ОБ УГРОЗАХ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ РОССИЙСКОЙ ФЕДЕРАЦИИ В СФЕРЕ СОЗДАНИЯ И ПРИМЕНЕНИЯ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

ДАЛИ Ф.А.

Академия криптографии Российской Федерации

ОБЩИЕ ПОНЯТИЯ

Общее понимание ИИ

Имитация когнитивных функции человека

В научном смысле ИИ

Аппаратное или программное **моделирование**

В техническом смысле ИИ

Программно-аппаратный **продукт** с конкретным набором функций

Информационная система с ТИИ

параметры системы настраиваются с помощью моделей ИИ - программ, прошедших «обучение»

01

управление и обеспечение информационной безопасности

02

оценка рисков, возникающих при применении ТИИ

03

определение уровня качества результатов применения ТИИ

04

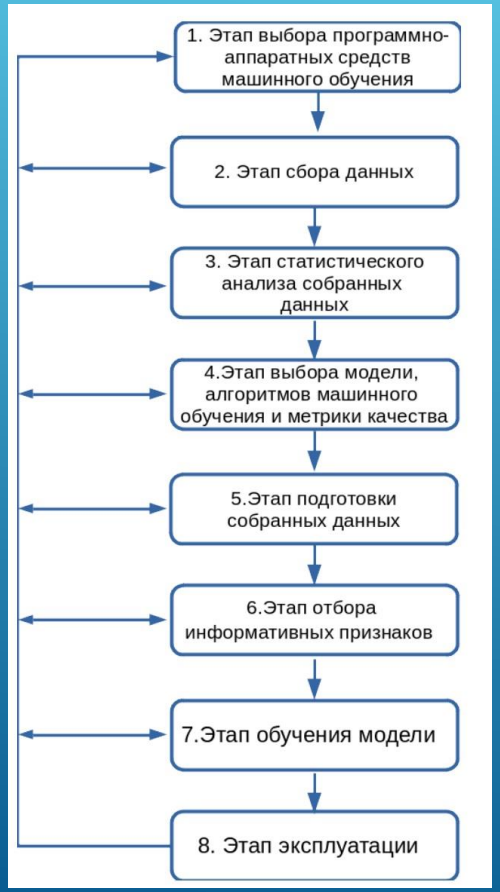
этические аспекты применения ТИИ

ФОРМАЛИЗАЦИЯ ОБЪЕКТА ЗАЩИТЫ

- Нейросетевые алгоритмы машинного обучения, предназначенные для решения одной или нескольких статистических задач (средство ИИ)

- Основные проблемы:
  - алгоритм решения задачи формируется одновременно с получением результата ее решения (обучение ИИ)
  - возможность реализации «логических» атак через содержание обрабатываемых данных

ЭТАПЫ ЖИЗНЕННОГО ЦИКЛА МОДЕЛИ ИИ



Влияние на данные

Влияние на обучение

Анализ предсказаний модели ИИ

Противоправные действия при

01 сборе и преобразовании данных

02 доступе к данным, что позволяет извлечь из них чувствительную информацию

03 настройке процесса обучения позволяет повлиять на качество модели в целом

04 извлечении информации из модели в процессе ее эксплуатации

УГРОЗЫ ИБ ТИИ И ЦЕЛИ НАРУШИТЕЛЯ

Цели нарушителя

Целостность моделей и данных

Доступность моделей и данных

попытки повлиять на предсказания модели в своих интересах

Конфиденциальность моделей и данных

попытки получить оберегаемую информацию либо о самой модели, либо о данных, на которых она обучалась

Угрозы ИБ

- обучение нежелательному поведению (модификация модели машинного обучения путем искажения (отравления) данных
- хищение или подмена обучающих данных
- введение модели в заблуждение (состязательные атаки)
- дрейф данных (непредсказуемое бесконечное преобразование данных)
- выбросы данных (нахождение отдельных данных далеко за пределами большинства точек данных в заданном наборе
- цифровая подмена фото- и видеоизображений человека с наложением аудиозаписей (дипфэйки)

АТАКИ, РЕАЛИЗУЮЩИЕ УГРОЗЫ ЦЕЛОСТНОСТИ МОДЕЛЕЙ И ДАННЫХ

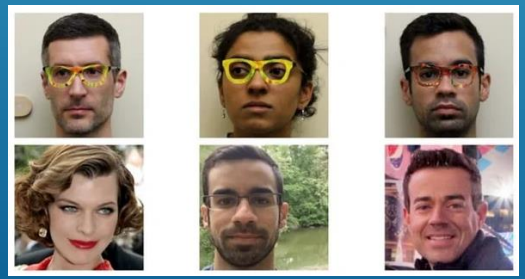
Векторы атак

«отравление»

на запрос «гвозди», сеть Кандинский 2.0 показывает наличие внутренней трансляции запроса на английский язык

«уклонение»

после наложения на изображение шума, образ классифицируется неверно, ниже пример использования оправы очков, чтобы выдать себя за другого человека



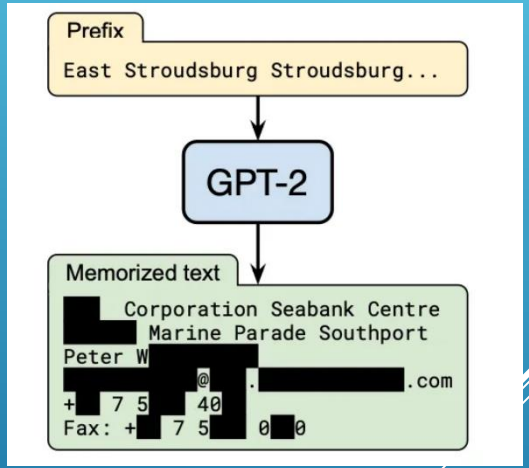
«дипфейки»

прогресс качества синтетических изображений несуществующих людей по годам

АТАКИ, РЕАЛИЗУЮЩИЕ УГРОЗЫ КОНФИДЕНЦИАЛЬНОСТИ МОДЕЛЕЙ И ДАННЫХ

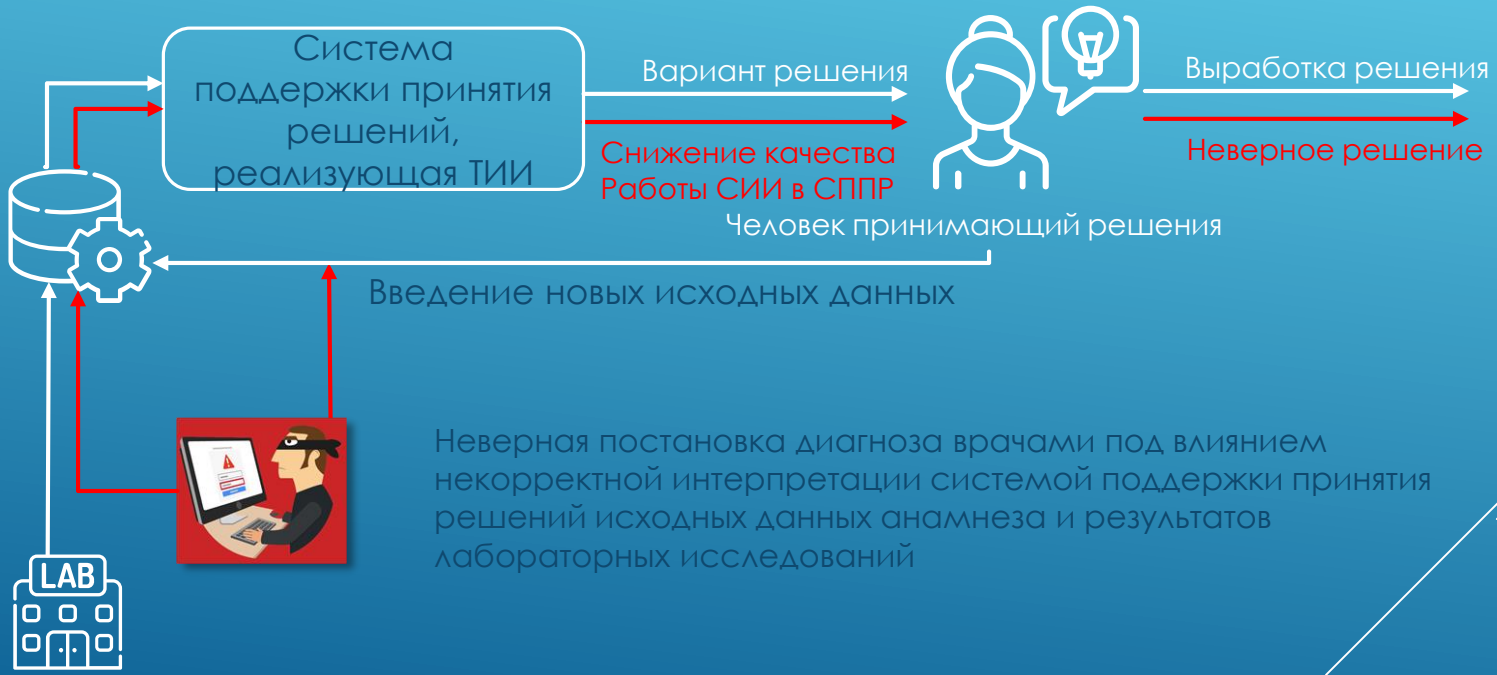
Обученная модель ИИ является контейнером, содержащим в сжатом, архивированном виде данные, использовавшиеся при ее обучении

Слева в каждом блоке – исходные изображения, справа – синтезированные из выходного вектора атакуемой нейронной сети VGGfacenet с помощью генеративно-состязательной сети



Пример извлечения информации из большой языковой модели GPT-2

АТАКИ, РЕАЛИЗУЮЩИЕ УГРОЗЫ ДОСТУПНОСТИ МОДЕЛЕЙ И ДАННЫХ



Неверная постановка диагноза врачами под влиянием некорректной интерпретации системой поддержки принятия решений исходных данных анамнеза и результатов лабораторных исследований

ВНЕДРЕНИЕМ ТИИ В РОССИЙСКОЙ ФЕДЕРАЦИИ

В настоящее время внедрение и использование ТИИ в Российской Федерации определяется следующими документами

Указ Президента Российской Федерации от 10.10.2019 № 490 (обновлена 15.02.2024 № 124)	«Национальная стратегия развития искусственного интеллекта на период до 2030 года»
Распоряжение Правительства Российской Федерации от 19.08.2020 № 2129-р	«Концепция развития регулирования в сфере технологий искусственного интеллекта и робототехники до 2024 года»
Распоряжение Правительства Российской Федерации от 05.10.2023 № 2715-р	«Дорожная карта» развития высокотехнологичного направления «Искусственный интеллект» на период до 2030 года»
Указ Президента Российской Федерации от 21.07.2020 № 474	Федеральный проект «Искусственный интеллект» национальной программы «Цифровая экономика Российской Федерации»
Утвержден Президентом Российской Федерации 29.01.2023 № Пр-172	Отдельные поручения Президента Российской Федерации и Правительства Российской Федерации
Внутренние поручения	Отдельные планы цифровой трансформации министерств, ведомств, государственных корпораций

Приказом Росстандарта создан Технический комитет по стандартизации «Искусственный интеллект» (ТК 164)

ОТДЕЛЬНЫЕ МЕРОПРИЯТИЯ, НАПРАВЛЕННЫЕ НА ОБЕСПЕЧЕНИЕ БЕЗОПАСНОСТИ ПРИМЕНЕНИЯ ТИИ

01

Федеральный проект

- Разработка научной базы противодействия новым угрозам ИБ в области ИИ
- Разработка Требований по обеспечению информационной безопасности в информационных системах, реализующих ТИИ

02

Дорожная карта

- Формирование Перечня областей применения СИИ, в которых может быть нанесен ущерб безопасности Российской Федерации
- Разработка принципов регулирования жизненного цикла СИИ из указанного Перечня

03

Национальная стратегия

Подготовка к созданию правовых оснований для внедрения ТИИ в области из Перечня, при условии их соответствия Требованиям

СПАСИБО ЗА ВНИМАНИЕ!