Yandex Cloud

Безопасность ИИ-систем в Yandex Cloud

Обзор возможностей обеспечения информационной безопасности сервисов ИИ

yandex.cloud

Вызовы безопасности ИИ-систем

Современные ИИ-системы не только ускоряют бизнес-процессы, но и открывают новые угрозы безопасности, которые не учитывают классические методы защиты. Кроме того, возникают этические и правовые аспекты, которые необходимо решить при разработке и внедрении систем ИИ.

Команда Yandex Cloud работает над созданием инструментов и экспертизы встроенных в Yandex AI Studio для упрощения решения этих задач.

Новые угрозы информационной безопасности

Появляются угрозы позволяющие получить несанкционированный доступ к данным и использование недекларированных возможностей систем. Примеры:

Data Poisoning

В даркнете появляются предложения по заражению открытых датасетов.

Утечка моделей

Компании теряют IP: модели крадут, запрашивая их через открытые интерфейсы.

Adversarial Malware

Заражённые изображения, которые выглядят безобидно, но вызывают сбои в работе ML-моделей.

Model Inversion & Privacy Leakage

Злоумышленники извлекают персональные данные из обученных моделей.

Fake Model Deployment

Подмена модели на «клон» с бэкдором в продакшне, особенно в опенсорс-цепочках.

LLM Prompt Injection

Новые атаки на контекст и подсказки, позволяющие обойти ограничения генерации.

Этические и правовые аспекты

Технологии ИИ могут применены для различных недобросовестных целей, кроме того, необходимо поддерживать высокий уровень доверия к искусственному интеллекту как среди пользователей, так и на уровне общества.

Также возникают задачи обеспечения требований приватности и защиты данных включая требования Федерального закона № 152-ФЗ «О персональных данных» и других.

Возможности обеспечения безопасности Yandex AI Studio

На данный момент инструменты создания ИИ на базе AI Studio позволяют получить слой безопасности на основе следующих возможностей:

Возможность	Какая проблема решается	Название функциональности
Контроль потоков данных и запрет обучения на данных пользователя	Контроль данных	Запрет логирования
Изоляция данных с помощью приватных соединений для пользователей	Изоляция данных	Private endpoints
Обозреваемость и мониторинг логов и событий безопасности	Мониторинг безопасности	Audit Trails
Контроль запросов и выводов с помощью встроенной этики	Этика ИИ	Guardrails
Точечный контроль доступа к ресурсам на основе ролей	Управление доступом	Identity and Access Management
Контроль действий сотрудников Yandex Cloud на уровне гипервизоров и баз данных	Контроль подрядчика	Access Transparency
Защищенные и выделенные каналы доступа к инфраструктуре ИИ	Изоляция данных	Cloud Interconnect
Возможность шифрования каналов с помощью сертифицированных криптошлюзов на основе алгоритмов ГОСТ	Шифрование данных	ГОСТ VPN

.

Экспертиза по безопасности ИИ

Команда Yandex Cloud постоянно уделяет внимание и силы на развитие внутренней экспертизы обеспечения безопасности ИИ-систем в том числе используемых для внутренних задач безопасности с учетом современных угроз безопасности и постоянно делиться с сообществом ИБ,

В том числе, была разработана методология которая помогает системно внедрять безопасность на каждом этапе жизненного цикла ИИ:
AI Secure Agentic Framework Essentials (AI-SAFE)

AI Secure Agentic Framework Essentials (AI-SAFE)
В данном документе мы собрали примеры угроз и описали риски, которые могут возникнуть на пяти уровнях работы с генеративными технологиями: от интерфейса до инфраструктуры и оркестрации. Для каждой угрозы приводятся рекомендации и оценка рисков.

Соответствие требованиям регуляторов и стандартов

152-ФЗ

Сервис Yandex Foundation Models прошёл независимый аудит безопасности. Это заключение подтверждает, что при работе с компонентами сервиса обеспечено соответствие требованиям федерального закона от 27 июля 2006 года № 152-ФЗ «О персональных данных» согласно приказу ФСТЭК России от 18 февраля 2013 года № 21.

ISO 42001

Стандарт ISO/IEC 42001:2023 определяет требования к системе управления искусственным интеллектом (ИИ). Соответствие стандарту помогает организациям минимизировать риски, связанные с ИИ, повысить прозрачность и доверие к ИИ-решениям, оптимизировать процессы разработки и использования ИИ, а также обеспечить соответствие регуляторным и этическим требованиям.

Совместная ответственность за безопасность

Безопасность систем, использующих облачные сервисы, требует разделения ответственности между клиентом (владельцем конечной системы) и провайдером (владельцем облачной инфраструктуры)

Область	**Yandex Cloud (провайдер)**	**Клиент**
Физическая безопасность	Серверное оборудование. Защита дата-центров, физической инфраструктуры	—
Инфраструктура платформы	Отказоустойчивость платформы, защита сети, гипервизоров и компонентов инфраструктуры	—
Сервисы и компоненты AI Studio	Безопасность и доступность сервиса. Реализация компонентов Model Gallery, Agent Atelier, MCP Hub, API (Responses, Realtime, Vector Store)	—
ML-модели и их работа	Обеспечение безопасной работы моделей Alice AI LLM, YandexGPT, YandexART, опенсорсные SOTA- и мультимодальные модели и дообученные пользователем модели LoRA. Обновления и поддержка моделей. Библиотека промптов	Выбор подходящей модели, настройка параметров, дообучение моделей на своих датасетах, промпт-инжиниринг
Доступность и целостность данных	Мониторинг, логирование, резервное копирование данных платформы	—

Область	Yandex Cloud (провайдер)	Клиент
Управление доступом	Identity and Access Management. Yandex Identity Hub. Механизмы аутентификации, реализация ролевой модели	Назначение прав пользователям, управление учетными записями, контроль доступа к ресурсам
Данные клиента	Защита хранилища данных, шифрование при передаче и хранении, безопасность на уровне инфраструктуры платформы	Классификация данных, разграничение доступа, защита конфиденциальной информации
Разработка агентов и приложений	Предоставление API, Yandex Cloud ML SDK, Agent Atelier (среда разработки агентов) и других инструментов	Разработка логики агентов, настройка workflows, тестирование промптов, интеграция в бизнес-процессы
Дообучение моделей	Возможность дообученния моделей на клиентских датасетах. Инфраструктура Tuning API, среда для обучения	Подготовка датасетов, выбор параметров дообучения, контроль качества результатов
Интеграция с внешними системами	Предоставление MCP Hub, Vector Store API, Yandex Search API, Vision OCR, Translate API, SpeechKit и другие	Настройка подключений к корпоративным системам, безопасность интеграций,

Область	Yandex Cloud (провайдер)	Клиент
		управление внешними инструментами
Мониторинг приложений	Мониторинг инфраструктуры платформы и её компонентов	Мониторинг прикладного уровня, логирование действий агентов, анализ использования
Compliance и регуляторные требования	Соответствие регуляторным требованиям и международным стандартам: 152-ФЗ, ISO/IEC 27001, ISO/IEC42001 Предоставление информации о мерах безопасности клиентам	Разработка собственной политики информационной безопасности. Соблюдение отраслевых и регуляторных требований, аудит безопасности и использования данных