Доверенный искусственный интеллект: вызовы и перспективы

Арутюн Аветисян
директор ИСП РАН
академик РАН
arut@ispras.ru
14 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

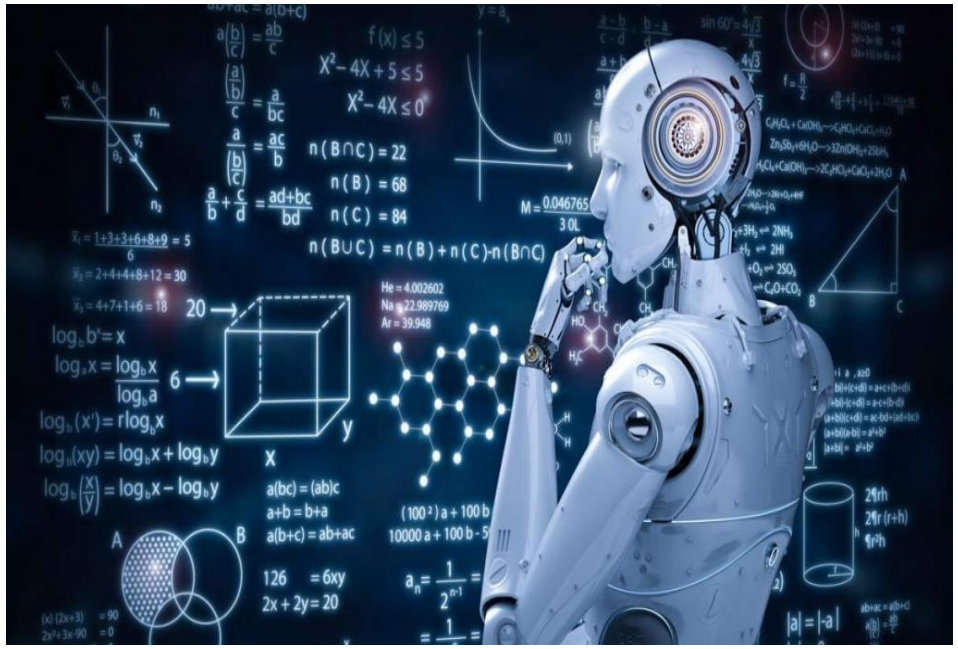
2011 – IBM Watson выиграл шоу Jeopardy! («Своя игра»)

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).



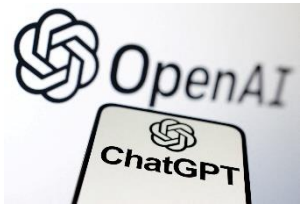
PyTorch

TensorFlow

ИИ повсюду

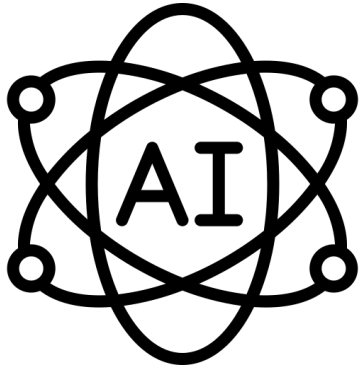
ИСП РАН

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

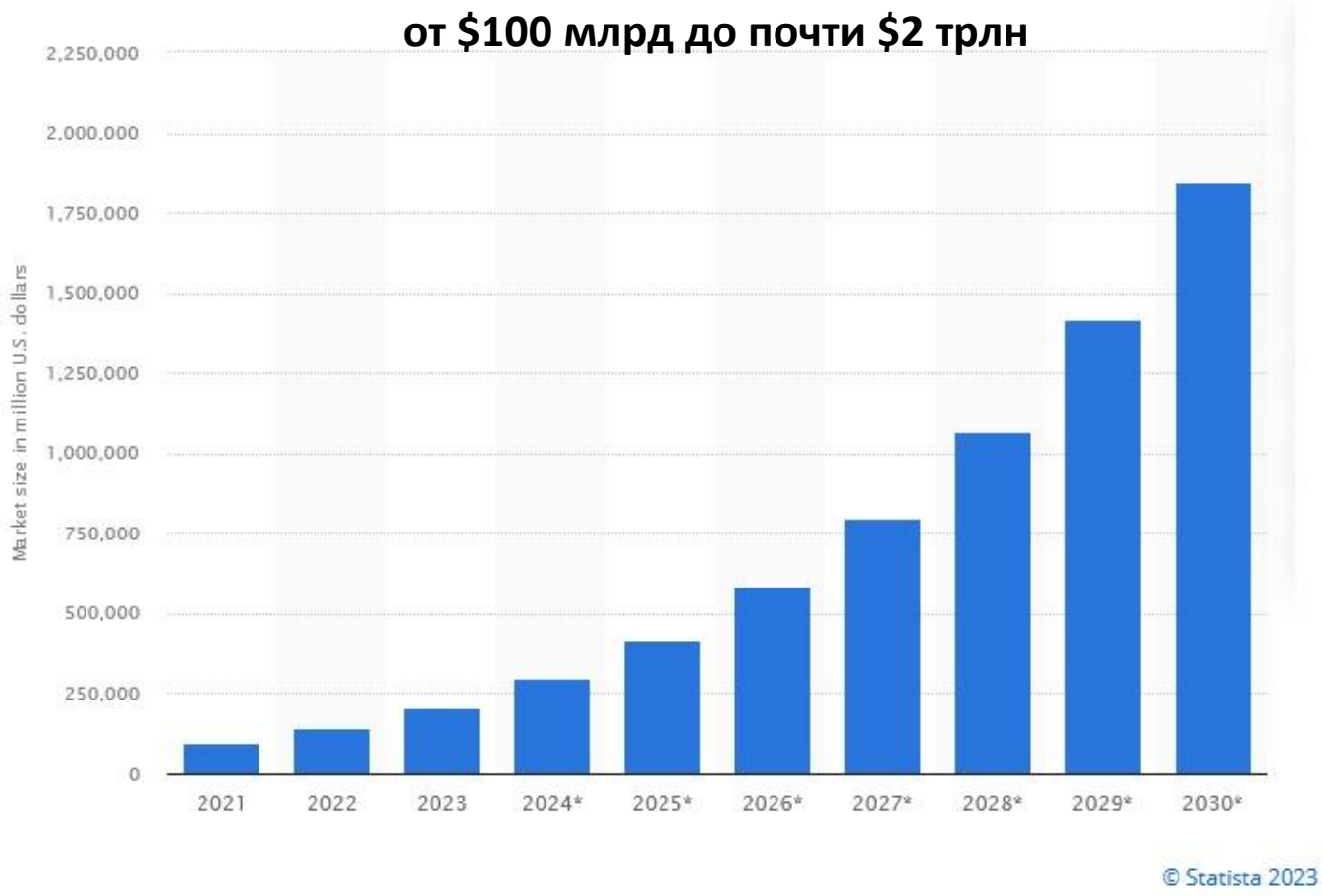
ИИ: рост рынка и повышение продуктивности

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

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

Искусственный интеллект: 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

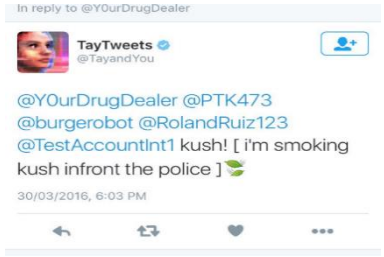
КАК ЗАЩИТИТЬСЯ ОТ ОШИБОК СЛАБОГО ИИ?

Слабый ИИ: угрозы

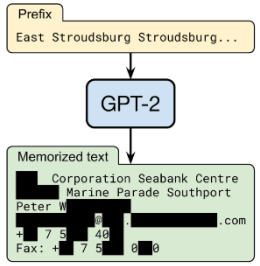
ИСП РАН

Атаки уклонения на системы обнаружения объектов

Неконтролируемое поведение дообучаемых чат-ботов



Атаки с извлечением конфиденциальных данных из обученных моделей



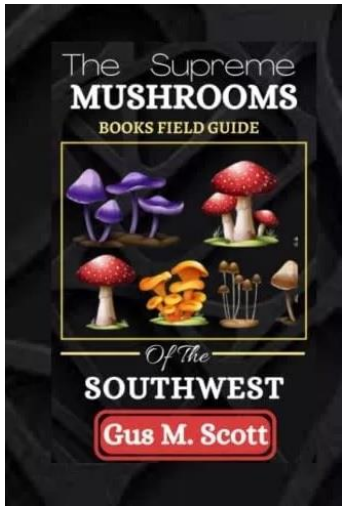
ДТП с участием беспилотных автомобилей

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.

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



КСТАТИ: 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

Как создаётся доверенное ПО в целом?

ИСП РАН

США: Разработка стандартов Common Criteria (институт NIST: National Institute of Standards and technology), 1999
Жизненный цикл разработки безопасного ПО, Microsoft, 2004

Россия: ГОСТ Р 56939-2016, 2016 (разрабатывается новая версия)
ГОСТы по процессам и инструментам (статический анализ, безопасный компилятор), 2023

ЕС: The Cybersecurity Act (EU 881 / 2019), система сертификации ПО, сервисов и процессов

Китай: стандарты по кибербезопасности от национального комитета ТК260, 19 стандартов в 2023

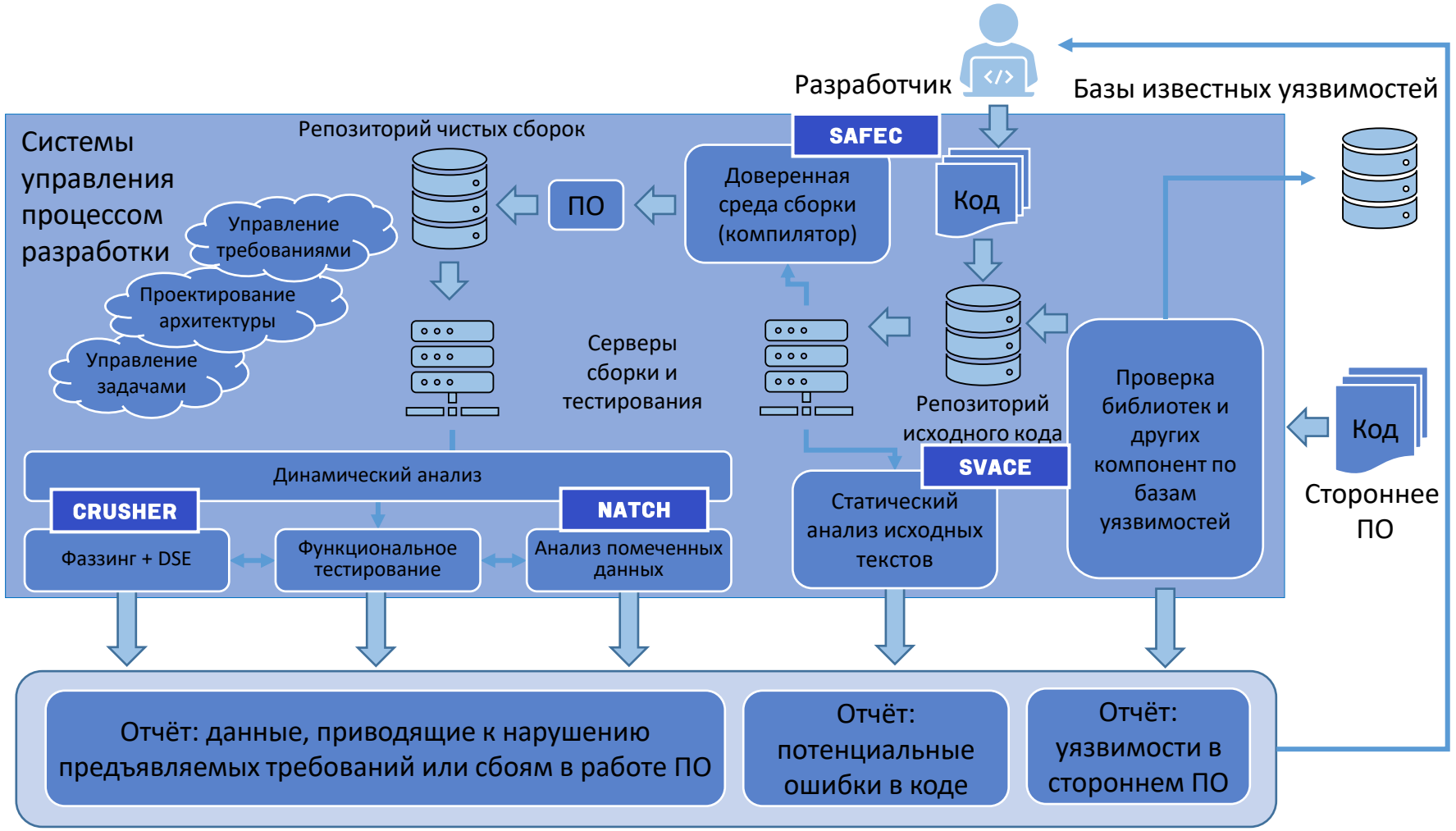
Нужно обеспечить жизненный цикл разработки доверенного ПО:

Secure Development Lifecycle, SDL

Пример SDL-цикла

Обучение	Требования	Проектирование	Реализация	Проверка	Выпуск	Реагирование
Обучение основам безопасности	Задание требований безопасности Создание контрольных условий качества и панелей ошибок Оценка рисков безопасности и конфиденциальности	Задание требований проектирования Анализ возможных направлений для злоумышленных атак Моделирование рисков	Применение утвержденных инструментов Отказ от небезопасных функций Статический анализ	Динамический анализ Нечеткое тестирование Проверка возможных направлений для злоумышленных атак	Планирование реагирования на инциденты Окончательная проверка безопасности Архивация выпуска	Выполнение плана реагирования на инциденты

Схема SDL и наши технологии



>100 компаний используют технологии ИСП РАН

Безопасный компилятор
SAFEC

Инструмент статического анализа
SVACE

Комплекс динамического анализа
CRUSHER

Инструмент определения поверхности атаки
NATCH

SDL специализируется в каждой индустрии (авиация, автомобили, космос, госсектор) под нормативные документы и специфику ПО отрасли

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)

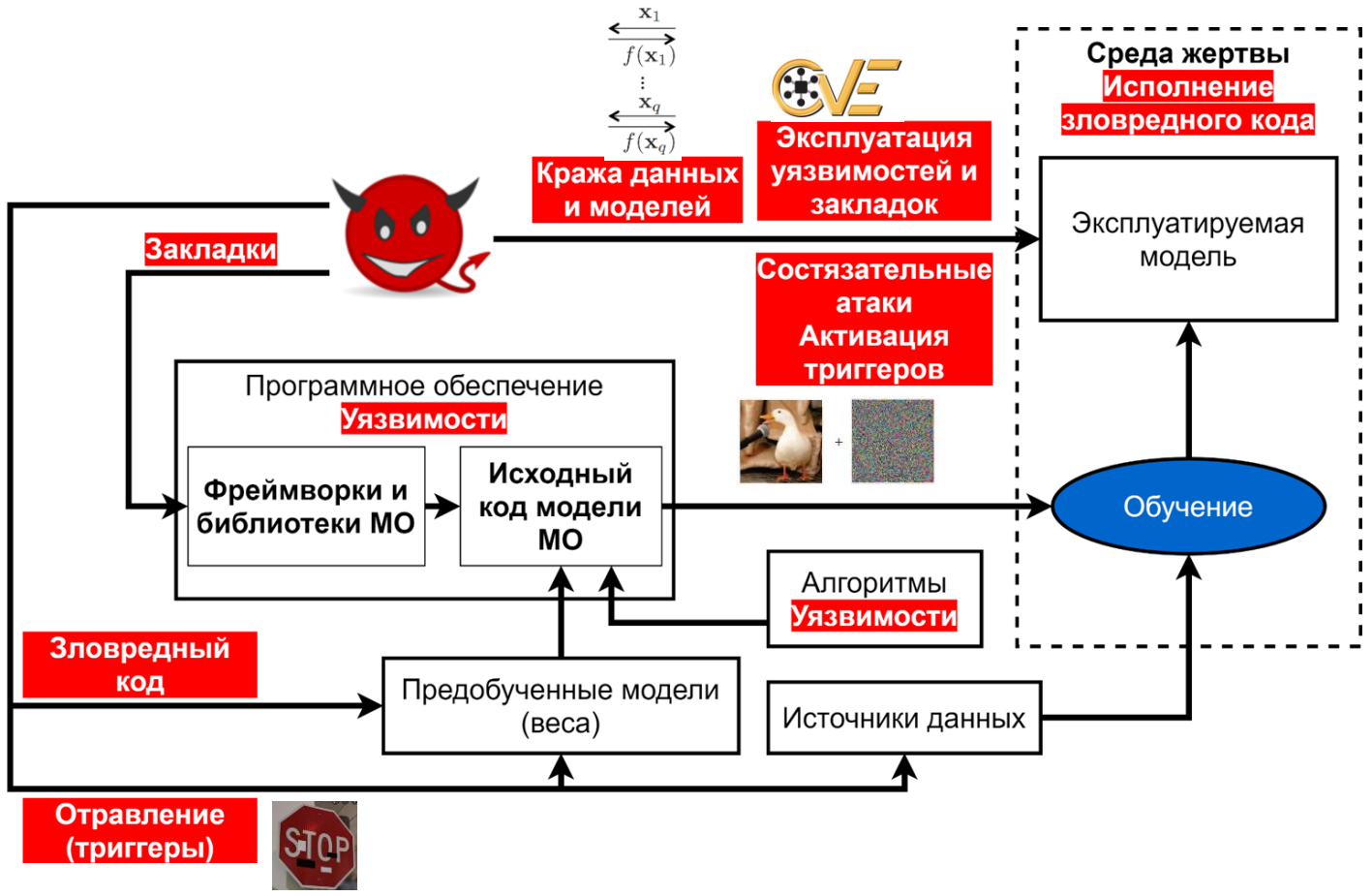
…

SDL в ИИ: что нужно?

ИСП РАН

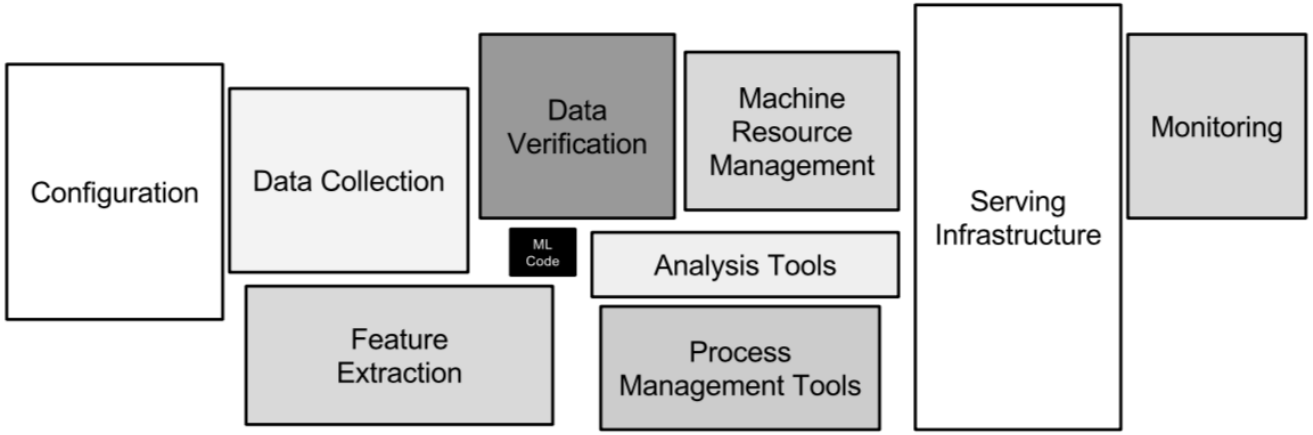
Угрозы **на разных уровнях:**

- **Исходный код инфраструктур машинного обучения** (уязвимости, закладки)
- **Данные** (отравление данных, кража из облачных сред)
- **Алгоритмы** (предобученные модели с закладками или вредоносным ПО)



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**

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

ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР ДОВЕРЕННОГО ИИ

ИСП РАН

ИЦДИИ ИСП РАН создан в 2021 году по инициативе Минэкономразвития

Основной продукт Центра:
Облачная платформа для анализа и разработки доверенных систем, использующих технологии ИИ

Ошибки в исходном коде фреймворков (TensorFlow, PyTorch)

- Классические SDL-технологии с глубоким анализом кода
- **>60** патчей поданы и приняты в сообщество. Доверенные версии фреймворков включены в «Kaspersky Machine Learning for Anomaly Detection» 3.0

Защита от атак через закладки

- Атака на классификаторы текста через изменение порядка слов
- Защита от скрытого отравления картинок

МЛ-модели: состязательные атаки

- Атака на определение объектов через генеративные состязательные сети и диффузионные модели
- Ведутся исследования:
  - ✓ Алгоритмы самостоятельного обучения для расширения данных обучения и большей надежности
  - ✓ Защита от атак через «состязательные картинки» на футболках

Модель долгосрочного устойчивого 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~~

***Создание репозитория доверенных решений поддержано на стратегической сессии «Развитие искусственного интеллекта» под председательством главы правительства РФ Михаила Мишустина в сентябре 2023 года**

4 декабря 2023 года – 75 лет ИТ в России

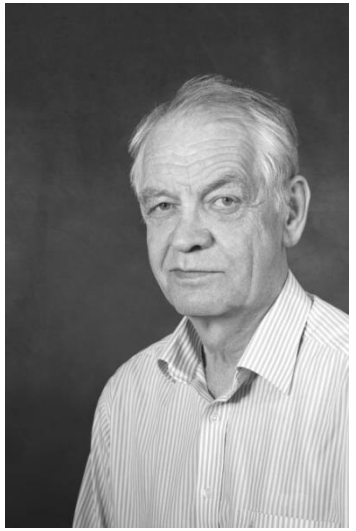
ИСП РАН

С.А. Лебедев

В.А. Мельников

БЭСМ-6 в музее науки Лондона

В.П. Иванников



Л.Н. Королёв

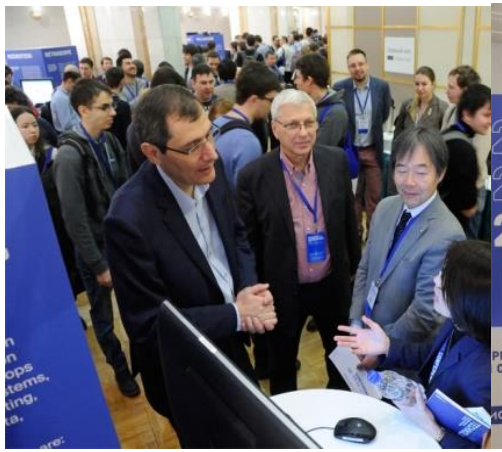
«Если мы глубже разберёмся в этом эпохальном советском суперкомпьютере, это позволит пересмотреть заявления времён холодной войны об отставании русской технологии, а также подтвердить или развеять мифы о технологическом совершенстве наших союзников».

Doron Swade, senior curator of computing and information technology

2018: 70-летие ИТ в России и пост-советских странах

2023: 75-летие ИТ в России и пост-советских странах

2024: 30-летие ИСП РАН



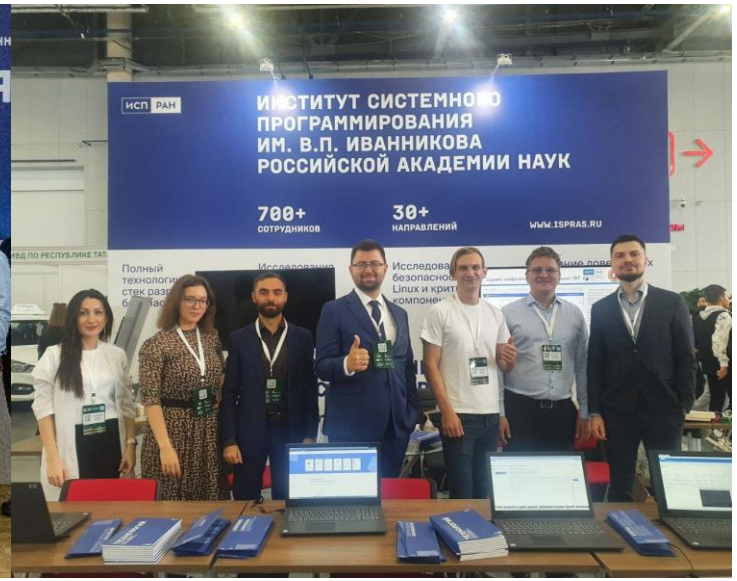
Приходите на нашу конференцию!

ИСП РАН

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

Москва, РАН

4-5 декабря 2023 года



Спасибо!

ИСП РАН