ЦЕНТРАЛЬНЫЙ БАНК РОССИЙСКОЙ ФЕДЕРАЦИИ
(БАНК РОССИИ)

Методические рекомендации Банка России по обеспечению информационной безопасности при разработке и применении искусственного интеллекта на финансовом рынке

16.06.2026 № 3-МР

Глава 1. **Общие положения**

1.1. Настоящие Методические рекомендации разработаны в целях обеспечения единства подходов кредитных организаций, иностранных банков, осуществляющих деятельность на территории Российской Федерации через свои филиалы, некредитных финансовых организаций, лиц, оказывающих профессиональные услуги на финансовом рынке, субъектов национальной платежной системы (далее – организации) к реализации принципа безопасности, надежности и эффективности в части обеспечения информационной безопасности при разработке и применении искусственного интеллекта (далее – ИИ) на финансовом рынке в соответствии с Кодексом этики в сфере разработки и применения искусственного интеллекта на финансовом рынке[1] (далее – Кодекс этики).

1.2. Настоящие Методические рекомендации содержат рекомендации по минимизации рисков, связанных с нарушением информационной безопасности и операционной надежности при разработке и применении ИИ на финансовом рынке, разработке модели угроз безопасности информации системы ИИ и определении состава и содержания

[1] Информационное письмо Банка России о Кодексе этики в сфере разработки и применения искусственного интеллекта на финансовом рынке от 09.07.2025 № ИН-016-13/91.

мер защиты системы ИИ, разработке политики обеспечения информационной безопасности при разработке и применении ИИ, а также рекомендации по обеспечению информационной безопасности при использовании сервисов ИИ поставщиков услуг и (или) компонентов технологий ИИ с открытым исходным кодом.

1.3. В целях настоящих Методических рекомендациях используются следующие термины и определения:

галлюцинации ИИ – результат генерации выходных данных, которые содержат бессмысленные, некорректные или ошибочные суждения, которые могут выглядеть правдоподобными;

дрейф данных – изменение характеристик и свойств данных со временем, которое может привести к снижению точности и эффективности моделей ИИ, основанных на таких данных, а также некорректным выходным данным и решениям;

прямое внедрение запроса – вредоносный запрос непосредственно к модели ИИ, в том числе через различные интерфейсы взаимодействия;

непрямое внедрение запроса – запрос, инициирующий обращение модели ИИ к внешним источникам, которые могут контролироваться нарушителем (веб-сайты, файлы и прочее);

«отравленный» набор данных – набор данных, подвергнутый умышленной модификации, использование которого влечет нарушение функционирования модели ИИ.

Термины «искусственный интеллект» («ИИ»), «технологии ИИ», «модель ИИ», «набор данных» используются в значениях, установленных Национальной стратегией развития искусственного интеллекта на период до 2030 года[2].

Термины «система ИИ», «объяснимость», «предсказуемость», «надежность» используются в значениях, установленных национальным

[2] Утверждена Указом Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации».

стандартом Российской Федерации ГОСТ Р 71476-2024 (ИСО/МЭК 22989:2022) «Искусственный интеллект. Концепции и терминология искусственного интеллекта».

Термин «качество» используется в значении, установленном национальным стандартом Российской Федерации ГОСТ Р 59898-2021 «Оценка качества систем искусственного интеллекта. Общие положения».

Глава 2. **Риски ИИ, связанные с нарушением информационной безопасности и операционной надежности**

2.1. В целях соблюдения принципа безопасности, надежности и эффективности в части обеспечения информационной безопасности при разработке и применении ИИ на финансовом рынке, а также с учетом принципа ответственного управления рисками организациям при осуществлении деятельности на финансовом рынке с использованием технологий ИИ рекомендуется учитывать следующие риски ИИ, связанные с нарушением информационной безопасности и операционной надежности (далее – риски информационной безопасности ИИ):

2.1.1. Риски, связанные с управлением данными, обусловленные использованием для обучения модели ИИ «отравленных» наборов данных, неактуальных и (или) некорректных наборов данных, влекущие за собой нарушение функционирования системы ИИ.

2.1.2. Риски нарушения конфиденциальности данных, обусловленные в том числе реализацией угроз безопасности технологий ИИ, влекущие за собой доступность информации лицам, не имеющим право на ее получение, нарушение авторских прав и (или) интеллектуальной собственности.

2.1.3. Риски нарушения функционирования модели ИИ, обусловленные реализацией угроз безопасности технологий ИИ, влекущие за собой снижение качества системы ИИ, в том числе галлюцинации ИИ, дрейф данных и иные

варианты деградации модели ИИ, приводящие к некорректным выводам и решениям модели ИИ.

2.1.4. Риски отсутствия достаточной объяснимости и (или) предсказуемости действий модели ИИ, обусловленные сложностью интерпретации результатов исполнения модели ИИ, приводящие к некорректным выводам и решениям модели ИИ.

2.1.5. Риски, связанные с привлечением поставщиков услуг и (или) использованием компонентов технологий ИИ с открытым исходным кодом, обусловленные умышленными или неосторожными действиями сотрудников поставщика услуг, наличием слабостей и (или) уязвимостей в таких компонентах.

2.1.6. Риски, связанные с нарушением операционной надежности организации, обусловленные умышленными или неосторожными действиями в отношении системы ИИ, приводящие к прерыванию процессов основной деятельности организации.

2.2. Организациям рекомендуется учитывать риски информационной безопасности ИИ в следующих случаях:

при определении целесообразности разработки и (или) применения ИИ в отдельных задачах или бизнес-процессах в целом;

при решении вопроса об использовании конкретных технологий ИИ;

при обеспечении информационной безопасности;

при осуществлении управления рисками ИИ.

2.3. Организациям рекомендуется проводить оценку рисков информационной безопасности ИИ в том числе на основании факторов риска, указанных в пункте 6.7 Кодекса этики.

2.4. При оценке рисков организациям рекомендуется учитывать, что реализация рисков информационной безопасности ИИ может в том числе повлечь за собой:

нарушение прав, свобод и законных интересов физических лиц;

нарушение операционной деятельности организации, в том числе в результате неверного управленческого решения, принятого с использованием ИИ;

причинение убытков;

причинение вреда деловой репутации организации;

нарушение функционирования информационных систем и (или) сервисов иных организаций, в том числе в рамках цепочки поставок;

нарушение стабильности финансовой системы.

2.5. В случае использования ИИ для выполнения операций в автоматическом режиме в критически важных процессах (например, в платежных процессах, процессах учетных систем, которые отражают факты основной деятельности организации), когда риски информационной безопасности ИИ оценены организацией как высокие, организации рекомендуется реализовать валидацию результатов операций, выполненных ИИ в автоматическом режиме, человеком с возможностью изменения таких результатов.

Глава 3. **Разработка модели угроз безопасности информации системы ИИ и определение состава и содержания мер защиты системы ИИ**

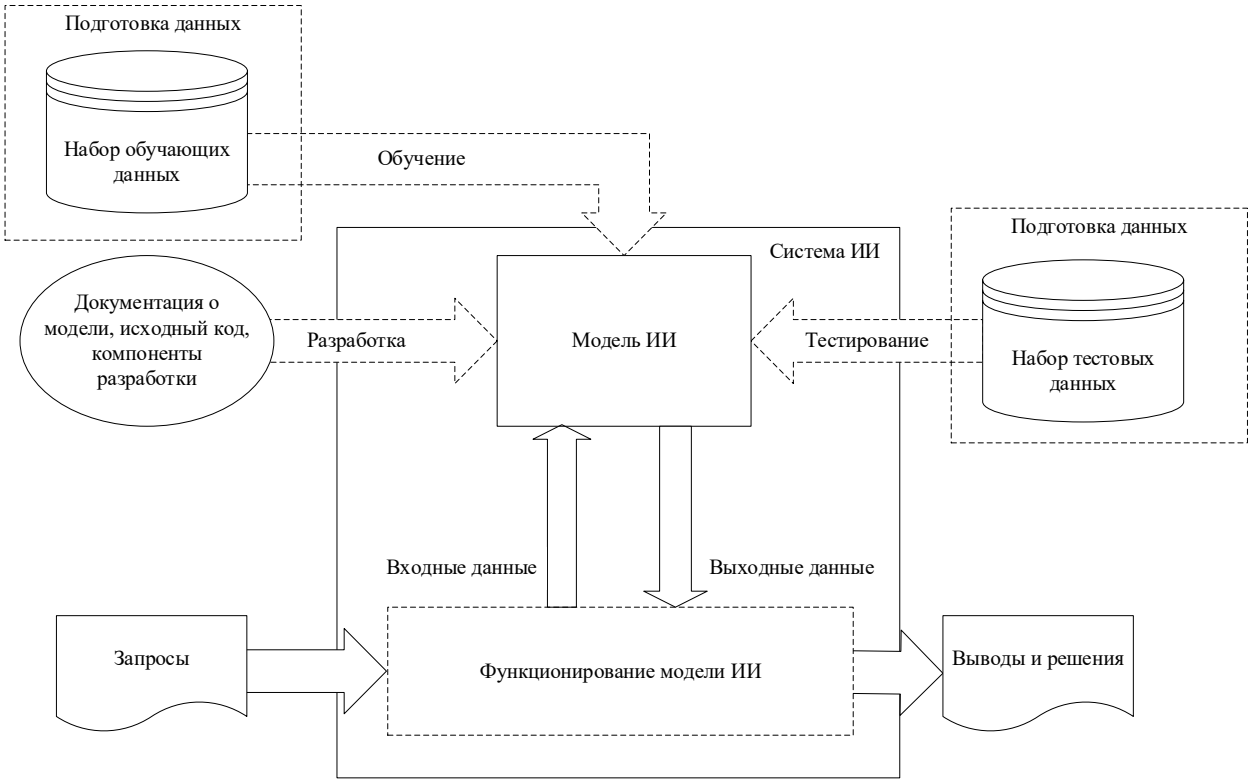
3.1. В целях определения актуальных угроз безопасности информации системы ИИ организациям рекомендуется разработать модель угроз безопасности информации системы ИИ (далее – Модель угроз).

3.2. При разработке Модели угроз организациям рекомендуется использовать Методику оценки угроз безопасности информации[3] и учитывать следующее:

3.2.1. Результаты оценки рисков информационной безопасности ИИ.

[3] Для целей настоящих Методических рекомендаций используется «Методический документ. Методика оценки угроз безопасности информации» (утвержден ФСТЭК России 05.02.2021).

3.2.2. Этапы разработки и применения ИИ (подготовка данных, разработка модели ИИ, обучение и тестирование модели ИИ, функционирование модели ИИ), функциональное представление которых приведено на схеме:



3.2.3. Возможные угрозы безопасности информации, специфичные для технологий ИИ, включая угрозы:

нарушения функционирования («обхода») средств, реализующих технологии ИИ;

искажения («отравления») обучающих данных;

раскрытия информации о модели ИИ;

хищения обучающих данных;

модификации модели ИИ, в том числе изменения архитектуры, последовательности взаимодействий;

приведения модели ИИ в состояние «отказ в обслуживании»;

манипуляции поведением модели ИИ;

подмены модели ИИ.

3.2.4. Актуальность как внешнего, так и внутреннего нарушителя.

3.2.5. Структурно-функциональные характеристики системы ИИ в информационной инфраструктуре организации.

3.2.6. Угрозы безопасности информации, возможные для информационной инфраструктуры организации, на базе которой функционирует система ИИ.

3.3. Организациям рекомендуется учитывать следующие возможные способы реализации угроз безопасности информации, специфичные для технологий ИИ, в частности:

атака с использованием фаззинга[4] (ввод различных входных данных в целях выявления уязвимостей модели ИИ и (или) приведения модели ИИ в состояние «отказ в обслуживании»);

внедрение бэкдора (закладки) в модели ИИ в целях нарушения ее функционирования или достижения иных негативных последствий в соответствии с ожиданиями нарушителя;

модификация алгоритма обучения (нарушение логического алгоритма обучения модели ИИ, приводящее к ее модификации);

извлечение данных (получение нарушителем конфиденциальной информации путем взаимодействия с моделью ИИ);

вредоносная «инъекция» (модификация или обход системных инструкций модели ИИ с использованием специально подготовленных входных данных в целях изменения поведения модели ИИ, осуществленные путем прямого или непрямого внедрения запроса);

атака типа «губка» (ввод вредоносных входных данных в модель ИИ, обработка которых приводит к снижению производительности вычислительных ресурсов);

---

[4] Данный способ может являться промежуточным шагом при реализации более вредоносных способов воздействия.

«отравление» обучающих данных (модификация набора обучающих данных в целях нарушения функционирования модели ИИ, в том числе модификация данных в наборах данных, добавление вредоносных данных, удаление данных);

«состязательные атаки» (атаки, направленные на получение неверных выводов и решений модели ИИ путем ввода искаженных, в том числе «отравленных», входных данных).

3.4. Для определения возможных сценариев реализации угроз безопасности информации, специфичных для технологий ИИ, организациям рекомендуется использовать, в частности, тактики и соответствующие им техники, указанные в приложении 1 к настоящим Методическим рекомендациям.

3.5. Для определения соотношения возможных угроз безопасности, специфичных для технологий ИИ, и этапов разработки и применения ИИ организациям рекомендуется руководствоваться соотношением таких этапов и угроз, указанным в приложении 2 к настоящим Методическим рекомендациям.

3.6. Для определения соотношения возможных угроз информационной безопасности, специфичных для технологий ИИ, и возможных способов реализации (возникновения) указанных угроз организациям рекомендуется руководствоваться соотношением угроз и способов, указанным в приложении 3 к настоящим Методическим рекомендациям.

3.7. В целях нейтрализации актуальных угроз информационной безопасности, специфичных для технологий ИИ, организациям рекомендуется реализовать процесс «Безопасность и защита данных»[5], разделив его на следующие подпроцессы, соответствующие этапам разработки и применения ИИ:

обеспечение безопасности при подготовке данных;

---

[5] Национальный стандарт Российской Федерации ГОСТ Р 71539-2024 (ИСО/МЭК 5338:2023) «Искусственный интеллект. Процессы жизненного цикла системы искусственного интеллекта».

обеспечение безопасности при разработке модели ИИ;

обеспечение безопасности при обучении и тестировании модели ИИ;

обеспечение безопасности при функционировании модели ИИ.

3.8. Для каждого из подпроцессов, указанных в пункте 3.7 настоящей главы, организациям рекомендуется определить меры защиты, направленные на нейтрализацию актуальных угроз безопасности технологий ИИ, с учетом мер защиты, указанных в приложении 4 к настоящим Методическим рекомендациям. При этом организациям рекомендуется соблюдать пропорциональность принимаемых мер защиты выявленным рискам и масштабам последствий от их реализации.

3.9. При соотношении мер защиты, направленных на нейтрализацию актуальных угроз безопасности технологий ИИ, и актуальных угроз безопасности технологий ИИ организациям рекомендуется руководствоваться соотношением мер защиты и возможных угроз безопасности, указанным в приложении 3 к настоящим Методическим рекомендациям.

3.10. Организациям рекомендуется определять в проектной и эксплуатационной документации сведения о реализации мер защиты систем ИИ, а также проводить оценку эффективности реализованных мер защиты.

Глава 4. **Политика обеспечения информационной безопасности при разработке и применении ИИ**

4.1. В целях минимизации рисков, связанных с разработкой и применением ИИ, организациям рекомендуется разработать или дополнить политику обеспечения информационной безопасности при разработке и применении ИИ (далее – Политика) с учетом основных положений, указанных в приложении 5 к настоящим Методическим рекомендациям, в виде отдельного документа или составной части документа по обеспечению информационной безопасности в организации, а также реализовать Политику в организации.

4.2. Организациям рекомендуется возложить разработку Политики и контроль за ее реализацией на заместителя руководителя организации, ответственного за обеспечение информационной безопасности в организации, в том числе ответственного за обнаружение, предупреждение, ликвидацию последствий компьютерных атак и реагирование на компьютерные инциденты[6].

4.3. Организациям рекомендуется разработать внутренние документы, предусматривающие порядок контроля за реализацией требований Политики, в том числе аудита, а также порядок информирования органов управления организации о нарушениях требований Политики.

Глава 5. Обеспечение информационной безопасности при использовании сервисов ИИ поставщиков услуг и (или) компонентов технологий ИИ с открытым исходным кодом

5.1. В целях минимизации рисков информационной безопасности при использовании сервисов ИИ поставщиков услуг и организации работ с цепочками поставок организациям рекомендуется руководствоваться положениями стандарта Банка России «Обеспечение информационной безопасности организаций банковской системы Российской Федерации. Управление риском нарушения информационной безопасности при аутсорсинге» СТО БР ИББС-1.4-2018, введенном в действие приказом Банка России от 06.03.2018 № ОД-568.

5.2. В целях обеспечения безопасного использования данных, моделей ИИ поставщиков услуг и (или) компонентов технологий ИИ с открытым исходным кодом организациям рекомендуется обеспечить

---

[6] Указанная роль предусмотрена постановлением Правительства Российской Федерации от 15.07.2022 № 1272 «Об утверждении типового положения о заместителе руководителя органа (организации), ответственном за обеспечение информационной безопасности в органе (организации), и типового положения о структурном подразделении в органе (организации), обеспечивающем информационную безопасность органа (организации)».

доверие к таким данным, моделям и (или) компонентам в отношении информационной безопасности (далее – доверие) в соответствии с национальным стандартом Российской Федерации ГОСТ Р 59276-2020 «Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения».

5.3. В рамках обеспечения доверия организациям рекомендуется разрабатывать собственные методики оценки доверия в отношении безопасности данных, моделей ИИ поставщиков услуг и (или) компонентов технологий ИИ с открытым исходным кодом (далее – методика оценки доверия).

5.4. При разработке методики оценки доверия организациям рекомендуется учитывать в том числе следующие факторы оценки:

наличие оценки соответствия объектов информационно-коммуникационной инфраструктуры, на базе которых размещаются наборы данных и (или) модель ИИ, требованиям безопасности информации, установленным законодательством Российской Федерации, в том числе нормативными актами Банка России;

участие модели ИИ в программе поиска уязвимостей (Bug Bounty);

наличие Модели угроз, учитывающей возможности как внешнего, так и внутреннего нарушителя;

реализация мер защиты для модели ИИ, определенных в проектной и эксплуатационной документации;

регулярное предоставление поставщиком услуг отчетов об аудите и проверках состояния информационной безопасности информационно-коммуникационной инфраструктуры и сервисов систем ИИ;

регулярное предоставление поставщиком услуг отчетов об аудите лицензий;

наличие спецификации программного обеспечения;

регулярное предоставление поставщиком услуг отчетов об анализе уязвимостей и тестировании на проникновение;

осуществление постоянного мониторинга информационно-коммуникационной инфраструктуры поставщика услуг, обеспечение реагирования на инциденты защиты информации и своевременного информирования организации о зарегистрированных инцидентах;

наличие подтверждения внедрения и использования процессов безопасного жизненного цикла[7] при разработке систем и (или) сервисов ИИ;

наличие оценки соответствия или сертификации по требованиям безопасности информации для процессов безопасной разработки модели ИИ;

реализация требований безопасности к использованию компонентов и (или) моделей ИИ с открытым исходным кодом, в том числе ведение их учета с указанием используемых агентов, плагинов, интерфейсов взаимодействия и иных сведений;

предоставление результатов отслеживания источников данных;

наличие оценки рисков информационной безопасности, проведенной организацией самостоятельно.

5.5. При использовании данных и (или) моделей ИИ поставщиков услуг организациям рекомендуется обеспечить целостность таких данных и моделей по всей цепочке поставки путем использования средств контроля целостности, прошедших процедуру оценки соответствия в системе сертификации ФСТЭК России.

5.6. В целях обучения модели ИИ поставщиком услуг с использованием наборов данных, принадлежащих организации, организациям рекомендуется передавать поставщику услуг «очищенные» наборы данных или передавать синтетические, обезличенные, специально подготовленные данные.

5.7. При заключении договора на предоставление сервисов ИИ с поставщиком услуг организациям рекомендуется включать в такой договор положения об ответственности поставщика услуг за нарушения

---

[7] Национальный стандарт Российской Федерации ГОСТ Р 56939-2024 «Защита информации. Разработка безопасного программного обеспечения. Общие требования», методический документ Банка России («Профиль защиты прикладного программного обеспечения автоматизированных систем и приложений кредитных организаций и некредитных финансовых организаций»).

информационной безопасности предоставляемых сервисов ИИ, а также обязанность поставщика услуг по своевременному информированию о выявленных уязвимостях и инцидентах информационной безопасности.

Глава 6. Заключительные положения

Настоящие Методические рекомендации подлежат размещению на официальном сайте Банка России в информационно-телекоммуникационной сети «Интернет».

Заместитель Председателя
Банка России Г.А. Зубарев

Приложение 1
к Методическим рекомендациям Банка России
по обеспечению информационной безопасности
при разработке и применении искусственного
интеллекта на финансовом рынке

Возможные тактики и техники реализации (возникновения) угроз информационной безопасности технологий ИИ

№ п/п	Тактика	Техника	Описание техники
1	Сбор информации о системе ИИ	Сбор общедоступной информации о результатах анализа известных уязвимостей моделей ИИ	Нарушитель имеет возможность провести поиск известных уязвимостей модели ИИ в открытых источниках
		Сбор информации в открытых репозиториях	Нарушитель имеет возможность осуществить поиск информации в открытых репозиториях приложений во время подготовки атаки
		Сбор информации о модели	Нарушитель имеет возможность выявить информацию о модели ИИ. Такая информация может быть получена из документации либо с использованием тщательно составленных образцов данных и анализа ответов модели ИИ. Нарушитель формирует способы атаки исходя из определенного вида модели ИИ
2	Получение первоначального доступа к компонентам системы ИИ	Компрометация цепочки поставок модели ИИ	Нарушитель может получить первоначальный доступ к системе, скомпрометировав компоненты цепочки поставок. Примеры компрометаций: компрометация аппаратного обеспечения;

№ п/п	Тактика	Техника	Описание техники
			компрометация данных; компрометация программного обеспечения; компрометация модели ИИ; компрометация сторонних сервисов
		Публикация вредоносных наборов данных	Нарушитель имеет возможность осуществить «отравление» данных и опубликовать их на общедоступных ресурсах, для того чтобы в набор обучающих данных были включены «отравленные» данные
		Модификация данных в наборах данных	Нарушитель имеет возможность изменить данные и (или) метки в наборе обучающих данных, что приведет к неправильному обучению модели и неверным выводам и решениям
		Добавление вредоносных данных	Нарушитель имеет возможность включить в набор обучающих данных поддельные или ложные примеры, что может привести к неправильному обучению модели
		Удаление данных	Нарушитель имеет возможность удалить важные примеры из набора обучающих данных, что может снизить устойчивость модели ИИ и нарушить ее функционирование
3	Внедрение и исполнение вредоносного программного обеспечения в системе ИИ	Обход системных инструкций	Нарушитель имеет возможность создать входные данные, которые могут обойти системные инструкции модели ИИ
		Модификация данных в наборах данных	Нарушитель имеет возможность изменить данные и (или) метки в наборе обучающих данных, что приведет к неправильному обучению модели и неверным выводам и решениям
		Добавление вредоносных данных	Нарушитель имеет возможность включить в набор обучающих данных поддельные или ложные примеры, что может привести к неправильному обучению модели

№ п/п	Тактика	Техника	Описание техники
		Удаление данных	Нарушитель имеет возможность удалить важные примеры из набора обучающих данных, что может снизить устойчивость модели ИИ и нарушить ее функционирование
4	Закрепление (сохранение доступа) в системе ИИ	Создание бэкдора (закладки) в модели ИИ	Нарушитель имеет возможность внедрить бэкдор (закладки) в модель ИИ. Модель ИИ с внедренным бэкдором (закладками) работает в соответствии с ожидаемым поведением в стандартных условиях, но при использовании бэкдора (закладок) будет формировать результат, ожидаемый нарушителем
		Модификация алгоритма обучения модели ИИ	Нарушитель имеет возможность модифицировать алгоритм обучения, внедрив уязвимости и (или) изменив параметры
5	Сбор и вывод из системы ИИ информации, необходимой для дальнейших действий при реализации угроз безопасности информации или реализации новых угроз	Кража информационных ресурсов	Нарушитель имеет возможность похитить информационные ресурсы, используемые при разработке и эксплуатации модели ИИ: документацию о модели ИИ, программный код, наборы обучающих и тестовых данных, входные и выходные данные, а также иные сведения, используемые в технологиях ИИ
		Перехват потоков данных	Нарушитель имеет возможность перехватить поток данных между легитимным пользователем и моделью ИИ
6	Несанкционированный доступ и (или) воздействие на информационные ресурсы или компоненты систем ИИ, приводящие к негативным последствиям	Отказ в обслуживании при манипуляции данными	Нарушитель, имеет возможность манипулировать данными в наборах обучающих данных, формировать вредоносные запросы (например, атака типа «губка») или управлять потоком запросов в модель ИИ, что может привести к состоянию «отказ в обслуживании» (повышение нагрузки на вычислительные ресурсы и увеличение времени, затраченного на проверку и исправление неправильных выводов)

№ п/п	Тактика	Техника	Описание техники
		Кража модели	Нарушитель имеет возможность создать копию модели ИИ путем копирования результатов входных и выходных данных целевой модели и последующего обучения копии модели на основе полученных результатов
		Инверсия модели	Нарушитель имеет возможность извлечь конфиденциальную информацию из набора обучающих данных путем анализа выходных данных модели ИИ и восстановления информации
		Обход модели ИИ	Нарушитель имеет возможность создать состязательные атаки, при которых используются специальные входные данные для модели ИИ, созданные путем добавления специально подготовленного «шума» к исходным входным данным в целях искажения выходных данных модели. Эффекты могут быть различными: неправильная классификация, отсутствие обнаружения и иные

Приложение 2
к Методическим рекомендациям Банка России
по обеспечению информационной безопасности
при разработке и применении искусственного
интеллекта на финансовом рынке

Соотношение этапов разработки и применения ИИ и возможных угроз безопасности, специфичных для технологий ИИ

Угрозы безопасности, специфичные для технологий ИИ	Этапы разработки и применения ИИ			
	подготовка данных	разработка	обучение и тестирование	функционирование
Нарушение функционирования («обхода») средств, реализующих технологии ИИ				V
Искажение («отравление») обучающих данных	V			
Раскрытие информации о модели ИИ				V
Хищение обучающих данных	V			V
Модификация модели ИИ, в том числе изменение архитектуры, последовательности взаимодействий		V	V	
Приведение модели ИИ в состояние «отказ в обслуживании»				V
Манипуляция поведением модели ИИ				V
Подмена модели ИИ		V		V

Приложение 3
к Методическим рекомендациям Банка России
по обеспечению информационной безопасности
при разработке и применении искусственного
интеллекта на финансовом рынке

Соотношение возможных угроз информационной безопасности технологий ИИ и возможных способов реализации таких угроз и мер защиты

№ п/п	Возможная угроза безопасности технологий ИИ	Описание возможной угрозы безопасности технологий ИИ	Возможные способы реализации угроз	Меры защиты
1	Угроза нарушения функционирования («обхода») средств, реализующих технологии ИИ	Угроза заключается в возможности выполнить состязательную атаку в отношении модели ИИ в целях получения ошибочного вывода или решения	Атака с использованием фаззинга. Состязательные атаки	Контроль и очистка аномалий во входных и выходных данных
				Шифрование входных и выходных данных
				Использование методов повышения устойчивости модели ИИ к состязательным атакам
				Ограничение потоков данных
2	Угроза искажения («отравления») обучающих данных	Угроза заключается в возможности изменить набор обучающих и (или) тестовых данных таким образом, чтобы модель ИИ принимала ошибочные выводы или решения, соответствующие ожиданиям нарушителя	«Отравление» набора обучающих данных. Бэкдоры (закладки)	Тестирование модели ИИ на предмет «отравления»
				Контроль и очистка наборов обучающих и тестовых данных от аномалий в данных
				Использование методов повышения устойчивости модели ИИ к состязательным атакам (ансамблевые методы)
				Шифрование данных, передаваемых за пределы контролируемой зоны

№ п/п	Возможная угроза безопасности технологий ИИ	Описание возможной угрозы безопасности технологий ИИ	Возможные способы реализации угроз	Меры защиты
		Под изменением набора обучающих данных понимается: модификация меток в наборах данных; добавление новых и (или) «мусорных» данных; модификация данных; удаление данных		Обеспечение целостности и проверка подлинности наборов обучающих и тестовых данных
				Отслеживание и документирование изменений в наборах обучающих и тестовых данных
3	Угроза раскрытия информации о модели ИИ	Угроза заключается в возможности раскрыть информацию о модели ИИ в результате утечки, в том числе через выходные данные, всей или отдельной информации о ней, включая сведения об архитектуре и параметрах обучения модели ИИ	Атака с использованием фаззинга. Извлечение данных. Вредоносная «инъекция»	Шифрование данных, передаваемых за пределы контролируемой зоны
				Контроль и очистка аномалий во входных и выходных данных
				Регистрация и контроль событий, связанных с входными и выходными данными
				Использование «цифровых меток»
				Определение и мониторинг показателей штатного функционирования модели ИИ
				Тестирование на проникновение
				Ограничение потоков данных
4	Угроза хищения обучающих данных	Угроза заключается в возможности получения несанкционированного доступа к обучающим и (или) тестовым данным, в том числе через выходные данные	Атака с использованием фаззинга. Извлечение данных.	Использование методов федеративного обучения (в применимых случаях)
				Контроль и очистка аномалий во входных и выходных данных
				Регистрация и контроль событий, связанных с входными и выходными данными

№ п/п	Возможная угроза безопасности технологий ИИ	Описание возможной угрозы безопасности технологий ИИ	Возможные способы реализации угроз	Меры защиты
			Вредоносная «инъекция»	Шифрование данных, передаваемых за пределы контролируемой зоны
				Использование протоколов конфиденциального вычисления
				Использование методов обезличивания персональных данных и маскирования иной информации ограниченного доступа
				Ограничение потоков данных
				Определение и мониторинг показателей штатного функционирования модели ИИ
				Гарантированное стирание данных
				Использование «цифровых меток»
5	Угроза модификации модели ИИ, в том числе изменения архитектуры, последовательности взаимодействий	Угроза заключается во внесении несанкционированных изменений в модель ИИ путем использования скрытых возможностей в компонентах, которые применяются для разработки, обучения и эксплуатации модели ИИ	Бэкдор (закладки) Модификация алгоритма обучения	Анализ уязвимостей используемых компонентов
				Контроль целостности и проверка подлинности модели ИИ и (или) программного кода
				Отслеживание и документирование изменений в программные коды и документацию о модели ИИ
6	Угроза приведения модели ИИ в состояние «отказ в обслуживании»	Угроза заключается в возможности приведения модели ИИ в состояние «отказ в обслуживании» или снижении ее производительности путем	Атака типа «губка» Атака с использованием фаззинга	Контроль и очистка аномалий во входных и выходных данных
				Определение и мониторинг показателей штатного функционирования модели ИИ
				Регистрация и контроль событий, связанных с входными и выходными данными

№ п/п	Возможная угроза безопасности технологий ИИ	Описание возможной угрозы безопасности технологий ИИ	Возможные способы реализации угроз	Меры защиты
		манипуляции над входными данными		Ограничение параметров запроса
7	Угроза манипуляции поведением модели ИИ	Угроза заключается в возможности использовать вредоносные запросы таким образом, чтобы поведение модели ИИ соответствовало ожиданиям нарушителя, в том числе для выполнения нелегитимного кода моделью ИИ	Вредоносная «инъекция»	Контроль и очистка аномалий во входных и выходных данных
				Регистрация и контроль событий, связанных с входными и выходными данными
				Определение и мониторинг показателей штатного функционирования модели ИИ
				Тестирование на проникновение
8	Угроза подмены модели ИИ	Угроза заключается в возможности нарушителя подменить модель ИИ таким образом, чтобы контролировать ее поведение или получать нежелательные результаты	Угроза возможна при наличии у нарушителя непосредственного доступа к модели ИИ	Регистрация и контроль пар входных и выходных данных
				Определение и мониторинг показателей штатного функционирования модели ИИ
				Использование «цифровых меток»

Приложение 4
к Методическим рекомендациям Банка России
по обеспечению информационной безопасности
при разработке и применении искусственного
интеллекта на финансовом рынке

Меры защиты,
направленные на нейтрализацию актуальных угроз безопасности технологий ИИ

№ п/п	Наименование меры защиты	Описание меры защиты	Объект применения
Подпроцесс «Обеспечение безопасности при подготовке данных»			
1	Контроль и очистка наборов обучающих и тестовых данных от аномалий в данных	Параметры контроля определяются исходя из особенностей функционирования модели ИИ. Примерный перечень мероприятий по контролю аномалий в данных: отслеживание происхождения данных; проверка на наличие неверных данных; проверка на наличие несогласованных защищаемых данных; проверка на предмет корректности меток данных; структурный контроль данных; форматно-логический контроль данных. Примерный перечень мероприятий по очистке аномалий в данных: удаление неверных и несогласованных защищаемых данных; коррекция меток данных;	Наборы обучающих данных, наборы тестовых данных

№ п/п	Наименование меры защиты	Описание меры защиты	Объект применения
		очистка аномалий в данных по результатам структурного и форматно-логического контроля данных. Контроль и очистка аномалий в данных в наборах обучающих данных производится непосредственно перед обучением модели ИИ. Контроль и очистка аномалий в данных в наборах тестовых данных производится непосредственно перед тестированием модели ИИ. Рекомендуется проводить повторный контроль после очистки данных в целях выявления возможных аномалий в данных, возникших в процессе очистки	
2	Шифрование данных, передаваемых за пределы контролируемой зоны	В целях обеспечения конфиденциальности и целостности данных рекомендуется производить шифрование защищаемой информации, передаваемой за пределы контролируемой зоны. Алгоритмы шифрования определяются организацией самостоятельно	Данные в наборах обучающих и тестовых данных
3	Контроль целостности и проверка подлинности наборов обучающих и тестовых данных	Рекомендуется проводить контроль целостности и проверку подлинности наборов обучающих и тестовых данных	Наборы обучающих и тестовых данных
4	Отслеживание и документирование изменений в наборы обучающих и тестовых данных	В целях управления изменениями в наборах обучающих и тестовых данных рекомендуется реализовать процессы отслеживания и документирования изменений с обеспечением воспроизводимости и идентификации таких изменений	Наборы обучающих и тестовых данных
5	Использование методов обезличивания персональных данных и маскирования иной информации ограниченного доступа	Рекомендуется реализовывать методы обезличивания персональных данных, соответствующие требованиям регулирующих и надзорных органов, а также маскирования иной информации ограниченного доступа	Наборы обучающих и тестовых данных
6	Гарантированное стирание данных	Гарантированное стирание данных производится в целях исключения возможности несанкционированного восстановления и использования применяемых для обучения модели ИИ данных	Наборы обучающих и тестовых данных

№ п/п	Наименование меры защиты	Описание меры защиты	Объект применения
Подпроцесс «Обеспечение безопасности при разработке модели ИИ»			
7	Анализ уязвимостей модели ИИ, программного кода и используемых компонентов	Анализ уязвимостей проводится для всех компонентов, применяемых на всех этапах жизненного цикла модели ИИ. Под анализом уязвимостей также понимается анализ кода (статический/динамический/композиционный), использование сканеров уязвимостей и поиск уязвимостей в различных источниках и справочниках (Банке данных угроз безопасности информации Федеральной службы по техническому и экспортному контролю (БДУ ФСТЭК России), CVE (Common Vulnerabilities and Exposures), информационных ресурсах официальных разработчиков моделей ИИ и (или) их компонентов и других)	Компоненты разработки
8	Обеспечение целостности криптографическими методами и проверка подлинности модели ИИ и (или) программного кода	Криптографические методы обеспечения целостности модели ИИ и (или) программного кода определяются организацией самостоятельно	Модель ИИ, программный код
9	Отслеживание и документирование изменений в программные коды и документацию о модели ИИ	В целях управления изменениями, в том числе несанкционированными изменениями, в программном коде и документации модели ИИ рекомендуется реализовать процессы отслеживания и документирования изменений с обеспечением воспроизводимости и идентификации таких изменений	Программный код, документация о модели ИИ
Подпроцесс «Обеспечение безопасности при обучении и тестировании модели ИИ»			
10	Использование методов повышения устойчивости модели ИИ к состязательным атакам	При обучении модели ИИ рекомендуется применять один или несколько методов повышения устойчивости модели ИИ к состязательным атакам. К данным методам относятся: состязательное обучение: в рамках состязательного обучения необходимо использовать примеры состязательных атак при обучении модели ИИ, которые позволят ей обучиться классифицировать состязательные атаки;	Модель ИИ

№ п/п	Наименование меры защиты	Описание меры защиты	Объект применения
		защитная дистилляция: метод заключается в обучении основной модели ИИ, при котором основная модель ИИ учитывает выводы и решения, полученные от модели-учителя. Модель-учитель необходимо обучить на наборе обучающих данных, включающем в себя примеры состязательных атак. Количество примеров состязательных атак в наборе обучающих данных должно определяться организацией самостоятельно; ансамблевые методы: метод заключается в увеличении обобщающей способности модели ИИ путем комбинирования нескольких моделей ИИ. Устойчивость модели ИИ к состязательным атакам повышается путем усреднения результатов нескольких моделей ИИ. При этом необходимо предусмотреть использование различных типов моделей ИИ или различных параметров обучения для каждой модели ИИ в ансамбле, чтобы гарантировать их разнообразие. Также ансамблевые методы позволяют повысить устойчивость модели ИИ, в случае если обучающие данные были «отравлены»	
11	Использование протоколов конфиденциального вычисления	Протоколы конфиденциального вычисления применяются в целях сохранения конфиденциальности данных при обучении модели ИИ на базе вычислительных ресурсов сторонней организации	Модель ИИ
12	Использование методов федеративного обучения	Децентрализованное обучение, применяемое в федеративном обучении моделей ИИ, позволяет сохранить конфиденциальность данных, исключая их передачу третьему лицу	Модель ИИ
13	Использование «цифровых меток»	Встраивание различных «цифровых меток» путем маркировки данных (присваивание определенных маркеров данным). «Цифровые метки» рекомендуется внедрять для различных типов данных (изображения, аудио и видео, текст). Способы и область использования «цифровых меток» определяются организацией самостоятельно	Выходные данные, наборы обучающих и тестовых данных
14	Тестирование на предмет «отравления»	Рекомендуется проводить тестирование модели ИИ в целях выявления случаев «отравления» набора обучающих данных непосредственно после ее обучения. Для тестирования модели ИИ рекомендуется применять набор тестовых данных, который должен содержать данные, отличные	Набор обучающих и тестовых данных

№ п/п	Наименование меры защиты	Описание меры защиты	Объект применения
		от содержащихся в наборе обучающих данных. При этом рекомендуется обеспечить раздельное друг от друга хранение набора обучающих данных и набора тестовых данных с применением мер разграничения доступа. При выявлении факта «отравления» набора обучающих данных рекомендуется выполнить мероприятия по очистке аномалий в данных в указанном наборе, а в случае невозможности проведения очистки произвести смешивание «отравленного» набора обучающих данных с проверенными и достоверными данными путем их добавления в набор. После выполнения указанных мероприятий необходимо провести повторное тестирование модели ИИ и повторять процесс до тех пор, пока факт «отравления» набора обучающих данных не перестанет выявляться в результате тестирования модели ИИ	
15	Тестирование на проникновение	Тестирование на проникновение проводится в отношении модели ИИ. Рекомендуется применять способы реализации угроз безопасности технологий ИИ, описанных в главе 3 настоящих Методических рекомендаций	Модель ИИ
16	Периодическое дообучение модели	Рекомендуется проводить периодическое дообучение модели ИИ, направленное на адаптацию модели ИИ к новым угрозам безопасности технологий ИИ. Дообучение происходит с использованием обновленных и проверенных наборов обучающих данных. Рекомендуется при каждом дообучении фиксировать версию модели ИИ, а также для каждой версии использовать механизмы контроля целостности и проверки подлинности	Набор обучающих данных, модель ИИ
Подпроцесс «Обеспечение безопасности при функционировании модели ИИ»			
17	Шифрование входных и выходных данных	В целях обеспечения конфиденциальности и целостности входных и выходных данных рекомендуется производить их шифрование	Входные и выходные данные
18	Контроль и очистка аномалий во входных и выходных данных	Параметры контроля определяются исходя из особенностей функционирования модели ИИ. Примерный перечень мероприятий по контролю аномалий в данных:	Входные и выходные данные

№ п/п	Наименование меры защиты	Описание меры защиты	Объект применения
		проверка на наличие нежелательной информации в данных, в том числе слова, выражения, специальные символы, вредоносные «инъекции»; проверка на несогласованное наличие защищаемой информации; проверка формата файлов; контроль структуры данных; логический контроль данных. Примерный перечень мероприятий по очистке аномалий в данных: удаление лишних символов; заполнение пропущенных значений; очистка от нежелательной и несогласованной защищаемой информации. Рекомендуется проводить повторный контроль после очистки данных в целях выявления возможных аномалий, возникших в процессе очистки	
19	Регистрация и контроль событий, связанных с входными и выходными данными	Параметры регистрации и контроля определяются организацией. Примерный перечень параметров, подлежащих регистрации и контролю: идентификаторы запросов и выводов (входных и выходных данных соответственно); время поступления запросов и выводов; тип данных (текст, изображения и другое); запросы (входные данные) и выводы (выходные данные); идентификатор пользователя; время обработки запроса; статус обработки (успешная обработка, ошибка)	Входные и выходные данные
20	Ограничение параметров потоков входных данных	Ограничение общего количества, частоты и размера запросов к модели ИИ реализуется в целях предотвращения избыточной вычислительной нагрузки на модель ИИ	Модель ИИ

№ п/п	Наименование меры защиты	Описание меры защиты	Объект применения
21	Определение и мониторинг показателей штатного функционирования модели ИИ	Рекомендуется определить параметры для мониторинга показателей в целях обеспечения штатного функционирования модели ИИ (производительность, смещение распределения входных и выходных данных, качество данных, поведение модели). Представленные показатели рекомендуется дополнять самостоятельно	Модель ИИ

Приложение 5
к Методическим рекомендациям
Банка России по обеспечению
информационной безопасности
при разработке и применении
искусственного интеллекта на
финансовом рынке

Основные положения политики обеспечения информационной безопасности при разработке и применении технологий ИИ

1. Целевые свойства информационной безопасности разрабатываемых и (или) применяемых систем ИИ рекомендуется определить в Политике с учетом их назначения и ценности (значимости) обрабатываемой в них информации:

1.1. Целостность. Рекомендуется закрепить необходимость применения мер, направленных на сохранение целостности программного обеспечения и наборов данных, связанных с разработкой и применением ИИ, а также входных и выходных данных системы ИИ, в том числе при передаче выходных данных системы ИИ лицу, принимающему решение в рамках отдельно взятого бизнес-процесса, технологического участка организации.

1.2. Конфиденциальность. Рекомендуется определить принципы обработки и уничтожения данных в целях обеспечения конфиденциальности информации в соответствии с требованиями законодательства Российской Федерации, нормативных правовых актов федеральных органов исполнительной власти, а также нормативных актов Банка России.

1.3. Доступность. Рекомендуется закрепить необходимость применения мер, которые направлены на сохранение доступности технологий ИИ и предотвращение потери доступности системы ИИ или компонента системы, использующей ИИ.

1.4. Киберустойчивость. Рекомендуется закрепить необходимость использования при разработке и применении ИИ комплекса технологических

и технических решений защиты информации, обеспечивающих устойчивость и работоспособность системы ИИ в случае наступления инцидентов защиты информации, а также необходимость проведения тестирования red team[8].

1.5. Достаточная объяснимость и (или) предсказуемость. Рекомендуется закрепить необходимость применения механизмов интерпретации поведения модели ИИ в целях обеспечения достаточной объяснимости и (или) предсказуемости действий модели ИИ (в релевантных случаях).

2. Минимальные права доступа. Рекомендуется закрепить необходимость реализации принципа, направленного на предоставление пользователям и (или) эксплуатационному персоналу минимально необходимых прав при разработке и (или) эксплуатации систем ИИ.

3. Минимальные персональные данные. Рекомендуется закрепить принцип использования минимальных персональных данных, на обработку которых в соответствии с требованиями законодательства Российской Федерации получено согласие субъекта персональных данных, в целях разработки и (или) применения систем ИИ.

4. Безопасная разработка. Рекомендуется закрепить необходимость адаптации процесса безопасной разработки программного обеспечения в организации под процесс создания системы ИИ с учетом специфики ее разработки, подготовки набора данных, обучения модели ИИ, тестирования и эксплуатации системы ИИ, необходимости постоянного мониторинга результатов ИИ на этапе эксплуатации, а также с учетом положений национального стандарта Российской Федерации ГОСТ Р 71539-2024 (ИСО/МЭК 5338:2023) «Искусственный интеллект. Процессы жизненного цикла системы искусственного интеллекта» и национального стандарта Российской Федерации ГОСТ Р 70889-2023 (ИСО/МЭК 8183:2023)

---

[8] Под тестированием red team понимается симуляция действий нарушителя безопасности в контролируемых условиях, в том числе попыток реализации компьютерных атак в отношении объектов информатизации, входящих в критичную архитектуру, в соответствии с заранее определенными сценариями.

«Информационные технологии. Искусственный интеллект. Структура жизненного цикла данных».

5. Повышенная осведомленность. Рекомендуется закрепить необходимость повышения уровня знаний работников организации об особенностях безопасной разработки и применения систем ИИ, актуальных угрозах безопасности технологий ИИ и мерах защиты.

6. Полномочия работников. Рекомендуется закрепить необходимость определения полномочий, ролей, зон ответственности работников организации при разработке и (или) применении систем ИИ в части обеспечения информационной безопасности.

7. Комплаенс. Рекомендуется закрепить необходимость обеспечения соответствия систем ИИ национальным и (или) международным стандартам, методическим документам, отражающим лучшие практики обеспечения информационной безопасности систем ИИ.

8. Реализация Политики. Рекомендуется закрепить необходимость документирования механизмов реализации Политики в отношении применения технологий ИИ в организации в части мониторинга выполнения требований, аудита и информирования о ее соблюдении.

9. Ответственность. Рекомендуется закрепить необходимость определения ответственности за невыполнение требований Политики.

10. Общедоступные сервисы ИИ. Рекомендуется закрепить необходимость определения требований по использованию в организации общедоступных сервисов ИИ в целях обеспечения минимизации рисков нарушения информационной безопасности.

11. Использование моделей и компонентов технологий ИИ с открытым исходным кодом. Рекомендуется закрепить необходимость разработки требований информационной безопасности к использованию моделей ИИ и компонентов технологий ИИ с открытым исходным кодом в целях обеспечения надежности, качества и доверия к ИИ в части обеспечения информационной безопасности.

12. Оценка и учет моделей и компонентов технологий ИИ с открытым исходным кодом. Рекомендуется закрепить необходимость проведения оценки рисков использования моделей ИИ и компонентов технологий ИИ с открытым исходным кодом в процессах организации, а также необходимость ведения учета таких моделей и компонентов с указанием используемых агентов, плагинов, интерфейсов взаимодействия и иных необходимых сведений.

13. Маркировка. Рекомендуется закрепить необходимость использования технологий маркировки выходных данных ИИ в целях обеспечения информационной безопасности.

14. Сокращение сведений. Рекомендуется закрепить принцип минимизации сведений в общедоступных источниках информации, в том числе в репозиториях, материалах конференций и статьях, об особенностях применяемых в организации моделей ИИ в целях предотвращения использования таких сведений при подготовке атак.

15. Инциденты защиты информации. Рекомендуется закрепить необходимость регистрации и реагирования на инциденты защиты информации и инциденты операционной надежности в системах ИИ.

16. Нештатные ситуации. Рекомендуется закрепить необходимость разработки плана восстановления операционной надежности и порядка действий работников организации в случае возникновения нештатных ситуаций при эксплуатации систем ИИ, в том числе действий по аварийной остановке системы ИИ, в целях оперативного реагирования на нештатные ситуации, связанные с технологиями ИИ.

17. Аутсорсинг. Рекомендуется закрепить необходимость проведения оценки доверия к сторонним данным и (или) моделям ИИ в части информационной безопасности и оценки рисков нарушения информационной безопасности в случае применения организацией сервиса ИИ поставщика услуг с учетом ценности (значимости) обрабатываемой информации и

технологического процесса, в котором будут использованы результаты применения ИИ.

18. Пересмотр Политики. Рекомендуется закрепить необходимость пересмотра Политики, мер защиты, направленных на нейтрализацию актуальных угроз безопасности технологий ИИ, рисков ИИ, связанных с нарушением информационной безопасности и операционной надежностью, и возможных угроз информационной безопасности, специфичных для технологий ИИ. Такой пересмотр рекомендуется осуществлять на периодической основе с учетом внешних и внутренних факторов влияния, в том числе в связи с развитием технологий ИИ, тактик и техник нарушителей информационной безопасности.