Координационный центр СПбГУПТД

Искусственный интеллект в киберберпространстве: оружие злоумышленника и щит правозащитника

AI

киберугрозы

29/10/2025

План лекции

/Основы концепции AI в киберпространстве

/Виды угроз

/ИИ в руках злоумышленника

/Инструменты защиты

ВВЕДЕНИЕ:

Основные понятия: ИИ, киберпространство

Ситуация сегодня: искусственный интеллект в киберпространстве

Реальный кейс:
Anthropic в своём августовском отчете «Threat Intelligence Report» описывает кампанию (кодовое имя GTG-2002), где злоумышленник использовал Claude / Claude Code для автоматизации разведки, похищения данных и подготовки вымогательских операций.

В результате
Пострадали по меньшей мере 17 организаций
Сумма выкупа доходила до $500 000

Отчёты описывают сценарии, где модель выполняла роль «ассистента» или «оркестратора» атак: подбор целей, приоритизация систем, подготовка команд/скриптов, выработка вымогательных текстов

https://www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pdf?

Что имеем сегодня: искусственный интеллект в киберпространстве

• **WormGPT / FraudGPT** — это «тёмные» LLM/сервисы, о которых писали исследователи ещё в 2023–2025 гг.; их назначение — облегчать фишинг/мошенничество и генерацию вредоносного содержимого.

• **PromptLock** — обнаружен/описан исследователями ESET как пример ранней «AI-powered» ransomware (proof-of-concept / work-in-progress в 2025). При этом некоторые разборы указывали, что часть подобных образцов — исследовательские демонстрации; важно разграничивать «доказанные активные кампании» и «proof-of-concept».

• **LameHug** — анализы/блоги (Picus, Splunk и др.) документируют infostealer/malware, интегрирующий LLM-вызовы для динамической генерации команд.

Ситуация сегодня

ИИ снижает порог входа для сложных атак — менее квалифицированные злоумышленники получают инструменты автоматизации, масштабирования и адаптации атак; при этом сами приёмы (фишинг, шифровальщики) часто остаются теми же, но реализованы более быстро и изощрённо

Статистика: рост киберугроз в 2025 году

Фишинг — масштабная массовая угроза
В 2024 году решения Kaspersky заблокировали около 893 миллионов фишинговых попыток — это рост на ~26% по сравнению с 2023 годом.

Социальная инженерия — ведущий вектор атак на бизнес
По данным Positive Technologies, в 2024 году социальная инженерия использовалась в 57% успешных атак на финансовые организации — что подтверждает роль фишинга/предтекстинга в компрометации критичных систем.

Оценки масштаба инцидентов в 2025
Частные отчёты указывают на десятки тысяч инцидентов (например, порядка **~63 000** инцидентов в 2025 в одной из выборок

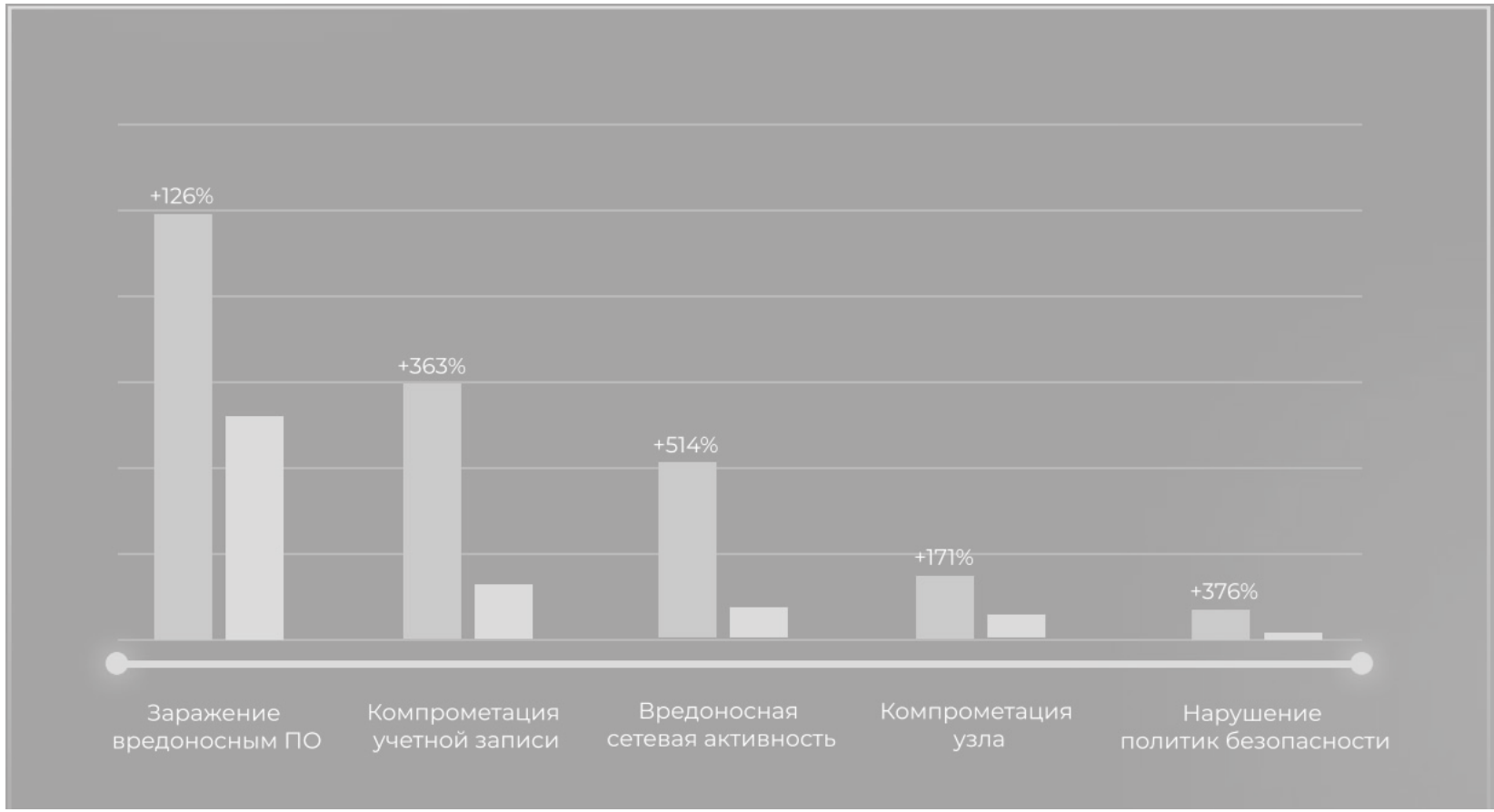
Статистика: рост киберугроз в 2025 году

Среднее количество атак на организацию выросло +47% в первом квартале 2025 года.

В России: зарегистрировано ~640 тыс. киберпреступлений за 2024 год, ущерб ~₽170 млрд.

По оценке СБЕР, убытки от кибермошенничества в России в 2025 году могут превысить $4,2 млрд.

72 % организаций заявили о росте кибер рисков. World Economic Forum



Основные понятия

КИБЕРПРОСТРАНСТВО

- Совокупность цифровой инфраструктуры: сети, серверы, устройства, сервисы и данные.
- Место взаимодействия пользователей, приложений и машин.
- Включает публичные интернет-ресурсы и приватные корпоративные сети.

Основные понятия

ИИ В КИБЕРПРОСТРАНСТВЕ

- ИИ — набор методов, позволяющих машинам выполнять задачи, требующие интеллекта (NLP, CV, принятие решений).
- Машинное обучение (ML) — подмножество ИИ; глубокое обучение (DL) — подмножество ML.

Примеры: модели, генерирующие текст (LLM), распознавание образов, аномалий.

Виды атак: обзор

Массовые: фишинг, массовый malware, спам.

Вредоносное ПО: ransomware, infostealers, бэкдоры.

Целевые (APT): длительные, сложные кампании против конкретных целей.

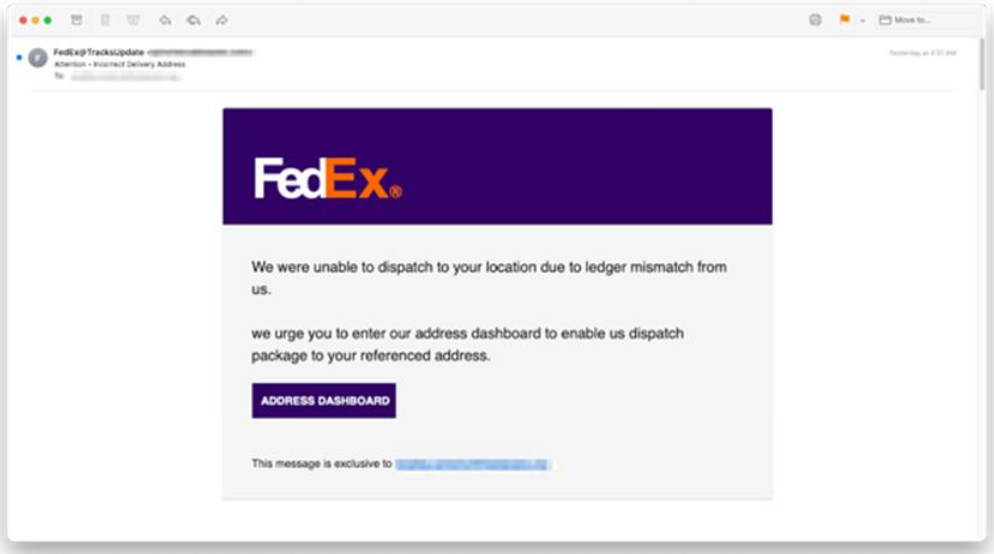
Сетевые атаки: DDoS, MITM, эксплойты.

Социальная инженерия: фишинг, vishing, предтекстинг.

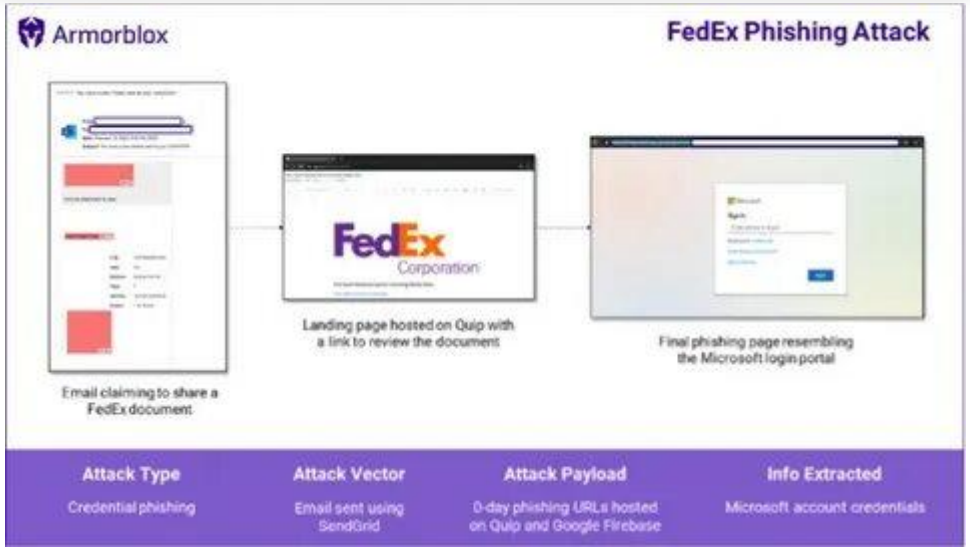
Supply-chain attacks: атаки через поставщиков/обновления.

Массовые атаки: фишинг

Массовые: фишинг, массовый malware, спам.
Цель: большое количество случайных жертв



Пример: Массовая фишинговая рассылка с поддельными уведомлениями от банков или почтовых служб, целью которой является кража учетных данных пользователей.



Конкретный случай: была зафиксирована массовая рассылка фишинговых писем, имитирующих сообщения от службы доставки FedEx, с целью установки вредоносного ПО на компьютеры пользователей.

Вредоносное ПО

Вредоносное ПО: ransomware, infostealers, бэкдоры (backdoor).

Пример: Атака с использованием шифровальщика WannaCry, который зашифровал данные на сотнях тысяч компьютеров по всему миру и требовал выкуп в биткоинах.

Конкретный случай: Шифровальщик NotPetya, замаскированный обновление бухгалтерского ПО, парализовал работу многих компаний, нанеся ущерб на миллиарды долларов.

Целевые (APT)

Целевые (APT): длительные, сложные кампании против конкретных целей.

Advanced Persistent Threat = длительная, целенаправленная кампания

Мотивы: шпионаж, саботаж, компрометация, экономическая выгода

Модель: разведка → проникновение → закрепление → сбор → эксфильтрация

Конкретный случай: APT32 (OceanLotus), действующая из Вьетнама, атаковала иностранные компании, чтобы получить коммерческую информацию и доступ к внутренним системам.

Данные исследования positive security: https://www.ptsecurity.com/research/analytics/apt-attacks-government-2019/#id1

Искусственный интеллект в киберберпространстве: оружие злоумышленника и щит правозащитника

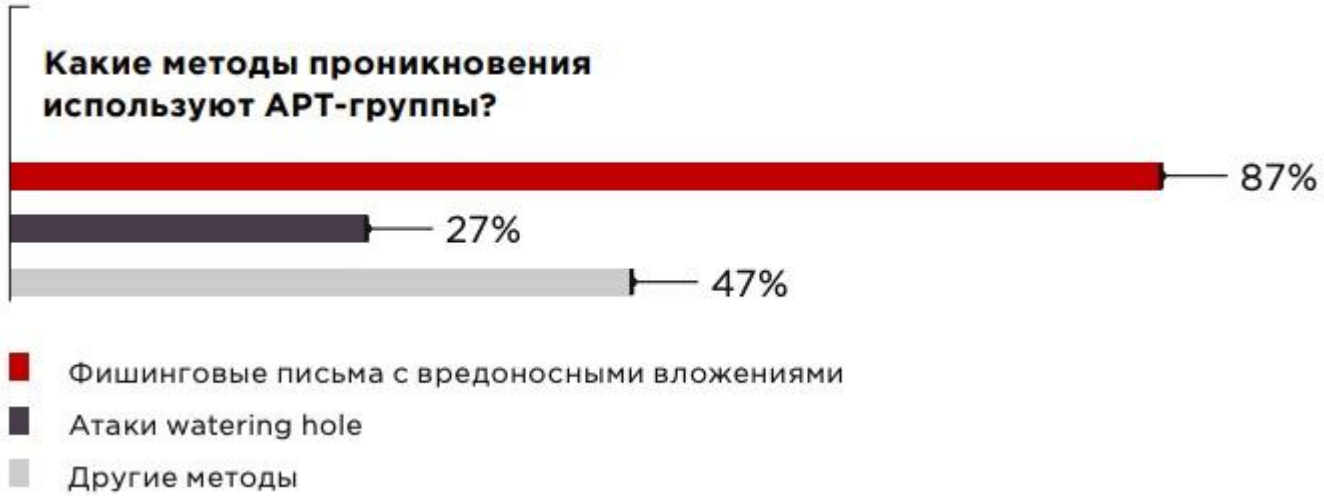
Целевые (APT)

87% APT-группировок начинают атаки на госучреждения с целенаправленного фишинга

каждый третий получатель фишингового письма запускает вредоносный файл

Данные исследования positive security: https://www.ptsecurity.com/research/analytics/apt-attacks-government-2019/#id1

Целевые (APT): кроме фишинга



Drive-by-compromise

Это техника атаки, при которой вредоносные программы незаметно загружаются на компьютер жертвы при посещении скомпрометированных ресурсов. Типовая схема атаки: сотрудник попадает на зараженный сайт, откуда вредоносный скрипт перенаправляет его на сервер злоумышленника.

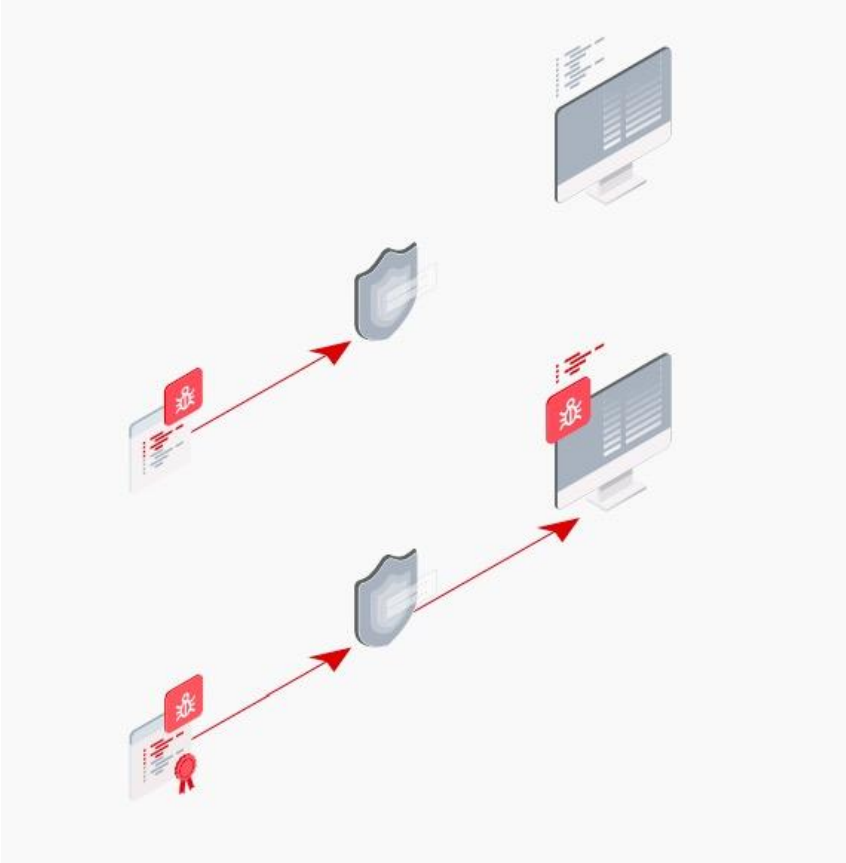
27% группировок применяли эту технику в отдельных кампаниях

Данные исследования positive security: https://www.ptsecurity.com/research/analytics/apt-attacks-government-2019/#id1

Почему антивирус может не сработать против APT

Code signing

Средства защиты лояльно относятся к файлам с цифровыми подписями, считая их доверенными. Подписав зловред, злоумышленники рассчитывают обмануть механизмы безопасности. Сертификаты для таких целей продаются в дарквебе. Они могут попадать в руки злоумышленников в результате атак supply chain. Так, злоумышленники взломали сервер ASUS и распространяли под видом обновлений вредоносное ПО, подписанное сертификатом компании.



Две трети (67%) APT-группировок удаляют вредоносные файлы, которыми больше не пользуются. Например, атакующая российские правительственные ресурсы группировка Cloud Atlas использует вредонос VBShower, который умеет удалять временные файлы, свидетельствующие о присутствии злоумышленников.

Данные исследования positive security: https://www.ptsecurity.com/research/analytics/apt-attacks-government-2019/#id1

Техники закрепления в инфраструктуре

93%
кодируют и шифруют вредоносный код
Obfuscated Files or Information

67%
используют бестелесный вредоносный код
Process Injection

40%
проверяют наличие песочницы
Virtualization/Sandbox Evasion

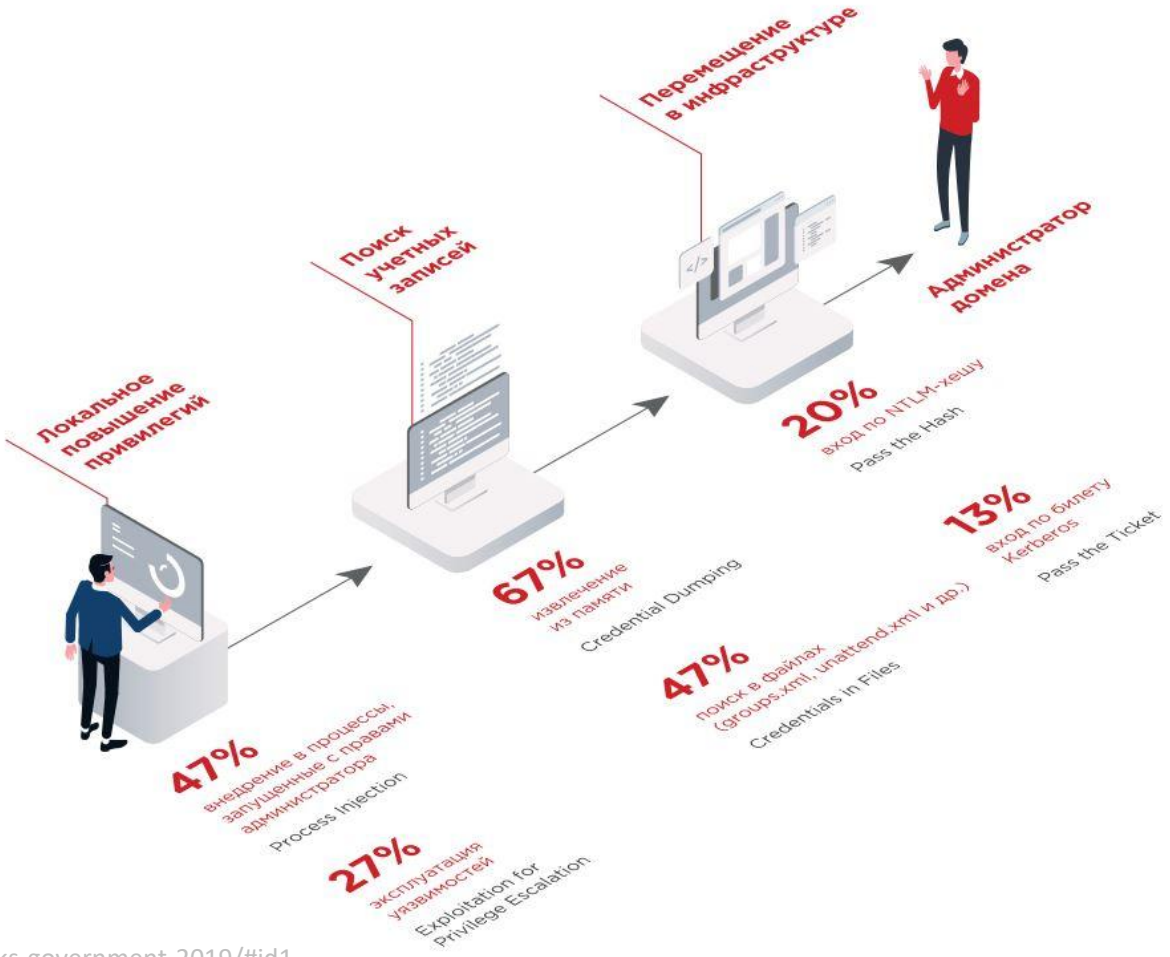
27%
подписывают вредоносные файлы цифровой подписью
Code Signing

Данные исследования positive security: https://www.ptsecurity.com/research/analytics/apt-attacks-government-2019/#id1

Искусственный интеллект в киберберпространстве: оружие злоумышленника и щит правозащитника

Какая информация может заинтересовать злоумышленников

- Персональные данные сотрудников и других граждан
- Сведения в области внешней политики и экономики
- Научно-исследовательские и проектные работы
- Финансовая отчетность

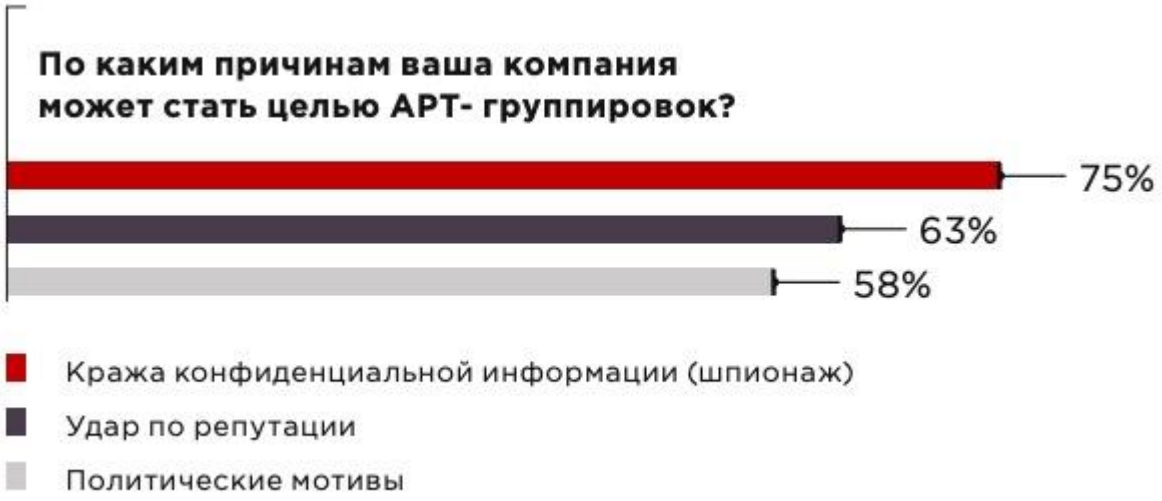


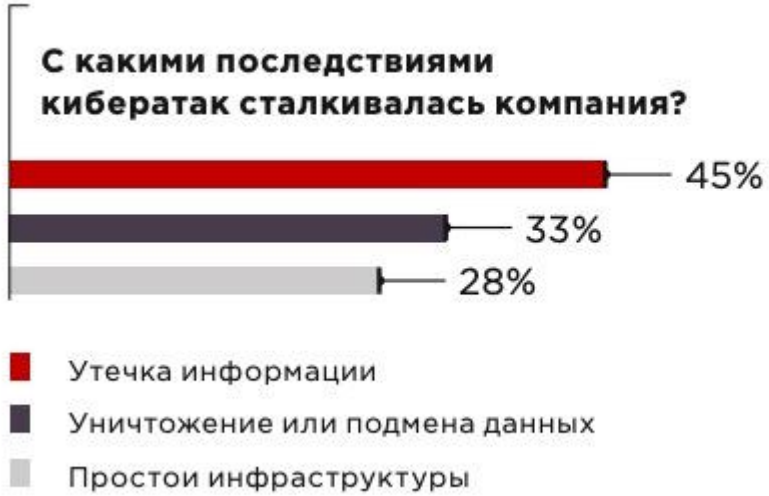
60%
APT-группировок, атакующих государственные организации, используют PowerShell

Что такое PowerShell
PowerShell — это инструмент администрирования и среда выполнения сценариев для ОС Windows. Может использоваться взломщиками на всех этапах атаки.

Данные исследования positive security: https://www.ptsecurity.com/research/analytics/apt-attacks-government-2019/#id1

Госучреждения не готовы к APT – атакам



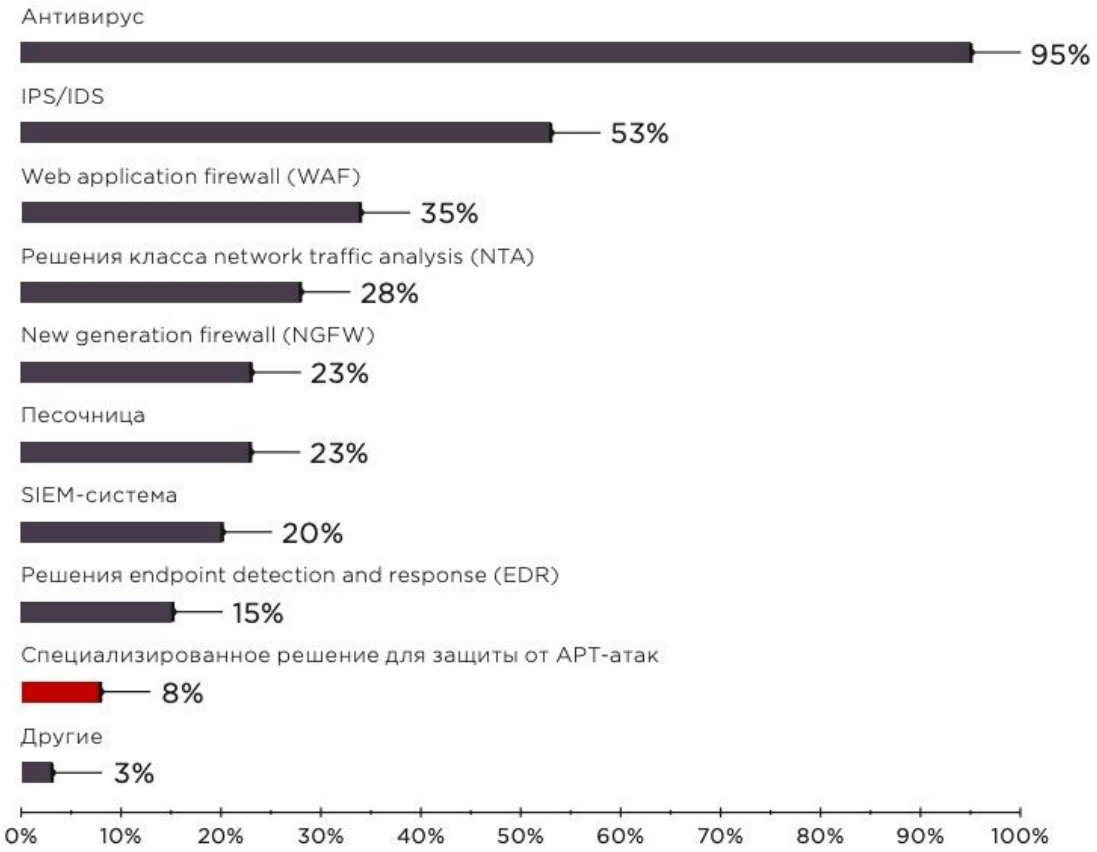


Данные исследования positive security: https://www.ptsecurity.com/research/analytics/apt-attacks-government-2019/#id1

68%
APT-группировок, нацеленных на Россию, атакуют государственные учреждения

Искусственный интеллект в киберберпространстве: оружие злоумышленника и щит правозащитника

Используемые средства защиты



Данные исследования positive security: https://www.ptsecurity.com/research/analytics/apt-attacks-government-2019/#id1

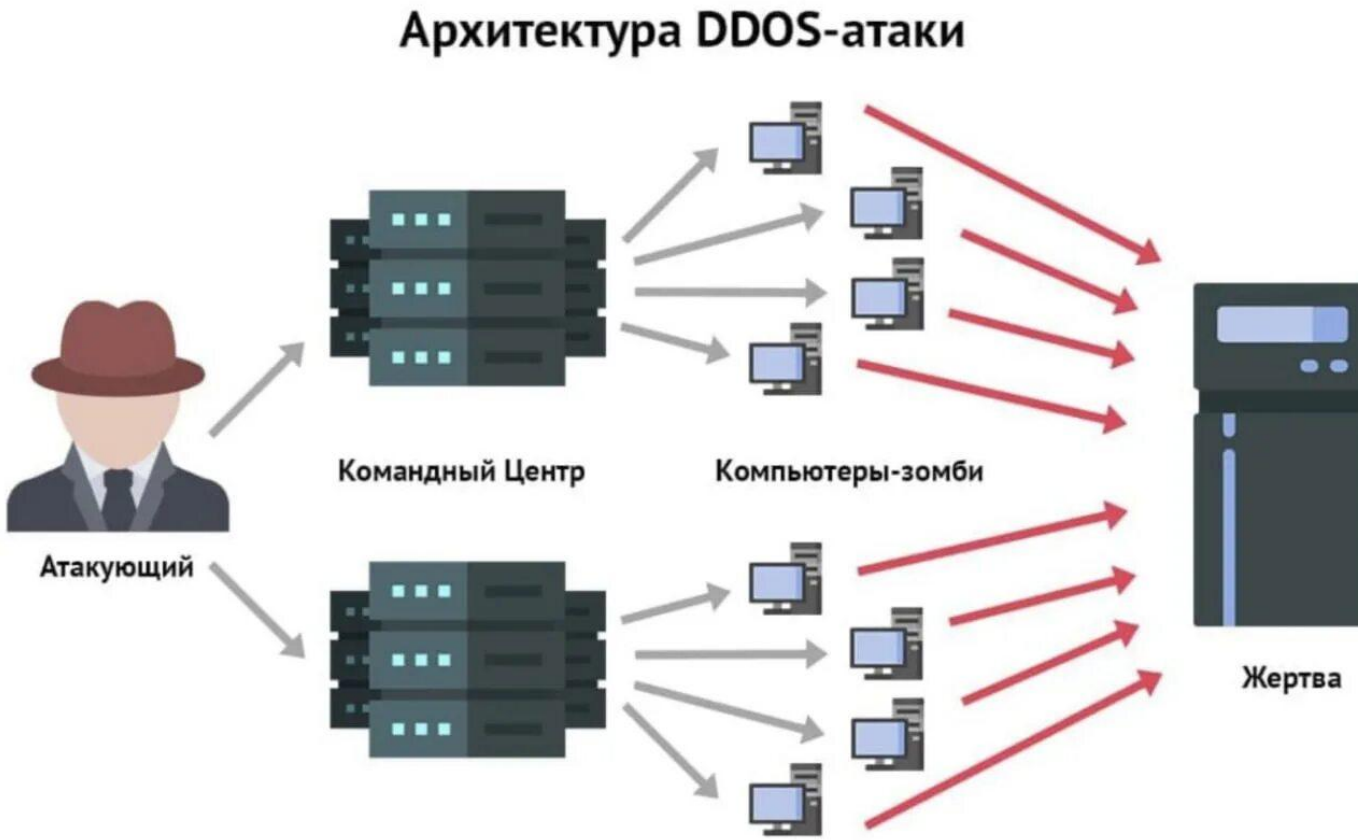
Искусственный интеллект в киберберпространстве: оружие злоумышленника и щит правозащитника

Сетевые атаки и эксплуатация уязвимостей (Ddos)

DDoS на GitHub (2018)— крупнейшая зафиксированная атака (1,35 Тбит/с)
— использован уязвимый сервис memcached

Heartbleed (2014)— уязвимость в OpenSSL
— позволяла похищать ключи и данные из памяти серверов, затронула миллионы сайтов, в том числе банковские и правительственные

DDoS-атака (Distributed Denial of Service) — это онлайн-атака на системы компании, при которой злоумышленники посылают огромное число запросов



Динамика Ddos-атак 2025

За 9 месяцев 2025 г. — в среднем 1,2 тыс. DDoS-атак на компанию

Снижение на 32% по сравнению с 2024 г.

Всего зафиксировано более 560 тыс. атак

Наибольшая активность: телеком, ИТ, финансы, госсектор

Средняя мощность — до 1,6 Гбит/сек, максимум — 1522 Гбит/сек

География атак

Москва — 282 тыс. атак

Приволжский округ — 44,8 тыс.

Урал — 41,3 тыс.

Южный округ — 26,3 тыс.

Рост атак на Северо-Западе (+10%)

Эволюция атакующих стратегий

DDoS теряет эффективность → атаки становятся прицельными

Комбинирование с веб-атаками и APT-группировками

Цель — проникновение и разрушение инфраструктуры

Рост зрелости ИБ-компаний снижает «массовость» атак

Социальная инженерия (deep fake)

Дипфейк (англ. deepfake) — технология синтеза медиа-контента с помощью алгоритмов искусственного интеллекта, чаще всего нейросетей. Позволяет создавать реалистичные фальшивые видео, изображения и аудиозаписи, в которых человек говорит или делает то, чего на самом деле никогда не происходило.

Deepfake: видео/аудио/изображения, сгенерированные нейросетями

Типы: замена лиц, аудио-дипфейки, фотореалистичные портреты

Практическая опасность: мошенничество, фейковые заявления руководителей, компрометация репутации

Социальная инженерия (deep fake)

Примеры: имитация голоса, изображений коллег/родственников, знакомых
Дипфейки используются в мошенничестве, шантаже, дезинформации
Каналы распространения: мессенджеры, соцсети, звонки, видеозвонки

Статистика по России

52% россиян знакомы с термином «дипфейк», 48% — нет

Почти 1/3 сотрудников крупных компаний сталкивались с мошенничествами с дипфейками

Форматы столкновений: развлекательный контент (64%), реклама (27%)

Подмена знакомых — 19%, подмена руководителей/коллег — 18%

Частые каналы: Telegram, WhatsApp, соцсети

Как распознавать дипфейки: практическая диагностика

Визуальные артефакты: неестественное моргание, искажения волос, неровные тени

Аудио-признаки: странные паузы, одинаковая интонация, шум/артефакты в конце фраз

Контекст и верификация: проверьте факт через независимый канал (корп. почта/линейный менеджер)

Технические проверки: обратный поиск изображения, метаданные, проверка цифровых подписей/маркировки

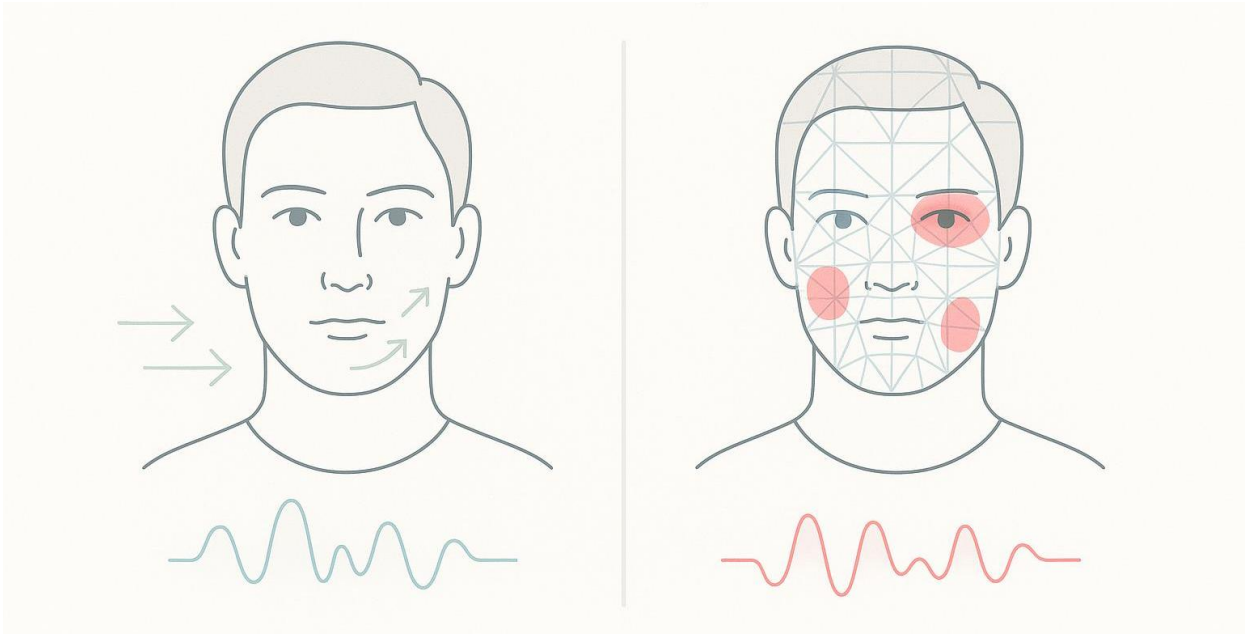
Как ИИ распознает deepfake

ИИ способен анализировать мельчайшие несоответствия: мимику, тени, частоты звука

Модели-детекторы обучаются на миллионах пар «реальное/синтетическое»

Методы: анализ микровыражений, ритма моргания, спектр голоса, следы GAN-генерации

Используются сверточные нейросети (CNN), трансформеры и модели спектрального анализа



ИИ-детекторы

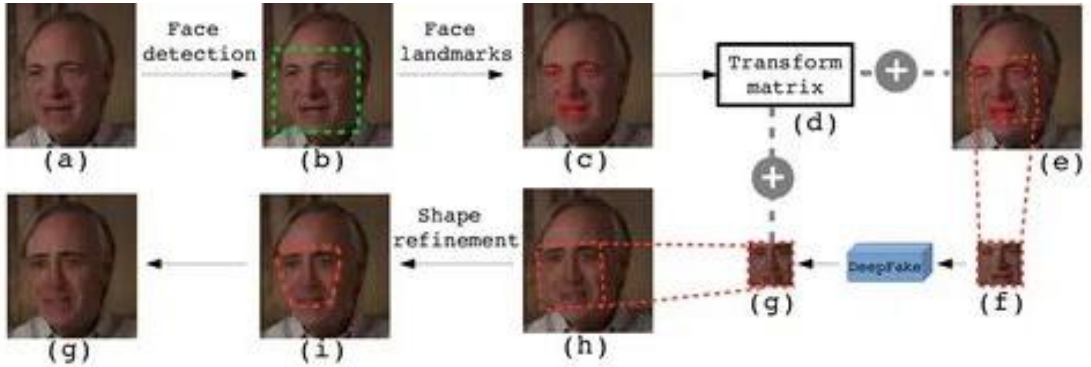
Пиксельный и частотный анализ (ELA, спектр Фурье)

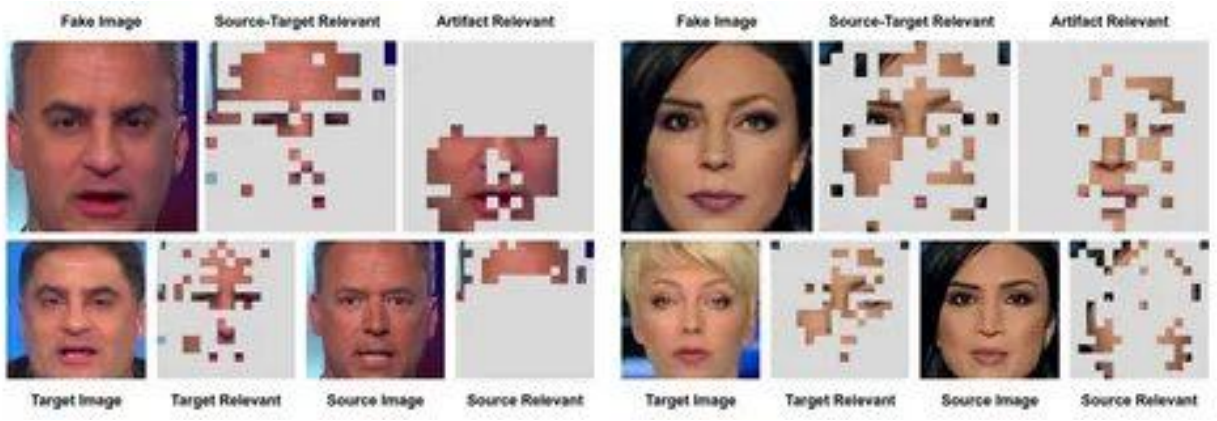
Optical flow и микросдвиги лица

Нейроморфные детекторы (Spiking Neural Networks)

Проверка метаданных и Camera Fingerprint

Комбинация анализа изображений, видео, аудио и контекста





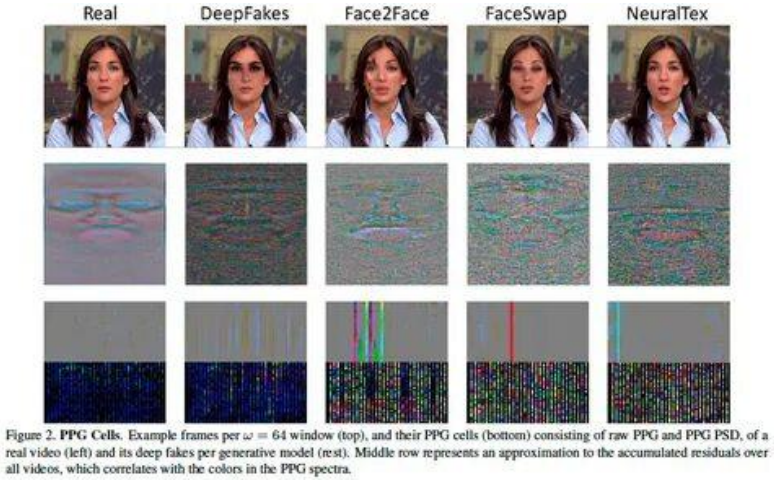


Figure 2. **PPG Cells.** Example frames per $\omega = 64$ window (top), and their PPG cells (bottom) consisting of raw PPG and PPG PSD, of a real video (left) and its deep fakes per generative model (rest). Middle row represents an approximation to the accumulated residuals over all videos, which correlates with the colors in the PPG spectra.

ИИ-детекторы

ИИ-детекторы первого поколения

Обучены на FaceForensics++ и DFDC

Ищут статистические артефакты (JPEG, warping, кожа)

Выдают вероятность фейка (ensemble-модели)

ИИ-детекторы нового поколения

Трансформеры и state-space модели

Анализ видео + аудио + метаданных

Примеры: DeepFake-MAMBA, X-Detector-T5

Проверка согласованности между модальностями

Платформы и сервисы

Reality Defender, Sensity AI, Deepware Scanner

FaceOnLive, Intel FakeCatcher, Microsoft Video AI Authenticator

WeVerify (InVID), AI or Not, DeepFake-o-meter

Resemble AI (аудио), Hume AI, DeepfakeDetect.A

IMobile: FakeShield, DeepCheck Mobile, TruthLens

Криптография и защита контента «с рождения»

C2PA и Content Credentials

SynthID (Google), CAI (Adobe, NYT, Twitter)

Аппаратные водяные знаки и шумовые паттерны

Проверка подлинности на уровне устройства

Итого: как ИИ используется во всех типах атак

Вид атаки	Как ИИ используется злоумышленниками
Массовые (фишинг, спам, malware)	Генерация фишинговых писем, адаптивные тексты, создание реалистичных доменов
Вредоносное ПО (ransomware, infostealers)	Автоматический подбор уязвимостей, уклонение от антивирусов
Целевые (APT)	Анализ инфраструктуры жертвы, социальное профилирование, выбор слабых звеньев
Сетевые атаки (DDoS, MITM, эксплойты)	Оптимизация маршрутизации бот-сетей, автоматический поиск уязвимостей
Социальная инженерия (включая deepfake)	Имитация личности, подделка голоса/видео, адаптивное убеждение

ИИ как инструмент защиты

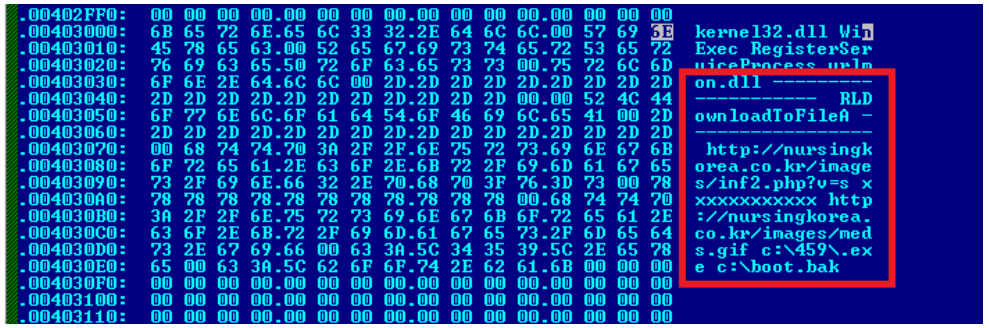
Вид защиты	Пояснение
ML-based Threat Detection	Как ИИ выявляет неизвестные вирусы и эксплойты по аномалиям, а не сигнатурам
Поведенческая аналитика (UBA)	Анализ действий пользователей и устройств
Антифишинг и контент-анализ	Модели NLP выявляют фишинг и поддельные письма
Автоматизация реагирования (SOAR)	ИИ сам закрывает инциденты, блокирует сессии, уведомляет SOC
Threat Intelligence с ИИ	ИИ агрегирует и анализирует данные о киберугрозах с глобальных источников
Применение ИИ в цифровой криминалистике	Восстановление цепочки атаки, корреляция журналов, идентификация атакующих

ML-based Threat Detection — ИИ против неизвестных вирусов

Сигнатуры — заранее известные "отпечатки" вирусов.
Аномалии — отклонения от нормальной работы системы.

ИИ анализирует поведение программ, выявляя вредоносные, даже если они новые.

Пример: Microsoft Defender, CrowdStrike Falcon.



Поведенческая аналитика (UBA)

User and Entity Behavior Analytics

ИИ запоминает, как обычно ведёт себя сотрудник/пользователь

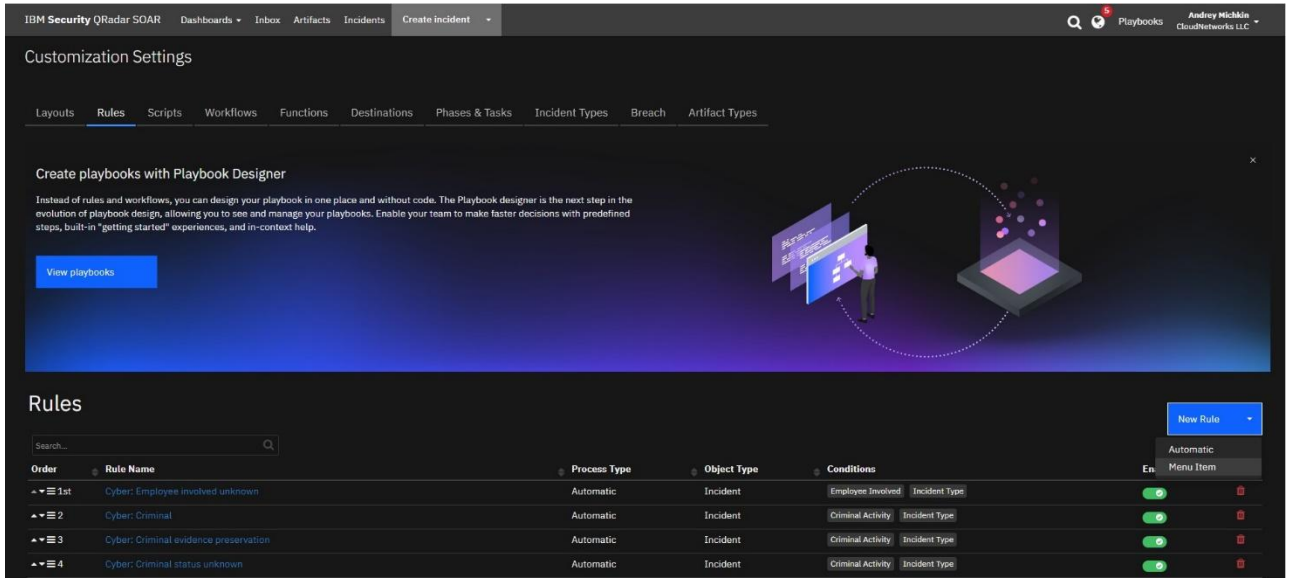
Пример: Splunk, Azure Sentinel

SOAR — Автоматизация реагирования

Security Orchestration, Automation and Response

ИИ автоматически блокирует учётные записи, IP и процессы.
Уведомляет SOC (центр безопасности).

Пример: IBM QRadar SOAR, Cortex XSOAR.



Threat Intelligence с ИИ

ИИ собирает и анализирует данные о глобальных киберугрозах

Анализ даркнета, форумов, логóв, соцсетей.

Выявление кампаний, инфраструктуры атакующих.

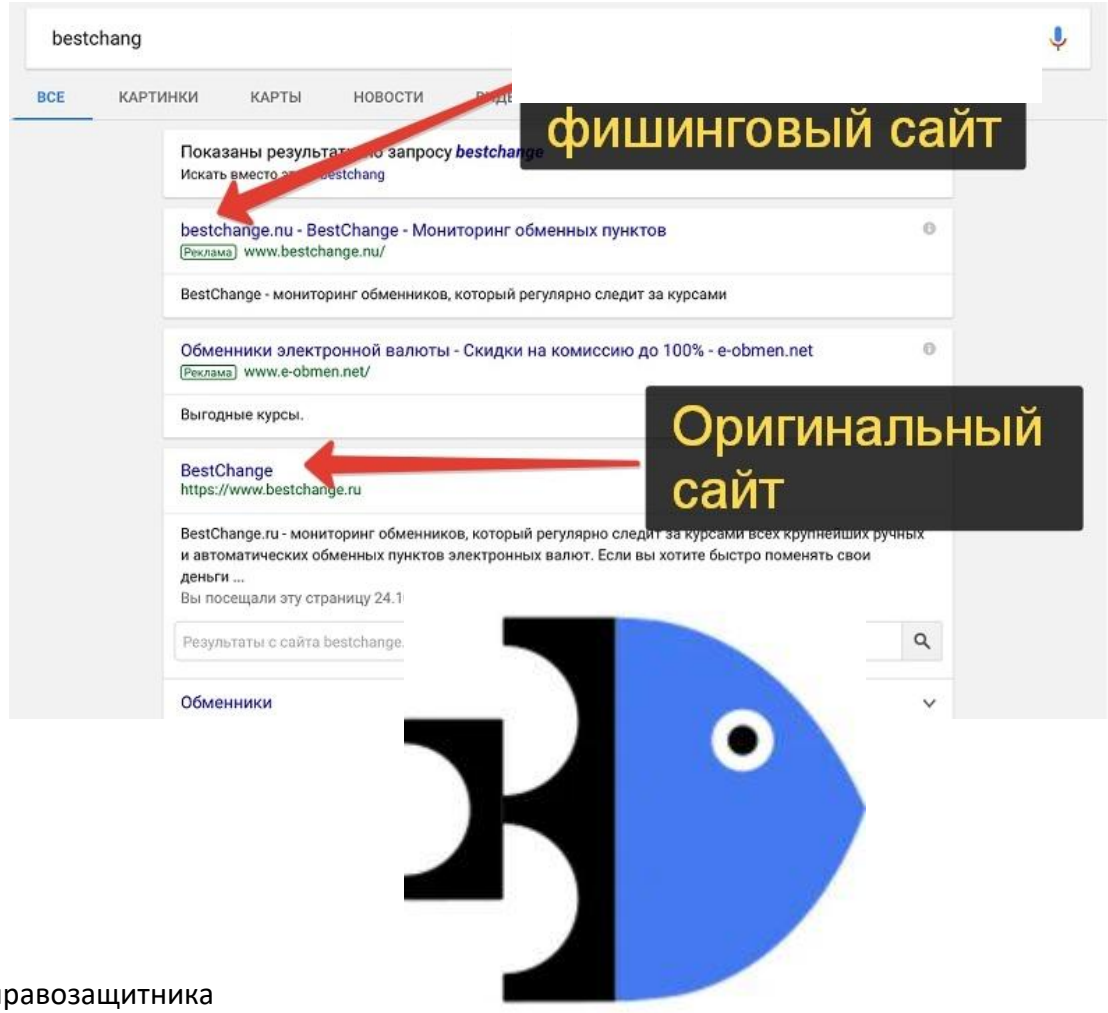
Пример: IBM X-Force, Mandiant.

Антифишинг и контент-анализ

Анализ текста с помощью NLP (Natural Language Processing)

Поиск подозрительных шаблонов, ссылок, манипулятивного стиля.
Распознаёт аудио- и видеодипфейки.

Пример: Google Safe Browsing, CheckPhish.



Сентимент-анализ и NLP

Обработка естественного языка (NLP (Natural Language Processing)) — это область искусственного интеллекта, которая занимается взаимодействием между компьютерами и человеческим языком. Она включает в себя анализ, понимание и генерацию текстов на естественном языке.

ИИ распознаёт фишинг по смыслу, стилю и структуре текста

- ИИ анализирует лексику, тон, эмоции, структуру предложений.
- Сравнивает с базой фишинговых паттернов.
- Реагирует на подозрительные просьбы и манипуляции.

основные задачи ИИ в защите

Что делает ИИ?	Как это работает?
Обнаружение неизвестных угроз	ML-модели анализируют поведение систем и выявляют аномалии без сигнатур.
Распознавание фишинга	NLP и CV-модели анализируют текст, структуру сайтов и логотипы, выявляют подделки.
Поведенческая аналитика пользователей	ИИ отслеживает действия пользователей и устройств, замечает необычную активность.
Автоматизация реагирования (SOAR)	ИИ сам блокирует угрозы, изолирует заражённые узлы и уведомляет операторов.
Анализ угроз (Threat Intelligence)	ИИ агрегирует данные из даркнета и форумов, строит карту атакующих кампаний.
Этический контроль ИИ	Внедрение explainable AI (объяснимого ИИ), чтобы снизить риск ложных срабатываний и утечек данных.

Этические и практические вызовы

Ложные срабатывания: системы распознавания лиц и анализа поведения могут ошибочно идентифицировать человека или неверно интерпретировать действия.

Приватность и доверие: сбор и хранение биометрических и поведенческих данных требует строгой защиты.

Асимметрия технологий: ИИ, применяемый для защиты, может использоваться и для манипуляций — например, в создании фейков, подмене лиц или распространении провокационного контента.

ИИ — не замена экспертам, а их «помощник», который берет на себя рутину и обработку big data.

Кафедра информационных технологий

Искусственный интеллект в киберберпространстве: оружие злоумышленника и щит правозащитника

AI

киберугрозы

29/10/2025