ИИ-технологии в системе безопасности Яндекса и защита нейросетей

Как технологии искусственного интеллекта Яндекса обеспечивают защиту на трёх уровнях: инфраструктуры, сервисов и пользователей.



2025 год

Оглавление

1. **Введение**
2. **ИИ как основа безопасности экосистемы**
   - Роль ИИ в системе защиты Яндекса
   - Интеграция ИИ в жизненный цикл разработки
3. **ИИ для мониторинга инфраструктуры**
   - Мониторинг и реагирование
   - Роль ИИ-технологий в SOC
   - Защита от DDoS: система «Антиробот»
4. **Безопасность ИИ моделей**
   - Безопасность данных и моделей
     - Подготовка данных и этичность моделей
     - Защита процесса обучения
     - Безопасная эксплуатация и человеческий контроль
   - Наш подход к безопасности ИИ
   - Принципы ИИ Яндекса
   - Сертификаты и аудиты
   - «Охота за ошибками»
5. **Защита пользователей: ИИ в сервисах**
   - Безопасность аккаунтов
   - Безопасность в интернете
   - Защита от телефонного мошенничества
   - Защита общения
   - Безопасная коммерческая среда
   - Защита бизнеса
6. **Обучение сотрудников и пользователей**
   - Культура безопасности в компании
   - Просветительские инициативы
     - «Урок цифры»
     - «Цифровой ликбез»
     - Безопасность для младших классов
     - Защита старшего поколения
7. **Заключение**

Введение

Цифровой мир усложняется: атаки становятся всё изощрённее, а обнаружить их всё труднее. Классические методы защиты, которые опираются на уже известные признаки вредоносной активности, зачастую не успевают реагировать на новые угрозы.

Современная ИТ-безопасность требует подхода, который позволяет не просто своевременно выявлять инциденты, а предугадывать действия злоумышленников.

Технологии искусственного интеллекта лежат в основе нашей системы безопасности — они помогают Яндексу действовать на опережение. Алгоритмы в реальном времени анализируют как исходный код сервисов, так и события в инфраструктуре. Они моментально фиксируют аномалии — от возможных уязвимостей на этапе создания сервисов до нестандартной попытки входа в Яндекс ID и всплесков подозрительного трафика. Это позволяет устранять риски и блокировать атаки до того, как они нанесут ущерб.

ИИ помогает защищать пользователей — например, предотвращать переход на вредоносные сайты и предупреждать о мошеннических звонках.

Надёжная защита начинается с безопасности самих технологий, поэтому устойчивость к атакам заложена в архитектуру наших ИИ-сервисов ещё на этапе проектирования. Мы регулярно организуем аудиты на устойчивость ИИ-моделей к манипуляциям, проводим стресс-тесты алгоритмов, стремимся к тому, чтобы решения, которые принимает ИИ, были безопасными, и делимся своим опытом с рынком.

В этом отчёте мы расскажем, как наши технологии искусственного интеллекта обеспечивают защиту инфраструктуры и пользователей. Отдельное внимание мы уделим тому, как обеспечивается безопасность самих ИИ-моделей.

ИИ как основа безопасности экосистемы

Наша главная задача — обеспечить непрерывную и надёжную защиту пользователей и инфраструктуры. ИИ-технологии позволяют построить самообучающуюся систему безопасности, в которой скорость работы, глубина анализа и масштаб охвата защитных механизмов в разы выше, чем в системах предыдущих поколений.

- Роль ИИ в системе защиты Яндекса
- Интеграция ИИ в жизненный цикл разработки

Роль ИИ в системе защиты Яндекса

В контуре безопасности Яндекса искусственный интеллект работает совместно с нашими специалистами. Он снимает рутинную нагрузку, помогает быстрее находить аномалии и сфокусироваться на реальных угрозах. В рамках этого подхода ИИ берёт на себя четыре ключевые функции.

Непрерывный мониторинг

Мы непрерывно контролируем состояние инфраструктуры и сервисов в фоновом режиме, получая полную и актуальную информацию в любое время. Нейросети позволяют автоматизировать реагирование на инциденты и потенциальные факторы риска — ещё до того, как ими могут воспользоваться злоумышленники.

Прогнозирование

Центр мониторинга безопасности (SOC) использует ИИ-технологии для анализа данных о новых киберугрозах в мире. Алгоритмы моделируют, как новые методы атак могут быть применены к инфраструктуре Яндекса, и ищут потенциально слабые места. Это позволяет перейти от реагирования на инциденты к их предотвращению.

Автоматизированный анализ и приоритизация

В огромном потоке данных важно отделять реальные угрозы от информационного шума. ИИ-инструменты помогают автоматически обрабатывать находки сканеров: классифицировать их, определять наиболее значимые аномалии и быстрее реагировать на возможные угрозы. Это позволяет экспертам не тратить время на ручной разбор ложных или тривиальных срабатываний, а фокусироваться на реальных угрозах в ключевых точках инфраструктуры.

Масштабирование

ИИ выступает в роли виртуального аналитика, способного обрабатывать огромные объёмы данных и обеспечивать непрерывную защиту сервисов даже при пиковых нагрузках на инфраструктуру.

Интеграция ИИ в жизненный цикл разработки

> Служба информационной безопасности Яндекса (СИБ) работает в связке с командами разработчиков на всех этапах создания продуктов:

1. На этапе идеи помогает спроектировать безопасную архитектуру.
2. В процессе разработки проводит промежуточные аудиты и проверки кода.
3. Сопровождает проект вплоть до запуска, давая «зелёный свет» на релиз.

Внедрение ИИ-технологий позволяет эффективно сочетать возможности нейросетей и человеческий опыт: алгоритмы обрабатывают большие массивы данных, а специалисты ставят задачи ИИ, совершенствуют систему защиты продуктов и отвечают за их качество. Это позволяет повысить надёжность сервисов и быстрее находить возможные уязвимости.



Мы интегрировали ИИ-технологии в каждый этап разработки новых продуктов.

Проектирование и анализ

> В 2025 году мы запустили ИИ-инструмент NeuroSecReview, который подключается на самом раннем этапе создания сервиса и помогает закладывать безопасную основу ещё до написания первой строчки кода.

NeuroSecReview анализирует описание будущей архитектуры, проверяет полноту исходных данных и автоматически формирует чеклисты требований безопасности, специфичных для конкретного сервиса.

В 3 раза

ускорилась проверка безопасности архитектуры благодаря NeuroSecReview

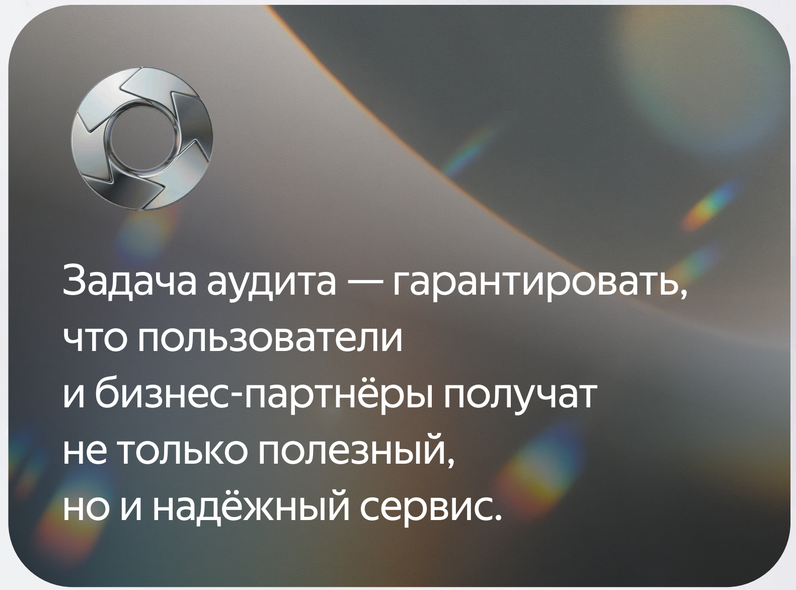
Разработка и тестирование

> Гибкие настройки наших сканеров позволяют проверять код на любом этапе разработки. Это помогает адаптировать проверки под размер проекта и активность команды: мы можем сканировать каждое отдельное изменение в коде в реальном времени, проверять крупные обновления перед их слиянием с основным проектом или запускать регулярный аудит для больших и стабильных систем.

Чтобы освободить инженеров от рутинного разбора отчётов от сканеров, мы используем ИИ-помощника. Он берёт на себя первичную обработку данных и автоматически:

1. Проверяет находки, отделяя реальные уязвимости от ложных срабатываний сканеров.
2. Анализирует уникальность находки: сверяется с базой известных инцидентов, чтобы не создавать дубли задач и проверок.
3. Даёт рекомендации: генерирует техническую документацию с описанием проблемы и предлагает готовые инструкции по исправлению кода, что ускоряет устранение рисков.

Перед релизом все продукты и обновления проходят обязательные аудиты безопасности. Эксперты СИБ проверяют сервис на устойчивость к атакам и убеждаются, что все архитектурные требования к безопасности выполнены, а уязвимости, найденные сканерами, устранены.



Эксплуатация и мониторинг

> Когда сервис становится доступен пользователям, включаются дополнительные механизмы активной защиты:

На этом этапе ИИ-инструменты анализируют отчёты об угрозах, прогнозируют развитие атак и ищут дубликаты в отчётах внешних исследователей. Это позволяет специалистам фокусироваться только на новых и критичных угрозах.



SOC (Центр мониторинга)

Круглосуточно отслеживает инциденты.



Система «Антиробот»

Отражает DDoS-атаки и защищает от других автоматизированных угроз, например ботов, сканеров, массовых регистраций аккаунтов и брутфорса — попыток перебора паролей.



Программа «Охота за ошибками»

Мы принимаем отчёты о возможных уязвимостях от внешних исследователей.

ИИ для мониторинга инфраструктуры

Для надёжной работы сервисов Яндекса мы выстроили масштабную систему защиты, которая непрерывно анализирует миллионы событий в секунду в нашей ИТ-инфраструктуре. При таких объёмах критически важна скорость реакции. Поэтому искусственный интеллект стал ключевым элементом центра мониторинга: он выступает первой линией защиты, автоматизирует обработку данных и отсеивает информационный шум.

- **Мониторинг и реагирование**
- **Роль ИИ-технологий в SOC**
- **Защита от DDoS: система «Антиробот»**

Мониторинг и реагирование

Специалисты Центра мониторинга безопасности (SOC) разрабатывают и внедряют «детекты» — алгоритмы, реагирующие на признаки кибератак или аномальной активности в инфраструктуре. Каждое срабатывание алгоритмов создаёт событие, требующее анализа и ответных действий.

Обычно события последовательно обрабатывают инженеры SOC: сначала они отсеивают типовые срабатывания, затем передают экспертам сложные инциденты для расследования. Мы изменили этот подход — теперь первой линией поддержки для обработки событий выступает искусственный интеллект.

Роль ИИ-технологий в SOC

> ИИ берёт на себя рутину по агрегации данных, анализирует историю активности и оценивает степень угрозы, превращая срабатывание алгоритма в развёрнутую готовую сводку для аналитика. Это позволяет нам решать три задачи:

- Повысить точность детектирования, охватив большее число вариантов атак.
- Повысить скорость реакции на задачи и сразу отправлять готовый отчёт для расследования.
- Снизить нагрузку на инженеров, автоматизируя рутинные этапы первичного анализа.

Роль ИИ-технологий в SOC

> Для повышения эффективности SOC в 2025 году мы внедрили ряд специализированных ИИ-инструментов:

Около 74%

правил ИИ-суммаризатора внедряют в работу

Нейроанализ

Автоматизирует многие сценарии обработки событий с помощью больших языковых моделей (LLM). ИИ-инструмент может либо полностью разобрать инцидент, либо предоставить дополнительный контекст аналитику.

LLM-анализатор файлов

Обеспечивает быстрый анализ подозрительных файлов внутри инфраструктуры, выдавая оценку риска и описание вредоносного поведения.

Суммаризатор ложных срабатываний

Выявляет повторяющиеся ложные срабатывания и предлагает готовые правила для их исключения.
Это позволяет снизить нагрузку на команду реагирования и сосредоточиться на расследовании реальных угроз.

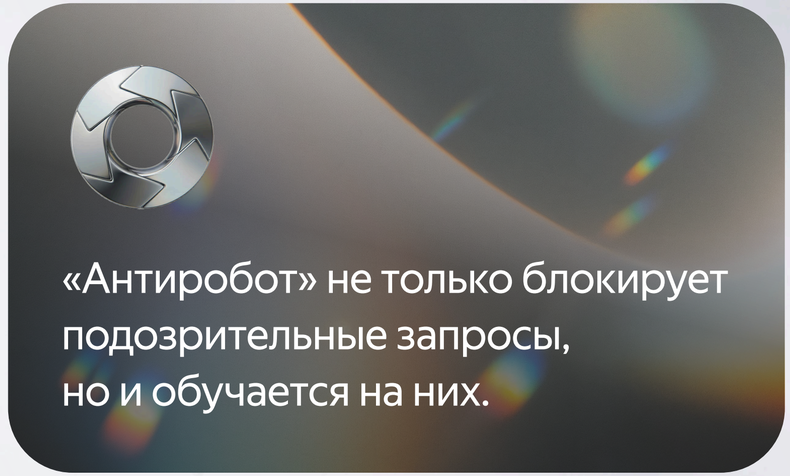
Защита от DDoS: система «Антиробот»

Для защиты инфраструктуры от DDoS-атак мы используем собственную разработку — систему «Антиробот». Это ИИ-сервис, который в режиме реального времени анализирует входящий интернет-трафик на сервисах Яндекса, с высокой точностью отделяя поведение реальных пользователей от активности роботов (ботнетов).

Более 99,9% атак на сервисы Яндекса отражаются «Антироботом» автоматически, без участия инженеров. Ежемесячно ИИ предотвращает в среднем 70 сложных DDoS-атак уровня L7, нацеленных на нарушение работы сервисов. В пиковые месяцы их число достигает 150, в спокойные составляет 50–60.

Защита от DDoS: система «Антиробот»

После нейтрализации угрозы мы разбираем и анализируем каждую DDoS-атаку. 75% инцидентов проходят автоматический анализ с помощью ИИ: система классифицирует методы атакующих и передаёт данные в обучающую выборку.
Это позволяет в короткие сроки адаптироваться к новым типам угроз, так как злоумышленники часто модифицируют способы и направления атак.



866
сложных DDoS-атак заблокировал «Антиробот» в 2025 году

До 150
DDoS-атак предотвращает ИИ в пиковые месяцы

70
DDoS-атак в среднем предотвращает ИИ каждый месяц

Безопасность ИИ-моделей

Мы обеспечиваем строгий контроль безопасности самих ИИ-технологий. Наши модели проходят комплексные проверки на устойчивость к атакам и манипуляциям задолго до того, как становятся частью продуктов для пользователей.

- Безопасность данных и моделей
- Сертификаты и аудиты
- Наш подход к безопасности ИИ
- «Охота за ошибками»
- Принципы ИИ Яндекса

Безопасность данных и моделей

> Яндекс развивает семейство собственных генеративных ИИ-моделей Alice AI.
> Это модели для работы с текстом (Alice AI LLM), изображениями (Alice AI ART) и мультимодальными данными (Alice AI VLM).
> Самые новые и мощные модели Яндекса лежат в основе нейросети Алиса AI.

Мы выстроили многоуровневый контроль жизненного цикла наших ИИ-моделей. Он охватывает все этапы: от сбора данных для обучения до защиты нейросетей в продуктах.

Подготовка данных и этичность моделей

> Мы обучаем модели следовать требованиям безопасности в несколько этапов. На первом этапе ответы, затрагивающие потенциально опасные или чувствительные темы, переформулируются в корректные и используются как эталонные примеры.

Например, модель должна учитывать многоконфессиональность, избегать дискриминации по гендерному и другим признакам, приводить в ответах различные мнения, а по медицинским вопросам — предлагать обратиться к специалистам и предостерегать пользователей от принятия решений исключительно на основе ответов ИИ. На втором этапе подключается отдельная модель-оценщик, которая поощряет наиболее безопасные и этичные формулировки. Такой подход позволяет создавать не только технологически эффективные, но и этичные алгоритмы.

При ответе на часть запросов пользователей, если это необходимо, модель может обращаться к открытым источникам в интернете или к информации, предоставленной пользователем, чтобы обогатить свой ответ.

Защита процесса обучения

> Процесс обучения защищён от потенциальных угроз, таких как «отравление» данных — внедрение вредоносной информации в обучающие данные, чтобы исказить логику работы алгоритмов.

На каждом этапе обучения применяются свои защитные механизмы:

- **Валидация.** Мы применяем строгие процедуры проверки и обработки входных данных.
- **Хранение.** Обученные модели находятся в специализированных защищённых хранилищах.
- **Развёртывание.** Внедрение моделей в инфраструктуру проходит в контролируемой среде с соблюдением принципа «нулевого доверия» (Zero Trust). Доступ к донастройке ИИ-модели закрыт по умолчанию и предоставляется лишь конкретным сотрудникам после строгой многофакторной верификации — и только на время выполнения задачи.

Безопасная эксплуатация и человеческий контроль

> У нас действует собственный стандарт по защите облачной инфраструктуры: вся информация о запросах пользователей и загружаемых файлах обрабатывается в закрытом техническом контуре компании. Яндекс использует для обучения моделей только обезличенную информацию.

Важнейшим элементом безопасности остаётся человеческий контроль. В критичных задачах, таких как генерация ответов на чувствительные темы, все новые модели проходят обязательное тестирование на безопасность как в процессе обучения, так и перед внедрением. Например, для тренировки нейросетей для ответов в сфере образования мы привлекаем преподавателей и педагогов.

Наш подход к безопасности ИИ

> Все технологии Яндекса, включая генеративные модели, создаются и развиваются с учётом принципов безопасной разработки. Мы интегрировали проверки безопасности в каждый этап жизненного цикла ИИ-системы — от сбора данных и обучения до внедрения в продукты.

Наш подход базируется на четырёх ключевых элементах

Аудит архитектуры сервиса с ИИ-моделями

Команда безопасности подключается к работе над сервисами с использованием генеративных ИИ-моделей Alice AI ещё на стадии идеи. Эксперты изучают будущую архитектуру сервиса, проверяют её на соответствие внутренним стандартам и отсутствие архитектурных уязвимостей. Это позволяет устранить фундаментальные риски до того, как начнётся разработка.

Также команда проводит оценку безопасности при интеграции ИИ-моделей в уже существующие сервисы — проверяют архитектуру и код сервиса для безопасного внедрения ИИ.

Соблюдение законодательства

Яндекс строго соблюдает требования законодательства РФ в области защиты информации и персональных данных, интегрируя правовые нормы в технические регламенты разработки ИИ.

Контроль качества и мониторинг

Мы следим за работой моделей и отслеживаем качество ответов по системе метрик и тестов. Любые ошибки и аномалии в работе нейросетей отслеживаются в двух режимах:

- Автоматически — через системы технического мониторинга.
- Через обратную связь — с помощью сигналов от пользователей и разметчиков данных.

Яндекс также регулярно проводит внутренние аудиты защиты сервисов, изучая новые способы атак на нейросети и разрабатывая меры противодействия.

Моделирование угроз

Мы адаптировали метод моделирования угроз под специфику ИИ-систем, чтобы просчитать возможные варианты характерных атак на нейросети. Это позволяет:

- Заранее обнаруживать специфические уязвимости, свойственные нейросетям.
- Прогнозировать возможные векторы нападения — например, атаки на обучающие данные или попытки обмануть ИИ-модель искажёнными промптами.
- Предотвращать инциденты ещё до их возникновения, тестируя модель в теоретических сценариях.

Принципы ИИ Яндекса

Мы развиваем искусственный интеллект и внедряем его в наши продукты, чтобы делать жизнь людей более удобной и давать им новые возможности. При этом мы следуем этическим принципам, которые сформулировали исходя из своего опыта создания и применения технологий. Эти принципы задают рамки безопасного развития ИИ, который будет приносить пользу людям, и объясняют работу наших ИИ-сервисов.

Принципы ИИ Яндекса

Польза и ответственность

Мы применяем искусственный интеллект только тогда, когда это приносит явную пользу. Перед внедрением ИИ мы оцениваем потенциальные риски и минимизируем возможные негативные последствия.

Предотвращение манипуляций

Мы разрабатываем и настраиваем ИИ-сервисы таким образом, чтобы исключить возможность их использования для манипуляции мнением пользователей.

Объективность и непредвзятость

Наши ИИ-сервисы не выражают чьё-либо мнение. Мы стремимся, чтобы их ответы были объективными и свободными от предвзятости.

Данные для обучения

Для обучения искусственного интеллекта можно и нужно использовать все общедоступные данные. При этом для отдельных сервисов мы донастраиваем алгоритмы на данных, наиболее подходящих для этих сервисов.

Принципы ИИ Яндекса

Прозрачность использования

Мы открыто сообщаем пользователям, когда они взаимодействуют с искусственным интеллектом. Мы разъясняем возможности и ограничения наших сервисов и маркируем созданный ИИ-контент, если это необходимо.

Чувствительные темы и независимая оценка

Принимая решения об использовании ИИ в сервисах, которые затрагивают чувствительные для людей темы, мы опираемся на мнение независимых экспертов и организаций, чтобы наши решения были максимально взвешенными.

Развитие через обучение и инновации

Мы постоянно совершенствуем ИИ-технологии, извлекая уроки из обратной связи и ошибок, и применяем полученный опыт для улучшения текущих и создания новых решений.

Сертификаты и аудиты

> В 2025 году Яндекс первым в России получил международный сертификат ISO 42001. Это первый глобальный стандарт, который регламентирует подходы к безопасному и ответственному созданию и использованию ИИ-моделей.

Мы развиваем искусственный интеллект и внедряем его в наши продукты, чтобы делать жизнь людей более удобной и давать им новые возможности. При этом мы следуем этическим принципам, которые сформулировали исходя из своего опыта создания и применения технологий. Эти принципы задают рамки безопасного развития ИИ, который будет приносить пользу людям, и объясняют работу наших ИИ-сервисов.

Для Яндекса сертификация по ISO 42001 — это независимое подтверждение того, что наши внутренние процессы управления разработкой и внедрения ИИ соответствуют лучшим мировым практикам.

Получение сертификата — результат аудита, который подтвердил, что процессы разработки и предоставления Alice AI соответствуют всем требованиям стандарта, в том числе в области этики и безопасности. Проверка охватила не отдельные элементы, а весь жизненный цикл технологии:

1. **Этап создания**

   От определения целей модели и принципов отбора обучающих данных до проектирования архитектуры.

2. **Внедрение**

   Процессы интеграции нейросети в сервисы и механизмы безопасных обновлений.

3. **Эксплуатация**

   Обеспечение защиты данных при работе с моделью, мониторинг инцидентов и работа с обратной связью от пользователей.

«Охота за ошибками»

«Охота за ошибками» — это программа Яндекса по премированию этичных хакеров, которая существует с 2012 года. В рамках программы мы приглашаем внешних исследователей проверять наши сервисы на прочность и платим вознаграждение за найденные уязвимости. «Охота за ошибками» позволяет усиливать безопасность и надёжность сервисов.



Направление генеративных сетей

В 2025 году Яндекс запустил направление программы, связанное с генеративными нейросетями. Оно охватывает технические уязвимости в семействе генеративных моделей Яндекса и сопутствующей инфраструктуре.

Технические уязвимости — это ошибки, способные привести к сбою модели, изменить её поведение, повлиять на работу смежных сервисов или раскрыть чувствительную информацию. К чувствительной информации относятся внутренняя конфигурация модели и данные обучения.

1 млн ₽

максимальная награда за найденную критическую уязвимость в генеративных ИИ. Размер награды зависит от серьёзности ошибки и простоты её воспроизведения.

ИИ-помощник «Охоты за ошибками»

> Чтобы ускорить работу с отчётами от исследователей, мы разработали собственного ИИ-помощника. Он анализирует базу всех ранее принятых и отклонённых отчётов.

Когда поступает новый отчёт, ИИ-помощник быстро находит похожие случаи и показывает их инженеру. Специалист сразу видит контекст, может быстро определить, зарегистрирована ли эта уязвимость, и принять решение о награде.

Защита пользователей: ИИ в сервисах

- Безопасность аккаунтов
- Защита общения
- Безопасность в интернете
- Безопасная коммерческая среда
- Защита от телефонного мошенничества
- Защита бизнеса

Безопасность аккаунтов

Аккаунт Яндекс ID — это ключ ко всей экосистеме Яндекса. В 2025 году мы усилили систему защиты данных и улучшили алгоритмы машинного обучения. Они анализируют паттерны поведения пользователя и контекст авторизации, блокируя аномальные запросы.

>110 млн

подозрительных попыток входа заблокировали в 2025 году

>11 млн

аккаунтов пользователей защитили от попыток взлома в 2025 году

Безопасность в интернете

Встроенные нейросети в Яндекс Браузере определяют признаки фишинговых страниц в реальном времени. Технология «Нейропротект» объединяет ИИ-технологии, которые проверяют подлинность сайтов, защищают онлайн-платежи, блокируют загрузку вредоносных файлов и нежелательные звонки.

9,1 млн

пользователей предупредил Браузер о переходе на опасную страницу в 2025 году

12,4 млн

скачиваний опасных файлов было заблокировано в 2025 году

Безопасность в интернете

В 2025 году ГК Softline провела исследование качества защиты от фишинга в современных браузерах. Специалисты протестировали работу встроенных защитных механизмов — их способность распознавать мошеннические сайты, которые используются для кражи личных данных.

Яндекс Браузер занял первое место в исследовании: он выявил в 8 раз больше фишинговых сайтов, чем другие браузеры. Из контрольной выборки в 100 опасных сайтов десктопная версия обнаружила 85, а мобильная — более 75 сайтов.

Защита от телефонного мошенничества

> Наш автоматический определитель номера использует нейросети, которые учитывают не только отзывы пользователей, но и свыше 300 поведенческих факторов, например массовые одновременные звонки с одного номера.

В 2025 году автоматический определитель номера обработал 1 млрд телефонных звонков с незнакомых номеров. Среди них 454 млн были классифицированы как нежелательные или мошеннические.

Среди факторов, которые анализируют нейросети, — время начала и завершения звонка, его продолжительность и информация о том, кто завершил вызов. Эти данные позволяют нейросетям делать выводы о качестве звонков. Например, если пользователи массово сбрасывают входящий вызов через несколько секунд, такие звонки маркируются нежелательными. При этом данные пользователей остаются в безопасности: алгоритмы работают исключительно с обезличенной информацией.

Защита от телефонного мошенничества

В 2025 году Яндекс первым в России внедрил новый механизм определения звонков в реальном времени на iOS. Он позволяет использовать собственную онлайн-базу Яндекса из 70 млн номеров, а не локальную, ограниченную 2 млн номеров. Благодаря этому точность определения спама и мошеннических звонков на iPhone достигла 93%.

59,7 млн
нежелательных звонков было автоматически заблокировано в 2025 году

На 40%
выросло качество определения мошеннических звонков с помощью ИИ

Почти в 7 раз
быстрее определитель оценивает новые номера

1,5 минуты
требуется для оценки нового номера

Защита общения

В 2025 году специалисты Яндекс 360 обновили технологии «Спамообороны», работающей на основе сложных алгоритмов и нейросетей. Благодаря этому защита пользователей Почты от фишинга и сложных атак стала ещё надёжнее.

Были добавлены функции распознавания поддельных логотипов (CV-антифишинг) и текста в изображениях (OCR). Новые технологии позволяют бороться с «картиночным спамом», когда мошенники заменяют текст в письме на изображение, и не дают маскировать вредоносные письма под официальные сообщения, например, от компаний и госслужб. Дополнительно «Спамооборона» считывает текст и ссылки не только с картинок, но и из вложенных PDF-документов, чтобы лишить спамеров возможности обмануть фильтры, спрятав ссылки внутри файлов.

93,6 млрд

писем обработала Яндекс Почта в 2025 году

≈ 4,9 млрд

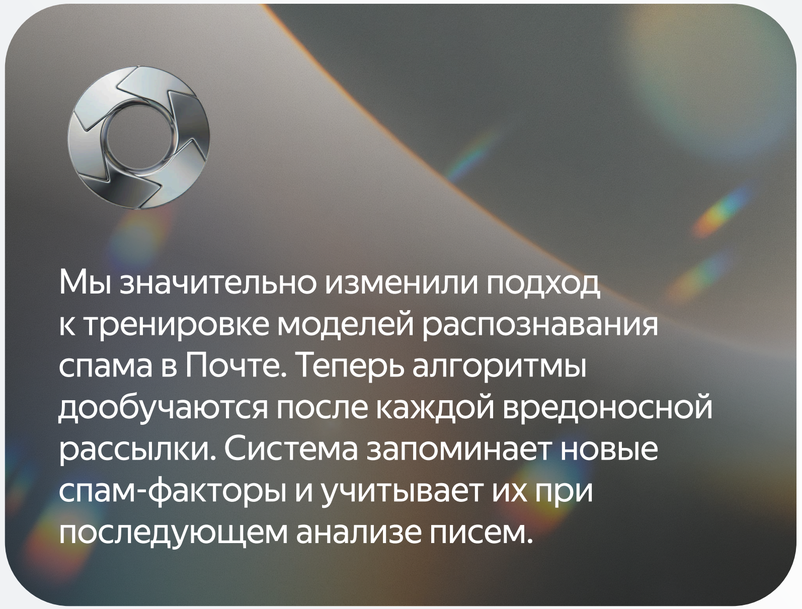
вредоносных писем заблокировали нейросети на этапе обработки

11,9 млрд

писем были отправлены в «Спам»

Защита общения

В 2025 году появился механизм повторной проверки писем ReCheck. Он позволяет выявлять потенциально опасные письма уже после доставки и нейтрализовать их. Если информация о вредоносности письма поступила с задержкой, ReCheck находит его у пользователя и автоматически изолирует — например, переносит в «Спам».



Безопасная коммерческая среда

Основа любых сервисов, связанных с финансами и онлайн-покупками, — это доверие пользователей. Мы используем ИИ-технологии, чтобы взаимодействие с бизнесом в экосистеме Яндекса было прозрачным и надёжным. Наши алгоритмы действуют на опережение, пресекая возможное мошенничество, и защищают от рисков как самих покупателей, так и добросовестных партнёров.

Реклама

В 2025 году Яндекс улучшил процесс поиска недобросовестных рекламодателей с помощью ИИ.

Алгоритмы научились работать на опережение и предугадывать поведение мошенников. Например, система распознаёт сценарии, когда рекламодатель-мошенник подменяет обычный сайт на фишинговый уже после прохождения модерации, и блокирует такие аккаунты до того, как сайт увидят пользователи.

Благодаря новым ИИ-технологиям для модерации рекламы мы в 10 раз сократили количество переходов пользователей по фишинговым объявлениям.

120 000+

аккаунтов недобросовестных рекламодателей заблокировано в 2025 году

Яндекс Маркет

> В 2025 году маркетплейс внедрил несколько ИИ-технологий для проверки логистических цепочек и контроля качества товаров. Нейросети помогают выявлять мошеннические схемы доставки и быстрее находить контрафакт.

В 2,3 раза

быстрее обрабатываются жалобы с помощью ИИ

В 1,5 раза

сократились затраты на модерацию

Яндекс Маркет

Антифрод-система для логистики

Алгоритм анализирует данные о перемещении товара со склада к покупателю и соотносит их с обычным сценарием движения. Если система обнаруживает отклонение от нормы, например смену маршрута или задержку доставки, она сообщает в службу безопасности. Таким образом ИИ помогает отслеживать признаки мошенничества и остановить подозрительные операции до того, как они нанесут финансовый ущерб покупателям или партнёрам.

Борьба с подделками

В 2025 году на Маркете была внедрена система обработки жалоб на базе машинного обучения. Она автоматически обрабатывает жалобы на возможный контрафакт и классифицирует их.

Алгоритмы распознают суть сообщений даже при сложной или некорректной формулировке, собирают информацию о продавце и передают отчёт специалистам контроля качества. Это ускоряет удаление некачественных товаров с витрины маркетплейса.

Защита бизнеса

Мы не только разрабатываем ИИ-технологии и внедряем их в свои сервисы, но и делимся этими решениями с рынком. Клиенты Yandex B2B Tech получают доступ к тем же инструментам для защиты своей критической инфраструктуры и чувствительных данных.



Yandex Cloud

Масштаб киберугроз непрерывно растёт. Ежедневно аналитики команды безопасности Yandex Cloud обрабатывают огромные массивы данных с помощью собственной SIEM-системы, которая собирает и анализирует данные, связанные с безопасностью ИТ-инфраструктуры.

Исследование наших экспертов показывает, что наиболее привлекательные цели для злоумышленников — ритейл (34%), ИТ-компании (31%) и промышленность (22%). Почти треть угроз связана с кражей учётных данных, фишингом и ошибками конфигурации инфраструктуры.

103 млрд

событий безопасности ежедневно обрабатывают аналитики Yandex Cloud

Yandex Cloud

> Yandex Cloud защищает клиентов с помощью комплекса собственных ИИ-сервисов безопасности.

на 86%

сократилось время обработки ложных срабатываний

Мультиагентная ИИ-система для SOC

Чтобы снизить нагрузку на своих аналитиков, Yandex Cloud внедрил мультиагентную систему на базе технологий Yandex AI Studio. Она помогает минимизировать влияние ложных срабатываний и значительно снизить количество рутинных задач у аналитиков центра мониторинга (SOC). Эта разработка доступна внешним компаниям внутри сервиса Yandex Cloud Detection and Response.

Модуль ML WAF

В сервисе для защиты от DDoS, веб-атак и ботов Smart Web Security появился модуль ML WAF. Это технология защиты на основе машинного обучения, которая охватывает более широкий спектр угроз и позволяет блокировать сложные веб-атаки, не покрываемые классическими сигнатурными методами.

Умный контроль данных в Security

Дополнительным слоем защиты стал сервис Security Deck с модулем контроля данных DSPM (Data Security Posture Management). С помощью ИИ-моделей он автоматически классифицирует данные клиента. Это даёт бизнесу точное понимание, где хранятся наиболее ценные данные, и позволяет выстроить их приоритетную защиту для предотвращения утечек.

Yandex Cloud

> В 2025 году эксперты Yandex B2B Tech выпустили руководство по снижению рисков при разработке и внедрении ИИ-агентов. В документе описаны модели угроз для больших языковых моделей (LLM) и рекомендации по созданию собственных мультиагентных систем.

Обучение сотрудников и пользователей

Чтобы защита всегда была на шаг впереди, мы формируем культуру безопасности — как внутри Яндекса, так и в обществе.

- Культура безопасности в компании
- Просветительские инициативы

Культура безопасности в компании

В Яндексе информационная безопасность — это не просто набор технологий, но и важнейший элемент корпоративной культуры. Мы осознаём, что надёжность цифровых решений напрямую зависит от компетенций людей, которые их создают. Поэтому мы выстраиваем систему непрерывного обучения и программу подготовки сотрудников.

Каждый новый сотрудник, независимо от должности, проходит обязательный вводный курс по информационной безопасности. Это обеспечивает единый стандарт «цифровой гигиены» во всей компании.

Для ключевых ролей — разработчиков, менеджеров продуктов, руководителей проектов — предусмотрено углублённое обучение по безопасности с учётом их профессиональных задач. Курсы построены на основе реальных сценариев, с которыми команды сталкивались на этапах создания архитектуры и запуска сервисов.

В 2025 году мы обновили образовательные программы для сотрудников, добавив блок, посвящённый технологиям искусственного интеллекта. В нём мы обучаем безопасному применению нейросетей и рассказываем, как мошенники используют ИИ в своих интересах и как от этого защититься.

Для разработчиков есть расширенный курс с разделом, посвящённым безопасной разработке ИИ.

Просветительские инициативы

Яндекс развивает образовательные проекты, посвящённые ИИ-инструментам, адаптируя сложные темы информационной безопасности для разных аудиторий — от первоклассников до людей старшего возраста.

«Урок цифры»

В 2025 году Яндекс провёл всероссийский урок на платформе «Урок цифры». Он был посвящён эволюции нейросетей и цифровой безопасности. Ученики 1–11 классов узнали, как развивались нейросети, какие технологии лежат в основе ИИ-ассистентов, как распознать дипфейки и как искусственный интеллект помогает человеку решать разные задачи — от генерации текстов до анализа изображений.

2,6 млн

школьников приняли участие в «Уроке цифры»

«Цифровой ликбез»

> В 2025 году Яндекс стал партнёром просветительского проекта «Цифровой ликбез». Эксперты компании в области искусственного интеллекта подготовили онлайн-урок о дипфейках и о том, как их вовремя распознать.

Урок ориентирован на детей от 6 лет и старше, но будет полезен и взрослым — например, родителям или педагогам, которые рассказывают детям о цифровой безопасности.

836 000

учеников по всей стране прошли урок к концу 2025 года

Безопасность для младших классов

> В 2025 году Яндекс обновил бесплатный курс «Безопасность в интернете» для школьников 1–4 классов. В программу добавлены два новых урока — «Что такое дипфейки» и «Как обманывают в онлайн-играх».

Они помогут детям узнать, как отличить поддельный контент, какие уловки мошенники используют в онлайн-играх и как защититься от возможных угроз. Материалы курса доступны на платформе Яндекс Учебника.

43,7 тыс.

детей прошли курс к концу 2025 года

Защита старшего поколения

> Яндекс совместно с проектом «Московское долголетие» провёл образовательный интенсив о кибербезопасности и нейросетях для людей старшего возраста. Лекции о цифровых угрозах и искусственном интеллекте прошли в 11 московских центрах.

Эксперты по кибербезопасности из Яндекса рассказали участникам о современных схемах мошенничества с использованием ИИ — от фишинга и социальной инженерии до дипфейков.

Заключение

В 2025 году ИИ-технологии стали фундаментом архитектуры безопасности Яндекса. В отчёте мы показали, как выстроена экосистема защиты Яндекса: от внедрения нейросетей для мониторинга инфраструктуры до развития цифровой грамотности пользователей.

Мы превратили сложные алгоритмы в простые и понятные функции для людей и бизнеса. Миллионы предотвращённых взломов аккаунтов, миллиарды отфильтрованных спам-писем, быстрая блокировка мошеннических звонков и этичная работа ассистентов — это реальный результат применения ИИ.

Киберугрозы продолжат эволюционировать: злоумышленники тоже активнее применяют новые технологии. Наша стратегия остаётся неизменной — действовать на опережение. Мы продолжим строить экосистему, где сервисы остаются удобными, а защита — незаметной и надёжной.