Арутюн Аветисян
директор ИСП РАН
академик РАН
arut@ispras.ru
13 февраля 2024 г.

Доверенный искусственный интеллект

c·news

ИСП РАН 30

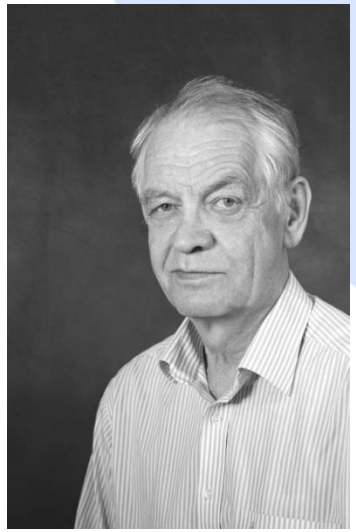
Основа ИСП РАН – постоянно развивающаяся научная школа

С.А. Лебедев

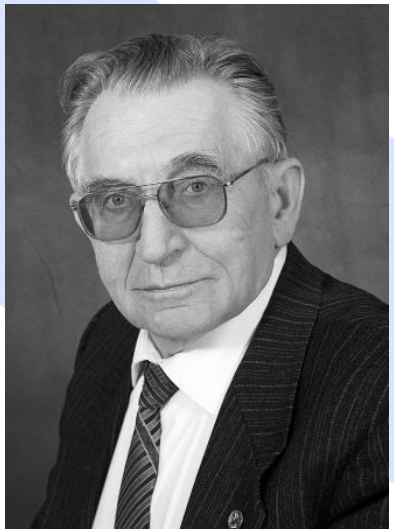
В.А. Мельников

БЭСМ-6 в Музее науки Лондона

В.П. Иванников



Л.Н. Королёв



- Фундаментальные исследования и разработка инновационных технологий
- Три кафедры системного программирования (ВМК МГУ, МФТИ, ФКН ВШЭ)
- Лаборатории под научным руководством ИСП РАН (Ереван, Великий Новгород, Орёл)
- Более 700 сотрудников
- Технологии ИСП РАН внедрены более чем в 100 компаниях

В честь исторической даты:
4-5 декабря 2023
Открытая конференция ИСП РАН, >1000 участников

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

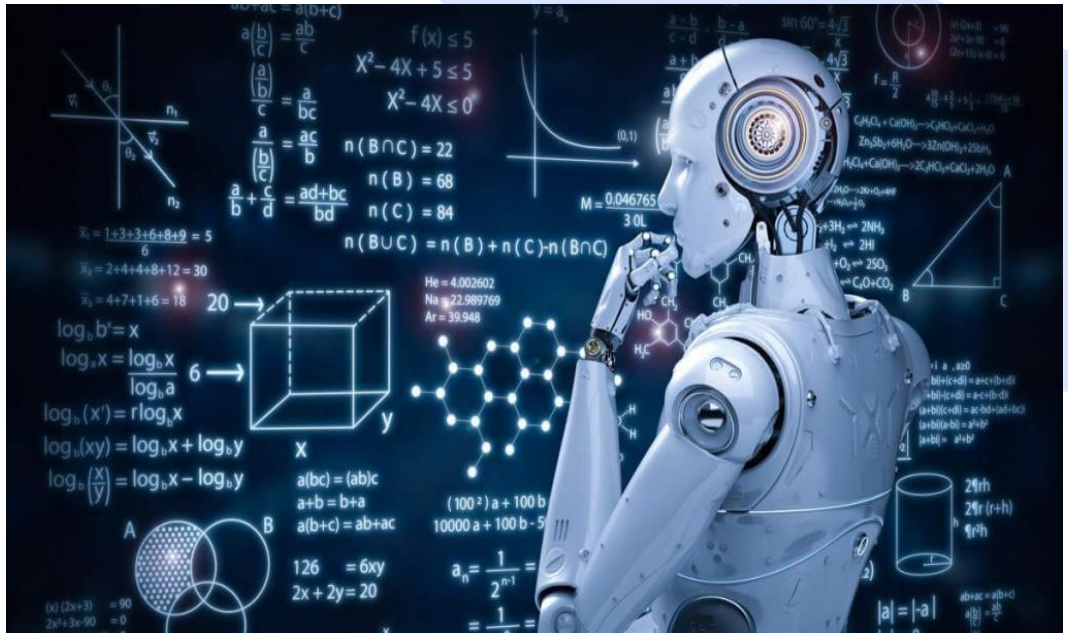
2011 – IBM Watson выиграл шоу Jeopardy! («Своя игра»)

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

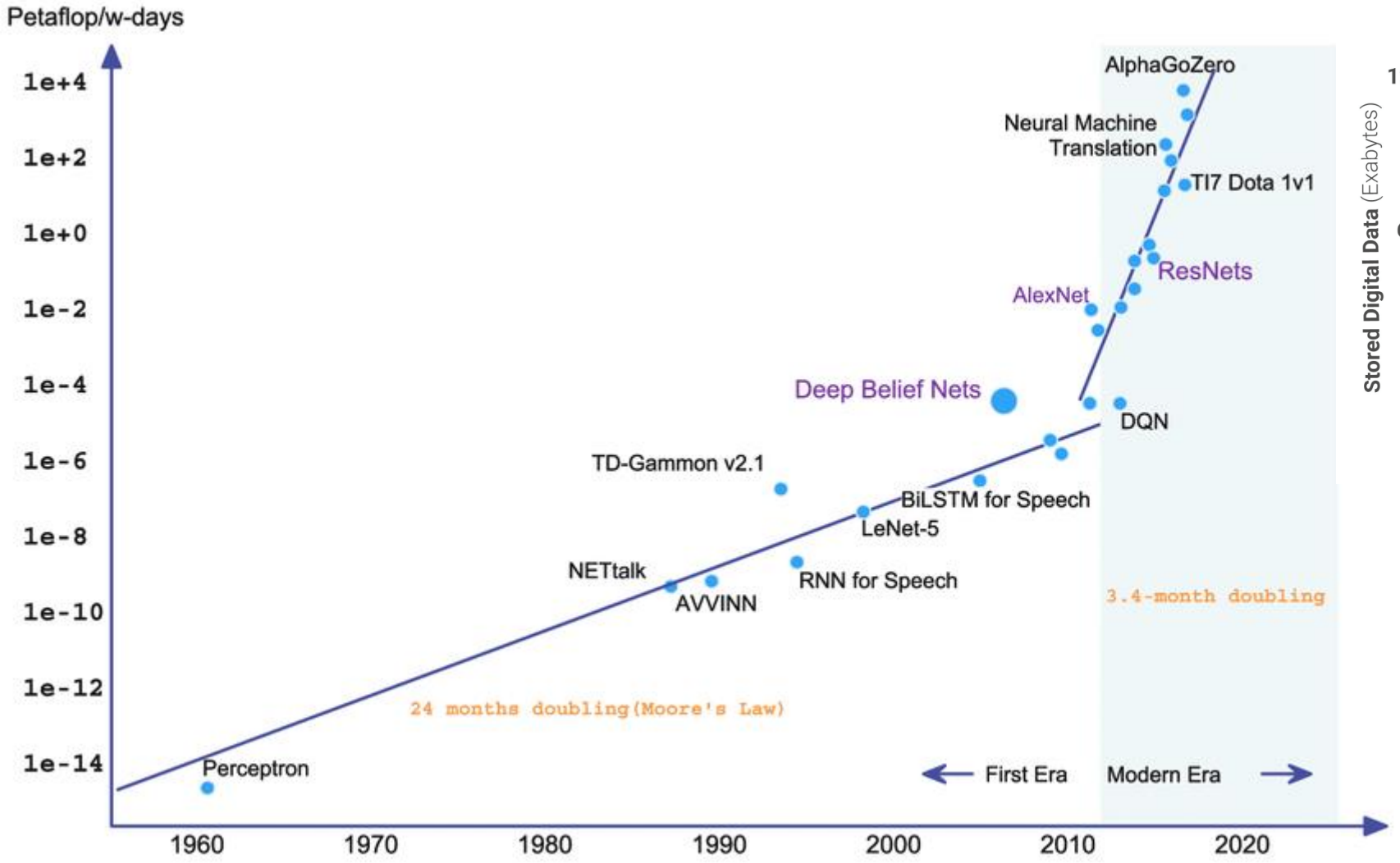
2022 – по н.в.: появление и развитие «больших моделей»: Open AI ChatGPT, YandexGPT2, RuGPT3 (Сбер) и другие.

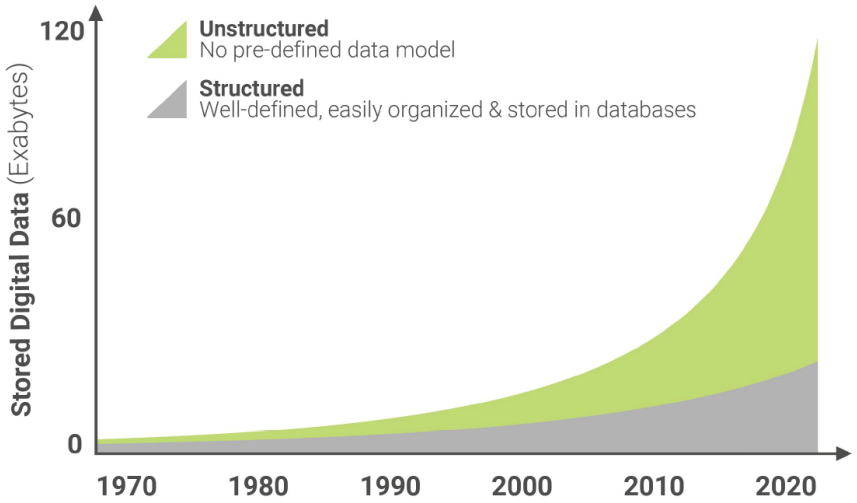


Переход от моделирования к решению задач по аналогии

Вычислительные ресурсы и большие данные: двигатели развития ИИ

ИСП РАН





- **Рост вычислительной мощности**
- **Рост объёма неструктурированных и структурированных данных**

https://www.researchgate.net/figure/Two-distinct-eras-of-computation-usage-in-training-AI-systemscredit-OpenAI-48_fig3_346359064

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
- **Представляет собой технологию без субъектности**

СИЛЬНЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ (КОГДА?)

Strong AI, General AI

Методы: ?

- Делает интеллектуальные выводы
- Решает задачи на уровне человека
- Использует стратегии, планирует действия
- Функционирует в условия неопределенности
- Общается на естественном языке
- Способен к абстрактному мышлению
- **Пока не существует**

Автор: Chris Noessel

Регулирование: Россия, Китай

ИСП РАН

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; необходимость разработки безопасных технологий

ГОСТ 59921. Системы ИИ в клинической медицине (Россия, 2022)

Задает требования к клиническому тестированию ИИ-систем на основе глубоких нейронных сетей

Стандарты ИИ в Китае находятся в разработке

Создаются национальным комитетом ТК260

2019:
Национальная стратегия развития ИИ на период до 2030 года утверждена Указом Президента РФ №490.

Отсутствие понимания того, как искусственный интеллект достигает результатов, является одной из причин низкого уровня доверия к современным технологиям искусственного интеллекта и может стать препятствием для их развития.

http://static.kremlin.ru/media/events/files/ru/AH4x6HgKWANwVtMOfPDhcbRpvd1HCCsv.pdf

2021:
Шесть исследовательских центров открыты в России
(включая Исследовательский центр доверенного искусственного интеллекта ИСП РАН)

2023:
В России учреждены еще 6 исследовательских центров ИИ

Регулирование ИИ: Европа, США

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

EU AI Act (Евросоюз, предварительно одобрен в 2023)

- ✓ Предлагает разделить все системы с ИИ на три категории: с неприемлемыми рисками, высокими и низкими. Первые запретят, вторые должны будут соответствовать «определённым юридическим требованиям», третьи не будут регулировать.

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 др.

Executive Order on Safe, Secure, and Trustworthy AI (США, 2023)

- ✓ Устанавливает новые стандарты в сфере безопасного развития ИИ
- ✓ Содержит поручения для ведомств и разработчиков (например, разработчики ряда значимых систем обязаны делиться с правительством результатами тестов на безопасность продуктов)

2023: Агентство национальной безопасности США объявило о создании AI Security Center

2023: Национальный научный фонд США объявил о создании семи исследовательских институтов ИИ.
Один из них:
NSF Institute for Trustworthy AI in Law & Society (TRAILS)

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

Регулирование: общемировое
Путь к доверенному ИИ

«Хиросимский процесс ИИ» был учреждён на саммите G7 19 мая 2023 года с целью содействия развитию передовых систем ИИ на глобальном уровне. 30 октября лидеры G7 поддержали Международный кодекс поведения и Руководящие принципы для организаций, разрабатывающих передовые системы ИИ. Выдержки из Кодекса («направлен на продвижение безопасного, надежного и заслуживающего доверия ИИ во всем мире»):

- ✓ «Тестирование и меры по смягчению последствий атак должны быть направлены на обеспечение надежности, безопасности и защищенности систем на протяжении всего их жизненного цикла – чтобы системы не содержали в себе непредсказуемые риски».
- ✓ «Для обеспечения такого тестирования разработчики должны добиваться полной прозрачности – документировать используемые наборы данных, процессы и решений, принятые в ходе разработки системы. Кроме того, нужно поддерживать регулярно обновляемую техническую документацию».
- ✓ «Организациям следует создать или присоединиться к процессам для разработки, продвижения и принятия, где это необходимо, общих стандартов, инструментов, механизмов и лучших практик для обеспечения безопасности, надежности и достоверности передовых систем ИИ».
- ✓ «Организациям также следует стремиться к разработке инструментов или интерфейсов, дающих возможность пользователям определить, был ли данный контент создан с помощью продвинутой системы ИИ – например, через применение водяных знаков. Организациям следует сотрудничать и инвестировать в исследования, если нужно, чтобы продвинуться в этой области».
- ✓ «Организациям рекомендуется обмениваться исследованиями и лучшими практиками по снижению рисков».

Примеры принципиальной уязвимости технологий ИИ

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 беспристрастно, 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.

ДТП с участием беспилотных автомобилей

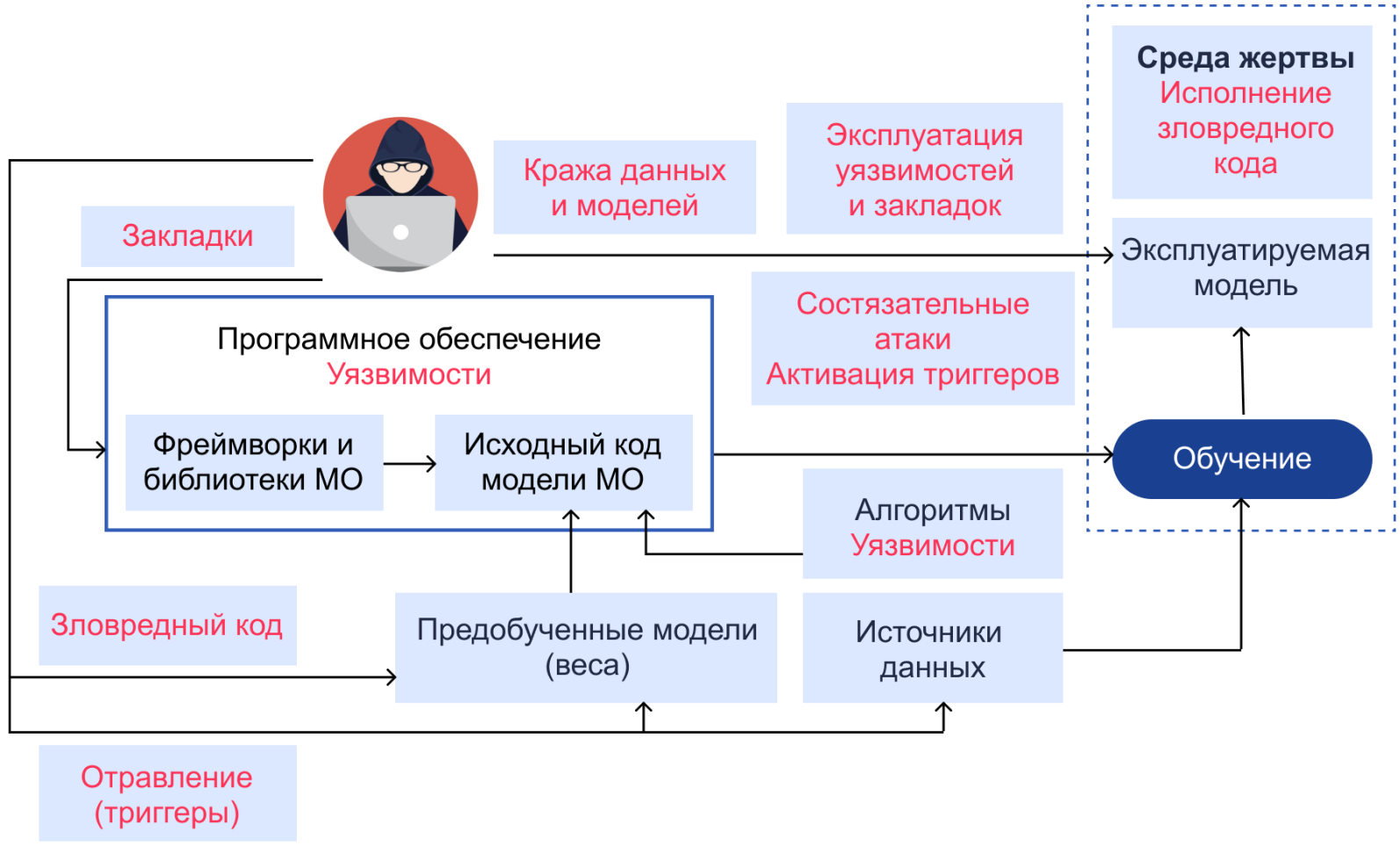
Пример I: состязательные атаки. Из-за наклеек дорожный знак STOP может стать нераспознаваемым для беспилотного автомобиля.

Пример II: неадекватная реакция беспилотного автомобиля на ДТП. В 2023 в США автомобиль Cruise наехал на пешехода, протащил его на 6 метров вперед и остановился, не съезжая. Человек выжил, но получил серьезные травмы. 950 машин Cruise отозваны для обновления ПО.

Основной источник уязвимостей связан с данными

ИСП РАН

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



Для решения этих проблем на базе ИСП РАН по инициативе Минэкономразвития в 2021 был создан Исследовательский центр доверенного ИИ (ИЦДИИ).

Множество проектов по развитию доверенного ИИ

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**:

NextAttack (University of Virginia)
Foolbox (University of Tuebingen)
CleverHans (CleverHans Project)
…

- **Определение смещения модели**:

Aequitas (Университет Чикаго)
Fairlearn (Microsoft)
…

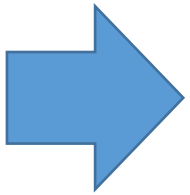
ПРОБЛЕМА
Отсутствие общей среды для прозрачного одновременного использования разных инструментов

Примеры новых постановок задач,
связанных со спецификой данных

ИСП РАН

Если украли модель или данные, как доказать их принадлежность?

Если нужно обучаться на общих моделях, как сохранить приватность данных?



Федеративное обучение — это метод распределённого машинного обучения, который позволяет обучать модели на нескольких устройствах без обмена образцами данных.

Цифровые водяные знаки (ЦВЗ) – специальные метки, встраиваемые в цифровой контентс целью защиты авторских прав и подтверждения целостности самого документа.

Мировой тренд:
ЦВЗ для защиты потребителей ИИ-контента от фейков.
Потенциальная криминализация контента без ЦВЗ

Исследовательский центр доверенного ИИ: создан в 2021 по инициативе Минэкономразвития

ИСП РАН

Цель работы Центра:
создание научно-технологической базы обеспечения необходимого уровня доверия систем с использованием ИИ

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

Направления работ:

- **Защита моделей машинного обучения от состязательных атак;**
- **Защита от атак через закладки;**
- **Поиск и исправление ошибок в исходном коде фреймворков машинного обучения (>60 патчей поданы и приняты в основные ветки TensorFlow и PyTorch);**
- **И многое другое.**

Отдельные прорывные технологии – необходимы, но не достаточны.

Нужны модели, обеспечивающие долгосрочное устойчивое развитие и технологическую независимость отрасли ИТ, а значит, и страны в целом.

Пример глобальной модели 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

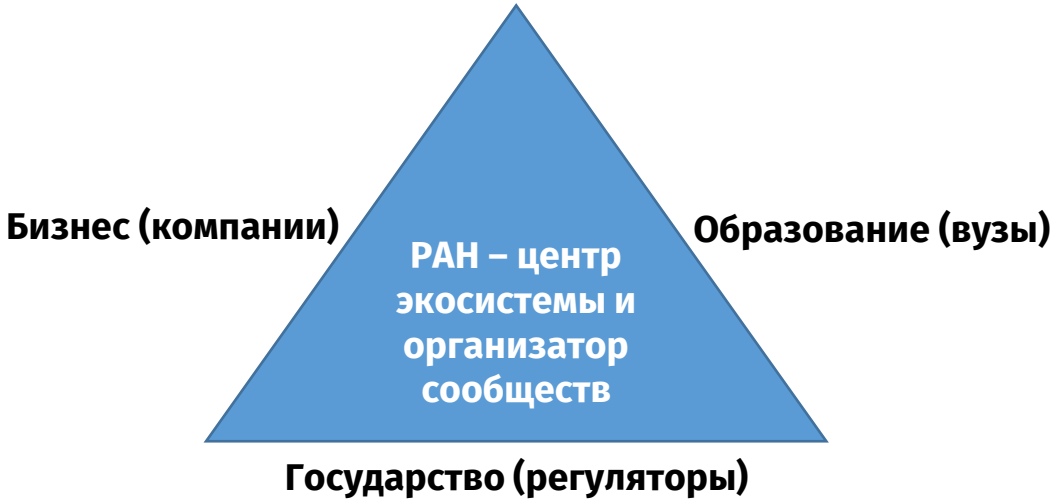
Проблемы

~~Технологические риски~~

~~Кадровые риски~~

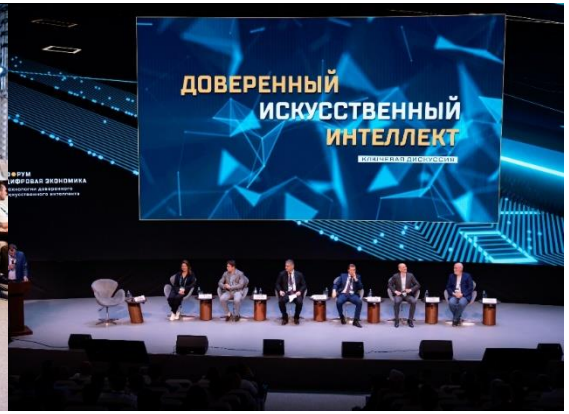
~~Политические риски~~

Единый язык науки, бизнеса и образования



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

Оргкомитет конференции: академики РАН В.А. Лекторский, Т.Я. Хабриева, Д.В. Ушаков и др.



Заседание Научно-консультативного совета ООН РАН на тему «Закат общества конкуренции и коллаборативное преимущество» (2023). Академики РАН Г.А. Тосунян и А.А. Гусейнов

… и многие другие мероприятия

Предложения по масштабированию и ускоренному развитию

ИСП РАН

1. **Развернуть Комплексную научно-техническую программу/проект (КНТП)** нацеленную на исследование перспективных подходов к обеспечению кибербезопасности и создание интеллектуальных технологий и инструментальных средств, обеспечивающих минимизацию угроз безопасности, связанных с ошибками, включая новые виды уязвимостей и рисков связанных с использованием технологий ИИ.

2. Определить, что важным механизмом реализации КНТП является **создание репозиториев доверенных средств ИИ и инструментов обеспечения доверия.**

3. **Развивать нормативное регулирование ИИ** в Российской Федерации, которое в зависимости от применения предусматривает, как возможности саморегулирования, так и обязательную государственную сертификацию на основе высокотехнологичных программных средств.

4. **Расширить подготовку специалистов** высшей квалификации по специальности «Кибербезопасность».

Спасибо!