Безопасность ИИ

Встреча рабочей группы

Управление информационной безопасности,
Ассоциация ФинТех

12 февраля 2025

01

Итоги работы за 2024 год

Результаты 2024

1. Сформирован перечень угроз информационной безопасности ИИ при разработке и применении ИИ в финтех-организациях.
2. Предложен классификатор угроз по этапам жизненного цикла решения на основе ИИ.
3. Разработан фреймворк ИБ ИИ для связки угроз и мер защиты (в формате технических мер, компетенций и процессов).
4. Сформирован набор компетенций в области безопасности ИИ для подразделений безопасности для минимизации угроз ИБ ИИ.

Индустриальный фреймворк по ИИ

Индустриальный фреймворк безопасного применения ИИ:

- Набор принципов, практик, подходов и методов для минимизации рисков и обеспечения безопасного применения решений на основе ИИ в финтехе.
- Включает в себя **технологические инструменты (Технологии)**, **практики (Процессы)** и **компетенции (Люди).**

Для чего нужен фреймворк:

- консолидация лучшего опыта и практик по минимизации угроз ИБ ИИ.
- обеспечение диалога финтеха с технологическими организациями и разработчиками инструментов.
- систематизация подходов к анализу безопасности ИИ-решений.

Индустриальный фреймворк по ИИ. Как выстроен?

Угрозы

- Набор угроз ИБ ИИ

Домен

- Планирование
- Обеспечение данными
- Разработка модели
- Развертывание
- Эксплуатация

Объект воздействия

- Данные
- Модель
- Инфраструктура

Меры минимизации

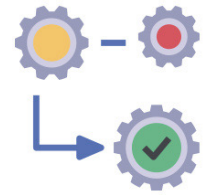
- Процессы
- Технологии
- Люди

Индустриальный фреймворк по ИИ. Типы мер.

Угрозы

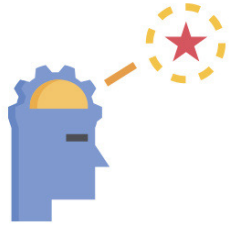
↑

Меры минимизации



ПРОЦЕССЫ

- Внедрение дополнительных процессов проверки качества моделей
- Внедрение процессов контроля за тюнингом модели
- и т.д.



ТЕХНОЛОГИИ

- Внедрение инструментов отслеживания вывода моделей
- Внедрение инструментов выявления галлюцинаций моделей
- и т.д.

ПЕРСОНАЛ

- Повышение экспертизы сотрудника в определенной области
- Внедрение отдельных ролей в ИБ (например, промт-инженер)
- и т.д.

Индустриальный фреймворк по ИИ. Пример

Домен	Угроза	Объект воздействия (данные, модель, инфраструктура)	Описание мер минимизации	Меры минимизации		
				Люди	Технологии	Процессы
Обеспечение данными	Отсутствие классификации данных	Данные	1. Внедрение многоуровневой системы классификации данных с четкими категориями, основанными на важности и конфиденциальности данных. 2. Обучение сотрудников на регулярной основе для повышения квалификации в правильной классификации данных. 3. Создание регламентов и автоматических инструментов для облегчения классификации новых данных в процессе их поступления. 4. Регулярный аудит классификации данных для выявления несоответствий и корректировки процесса. 5. Интеграция классификации данных в системы управления данными для автоматической сортировки по категориям.	1. Специалисты по управлению данными и ML-инженеры, владеющие методологиями классификации и принципами защиты конфиденциальных данных. 2. Аналитики данных, способные правильно применять правила классификации к вновь поступающим наборам данных. 3. Офицеры по комплаенс, осуществляющие регулярный аудит и корректировку схем классификации.	1. Системы DLP для контроля данных 2. Средства каталогизации данных для управления данными и тегами. 3. Инструменты для автоматизации классификации	1. Пересмотр политик классификации данных в соответствии с требованиями бизнеса и регуляторов. 2. Обучение персонала принципам и практикам классификации данных. 3. Автоматизированные процедуры маркировки вновь поступающих данных и периодический аудит корректности меток.
	Использование данных низкого качества	Данные	1. Внедрение автоматизированных систем очистки данных для выявления и исправления ошибок, пропусков или аномалий. 2. Проведение регулярных проверок и тестов качества данных перед их использованием в моделях ИИ. 3. Разработка стандартов качества данных и внедрение мониторинга для выявления несоответствий. 4. Обеспечение прозрачности источников данных с проверкой их достоверности перед их включением в обучение. 5. Создание резервных систем для хранения качественных данных, на которые можно ссылаться при возникновении проблем с основным источником. 6. Применение механизмов обратной связи для пользователей и партнеров, чтобы оперативно выявлять и устранять проблемы с качеством данных.	1. Специалисты по качеству данных, владеющие методами очистки, валидации и нормализации данных. 2. ML-инженеры, способные интегрировать фильтры и проверки качества на этапе подготовки данных. 3. Аналитики и кураторы данных, ответственные за мониторинг и документирование источников данных	1. Инструменты оценки качества данных 2. Средства очистки и нормализации данных 3. Инструменты мониторинга качества данных	1. Внедрение стандартов качества данных и регулярный аудит источников. 2. Непрерывный мониторинг качества данных с уведомлениями при нарушениях критериев качества. 3. Процессы обратной связи от пользователей и партнёров для оперативного устранения проблем.
	Отсутствие версионирования данных	Данные	1. Внедрение системы контроля версий данных с возможностью восстановления предыдущих версий и отслеживания всех изменений. 2. Автоматизация резервного копирования данных с раздельным хранением разных версий для предотвращения потерь. 3. Введение строгих протоколов на внесение изменений в данные, с автоматическим отслеживанием и документированием всех изменений. 4. Интеграция системы управления версиями данных с основными бизнес-процессами для своевременной реакции на возможные ошибки. 5. Регулярное тестирование систем версионирования на устойчивость и правильность работы.	1. Специалисты по данным, владеющие системами контроля версий данных 2. MLOps-специалисты, умеющие настраивать процессы CI/CD с версионированием. 3. Специалисты ИБ, определяющие политики версионирования и хранения исторических наборов данных.	1. Системы версионирования данных 2. Хранилища с поддержкой версионирования 3. Средства шифрования и цифровой подписи для отслеживания изменений данных	1. Бэкап и проверка целостности версионных репозиториев. 2. Периодическое тестирование и валидация систем версионирования.

02

Предложения на 2025 год

Предложения на 2025 год

1. Сформировать карту рисков ИБ ИИ для финтех-организаций.
2. Доработать фреймворк (расширить перечень мер защиты техническими мерами - инструментами, компетенциями сотрудников, процессами).
3. Сформировать базу MLSecOps инструментов и кейсов применения мер защиты ИИ на основе экспертизы участников АФТ (для развития центра компетенции АФТ по ИИ).

03

Формирование карты рисков

Формирование карты рисков. Для чего?

1. Выявление и классификация рисков

Определение ключевых рисков, связанных с использованием ИИ в финансовой сфере, в соответствии с требованиями нормативных актов.

2. Оценка уровня риска

Разработка количественных и качественных показателей для оценки вероятности возникновения и уровня воздействия выявленных рисков.

3. Приоритизация рисков

Визуализация рисков с учетом их значимости, позволяющая определить приоритеты для минимизации угроз.

4. Разработка мер управления рисками

Определение стратегий снижения рисков, включая меры контроля, мониторинга и реагирования.

5. Интеграция в процесс управления ИБ

Встраивание карты и матрицы рисков в общую систему управления информационной безопасностью организации.

Сформированная карта рисков в соотнесении с перечнем угроз ИБ ИИ даст возможность участникам АФТ повысить качество и скорость проведения оценки рисков за счет консолидированной экспертизы представителей отрасли.

Управление рисками ИБ

Зарубежные

- Стандарт ISO/IEC 27005 «Менеджмент риска информационной безопасности»
- NIST SP 800-37 Фреймворк управления рисками для ИС и организаций: жизненный цикл систем для обеспечения безопасности и конфиденциальности
- NIST SP 800-30 Руководство по проведению оценки риска
- NIST SP 800-53 Контроль безопасности и конфиденциальности в ИС и организациях

Отечественные

ГОСТ Р ИСО/МЭК 27005-2021:

Русскоязычная адаптация международного стандарта ISO/IEC 27005. Содержит описание процессов управления рисками информационной безопасности.

Специфические требования для финсектора:

- Положение Банка России № 716-П
- Рекомендации Банка России по применению подходов к оценке рисков

Виды операционного риска (по Положению Банка России № 716-П)

Риск информационной безопасности: угрозы, обусловленные недостатками в процессах обеспечения информационной безопасности, включая технологические мероприятия и программное обеспечение.

Риск информационных систем: отказы или нарушения функционирования информационных систем, а также несоответствие их функциональных возможностей потребностям кредитной организации.

Правовой риск: возникновение убытков вследствие несоблюдения кредитной организацией требований правовых актов, договорных обязательств, а также вследствие нарушения правовых норм, регулирующих деятельность кредитной организации.

Риск ошибок в управлении проектами: недостатки и нарушения в процессах управления проектной деятельностью, направленной на изменение систем функционирования и поддержания работоспособности кредитной организации.

Риск ошибок в управленческих процессах: недостатки и нарушения внутренних процессов, а также принятия решений по банковским операциям и внутрихозяйственной деятельности.

Риск ошибок в процессах внутреннего контроля: недостатки и нарушения системы внутреннего контроля, включая несоблюдение правил противодействия легализации доходов, полученных преступным путем, и финансированию терроризма.

Модельный риск: риск ошибок в процессах разработки, проверки, адаптации, приёмки и применения методик количественных и качественных моделей оценки активов, рисков и иных показателей, используемых при принятии управленческих решений.

Риск потерь средств клиентов и третьих лиц: вследствие нарушения кредитной организацией кодексов профессиональной этики, рыночных практик и правил поведения при продаже финансовых инструментов и услуг.

Операционный риск платежной системы: риск возникновения убытков или иных негативных последствий для участников платежной системы вследствие недостатков или ошибок в процессах, системах, человеческих действиях или внешних событиях, связанных с функционированием платежной системы.

Обработка рисков

Принятие.
Руководство берет на себя ответственность и принимает риски, инфраструктура организации продолжает функционировать.
Избегание.
Риск устраняется путем устранения причин его возникновения или существования.
Минимизация/Принятие мер.
В организации применяются компенсирующие организационные и технические меры для снижения влияния риска до допустимого уровня.
Разделение.
Стратегия подразумевает распределение угроз и ответственности между разными уровнями инфраструктуры — или привлечение третьей заинтересованной стороны для уменьшения влияния риска.
Передача.
Ответственность за реализацию и последствия риска полностью передается на сторону, способную обработать риск эффективнее. В качестве примера можно выделить страхование рисков или привлечение сторонних сервисов, например SOC.

Пример соотнесения угроз ИИ с рисками (в соотв. с Положением 716-П)

Злоупотребление доступом к финансовым данным

- **Риск информационной безопасности**: Несанкционированный доступ к данным из-за недостатков в защите информации.
- **Риск ошибок в процессах внутреннего контроля**: Недостаточная система внутреннего контроля может допустить злоупотребления.
- **Правовой риск**: Нарушение законодательства о защите данных (например, персональных данных) может повлечь юридические последствия.

Галлюцинации (ошибки генеративных моделей)

- **Риск ошибок в управленческих процессах**: Использование ошибочных данных для принятия решений.
- **Риск информационных систем**: Некорректная работа алгоритмов может быть связана с недостатками в их реализации или интеграции.

Использование устаревших данных

- **Риск ошибок в управленческих процессах**: Ошибки при принятии решений из-за использования устаревшей информации.
- **Риск информационных систем**: Несвоевременное обновление данных может быть следствием недостатков в информационных системах.

Манипуляция рыночной информацией с помощью LLM

- **Риск информационной безопасности**: Использование генеративных моделей для манипуляций может угрожать целостности рыночной информации.
- **Риск потерь средств клиентов и третьих лиц**: Такие манипуляции могут повлиять на доверие клиентов и партнеров.
- **Правовой риск**: Нарушение законодательства о манипуляциях на рынке.

Отсутствие разграничения прав доступа

- **Риск информационной безопасности**: Недостатки в разграничении доступа создают угрозы утечки данных.
- **Риск ошибок в процессах внутреннего контроля**: Проблемы с разграничением доступа указывают на слабости системы внутреннего контроля.
- **Риск ошибок в управленческих процессах**: Могут возникнуть ошибки в процессах, связанных с обеспечением безопасности данных.

Угроза безопасности цепочки поставок

- **Риск информационной безопасности**: Уязвимости в цепочке поставок могут быть использованы для компрометации информационных активов.
- **Риск нарушений операционной устойчивости**: Проблемы с поставщиками могут нарушить стабильность операций.
- **Правовой риск**: Несоответствие поставщиков регуляторным требованиям может привести к санкциям.

04

Формирование перечня мер защиты

Меры защиты в фреймворке ИБ ИИ

Для чего добавлять и детализировать меры защиты?

1. Идентификация новых угроз и средств защиты от них

Анализ актуальных атак на ИИ-системы, включая подмену данных, эксплуатацию уязвимостей моделей и атак на конфиденциальность.

2. Оценка применимости

Определение технологической решений, их совместимости с регуляторными требованиями и возможностью интеграции в финансовые ИИ-системы.

3. Создание методологии выбора решений

Разработка критериев и рекомендаций по применению технологических решений в зависимости от специфики ИИ-моделей, архитектур и уровней критичности данных.

Технические меры (пример)

Отсутствие классификации данных	
Инструменты и меры:	1. Frameworks: Внедрение корпоративных политик классификации данных (например, выделение уровней: общедоступные, для внутреннего использования, конфиденциальные, строго конфиденциальные). 2. DLP-системы (Data Leak Prevention) — Обнаруживает и классифицирует конфиденциальные данные для предотвращения их несанкционированного распространения (утечку) 3. Интеграция с MDM (Master Data Management): Средства управления данными для централизованного контроля классификации.
Использование данных низкого качества	
Инструменты и меры:	1. Data Quality Assessment Tools: Great Expectations, Soda, Tecton или другие решения для профилирования и валидации данных. 2. Автоматизированный мониторинг качества: Метрики качества (missing values, outliers) с триггерами оповещений при отклонениях.
Отсутствие версионирования данных	
Инструменты и меры:	1. DVC (Data Version Control) или Git LFS для версионирования больших наборов данных. 2. MLflow Data Registry, LakeFS, Delta Lake для отслеживания версий в хранилищах данных. 3. Контрольные суммы (hash-суммы) для валидации неизменности версий.

Компетенции (пример)

Роль	Компетенции:	Задачи:
Инженер по информационной безопасности	1. Знание принципов сетевой и прикладной безопасности, протоколов шифрования. 2. Опыт работы с SIEM/SOAR, WAF, IAM-системами. 3. Навыки анализа уязвимостей, проведения пентестов, использования DLP и систем обнаружения аномалий.	1.Настройка систем защиты API, управление доступами к данным и моделям. 2. Обнаружение и предотвращение кибератак (DDoS, отравление данных, проникновения). 3.Регулярный аудит конфигураций, обновление политики безопасности согласно текущим угрозам.
Специалист по управлению данными и качеством	1. Понимание принципов классификации данных, требований к их качеству и правовых норм обработки 2.Опыт работы с инструментами профилирования данных, оценкой качества, DVC для версионирования. 3.Знание методологий очистки, нормализации, а также мониторинга целостности данных	1. Разработка и внедрение стандартов качества данных, регулярная проверка на наличие пропусков, ошибок и устаревших данных. 2. Обеспечение соответствия данных регуляторным требованиям. 3.Введение механизмов классификации, валидации
MLSecOps-инженер	1. Глубокое понимание CI/CD и DevOps-практик применительно к ML-моделям. 2. Знание инструментов контейнеризации и оркестрации. 3. Умение настраивать конвейеры обучения, тестирования и развертывания моделей с фокусом на безопасность.	1. Автоматизация развёртывания моделей с учётом требований безопасности и регуляторных норм. 2. Настройка мониторинга производительности и стабильности моделей. 3. Регулярная проверка устойчивости конвейера к сбоям и угрозам

Пример доработки фреймворка

Угроза	Объект воздействия (данные, модель, инфраструктура)	Описание мер минимизации
Отсутствие классификации данных	Данные	1. Внедрение многоуровневой системы классификации данных с четкими категориями, основанными на важности и конфиденциальности данных. 2. Обучение сотрудников на регулярной основе для повышения квалификации в правильной классификации данных. 3. Создание регламентов и автоматических инструментов для облегчения классификации новых данных в процессе их поступления. 4. Регулярный аудит классификации данных для выявления несоответствий и корректировки процесса. 5. Интеграция классификации данных в системы управления данными для автоматической сортировки по категориям.

Меры минимизации		
Люди	**Технологии**	**Процессы**
1. Специалисты по управлению данными и ML-инженеры, владеющие методологиями классификации и принципами защиты конфиденциальных данных. 2. Аналитики данных, способные правильно применять правила классификации к вновь поступающим наборам данных. 3. Офицеры по комплаенс, осуществляющие регулярный аудит и корректировку схем классификации.	1.Системы DLP для контроля данных 2.Средства каталогизации данных для управления данными и тегами. 3.Инструменты для автоматизации классификации **4. Инструмент 1** **5. Инструмент 2** **6. Инструмент 3**	1. Пересмотр политик классификации данных в соответствии с требованиями бизнеса и регуляторов. 2. Обучение принципам и практикам классификации данных. 3.Автоматизированные процедуры маркировки вновь поступающих данных и периодический аудит корректности меток.

05

Предложения от отрасли

Предложения к подходам к обеспечению ИБ ИИ

1. Направить от экспертов Ассоциации предложения в Банк России по подходам к обеспечению безопасности применения ИИ на финансовом рынке и формированию подходов по регулированию вопросов ИБ искусственного интеллекта в финансовом секторе, чтобы

- подчеркнуть актуальность данной темы для финансового и финтех-секторов,
- экспертам отрасли включиться на ранней стадии в процесс формирования практической составляющей методологии регулятора,
- предложить регулятору взвешенный подход между безопасностью и инновациями, подкрепленный практическими рекомендациями, проработанными на площадке АФТ.

2. Перечень предложений в виде письма направить в конце 1 квартала 2025 года.