Доверенный искусственный интеллект

Арутюн Аветисян
директор ИСП РАН
академик РАН
arut@ispras.ru
17 октября 2023 г.

ИСП РАН

Искусственный интеллект: тогда и сейчас

ИСП РАН

В 1956 появился термин «искусственный интеллект».
Прошло чуть больше 40 лет и…

1997 – IBM Deep Blue выиграл в шахматы у Гарри Каспарова

2002 – первый робот-пылесос

2010 – база данных ImageNet, разметка данных обычными людьми. 14 млн изображений, 20 тысяч категорий

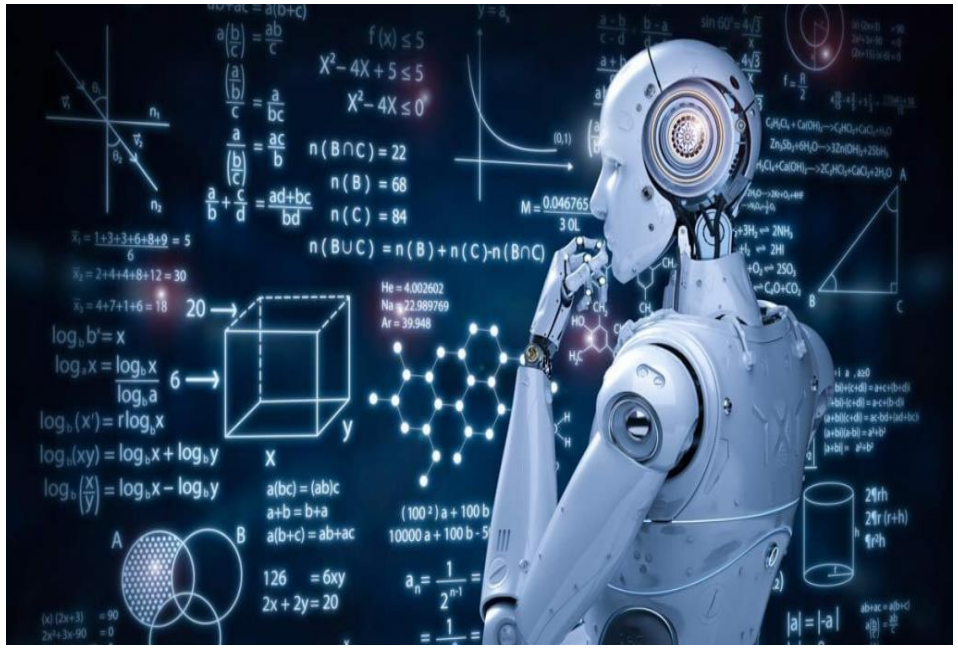
2011 – IBM Watson выиграл шоу Jeopardy! («Своя игра»)

2011 – персональный ассистент в смартфоне (Siri)

2016 – AlphaGO выиграла у профессионального игрока в Го

2016 – Google Translate начинает использовать нейронный машинный перевод для 8 языков

2022 – выпущен ChatGPT от OpenAI. За 2 месяца число пользователей достигло 100 миллионов (это рекорд)



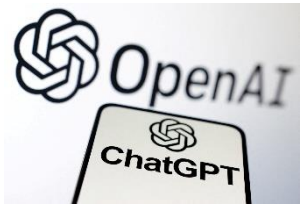
PyTorch

TensorFlow

ИИ повсюду

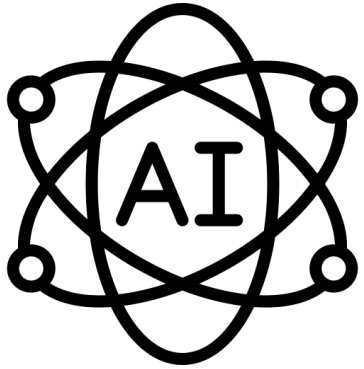
ИСП РАН

Большие языковые модели и приложения
ChatGPT
DeepL Translate
Google Assistant



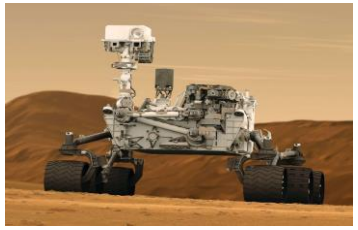
Медицина
Компьютерная диагностика
Подбор лечения.
Фитнес-браслеты, глюкометры …

Транспорт
Беспилотные автомобили



Системы безопасности
Распознавание лиц с помощью компьютерного зрения

Финансы
Обнаружение мошенничества и отмывания денег



Исследование космоса
Автономная космическая навигация (роботы на Марсе)

Торговля
Рекомендации в ритейле: Amazon, Lamoda
Роботизация складского бизнеса: Walmart

Промышленность
Роботизация производства

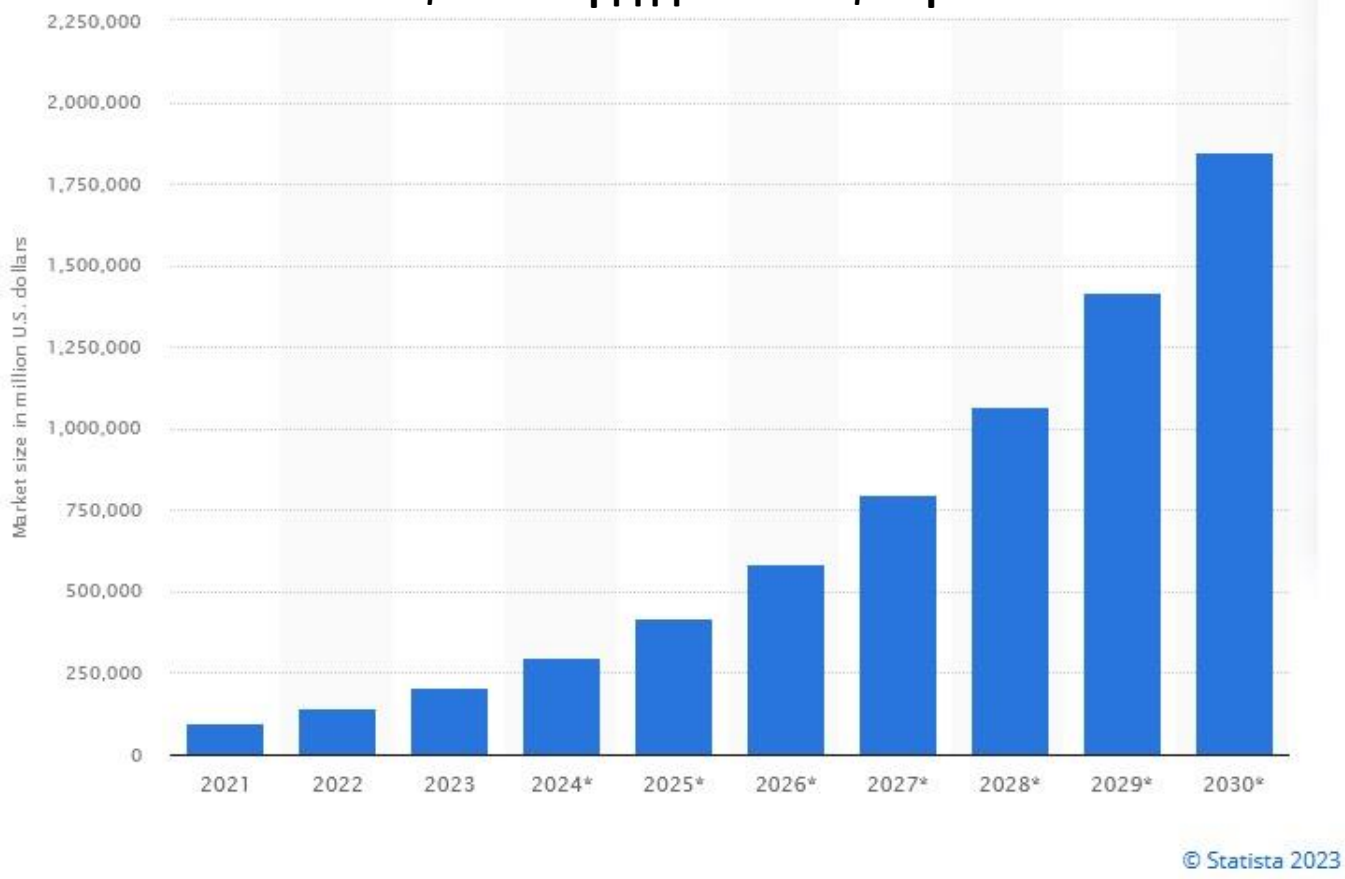
ИИ: рост рынка и повышение продуктивности

ИСП РАН

Банк Goldman Sachs (2023):

если 50% компаний во всем мире внедрят ИИ, то годовой глобальный ВВП вырастет на 7% в течение следующих 10 лет

Прогнозируемый рост глобального рынка ИИ в 20 раз за 10 лет: от $100 млрд до почти $2 трлн



https://www.bloomberg.com/news/articles/2023-03-27/goldman-says-ai-will-spur-us-productivity-jump-global-growth

ИИ: в чём сила?

ИСП РАН

СЛАБЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ (СЕЙЧАС)

Weak AI, Narrow AI

Методы: машинное обучение, глубокое обучение, нейронные сети

Может решать только те задачи, для которых он запрограммирован. Извлекает информацию из ограниченного набора данных. Если данные искажены, может выдавать необъективный (неэтичный, дискриминационный) результат. Уязвим для предвзятостей и ошибок.

СИЛЬНЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ (КОГДА?)

Strong AI, General AI

Методы: **?**

Делает интеллектуальные выводы. Решает задачи на уровне человека. Использует стратегии, функционирует в условия неопределенности, общается на естественном языке, планирует действия.

Автор: Chris Noessel

Слабый ИИ: примеры проблем

ИСП РАН

Использование дискриминирующих алгоритмов

?

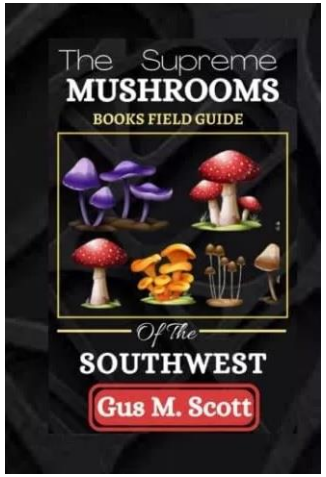
Пример (Reuters, 2018): в Amazon создали модель для выбора кандидатов на должности разработчиков. Однако потом выяснили, что система не оценивает кандидатов гендерно-нейтрально, т.к. она была обучена на данных за 10 лет, и в основном резюме были от мужчин.

ДТП с участием беспилотных автомобилей

Пример: состязательные атаки. Из-за наклеек дорожный знак STOP может стать нераспознаваемым для беспилотного автомобиля

Безответственное использование генеративных сетей

Пример (The Guardian, 2023): в продаже появились книги по сбору грибов, написанные ChatGPT. Специалисты не рекомендуют грибникам их использовать, т.к. в книгах есть ошибки



<u>КСТАТИ</u>: компании, которые занимаются ИИ в США (OpenAi, Meta.Platforms, Alphabet и др.), уже пообещали наносить «водяные знаки» на контент, созданный ИИ

Как защититься?

ИСП РАН

Нужны:
сообщество, стандарты и инструменты,
позволяющие обеспечить жизненный цикл
разработки доверенных (безопасных) технологий
искусственного интеллекта

ОБЫЧНЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

ДОВЕРЕННЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Регулирование: всё только начинается

ИСП РАН

Whitepaper on AI: A European approach (Евросоюз, 2020)

- ✓ Объясняет важность ИИ и призывает к его оптимизации и развитию экосистемы
- ✓ Инициирует работу над нормативной базой ИИ и определяет ключевые требования: безопасные обучающие данные без дискриминации; надежность и воспроизводимость; контроль человека над ИИ; защита биометрических данных

Кодекс этики в сфере ИИ (Россия, 2021)

- ✓ Разработан при участии АЦ при Правительстве, Минэкономразвития России, а также около 500 экспертов академического и бизнес-сообщества
- ✓ Подчеркивает приоритет прав человека; ответственность человека за действия ИИ; потребность в безопасности и защищенности данных

AI Bill of Rights (США, 2022)

- ✓ Разработан компаниями, общественными организациями и экспертными группами
- ✓ Формулирует пять принципов создания и использования систем ИИ, в числе которых: разработка безопасных и эффективных систем; отсутствие алгоритмической дискриминации; обеспечение конфиденциальности данных и контроль пользователя за тем, как используются его данные и др.

И ДРУГИЕ ДОКУМЕНТЫ:

- NIST AI Risk Management Framework (NIST: National Institute of Standards and technology, США)
- MITRE ATLAS, Adversarial Threat Landscape for Artificial-Intelligence Systems

Центры по ИИ и безопасности:

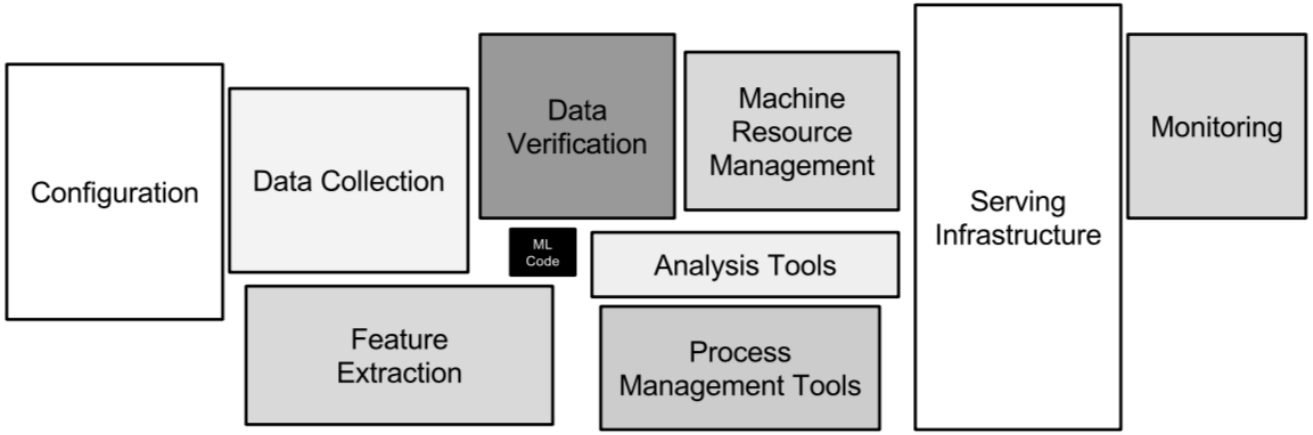
Шесть исследовательских центров ИИ (Россия, 2021)

AI Security Center (США, в 2023 объявлено об учреждении)

…

Интеллектуальные системы: структура и доверие

ИСП РАН



- Модели машинного обучения – центральная, но не самая большая часть интеллектуальных систем
- Понятие «доверие» к программным системам определяется национальными стандартами
  - ГОСТ Р 56939-2016
  - приказ ФСТЭК №131 от 30 июля 2018 года
- Принципиальное отличие «интеллектуальных систем» – информация содержится не в программном коде, а в данных
- Для обеспечения доверия к интеллектуальным системам **отсутствует научно-технологическая база**

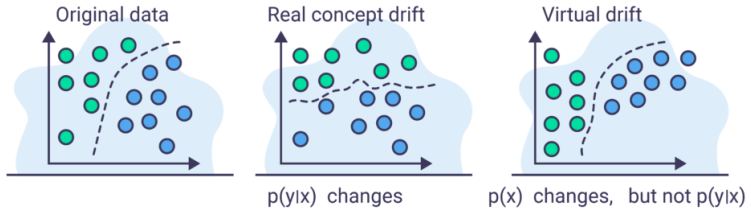
*D. Sculley et.al. (Google) Hidden Technical Debt in Machine Learning Systems. NIPS 2015

Проблемы доверия

ИСП РАН

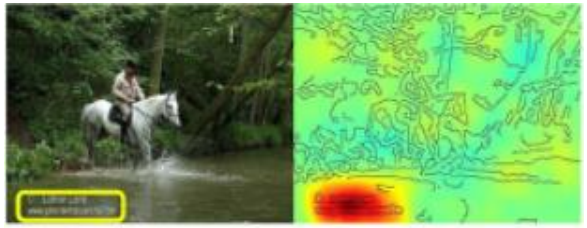
Проблемы разработки и эксплуатации

- Переобучение
- Сдвиги в данных в течение эксплуатации



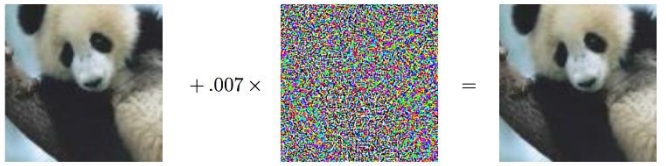
Предвзятость моделей

Интерпретация результатов

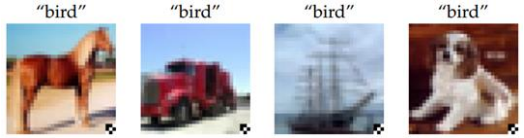


Безопасность

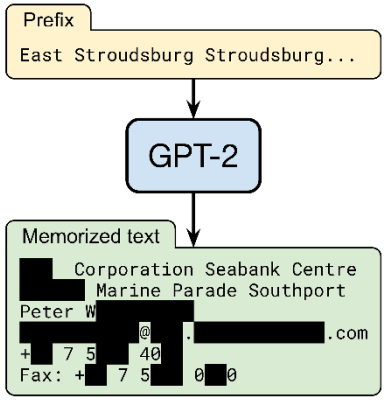
- Состязательные атаки для манипуляции



- Встраивание закладок на этапе обучения



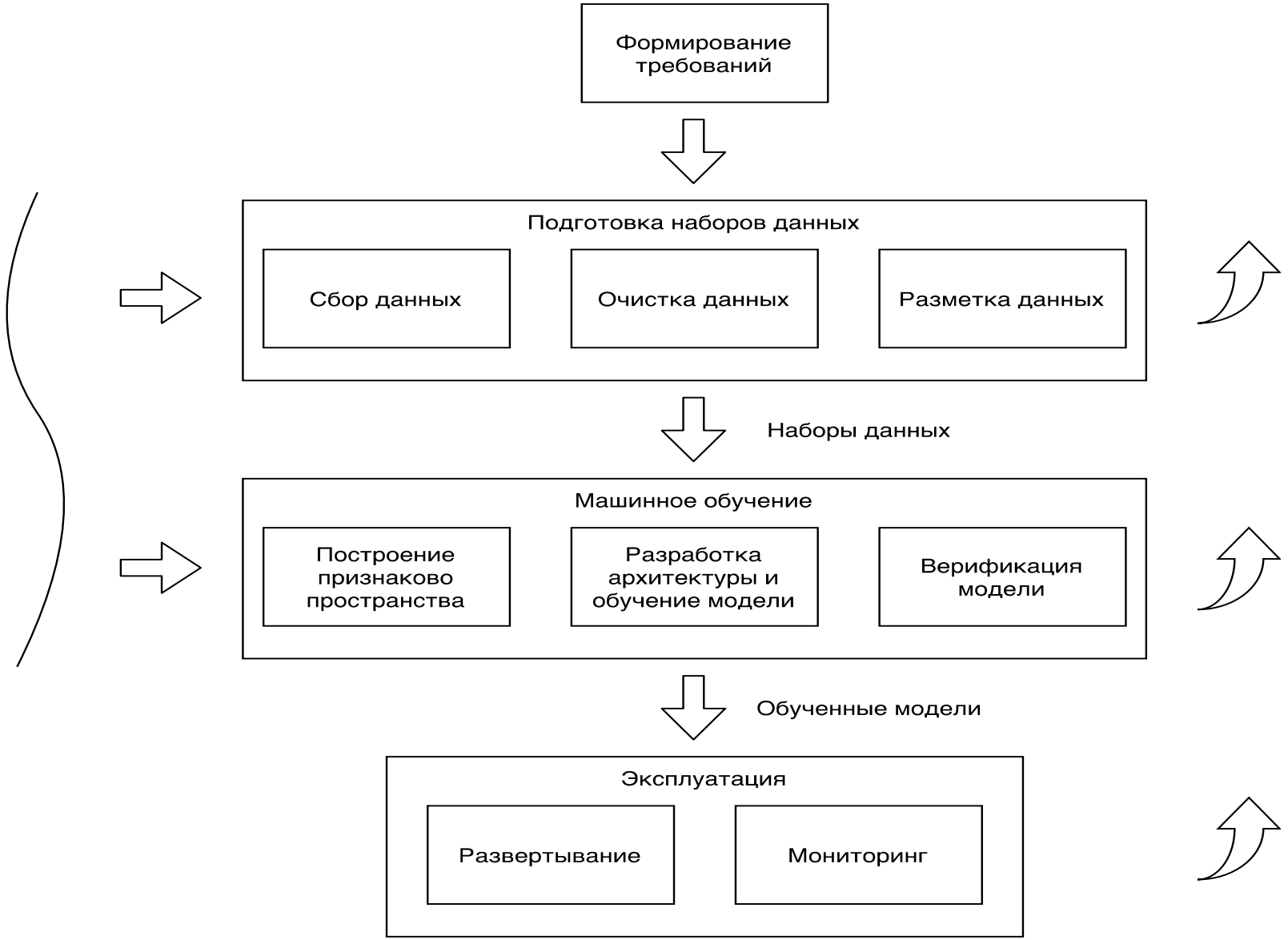
- Извлечение конфиденциальных данных из обученных моделей



- «Кража» самих моделей

База **реальных инцидентов** с ИИ:
https://incidentdatabase.ai/

Жизненный цикл машинного обучения



Вызов I: устаревание моделей

ИСП РАН

- Мир постоянно меняется, поэтому модель начинает устаревать с момента окончания разработки
- Модели нужно периодически или постоянно дообучать
  - Это сложно согласуется с ГОСТами, предполагающими «водопадную» модель жизненного цикла разработки изделий (нет возврата от эксплуатации к разработке)
- Вариант решения: включать методики и **инфраструктуру** для дообучения в готовое изделие
  - Открытый вопрос: как в таком изделии обеспечить безопасность функционирования? Сертификация?
  - Какое оборудование должно быть включено в изделие дополнительно, как эффективно его использовать?

Вызов II: инфраструктура разработки

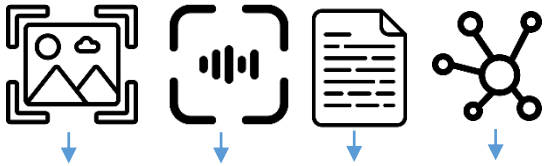
ИСП РАН

- **Продуктивность**
  - Удобство решения типовых задач
  - Простота доступа к инфраструктуре
  - Доверенный репозиторий для обмена решениями
- **Эффективность**
  - Обучение:
    - Распределенное обучение
    - Задействование оборудования из различных дата-центров
  - Эксплуатация:
    - Использование отечественного оборудования
    - Оптимизация загрузки (inference server)
- **Безопасность**

Составляющие инфраструктуры

- Нейросетевые фреймворки
- Библиотеки алгоритмов
- Библиотеки доверенных предобученных моделей
- Инструменты уменьшения моделей
- Инструменты MLOps
- Инструменты разметки данных
- Инструменты аугментации данных
- Системы автоматического обучения (AutoML)
- Библиотеки алгоритмов распределенного обучения
- Инструменты верификации моделей
- Инструменты оптимизации исполнения моделей (Inference Server)
- Средства интерактивного прототипирования
- Средства совместной разработки
- Инструменты мониторинга работы ПО
- Средства виртуализации и оркестрации
- Инструменты обеспечения жизненного цикла безопасной разработки ПО
- Системы распределенной обработки данных
- Инструменты очистки данных

Вызов III: обеспечение безопасности



Доверенные прикладные системы

Методики и ПО для разработки и оценки доверенных систем, использующих ИИ

ПО для выявления и противодействия угрозам, специфичным для ИИ

Доверенные версии фреймворков машинного обучения

Доверенные приватные облака (ПО, ПАК)

Подготовка датасетов

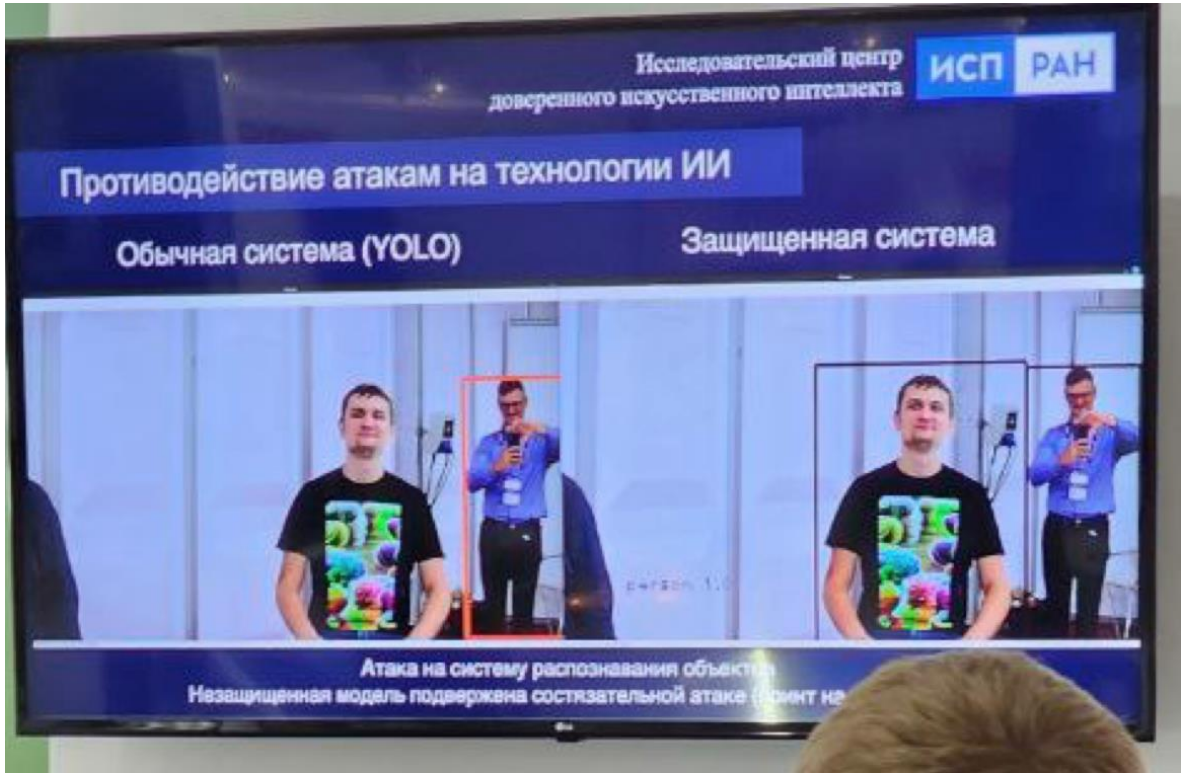
Обучение моделей

Дообучение моделей

Функционирование

- Доверенные фреймворки и библиотеки машинного обучения
- Инструменты проверки наличия аномалий в наборах данных
- Инструменты оценки устойчивости обученных моделей к атакам
- Инструменты для повышения доверия к предобученным моделям
- Методы защиты моделей от атак на этапе эксплуатации
- Методы объяснения моделей
- Методы обнаружения дрейфа данных
- Методы выявления предвзятости моделей

Пример: состязательные атаки



- **Осуществление атак**
  - Как понять, какая модель использовалась?
  - Какие есть гарантии на надежность атаки?
    - Патч-атаки в трехмерном пространстве (как обклеить 3D-объект наклейками так, чтобы он со всех сторон не распознавался)
    - Атаки на алгоритмы компьютерного зрения при захвате изображения в разных спектрах
- **Защита от атак**
  - Устойчивость защиты
  - Общая надежность защищенной интеллектуальной системы

Open source для обеспечения доверия

ИСП РАН

Активные исследования начались в 2017 году

LINUX FOUNDATION, основные проекты:
Adversarial Robustness Toolbox (ART)
AI Explainability 360
AI Fairness 360

Linux Foundation также поддерживает проекты различных компаний, нацеленные на:

- **Анализ уязвимостей моделей и повышение безопасности их использования**:

AdvBox (Baidu)
Advertorch (RBC Capital)
Foolbox (University of Tuebingen)
CleverHans (CleverHans Project)

- **Определение смещения модели**:

Aequitas (Университет Чикаго)
Audit AI (Pymetrics)
DeepLIFT (Стэнфордский университет)
Fairlearn (Microsoft)

ПРОБЛЕМА
Отсутствие общей среды для прозрачного одновременного использования разных инструментов

- Доверенный ИИ – перспективное и активно развивающееся направление
- Основная задача – создание программных инструментов и методик (основанных на возможностях этих инструментов) для обеспечения доверия к интеллектуальным системам
- Эту задачу мы успешно решаем в **Исследовательском центре доверенного искусственного интеллекта ИСП РАН**

Исследовательский центр доверенного ИИ ИСП РАН

ИСП РАН

ОСНОВАН В 2021

ЦЕЛЬ

Создание методик и соответствующих программных и аппаратно-программных платформ для разработки и верификации технологий искусственного интеллекта (ИИ) с требуемым уровнем доверия

ЗАДАЧИ

- ✓ фундаментальные и прикладные исследования в области доверенного искусственного интеллекта
- ✓ создание масштабируемой (облачной) платформы, включающей специализированные инструменты разработки доверенных систем ИИ, а также соответствующих ПАК
- ✓ подготовка высококвалифицированных кадров через реализацию новых учебных курсов по тематике Центра и вовлечение студентов и молодых специалистов во все активности Центра, в том числе разработку прикладных доверенных систем ИИ
- ✓ создание и модерирование распределенного сообщества ведущих российских ученых по тематике Центра и экосистемы потребителей на основе принципа равнодоступности результатов работы Центра

Экосистема Центра доверенного ИИ

ИСП РАН

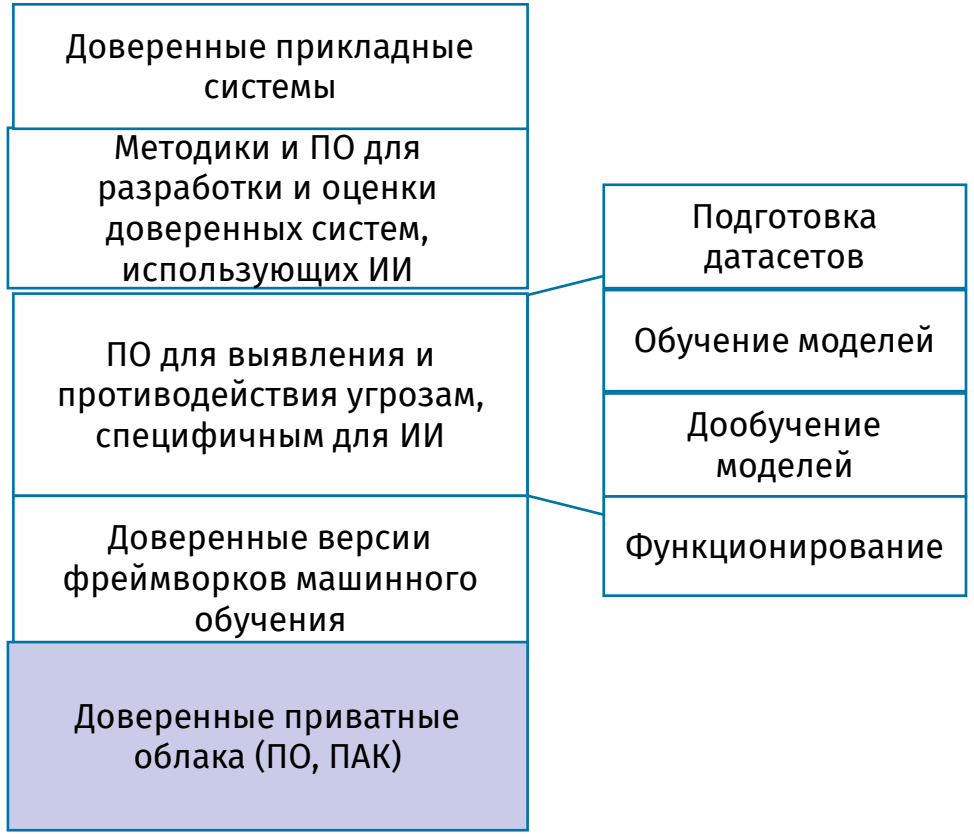
- ✓ **Работы синхронизированы с НИР, проводимыми Академией Криптографии РФ**
- ✓ **ФСТЭК России информируется о текущих исследованиях и промежуточных результатах**

Основной продукт ИЦДИИ

ИСП РАН

Облачная платформа для анализа и разработки доверенных систем, использующих технологии ИИ

Программные инструменты и методики для противодействия принципиально новым угрозам, возникающим на всех этапах жизненного цикла технологий ИИ



- Доверенные фреймворки и библиотеки машинного обучения
- Инструменты проверки наличия аномалий в наборах данных
- Инструменты оценки устойчивости обученных моделей к атакам
- Инструменты для повышения доверия к предобученным моделям
- Методы защиты моделей от атак на этапе эксплуатации
- Методы объяснения моделей
- Методы обнаружения дрейфа данных
- Методы выявления предвзятости моделей

KASPERSKY lab ЕС-ЛИЗИНГ IPC InterTrust.com ТЕХНОПРОМ

Доверенные фреймворки машинного обучения

ИСП РАН

PyTorch **7** млн строк кода **107** тыс проектов (Github)

TensorFlow **10** млн строк кода **137** тыс проектов (Github)

С августа 2022 Центром найдено и исправлено 57 ошибок

Специфика внедрения

- Непрерывная работа по выявлению дефектов фреймворков
- Синхронизация с оригинальными открытыми версиями

Проблемы: масштабируемость

- Индивидуальная сборка и поддержка фреймворков и библиотек с учетом специфики каждого заказчика
- Нехватка кадров. Формирование команд из уникальных специалистов: от DevOps до разработчиков нового функционала

2022: результат

- Уменьшение числа угроз безопасности систем ИИ за счет использования доверенных фреймворков машинного обучения
  - Апробация на решениях индустриальных партнеров
  - Внедрено в «Kaspersky Machine Learning for Anomaly Detection» v. 3.0

KASPERSKY lab, ЕС-лизинг, IPC InterProCom, ТЕХНОПРОМ

- Обеспечение технологической независимости

  Формирование коллектива, который может оперативно исправлять уязвимости в базовом ПО для машинного обучения

2023: план

- Инициативная сертификация для нового заказчика АО «КТ – Беспилотные системы»

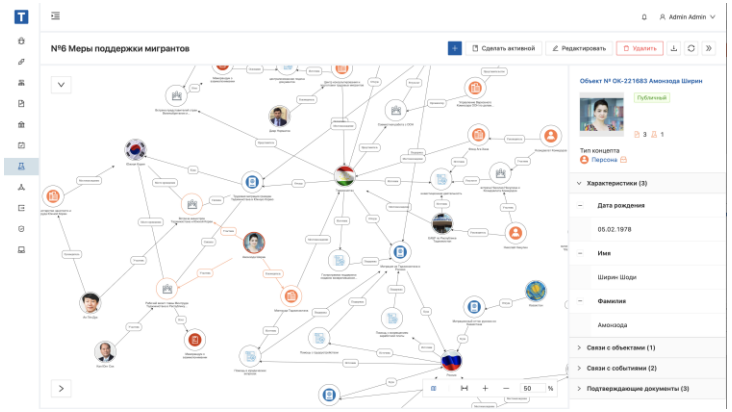
КРОНШТАДТ КТ

- Отчуждаемая методика разработки доверенных фреймворков машинного обучения
- Новые версии доверенных фреймворков

Доверенная версия платформы Talisman

ИСП РАН

Соответствующая критериям доверия* к системам, использующим технологии искусственного интеллекта



Платформа для построения интеллектуальных информационно-аналитических систем

В реестре российского ПО

НДВ2
сертификат

300+ млн руб.
Внедрения,
в частности продажа лицензий по гособоронзаказу

>50 моделей машинного обучения
анализ текста, изображений, видео, сетей

Решаемые задачи

Применение разработанных технологий и методик к платформе «Talisman», в том числе апробация и доработка фреймворков, уточнение критериев доверия с учетом специфики платформы

- Доверенные фреймворки
- Анализ используемых датасетов
- Анализ моделей машинного обучения

2023: план

- Доверенная версия технологии «Talisman»
  Апробация и внедрение индустриальными партнерами

IPC InterProCom ЕС-лизинг

- Подтверждение работоспособности методик и инструментария, разрабатываемых Центром

* Разработаны в рамках мероприятия 46 Программы Центра, ожидается официальный документ от ФСТЭК РФ в 2023 году

Не только ИЦДИИ

В ИСП РАН реализуются также и другие проекты в области ИИ, в частности:

Docmarking

Уникальная система внедрения цифровых водяных знаков (меток) в текстовые документы

- ✓ Позволяет создавать едва отличимые от оригинала цифровые и физические копии документов, идентифицирующие пользователей
- ✓ Базируется на результатах исследований в областях стеганографии, цифровой обработки изображений и машинного обучения
- ✓ В основе системы маркирования лежат методы поиска и классификации текста на изображениях, используются статистические особенности изображений документов

ECGHub

Система разметки 12-канальных ЭКГ и нейросетевые модели классификации патологий

- ✓ Позволяет предсказывать наличие или отсутствие ряда патологий, а также выполнять и верифицировать синдромальную разметку ЭКГ
- ✓ Базируется на результатах исследований в областях цифровой обработки сигналов и алгоритмов машинного обучения
- ✓ В основе системы классификации патологий лежат глубокие нейронные сети

Перспективное развитие технологий доверенного ИИ: что мы предлагаем?

Актуальная модель разработки на основе open source

ИСП РАН

Проблемы

!!! Технологические риски: недостаток доверия (отсутствие открытых инструментов анализа необходимого качества)

! Кадровые риски: дублирование работ в компаниях, нехватка квалифицированных экспертов

! Политические риски: проблемы с доступом к открытым технологиям

Эта модель не обеспечивает необходимый уровень доверия и не гарантирует долгосрочное устойчивое развитие

Модель развития, используемая ИСП РАН

ИСП РАН

Международные сообщества разработчиков открытых проектов

GitHub · SOURCEFORGE · GitLab · Bitbucket

Синхронизация с сообществами

Доверенная экосистема ИСП РАН (+ репозиторий)

Технологический центр исследования безопасности ядра Linux
Исследовательский центр доверенного ИИ
Центр исследования критических компонентов
Применяется полный стек технологий анализа (ИСП РАН)

Исследования | **Кадры**

Академическое сообщество

КОД безопасности → + новый функционал → **Продукты**

kaspersky → + новый функционал → **Продукты**

...и другие → + новый функционал → **Продукты**

Эффективность
Продуктивность
Доверие

Результаты:
- ✓ Открытое академическое сообщество
- ✓ Единый язык науки, образования и индустрии
- ✓ Доверенные базовые технологии, доступные для всех

Проблемы

~~Технологические риски~~
~~Политические риски~~

Ограниченное число проектов («кадровый голод»)

Пример: исследование безопасности системного ПО

ИСП РАН

Технологический центр исследования безопасности ядра Linux и сопутствующие проекты:

– начало исследований с ядра Linux
– дальнейшие критические компоненты: qemu, libvirt, nginx, openssl, Node.js, .NET6, UEFI и др.
– 50+ российских компаний, в том числе:
  вендоры дистрибутивов на основе Linux;
  разработчики платформ виртуализации, файерволлов, антивирусов и пр.

Используемые технологии

- ✓ Статический анализ (Svace)
- ✓ Системное и модульное тестирование (kernel-ci, LAVA, ...)
- ✓ Фаззинг (syzkaller)
- ✓ Архитектурный анализ для определения поверхности атаки (Natch)
- ✓ Динамический анализ помеченных данных (Блесна)

Результаты 2021-2023 гг.

Доверенная версия ядра Linux на основе Linux 5.10:

- ✓ 215 патчей включены в репозиторий ядра
- ✓ статический анализ: проанализировано 15 000 предупреждений, из них перепроверено 50+%
- ✓ фаззинг: 23 ошибки исправлены нами,

30 ошибок исправлены портированием патчей с более новых веток ядра на 5.10

Глобальная модель долгосрочного развития

ИСП РАН

Глобальный вызов – долгосрочное устойчивое развитие доверенного открытого ПО

Глобальная цель – технологическая независимость для всех

Результаты:

✓**Необходимый уровень доверия** без потери конкурентоспособности (эффективности и продуктивности)

✓**Открытое академическое сообщество** квалифицированных экспертов

✓**Полный контроль** над кодовой базой без каких-либо ограничений

Проблемы

~~Технологические риски~~
~~Кадровые риски~~
~~Политические риски~~

Больше информации – на наших конференциях!

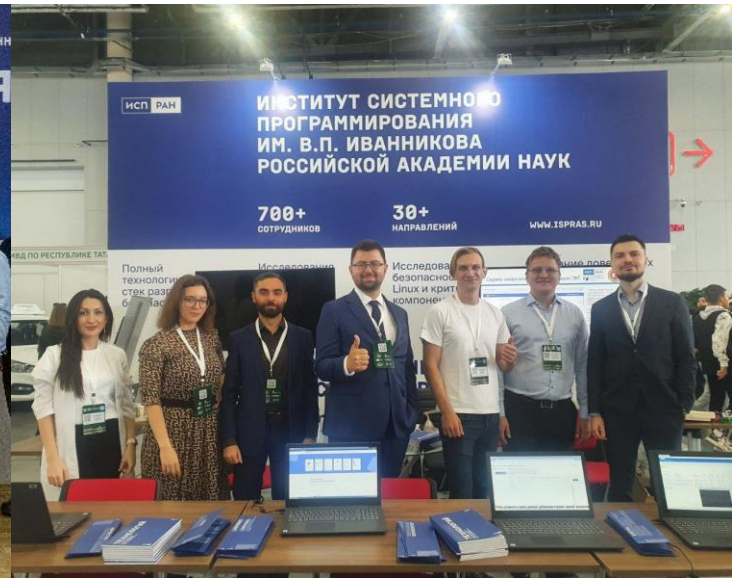
ИСП РАН

СКОРО

Открытая конференция ИСП РАН, посвящённая 75-летию отечественных информационных технологий

Москва, РАН

4-5 декабря 2023 года



Спасибо!