AI SECURITY

В ФИНТЕХЕ

АССОЦИАЦИЯ ФИНТЕХ × SWORDFISH SECURITY

БЕЗОПАСНОСТЬ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ФИНТЕХЕ

ОБ ИССЛЕДОВАНИИ

Марианна Данилина

Руководитель Управления стратегии,
исследований и аналитики,
Ассоциация ФинТех

Современная стратегия AI Security строится на постоянном развитии компетенций сотрудников и формировании культуры осведомленности на всех уровнях организации. И хотя ИИ автоматизирует многие процессы, в том числе в кибербезопасности, все-таки роль специалистов ИБ станет в будущем еще значительнее – из содействующей она трансформируется в стратегическую. Взаимное проникновение ИИ в направлении безопасности оставит роль руководителя ИБ «у штурвала»: именно он задаст курс, примет ключевые решения для безопасной и корректной работы решений на базе ИИ.

Александр Товстолип

Руководитель Управления
информационной безопасности,
Ассоциация ФинТех

Искусственный интеллект трансформирует подход к кибербезопасности. Мы знаем как важна для организаций защита и цифровая гигиена, поэтому провели исследование и оценили уровень готовности к внедрению практик безопасности искусственного интеллекта в финтехе.

Модель угроз меняется, поверхность атак становится иной, поэтому уже сегодня нужно обновлять традиционные стратегии и подходы. Уверены, что данное исследование позволит вам взглянуть на роль ИБ-департаментов с точки зрения новых смыслов и перспектив.

Юрий Сергеев

Директор по стратегии,
Генеральный Партнер,
Swordfish Security

Разработка безопасного ПО трансформируется в связи с развитием искусственного интеллекта. AI открывает как новые возможности для ускорения всех процессов, так и форсированно формирует новые угрозы для людей и организаций. Высокая ценность ИИ очевидна, но вместе с тем растет запрос на его защиту и кибербезопасность. AI Security – новый виток развития проектирования ИИ, при котором меры безопасности интегрированы в его архитектуру и программный код.

Это исследование позволит читателям изучить направление AI Security в финтех-отрасли, а потенциальные угрозы поможет преобразовать в новые возможности.

СОДЕРЖАНИЕ

КЛЮЧЕВЫЕ ВЫВОДЫ

81%

организаций определяют машинное обучение (ML) и большие языковые модели (LLM) как ядро своих ИИ-стратегий.

В то время как эти технологии становятся стандартом, рынок демонстрирует в том числе растущий спрос на узкоспециализированные решения в области ИИ. Так, глубокое обучение (Deep Learning) находит применение в 51% компаний, обработка естественного языка (NLP) – в 49%, а компьютерное зрение – в 37%. Компании движутся от пилотных внедрений к построению комплексной ИИ-экосистемы, способной решать специализированные бизнес-задачи.

75%

финтех-организаций выделяют утечку конфиденциальных данных в качестве ключевой угрозы безопасности при использовании ИИ.

Среди других значимых рисков респонденты выделяют некорректную работу моделей, несанкционированный доступ к ИИ-системам и инъекции промптов, когда атакующий встраивает в запрос скрытые команды или искажает контекст, заставляя модель выполнять неавторизованные действия (например, отправлять письма, удалять данные) или игнорировать установленные ограничения.

Для минимизации этих рисков финтех-организации применяют различные меры: используют гибридный подход к оценке рисков, разрабатывают внутренние политики безопасности, интегрируют инструменты защиты, проводят аудиты и тестирования ИИ-систем.

>25%

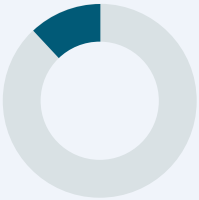
респондентов уже столкнулись несколько раз с инцидентами безопасности, связанными с атаками на ИИ-системы.

Угрозы реальны и регулярны – атаки ИИ не гипотетические, а повторяются с ощутимой частотой. Это не «единичные случаи», а системный риск. В то время как одна часть рынка только начинает знакомиться с угрозами ИИ (порядка 60%), другая – уже накапливает практический опыт противодействия атакам.

**По данным опроса Ассоциации ФинТех, декабрь 2025 года*

62% **организаций применяют гибридный подход к оценке результатов и выводов в работе с ИИ-решениями.**

В целом, доверяя результатам работы и анализа ИИ-систем, они сохраняют ручной контроль над критически важными бизнес-процессами и в обязательном порядке проводят дополнительную проверку выводов моделей. При этом лишь восьмая часть компаний (12%) готова продемонстрировать высокий уровень доверия к ИИ. Такой значительный разрыв подсвечивает ключевую проблему современного этапа развития технологий: острую необходимость в валидации качества работы ИИ-систем, надежности и, что особенно важно, объяснимости алгоритмов.



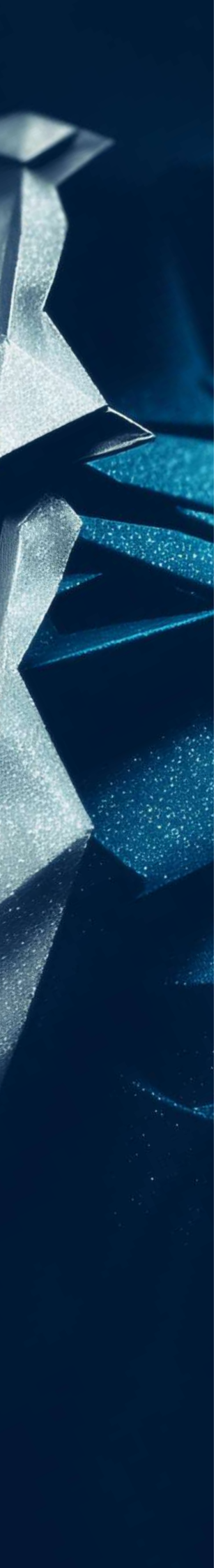
12% **компаний внедряют системную оценку этических рисков ИИ.**

Большинство организаций находятся на стадии обсуждения и подготовки этических принципов в области ИИ, ограничиваясь эпизодическими, а не регулярными практиками. Формируется значительный разрыв между осознанием важности этических аспектов и выстраиванием полноценных процедур.

74% **компаний концентрируют защитные меры на этапе подготовки данных и тестирования моделей.**

Параллельно наблюдается рост внимания к безопасности на всех стадиях жизненного цикла ИИ: организации усиливают контроль как на этапе разработки, так и в процессе эксплуатации, формируя практику непрерывного мониторинга безопасности.

**По данным опроса Ассоциации ФинТех, декабрь 2025 года*



>80%

компаний планируют масштабные обучающие программы, определяя развитие культуры безопасности ИИ как ключевой приоритет на 2026 год.

Второй по значимости задачей становится формирование внутренних политик AI Security, что отражает переход от точечных мер к системному управлению рисками.

Технологический контур безопасности также будет усилен: компании намерены активнее внедрять специализированные инструменты защиты и проводить регулярный аудит систем.

Такой комплексный подход, объединяющий образовательные, организационные и технические меры, позволит организациям поэтапно выстраивать зрелую инфраструктуру безопасности искусственного интеллекта.

60%

компаний уже внедряют комплексные меры по защите своих ИИ-систем.

В число наиболее распространенных практик входят:

- регулярный мониторинг и аудит работы моделей
- активное тестирование на уязвимости
- разработка внутренних стандартов и политик безопасности
- обязательное обучение сотрудников основам AI Security

При выборе инструментов приоритет отдается специализированным платформам безопасности для ИИ. Параллельно с этим набирает силу тренд на развитие собственных разработок – это говорит не только о дефиците готовых решений на рынке, но и о растущей потребности бизнеса в максимально адаптированных системах защиты.

**По данным опроса Ассоциации ФинТех, декабрь 2025 года*

56,8

МЛРД РУБЛЕЙ

инвестиции финансового сектора во внедрение и использование искусственного интеллекта (ИИ) в 2024 году*.

*По данным Интерфакс: *interfax.ru*

ПЕРИМЕТР ИССЛЕДОВАНИЯ

ОПИСАНИЕ ИССЛЕДОВАНИЯ

Цель:
анализ текущего состояния и перспектив развития безопасности искусственного интеллекта в российских финтех-компаниях.

Выборка исследования состоит из крупнейших банков России, страховых и технологических компаний – членов Ассоциации ФинТех.

Участники Ассоциации ФинТех:

ХАРАКТЕРИСТИКА РЕСПОНДЕНТОВ

Роли участников:

14%
CISO/CIO/CTO

25%
Руководитель ИТ/ИБ-проекта

25%
AI-инженер/Data Science инженер

36%
Другие роли в компании в области ИБ

ПРОФИЛЬ РЕСПОНДЕНТОВ

Сегментация направлений в организациях:

50% Банковские услуги и платежные системы

39% Страхование

38% Инвестиции

37% Инвестиционные платформы

25% Кредитование

ВЫЗОВЫ РАЗВИТИЯ ИИ

Развитие искусственного интеллекта стало одним из главных технологических сдвигов десятилетия – и, пожалуй, самым противоречивым. С одной стороны, ИИ – это инструмент колоссального усиления человеческих возможностей, ускоряющий анализ, принятие решений и научные открытия. С другой – его повсеместное внедрение несёт новые типы рисков, не сводимых к традиционным угрозам информационной безопасности.

Современные модели ИИ обладают свойствами, которые делают их принципиально иными объектами защиты. Они учатся на данных, чья природа может быть частично неизвестна, формируют выводы, механизм которых непрозрачен даже для разработчиков и взаимодействуют с пользователями в реальном времени, что создаёт возможность эксплуатационных атак – от промпт-инъекций до утечек конфиденциальной информации. В отличие от классических ИТ-систем, в ИИ невозможно «заморозить» состояние – каждая новая версия модели и каждый дополнительный обучающий набор изменяют её поведение, что делает процесс защиты не точечным, а непрерывным.

К технологическим рискам добавляются этические и социальные вызовы, которые напрямую влияют на доверие к ИИ-системам. Проблема «чёрного ящика», предвзятость моделей, возможность манипулирования результатами, утрата контроля автономных систем – всё это превращает безопасность ИИ в вопрос не только инженерный, но и цивилизационный. Ошибка алгоритма в медицине, судебной системе или финансовом секторе становится не просто техническим инцидентом, а событием с последствиями для человеческих жизней и институтов доверия.

Этические вызовы и задачи в связи с развитием и внедрением ИИ *

Предвзятость и недискриминация	Подотчетность и подконтрольность ИИ человеку	Равенство при распределении благ от ИИ	Вопросы трудоустройства и безработицы	Защита данных и конфиденциальность, подотчетность и подконтрольность ИИ человеку
Прозрачность и объяснимость алгоритмов ИИ	Надежность	Автономия человека и его свободного выбора	Влияние ИИ на поведение человека и межличностное взаимодействие	Общее обеспечение гарантий основных прав человека в контексте внедрения ИИ

*ИИ требует **новой логики безопасности** – такой, которая сочетает в себе принципы киберзащиты, управления рисками и этического регулирования. Развитие технологий невозможно остановить, но можно создать процессы, способные сделать это развитие контролируемым, устойчивым и предсказуемым.*

** По данным Белой книги в сфере искусственного интеллекта: ethics.a-ai.ru*

ЧТО ТАКОЕ AI SECURITY И ПОЧЕМУ ЭТО ВАЖНО

> *Успех в создании искусственного интеллекта был бы величайшим достижением в истории человечества, но, к сожалению, он также может стать и последним*
>
> *Стивен Хокинг*

Безопасность искусственного интеллекта становится ключевым фактором, определяющим уровень доверия к решениям на его основе. По аналогии с обязательной проверкой безопасности автомобилей, самолетов или любых фундаментальных строений, все продукты на базе ИИ должны проходить тщательную проверку и гарантировать высокий уровень защищенности.

Согласно данным ВЦИОМ, в отрасли финансовых услуг уровень внедрения ИИ в 2024 году составляет порядка 63% и растет год к году*.

Что такое AI Security?

В современной практике существует два значения понятия AI Security.

1. AI Security – это обеспечение и защита систем искусственного интеллекта от угроз, атак и злоупотреблений. Она включает в себя методы обнаружения уязвимостей, предотвращения атак, а также обеспечение конфиденциальности и целостности данных, используемых в ИИ. В текущем исследовании под «AI Security» подразумевается данный термин.

2. AI Security – это интеграция искусственного интеллекта в сферу кибербезопасности с целью автоматизации и повышения скорости выполнения процессов защиты.

**По данным ВЦИОМ: ok.wciom.ru*

AI Security как ответ на угрозы

AI Security формируется как новое направление, объединяющее принципы кибербезопасности, инженерии надёжных систем и управления рисками ИИ.

Такой подход учитывает, что уязвимости могут возникнуть на любом уровне – в данных, алгоритмах, цепочке поставок или механизмах интеграции. Безопасность становится не конечной целью, а свойством системы. В этом контексте AI Security – не отдельный проект, а функция зрелости организации. Она требует чёткого распределения ролей, понимания взаимосвязей между технологиями и людьми, а также внедрения механизмов контроля, соответствующих динамике развития ИИ.

Встраивая принципы безопасности на уровне данных, моделей и эксплуатационной среды, организации получают не просто защищённый продукт, а надёжную основу для доверия и инноваций.

Как сбалансировать развитие искусственного интеллекта и его безопасность?

Искусственный интеллект стремительно становится главным катализатором технологического и экономического роста. Однако вместе с масштабом применения растёт и спектр рисков – от уязвимостей в цепочках поставок моделей до непредсказуемого поведения генеративных систем. В условиях усиливающегося регулирования и общественного внимания бизнесу необходимо научиться совмещать две, на первый взгляд, противоположные цели: ускорять инновации и одновременно обеспечивать безопасность.

Речь идёт не о дополнительном контроле, а о проектировании ИИ-систем таким образом, чтобы их можно было безопасно масштабировать, тестировать и разворачивать без потери скорости. Ключевой принцип – автоматизация доверия: безопасность интегрируется в каждую стадию жизненного цикла разработки больших языковых моделей через код, тесты и политики, а не через ручные согласования. Это позволяет развивать десятки ИИ-инициатив параллельно, сохраняя высокий уровень надёжности.

В ближайшие годы именно способность сочетать скорость и защищенность станет ключевым фактором конкурентоспособности в ИИ-экономике. Компании, которые встроят безопасность в инновационный контур, смогут масштабировать свои доверенные решения быстрее, завоёвывая доверие клиентов, регуляторов, партнеров и инвесторов.

Александр Пинаев

Генеральный директор
ГК Swordfish Security

ПРИНЦИП «ЭКСПЕРТИЗА – ТЕХНОЛОГИИ – ПРОЦЕССЫ»

Любая система безопасности держится на трёх взаимосвязанных опорах – экспертизе, технологиях и процессах. Этот принцип, сформировавшийся в практике DevSecOps и управления рисками, полностью сохраняет актуальность и в контексте AI Security. Более того, именно в сфере искусственного интеллекта он приобретает особую остроту: уязвимость может возникнуть не только из-за сбоя алгоритма или ошибки конфигурации, но и вследствие человеческого непонимания, неверной интерпретации данных или отсутствия процедурного контроля.

01

Экспертиза – центральный элемент этой триады. От уровня осведомлённости и компетенций специалистов зависит способность организации распознавать риски и адекватно на них реагировать. Инженеры, работающие с моделями, должны понимать не только принципы обучения, но и угрозы, связанные с утечкой данных, подменой моделей или предвзятостью выборки. Руководители – осознавать, что ошибки в управлении безопасностью ИИ несут не только финансовые, но и репутационные, а порой и правовые последствия. Культура безопасности становится такой же частью корпоративной зрелости, как и культура экспериментов.

02

Технологии обеспечивают защиту на уровне инфраструктуры, данных и моделей. Это инструменты для контроля цепочки поставок данных и моделей, платформы для отслеживания обучения (ML-monitoring), решения для обнаружения аномалий и предотвращения атак на модель, механизмы обеспечения конфиденциальных вычислений (Confidential Computing) и управления доступом. Но сами по себе технологии не создают безопасность – они работают только в связке с процессами и осознанным применением.

03

Процессы связывают экспертизу и технологии в устойчивую систему. Именно процессы задают повторяемость, подотчётность и прозрачность: как принимаются решения о внедрении моделей, кто отвечает за проверку данных, каким образом проводится аудит, в какие циклы встроено тестирование на устойчивость к атакам. Без процессного каркаса даже самые совершенные инструменты превращаются в набор формальных средств без реальной эффективности.

ФРЕЙМВОРКИ КАК ОСНОВА МОДЕЛИ ПРОЦЕССОВ

Когда принципы «экспертиза – технологии – процессы» превращаются в ежедневную практику, возникает следующий вызов: **как сделать эту практику управляемой и измеримой?** Фреймворки становятся тем самым структурным каркасом, который помогает организациям перевести разрозненные меры безопасности в системную модель зрелости. Они задают общий язык для всех участников – от инженеров до руководителей – и позволяют выстроить единое понимание того, где организация находится сейчас и к какому уровню зрелости стремится.

Какую роль играют фреймворки в формировании безопасности ИИ?

Фреймворки безопасности ИИ играют ключевую роль в формировании устойчивых и надежных систем – они задают общий язык, стандарты и принципы, которые позволяют компаниям и исследователям двигаться в одном направлении. По сути, это структурированные карты, помогающие корректно оценивать риски, обеспечивать прозрачность моделей, отслеживать цепочку данных, контролировать поведение систем и минимизировать вероятность некорректных или опасных сценариев. Без них индустрия действовала бы интуитивно и фрагментарно, что неизбежно привело бы к ухудшению качества, снижению доверия и увеличению рисков.

Важно и то, что фреймворки создают основу для диалога между техниками, регуляторами и обществом. Они помогают объяснить, как принимаются решения, какие меры контроля и аудита существуют, и почему система может считаться безопасной. В долгосрочной перспективе именно согласованные и адаптивные фреймворки станут фундаментом этичной, защищённой и масштабируемой экосистемы искусственного интеллекта – экосистемы, которая развивается быстро, но при этом не теряет из виду безопасность и общественные ценности.

Сергей Демидов

Директор департамента операционных рисков, информационной безопасности и непрерывности бизнеса, Группа «Московская Биржа»

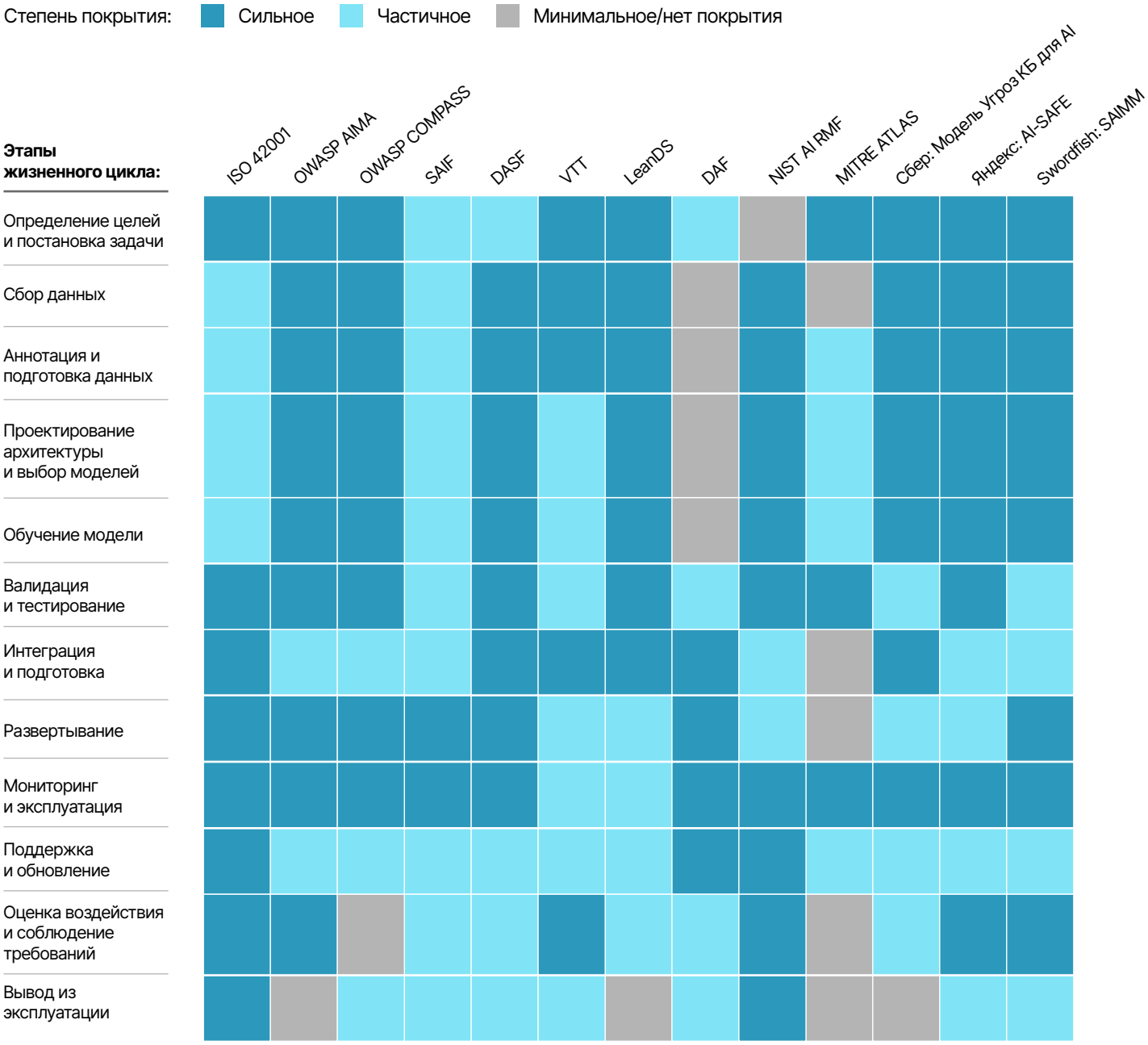
Фреймворки как основа модели процессов

Фреймворк/ Стандарт	Фокус	Уровень детализации	Кому подходит
ISO/IEC 42001:2023	Система управления ИИ, процессы, политика, соответствие требованиям	Высокий Процессный и управленческий	Крупные компании, госорганизации, компании с требованием сертификации
OWASP COMPASS	Управление безопасностью ИИ: процессы, разработка и эксплуатация	Высокий Много концепций, меньше практики	Командам разработки и эксплуатации ИИ, DevSecOps
Google SAIF	Безопасность AI/ ML, защита от угроз, принципы «secure-by-default»	Средний Структурные принципы, практические рекомендации	ИТ-компании и облачные провайдеры, команды безопасности, пользователи продуктов компании
Databricks DASF	Безопасность AI/ML, риски по жизненному циклу, тех. контроли	Высокий Технические, практические меры	Инженеры MLOps, ML, пользователи продуктов компании
VTT Framework	Управление рисками ИИ, ответственность, устойчивость	Средний Процессы, риск-ориентированный подход	Европейские компании, исследовательские проекты, R&D
LeanDS Maturity Index	Зрелость процессов, анализ данных и ML, оценка эффективности	Средний Опросники, модели оценки зрелости	Команды по анализу данных, стартапы, компании, которым нужно оценить зрелость процессов
DevSecOps Assessment Framework (DAF)	Встраивание безопасности в DevOps/MLOps	Высокий Процессно-технический	DevOps/MLOps-команды, компании с CI/CD и частыми релизами

Фреймворк/ Стандарт	Фокус	Уровень детализации	Кому подходит
NIST AI RMF	Управление рисками ИИ, надежность, безопасность, прозрачность	Высокий Структурный, риск-ориентированный подход)	Широкий круг: от госкорпораций до стартапов для формализации управления рисками
MITRE ATLAS	Карта атак и угроз	Высокий Технический, атакующие сценарии	Взлом и пентест, специалисты по кибербезопасности, исследователи ИИ
OWASP: AI Maturity Assessment (AIMA)	Оценка зрелости процессов AI Security	Средний Оценка зрелости процессов AI Security и уровней рисков	CISO, риск- и комплаенс-офицеры
Сбер: модель угроз для кибер-безопасности AI	Угрозы ИИ на всех этапах жизненного цикла PredAI и GenAI	Высокий Угрозы, последствия, объекты воздействия	Компании, задействованные в любых этапах ЖЦ ИИ — от моделирования до внедрения AI-продуктов
Яндекс: AI Secure Agentic Framework Essentials (AI-SAFE)	Безопасность и контроль поведения ИИ-агентов	Средний Практические принципы безопасного поведения AI-агентов, базовый набор требований для внедрения безопасных ИИ-систем.	Команды внедрения и DevOps
Swordfish: Secure AI Maturity Model (SAIMM)	Зрелость процессов обеспечения безопасности ИИ-систем, оценка эффективности	Высокий Процессный, практические меры	Организации, которые разрабатывают, эксплуатируют или закупают ИИ-системы

Матрица «Фреймворки безопасности VS Жизненный цикл ИИ-систем»

Маппинг фреймворков/стандартов на этапы жизненного цикла разработки ИИ-систем



Однако фреймворк сам по себе не решает проблему. Он задаёт рамку, но не наполняет её содержанием – это задача самой организации. Независимо от того, используется ли международный стандарт или национальная методика, любой фреймворк должен быть адаптирован под собственные процессы, культуру и зрелость. Только в этом случае он перестаёт быть набором рекомендаций и становится рабочим инструментом управления безопасностью.

Выбор фреймворка для управления безопасностью ИИ-систем

Выбор фреймворка – это не поиск универсального стандарта, а определение точки отсчёта для собственного пути. Ни один документ, даже самый детальный, не может в полной мере описать уникальные риски, процессы и культуру конкретной организации. Поэтому ключевая задача – не просто выбрать фреймворк, **а понять, какой из них способен стать частью вашей системы управления безопасностью ИИ.**

Отправной точкой служит **уровень зрелости организации**. Если процессы в сфере DevSecOps и управления рисками уже развиты, логично опираться на технически насыщенные модели вроде OWASP AI Security или Databricks DASF. Если же компания находится в стадии формирования подходов к ответственному использованию ИИ, стоит начать с рамок управленческой зрелости – таких как NIST AI RMF или AIMA.

Фреймворки для управления безопасностью ИИ-систем

ISO/IEC 42001 – управление и ответственность в сфере ИИ

ISO/IEC 42001 , Information technology — Artificial intelligence — Management system, IDT — международный стандарт, определяющий требования к системе менеджмента искусственного интеллекта (AIMS). Он задаёт единые правила
для ответственного, безопасного и прозрачного внедрения ИИ в организациях.

Стандарт охватывает процессы по управлению рисками, качеством данных, надежностью моделей, мониторингом, документацией и соблюдением этических принципов. Он помогает компаниям внедрять ИИ контролируемо, снижая операционные, юридические и репутационные риски.

NIST AI RMF и MITRE ATLAS – акцент на рисках и безопасности

NIST AI RMF формирует основу для управления рисками ИИ на всех стадиях жизненного цикла. Он акцентирует внимание на валидации, мониторинге и оценке воздействия ИИ-систем.

Фреймворк **MITRE ATLAS**, напротив, делает упор на фазу тестирования, защиту инфраструктуры и угрозы эксплуатации, оставаясь слабее в части ранних этапов – сбора и аннотации данных.

Оба документа применимы на поздних стадиях жизненного цикла, когда система уже развернута и необходимо обеспечить устойчивость к сбоям, атакам и нарушениям доверия.

Сбер: модель угроз для кибербезопасности AI

Модель угроз Сбера формирует системную основу для оценки рисков, связанных с безопасностью предиктивных и генеративных ИИ-систем. Она охватывает весь жизненный цикл ИИ — от подготовки данных и обучения до внедрения и эксплуатации моделей.

Модель структурирует типовые угрозы, затрагивающие данные, модели, инфраструктуру и процессы, и учитывает такие аспекты, как конфиденциальность, целостность и доступность. Подход помогает выявлять уязвимости, понимать возможные последствия и выбирать меры защиты.

SAIF и DASF – инженерные и прикладные подходы

SAIF (Secure Agentic Framework) описывает меры инженерной безопасности для этапов развёртывания, эксплуатации и поддержки, где особенно важен контроль автономных решений и их взаимодействия с внешними системами.

DASF (Databricks & AI Security Framework) фокусируется на этапах подготовки и аннотации данных, проектировании архитектуры и построении моделей, обеспечивая целостность и защищённость обучающей базы.

LeanDS, VTT и OWASP COMPASS – методические и практические стандарты

LeanDS, VTT и **OWASP COMPASS** не столько задают строгие требования к безопасности, сколько формируют методическую основу зрелости процессов и внедрения практик безопасного ИИ.

LeanDS помогает выстраивать гибкие и итерационные подходы к разработке ИИ, VTT объединяет академический и инженерный опыт управления рисками, а OWASP COMPASS предлагает практические ориентиры по governance, разработке и эксплуатации ИИ-систем.

Эти стандарты полезны как инструменты выравнивания процессов, внедрения безопасных практик и повышения зрелости организаций в области ИИ.

OWASP AIMA – зрелость и управляемость ИИ

AIMA (AI Maturity Assessment) оценивает уровень готовности компании к безопасному и ответственному использованию ИИ.

Она рассматривает зрелость по направлениям стратегия, данные, технологии, комплаенс и этика, помогая выявлять слабые места и формировать план развития AI Governance.

Фреймворк применим как инструмент диагностики и управленческого самоаудита.

Яндекс: AI-SAFE – безопасность генеративных и автономных ИИ

AI-SAFE (Яндекс) ориентирован на защиту генеративных моделей и LLM-систем.

Он охватывает риски утечек данных, промпт-инъекций и непредсказуемого поведения моделей, а также вводит меры интерпретируемости и ограничений на уровне исполнения.

AI-SAFE применим для корпоративных и клиентских сервисов, использующих автономных ассистентов и генеративные ИИ-решения.

Swordfish: SAIMM – зрелость процессов безопасности

Swordfish: Secure AI Maturity Model (SAIMM) объединяет управленческие и инженерные подходы для оценки зрелости и эффективности процессов безопасности ИИ.

Он рассматривает безопасность не как набор отдельных мер по защите ИИ-моделей и ИИ-агентов, а как интегрированную систему практик в рамках AISecOps, позволяя формировать долгосрочную стратегию развития AI Security.

Что такое «подходящий фреймворк»?

Важно помнить, что фреймворк – это не чек-лист, а карта местности, которая помогает ориентироваться, но не гарантирует успеха, если двигаться без осмысления контекста. Любой фреймворк должен быть встроен в реальные процессы организации: в процедуры обучения, корпоративную систему контроля рисков. Только тогда он перестаёт быть внешним требованием и становится инструментом развития. В конечном счёте, правильный выбор – это не про «какой фреймворк лучше», а про то, как организация выстраивает свою безопасность. Стандарты задают направление, но движение по этому пути требует осознанности, согласованности действий и готовности к постоянной адаптации – ведь вместе с самим искусственным интеллектом будет меняться и природа угроз.

ОСВЕДОМЛЁННОСТЬ И ОБУЧЕНИЕ

Рост внедрения искусственного интеллекта радикально меняет саму природу разработки безопасных систем. Рынок разработки ПО сегодня переживает трансформацию, сопоставимую по масштабу с переходом от ручного кодирования к DevOps, а затем к автоматизированным CI/CD-конвейерам. Появление ИИ-инструментов ускорило многие этапы жизненного цикла разработки, но при этом не снизило роль человека – напротив, она стала более стратегической.

Поскольку машины теперь способны быстро писать код, ценность инженера-человека смещается в сторону тех областей, где **требуются системное мышление и ответственность:** архитектура решений, управление техническим долгом, обеспечение безопасности и соответствие нормативным требованиям.

На этом фоне индустрия сталкивается с дефицитом квалифицированных кадров. Компании по всему миру испытывают трудности в поиске инженеров, обладающих компетенциями на стыке ИИ, безопасности и управления рисками.

Даже опытные профессионалы вынуждены постоянно обновлять знания – ведь угрозы эволюционируют быстрее, чем учебные программы. Традиционные подходы к повышению квалификации – сертификация, курсы, тренинги – больше не успевают за скоростью появления новых инструментов и векторов атак. Это делает **непрерывное обучение и развитие осведомлённости** ключевыми элементами стратегии AI Security.

Компетенции в области AI Security

MLSecOps-инженер

КОМПЕТЕНЦИИ:

- Глубокое понимание CI/CD и DevOps-практик применительно к ML-моделям.
- Знание инструментов контейнеризации и оркестрации.
- Умение настраивать конвейеры обучения, тестирования и развертывания моделей.

ЗАДАЧИ:

- Автоматизация развёртывания моделей с учётом требований безопасности.
- Настройка мониторинга производительности и стабильности моделей.
- Регулярная проверка устойчивости к сбоям и угрозам.

Инженер по информационной безопасности

КОМПЕТЕНЦИИ:

- Знание принципов сетевой и прикладной безопасности, протоколов шифрования.
- Опыт работы с SIEM/SOAR, WAF, IAM-системами.
- Навыки анализа уязвимостей, проведения пентестов, использования DLP и систем обнаружения аномалий.

ЗАДАЧИ:

- Настройка систем защиты API, управление доступами к данным и моделям.
- Обнаружение и предотвращение кибератак (DDoS, отравление данных, проникновения).
- Регулярный аудит конфигураций, обновление политики безопасности согласно текущим угрозам.

Специалист по управлению данными и качеством

КОМПЕТЕНЦИИ:

- Понимание принципов классификации данных, требований к их качеству и правовых норм обработки.
- Опыт работы с инструментами профилирования данных, оценкой качества.
- Знание методологий очистки, нормализации, а также мониторинга целостности данных.

ЗАДАЧИ:

- Разработка и внедрение стандартов качества данных, регулярная проверка на наличие пропусков и ошибок.
- Обеспечение соответствия данных регуляторным требованиям.
- Введение механизмов классификации, валидации.

Аналитик по рискам и комплаенсу

КОМПЕТЕНЦИИ:

• Понимание нормативно-правовых актов, регулирующих использование ИИ (законодательство о данных, отраслевые стандарты).

• Навыки проведения оценок рисков, подготовки отчётности для аудиторов и регуляторов.

• Умение оценивать юридические и репутационные последствия решений ИИ.

ЗАДАЧИ:

• Мониторинг изменений в законодательстве, адаптация внутренних политик под новые требования.

• Подготовка документов для аудита, взаимодействие с юридическим отделом.

• Анализ рисков, связанных с автономностью моделей, использование результатов для совершенствования методологий разработки и эксплуатации.

Этический консультант по ИИ

КОМПЕТЕНЦИИ:

• Понимание этических и социальных аспектов внедрения ИИ: недопущение дискриминации, предвзятости.

• Способность внедрить в организационную культуру принципы ответственного использования ИИ.

ЗАДАЧИ:

• Разработка этических норм и стандартов для использования ИИ.

• Оценка моделей на предмет справедливости, объяснимости и прозрачности решений.

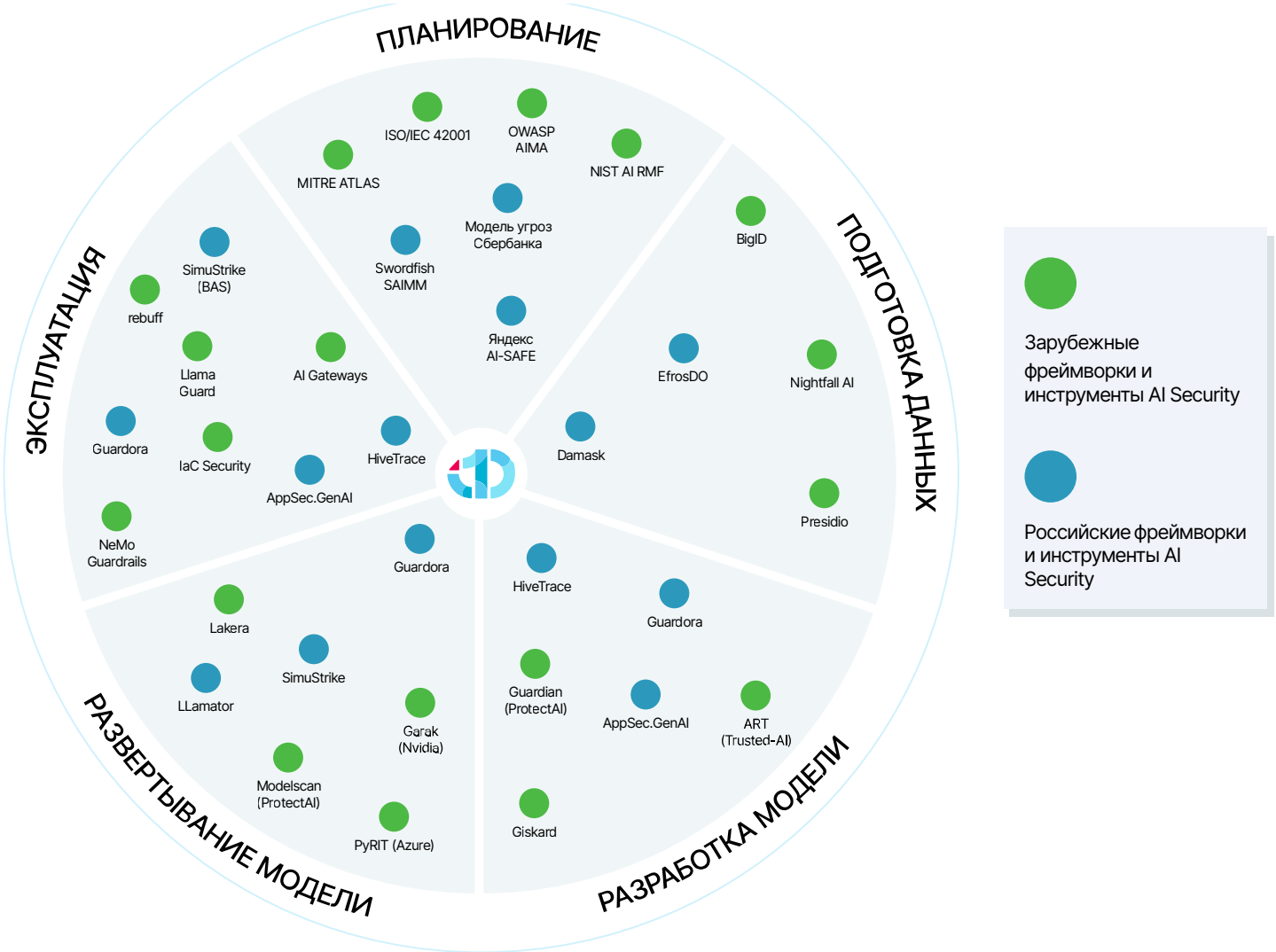
• Обучение команд принципам ответственного ИИ, контроль соблюдения этических норм.

КАРТА ИНСТРУМЕНТОВ AI SECURITY

Для перехода от концепций к практике, организациям необходима наглядная связь между задачами безопасности и конкретными инструментами, которые их решают. Эту функцию выполняют **технологические карты AI Security** – визуальные схемы, сопоставляющие угрозы, этапы жизненного цикла ИИ и доступные средства защиты.

Такие карты позволяют систематизировать меры безопасности: какие инструменты применяются при сборе и разметке данных, какие – при обучении и развертывании моделей, а какие обеспечивают защиту во время эксплуатации.

КАРТА ИНСТРУМЕНТОВ AI SECURITY от Ассоциации ФинТех и Swordfish Security*



*По данным Ассоциации ФинТех, Swordfish Security

Специализированные инструменты карты AI Security

Этап жизненного цикла ИИ	Средства защиты и их назначение	
	Зарубежные	Российские
1 **ПЛАНИРОВАНИЕ** Определение угроз и мер защиты на этапах проектирования и жизненного цикла ИИ-систем. Моделирование рисков, выбор стандартов и методик безопасной разработки.	• MITRE ATLAS – каталог угроз и сценариев атак на ИИ-системы; используется для моделирования угроз и определения векторов атак на этапе проектирования. • ISO/IEC 42001:2023 стандарт системы управления ИИ; требования к ответственному, безопасному и прозрачному использованию ИИ. • NIST AI RMF модель управления рисками ИИ; принципы оценки, снижения и мониторинга рисков на всех этапах жизненного цикла. • OWASP AI Maturity Assessment (AIMA) модель оценки зрелости безопасности ИИ, определяющая уровень процессов, рисков и защитных практик.	• Сбер: модель угроз для кибербезопасности AI – каталог ключевых угроз для ИИ-систем, используемый для оценки рисков и выбора мер защиты. • Яндекс: AI Secure Agentic Framework Essentials (AI-SAFE) фреймворк для безопасной разработки ИИ-систем; включает чек-листы и практики по управлению рисками и защите данных. • Swordfish: SAIMM модель зрелости защищенности искусственного интеллекта; предоставляет организациям единый ориентир и методологическую основу для оценки и повышения уровня безопасности в сфере ИИ.
2 **ОБЕСПЕЧЕНИЕ ДАННЫМИ** Обнаружение, классификация и анонимизация данных	• BigID – Data Intelligence Platform, обнаружение и контроль чувствительных данных. • Nightfall AI – защита от утечек текстов и кода (API, Slack, GitHub). • Presidio – open-source библиотека от Microsoft для деидентификации (PII detection + anonymization).	• Damask (АО «Дамаск Цифровая Безопасность») – динамическая подмена данных (токенизация/ детокенизация) для безопасного использования конфиденциальной информации. • EfrosDO – комплексная защита корпоративных данных и процессов, контроль утечек.

Этап жизненного цикла ИИ	Средства защиты и их назначение	
	Зарубежные	Российские
3 **РАЗРАБОТКА МОДЕЛИ** Проверка устойчивости и безопасности моделей	• ProtectAI Guardian – аудит ML-проектов, поиск уязвимостей, управление безопасностью моделей. • Giskard – тестирование моделей на bias и уязвимости. • ART (Adversarial Robustness Toolbox, Trusted-AI) – библиотека атак/защит против adversarial input.	• HiveTrace – тестирование безопасности моделей и LLM-поведения; выявление prompt-инъекций и аномалий при обучении. • Guardora – конфиденциальное обучение (privacy-preserving ML), защита тренировочных данных и алгоритмов. • AppSec.GenAI – анализ защищенности ИИ-моделей, предотвращение prompt-инъекций и утечек данных.
4 **РАЗВЁРТЫВАНИЕ МОДЕЛИ** Безопасное внедрение и защита API/ инференса	• Lakera – мониторинг и защита LLM от prompt-инъекций. • ModelScan (ProtectAI) – поиск уязвимостей в ML-моделях при релизе. • Garak (NVIDIA) – фреймворк тестирования безопасности LLM. • PyRIT (Azure) – проверка устойчивости AI-сервисов Microsoft.	• SimuStrike (Газинформсервис, LLM) – автоматизированный пентест (BAS) внешней и внутренней инфраструктуры, проверка защищённости окружения перед релизом модели. • Guardora – безопасный инференс, защита каналов передачи данных, интеграция криптографических средств. • LLamator – мониторинг поведения LLM, обнаружение аномалий и фильтрация вредоносных запросов.
5 **ЭКСПЛУАТАЦИЯ** Непрерывный мониторинг, фильтрация запросов/ответов, защита от атак в продакшн.	• Guardrails.ai и NeMo Guardrails (NVIDIA) – политика безопасных ответов, ограничение поведения LLM. • LlamaGuard – фильтрация токсичных/чувствительных запросов. • Rebuff – обнаружение prompt-инъекций. • ProtectAI / Cisco AI Validation – управление рисками и мониторинг моделей.	• HiveTrace – мониторинг поведения модели, выявление аномалий и атак во время работы. • SimuStrike (BAS) – непрерывный сценарный тест (red-team/blue-team) для проверки защищённости инфраструктуры. • AppSec.GenAI – анализ защищенности ИИ-моделей, предотвращение prompt-инъекций и утечек данных. • Guardora – безопасное дообучение (файнтьюнинг) ML-моделей в контуре клиента без передачи конфиденциальных данных.

Классы инструментов информационной безопасности

	ПЛАНИРОВАНИЕ	ОБЕСПЕЧЕНИЕ ДАННЫМИ	РАЗРАБОТКА МОДЕЛИ	РАЗВЁРТЫВАНИЕ МОДЕЛИ	ЭКСПЛУАТАЦИЯ
SGRC	✓	✓	✓	✓	✓
DLP					✓
IDM / IAM		✓	✓	✓	✓
Шифрование		✓	✓	✓	✓
SAST/SCA			✓	✓	
EDR		✓	✓	✓	✓
TSPM			✓	✓	
CI/CD Security			✓	✓	
WAF			✓	✓	✓
IDS / IPS				✓	✓
SIEM/SOAR			✓	✓	✓
UEBA					✓
ML Monitoring				✓	✓
AI-BOM			✓	✓	

Класс	Расшифровка
SGRC	Security Governance, Risk & Compliance – управление политиками, рисками и соответствием требованиям ИБ.
DLP	Data Leak Prevention – предотвращение утечек данных.
IDM / IAM	Identity Management\Identity & Access Management – управление идентификацией и контролем доступа пользователей.
Шифрование	Криптографическая защита данных (в покое и при передаче).
SAST/SCA	Static Application Security Testing/Software Composition Analysis – статический анализ и композиционный анализ.
EDR	Endpoint Detection & Response – защита и мониторинг конечных точек (серверов и рабочих станций разработчиков).
TSPM	Trustware Security Posture Management – анализ и управление безопасностью ИИ.
CI/CD Security	Защита конвейеров сборки и доставки ПО.
WAF	Web Application Firewall – файрвол веб-приложений.
IDS / IPS	Intrusion Detection / Prevention Systems – системы обнаружения и предотвращения вторжений.
SIEM/SOAR	Security Information & Event Management / Security Orchestration, Automation & Response – сбор событий, корреляция, реагирование.
UEBA	User and Entity Behavior Analytics – поведенческий анализ пользователей/систем.
ML Monitoring	Мониторинг поведения ML-моделей и аномалий.
AI-BOM	AI Bill of Materials – перечень компонентов, данных и зависимостей ИИ-модели для аудита, прозрачности и управления рисками цепочки поставок ИИ.

Безопасный искусственный интеллект – это не технология, а экосистема. Она объединяет людей, процессы и инструменты, опираясь на культуру осознанного управления рисками. Фреймворки задают структуру, обучение формирует компетенции, а технологические карты превращают принципы в конкретные действия. В совокупности эти элементы образуют систему, в которой безопасность становится естественной частью жизненного цикла ИИ.

Построение AI Security – это не финальная цель, а путь к зрелости, на котором организация учится сочетать инновации и ответственность, открытость и защиту, скорость и надёжность. Именно в этом равновесии и заключается основа доверия к искусственному интеллекту будущего.

Что является основой для создания доверенных технологий ИИ?

Современные системы искусственного интеллекта требуют не только точных моделей, но и комплексной защиты на всех этапах их жизненного цикла – от подготовки данных и проектирования архитектуры до развертывания и мониторинга деятельности системы с контролем точности и выдаваемой информации.

Сегодня ключевыми стали технологии безопасной разработки, инструменты проверки датасетов на токсичность и утечку, средства анализа уязвимостей моделей и механизмы контроля цепочек поставок. Кроме того, очень плотно расположены вопросы ИБ и функциональной надёжности, что формирует понятие «доверенные технологии ИИ». Именно такие решения создают фундамент доверия к ИИ и позволяют организациям внедрять его без риска.

Дмитрий Служеникин

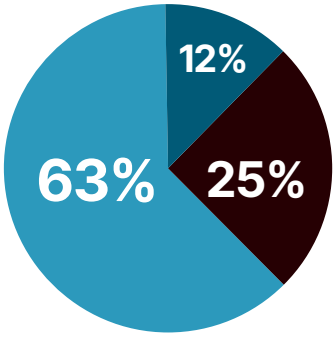
Секретарь Консорциума,

Консорциум исследований безопасности технологий искусственного интеллекта

МНЕНИЕ УЧАСТНИКОВ ФИНАНСОВОГО РЫНКА ОТНОСИТЕЛЬНО ПРОЦЕССА «AI SECURITY»

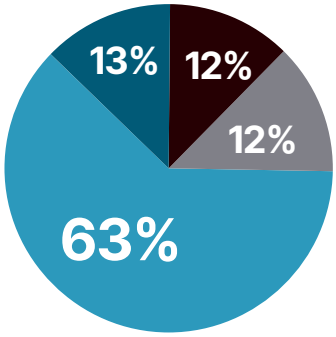
Использование ИИ-технологий

Количество сотрудников, занимающихся вопросами ИИ/машинного обучения:



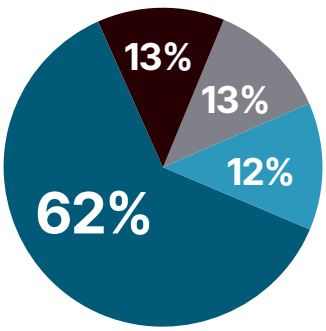
- Крупные команды (более 20 человек)
- Средние команды (от 5 до 20 человек)
- Малые команды (менее 5 человек)

Типы команд, ответственных за безопасность ИИ-систем:



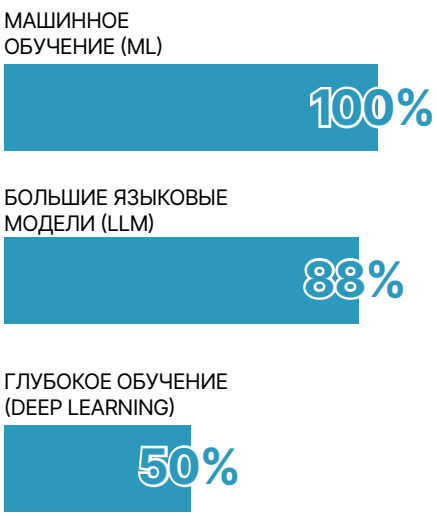
- Совместная ответственность нескольких команд
- Команда ИБ (без явного фокуса на ИИ)
- Другой тип команды
- Нет ответственного лица

Доля компаний, доверяющих результатам и выводам ИИ:



- Очень доверяют
- Скорее доверяют
- Доверяют ниже среднего
- Не применимо/не используют ИИ

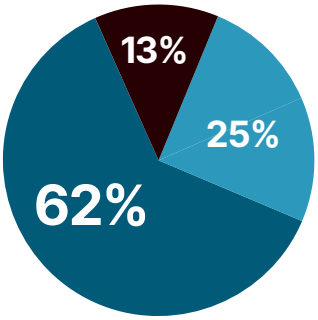
Ключевые технологии ИИ в финтех-организациях:



**По данным опроса Ассоциации ФинТех, декабрь 2025 года*

Безопасность ИИ-систем

Доля компаний, которые столкнулись с инцидентами безопасности ИИ в 2025 году:



- Столкнулись несколько раз
- Инцидентов не было
- Не располагают информацией

Ключевые угрозы безопасности ИИ, которые отмечают представители финтех-компаний:

УТЕЧКИ КОНФИДЕНЦИАЛЬНЫХ ДАННЫХ — 75%

ОШИБКИ И ПРЕДВЗЯТОСТЬ МОДЕЛЕЙ — 38%

НЕСАНКЦИОНИРОВАННЫЙ ДОСТУП К AI — 36%

НЕКОРРЕКТНАЯ ОБРАБОТКА ВЫХОДНЫХ ДАННЫХ — 33%

ИНЪЕКЦИИ ПРОМПТОВ — 25%

ЧРЕЗМЕРНО РАСШИРЕННЫЕ ПРАВА АГЕНТОВ — 25%

Этапы жизненного цикла ИИ-моделей, на которых реализуются меры безопасности:

1. На этапе подготовки и обработки данных — 75%
2. Во время обучения модели — 36%
3. При тестировании и валидации — 63%
4. При развертывании в продакшн — 50%
5. При эксплуатации ИИ — 34%
6. Не реализуем меры безопасности — 13%

Популярность инструментов, используемых для безопасности ИИ:

ИНСТРУМЕНТЫ МОНИТОРИНГА И АУДИТА МОДЕЛЕЙ — 63%

РЕШЕНИЯ ПО ЗАЩИТЕ ДАННЫХ (НАПРИМЕР, ШИФРОВАНИЕ, ТОКЕНИЗАЦИЯ) — 36%

СОБСТВЕННЫЕ ВНУТРЕННИЕ РАЗРАБОТКИ — 50%

НЕ ПРИМЕНЯЕМ ИНСТРУМЕНТЫ AI SECURITY — 14%

СПЕЦИАЛИЗИРОВАННЫЕ ПЛАТФОРМЫ AI SECURITY — 37%

ДРУГИЕ — 13%

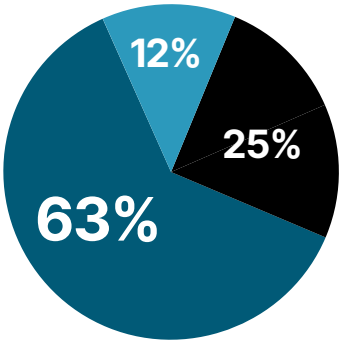
**По данным опроса Ассоциации ФинТех, декабрь 2025 года*

Интегрируются ли решения AI Security в существующую ИТ-инфраструктуру финтех-компании?

75%
финтех-компаний успешно интегрировали

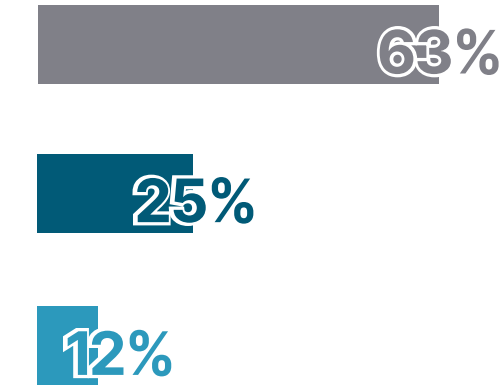
25%
финтех-компаний только планируют интеграцию

Используют ли финтех-компании подходы к обеспечению устойчивой модели к атакам?



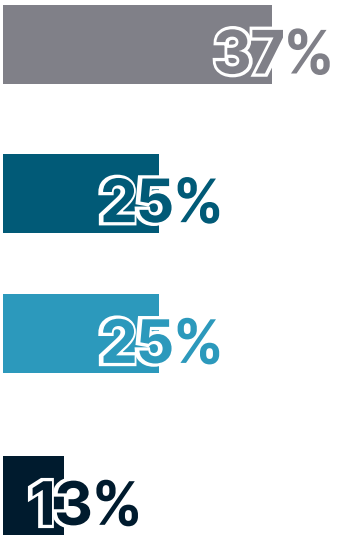
- Регулярно проводят тестирование на устойчивость
- Планируют внедрение
- Не применяют такие практики

Оценивают ли финтех-компании зрелость процессов AI Security?



- Используют модели зрелости
- Частично адаптируют внутренние подходы
- Рассматривают возможность

Проводят ли финтех-компании проверку данных на наличие вредоносных примеров?

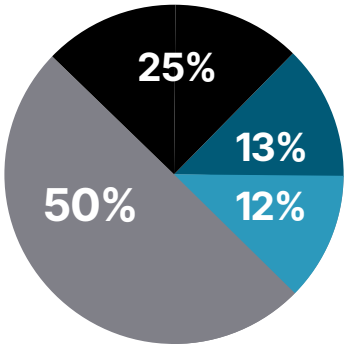


- Проверка вручную по выборке
- Автоматическая проверка при загрузке
- Нет проверки
- Планируется внедрение

**По данным опроса Ассоциации ФинТех, декабрь 2025 года*

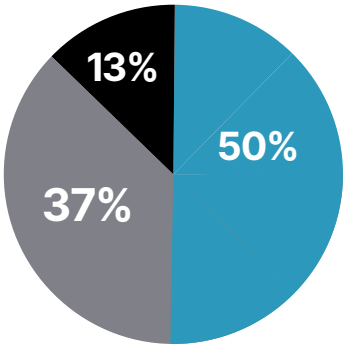
Этика и ответственность при использовании ИИ-систем

Оценивают ли финтех-компании этические риски использования ИИ?



- Да, оценка встроена в разработку моделей
- Да, проводится эпизодически
- Нет, не оценивают
- Нет, но обсуждается внедрение

Проводят ли финтех-компании аудит моделей на предмет соответствия политике безопасности и этики?



- Да, по внутренним стандартам
- Нет, но планируют
- Нет, не проводят

Планы финансовой индустрии по развитию инструментов AI Security на 2026 год:

1. Повышение осведомленности и обучение сотрудников — **88%**

2. Разработка внутренней политики по AI Security — **75%**

3. Интеграция инструментов защиты — **63%**

4. Проведение аудита или тестирования ИИ-систем — **62%**

**По данным опроса Ассоциации ФинТех, декабрь 2025 года*

РЕКОМЕНДАЦИИ

Обеспечение безопасности искусственного интеллекта является не просто технической задачей, а сложной, многогранной проблемой, от решения которой зависит устойчивое и ответственное развитие технологий в долгосрочной перспективе.

Для внедрения успешной стратегии внедрения AI Security компании должны придерживаться следующих принципов развития ИБ:

01 ПОВЫШЕНИЕ КОМПЕТЕНЦИЙ

Человеческий фактор остается критически важным звеном. Требуется целенаправленная работа по обучению и повышению осведомленности всех причастных специалистов – от разработчиков и data-саентистов до руководителей и конечных пользователей – в вопросах кибербезопасности, этики ИИ и управления рисками.

02 ВЫСТРАИВАНИЕ БЕЗОПАСНЫХ И ДОВЕРЕННЫХ ПРОЦЕССОВ

Безопасность должна быть встроена не только в код, но и в организационные процедуры. Это включает внедрение сквозных фреймворков управления рисками, регулярный аудит моделей, создание систем документирования и отслеживаемости (MLOps/LLMOps), а также формирование прозрачных и подотчетных процессов принятия решений, что является основой для доверия к технологиям ИИ со стороны общества, бизнеса и регуляторов.

03 РАЗВИТИЕ ИНСТРУМЕНТОВ И ТЕХНОЛОГИЙ ЗАЩИТЫ ИИ

Безопасность должна быть встроена не только в код, но и в организационные процедуры. Это включает внедрение сквозных фреймворков управления рисками, регулярный аудит моделей, создание систем документирования и отслеживаемости (MLOps/LLMOps), а также формирование прозрачных и подотчетных процессов принятия решений, что является основой для доверия к технологиям ИИ со стороны общества, бизнеса и регуляторов.

Таким образом, безопасность ИИ – это непрерывный динамический процесс, требующий тесной интеграции технологических решений, управленческих практик и кадрового потенциала. Только такой сбалансированный и проактивный подход позволит раскрыть колоссальный потенциал искусственного интеллекта, минимизируя сопутствующие риски и обеспечивая его надежное и ответственное использование.

Над исследованием работали:

Ассоциация ФинТех

Марианна Данилина
Руководитель Управления стратегии, исследований и аналитики

Александр Товстолип
Руководитель Управления информационной безопасности

Мария Чернышева
Ведущий бизнес-аналитик

Сергей Лапин
Эксперт Управления информационной безопасности

Привлеченные эксперты, ГК Swordfish Security:

Александр Пинаев
Генеральный директор ГК Swordfish Security

Юрий Сергеев
Директор по стратегии, Генеральный Партнер

Антон Башарин
Старший управляющий директор

Юрий Шабалин
Старший директор по развитию ИИ-технологий

Альбина Аскерова
Руководитель направления по взаимодействию с регуляторами

Привлеченные эксперты:

Сергей Демидов
Директор департамента операционных рисков, информационной безопасности и непрерывности бизнеса, Группа «Московская Биржа»

Дмитрий Служеникин
Секретарь Консорциума исследований безопасности технологий искусственного интеллекта

Борис Захир
Главный эксперт, Департамент кибербезопасности, Сбер

Дизайн:

Александра Щедрина
Креативный директор, Ассоциация ФинТех

Татьяна Симчук
Дизайнер, Ассоциация ФинТех

АССОЦИАЦИЯ ФИНТЕХ

ИССЛЕДОВАНИЯ & АНАЛИТИКА

research.analytics@fintechru.org

АССОЦИАЦИЯ ФИНТЕХ

ТЕЛЕГРАМ-КАНАЛ АФТ

WWW.FINTECHRU.ORG

Ассоциация ФинТех *основана в конце 2016 г. по инициативе Банка России и ключевых участников отечественного финансового рынка. Это уникальная площадка для конструктивного диалога регулятора с представителями бизнеса.*

Здесь формируется экспертная оценка инновационных технологий с учетом международного опыта, а также разрабатываются концепции финансовых технологий и подходы к их внедрению.