. В случае внесения изменений в модели ИИ валидация может проводиться несколько раз в год

5.4. Мониторинг качества искусственного интеллекта

Если в результате мониторинга показателей выявлено, что целевые показатели модели не достигаются, команда разработки обеспечивает доработку модели с привлечением владельца бизнес-кейса.

В рамках этапа мониторинга может также проводиться периодическая валидация модели. В случае если по итогам периодической валидации уровень качества модели признан недостаточным для ее применения в процессах банка, команда разработки может провести доработку модели и повторно направить ее на рассмотрение в подразделение валидации. Если по результатам повторного проведения валидации уровень качества модели также признан недостаточным, осуществляется вывод модели из промышленной эксплуатации.

Стандарт управления жизненным циклом моделей искусственного интеллекта



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

В результате мониторинга модели обеспечивается соответствие целевых показателей и показателей качества модели базовым значениям и достижение бизнес-результата для банка от применения модели в процессе владельца бизнес-кейса. Мониторинг целевых показателей и показателей качества модели продолжается до вывода модели из эксплуатации.

6.4. Учет применяемых моделей искусственного интеллекта

С целью ведения централизованного учета применяемых моделей ИИ в банке разработана и внедрена ИС RAISA (Система управления моделями и приложениями технологий ИИ RAISA – RSHB AI Systems and Application). ИС RAISA является основным инструментом по разработке, обучению и хранению моделей машинного обучения и ИИ в банке, на базе которой создан реестр моделей.

Реестр включает в себя поля, содержащие в себе исчерпывающую информацию о модели в рамках всего жизненного цикла:

- ID и наименование модели;
- версия, статус и тип модели;
- целевые показатели модели;
- документы, формируемые на каждом из этапов жизненного цикла модели;
- описание подхода к оценке рисков, даты проведения валидации и актуализации информации и др.

6.6. Оценка и присвоение уровня риска искусственного интеллекта

Для моделей с оценкой уровня значимости «высокий» и «средний» обязательным является утверждение модели оценки риска перед внедрением.

Построение жизненных циклов моделей искусственного интеллекта в системах принятия решений по страхованию



СПРАВЕДЛИВОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 3.3. Недопустимость использования дискриминационных факторов
- пункт 5.2. Проверка качества наборов данных
- пункт 5.3. Проверка качества ИИ
- пункт 5.4. Мониторинг качества ИИ
- пункт 5.7. Обеспечение непрерывности деятельности

О сервисе

Модели ИИ интегрированы в контур андеррайтинга для оценки рисков при пролонгации страховых полисов ОСАГО и их корректной тарификации. Ежедневно они обрабатывают сотни тысяч запросов.

Высокая нагрузка на сервис делает критически важным обеспечение непрерывного цикла управления моделями ИИ – от проектирования до вывода из эксплуатации. В связи с этим в основе системы принятия решений в моторных видах страхования для физических лиц (ОСАГО, КАСКО) на базе моделей ИИ лежит сквозной контроль жизненного цикла ИИ-решений.

Соблюдение рекомендаций Кодекса этики

3.3. Недопустимость использования дискриминационных факторов

На этапе аналитики и обучения моделей ИИ контролируется смещение (bias) во внутренних (например, портрет клиента, история страхования, сведения о транспортном средстве) и внешних (например, история эксплуатации транспортного средства) данных, входящих в обучающую выборку. Это позволяет избегать дискриминационных решений и необоснованного завышения стоимости страховых полисов.

5.2. Проверка качества наборов данных

Высокая интенсивность запросов делает невозможным ручной процесс проверки каждого из них. В связи с этим организована регламентная процедура проверки и защиты наборов данных перед обучением или обновлением моделей ИИ.

Построение жизненных циклов моделей искусственного интеллекта в системах принятия решений по страхованию



СПРАВЕДЛИВОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

В рамках этой процедуры проводится проверка качества обучающей выборки, при которой учитывается уникальность записей, применяемые фильтры и временной диапазон. Для каждого источника данных также проводятся проверки по следующим показателям:

- полнота заполнения;
- изменение полноты заполнения с течением времени;
- смещение распределения с течением времени;
- сравнение с предыдущими выгрузками.

По результатам выделяются показатели, отклонения в значении которых превышают допустимые пределы.

При использовании для формирования решения по цене внешних данных (например, истории использования автомобиля) проводится регулярная проверка их качества, в рамках которой:

- во внимание принимается доля успешных ответов источника, полнота заполнения, стабильность распределений во времени;
- проводится регулярная оценка значимости источников внешних данных и степени их влияния на моделируемый процесс;
- при предоставлении разными источниками однотипных данных проводится их сопоставление.

5.3. Проверка качества искусственного интеллекта

Перед выводом модели ИИ в продуктивную среду проводится тестирование на реальном потоке данных в фоновом режиме без влияния на клиентов. Это позволяет оценить надежность алгоритма в условиях, приближенных к реальным. По результатам проводится анализ и дается оценка целесообразности вывода в продуктивную среду.

Построение жизненных циклов моделей искусственного интеллекта в системах принятия решений по страхованию



СПРАВЕДЛИВОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

5.4. Мониторинг качества искусственного интеллекта

Осуществляется мониторинг качества принятых моделями ИИ решений с учетом запланированного диапазона, в том числе непрерывный мониторинг дрейфа данных и метрик качества стабильности данных. В частности, для этого проводится регулярный мониторинг совпадения распределений каждого из показателей при реальном использовании с обучающей выборкой. При отклонении сверх допустимых пределов автоматически отправляется уведомление разработчикам моделей ИИ с сигналом о необходимости вмешаться.

Модели ИИ также регулярно обновляются с использованием свежих данных. При этом реализуется следующее:

- качество моделей ИИ оценивается на основе технических и бизнес-метрик;
- при значимом дрейфе выпускается обновленная версия модели ИИ, и отдельно рассчитывается влияние обновления на изменение тарифов (или других моделируемых величин);
- после получения автоматического уведомления с рассчитанными показателями владельцы продуктов по моторным видам страхования принимают решение о необходимости запуска обновлений и времени их применения.

5.7. Обеспечение непрерывности деятельности

В условиях большого потока поступающих запросов (более 200 в минуту) при недоступности одного или нескольких источников данных в организации применяются упрощенные модели, использующие для прогнозирования сокращенный перечень показателей. При полной недоступности моделей ИИ применяется алгоритм принятия решений, основанный на списке бизнес-правил. Процедуры бесшовного переключения между моделями, а также механизмы возврата к ручному андеррайтингу четко регламентированы.

Единый подход к внедрению и эксплуатации ИИ-решений



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 5.5. Обеспечение информационной безопасности
- пункт 5.6. Обеспечение конфиденциальности информации
- пункт 6.2. Организация управления рисками при разработке и применении ИИ
- пункт 6.3. Процессы управления рисками ИИ
- пункт 6.4. Учет применяемых моделей ИИ
- пункт 6.5. Выявление рисков ИИ
- пункт 6.8. Мониторинг и контроль рисков ИИ
- пункт 6.11. Ведение базы реализовавшихся рисков ИИ

О сервисе

СУР ИИ в организации охватывает клиентские и внутренние ИИ-сервисы, включая речевую аналитику, виртуального ассистента и голосового помощника.

Правила управления рисками ИИ интегрируются в общую СУР и используются на разных этапах жизненного цикла модели, включая инициирование, сбор и обработку данных, разработку и тестирование, внедрение, эксплуатацию, мониторинг и вывод из эксплуатации.

Ключевые элементы СУР ИИ:

- централизованный учет моделей ИИ и случаев их применения;
- оценка уровня риска с учетом влияния на клиента, использования персональных данных, объяснимости решений и критичности процесса;
- валидация качества модели ИИ перед внедрением и регулярный мониторинг после запуска;
- контроль входных данных, базы знаний, сценариев и изменений модели;
- регистрация инцидентов и риск-событий, связанных с применением ИИ.

Соблюдение рекомендаций Кодекса этики

6.2. Организация управления рисками при разработке и применении искусственного интеллекта

Порядок управления рисками ИИ в организации интегрирован в общую СУР. Он определяет роли участников, критерии отнесения решений к ИИ, требования к согласованию, контролю качества, информационной безопасности и мониторингу после внедрения.

Единый подход к внедрению и эксплуатации ИИ-решений



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

6.3. Процессы управления рисками искусственного интеллекта

СУР ИИ включает учет моделей, выявление рисков, оценку уровня риска, мониторинг, минимизацию рисков, реагирование на риск-события и ведение базы таких событий. Для клиентских сервисов дополнительно учитываются возможность обращения к сотруднику, прозрачность применения ИИ и влияние ошибки на клиента.

6.4. Учет применяемых моделей искусственного интеллекта

Для ИИ-решений ведется централизованный реестр, в котором фиксируются назначение модели, владелец процесса, используемые данные, поставщик или разработчик, уровень риска, дата внедрения, ключевые метрики качества и ответственные лица.

6.5. Выявление рисков искусственного интеллекта

По каждому значимому ИИ-решению документируются сценарии применения, источники данных, ограничения модели, порядок контроля качества, правила эскалации на сотрудника и условия, при которых решение не должно использоваться без дополнительной проверки человеком.

5.5. Обеспечение информационной безопасности

В рамках СУР ИИ отдельно оцениваются риски нарушения информационной безопасности и операционной надежности: несанкционированное изменение сценариев, недоступность сервиса, ошибки интеграций, некорректная обработка запросов и доступ к данным сверх полномочий.

5.6. Обеспечение конфиденциальности информации

Для ИИ-решений определяются требования к обработке данных ограниченного доступа: состав данных, допустимые цели использования, круг пользователей, правила хранения, обезличивания при необходимости и ограничения на передачу информации во внешние сервисы.

6.8. Мониторинг и контроль рисков искусственного интеллекта

После запуска проводится регулярный мониторинг качества модели: анализ ошибок, жалоб, инцидентов, изменения клиентского поведения и актуальности данных. При ухудшении показателей качества модель или сценарии дорабатываются, а при необходимости их применение временно ограничивается.

6.11. Ведение базы реализовавшихся рисков искусственного интеллекта

Инциденты, связанные с некорректной работой ИИ, ошибочной маршрутизацией, неверным ответом, нарушением доступности или рисками информационной безопасности, регистрируются с описанием причины, последствий, корректирующих мероприятий и статуса исполнения.

Порядок управления модельным риском



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 5.3. Проверка качества ИИ
- пункт 5.4. Мониторинг качества ИИ
- пункт 5.7. Обеспечение непрерывности деятельности
- пункт 6.1. Управление рисками ИИ в рамках общей СУР
- пункт 6.2. Организация управления рисками при разработке и применении ИИ
- пункт 6.4. Учет применяемых моделей ИИ
- пункт 6.5. Выявление рисков ИИ
- пункт 6.6. Оценка и присвоение уровня риска ИИ
- пункт 6.9. Минимизация выявленных рисков ИИ
- пункт 6.11. Ведение базы реализовавшихся рисков ИИ

О сервисе

В ВТБ жизненный цикл моделей ИИ интегрирован в единый производственный процесс разработки и внедрения информационных технологий. Этот процесс учитывает действующие в банке нормативные документы по управлению модельным риском.

Ежегодно в ВТБ реализуется около 400 модельных инициатив для различных бизнес-процессов, поэтому такой подход, охватывающий не только регуляторные модели, позволяет существенно снизить общий уровень модельного риска и повысить культуру ответственного использования ИИ.

Соблюдение рекомендаций Кодекса этики

6.1. Управление рисками искусственного интеллекта в рамках общей системы управления рисками

Управление рисками при разработке и применении ИИ осуществляется в рамках общей системы управления модельным риском. Этапы жизненного цикла моделей ИИ интегрированы с процессами внедрения информационных систем банка.

6.2. Организация управления рисками при разработке и применении искусственного интеллекта

Управление рисками ИИ осуществляется в соответствии с утвержденным Порядком управления модельным риском (ПУМР).

Порядок управления модельным риском



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

ПУМР содержит следующие ключевые разделы:

- цели и задачи, основные принципы управления модельным риском;
- классификация и жизненный цикл моделей в контексте модельного риска;
- этапы и участники процесса управления модельным риском;
- положение об учете величины модельного риска для определения экономического капитала организации;
- принципы независимого внутреннего контроля и отчетности по модельному риску.

6.4. Учет применяемых моделей искусственного интеллекта

Информация об используемых в банке моделях содержится в реестре Системы управления модельным риском. Реестр содержит карточки эксплуатируемых моделей:

- название;
- тип модели в соответствии с принятой в банке классификацией;
- подразделение – владелец модели;
- подразделение – разработчик модели;
- область применения и охват;
- дата ввода/вывода модели из эксплуатации;
- результаты проверок модели;
- история изменений модели в течение всего жизненного цикла.

6.5. Выявление рисков искусственного интеллекта

Регламентируются процессы всех этапов жизненного цикла модели, включая заявку на разработку, валидацию, эксплуатацию, мониторинг и вывод модели из эксплуатации.

Регламентация осуществляется как на уровне верхнеуровневых внутрибанковских нормативных документов, так и на уровне методологических указаний.

Установлены требования к минимальной документации по разработке и изменению моделей, а также требования к проведению и документированию испытаний при их внедрении.

По результатам разработки моделей формируются отчеты, позволяющие понять и воспроизвести логику этого этапа.

Порядок управления модельным риском



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Для разработки ИИ решений созданы отдельные среды и внедрен MLOps конвейер технологической выкладки моделей в общий промышленный ИТ-ландшафт.

Этапы разработки и возникающие в ходе нее артефакты отслеживаются, сохраняются и версионизируются в Системе управления моделями. Это позволяет вести детальный сквозной учет всей истории разработки и эксплуатации моделей.

6.6. Оценка и присвоение уровня риска искусственного интеллекта

На этапе разработки каждой модели присваивается агрегированный уровень значимости (высокая, средняя, низкая).

Значимость устанавливается в зависимости от уровня:

- материальности (оценивается масштаб использования модели или охват последствий реализации модельного риска на основании ряда показателей, характеризующих долю охвата);
- величины возможных убытков;
- степени регуляторного надзора (определяется по обязательности согласования использования модели с регулятором, наличию или отсутствию обязательных к исполнению регуляторных требований к модели).

На этапах валидации и эксплуатации проводится оценка уровня риска ИИ. При этом учитывается не только уровень риска самой модели, но и конкретный случай ее применения.

При оценке и присвоении уровня риска учитываются, в том числе:

- риск потерь ввиду некорректной постановки задачи моделирования;
- риск потерь в связи с низкой точностью модели (недооценка или переоценка риска);
- риск потерь ввиду некорректной обработки входных данных из-за ошибки в коде;
- риск потерь ввиду некорректного внедрения;
- риск потерь ввиду ошибок на источнике данных или низкого качества данных, которые используются в модели;
- риск потерь ввиду некорректного использования результатов модели;
- риск ошибок модели ввиду отсутствия процесса контроля за работой модели.

В рамках общепанковских процедур по управлению операционным риском по перечисленным факторам проводится регулярная самооценка операционного риска подразделением, ответственным за процессы разработки и внедрения ИИ, автоматизированный учет и контроль рисков и рисков событий в специализированной информационной системе, определение мер реагирования и необходимости в изменении внутренних нормативных документов.

Порядок управления модельным риском



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

6.9. Минимизация выявленных рисков искусственного интеллекта

Помимо прочих мер по минимизации рисков ИИ банк разрабатывает специальные меры по обеспечению надежности при разработке и внедрении цифровых помощников на базе больших языковых моделей, в том числе ИИ-агентов.

Сервис выявления галлюцинаций в ответах больших языковых моделей

Сервис направлен на минимизацию операционных, в том числе технологических, рисков при использовании больших языковых моделей в бизнес-процессах. Он выступает в качестве встроенного механизма контроля качества, который выявляет галлюцинации и непрерывно валидирует фактологическую точность ответов моделей на этапе их тестирования и эксплуатации. Сервис сверяет ответ модели с исходным контекстом при помощи отдельно и независимо обученной модели.

Сервис на уровне предложений выявляет следующие виды галлюцинаций:

- противоречие – если сгенерированный текст противоречит источникам;
- ответ без контекста – текст, сгенерированный без опоры на источник;
- неподтвержденный ответ – текст, сгенерированный на основе источника, но включающий в себя дополнительную информацию, не содержащуюся в документе;
- ответ не на тот вопрос – модель отвечает не на вопрос пользователя;
- частичный ответ – ответ модели относится лишь к части объектов, интересующих пользователя. Эта категория применима, например, если вопрос пользователя задан об условиях автокредитов в целом, а модель предоставляет информацию только об условиях конкретной программы, явно не уточняя этого;
- ложный отказ – модель сообщает, что интересующая пользователя информация не найдена несмотря на то, что найденные документы ее содержат

Порядок управления модельным риском



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

6.11. Ведение базы реализовавшихся рисков искусственного интеллекта

Информация об инцидентах, в том числе по модельному риску, включая риск ИИ, содержится в автоматизированной системе управления операционным риском. Эти инциденты подлежат обязательному расследованию.

5.3. Проверка качества искусственного интеллекта

Для моделей высокой значимости в обязательном порядке предусмотрено проведение первичной валидации.

Валидация осуществляется независимым подразделением, отвечающим за процессы управления модельным риском, и не может быть заменена тестированием, которое проводит команда – разработчик модели.

5.4. Мониторинг качества искусственного интеллекта

В банке создана единая платформа мониторинга моделей, которая позволяет контролировать метрики качества модели и входящего потока данных в процессе эксплуатации.

На этапах разработки и валидации определяются и утверждаются метрики количественной оценки качества модели, а также пороговые значения для них. Данные метрики в дальнейшем используются при постановке модели на автоматический мониторинг и рассылки сигналов владельцу и разработчику модели при превышении пороговых значений.

Сервис инвестиционного консультирования «Интеллект»

«Интеллект» является программой автоконсультирования и автоследования, которая предоставляет индивидуальные инвестиционные рекомендации (ИИР). Сервис доступен в приложении «ВТБ Онлайн» и «ВТБ Мои Инвестиции» с функционалом подключения стратегий. Перечень стратегий, доступных для подключения, приводится в интернет-банке ВТБ Онлайн и на официальном сайте банка.

Сервис позволяет привести портфель клиента по составу и долям к модельному портфелю стратегии. Также «Интеллект» автоматизированным способом преобразовывает предоставленные ИИР в поручение брокеру на совершение сделки с ценными бумагами без непосредственного участия клиента.



Порядок управления модельным риском



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Стратегия осуществляет от 1 до 4 ребалансировок в месяц. Не реже раза в месяц в рамках рабочей группы проводится детальный разбор предоставленных ИИ торговых рекомендаций. В рамках разбора проверяется корректность работы модели и качество инвестиционных рекомендаций. По итогам разбора принимается коллегиальное решение о валидности рекомендаций модели ИИ и продолжении ее работы либо об остановке эксплуатации и инициировании доработки модели

5.7. Обеспечение непрерывности деятельности

Для моделей высокой значимости обеспечена непрерывность деятельности.

Уровень критичности ИТ-сервиса, зависящего от функционирования ИИ, определяет поддержку, резервирование.

Заблаговременно проектируются и реализуются консервативные механизмы, позволяющие не останавливать процесс, использующий ИИ, в случае выявления некорректной работы ИИ-сервиса или его неработоспособности, а также минимизировать возможные потери от работы по обходному алгоритму.

Система управления рисками агентского ИИ



ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 6.1. Управление рисками ИИ в рамках общей СУР
- пункт 6.2. Организация управления рисками при разработке и применении ИИ
- пункт 6.6. Оценка и присвоение уровня риска ИИ
- пункт 6.9. Минимизация выявленных рисков ИИ

О сервисе

В организации сформирован единый подход к выявлению, оценке и минимизации рисков при разработке и внедрении агентских решений. Это позволяет заранее учитывать риски превышения автономности, некорректного поведения, галлюцинаций, промпт-инъекций, утечек данных, нарушения доступности и так далее.

Оценка рисков ИИ-агентов встроена в действующие процессы организации и проводится до промышленного запуска ИИ-агента с участием профильных подразделений: валидации, операционных рисков, технологического блока, кибербезопасности и юридической службы. Профильными подразделениями проводится оценка уровня риска, а также формируются предложения для владельцев агентских решений по минимизации рисков.

Такой подход снижает вероятность некорректных действий ИИ-агента, раскрытия конфиденциальной информации, предоставления недостоверных ответов или выполнения агентом действий за пределами допустимых полномочий.

Соблюдение рекомендаций Кодекса этики

6.1 Управление рисками искусственного интеллекта в рамках общей системы управления рисками

Таксономия риска ИИ-агентов позволяет встроить управление рисками агентских решений в общую СУР организации.

Agent Execution Framework

Внедрение ИИ-агентов проходит в рамках универсального Agent Execution Framework (AEF) – унифицированного набора инструментов и сервисов, который упрощает соблюдение и контроль требований к ИИ-агентам. Он является обязательным к использованию всеми командами и помогает обеспечить безопасность, надежность, архитектурную



Система управления рисками агентского ИИ



ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

согласованность и управление рисками при разработке и эксплуатации ИИ-агентов независимо от бизнес-сценария. В частности, AEF обеспечивает технологическую доступность к внутренним механизмам контроля.

В рамках оценки риска конкретного ИИ-агента определяется перечень необходимых мер митигации (в зависимости от уровня риска), в том числе защитные механизмы онлайн контроля «Guardrails». Эта информация передается по интеграции в AEF. В момент установки агента в промышленный контур в AEF отработывают контроли, которые показывают, исполнил ли владелец агента требования по митигации. В случае если меры митигации не учтены в логике работы агента, его работа будет ограничена

AEF Manager

Централизованная система контроля выполнения требований регуляторов и управления агентами



AEF

Agent execution framework

AEF Containers

Среда исполнения агентов, настроенная в соответствии с требованиями AEF и содержащая AI-агентов



AEF Operator

Инструменты деплоя и контролируемого изменения окружения агента



GuardRails

Меры митигации агентных рисков



AEF Controller

Сервисы обработки, фильтрации, нормализации данных о работе агентов



Система управления рисками агентского ИИ



ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

6.2. Организация управления рисками при разработке и применении искусственного интеллекта

Для оценки рисков ИИ-агентов определены ответственные подразделения, шкала уровней риска и матрица полномочий по принятию рисков.

6.6. Оценка и присвоение уровня риска искусственного интеллекта

Предусмотрена классификация рисков ИИ-агентов по ключевым группам: качество и функционал, надежность, безопасность данных и систем, правовые и регуляторные риски.

Группа «Качество и функционал»²:

Некорректное поведение и использование инструментов	Отклонение ИИ-агента в процессе работы от заданных разработчиками целей или неоптимальное решение поставленных задач
Превышение уровня автономности	Риск потерь в результате того, что ИИ-агенты получают слишком широкие права принятия решений без достаточного человеческого надзора
Деградация знаний и навыков ИИ-агента	Постепенное снижение способности агента адекватно обрабатывать и интерпретировать информацию, приводящее к утрате точности, надежности и стабильности принимаемых решений
Дезинформация и манипуляции ИИ	Предоставление ложной или вводящей в заблуждение информации
Галлюцинации ИИ-агентов	Генерация ИИ-агентом ложной, противоречивой или бессмысленной информации
Токсичность и дискриминация	Передача пользователю оскорбительного, небезопасного или неприемлемого контента
Низкое качество внедряемых ИИ-агентов	Отсутствие устоявшихся стандартов, методов и практик внедрения, эксплуатации, интеграции, совместимости и масштабируемости технологий ИИ

² Включается в состав операционного риска, модельного риска, риска поведения.

Система управления рисками агентского ИИ



ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Группа «Надежность»³:

Отказ инфраструктуры ИИ-платформ и сервисов	Отказы или нестабильная работа информационных систем, обеспечивающих функционирование ИИ-агентов
Взаимное влияние ИИ-решений на едином ландшафте	Например, вследствие нарушения профиля нагрузки, заикливания и др.

Группа «Безопасность данных и систем»⁴:

Утечки данных через ИИ-агентов	Несанкционированное раскрытие конфиденциальной информации, информации о среде исполнения и конфигурации ИИ-агента или мультиагентной системы (в том числе информации о среде исполнения, системном промпте, цели, функциях, бизнес-логике) третьим лицам в следствие промпт-атак
Несанкционированные действия ИИ-агента	Компрометация ИИ-агента в следствие промпт-атак, приводящая к выполнению опасных операций (в том числе внедрению и исполнению кода, удалению или модификации файлов и данных, нарушению изоляции среды, эскалации атак в системе)
Нарушение доступности или исчерпание лимитов	Нарушение стабильной работы информационных систем, ИИ-агентов, мультиагентных систем в следствие промпт-атак, приводящих к перегрузке, выводу из строя, исчерпанию вычислительных ресурсов или финансовых лимитов, созданию операционных помех
Нарушение мультиагентного взаимодействия	Нарушение функционирования группы ИИ-агентов вследствие промпт-атак, приводящее к передаче ложных данных между агентами, искажению их целей или распространению вредоносных команд, что вызывает сбой рабочего процесса мультиагентной системы
Нарушение функционирования ИИ-агента	Искажение логики, целей или поведения ИИ-агента в следствие промпт-атак, приводящее к его некорректной работе или отклонению от бизнес-сценариев

³ Включается в состав риска информационных систем.

⁴ Включается в состав риска информационной безопасности.

Система управления рисками агентского ИИ



ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Группа «Правовые и регуляторные риски»⁵:

Нарушение интеллектуальных прав, авторского права или политики лицензирования данных	Использование данных, интеллектуальной собственности или программных продуктов без наличия соответствующих разрешений или с нарушением условий лицензионных соглашений, а также нарушение требований законодательства в части оборота информации
Допущение ИИ-агентами правовых ошибок при функционировании	Например, неправильные юридические консультации или неверное составление документов

Оценка проводится по шкале уровня риска (низкий, средний, высокий, очень высокий), а интегральный рейтинг определяется по максимальному из выявленных рисков.

При формировании классификации рисков учитывались в том числе следующие источники информации:

- NIST AI Risk Mgmt. Framework: GenAI Profile;
- NIST Secure SW Dev Practices for GenAI;
- ISO 42001;
- ISO 23894;
- OWASP LLM Top-10;
- Google Secure AI Framework.

6.9. Минимизация выявленных рисков искусственного интеллекта

Для минимизации рисков предусмотрены практические меры контроля:

- привлечение человека для проверки и утверждения критичных решений ИИ-агента (human-in-the-loop);
- самопроверка модели (self-check / critic);
- привязка ответов агента к проверенным источникам данных (grounding и RAG);
- механизмы контроля входящих и исходящих данных (guardrails);

⁵ Включается в состав правового риска, регуляторного риска.

Система управления рисками агентского ИИ



ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

- проектирование системных промптов;
- ограничение частоты запросов (rate limiting);
- наблюдаемость и логирование.

Для промышленного запуска ИИ-агента требуется прохождение контроля качества (quality gate), подтверждающего наличие оценки и принятия риска в соответствии с утвержденной матрицей полномочий (определяет уровни принятия решения по рискам в зависимости от уровня риска).

Применение RAG

В банке разработана и применяется система GigaSearch (внутренняя разработка) для генерации ответов на основе информации, извлеченной из внешних источников (на базе RAG-механик). Она использует технологии обработки естественного языка и машинного обучения для поиска и предоставления релевантных данных. Система способна анализировать запросы пользователей, находить нужную информацию в базах данных и документах, а затем формировать точные и информативные ответы, значительно снижая риск дезинформации пользователей

Проектирование системных промптов

Системный промпт — это первая линия обороны ИИ-агента. Его правильное проектирование критически важно для защиты от промпт-атак, утечек информации и некорректного поведения агента.

Системный промпт должен удовлетворять ряду требований:

- неизменяемость системного промпта после релиза;
- запрет конфиденциальной информации (персональные данные, коммерческая тайна и т. п.);
- запрет скрытых символов;
- запрет описания информации об архитектуре;



Система управления рисками агентского ИИ



ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

- запрет описания бизнес-логики (например, описание правил валидации, формулы расчета, условий принятия решений и др.);
- явное определение роли, целей и задач ИИ-агента;
- явный список допустимых и запрещенных тем;
- явное ограничение языков взаимодействия;
- приоритет безопасности (безопасность имеет приоритет над полезностью).

Система мониторинга качества моделей искусственного интеллекта

Альфа Банк

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 5.4. Мониторинг качества ИИ
- пункт 6.4. Учет применяемых моделей ИИ
- пункт 6.6. Оценка и присвоение уровня риска ИИ

О сервисе

Система управления модельным риском в Альфа-Банке построена на четырех взаимодополняющих линиях защиты:

- A/B-тестирование;
- валидация;
- мониторинг;
- внутренний аудит.

Соблюдение рекомендаций Кодекса этики

5.4. Мониторинг качества искусственного интеллекта

A/B-тестирование

В банке проводится A/B-тестирование, которое позволяет проверить гипотезы о наличии финансового эффекта от модели, оценить прирост бизнес-метрик и принять решение о масштабировании.

Валидация

Валидация включает первичную проверку модели после ее разработки, ИТ-валидацию качества внедрения и ежегодные / периодические проверки в эксплуатации, что гарантирует объективную оценку качества и выявление ошибок на всех этапах жизненного цикла.

Непрерывный мониторинг

Непрерывный мониторинг осуществляется посредством автоматизированного анализа метрик моделей: владельцы продукта автоматически информируются об отклонениях метрик. Также предусмотрены периодические отчеты о деградации моделей и проведение сессий оценки критичности отклонений.

Система мониторинга качества моделей искусственного интеллекта

Альфа Банк

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Внутренний аудит

В ходе внутреннего аудита осуществляется независимый контроль соответствия процессов нормативным требованиям, документируются результаты и формируются рекомендации по улучшению методологий.

6.4. Учет применяемых моделей искусственного интеллекта

Применяемые модели ИИ учитываются в централизованном реестре. В нем содержится информация обо всех версиях моделей, бизнес-процессах, в которых они применяются, и риск-параметрах.

Реестр обеспечивает сертифицированный доступ к информации для бизнес-подразделений, подразделений по валидации и управлению рисками, позволяя контролировать жизненный цикл моделей и принимать обоснованные решения об их дальнейшем использовании.

6.6. Оценка и присвоение уровня риска искусственного интеллекта

Все модели документируются (цели, входные данные, критерии, ограничения) и проходят предварительную оценку рисков для клиентов, в том числе через проведение А/В тестирования, и автоматизированный мониторинг модельных метрик, что позволяет в кратчайшие сроки оценивать уровень критичности рисков и быстро корректировать работу систем ИИ.

Модель угроз кибербезопасности искусственного интеллекта



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 5.5. Обеспечение информационной безопасности
- пункт 5.6. Обеспечение конфиденциальности информации

О сервисе

В организации разработана комплексная модель угроз для кибербезопасности систем ИИ (МУ КБ ИИ). Она охватывает весь жизненный цикл системы ИИ – от подготовки данных и обучения модели до ее интеграции в конечное приложение.

МУ КБ ИИ описывает потенциальные риски для генеративного ИИ и предиктивного ИИ. Для каждой угрозы в ней определены возможные последствия и конкретные свойства информации, которые могут быть нарушены (конфиденциальность, целостность, доступность). МУ КБ ИИ также определяет объект атаки – то, на что именно может направить усилия злоумышленник. В их числе, например, обучающие датасеты или открытые модели.

Методология выступает практическим инструментом для команд, которым необходим структурированный подход к анализу рисков КБ ИИ. Она объединяет экспертизу внутренних специалистов с лучшими мировыми стандартами кибербезопасности (OWASP, MITRE, NIST и другие).

Соблюдение рекомендаций Кодекса этики

5.5. Обеспечение информационной безопасности

МУ КБ ИИ разделяет специфические угрозы ИИ по всем этапам жизненного цикла: от сбора и подготовки данных до инфраструктуры самой модели, приложений и ИИ-агентов. Документ предлагает готовую классификацию для структурирования и документирования угроз и позволяет системно выявлять и оценивать их на каждом шаге разработки и внедрения.



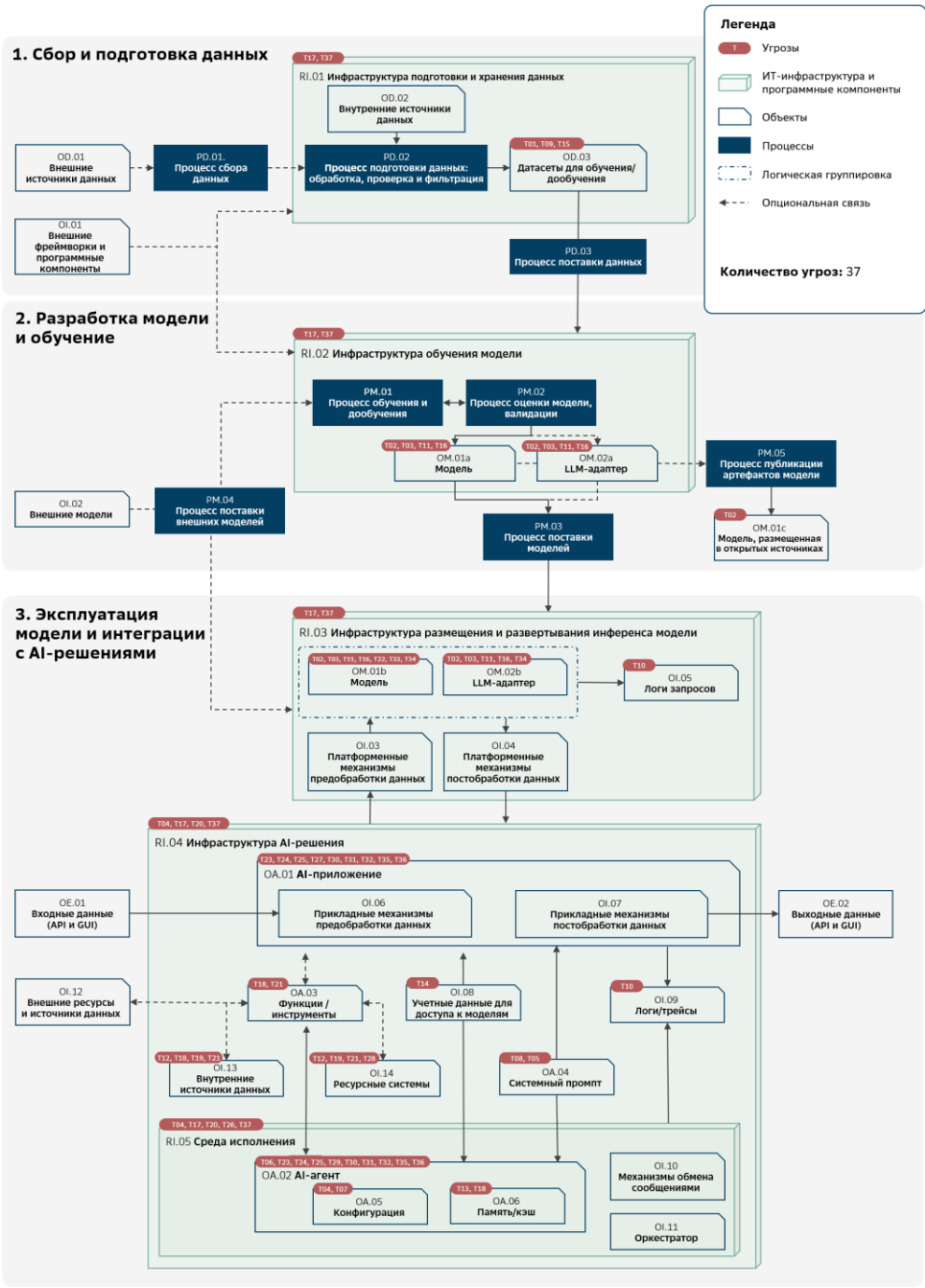
Модель угроз доступна
всем желающим

Модель угроз кибербезопасности искусственного интеллекта



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

Обобщенная схема объекта защиты и актуальных угроз для КБ ИИ на этапах
жизненного цикла



Модель угроз кибербезопасности искусственного интеллекта



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

Ключевые особенности модели угроз кибербезопасности ИИ:

- разделение на угрозы кибербезопасности ИИ и способы их реализации. Угрозы описывают негативные последствия, способы реализации – конкретные действия нарушителя. Это позволяет выстраивать многоуровневую защиту: блокировать техники атак и минимизировать последствия;
- привязка к цепочке атаки MITRE ATLAS. Способы реализации классифицированы по тактикам (от Reconnaissance до Exfiltration), угрозы – по типу последствий (Impact). Это обеспечивает единый язык с экспертами кибербезопасности;
- детализация нарушителей. Введены категории внешних и внутренних нарушителей с разным уровнем технических возможностей и осведомленности нарушителя о системе – от white-box (полное знание архитектуры, данных и параметров) до black-box (только публичные интерфейсы), с промежуточным grey-box (частичное знание), что позволяет точнее оценивать реализуемость угроз для каждого типа нарушителя;
- расширенный перечень объектов защиты (таких как системные промпты, LLM-адаптеры, память и кэш ИИ-агентов, механизмы обмена сообщениями в мультиагентных системах, среда исполнения ИИ-агентов, а также ресурсные системы ИИ-агентов (внешние сервисы и API). В модель включены компоненты, которые являются критическими точками атак. Их выделение позволяет покрыть угрозы, связанные с новой архитектурой ИИ-решений, и дать командам разработки и кибербезопасности конкретные ориентиры для выработки мер защиты.

В документе представлены 37 угроз кибербезопасности ИИ и 51 способ их реализации. Для каждого элемента указаны: описание, нарушаемые свойства информации и последствия, связанные объекты, этапы жизненного цикла, виды моделей (PredAI/GenAI) и конкретные роли сотрудников (например, владелец данных, разработчик (владелец) модели, разработчик (владелец) ИИ-решения, владелец ИТ-инфраструктуры, эксперт по кибербезопасности), отвечающие за меры защиты. Также приводятся виды нарушителей, способные реализовать каждую угрозу и каждый способ реализации.

Модель предназначена для разработчиков ИИ-решений, MLOps/DevOps-инженеров, архитекторов, экспертов по кибербезопасности и руководителей, отвечающих за управление рисками при внедрении ИИ-решений.

Модель угроз кибербезопасности искусственного интеллекта



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

5.6. Обеспечение конфиденциальности информации

МУ КБ ИИ фокусируется на специфических каналах, через которые из систем ИИ могут утекать данные ограниченного доступа:

- на этапе обучения: риск неконтролируемой загрузки секретов в тренировочные наборы данных;
- на этапе эксплуатации: утечки через побочные каналы и служебные системы. Например, компрометация журналов (логов) запросов к API, а также кража данных из внутренних источников – баз данных или векторных хранилищ систем RAG.

Противостоять несанкционированному распространению информации помогает комплекс технических мер:

- валидация логов: автоматическая очистка журналов обращений от чувствительных данных перед их сохранением или анализом.
- ограничение доступа к RAG: ограничение прав на чтение фрагментов корпоративной базы знаний, которую использует генеративный ИИ для поиска ответов.



Виртуальные ассистенты

На финансовом рынке виртуальные ассистенты все чаще применяются в клиентском сервисе. Организации интегрируют их в мобильные приложения и сайты, создавая альтернативу классическим каналам взаимодействия через чат поддержки или телефонный разговор. Сегодня в России уже 56% организаций, применяющих и тестирующих ИИ, используют его для взаимодействия с клиентами через голосовые помощники и чат-боты⁶

Важно учитывать

Виртуальные ассистенты позволяют оказывать клиентам доступную, быструю и персонализированную поддержку. В самой организации они снижают нагрузку на операторов и операционные затраты в целом. При этом ориентированность таких ИИ-сервисов на клиента требует учета специфических рисков.

МОГУТ УХУДИТЬ КЛИЕНТСКИЙ ОПЫТ

Низкое качество работы виртуальных помощников, в том числе неспособность предоставить нужную информацию и «хождение кругами», приводят к негативному восприятию ИИ-сервисов⁷. При этом клиентам важно иметь возможность обсудить свой вопрос с человеком, особенно если чат-боту не удастся найти решение самостоятельно: более 60% пользователей предпочли бы общаться с сотрудником организации, а не виртуальным ассистентом⁸.

По статистике, **более половины россиян** сталкиваются с ситуациями, когда они не могут определить, с кем они взаимодействуют – сотрудником или ИИ⁹

МОГУТ ДОПУСКАТЬ ОШИБКИ

Виртуальные ассистенты, как и другие ИИ-решения, предоставляют пользователям вероятностные ответы, которые могут быть неточными или ошибочными. Это может вводить потребителя в заблуждение относительно доступных для него условий оказания финансовых услуг.

Следуя совету чат-бота, клиент компании Air Canada приобрел билет и рассчитывал вернуть его частичную стоимость в будущем. Однако при попытке получить компенсацию представитель авиакомпании сообщил, что описанный ИИ-ассистентом порядок на самом деле не предусмотрен правилами авиакомпании. Установив факт введения потребителя в заблуждения, суд Канады обязал авиакомпанию выплатить компенсацию¹⁰

⁶ [Применение искусственного интеллекта на финансовом рынке: текущий статус и условия дальнейшего развития](#). Банк России. Ноябрь 2025.

⁷ [Chatbots in consumer finance](#). Consumer Financial Protection Bureau. June 2023.

⁸ [Gartner Survey Finds 64% of Customers Would Prefer That Companies Didn't Use AI For Customer Service](#). Gartner. July 2024.

⁹ [Чат-боты перестают удивлять: каждый третий россиянин оценивает их работу негативно](#) НАФИ. Сентябрь 2025.

¹⁰ [Moffatt v. Air Canada](#). Civil Resolution Tribunal. February 2024.

МОГУТ НЕ СПРАВИТЬСЯ СО СЛОЖНЫМ ВОПРОСОМ

ИИ-ассистенты, как правило, обучены на типовых сценариях и при возникновении нестандартной ситуации или необходимости учета всей истории взаимодействия клиента с финансовой организацией могут оказаться неспособны предложить корректное решение.

МОГУТ СТАНОВИТЬСЯ ИНСТРУМЕНТОМ МИССЕЛИНГА И НАВЯЗЫВАНИЯ КЛИЕНТАМ ФИНАНСОВЫХ УСЛУГ

Это может происходить как намеренно, например при разработке ИИ-ассистента или выборе формата диалогового окна, так и вследствие проявления свойств ИИ. Такие практики могут включать неоднократное предложение приобрести продукт без прямого вопроса клиента, игнорирование в советах нейтральных опций или позиционирование платного решения как естественный следующий шаг ¹¹.

Лучшие практики

Кодексом этики предусмотрен ряд мер, направленных на учет таких специфических рисков и обеспечение интересов клиентов при использовании виртуальных ассистентов (в том числе оценка удовлетворенности клиентов и мониторинг их жалоб, информирование о применении ИИ и раскрытие информации о его рисках, обеспечение возможности переключить общение на оператора). На практике перечисленные меры могут быть реализованы по-разному.

УВЕДОМЛЕНИЕ О ПРИМЕНЕНИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Уведомлять клиента о применении ИИ необходимо всегда, если только применение ИИ неочевидно из обстоятельств.



Даже если применение ИИ очевидно, клиент в любом случае должен получить ответ на свой прямой вопрос о том, с кем он взаимодействует

Следует проактивно раскрывать клиенту информацию о том, что диалог с ним ведет ИИ. Факт применения ИИ возможно обозначить разными способами:

- **В названии**

Например, «ИИ-помощник», «Виртуальный ассистент» и т. п.

¹¹ [Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design](#). Center for Democracy and Technology. May 2026.

- **В отдельном дисклеймере**

В интерфейсе сервиса или в статичном сообщении в диалоговом окне (например, «Сервис работает на базе ИИ»)

- **В сообщении клиенту:**

- в ответ на его прямой вопрос о природе собеседника в любое время сессии;

 **Ассистент в СберБанк Онлайн**

п. 4.2 

Пример ответа на прямой вопрос клиента

- сопровождение ответа клиенту указанием на применение ИИ.

 **Нейропомощник**

п. 4.2 

Пример визуальной метки о применении ИИ



Важно учитывать разные варианты вопроса клиента о том, с кем он взаимодействует. Например, чтобы клиент мог получить ответ как на вопрос «взаимодействую ли я с ИИ?», так и на вопрос «кто ты?»



При этом рекомендуется избегать случаев, когда для ознакомления с информацией о применении ИИ от пользователя требуется переходить в дополнительные разделы интерфейса. Например, недостаточно разместить такую информацию только в разделе часто задаваемых вопросов

Факт применения ИИ может обозначаться с помощью стилистических решений (например, возможно использовать маркировку, анимированный аватар, разделять ответы ИИ и сотрудника разными цветами, предусмотреть роботизированный тембр для голосового помощника и т. п.). Такие стилистические паттерны снижают когнитивную нагрузку на клиента, упрощая его восприятие и повышая удобство интерфейса. Однако стилистические решения следует рассматривать в качестве дополнительных (вторичных) инструментов, не заменяющих более прямого уведомления клиента о применении ИИ, исключающего возможные неоднозначные интерпретации.

УВЕДОМЛЕНИЕ О РИСКАХ И УСЛОВИЯХ ПРИМЕНЕНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

При предупреждении о рисках и условиях применения ИИ уровень информирования может отличаться в зависимости от содержания взаимодействия пользователя с ИИ.

- **Для высокорисковых запросов оправдано размещение отдельного предупреждения о рисках и условиях применения ИИ**
(например, инвестиционные советы, осуществление платежа, блокировка карты);
- **Для низкорисковых запросов достаточно размещение соответствующей информации в пользовательском соглашении или разделе часто задаваемых вопросов**
(например, информационные запросы об адресах банкоматов, графике работы офисов).

Клиенту нужно предоставлять полную и достоверную информацию о рисках ИИ и условиях его применения. На основе этой информации клиент должен иметь возможность принять информированное решение об использовании продуктов и услуг организации и осуществить возможные меры для учета раскрытых рисков ИИ (например, перепроверить корректность сгенерированной чат-ботом информации). При этом критерии, на основании которых определяется полнота раскрываемой информации, зависят от конкретного случая применения ИИ.

точка
банк ИИ-Ассистент

п. 4.3

Пример дисклеймера о рисках ИИ

ПЕРЕКЛЮЧЕНИЕ НА ОПЕРАТОРА

Клиенту следует обеспечить возможность переключить общение на оператора.

Переключение должно быть максимально бесшовным и наименее трудозатратным для клиента. Это подразумевает отсутствие у клиента необходимости использовать альтернативные способы связи с оператором (например, дополнительно обращаться в общую службу поддержки по телефону или адресу электронной почты организации).

Могут быть предусмотрены разные механизмы переключения на оператора:

- Переключение может быть предусмотрено **в интерфейсе**, позволяющем перевести общение в отдельный канал или подключить сотрудника к текущему диалогу;
- **Отдельное сообщение клиенту** о возможности перевести общение на оператора (например, вопрос: «Хотите перейти на сотрудника или продолжим общение?») и переключение в случае положительного ответа;

- Клиент уведомляется о возможности перейти на взаимодействие с оператором **в начале диалога**, и переключение происходит в ответ на соответствующий запрос в любой момент;



Важно учитывать разные варианты запроса клиента с просьбой переключить на оператора (например, «Оператор», «Хочу говорить с человеком», «Подключи специалиста» и т. п.)

- **Автоматическое переключение**, если вопрос не может быть качественно решен ИИ (например, если ИИ не знает ответа, диалог зашел в тупик, клиент неоднократно повторяет свой запрос и т. п.).



При переключении диалога на оператора следует передать ему резюме с запросом клиента об осуществленном пути его обработки ИИ. Это обеспечит улучшение клиентского опыта, поскольку клиент не будет вынужден повторять свой вопрос и цепочку рассуждений заново

точка банк ИИ-Ассистент

п. 2.3 

Пример переключения на оператора с саммаризацией

СОВЕРШЕНСТВОВАНИЕ СЕРВИСА И УЧЕТ ОБРАТНОЙ СВЯЗИ

Повысить качество оказания услуг с помощью ИИ позволяет анализ обратной связи от клиентов в форме оценки их удовлетворенности или их жалоб, а также организация внутреннего контроля качества оказываемых с помощью ИИ услуг.

- Оценить взаимодействие на основе обратной связи от клиента возможно следующими способами:
 - предложить клиенту оценить качество обслуживания по определенной шкале (например, в виде «звездочек» от 1 до 5 или знаков «палец вверх» и «палец вниз»);
 - задать отдельный вопрос о том, что не устроило клиента в работе сервиса, и предложить готовые варианты ответов (в случае отрицательной оценки);



Вопросы не должны быть навязчивыми, следует обеспечить максимальную краткость и простоту формулировок для удобства заполнения. Отвечая на вопросы, клиент должен потратить минимум времени

**Точка
банк** ИИ-Ассистент

п. 2.2 

Пример оценки взаимодействия

- проанализировать жалобы клиентов на сервис;
- Проактивная проверка качества сессий:
 - предусмотреть внутренний аудит качества: осуществлять выборочную проверку сессий с участием специалистов службы контроля качества с последующим анализом типичных сценариев и регулярным обновлением модели;

А Нейропомощник

п. 2.2 

Пример оценки взаимодействия

- организовать комплексный контроль качества сессий, предусматривающий автоматизированный мониторинг показателей, обработку обращений и выборочную проверку качества работы модели специалистами.

 Ассистент в СберБанк Онлайн



Пример проактивной оценки качества сессий

Ассистент в СберБанк Онлайн



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

ПРОЗРАЧНОСТЬ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 2.2. Повышение качества оказания услуг
- пункт 2.3. Предоставление клиентам возможности отказаться от взаимодействия с применением ИИ
- пункт 4.2. Информирование о применении ИИ

О сервисе

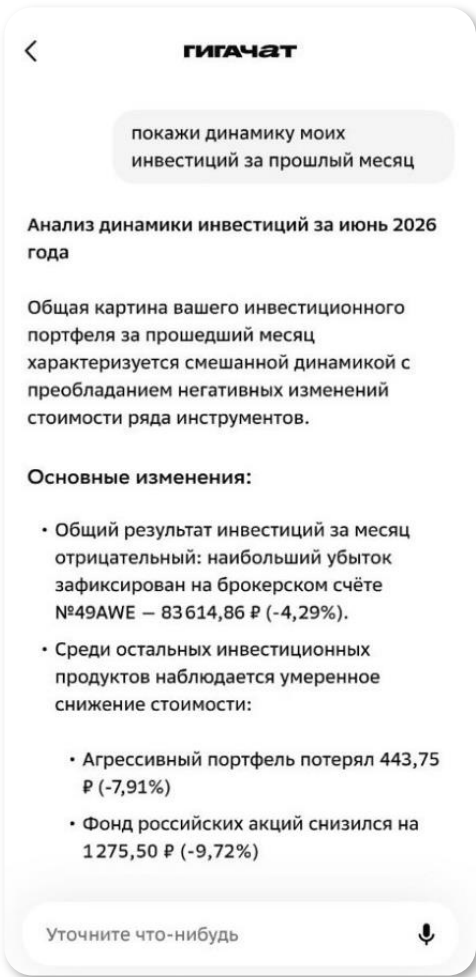
Ассистент встроен в поиск в СберБанк Онлайн и позволяет решать многие сценарии без дополнительных переходов. При этом сервис является не только навигационным инструментом для клиентов, но и позволяет им решать широкий спектр задач.

Ассистент может отвечать как на общие вопросы, например, про условия продуктов, так и на персонализированные, например, о счетах клиента.

При этом не требуется идеальная точность в формулировании вопросов: Ассистент может интерпретировать написанное и при необходимости уточнять.

Помимо этого, Ассистент может проанализировать траты, подсветить изменения в портфеле или объяснить на какие именно операции начисляется кэшбек, а на какие – нет и почему.

Таким образом, Ассистент расширяет возможности обычного интерфейса более гибкой аналитикой, которую можно обсудить. Если требуется продолжить сценарий, перейдя от получения информации к совершению действий, это можно сделать сразу в чате.



Технологии

Тип модели

Генеративный ИИ

Агентский ИИ

Модальность

Аудио

Естественный текст

Методы и техники

RAG

Ассистент в СберБанк Онлайн



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

ПРОЗРАЧНОСТЬ

Соблюдение рекомендаций Кодекса этики

2.2. Повышение качества оказания услуг

Несмотря на то, что чат-бот не принимает решений, влияющих на финансовые параметры или условия обслуживания, организацией предусмотрен внутренний аудит качества: осуществляется выборочная проверка сессий с участием специалистов службы контроля качества с последующим анализом типичных сценариев и регулярным обновлением модели.

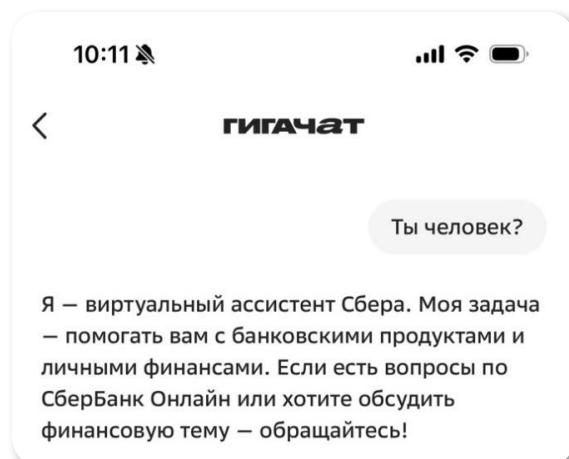
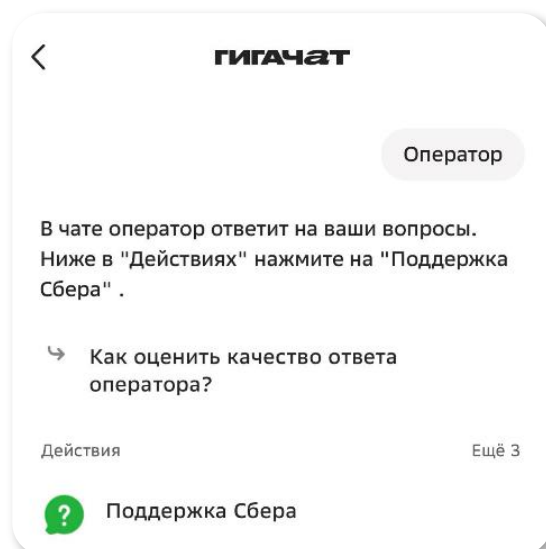
2.3. Предоставление клиентам возможности отказаться от взаимодействия с применением искусственного интеллекта

По прямому текстовому запросу пользователю предоставляется ссылка на чат поддержки, где возможно связаться с сотрудником.

4.2. Информирование о применении искусственного интеллекта

Ассистент не позиционируется как человек – внешний вид, стиль общения и структура меню однозначно указывают на его цифровую природу. В случае прямого вопроса пользователя он получает ответ о том, что общается с ИИ.

Дополнительно предусмотрено интерактивное меню, работающее в типовом формате, привычном для цифровых интерфейсов поддержки.



ИИ-Ассистент

ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

ПРОЗРАЧНОСТЬ

Точка

банк для предпринимателей

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 2.2. Повышение качества оказания услуг
- пункт 2.3. Предоставление клиентам возможности отказаться от взаимодействия с применением ИИ
- пункт 4.2. Информирование о применении ИИ
- пункт 4.3. Раскрытие информации об ИИ

О сервисе

ИИ-Ассистент отвечает на вопросы клиентов про сервисы экосистемы Точки и помогает с бизнес задачами, в том числе с анализом доходов и расходов, налоговой нагрузки и дебиторской задолженностью.

ИИ-Ассистент доступен клиентам в личном кабинете. По запросу пользователя ИИ-Ассистент детализирует данные по периодам, категориям и контрагентам, помогает разобраться в условиях продуктов и сервисов банка, а также консультирует по вопросам регистрации бизнеса: выбору организационно-правовой формы (ИП или ООО), системы налогообложения и кодов видов экономической деятельности.



Технологии

Тип модели

Генеративный ИИ

Модальность

Естественный текст

Структурированные данные

Методы и техники

RAG

Поддержка

ИИ-Ассистент

В чем преимущества патента и для кого он полезен?

ИИ-Ассистент

Преимущества патента и кому он полезен

Патент (ПСН) полезен ИП, у которых есть «патентные» виды деятельности в конкретном регионе и кому выгоднее платить **фиксированную стоимость патента**, чем налог с фактических доходов.

Преимущества ПСН

- Фиксированный налог — вы платите **стоимость патента**, а не процент от реально полученного дохода по этой деятельности
- Можно брать патент не на весь бизнес, а только на отдельный вид деятельности
- Можно совмещать с другими режимами (например, вести часть деятельности на УСН, часть — на ПСН): <https://allo.tochka.com/sno-ip>

Спросить ИИ-Ассистента

ИИ-Ассистент может ошибаться, рекомендуем проверять ответы.
[Условия использования](#)

ИИ-Ассистент

ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

ПРОЗРАЧНОСТЬ

Точка

банк для предпринимателей

Соблюдение рекомендаций Кодекса этики

2.2. Повышение качества оказания услуг

Клиент может оценить результаты взаимодействия с ИИ-Ассистентом. В случае неудовлетворительной оценки клиент может уточнить, что именно ему не понравилось в работе сервиса, из предложенных вариантов ответа.

2.3. Предоставление клиентам возможности отказаться от взаимодействия с применением искусственного интеллекта

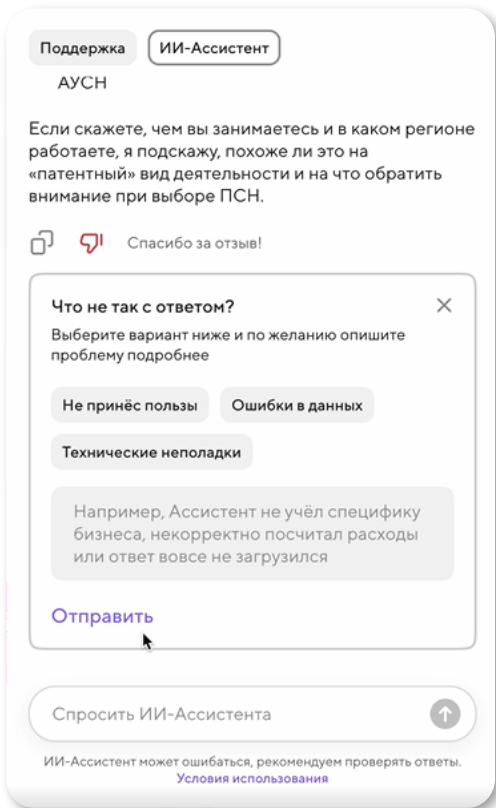
Интерфейс ИИ-Ассистента сделан таким образом, что у клиента сразу есть выбор: задать вопрос ИИ-Ассистенту или обратиться в чат «Поддержка» за помощью оператора. При такой реализации клиенту не нужно повторять ИИ-помощнику просьбу перевести его на оператора.

4.2. Информирование о применении искусственного интеллекта

Название сервиса (маркировка) «ИИ-Ассистент» формирует у пользователя четкое понимание используемой технологии.

4.3. Раскрытие информации об искусственном интеллекте

Закрепленный дисклеймер уведомляет пользователя о том, что ответы ИИ-Ассистента могут быть неполными и неточными, подсвечивая риск возможных ошибок.



Нейропомощник

Альфа Банк

ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

ПРОЗРАЧНОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 2.2. Повышение качества оказания услуг
- пункт 2.3. Предоставление клиентам возможности отказаться от взаимодействия с применением ИИ
- пункт 4.2. Информирование о применении ИИ
- пункт 4.3. Раскрытие информации об ИИ
- пункт 5.6. Обеспечение конфиденциальности информации
- пункт 6.9. Минимизация выявленных рисков ИИ

О сервисе

Нейропомощник помогает клиента предпринимателям быстро находить нужную информацию про сервисы и услуги Альфа-Банка, ориентироваться в интерфейсе, получать выписки, реквизиты, анализировать доходы и расходы.

Нейропомощник доступен в личном кабинете юридических лиц в приложении и на сайте. По текстовым запросам клиента определяет его намерение, извлекает релевантную информацию из внутренних баз знаний и формирует понятный ответ. В ряде сценариев помогает клиенту найти операцию, реквизиты контрагента.



Технологии

Тип модели

Генеративный ИИ

Модальность

Естественный текст

Методы
и техники

RAG

MCP

Дообучение (fine-tuning)

Нейропомощник

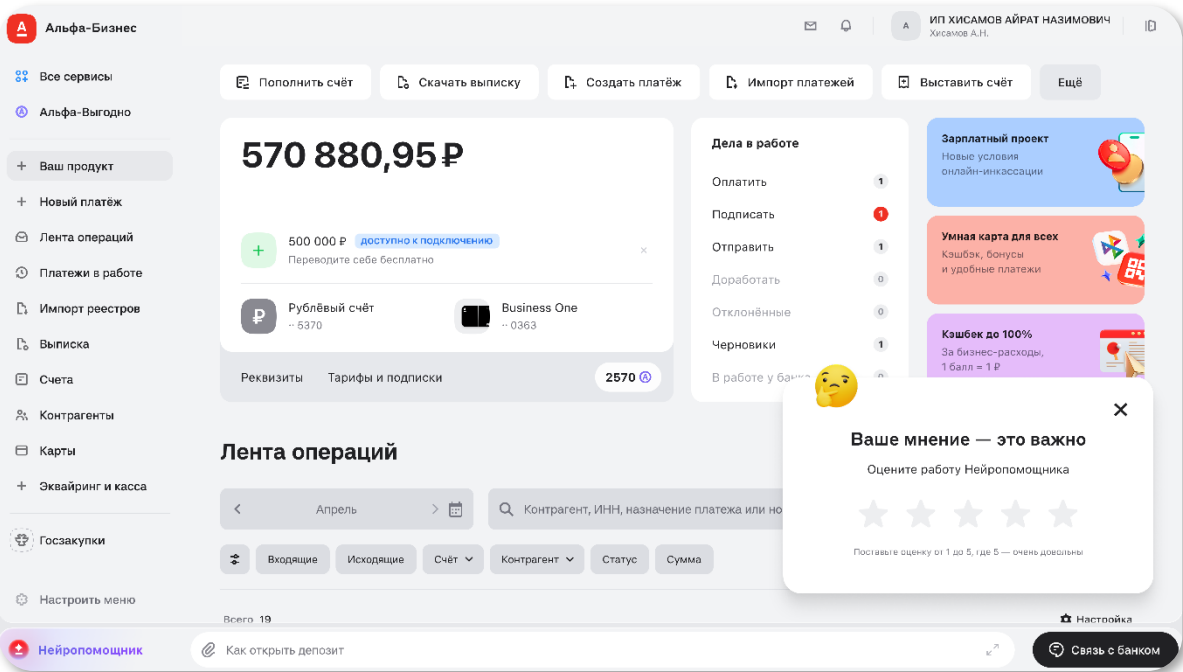
Альфа Банк

ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

ПРОЗРАЧНОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ



Соблюдение рекомендаций Кодекса этики

2.2. Повышение качества оказания услуг

Проводится регулярная оценка уровня удовлетворенности клиентов сервисом. Соответствующие данные автоматически собираются в аналитический дашборд, с помощью которого специалисты службы качества отслеживают динамику показателей и выявляют отклонения от установленных порогов.

Факт получения низкой оценки также выявляется автоматически, а информация об этом передается ответственным сотрудникам для отработки. Такой непрерывный цикл обратной связи обеспечивает постоянный контроль качества услуг, предоставляемых при помощи ИИ, и позволяет своевременно повышать их эффективность для корпоративных клиентов.

Нейропомощник

Альфа Банк

ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

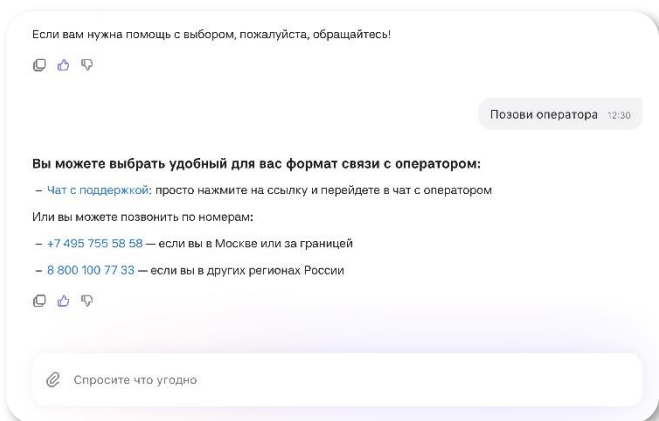
ПРОЗРАЧНОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

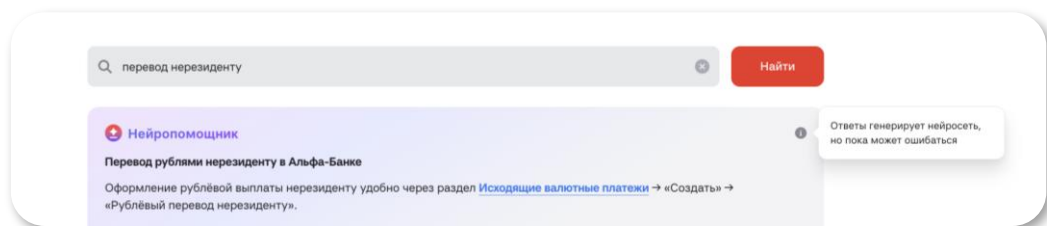
2.3. Предоставление клиентам возможности отказаться от взаимодействия с применением искусственного интеллекта

Клиент может в любой момент прекратить взаимодействие с ИИ и переключиться на общение с оператором по запросу.



4.2. Информирование о применении искусственного интеллекта

Пользователь явно понимает, что взаимодействует не с сотрудником банка, а с цифровым ИИ-помощником. Интерфейс, стиль коммуникации и сценарии взаимодействия прямо указывают на использование технологий ИИ. Также в интерфейсе сервиса каждый ответ сопровождается визуальной меткой (маркировкой), указывающей, что информация сформирована системой ИИ.



4.3. Раскрытие информации об искусственном интеллекте

В сервисе предусмотрено предупреждение о вероятности ошибок в ответах, сгенерированных нейросетью.

Нейропомощник

Альфа Банк

ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

ПРОЗРАЧНОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

5.6. Обеспечение конфиденциальности информации

Все данные, передаваемые в процессе взаимодействия с сервисом, проходят сквозное шифрование, маскируются и хранятся в защищенном банковском контуре, доступ к которому имеют только уполномоченные сотрудники.

При подготовке обучающих наборов данных применяется методика обезличивания, исключающая возможность восстановления персональных или корпоративных сведений.

Также система реализует политики контроля доступа, включая многофакторную аутентификацию и журналы аудита всех операций с конфиденциальной информацией.

6.9. Минимизация выявленных рисков искусственного интеллекта

Несмотря на высокий уровень автоматизации, ИИ-помощник не принимает юридически значимых решений без участия человека. В сложных случаях клиент может быть переведен на сотрудника банка, а критичные действия остаются под контролем специалиста.



Финансовые помощники в инвестициях

Финансовые ИИ-помощники становятся новым форматом сопровождения частного инвестора при принятии им инвестиционных решений. Такие сервисы помогают клиентам ориентироваться в рыночной информации, разбираться в финансовых терминах, оценивать возможные стратегии и подбирать инструменты с учетом индивидуальных целей и риск-профиля

Важно учитывать

Финансовые ИИ-помощники делают информацию доступнее для клиентов, обобщая различные данные и помогая сравнить инвестиционные продукты. Такие сервисы могут снижать барьер входа для начинающих инвесторов и повышать качество клиентского опыта за счет персонализации и интерактивного формата.

Финансовых ИИ-помощников в зависимости от уровня функциональности можно разделить на следующие типы:

- **консультационные ИИ-помощники**, к которым относятся информационные сервисы, предоставляющие клиенту общие данные в справочном формате, и программы автоконсультирования, формирующие и предоставляющие ИИР;
- **ИИ-помощники, обеспечивающие совершение сделок**, включая программы автоследования, которые преобразовывают предоставленную ИИР в поручение брокеру на совершение сделки с финансовыми инструментами, а также автономные ИИ-агенты.

МОГУТ ОСЛАБЛЯТЬ СПОСОБНОСТЬ КЛИЕНТА КРИТИЧЕСКИ ОЦЕНИВАТЬ ИНВЕСТИЦИОННЫЕ РЕШЕНИЯ

Формат общения с финансовым ИИ-помощником, приближенный к взаимодействию с человеком-консультантом, может повысить доверие клиента к предоставленным рекомендациями и снизить его критическое восприятие.

В одном из исследований приведены результаты онлайн-эксперимента с участием 1256 респондентов, которые взаимодействовали с ИИ-системой поддержки инвестиционных решений, различавшейся степенью имитации общения с человеком-консультантом, включая наличие имени, аватара и особенности стиля коммуникации. По итогам исследования установлено, что восприятие ИИ как более «человечного» влияет на оценку риска инвестиционной рекомендации.

Так, пользователи с низким уровнем финансовых знаний склонны воспринимать рекомендацию ИИ как менее рискованную из-за того, что более антропоморфная форма подачи информации вызывает больше доверия. Напротив, пользователи с высоким уровнем финансовых знаний были менее подвержены этому эффекту¹²

¹² Manuele Reani, Xiangyang He, Zuolan Bao. [Anthropomorphism on Risk Perception: The Role of Trust and Domain Knowledge in Decision-Support AI](#). February, 2026.

Если сервис уверенно отвечает на вопросы и прямо рекомендует ту или иную стратегию или продукт, клиент может воспринимать такие советы как объективные, точные и достаточные для принятия решения. Но это не означает, что клиент полностью осознает все последствия инвестирования. Например, клиент может излишне полагаться на объяснения и аналитические выводы финансового помощника, не перепроверяя предоставленную информацию.

МОГУТ ОШИБАТЬСЯ ИЛИ ОПИРАТЬСЯ НА УСТАРЕВШИЕ ДАННЫЕ

Финансовые ИИ-ассистенты работают с динамичными данными: котировками, новостями, отчетностью, аналитическими материалами и макроэкономическими показателями. Ошибка в источнике, задержка обновления, неверная интерпретация отчетности или «галлюцинация» модели могут привести к формированию убедительного, но некорректного вывода.

В 2023 году во время демонстрации Bing AI от Microsoft допустил ряд ошибок. Так, сервис должен был анализировать финансовую отчетность компаний Gap и Lululemon, но выдал некорректные финансовые показатели и выводы. В частности, он неверно указал операционную маржу обеих компаний и ошибочно интерпретировал прогноз продаж Gap¹³

МОГУТ СТАНОВИТЬСЯ ИНСТРУМЕНТОМ МИССЕЛИНГА И НАВЯЗЫВАНИЯ ПРОДУКТОВ

Финансовый ИИ-помощник может создавать риск мисселинга, оказывая как прямое, так и косвенное влияние на выбор клиента. Такой риск может проявляться не только в явном предложении приобрести конкретный продукт, но и в архитектуре сервиса: логике ранжирования, подборе примеров, формулировках ответов, составе доступных стратегий и сценариях продолжения диалога.

Отчет Международной организации комиссий по ценным бумагам 2025 года о цифровых практиках вовлечения клиентов показывает, что цифровой интерфейс инвестиционного сервиса сам по себе может влиять на поведение розничного инвестора. К таким практикам относятся в том числе уведомления, наджинг и геймификация, которые могут намеренно или ненамеренно вовлекать клиента в принятие инвестиционных решений. Подобные механизмы могут побуждать клиента торговать чаще, выбирать более рискованные продукты или менять инвестиционную стратегию без полного понимания рисков¹⁴

¹³ [Google wasn't the only one to make errors in its A.I. demo. Analysis finds Microsoft's Bing flubbed a string of financial figures](#). Fortune. February, 2023.

¹⁴ [Digital Engagement Practices \(DEPs\)](#). IOSCO. May 2025.

Лучшие практики

Кодексом этики предусмотрен ряд мер, которые направлены на минимизацию рисков применения финансовых ИИ-помощников клиентами, включая информирование клиентов о применении ИИ, раскрытие рисков и условий применения ИИ и проверку качества данных. На практике эти меры могут быть реализованы следующим образом.

ИНФОРМИРОВАТЬ КЛИЕНТА

В отличие от обычных виртуальных ассистентов финансовые ИИ-помощники в большинстве случаев консультируют клиентов по более чувствительным финансовым вопросам. В связи с этим особенно важно уведомлять клиентов о применении ИИ, а также сообщать о рисках следования советам таких сервисов. При этом чем выше влияние ИИ на инвестиционное решение клиента, тем более конкретным, заметным и своевременным должно быть уведомление.



При предоставлении клиентам информации целесообразно сообщать им о том, что ответы ИИ могут содержать ошибки и неточности

Ответы отдельных сервисов могут создавать впечатление ИИР, не являясь таковыми. Если ИИ-помощник не предоставляет ИИР, необходимо прямо информировать об этом клиента.



Уведомление о том, что сервис не предоставляет ИИР, возможно сопроводить дополнительными пояснениями для клиента. Например, о том, что сервис не дает конкретные рекомендации и «советы к действию», не выбирает «лучшие» предложения, предоставляя лишь информацию, имеющую аналитический и ознакомительный характер



FinGPT

п. 4.2-4.3 [↗](#)



Инвест-ассистент

п. 4.2-4.3 [↗](#)

Примеры информирования о природе, рисках и ограничениях сервиса

ИСПОЛЬЗОВАТЬ ДОВЕРЕННЫЕ ИСТОЧНИКИ ДАННЫХ

Для инвестиционного ИИ-ассистента качество данных является ключевым условием качества ответа. Сервис должен опираться на доверенные и контролируемые источники. Это могут быть как внешние, так и внутренние источники данных, прошедшие проверку на достоверность.

При этом используемые при предоставлении справочной информации данные следует поддерживать в актуальном состоянии и обеспечивать мониторинг своевременности и полноты их поступления в контур.

Также для инвестиционного сервиса особенно важно отделять факты от интерпретаций. Поэтому ответы клиенту следует сопровождать ссылками на источник информации, а аналитические выводы – указанием на допущения и ограничения. Если данные отсутствуют, устарели или противоречат друг другу, сервис может дополнительно сообщить об этом.



Снизить риск «галлюцинаций» возможно за счет того, что ИИ формирует ответ на основе данных не из собственной памяти, а из утвержденных источников (подход RAG). При этом сами источники и правила их обновления должны находиться под контролем финансовой организации



FinGPT

п. 6.9



Инвест-ассистент

п. 6.9

Пример использования доверенных источников данных

ПРЕДУПРЕЖДАТЬ МИССЕЛИНГ И НАВЯЗЫВАНИЕ ПРОДУКТОВ

При обращении к финансовому ИИ-помощнику за справочной информацией клиент рассчитывает на получение объективного и беспристрастного финансового совета. Если в этом случае конкретный сервис начинает продвигать собственные или партнерские продукты и услуги по логике, не основанной на их преимуществах в сравнении с представленными на рынке альтернативами, следует отдельно информировать клиента о рекламном характере подобных предложений. Это позволит обеспечить независимость сервиса и не будет вводить клиентов в заблуждение.

FinGPT



- ПРОЗРАЧНОСТЬ
- БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ
- ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

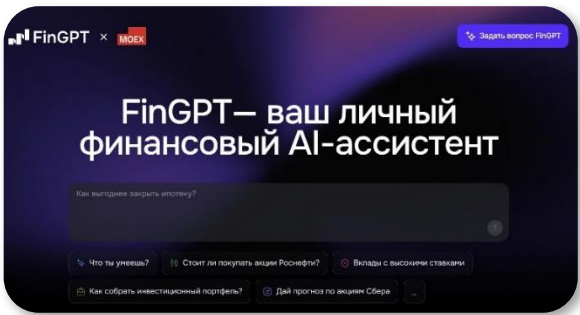
- пункт 4.2. Информирование о применении ИИ
- пункт 4.3. Раскрытие информации об ИИ
- пункт 5.3. Проверка качества ИИ
- пункт 5.4. Мониторинг качества ИИ
- пункт 5.5. Обеспечение информационной безопасности
- пункт 5.6. Обеспечение конфиденциальности информации
- пункт 6.9. Минимизация выявленных рисков ИИ

О сервисе

FinGPT – это ИИ-ассистент для розничных инвесторов, интегрированный в экосистему МосБиржи. Продукт доступен для клиентов на сайте moex.com, а также на сайте и в мобильном приложении Финуслуг.

Продукт построен как мультиагентная система, где каждый агент отвечает за свою узкую задачу (классификация и валидация запроса, сбор данных, расчеты, подготовка ответа и другое). Каждый агент работает на базе больших языковых моделей. В рамках каждого агента используется подход аугментации информации: модель не дообучается, а получает релевантный контекст в момент формирования ответа.

Учитывая массовый характер продукта и чувствительность финансовой тематики, в FinGPT реализован сквозной контур обеспечения безопасности, надежности и эффективности.



Технологии

- | | |
|------------------|---------------------------------|
| Тип модели | Генеративный ИИ
Агентский ИИ |
| Модалность | Естественный текст |
| Методы и техники | RAG (аугментация контекста) |

FinGPT



ПРОЗРАЧНОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Соблюдение рекомендаций Кодекса этики

4.2. Информирование о применении искусственного интеллекта

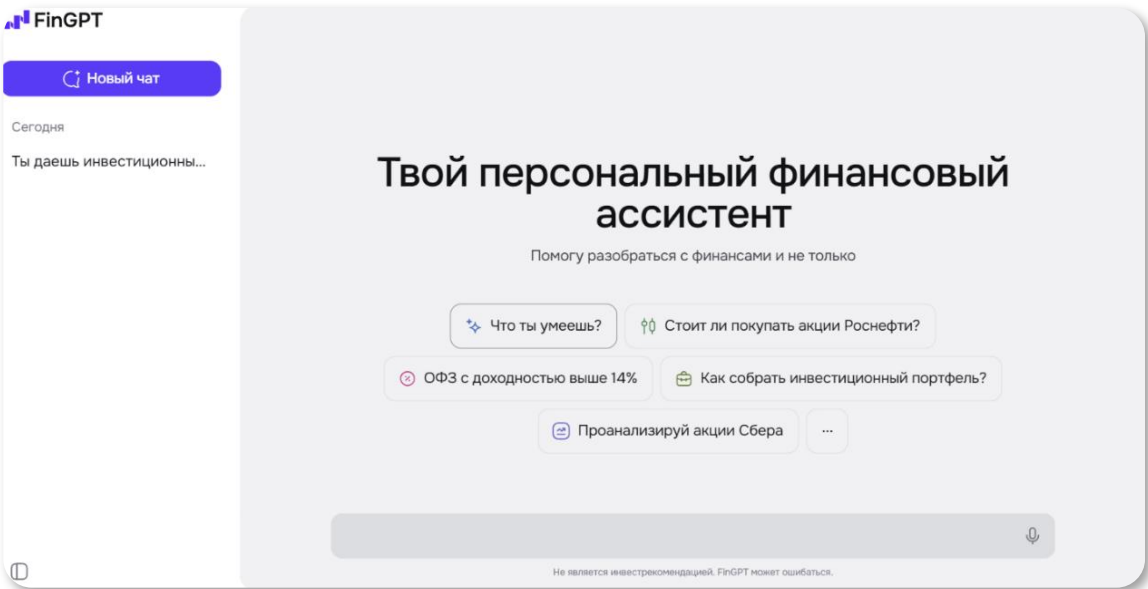
Название сервиса четко сообщает клиентам о применении ИИ. Интерфейс, стиль общения и функционал сервиса формируют у пользователя понимание, что он взаимодействует не с человеком.

**FinGPT – ваш личный
финансовый AI-ассистент**

4.3. Раскрытие информации об искусственном интеллекте

Пользователь неоднократно уведомляется о том, что ответы FinGPT не являются инвестиционной рекомендацией и остается риск ошибок:

- в закрепленном дисклеймере внизу окна чата на основной странице сервиса;
- в закрепленном дисклеймере внизу окна чата на странице чата;



FinGPT



- ПРОЗРАЧНОСТЬ
- БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ
- ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

- в рубрике «Часто задаваемые вопросы»;

Частые вопросы

Дает ли FinGPT инвестиционные рекомендации?

Нет. Сервис предоставляет только информационные материалы, которые не являются индивидуальной рекомендацией.

- в ответ на непосредственный вопрос пользователя в чате.

FinGPT

Новый чат

Сегодня

Ты даешь инвестиционны...

Ты даешь инвестиционные рекомендации?

Вот что я нашел на сайте Финуслуг и в интернете:

Нет, я не даю инвестиционные рекомендации. Я являюсь информационно-аналитическим ассистентом, моя задача — предоставлять сырые данные, статистику и факты, а также давать образовательные пояснения по финансовым темам.

Я не подбираю решения «под пользователя», не делаю выводы о том, какие варианты являются «лучшими», и не указываю, что именно стоит купить или продать. Все ответы носят исключительно ознакомительный и аналитический характер.

Подобный подход распространён в финансовой индустрии: многие платформы и сервисы отдельно подчёркивают в своих дисклеймерах, что предоставляемая информация не является советом к действию.

Примеры позиций финансовых платформ

Ниже приведена таблица с выдержками из дисклеймеров различных сервисов, которые обозначают свой контент как информационный или образовательный.

Компания	Основная позиция в дисклеймере
Stalder	Мы являемся зарегистрированным инвестиционным

Твой вопрос...

Не является инвестрекомендацией. FinGPT может ошибаться.

FinGPT



ПРОЗРАЧНОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

6.9. Минимизация выявленных рисков искусственного интеллекта

В FinGPT используются два класса данных, каждый со своим контуром контроля:

- внешние данные (новости, аналитика из открытых источников). Используется курируемый перечень доверенных источников, который поддерживается в актуальном состоянии. Каждый источник проходит проверку на достоверность публикуемой информации и на соответствие законодательству Российской Федерации. Источники, не прошедшие проверку, в контур не включаются;
- данные МосБиржи (биржевые котировки, корпоративная информация из Центра Корпоративной Информации, справочники эмитентов и инструментов). Дополнительно к штатным процедурам контроля качества корпоративных данных проводится проверка полноты и выявление неточностей применительно к задачам FinGPT, а также мониторинг своевременности и полноты поступления данных в контур ассистента.

Важная особенность подхода: модель не обучается на указанных данных – они передаются в модель в качестве контекста (аугментация). Это снижает риски «запоминания» некорректных данных моделью и позволяет оперативно актуализировать информационную базу

5.3. Проверка качества искусственного интеллекта

Для каждого узла/агента мультиагентной системы разработаны:

- собственный тестовый датасет, формируемый в том числе на основе реальных запросов пользователей и поддерживаемый в актуальном состоянии;
- набор метрик качества, отражающий специфику задачи конкретного агента.

При любом изменении компонента проводится регрессионное тестирование, позволяющее оценить влияние изменения как на качество работы отдельного узла, так и на качество работы системы в целом. Такой подход обеспечивает контролируемое развитие продукта без деградации качества ответов.

FinGPT



ПРОЗРАЧНОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

5.4. Мониторинг качества искусственного интеллекта

В FinGPT развернута система мониторинга, обеспечивающая сквозное отслеживание работы ИИ в режиме реального времени. Система фиксирует и визуализирует:

- пары «вопрос – ответ» и промежуточные шаги рассуждений агентов;
- ошибки на уровне отдельных узлов и системы в целом;
- скорость формирования ответа;
- нагрузку на систему.

В дополнение к трейсингу регулярно проводится полный или выборочный контроль качества реальных ответов FinGPT пользователям по двум независимым контурам:

- автоматизированный контур – оценка качества ответов по методике «модель в роли судьи» (LLM-as-a-judge): независимая языковая модель оценивает ответы по установленному набору критериев;
- экспертный контур – сотрудники-тестировщики вручную вычитывают и анализируют выборку ответов, выявляя случаи, которые не покрываются автоматической оценкой.

Совмещение двух контуров обеспечивает масштабируемость мониторинга и одновременно сохраняет качественный экспертный взгляд на поведение системы.

5.5. Обеспечение информационной безопасности

FinGPT функционирует в рамках стандартных требований МосБиржи к безопасности информационных систем. Дополнительно, с учетом специфики больших языковых моделей, реализован отдельный контур защиты на уровне пользовательского взаимодействия:

- система валидации входящих пользовательских запросов, направленная на противодействие специфичным угрозам большим языковым моделям: инъекция промптов (prompt injection), джейлбрейк (jailbreak), попытки извлечения системного промпта и попытки обхода ограничений ассистента;
- выявление запросов, ответ на которые может содержать признаки нарушения законодательства Российской Федерации или причинить вред пользователю. По таким запросам система не формирует ответ и возвращает пользователю корректное уведомление.

FinGPT



ПРОЗРАЧНОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

5.6. Обеспечение конфиденциальности информации

Защита данных ограниченного доступа в FinGPT обеспечивается на двух уровнях:

- организационный уровень: команда разработки работает в соответствии с практиками информационной безопасности МосБиржи, включая систему уровней доступа к пользовательским данным с принципом минимально необходимого доступа;
- архитектурный уровень: в продукте реализованы технические меры, исключающие возможность доступа одних пользователей к данным других пользователей через механизмы локальной (в рамках сессии) и глобальной (между сессиями) памяти ассистента. Изоляция пользовательских контекстов обеспечена на уровне дизайна системы, а не только на уровне контролей доступа.

Инвест-ассистент



ПРОЗРАЧНОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 4.2. Информирование о применении ИИ
- пункт 4.3. Раскрытие информации об ИИ
- пункт 5.4. Мониторинг качества ИИ
- пункт 6.9. Минимизация выявленных рисков ИИ

О сервисе

Инвест-ассистент – виртуальный помощник, предоставляющий консультации в части инвестиций, который:

- анализирует ценные бумаги и рынок: целевые цены, консенсус-прогнозы, финансовые показатели, дивиденды и свежие новости по компаниям;
- обучает: ассистент объясняет сложные термины и помогает сделать первые шаги в инвестициях;
- помогает с выбором ценных бумаг и с навигацией пользователя по рекомендациям аналитиков на основе его запроса.

В качестве источников используются, в том числе, следующие ресурсы:

- Trading API – получает базовую информацию о ценных бумагах;
- Fireg-core и OTC Market data – данные о купонах и параметрах облигаций;
- Fireg Advisory – консенсус-прогнозы от аналитиков;
- Feed и Comments – большие аналитические разборы и комментарии о плюсах и минусах ценных бумаг.

При этом инвест-ассистент не распоряжается финансами самостоятельно – он только дает рекомендации и предоставляет информацию.

Инвест-ассистент использует 2 модели:

- первая модель выступает в роли «диспетчера». Она отвечает за построение маршрута через агента: анализирует вопрос и понимает, какой именно инструмент или навык нужно задействовать для ответа;
- вторая модель (основная) используется для генерации текстовых ответов во всех инструментах ассистента, а также для проведения оценки качества работы.



Технологии

Тип модели

Генеративный ИИ

Модалность

Естественный текст

Методы и техники

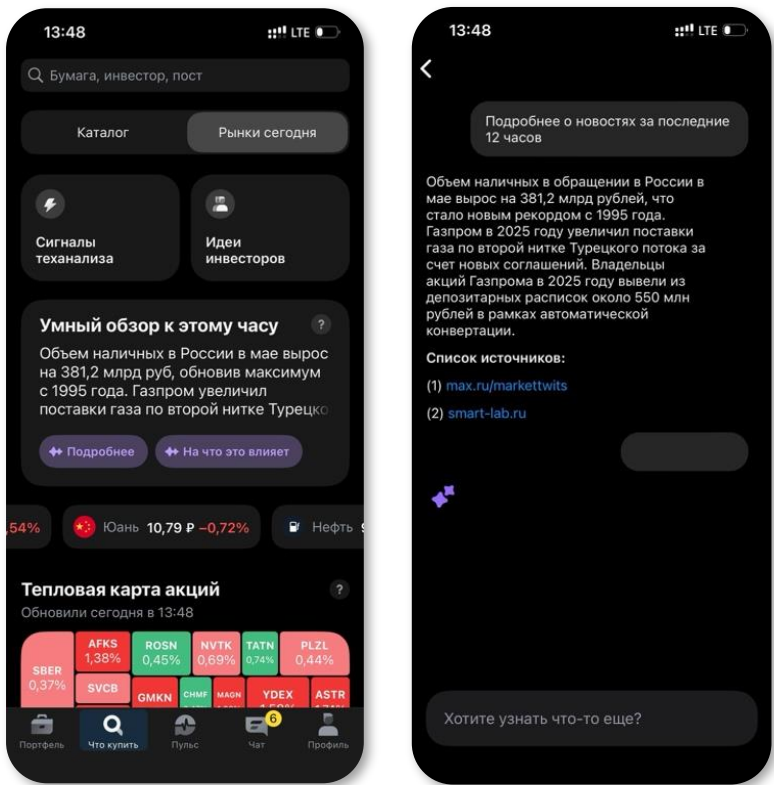
RAG

Дообучение (fine-tuning)

Инвест-ассистент



- ПРОЗРАЧНОСТЬ
- БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ
- ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ



Соблюдение рекомендаций Кодекса этики

4.2. Информирование о применении искусственного интеллекта

Ответы инвест-ассистента маркируются указанием на то, что они сгенерированы нейросетью и не являются ИИР.

4.3. Раскрытие информации об искусственном интеллекте

При генерации ответа, кроме указания на генерацию ответа нейросетью, дополнительно обращается внимание на риск неточностей и на то, что предоставляемые ответы не являются ИИР.

По данным на 3 июня 2026 года, ключевая ставка Банка России составляла **14,5%** годовых. [3]

По прогнозу ЦБ, в 2026 году средняя ключевая ставка будет в диапазоне **14–14,5%**, в 2027 году — **8–10%**. [2]

Ссылки на источники:

(2) [ria.ru](#)

(3) [cbr.ru](#)

Сгенерировано нейросетью, могут быть неточности

Не индивидуальная инвестиционная рекомендация

Инвест-ассистент



ПРОЗРАЧНОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

6.9. Минимизация выявленных рисков искусственного интеллекта

Все источники, используемые ассистентом, проходят через несколько этапов валидации:

- экспертная оценка аналитиков – основные данные (прогнозы, целевые цены, финансовые показатели) поступают от наших внутренних аналитиков и редакторов;
- регламент достоверности – существует строгая инструкция, которая запрещает ассистенту использовать непроверенные данные. Ответ считается достоверным, только если он опирается на допустимые источники или общепризнанные факты;
- техническая валидация – система автоматически проверяет соответствие ответа источникам. Если ассистент ссылается на документ, информация должна быть подтверждена соответствующим документом.

5.4. Мониторинг качества искусственного интеллекта

Специалисты регулярно оценивают ответы ассистента по критериям безопасности, достоверности и полезности, выявляя ошибки и дообучая модель.



Скоринг

Сегодня в России уже 62% кредитных организаций, применяющих и тестирующих ИИ, используют его в кредитном скоринге, а в отдельных видах кредитования заявки рассматриваются преимущественно с применением ИИ¹⁵. В этих условиях финансовые организации предоставляют клиентам возможность лучше оценить перспективы получения финансовых услуг, информируя потребителей об их кредитном рейтинге и условиях его формирования

Важно учитывать

Применение ИИ в скоринге позволяет анализировать большой объем разнообразных данных, находить в них сложные закономерности и тем самым повышать точность оценки кредитного риска. Кроме того, ИИ-решения автоматизируют процесс оказания финансовых услуг, сокращая время получения решения и снижая издержки организаций. При этом влияние такого применения ИИ на доступ граждан к финансовым услугам требует учета специфических рисков.

МОЖЕТ СНИЖАТЬ ИНФОРМИРОВАННОСТЬ

Получая итоговый результат без объяснения причин, клиент может не понимать, что именно повлияло на оценку его ситуации. Как следствие, ему трудно разобраться в логике решения, проверить его корректность и осознанно оценить собственное финансовое положение. Дополнительные риски возникают при применении сложных ИИ-решений, когда объяснимость затрудняется из-за технических особенностей моделей ИИ. В результате низкая информированность может снижать доверие к процессу принятия решений, в том числе в связи с возможным риском дискриминации.

В мире исследования показывают, что более 74% потребителей обеспокоены тем, что решения о кредитовании принимаются с помощью ИИ¹⁶. Прозрачность становится ключевым фактором доверия к финансовым организациям и их продуктам при оказании финансовых услуг: клиентам важно знать не только о применении ИИ, но и о том, какие данные при этом используются и как формируются решения¹⁷.

МОЖЕТ НЕ ПОКАЗЫВАТЬ, КАК УЛУЧШИТЬ СИТУАЦИЮ

Когда клиент узнает о низком рейтинге или отказе в предоставлении финансовой услуги, ему может быть непонятно, что делать дальше: какие действия могут повысить вероятность одобрения заявки, какие данные следует проверить, нужно ли обращаться в бюро кредитных историй или ждать изменения своей кредитной истории со временем. Даже при разъяснении обстоятельств, которые повлияли на решение, человек может не понимать, какие именно практические возможности для улучшения ситуации ему доступны.

¹⁵ [Применение искусственного интеллекта на финансовом рынке: текущий статус и условия дальнейшего развития](#). Банк России. Ноябрь 2025.

¹⁶ AI ambition meets consumer lending reality: What lenders need to know as borrower habits change. PwC. June 2026.

¹⁷ [Bridging the AI trust gap: the key to consumer confidence in financial services](#). RFI Global. March 2026.

Лучшие практики

Кодекс этики рекомендует финансовым организациям учитывать такие специфические риски и повышать осведомленность клиентов о принятии решений с применением ИИ таким образом, чтобы они могли получить положительное решение по своей заявке.

ИНФОРМИРОВАТЬ КЛИЕНТА О ПРИЧИНАХ

Следует обеспечить клиенту понимание, что влияет на решение, которое принимается в отношении него с использованием ИИ. Для этого можно обеспечить доступ клиента к оценке его кредитного рейтинга (скора) через отдельное приложение, раздел мобильного приложения или сайта организации.

При этом информирование клиента должно заключаться в предоставлении ему не только самого скор, но и сведений о том, как он был сформирован. Так, клиенту можно раскрывать сведения о:

- **Факторах, принимаемых во внимание при принятии решения**

Например, можно указывать на то, что скор формируется на основе конкретных обстоятельств из кредитной истории клиента, его дохода, текущих задолженностей, трудового стажа и т. п.

- **Данных, которые используются при принятии решения**

Информация может включать указание как на содержание данных, так и их источник (например, в формате «информация о просрочках – сведения из бюро кредитных историй»).

- **Влиянии конкретных факторов и данных на скор**

Факторы и данные можно приводить с указанием их влияния на итоговый скор в общем виде (например, цветовым решением или графическим обозначением) или с прямой оценкой их «веса» (в процентах).



Рекомендуется предоставлять клиенту как можно более полную информацию, сочетая раскрытие сведений о факторах, данных и их роли в формировании скор. Так, можно приводить факторы и соответствующие им источники данных, а также указывать на роль принятых во внимание обстоятельств (например, в формате «на снижение Вашего кредитного рейтинга в основном повлияли текущие задолженности, в том числе кредит на сумму X тысяч рублей (информация получена из бюро кредитных историй) – 50% от общего скор»)

Раскрытие информации о технических особенностях внутренних процессов работы ИИ может быть избыточным. Исследования показывают, что предоставление такой информации может положительно восприниматься клиентами, но при этом не давать им больше возможностей для оценки своего положения и перспектив получения финансовой услуги ¹⁸. Кроме того, раскрытие избыточной технической информации может создавать неоправданные риски мошенничества или атак на системы организации.



Информацию о принимаемых во внимание факторах и данных в любом случае следует предоставлять клиенту простым языком в легком для восприятия формате

 Сервис Финздоровье

п. 2.5 

 Сервис Кредитный рейтинг

п. 2.5 

Примеры информирования о факторах, учитываемых при формировании кредитного рейтинг

РАЗЪЯСНЯТЬ КЛИЕНТУ, КАК УЛУЧШИТЬ СИТУАЦИЮ

Раскрытие факторов и данных, принимаемых во внимание, следует сопровождать советами и разъяснениями, как можно повлиять на текущую ситуацию.



Важно не ограничиваться общими пояснениями и предлагать клиенту понятные и реализуемые шаги, которые помогут ему улучшить свой рейтинг (например, проверить и уточнить конкретные данные, реструктурировать задолженность в организации, снизить долговую нагрузку в определенной степени, сформировать историю своевременных платежей и т. п.)

¹⁸ [Credit where credit is due: how can AI's role in credit decisions be explained?](#) Financial Conduct Authority. February 2025.

Клиенту также должно быть понятно, что делать, если он считает текущий скор или информацию, на основе которой он сформирован, ошибочными. Ему следует пояснять, куда можно обратиться, чтобы проверить корректность данных, как подать запрос на пересмотр или уточнение сора, какие документы могут понадобиться и т. п. Предлагаемые действия по возможности должны быть доступны из того же интерфейса, в котором клиент узнает свой рейтинг.

При наличии дополнительных вопросов клиенту следует дать возможность обратиться к сотруднику, который сможет подробнее объяснить ситуацию и помочь с выбором дальнейших шагов.

 Сервис Финздорье	п. 2.5 
 Сервис Кредитный рейтинг	п. 2.5 

Примеры предоставления рекомендаций
по улучшению кредитного рейтинга

Сервис Финздорovie



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 2.5. Повышение осведомленности клиентов о принятии решений с применением ИИ

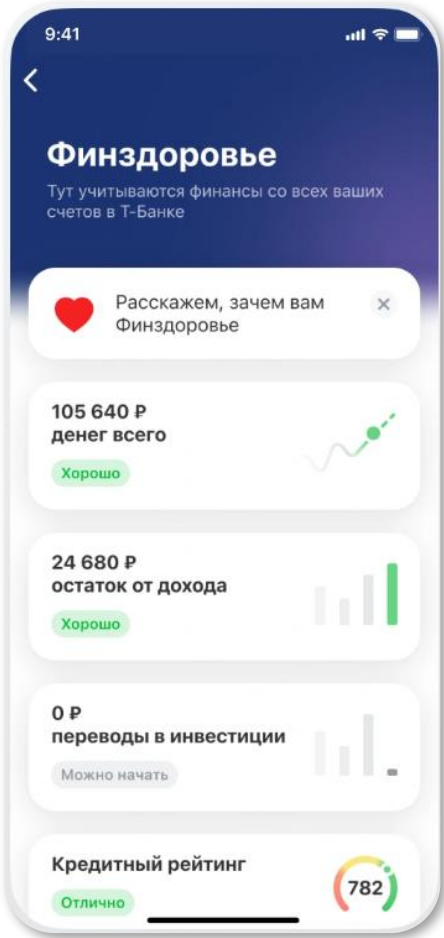
О сервисе

Сервис Финздорovie доступен клиентам в мобильном приложении банка. Если клиент дает согласие на запрос кредитной истории, в Финздорovie отображаются все займы, а также кредитный рейтинг, который показывает платежеспособность клиента и варьируется от 0 до 999 баллов. Чем выше кредитный рейтинг, тем выше вероятность получить финансовую услугу.

Кредитный рейтинг формируется с учетом информации из нескольких бюро кредитных историй, что позволяет клиенту получить более комплексную и точную оценку финансового положения, чем индивидуальный рейтинг субъекта кредитной истории.

Оценка является примерной, и при принятии решений каждый банк опирается на свои критерии. Тем не менее баллы помогают клиенту ориентировочно узнать шансы на одобрение заявки, и сервис может помочь клиенту улучшить свой профиль как заемщика.

Так, в Финздорovie можно увидеть слабые места кредитной истории, причины низкого рейтинга и получить персональные рекомендации по тому, как можно повысить свой рейтинг.



Технологии

- | | |
|------------------|--------------------------|
| Тип модели | Классический ML |
| Модальность | Структурированные данные |
| Методы и техники | Объяснимость (SHAP) |

Сервис Финздорovie



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

Соблюдение рекомендаций Кодекса этики

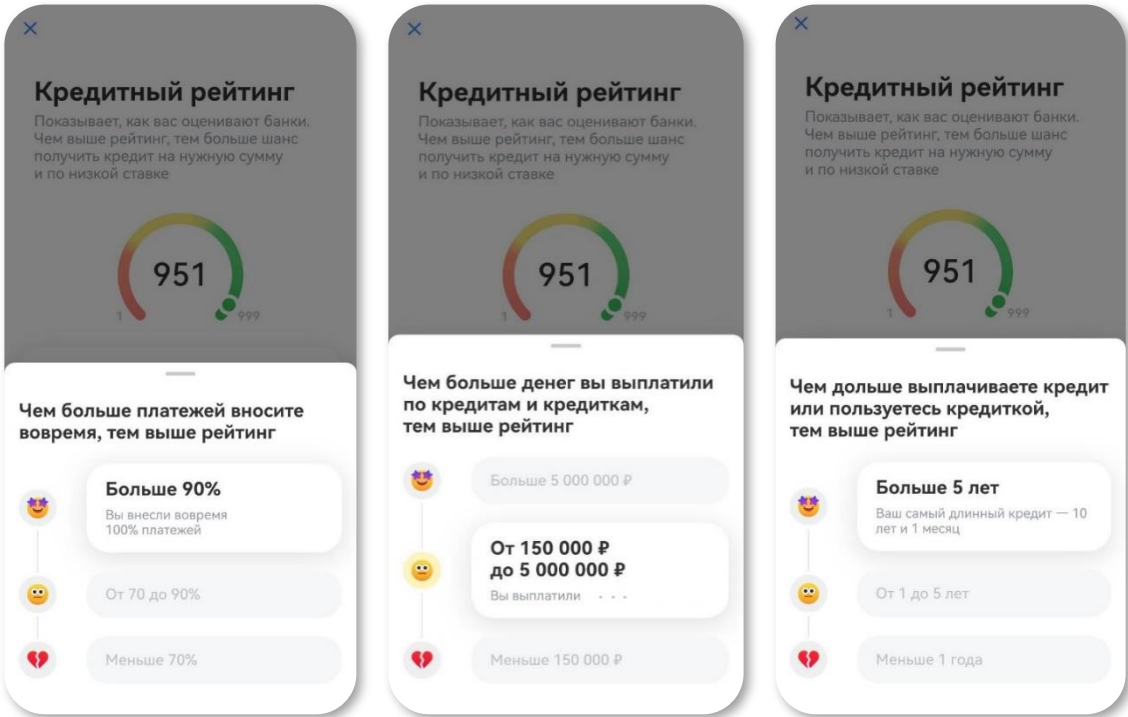
2.5. Повышение осведомленности клиентов о принятии решений с применением искусственного интеллекта

Финздорovie указывает клиенту на факторы, которые приводят к формированию положительного или отрицательного кредитного рейтинга, а также дает подробные рекомендации, позволяющие снизить влияние негативных факторов.

В частности, в отношении каждого клиента определяются факторы, оказывающие наибольшее негативное влияние на кредитный рейтинг, в том числе наличие просроченных платежей (и их давность), доля своевременных платежей, отсутствие кредитной истории, и предлагаются советы по повышению рейтинга.

Сервис может посоветовать клиенту обратиться в бюро кредитных историй для устранения ошибок, объяснить механику влияния просрочек на кредитный рейтинг и помочь подобрать вариант накопления платежей без просрочек либо указать на важность минимизации просроченной задолженности.

Если клиент видит ошибку в рейтинге или отображаемых данных, он может обратиться в поддержку. Сотрудники выяснят, была ли допущена ошибка в бюро кредитных историй или банке-источнике либо возникла ли техническая проблема на стороне сервиса. При выявлении ошибок рейтинг может быть пересчитан.



Сервис Финздоровье



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

Применение сервиса не только помогает повысить вероятность принятия положительного решения об оказании финансовых услуг, но и повышает прозрачность взаимодействия клиента с банком.

ФАКТ

В вашей истории осталась закрытая кредитка

18 ноября 2019 года вы закрыли кредитку с лимитом 100 000 ₽, но в кредитной истории она все еще отображается как открытая. Скорее всего, произошла техническая ошибка или банк не передал данные в НБКИ.

Эта ошибка может влиять на рейтинг и снижает шансы на получение новых кредитов.

Чтобы это исправить, заполните заявление в НБКИ

[Написать заявление в НБКИ](#)

ФАКТ

Чем меньше сумма просрочки, тем меньше она влияет на рейтинг

3 месяца назад вы просрочили платеж на 2 дня — это ухудшило рейтинг. Чтобы его восстановить, оплачивайте покупки кредиткой и возвращайте деньги вовремя.

Банк будет отправлять в БКИ информацию о внесенных платежах, и чем активнее вы будете пользоваться кредиткой, тем быстрее улучшится рейтинг.

Сервис Кредитный рейтинг



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 2.5. Повышение осведомленности клиентов о принятии решений с применением ИИ

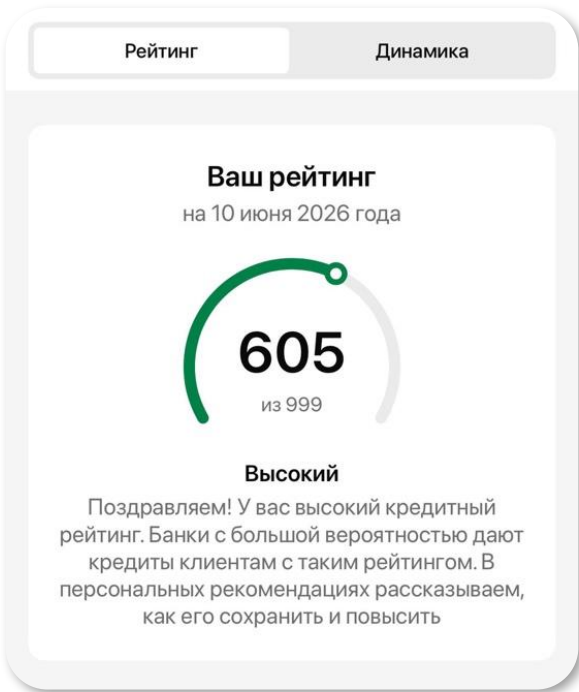
О сервисе

Модель ИИ применяется для оценки риска дефолта по всему потоку заявок на потребительские кредиты, кредитные карты и автокредиты.

Одновременно Газпромбанк предоставляет клиентам доступ к информационному сервису оценки кредитного рейтинга в цифровых каналах – Кредитный рейтинг. Он помогает заемщику учитывать свою кредитоспособность и работать над ее улучшением.

Так, сервис предоставляет клиенту его персональный кредитный рейтинг, который рассчитывается на основании собственных методик и отражает оценку уровня привлекательности физического лица как потенциального заемщика. Кредитный рейтинг выражается в баллах от 0 до 999: чем выше значение, тем потенциально выше шансы на одобрение заявки. Оценка, полученная клиентом в сервисе, не гарантирует положительного решения о заключении кредитного договора.

Клиенту также предоставляется информация о факторах формирования кредитного рейтинга и персональные рекомендации по его улучшению.



Технологии

Тип модели

Классический ML

Модальность

Структурированные данные

Сервис Кредитный рейтинг



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

Соблюдение рекомендаций Кодекса этики

2.5. Повышение осведомленности клиентов о принятии решений с применением искусственного интеллекта

Сервис указывает клиенту на факторы, которые учитываются при формировании кредитного рейтинга, и дает рекомендации, как получить хороший результат.

В частности, каждому клиенту демонстрируется результат оценки основных зон определения кредитного рейтинга: кредитная история, благонадежность и платежеспособность. В сервисе клиент также может ознакомиться с информацией о своих обязательствах на основе информации из бюро кредитных историй.

Кроме того, сервис дает клиенту индивидуальные рекомендации по улучшению и поддержанию кредитного рейтинга, а также предупреждает о рисках мошенничества.

Обновление кредитного рейтинга в сервисе доступно клиенту бесплатно раз в месяц, а также неограниченное количество раз в течение 10 дней после отказа в заключении договора потребительского кредита. В сервисе клиент может отслеживать историю изменений кредитного рейтинга.

Зоны влияния на рейтинг

Кредитная история

Вы установили самозапрет на оформление кредитов



Благонадежность

Не раскрывайте личные данные незнакомым людям, игнорируйте подозрительные звонки, письма и сообщения



Кредитная нагрузка

У вас отсутствует кредитная нагрузка



Как повысить рейтинг



Рекомендации по улучшению
вашего рейтинга



Сильные стороны
вашей кредитной истории



Анализ рисков операций

На финансовом рынке системы анализа рисков операций с применением ИИ становятся важным инструментом противодействия мошенническим и иным подозрительным операциям. Такие решения позволяют учитывать широкий набор признаков и поведенческих паттернов и в режиме реального времени анализировать транзакции, приостанавливая их при выявлении признаков нетипичных операций

Важно учитывать

Системы анализа рисков операций позволяют быстро обрабатывать большой объем операций и выявлять подозрительные транзакции. Они повышают эффективность мониторинга, защищая интересы клиента и организации, а также снижают операционную нагрузку на сотрудников.

МОГУТ ДОПУСКАТЬ ОШИБКИ

Использование ИИ для анализа рисков операций может существенно влиять на интересы клиентов. Для клиента негативные последствия могут возникать в случаях, когда система ошибочно квалифицирует операцию как рискованную (ложноположительное срабатывание, false positive). Это может приводить к задержке или блокировке операции, необходимости дополнительного подтверждения своих действий, невозможности своевременно распорядиться денежными средствами и общему ухудшению клиентского опыта.

Для финансовой организации большое количество неверных срабатываний может приводить к росту нагрузки на сотрудников, увеличению числа проверок, снижению операционной эффективности и необходимости регулярно корректировать параметры модели и правила мониторинга. Не менее значимы и ошибки, когда система не выявляет подозрительную операцию, что влечет регуляторный и финансовый риски.

Лучшие практики

Для защиты интересов клиентов и обеспечения эффективности работы систем ИИ, осуществляющих анализ операций, релевантен ряд положений Кодекса этики. В частности, это рекомендации, касающиеся пересмотра решения, принятого с применением ИИ, обоснованности использования персональных данных о клиенте, а также минимизации рисков ИИ. На практике они могут быть реализованы по-разному.

ОБЕСПЕЧИВАТЬ ЧЕЛОВЕЧЕСКИЙ НАДЗОР И ПЕРЕСМОТР РЕШЕНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Целесообразно сохранять возможность участия сотрудника организации в процессах принятия решений о блокировках, принимаемыми с применением ИИ. Так, текущий и последующий человеческий контроль может позволить своевременно отреагировать на ложноположительное срабатывание, снизив неблагоприятные последствия для клиента. При этом процедуры такого контроля должны определяться организацией самостоятельно с учетом внутренних правил управления рисками.

Последующую проверку решений ИИ, принятых в результате мониторинга, можно дополнять применением другой модели ИИ, которая будет проверять принятое решение. Такой подход может предусматривать функционал автоматической разблокировки, когда проверяющая модель ИИ выявляет ошибку в исходном решении о блокировке и самостоятельно пересматривает его. В этих процессах также целесообразно обеспечивать последующий контроль сотрудником организации.



В любом случае важно обеспечить для клиента возможность пересмотра сотрудником организации решения о блокировке, принятого с применением ИИ

Система автоматического анализа рисков операций
с функцией авторазблокировки и пост-контролем

п. 2.4 [↗](#)

Пример организации контроля человека за решениями ИИ

УЧИТЫВАТЬ ТОЛЬКО РЕЛЕВАНТНЫЕ ДАННЫЕ

Необязательное использование данных о клиенте может приводить к необоснованным блокировкам. Если высокой точности результатов удастся достичь только на основе транзакционных данных, то нецелесообразно использовать дополнительные персональные данные клиента. В то же время их использование с учетом требований законодательства может быть обосновано, если это существенно повышает качество ИИ.

ПОВЫШАТЬ КАЧЕСТВО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Поскольку решения ИИ при анализе операций могут существенно затрагивать базовые права и интересы клиента, особенно важно обеспечить высокий уровень качества ИИ.

В частности, в рамках СУР в организации, в том числе рисками ИИ, важно проводить тщательное тестирование и валидацию модели.

Не менее значим мониторинг после введения в эксплуатацию, среди возможных метрик которого – доля подтвержденных подозрительных от общего числа заблокированных операций (precision) или доля заблокированных от общего числа подтвержденных подозрительных операций (recall).



Целесообразно постоянно улучшать качество моделей, результат работы которых непосредственно влияет на финансовые интересы клиентов. При доработке крайне важно учитывать причины реализовавшихся риск-событий, в том числе ложноположительных и ложноотрицательных срабатываний

В анализе рисков операций по возможности следует использовать интерпретируемые модели ИИ, позволяющие оценить влияние отдельных признаков на решение ИИ. При этом соответствующую информацию не следует раскрывать третьим лицам, если есть разумные основания полагать, что она может быть использована для обхода и недобросовестного воздействия на систему принятия решений о блокировках.

Комплексная оценка риска транзакций в режиме реального времени



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

СПРАВЕДЛИВОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 2.2. Повышение качества оказания услуг
- пункт 2.4. Обеспечение возможности пересмотра решения, принятого с применением ИИ
- пункт 2.5. Повышение осведомленности клиентов о принятии решений с применением ИИ
- пункт 3.2. Обоснованность использования персональных данных о клиентах
- пункт 3.3. Недопустимость использования дискриминационных факторов
- пункт 3.4. Обеспечение недискриминационного характера набора данных
- пункт 5.3. Проверка качества ИИ
- пункт 5.4. Мониторинг качества ИИ
- пункт 6.10. Реагирование на реализовавшиеся риски ИИ

О сервисе

В процессах банка по выявлению подозрительных операций применяется целый ряд ИИ-решений. Антифрод-платформа Сбера обеспечивает комплексный анализ рисков, ежедневно обрабатывая более 750 миллионов операций клиентов. Система покрывает все каналы обслуживания – от мобильного приложения и интернет-банка до переводов через партнерские сервисы и кассовых операций в отделениях.

Каждое действие оценивается за десятки миллисекунд, что позволяет принять решение о блокировке до фактического списания средств, не создавая задержек для клиентов.

Система позволяет выявлять 99,9% мошеннических операций. Только за 2025 год платформа предотвратила хищения на сумму свыше 360 миллиардов рублей.

Ключевая особенность решения – способность мгновенно адаптироваться к новым схемам атак. Благодаря выстроенной инженерной инфраструктуре время вывода моделей и новых признаков в продуктивную среду сокращено до минимума.

Платформа применяет гибридную архитектуру:

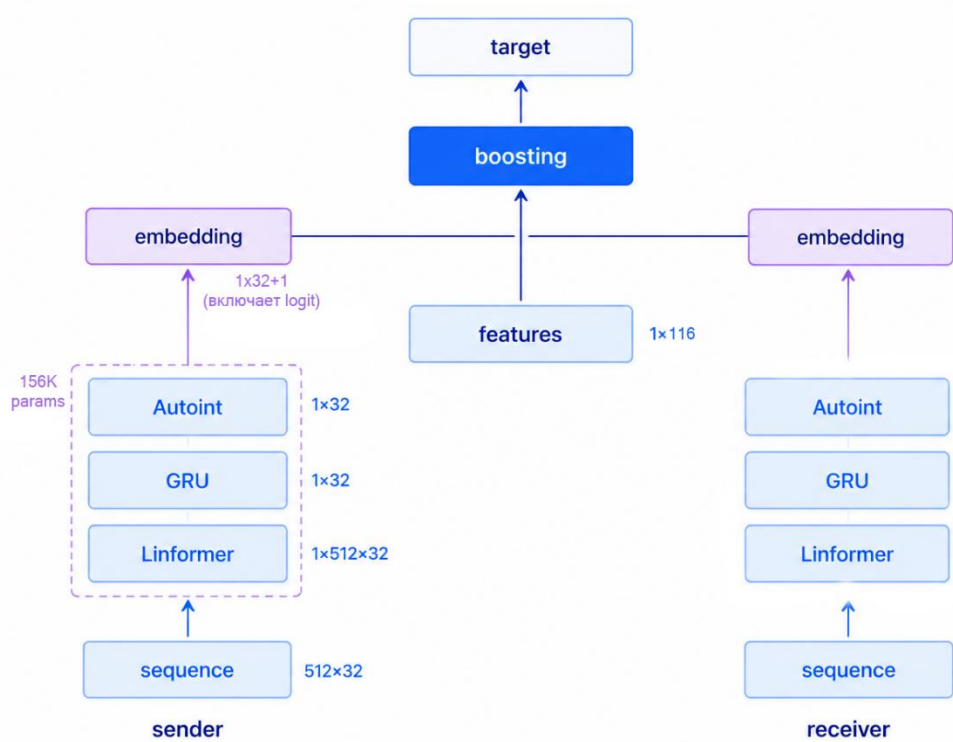
- классическое машинное обучение: быстрые ансамблевые модели обеспечивают мгновенный расчет вероятности мошенничества;
- глубокое обучение: нейросетевые компоненты обогащают расчеты глубокими векторными представлениями поведения клиентов – эмбедами. Это позволяет системе понимать скрытые паттерны, которые недоступны классическим алгоритмам.

Комплексная оценка риска транзакций в режиме реального времени



- ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ
- СПРАВЕДЛИВОСТЬ
- БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ
- ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Пример гибридной архитектуры моделей антифрода



Технологии

Тип модели

Классический ML

Модели глубокого обучения

Генеративный ИИ

Агентский ИИ

Модальность

Аудио

Естественный текст

Структурированные данные

Методы и техники

Дообучение (fine-tuning)

Объяснимость (SHAP)

Комплексная оценка риска транзакций в режиме реального времени



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

СПРАВЕДЛИВОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Соблюдение рекомендаций Кодекса этики

2.4. Обеспечение возможности пересмотра решения, принятого с применением искусственного интеллекта

При выявленном риске замораживается только конкретная операция, а не весь доступ к продуктам. Клиент получает звонок или сообщение в чате, где оператор (с учетом риск-маркеров) дает возможность подтвердить легитимность действия. Задача сотрудника – в ходе диалога определить, действует ли клиент самостоятельно или находится под влиянием мошенников.

Автономный агент контроля качества подтверждения операций

Для системного контроля действий сотрудников при подтверждении высокорисковых транзакций внедрен специальный ИИ-агент по контролю качества. Он работает в офлайн режиме и выполняет следующие задачи:

- анализ диалога: автоматически прослушивает записи разговоров;
- выявление ошибок: находит неточности в действиях оператора и нарушения регламента;
- оценка полноты: проверяет, все ли обязательные вопросы были заданы клиенту;
- соблюдение скрипта: оценивает корректность ведения диалога.

Такое решение позволяет масштабировать экспертный контроль на весь объем звонков без увеличения штата. Агент обеспечивает единообразие стандартов обслуживания и исключает человеческий фактор при оценке работы операторов.



Комплексная оценка риска транзакций в режиме реального времени



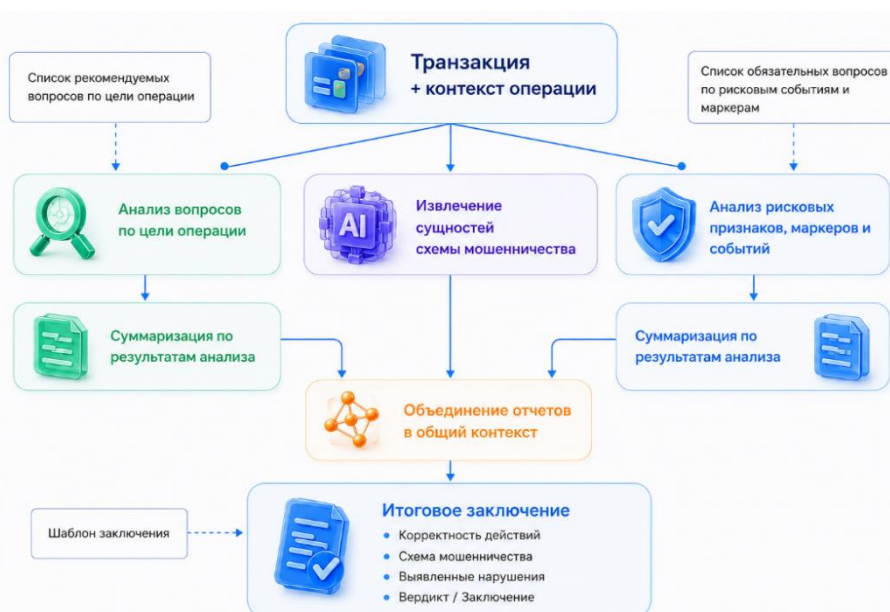
ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

СПРАВЕДЛИВОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Схема работы решения



Агент анализирует беседу по множеству параметров:

- соблюдение утвержденного скрипта;
- уместность и тактичность заданных вопросов;
- полнота выданных рекомендаций по безопасности

Семейство отчетов в Qlik Sense визуализирует эти критерии, делая логику оценки прозрачной и понятной для руководителей и самих операторов при разборе ошибок. Также разработано семейство интерактивных дашбордов. Они позволяют:

- точно выявлять нарушения в действиях конкретных сотрудников;
- отслеживать динамику качества обслуживания по разным командам;
- оперативно корректировать обучение персонала



Комплексная оценка риска транзакций в режиме реального времени



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

СПРАВЕДЛИВОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Система позиционируется и используется как инструмент для оперативного корректирования обучения персонала. Человек-супервизор использует выводы агента для точечного выявления нарушений и последующего разбора

2.2. Повышение качества оказания услуг

Для отслеживания удовлетворенности клиентов сервисами защиты регулярно проводятся опросы CSI (Customer Satisfaction Index). Внедрен агент, который в полностью автоматическом режиме обрабатывает результаты опросов (далее – CSI-агент):

- определяет реальную эмоцию пользователя и цель его комментария;
- группирует жалобы на конкретные проблемы (например, «сложная верификация», «долгое ожидание»);
- строит и анализирует связь между жалобой и конкретным алгоритмом фрод-мониторинга, торговой точкой, каналом подтверждения или регионом;
- автоматически выявляет факторы, которые сильнее всего влияют на негатив или позитив в динамике.

Ключевая особенность решения – способность анализировать тональность комментария независимо от выставленной клиентом оценки. CSI-агент корректно интерпретирует контекст, даже если формальный балл не отражает реальную эмоцию (например, клиент поставил 5 баллов из вежливости, но в тексте указал на неудобство).

Комплексная оценка риска транзакций в режиме реального времени



- ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ
- СПРАВЕДЛИВОСТЬ
- БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ
- ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Схема работы решения



Решение позволяет анализировать 100% ответов клиентов об удовлетворенности работой сервисов антифрода, а не только выборочные или негативные. Это обеспечивает полное и объективное понимание реальных потребностей клиентов, что является основой для системного повышения качества таких сервисов.

CSI-агент автоматически выявляет драйверы негатива и удовлетворенности, связывая жалобы с конкретными алгоритмами фрод-мониторинга, каналами или регионами. Это создает замкнутый цикл (feedback loop), где прямая обратная связь от клиентов становится сигналом для доработки и совершенствования моделей ИИ и алгоритмов фрод-мониторинга.

Комплексная оценка риска транзакций в режиме реального времени



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

СПРАВЕДЛИВОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

3.2. Обоснованность использования персональных данных о клиентах

3.3. Недопустимость использования дискриминационных факторов

Гибридная архитектура (ClassicML + DeepLearning) использует поведенческие эмбединги. Скоринг строится на паттернах транзакций и контексте операции, игнорируя статические демографические или социальные признаки клиента.

В свою очередь, CSI-агент формирует аналитические срезы по сегменту, полу, региону и другим признакам. Это позволяет объективно выявлять, не оказывают ли определенные антифрод-алгоритмы непропорционально негативного влияния на конкретные группы клиентов, обеспечивая справедливость и равномерное качество сервиса для всех категорий пользователей.

3.4. Обеспечение недискриминационного характера набора данных

Пайплайны валидации проверяют обучающие и рабочие выборки на отсутствие смещений (bias) и скрытых прокси-признаков, которые могли бы привести к дискриминации отдельных групп пользователей.

5.3. Проверка качества искусственного интеллекта

Проверка качества осуществляется в том числе за счет собственной платформы разработки. Ее ключевые элементы:

- централизованное хранилище признаков (Feature Store): обеспечивает расчет и гарантирует, что модель обучается и принимает решения на абсолютно одинаковых данных;
- автоматизированные пайплайны валидации: исключают ошибки при проверке качества новых версий;
- механизм автодообучения: позволяет выпускать обновления и новые признаки в продуктивную среду в максимально короткие сроки.

Платформа обеспечивает полное соответствие тестовой и промышленной сред, что позволяет мгновенно реагировать на появление новых схем мошенничества.

Комплексная оценка риска транзакций в режиме реального времени



- ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ
- СПРАВЕДЛИВОСТЬ
- БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ
- ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Система CoPilot фрод-аналитика

В процессе создания антифрод-алгоритмов используется CoPilot, закрывая полный цикл работы с алгоритмами:

- оценивает эффективность текущих алгоритмов;
- помогает корректировать существующие или генерировать совершенно новые алгоритмы;
- автоматически тестирует гипотезы и рассчитывает ожидаемый эффект до внедрения.

Фрод-аналитик ставит задачу на естественном языке, используя привычную терминологию предметной области. Агент самостоятельно разбивает ее на этапы, запрашивает нужные данные из хранилища, генерирует код алгоритма, тестирует его в безопасной среде и возвращает специалисту готовое решение с прогнозом ключевых метрик.



Благодаря автоматизации рутинного анализа время разработки и вывода новых защитных алгоритмов в продуктивную среду сократилось на 30–50%.

Ключевая особенность CoPilot – автоматическая проверка гипотез в изолированной среде до вывода в «прод». Агент рассчитывает ожидаемый эффект и прогнозирует метрики, что исключает риск внедрения нестабильных или неоптимальных алгоритмов, которые могли бы навредить клиентскому опыту.

Симуляция влияния новых алгоритмов на реальный поток данных через специализированные сервисы позволяет увидеть последствия еще до запуска. Агент прогнозирует метрики, минимизируя риски для реальных операций клиентов до того, как алгоритм будет внедрен в работу.

Комплексная оценка риска транзакций в режиме реального времени



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

СПРАВЕДЛИВОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

5.4. Мониторинг качества искусственного интеллекта

Механизм автоматического мониторинга непрерывно отслеживает смещение распределений (data drift / concept drift) и адаптирует веса под текущие реалии без остановки сервиса. Платформа автоматически проверяет модели на соответствие метрикам качества в меняющихся условиях рынка.

6.10. Реагирование на реализовавшиеся риски искусственного интеллекта

Любые отклонения в работе моделей оперативно выявляются. Встроенные механизмы быстрого отката, переобучения и дообучения минимизируют негативный эффект от возможных ошибок ИИ.

Помимо мониторинга технических отклонений существует мониторинг клиентской обратной связи, в рамках которого CSI-агент автоматически отслеживает динамику показателей и выявляет типовые сценарии недовольства. Это позволяет оперативно корректировать пороги срабатывания алгоритмов или логику взаимодействия, минимизируя репутационные и операционные риски, вызванные неоптимальной работой антифрод-системы.

2.5. Повышение осведомленности клиентов о принятии решений с применением искусственного интеллекта

Инструменты интерпретируемости (SHAP) помогают специалистам фрод-мониторинга объяснить клиенту причины запроса дополнительной верификации или блокировки транзакции простым языком, соблюдая принцип прозрачности принятия решений ИИ.

Система автоматического анализа рисков операций с функцией авторазблокировки и пост-контролем



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

СПРАВЕДЛИВОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 2.4. Обеспечение возможности пересмотра решения, принятого с применением ИИ
- пункт 3.2. Обоснованность использования персональных данных о клиентах
- пункт 6.9. Минимизация выявленных рисков ИИ
- пункт 6.10. Реагирование на реализовавшиеся риски ИИ

О сервисе

Система на базе ИИ анализирует операции клиентов в реальном времени для выявления рисков. Алгоритм изучает текст платежа, сумму, время и сопоставляет их с календарем (праздники, выходные) и историей трат.

При выявлении ложного срабатывания правил блокировки система самостоятельно отменяет ограничение (авторазблокировка), минимизируя неудобства для клиента.



Технологии

Тип модели

Классический ML

Модальность

Естественный текст

Структурированные данные

Методы и техники

Дообучение (fine-tuning)

Соблюдение рекомендаций Кодекса этики

2.4. Обеспечение возможности пересмотра решения, принятого с применением искусственного интеллекта

Реализован двухуровневый контроль за решениями ИИ:

- автоматический пересмотр: ИИ сам инициирует пересмотр решения о блокировке, если новые данные свидетельствуют об ошибке (авторазблокировка);
- человеческий контроль: каждое решение ИИ о разблокировке проверяется сотрудником комплаенса (пост-контроль). Сотрудник может отменить решение ИИ или инициировать дополнительную проверку. Таким образом, финальный контроль всегда остается за человеком, а клиент гарантированно получает решение сотрудника в спорной ситуации.

Система автоматического анализа рисков операций с функцией авторазблокировки и пост-контролем



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

СПРАВЕДЛИВОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

3.2. Обоснованность использования персональных данных о клиентах

На текущем этапе система принимает решения исключительно на основе транзакционных данных. Система не использует анкетные данные клиента (пол, возраст, доход) в том числе в связи с тем, что точность решений уже сейчас достигает почти 100%.

6.9. Минимизация выявленных рисков искусственного интеллекта

Внедрен процесс регулярного пост-контроля. Так, сотрудники проверяют выборку решений ИИ, выявляют ошибки и вносят корректировки в алгоритм, минимизируя риски.

6.10. Реагирование на реализовавшиеся риски искусственного интеллекта

Предусмотрен механизм обратной связи: если клиент не согласен с решением (даже после авторазблокировки или если ИИ не сработал), он может обратиться к оператору, который передаст информацию для дообучения модели.

Система фиксирует все случаи авторазблокировок как риск-события для последующего анализа.



Финансовая доступность

ИИ может выступать инструментом повышения финансовой доступности. Технология способна адаптировать финансовые сервисы под потребности различных групп клиентов, упростить сложную информацию, снизить языковые, сенсорные и когнитивные барьеры, а также создать новые форматы сопровождения клиента при оказании финансовых услуг

Важно учитывать

Сервисы на основе ИИ открывают новые возможности для повышения доступности финансовых услуг. Внедрение инклюзивных ИИ-сервисов позволяет финансовым организациям расширить клиентскую базу за счет более эффективного обслуживания социально незащищенных и уязвимых категорий клиентов, лиц, не обладающих достаточным опытом и знаниями, а также тех, для кого традиционные каналы взаимодействия могут быть менее удобными или доступными. Такие практики демонстрируют ответственное применение технологии ИИ финансовой организацией.

МОГУТ КАК ПОВЫШАТЬ, ТАК И ПОНИЖАТЬ ФИНАНСОВУЮ ДОСТУПНОСТЬ

Инклюзивные ИИ-сервисы позволяют адаптировать финансовые услуги и процесс коммуникации с учетом особенностей отдельных групп клиентов и тем самым повышают финансовую доступность. Учет факторов уязвимостей компенсирует ограничения, которые могут затруднять получение или понимание финансовой услуги.

В то же время факторы уязвимости клиентов могут стать инструментом мисселинга. ИИ может использовать сведения о возрасте, уровне финансовой грамотности, опыте инвестирования или поведенческих особенностях клиента для продвижения продуктов, которые не соответствуют его потребностям и способности оценить риски.

Например, пожилым клиентам могут предлагаться долгосрочные или низколиквидные инвестиционные продукты без учета их потребности в доступности средств, а клиентам без достаточного опыта и знаний – сложные финансовые продукты с непрозрачной структурой доходности и рисков ¹⁹

Лучшие практики

В рамках повышения финансовой доступности в Кодексе этики организациям рекомендуется при разработке и применении ИИ учитывать факторы уязвимости клиентов (возраст, образование, ограниченные возможности и другие) и возможное влияние указанных факторов на оказание таким клиентам услуг. Внедрение инклюзивных ИИ-сервисов может рассматриваться как лучшая практика ответственного следования этой рекомендации, поскольку предполагает не только учет уязвимостей клиентов, но и активное использование возможностей ИИ для снижения связанных с ними барьеров.

¹⁹ [Виды недобросовестного поведения участников финансового рынка](#). Банк России.

ПРИМЕНЯТЬ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ПОВЫШЕНИЯ ФИНАНСОВОЙ ДОСТУПНОСТИ

Инклюзивные ИИ-сервисы могут помочь клиенту лучше понимать финансовые продукты, безопасно пользоваться ими и получать услугу в удобном для него формате. Такие сервисы могут стать дополнением к традиционным каналам обслуживания.

Примеры инклюзивных ИИ-сервисов:

- **Сервисы саммаризации** позволяют выделить ключевые положения в договорах, правилах оказания услуг и иных финансовых документах, сокращая объем информации, который клиенту необходимо самостоятельно изучить. Такие сервисы могут обращать внимание клиента на комиссии, сроки, риски, ограничения и последствия нарушения обязательств.

 Сервис ИИ-саммаризации договоров



Пример сервиса саммаризации

Банк России о развитии практики

Банк России поддерживает развитие практик ответственного применения ИИ, которые помогают сделать условия финансовых продуктов более понятными для потребителя, обратить его внимание на существенные условия и риски и тем самым способствуют более осознанному финансовому выбору. Развитие подобных инструментов соответствует подходам Кодекса этики, включая человекоцентричность и прозрачность.

При этом принципиально важно сохранять **равные условия для участников финансового рынка, равноудаленность, технологическую нейтральность, достоверность, корректность и сопоставимость информации для граждан**. Аналогичные по целям и результату подходы могут реализовываться финансовыми организациями различными способами, при этом Банк России склоняется к целесообразности создания централизованного сервиса для финансовых организаций

- **Сервисы упрощения и адаптации текстов** преобразуют сложные юридические и финансовые формулировки в более понятный текст. Сервисы «простой язык» (plain language) помогают сделать информацию доступнее для широкого круга клиентов, а «легкое чтение» (easy-to-read) предполагают более глубокую адаптацию текста и формата подачи для клиентов, которым сложно воспринимать стандартные финансовые документы.
- **Сервисы для незрячих и слабовидящих клиентов** могут распознавать и озвучивать содержание документов, платежных реквизитов, квитанций, экранов мобильного приложения, интерфейсов банкоматов и иных визуальных элементов.

Сбер разработал сервис «легкое чтение», основанный на нейросетевой модели GigaChat²⁰. Сервис решает задачу перевода любых текстов на простой язык без потери смысла. В основе навыка лежат правила руководства по когнитивной доступности²¹.

Виртуальный ассистент Салют помогает клиентам Сбербанка с нарушением зрения голосом переводить деньги и оплачивать мобильную связь: он выбирает контакт, проговаривает детали операции и предлагает подтвердить или изменить их без использования скринридера²²

- **Сервисы для глухих и слабослышащих клиентов** могут обеспечивать автоматические субтитры, расшифровку аудиосообщений и видеоконсультаций, перевод речи сотрудника в текст в реальном времени.
- **Голосовые финансовые помощники** могут помогать пожилым клиентам и пользователям с низким уровнем цифровых навыков выполнять базовые операции, находить нужные разделы приложения, получать объяснения по продукту и обращаться за поддержкой в более привычном разговорном формате.
- **ИИ-помощники для детей и подростков** могут объяснять финансовые понятия простым языком, сопровождать первые платежи, показывать последствия финансовых решений, помогать формировать навыки безопасного обращения с деньгами и работать вместе с родительскими настройками контроля.



Пример ИИ-помощника для детей и подростков

²⁰ [GigaChat поможет адаптировать текст для понимания любого человека](#). Сбер. 2025.

²¹ [Доступные языки и инклюзивная коммуникация](#). Сбер.

²² [СберБанк Онлайн для незрячих](#).

УЧИТЫВАТЬ РИСКИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ФИНАНСОВОЙ ДОСТУПНОСТИ

При применении ИИ следует учитывать возможные риски для финансовой доступности. В частности, использование ИИ для принятия решений о территориальном размещении подразделений и платежных устройств кредитных организаций без учета социально значимых факторов может привести к перекосу в пользу экономически более привлекательных локаций и, как следствие, к снижению доступности финансовых услуг для уязвимых групп клиентов и жителей удаленных территорий²³.

²³ [Основные направления повышения доступности финансовых услуг в российской федерации на период 2025–2027 годов](#). Банк России. 2024.

Сервис ИИ-саммаризации договоров



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

ПРОЗРАЧНОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 2.6. Повышение финансовой доступности
- пункт 4.2. Информирование о применении ИИ
- пункт 5.3. Проверка качества ИИ

О сервисе

Для защиты потребителей от рисков, связанных со сложностью финансовых продуктов, Совкомбанк разработал ИИ-решение для анализа текстов финансовых документов.

Сервис работает как интеллектуальный ассистент, который анализирует юридически сложный договор и формирует для клиента сжатую и понятную памятку (до 2 страниц). В ней фокусируется внимание на существенных условиях, определяющих до 80% финансового результата продукта или возможных потерь. Модель выявляет повышенные риски продукта, скрытые комиссии и факторы недобросовестных практик, сравнивая договор с эталонными примерами.



Технологии

Тип модели

Генеративный ИИ

Модальность

Естественный текст

Методы и техники

Промпт-инжиниринг

Сервис ИИ-саммаризации договоров



- ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ
- ПРОЗРАЧНОСТЬ
- БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

Памятка по использованию

Описание продукта

1.1	Название продукта	Срочный вклад «Регулярные 9%»
1.2	Срок вклада	от 90 до 1095 дней Условия договора, п.п. 4
1.3	Сумма вклада	от 50 000 рублей. Вклады застрахованы государством (АСВ) на сумму до 1 400 000 рублей, включая все остатки по вкладам в одном банке. Условия договора, п.п. 2 и 12
1.4	Процентная ставка	Минимально гарантированная ставка — 8,6% годовых. Базовая ставка — от 8,6% до 12,1% годовых в зависимости от срока вклада. При выполнении условий по подписке «Халва Десятка» с выполнением оборота или «Оптимал» с выполнением расходных операций — ваша ставка может повыситься: +1% при сроках 3, 6, 12 месяцев; +2% при 24 месяцах; +5% при 36 месяцах. Повышение применяется с пересчётом процентов за весь срок. Надбавки не суммируются, применяется первая выполненная. При просрочке по другим договорам с банком повышение не действует. Условия договора, п.п. 5

Существенные условия

2.1	Как начисляются проценты	Проценты начисляются ежедневно на остаток средств на счёте вклада, но не причисляются к основной сумме. Выплаты происходят отдельно — на ваш текущий счёт. Условия договора, п.п. 6
2.2	Период начисления	Проценты будут перечисляться ежемесячно в день, соответствующий дате открытия вклада, на ваш текущий счёт. Условия договора, п.п. 6
2.3	Условия пополнения	Вы можете пополнять вклад в течение первых 10 календарных дней после открытия или пролонгации. После этого срока пополнение невозможно. Минимальная сумма пополнения — 1 000 рублей. Условия договора, п.п. 3
2.4	Что происходит с деньгами в день окончания срока	Если вы не заберёте деньги в день окончания срока, вклад автоматически продлится на условия вклада «Регулярный 9%» на момент пролонгации. Если пролонгирование прекратится банком, и от вас не будет запроса о возврате суммы, деньги перечислят на ваш текущий счёт. Условия договора, п.п. 9
2.5	Способы информирования	Банк сообщит вам о дате возврата вклада через интернет-банк, СМС, push-уведомления, по телефону, почте или при личном визите в отделение. Условия договора, п.п. 10

Отличительные особенности

3.1	Запрос дополнительных согласий	Вы даёте банку согласие: - на перевод денег между Вашими счётами в Совкомбанке при закрытии вклада; - на списание платы за услуги банка и иных обязательств; - на списание денег со счёта вклада для погашения задолженности по займам в МСК «Совком Экспресс» или МКК «Хоум Экспресс» при получении банком соответствующего требования. Условия договора, п. 6, 7, 8, 9
-----	--------------------------------	--

Важно запомнить

4.1	Исключения для получения оговорённой суммы вклада	Если у вас есть просроченная задолженность по любым договорам с банком или вы не выполните условия для повышения ставки хотя бы в одном отчётном периоде, то ставка по вкладу не повысится. Также, если сумма всех ваших вкладов в банке превышает 1 500 000 рублей, проценты по повышенной ставке начислят только на сумму в пределах 1 500 000 рублей, на остаток — по базовой ставке. Условия договора, п.п. 5
4.2	Если деньги понадобятся раньше срока	Досрочное снятие не предусмотрено, и если вам понадобится снять сумму раньше, проценты будут пересчитаны по ставке 0,01% годовых за весь период. Если вам уже выплачивали часть процентов, разница в пересчитанных процентах будет удержана из суммы вклада. Условия договора, п.п. 7
4.3	Срок вклада превышает 1 год	Если вы решите воспользоваться своими деньгами до окончания срока 1095 дней (36 месяцев), то проценты будут пересчитаны по ставке вклада до востребования — 0,01% годовых за весь период. Если ранее вам уже выплачивали проценты, разница будет удержана из суммы вклада. Условия договора, п.п. 4 и 7

Внимание: Данная памятка не является договором и не заменяет его.

Памятка подготовлена в справочных целях, то есть предназначена для предварительного ознакомления с ключевыми условиями договора. Полные условия определены в договоре. Памятка автоматически сформирована системой искусственного интеллекта.

Система искусственного интеллекта создаёт памятку на основе типовых условий договора с использованием единых правил их описания. Формулировки памятки могут не отражать все условия договора и характеристики финансового продукта.

С целью оценки всех правовых рисков, которые возникают у вас как у стороны заключаемого договора, необходимо ознакомление с полным текстом договора. Банк России определяет общие подходы к описанию условий договора в памятке и её достоверности, но не участвует в формировании конкретной памятки и не несёт ответственности за корректность отражения условий договора в памятке.

Ответственность по договору, а также в связи с его заключением несут стороны договора.

Сервис ИИ-саммаризации договоров



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

ПРОЗРАЧНОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

Соблюдение рекомендаций Кодекса этики

2.6. Повышение финансовой доступности

ИИ-памятка напрямую работает на повышение финансовой доступности: продукт становится интуитивно понятным для клиентов без специальных финансовых компетенций, включая пожилых людей и клиентов с низким уровнем финансовой грамотности. Резюме выдается в едином формате и обеспечивает сопоставимость условий разных банков – независимо от уровня финансовой грамотности клиента.

4.2. Информирование о применении искусственного интеллекта

Памятка позиционируется как отдельный информационный документ: клиент явно видит, что это резюме, а не договор. В документе указано, что он сформирован автоматически с применением технологий ИИ, договор остается первичным юридически значимым документом.

5.3. Проверка качества искусственного интеллекта

Зафиксирована полная критериальная модель экспертной оценки: пороговые значения корректности, релевантности, устойчивости ответов, проверка на тестовых наборах документов, а также нагрузочное тестирование.

СберКот



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 2.6. Повышение финансовой доступности
- пункт 5.2. Проверка качества наборов данных

О сервисе

СберКот – детский ИИ-ассистент в СберКидс, который помогает ребенку безопасно и понятно разбираться в деньгах, цифровых сервисах и повседневных вопросах.

В отличие от взрослого ассистента в СберБанк Онлайн СберКот работает с аудиторией, у которой сильно различаются словарный запас, стиль общения, уровень самостоятельности и способность формулировать запрос. Поэтому ключевая особенность кейса – не просто генерация ответов, а адаптация языка, тона и сценария диалога под возраст ребенка.

СберКот может объяснять финансовые понятия простым языком, помогать разобраться с накоплениями, тратами, переводами, безопасностью и базовыми правилами обращения с деньгами. При этом он не требует от ребенка точной формулировки вопроса: Ассистент умеет интерпретировать запрос, уточнять детали и вести диалог в привычной для ребенка форме.



Технологии

Тип модели	Генеративный ИИ
Модальность	Естественный текст
Методы и техники	RAG



СберКот



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

Соблюдение рекомендаций Кодекса этики

2.6. Повышение финансовой доступности

Кейс учитывает особенности детской аудитории (6-14 лет) как уязвимой группы пользователей. СберКот адаптирует тон, сложность объяснений и сценарии диалога под определенный возраст ребенка.

При этом учитывается, что дети 6–14 лет не являются одной однородной группой. У них сильно различаются стиль речи, словарный запас, интересы, восприятие юмора и ожидания от общения.

СберКот решает эту задачу через возрастную адаптацию сценариев и подачи:

6–8 лет – «старший друг»

СберКот помогает исследовать мир, объясняет простыми словами, поддерживает игровой контекст и создает ощущение безопасности.

9–12 лет – «наставник»

СберКот общается более живо, помогает разобраться в деньгах, покупках и цифровых привычках, говорит на равных, но сохраняет роль взрослого и надежного проводника.

13–14 лет – «равный собеседник»

СберКот поддерживает самостоятельность, объясняет сложные вещи без снисходительности, помогает разобраться в вопросе и не навязывает готовое решение.

5.2. Проверка качества наборов данных

Для ответов используются заранее подготовленные знания о продуктах, правилах безопасности, финансовой грамотности и сценариях СберКидс. Это важно при работе с детьми, поскольку ответы должны быть не только корректными, но и понятными, безопасными и соответствующими возрасту пользователя.



Внутренние сервисы

ИИ активно применяется финансовыми организациями во внутренних процессах. Такие сервисы позволяют в том числе осуществлять поиск по внутренним базам знаний, анализировать данные (документы, новости, клиентские обращения), готовить тексты, поддерживать разработку ПО и HR-процессы. Сегодня в России уже 51% организаций, применяющих и тестирующих ИИ, используют его в документообороте, при формировании отчетов и подготовке документов, 46% — в ИТ-направлении, включая разработку ПО, 40% — для работы с претензиями и обработки жалоб и 27% в HR-направлении²⁴.

Важно учитывать

Применение ИИ позволяет оптимизировать внутренние процессы финансовых организаций: ускорить выполнение рутинных задач, сократить время поиска информации, снизить нагрузку на сотрудников, а в некоторых случаях косвенно повысить скорость и качество оказания услуг клиентам.

КАК ПРАВИЛО, НИЗКОРИСКОВЫЕ, НО МОГУТ ВЛИЯТЬ НА ИНТЕРЕСЫ КЛИЕНТОВ

Внутренние ИИ-сервисы, как правило, относятся к низкорисковым, поскольку являются лишь вспомогательным инструментом для сотрудников организации.

Однако ключевое значение для оценки риска имеет не факт внутреннего использования ИИ, а контекст его применения. Так, риски могут повышаться, если внутренний ИИ-сервис начинает оказывать косвенное влияние на потребителей финансовых услуг или использовать их данные.

Лучшие практики

Кодекс этики предлагает финансовым организациям широкий перечень рекомендаций по ответственному управлению рисками. На практике им можно следовать по-разному с учетом вида и масштаба деятельности организации, ее организационной структуры, существующей СУР и характера оказываемых услуг.

УЧИТЫВАТЬ КОНТЕКСТ ПРИМЕНЕНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

При внедрении внутренних ИИ-сервисов финансовым организациям следует учитывать контекст их применения.

Например, если внутренний ИИ-сервис используется сотрудником как вспомогательный инструмент при оказании услуг клиенту (помогает подготовить ответ на обращение или разъяснить условия продукта), сотрудника следует уведомлять о рисках и условиях применения такого сервиса, включая возможность ошибок, неполных или устаревших ответов.

 Интеллектуальный внутренний ассистент «Сова» п. 4.2 

Пример информирования о применении ИИ

²⁴ [Применение искусственного интеллекта на финансовом рынке: текущий статус и условия дальнейшего развития](#). Банк России. Ноябрь 2025.

Также при обработке внутренним ИИ-сервисом клиентских данных (например, записи звонков, переписка, документы) следует обеспечить конфиденциальность информации, разграничение доступа и контроль за использованием таких данных.

 Речевая аналитика звонков

п. 5.6 

Пример обеспечения конфиденциальности данных

Интеллектуальный внутренний ассистент «Сова» для сотрудников банка



- ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ
- ПРОЗРАЧНОСТЬ
- БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ
- ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 2.2. Повышение качества оказания услуг
- пункт 4.2. Информирование о применении ИИ
- пункт 5.6. Обеспечение конфиденциальности информации
- пункт 6.10. Реагирование на реализовавшиеся риски ИИ

О сервисе

Сова – корпоративный ИИ-ассистент Совкомбанка, созданный для помощи сотрудникам в получении информации и решении рабочих вопросов. Ассистент доступен во всех корпоративных каналах связи и предоставляет ответы в режиме реального времени. Он заменяет длительный поиск в регламентах и обращения в службы поддержки, снижая нагрузку на сервисные подразделения.

Сова работает на базе локальной нейросети OWI – собственного защищенного решения банка. Чат-бот интегрирован с корпоративными базами знаний, порталами и учетными системами.

По данным внутреннего тестирования, Сова заменяет 60% обращений сотрудников в поддержку за консультацией у более опытных сотрудников, предоставляя ответ в среднем за 7 секунд.



Технологии

- Тип модели: Генеративный ИИ
- Модальность: Естественный текст
- Методы и техники: RAG (векторный поиск), Промпт-инжиниринг, Дообучение (fine-tuning)

Соблюдение рекомендаций Кодекса этики

2.2. Повышение качества оказания услуг

По завершении взаимодействия сотруднику предлагается оценить удовлетворенность взаимодействия с сервисом.

Интеллектуальный внутренний ассистент «Сова» для сотрудников банка



ЧЕЛОВЕКОЦЕНТРИЧНОСТЬ

ПРОЗРАЧНОСТЬ

БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

4.2. Информирование о применении искусственного интеллекта

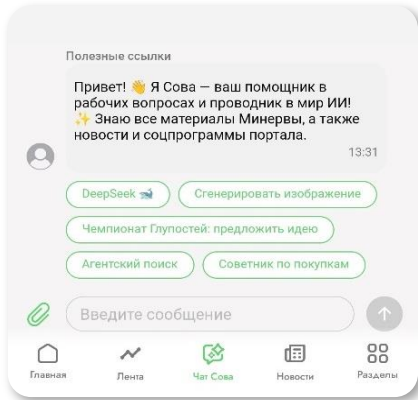
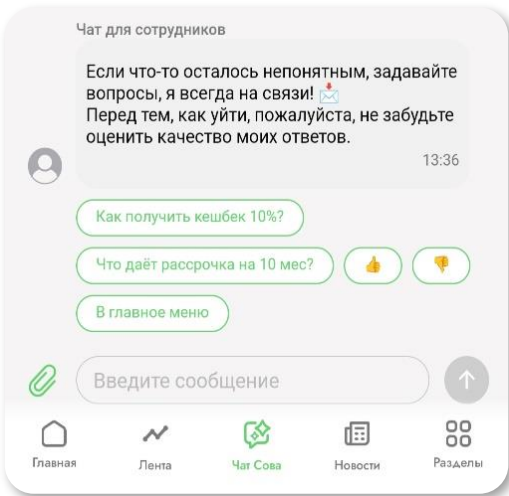
Интерфейс чат-бота, его наименование и формат ответов однозначно указывают на цифровую природу ассистента, позволяя сотруднику однозначно определить, что он взаимодействует с ИИ.

5.6. Обеспечение конфиденциальности информации

Сова работает на базе локального ИИ-решения OWI – данные сотрудников не передаются во внешние системы. Персональные кадровые данные (суммы премий, отпускных, больничных) доступны только в защищенном «приватном режиме»: история чата очищается в течение часа после сессии, что исключает несанкционированный доступ к персональным сведениям. Доступ к тематическим коллекциям фильтруется по ролям сотрудника в рабочих системах.

6.10. Реагирование на реализовавшиеся риски искусственного интеллекта

В системе реализован механизм оценки качества ответов ИИ. Сотрудники могут давать обратную связь на каждый сгенерированный текст, что позволяет команде разработчиков регулярно дообучать модель и повышать точность консультаций.



Речевая аналитика звонков



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Данный сервис оценивался на предмет соблюдения следующих рекомендаций Кодекса этики:

- пункт 5.3. Проверка качества ИИ
- пункт 5.6. Обеспечение конфиденциальности информации
- пункт 6.8. Мониторинг и контроль рисков ИИ

О сервисе

Речевая аналитика на основе ИИ применяется для анализа диалогов в телефонных продажах и клиентском обслуживании. Решение помогает выявлять эффективные практики коммуникации, контролировать соблюдение скриптов и быстрее находить обращения, требующие дополнительного внимания.

Основные задачи:

- контроль качества работы оператора при общении с клиентом, в том числе отработка возражений, соблюдение сценария диалога (скрипта);
- поиск негативных обращений (жалоб);
- выявление признаков мошеннических звонков и нестандартного поведения в диалоге;
- оценка результативности звонка и выделение успешных практик продаж.

При этом ИИ не заменяет сотрудника контроля качества и не оценивает работу оператора. Сервис предварительно размечает и приоритизирует диалоги, а итоговая оценка осуществляется ответственными сотрудниками.



Технологии

Тип модели

Модели глубокого обучения

Генеративный ИИ

Модальность

Естественный текст

Аудио

Речевая аналитика звонков



БЕЗОПАСНОСТЬ, НАДЕЖНОСТЬ И ЭФФЕКТИВНОСТЬ

ОТВЕТСТВЕННОЕ УПРАВЛЕНИЕ РИСКАМИ

Соблюдение рекомендаций Кодекса этики

5.3. Проверка качества искусственного интеллекта

Для модели установлены показатели качества: точность распознавания речи, корректность классификации тематики, полнота выявления негативных обращений и доля ложных срабатываний. На выборке диалогов проводится периодическая проверка качества, после чего сценарии и правила разметки уточняются.

5.6. Обеспечение конфиденциальности информации

Записи звонков, текстовые расшифровки и результаты анализа могут содержать персональные данные и иную информацию ограниченного доступа. В организации действует система технологических и организационных мер по обеспечению безопасности таких данных. Эти данные используются только уполномоченными сотрудниками для целей контроля качества, обучения и улучшения сервиса. Также установлены внутренние требования к хранению, доступу и нераспространению соответствующих данных. Действия с ними и настройками сервиса подлежат внутреннему контролю.

6.8. Мониторинг и контроль рисков искусственного интеллекта

Риски речевой аналитики контролируются через выборочную ручную проверку результатов, анализ ошибок классификации и контроль использования выводов системы. Такой подход позволяет не допускать автоматического применения некорректной оценки к сотрудникам или клиентам.

Приложение

Кодекс этики в сфере разработки
и применения искусственного
интеллекта на финансовом рынке

Кодекс этики в сфере разработки и применения искусственного интеллекта на финансовом рынке

1. Цели и принципы разработки и применения искусственного интеллекта на финансовом рынке

1.1. Целями настоящего Кодекса являются:

- 1) повышение доверия физических и юридических лиц (далее – клиенты) к применению искусственного интеллекта кредитными организациями, иностранными банками, осуществляющими деятельность на территории Российской Федерации через свои филиалы, некредитными финансовыми организациями, лицами, оказывающими профессиональные услуги на финансовом рынке, субъектами национальной платежной системы (далее – организации) при оказании клиентам услуг;
- 2) содействие развитию искусственного интеллекта на финансовом рынке и применению организациями доверенных технологий искусственного интеллекта;
- 3) минимизация рисков, связанных с разработкой и применением искусственного интеллекта на финансовом рынке (далее – риски искусственного интеллекта).

1.2. Для осуществления указанных целей при разработке и применении искусственного интеллекта организациям рекомендуется соблюдать следующие принципы:

- 1) человекоцентричность;
- 2) справедливость;
- 3) прозрачность;
- 4) безопасность, надежность и эффективность;
- 5) ответственное управление рисками.

1.3. Организациям рекомендуется обмениваться опытом, а также участвовать в разработке совместных документов и аналитических материалов, содействующих достижению целей настоящего Кодекса.

1.4. В случае привлечения при разработке и применении искусственного интеллекта организациями лиц, в отношении которых Банк России не осуществляет контроль (надзор), организациям рекомендуется обеспечивать соблюдение такими лицами настоящего Кодекса.

2. Принцип человекоцентричности

2.1. Организациям рекомендуется принимать следующие меры:

- 1) повышение качества оказания услуг;
- 2) предоставление клиентам возможности отказаться от взаимодействия с применением искусственного интеллекта;
- 3) обеспечение возможности пересмотра решения, принятого с применением искусственного интеллекта;
- 4) повышение осведомленности клиентов о принятии решений с применением искусственного интеллекта;
- 5) повышение финансовой доступности.

2.2. В рамках повышения качества оказания услуг организациям рекомендуется осуществлять оценку удовлетворенности клиентов, которым оказаны услуги с применением искусственного интеллекта, а также осуществлять контроль качества оказания таких услуг.	
2.3. В рамках предоставления клиентам возможности отказаться от взаимодействия с применением искусственного интеллекта организациям рекомендуется предоставить клиентам возможность взаимодействовать с сотрудником организации.	
2.4. В рамках обеспечения возможности пересмотра решения, принятого с применением искусственного интеллекта , организациям рекомендуется на основании запроса клиента организовать пересмотр такого решения сотрудником организации.	
2.5. В рамках повышения осведомленности клиентов о принятии решений с применением искусственного интеллекта организациям рекомендуется разъяснять действия, которые клиенту необходимо совершить для принятия искусственным интеллектом решения, на основании которого делается вывод о возможности оказания услуг. Организация вправе не предоставлять указанные разъяснения, если имеются разумные основания полагать, что такое информирование может снизить эффективность применения искусственного интеллекта.	
2.6. В рамках повышения финансовой доступности организациям рекомендуется при разработке и применении искусственного интеллекта учитывать факторы уязвимости клиентов (возраст, образование, ограниченные возможности и другие) и возможное влияние указанных факторов на оказание таким клиентам услуг.	

2. Принцип справедливости

3.1. Организациям рекомендуется принимать следующие меры:

- 1) обоснованность использования персональных данных о клиентах;
- 2) недопустимость использования дискриминационных факторов;
- 3) обеспечение недискриминационного характера набора данных.

3.2. В рамках обоснованности использования персональных данных о клиентах организациям рекомендуется при разработке искусственного интеллекта использовать персональные данные клиентов только в случаях, когда их использование требуется для существенного повышения эффективности применения искусственного интеллекта при оказании услуг.	
---	-------------------------------------

3.3. В рамках недопустимости использования дискриминационных факторов организациям рекомендуется при принятии решений с применением искусственного интеллекта не учитывать, в частности, следующие факторы:

- 1) национальная, языковая и расовая принадлежность;
- 2) принадлежность к политическим партиям, общественным объединениям;
- 3) вероисповедание и отношение к религии.



3.4. В рамках обеспечения недискриминационного характера наборов данных организациям рекомендуется при разработке искусственного интеллекта проверять полноту, актуальность и достоверность данных о социальных группах клиентов, при оказании услуг которым организация вправе применять искусственный интеллект.

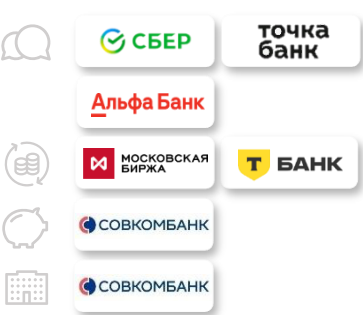


4. Принцип прозрачности

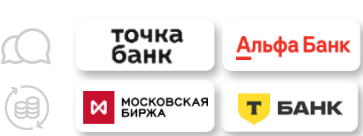
4.1. Организациям рекомендуется принимать следующие меры:

- 1) информирование о применении искусственного интеллекта;
- 2) раскрытие информации об искусственном интеллекте;
- 3) маркировка информации, созданной с применением больших генеративных моделей.

4.2. В рамках информирования о применении искусственного интеллекта организациям рекомендуется сообщать клиентам о применении искусственного интеллекта при оказании им услуг, если применение искусственного интеллекта неочевидно из обстоятельств.



4.3. В рамках раскрытия информации об искусственном интеллекте организациям рекомендуется предоставлять клиентам полную и достоверную информацию о рисках искусственного интеллекта и условиях применения организацией искусственного интеллекта.



4.4. В рамках маркировки информации, созданной с применением больших генеративных моделей, организациям рекомендуется маркировать сведения, в том числе изображения, аудио- и видеоматериалы, созданные с применением больших генеративных моделей, за исключением следующих случаев:

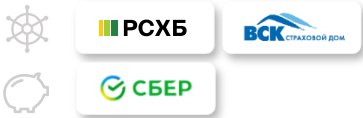
- 1) большие генеративные модели применяются только для редактирования сведений, которое не влечет существенного изменения содержания указанных сведений;
- 2) применение больших генеративных моделей очевидно из обстоятельств;
- 3) отсутствует риск причинения вреда клиентам.

5. Принцип безопасности, надежности и эффективности

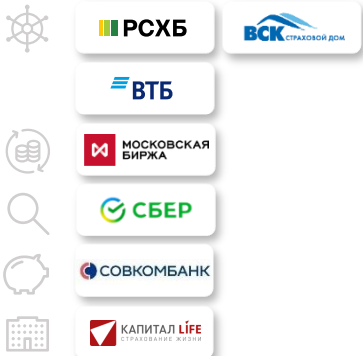
5.1. Организациям рекомендуется принимать следующие меры:

- 1) проверка качества наборов данных;
- 2) проверка качества искусственного интеллекта;
- 3) мониторинг качества искусственного интеллекта;
- 4) обеспечение информационной безопасности;
- 5) обеспечение конфиденциальности информации;
- 6) обеспечение непрерывности деятельности.

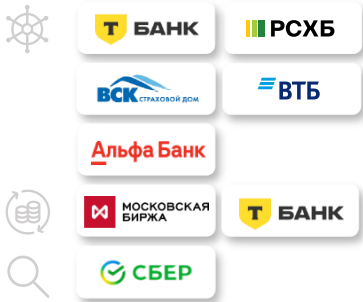
5.2. В рамках проверки качества наборов данных для разработки искусственного интеллекта организациям рекомендуется проверять используемые наборы данных, в том числе на предмет их достоверности, точности и полноты, собирать и хранить информацию об источниках наборов данных, разметке данных, исправлять неточности в наборах данных и осуществлять их актуализацию.



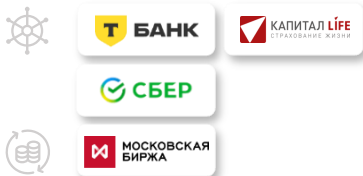
5.3. В рамках проверки качества искусственного интеллекта организациям рекомендуется установить показатели качества искусственного интеллекта и проверять при разработке искусственного интеллекта соответствие искусственного интеллекта установленным организацией показателям качества, в том числе посредством валидации, добровольной сертификации.



5.4. В рамках мониторинга качества искусственного интеллекта организациям рекомендуется регулярно осуществлять проверку применяемого искусственного интеллекта на предмет соответствия установленным организацией показателям качества.



5.5. В рамках обеспечения информационной безопасности организациям рекомендуется учитывать риски искусственного интеллекта, связанные с нарушением информационной безопасности и операционной надежностью, и осуществлять оценку достаточности имеющихся и планируемых к реализации мер противодействия угрозам информационной безопасности на всех этапах разработки и применения искусственного интеллекта, в том числе в отношении наборов данных.



5.6. В рамках обеспечения конфиденциальности информации организациям рекомендуется разработать систему технологических и организационных мер по обеспечению безопасности данных ограниченного доступа, используемых при разработке и применении искусственного интеллекта, в том числе меры, связанные с их обезличиванием и противодействием их несанкционированному распространению при применении больших генеративных моделей сотрудниками организации и ее клиентами.	
5.7. В рамках обеспечения непрерывности деятельности организациям рекомендуется обеспечить выполнение значимых функций, осуществляемых с применением искусственного интеллекта, сотрудниками организации в случае временной невозможности применения искусственного интеллекта.	
6. Принцип ответственного управления рисками	
6.1. Организациям рекомендуется организовать и осуществлять управление рисками при разработке и применении искусственного интеллекта в рамках общей системы управления рисками в организации.	
6.2. В рамках организации управления рисками при разработке и применении искусственного интеллекта организациям рекомендуется с учетом особенностей вида и масштаба ее деятельности, организационной структуры, существующей системы управления рисками и характера услуг, оказываемых организацией, разработать и утвердить правила управления рисками искусственного интеллекта, являющиеся составной частью применяемых документов об управлении рисками в организации, или в виде самостоятельного документа, а также назначить лицо или подразделение, ответственное за соблюдение указанных правил и соответствие их настоящему Кодексу.	
6.3. Организациям рекомендуется при управлении рисками искусственного интеллекта обеспечить следующие процессы: 1) учет применяемых моделей искусственного интеллекта; 2) выявление рисков искусственного интеллекта; 3) оценка и присвоение уровня риска искусственного интеллекта; 4) мониторинг и контроль рисков искусственного интеллекта; 5) минимизация выявленных рисков искусственного интеллекта; 6) реагирование на реализовавшиеся риски искусственного интеллекта (далее – риск-события); 7) ведение базы риск-событий.	



6.4. В рамках учета применяемых моделей искусственного интеллекта организациям рекомендуется вести централизованный учет моделей искусственного интеллекта и случаев их применения с учетом присвоенного уровня риска искусственного интеллекта.	   
6.5. В рамках выявления рисков искусственного интеллекта организациям рекомендуется документировать информацию о порядке разработки искусственного интеллекта, включая описание процесса разработки искусственного интеллекта, использованных при разработке наборов данных, сведения об условиях применения искусственного интеллекта, лицах, ответственных за применение искусственного интеллекта в организации, а также анализировать указанную информацию на предмет наличия рисков искусственного интеллекта.	  
6.6. В рамках оценки и присвоения уровня риска искусственного интеллекта организациям рекомендуется сформировать систему факторов риска и уровней риска, присваивать уровни риска разрабатываемому и применяемому искусственному интеллекту с учетом установленных факторов риска.	   
6.7. При присвоении уровня риска искусственного интеллекта рекомендуется учитывать следующие факторы риска: 1) сфера применения искусственного интеллекта (например, применение искусственного интеллекта при оказании услуг клиентам, в системах управления рисками, при управлении имуществом организации, в критически важных процессах организации, в процессах, для которых установлены требования операционной надежности); 2) использование при разработке искусственного интеллекта данных ограниченного доступа; 3) размер убытков или ущерб деловой репутации, которые могут быть причинены организации в случае реализации риска; 4) количество клиентов, при оказании услуг которым применяется искусственный интеллект; 5) объяснимость решений, принимаемых искусственным интеллектом; 6) применение искусственного интеллекта, разработанного третьим лицом; 7) использование наборов данных, полученных от третьих лиц, или наборов данных, находящихся в открытом доступе в информационно- телекоммуникационной сети «Интернет»; 8) наличие риск-событий, связанных с применением искусственного интеллекта.	 

6.8. В рамках мониторинга и контроля рисков искусственного интеллекта организациям рекомендуется регулярно проверять процессы, связанные с разработкой и применением искусственного интеллекта на предмет наличия рисков, а также осуществлять предварительный, текущий, последующий контроль рисков с учетом специфики конкретного случая применения искусственного интеллекта.	
6.9. В рамках минимизации выявленных рисков искусственного интеллекта организациям рекомендуется определять меры, направленные на снижение уровня риска искусственного интеллекта, и лиц, ответственных за реализацию указанных мер, а также обеспечивать контроль сотрудников организации за решениями искусственного интеллекта, которому присвоен высокий уровень риска.	
6.10. В рамках реагирования на реализовавшиеся риски искусственного интеллекта организациям рекомендуется определять типы риск-событий и принимать меры, направленные на минимизацию последствий риск-событий, а также обеспечить своим сотрудникам и клиентам возможность направления жалоб и сообщений о риск-событиях.	
6.11. В рамках ведения базы риск-событий организациям рекомендуется выявлять риск-события и осуществлять их регистрацию с описанием риск-события, обстоятельств, повлекших наступление риск-события, а также приведением сведений о мероприятиях, направленных на минимизацию последствий риск-события, и статусе исполнения таких мероприятий.	

