Арутюн Аветисян
директор ИСП РАН
академик РАН
arut@ispras.ru
11 декабря 2023 г.

Доверенный искусственный интеллект

300 лет Российская Академия Наук

ИСП РАН

1948: точка отсчёта

ИСП РАН

29 июня

Основан Институт точной механики и вычислительной техники (ИТМиВТ) Академии наук СССР

4 декабря

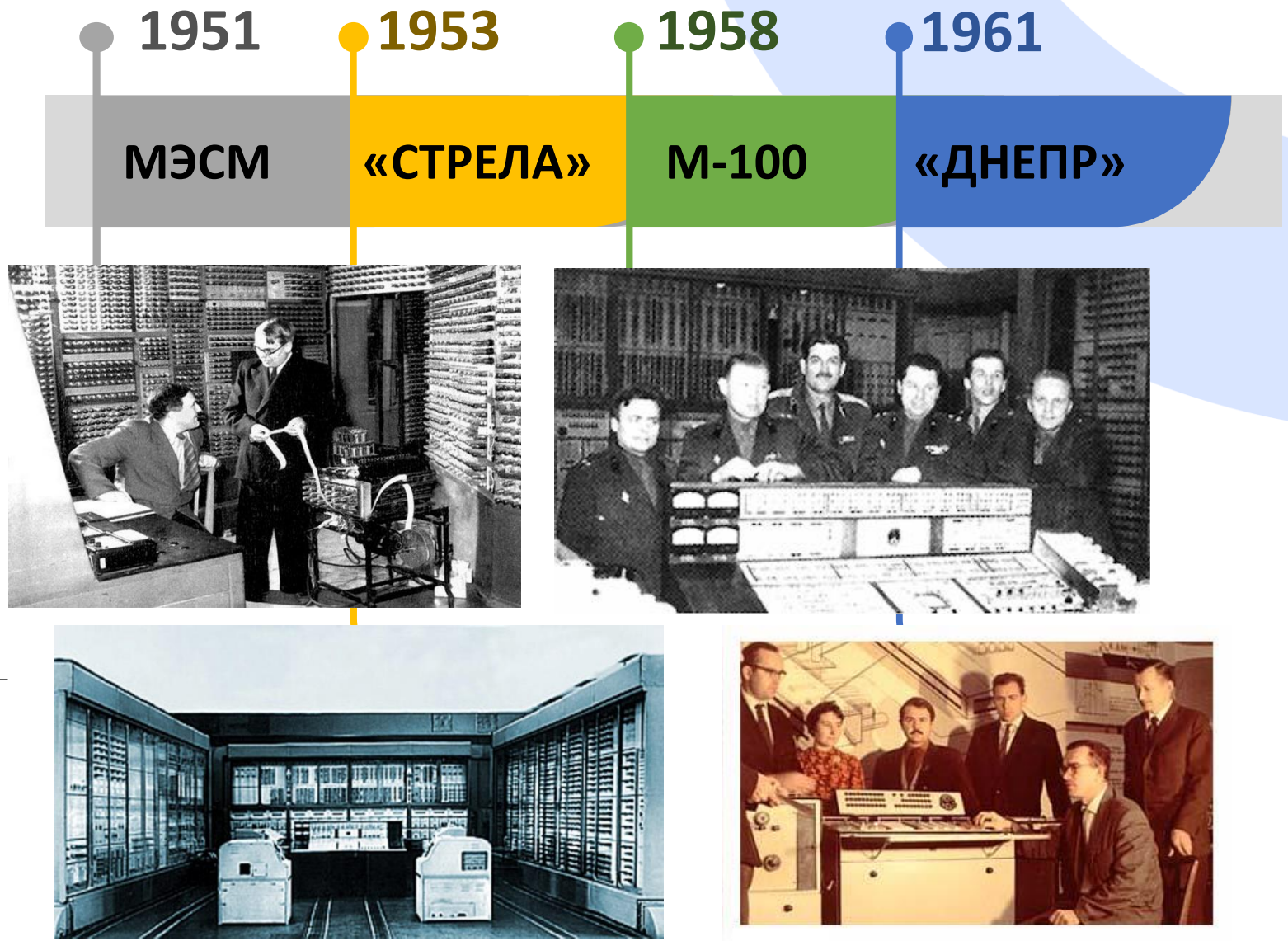
И.С. Брук и Б.И. Рамеев подали заявку на изобретение автоматической цифровой вычислительной машины

19 декабря

Основано Специальное конструкторское бюро №245 (СКБ-245, впоследствии НИЦЭВТ – Научно-исследовательский центр электронной вычислительной техники)

1949

Основана кафедра вычислительной математики мехмата МГУ (заведующий с 1952 – С.Л. Соболев, заместитель И.В. Курчатова)



1967: первый суперкомпьютер в СССР

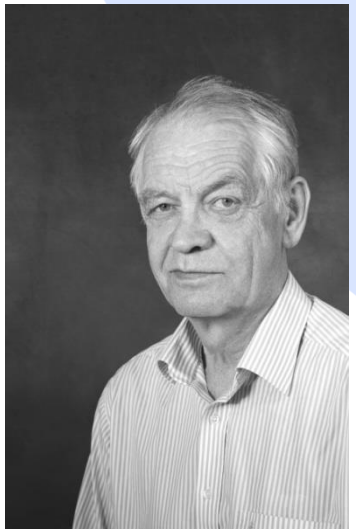
ИСП РАН

С.А. Лебедев

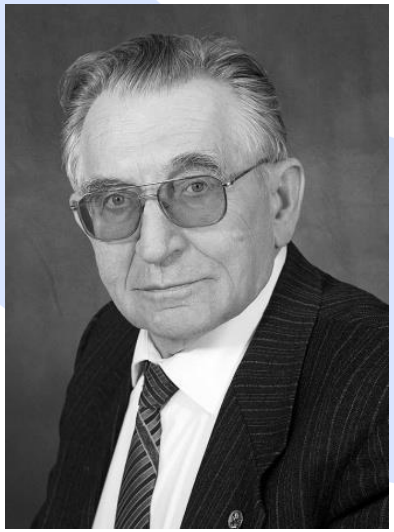
В.А. Мельников

БЭСМ-6 в Музее науки Лондона

В.П. Иванников



Л.Н. Королёв



БЭСМ-6 – первый суперкомпьютер в СССР

Скорость операций – 1 млн в секунду

Главный конструктор – С.А. Лебедев (ИТМиВТ)

Операционная система для БЭСМ-6 – Д-68

Разработана в ИТМиВТ (Л.Н. Королёв, В.П. Иванников*, А.Н. Томилин и др.)

**Основатель и первый директор ИСП РАН*

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

Доверенность – ключевой компонент качества ПО
в условиях ускоренного внедрения

ИСП РАН

Проблемы современных программно-аппаратных систем

- Эскалация размеров
- Сложность разработки и сборки
- Отсутствие изолированных систем

Необходимые качества современного ПО:

- **Эффективность**
- **Продуктивность**
- **Доверенность**

Немного статистики

GitHub-2023

>100 млн разработчиков (в 2016 – 5,8 млн)
284 млн активных репозиториев

ОС, фреймворки

PyTorch 7 млн строк кода
TensorFlow 10 млн строк кода
Astra Linux >150 млн строк кода

Большие данные

2020 – 64 зеттабайт (1 зеттабайт = 1 млрд ТВ)
2022 – 97 зеттабайт
2025 – 180 зеттабайт

Ошибки в ПО – основная причина уязвимости систем

ИСП РАН

Принципиальное наличие уязвимостей* в ПО и аппаратуре:

функциональные, архитектурные, программного кода/микрокода.

Размыты границы между ошибками программиста, закладками и НДВ

Утечка информации о уязвимостях

Хакеры «для интереса»

- Взломы ради известности
- Ограничены ресурсы
- Только известные эксплойты

Хакеры «для ущерба»

- Вандализм
- Ограничены ресурсы

Преступность

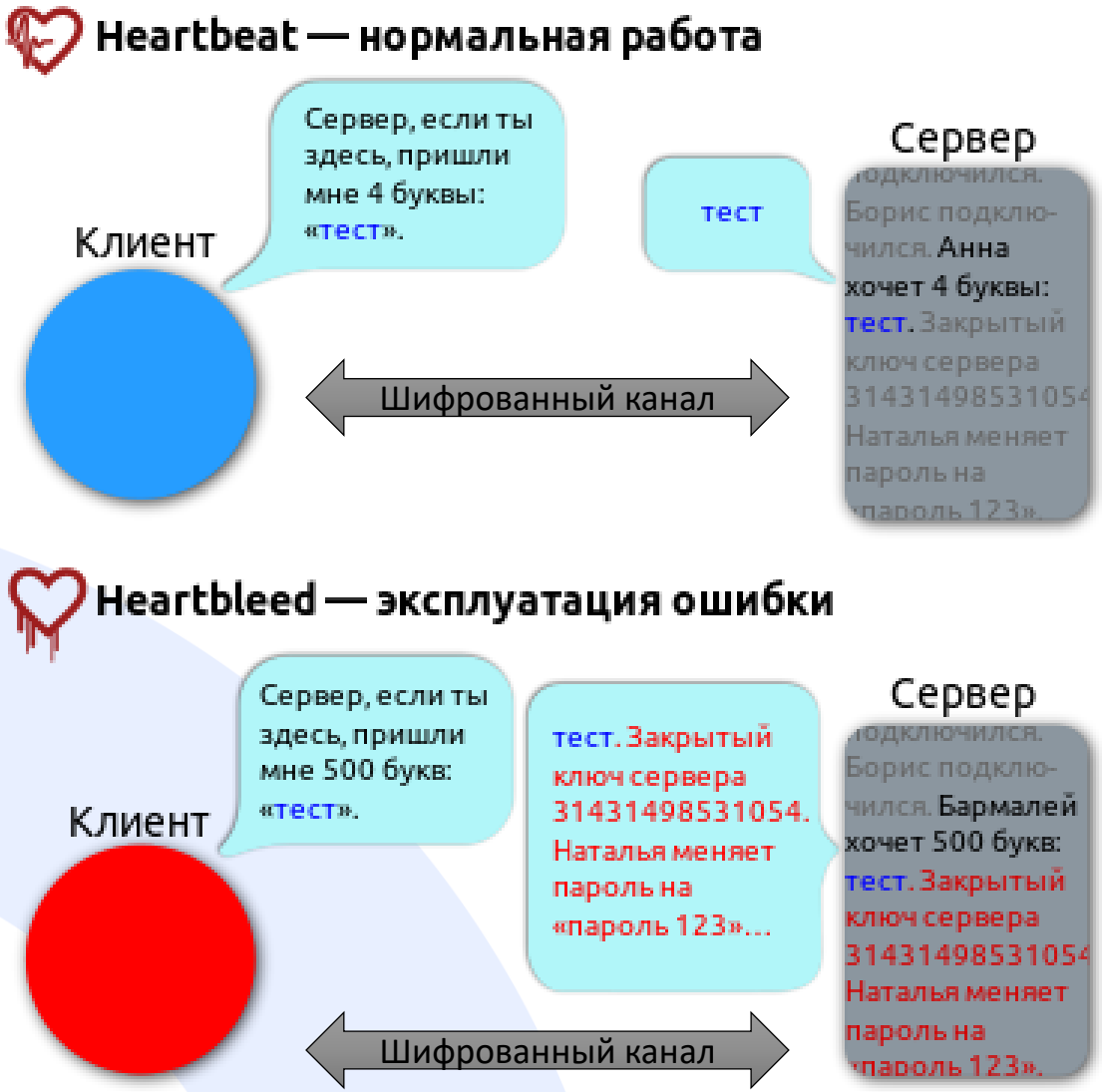
- Взлом ради прибыли
- Существенные ресурсы
- Синдикаты
- Специально разработанные программы для кражи данных

Спецслужбы

- Атаки на критические информационные инфраструктуры, промышленный шпионаж
- Практически неограниченные ресурсы
- Сложное ВПО и программы для взлома
- Постоянные угрозы

Рост ресурсов и сложности атак

Пример: уязвимость Heartbleed (библиотека OpenSSL)



500000 сайтов заражено
$500 млн потерь

- **Ошибка чтения данных за границей буфера: злоумышленник контролирует длину посланного текста**
- **Происходит утечка пользовательских данных**
- **Весь обмен данными строго следует зашифрованному протоколу**

Версия OpenSSL с уязвимостью была выпущена в марте 2012 года и обнаружена только через два года

Разработка безопасного ПО (РБПО): стандарты и требования

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

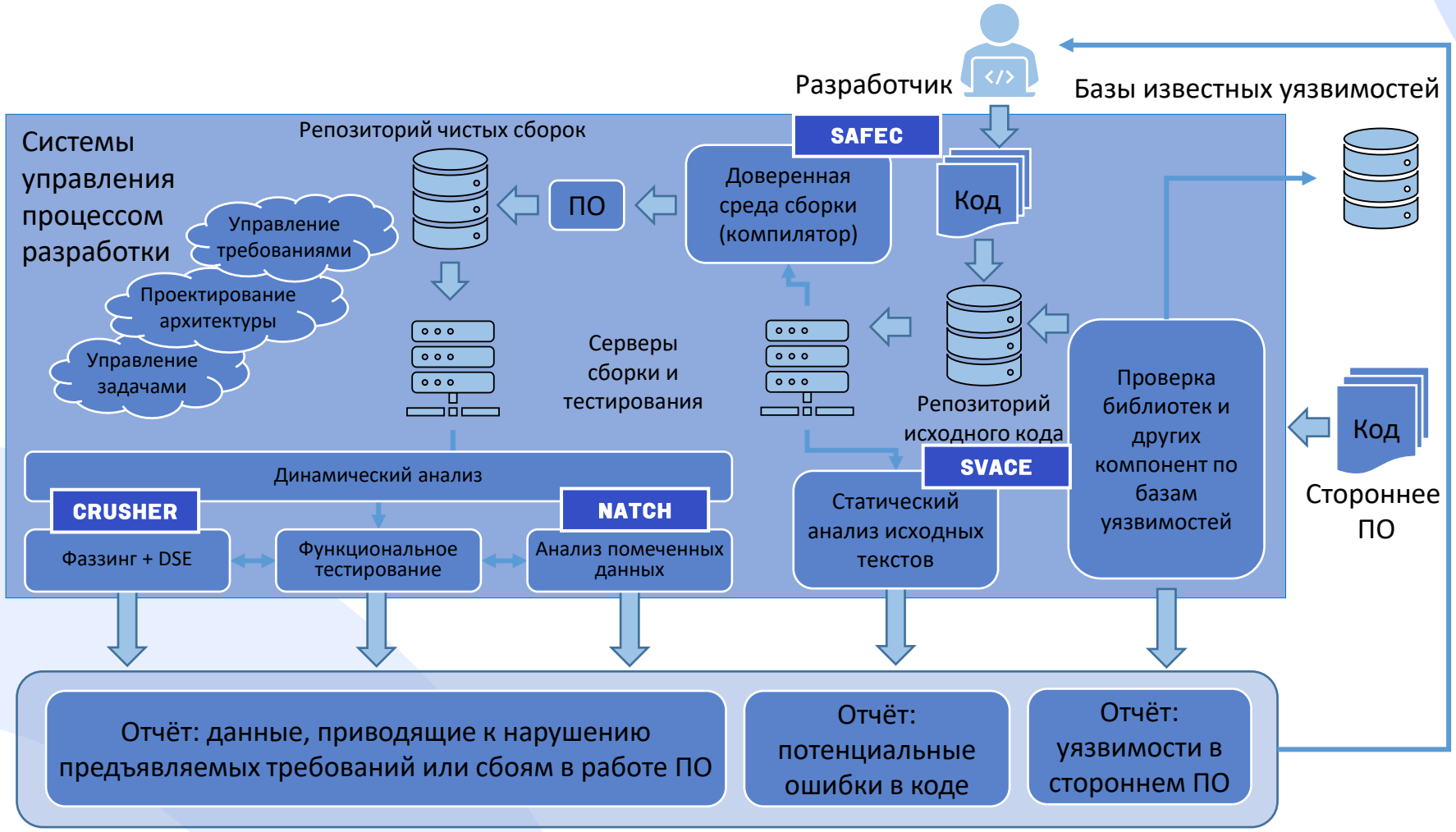
Китай: стандарты по кибербезопасности от национального комитета ТК260, 19 стандартов в 2023

- **Недостаточно использовать классические методы защиты (защита по периметру, проверка доступа, антивирусы и др.)**
- **Необходима разработка новых моделей, методов и технологий в области анализа и трансформации программ**

Найти и устранить максимальное количество уязвимостей в исполняемом коде во время **жизненного цикла разработки безопасного ПО**	**Обеспечить устойчивость** программной системы, затруднив эксплуатацию существующих уязвимостей или смягчив последствия их эксплуатации

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

РБПО: типовой конвейер



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

Технологический стек ИСП РАН развивается с 2002 г. на базе результатов фундаментальных научных исследований по контрактам с зарубежными и российскими компаниями.

РБПО: технологически независимый стек

ИСП РАН

Безопасный компилятор

SAFEC

kaspersky, Postgres PROFESSIONAL, ASTRA LINUX

Аналоги: компоненты компиляторов GCC/Clang, исследовательские компиляторы и рандомизаторы (CompCert C, kcc, Selfrando, Oxymoron, Shuffler)

Инструмент статического анализа

SVACE

SAMSUNG, Postgres PROFESSIONAL, HUAWEI, ..., kaspersky

Аналоги: Klocwork (Perforce, США), Coverity (Synopsys, США), Fortify (MicroFocus (прежде HP), США).
Открытые инструменты: Clang Static Analyzer, SpotBugs

Комплекс динамического анализа

CRUSHER

Postgres PROFESSIONAL, КОД безопасности, kaspersky, ASTRA LINUX

Аналоги: Peach Fuzzer (США, Peach Tech), Synopsys Defensics (Synopsys, США), MAYHEM (ForAllSecure, США).
Открытые инструменты: angr, American Fuzzy Lop, Driller (США)

Инструмент определения поверхности атаки

NATCH

kaspersky

Аналоги: инфраструктуры Panda, Decaf (без интроспекции), Panorama (исследовательский инструмент для Windows)

РБПО: перспективы развития*

ИСП РАН

SVACE:

- Использование больших языковых моделей для повышения точности анализа, фильтрации ложных срабатываний, поддержки новых языков
- Поддержка параллельного облачного анализа для масштабирования до анализа миллиардов строк кода

SAFEC:

- Интеграция технологий защиты и санитайзеров в компиляторы GCC/Clang
- Автоматизация проверки соответствия кода классам безопасности при помощи компиляторных технологий и ИИ

CRUSHER:

- Глубокий фаззинг сложных нейросетей и больших языковых моделей
- Использование ИИ для повышения качества мутаций и понимания документации для помощи фаззингу

NATCH:

- Автоматическое определение важнейших функций и частей сложных структур данных для интеллектуального фаззинга
- Внедрение в облако с интеграцией со статическим и динамическим анализом

**Без развития и передовых исследований технологии будут неконкурентоспособны*

Новая научная специальность ВАК «Кибербезопасность»

ИСП РАН

2018: принято решение Президиума РАН о новом научном направлении
2021: новая специальность ВАК «Кибербезопасность» утверждена Приказом Минобрнауки №118 (2021 г.)

Направления исследований:

- Анализ и систематизация уязвимостей
- Моделирование политик информационной безопасности, угроз и атак
- Методы, алгоритмы и средства пострелизного глубокого анализа защищенности ПО
- Методы интеграции средств защиты на уровне аппаратуры и на уровне ПО
- Интеллектуальный масштабируемый мониторинг инцидентов безопасности в распределенных программно-аппаратных системах
- Масштабируемые средства интеллектуального анализа данных и процессов в распределенных системах

И другие

Новый вызов: искусственный интеллект

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

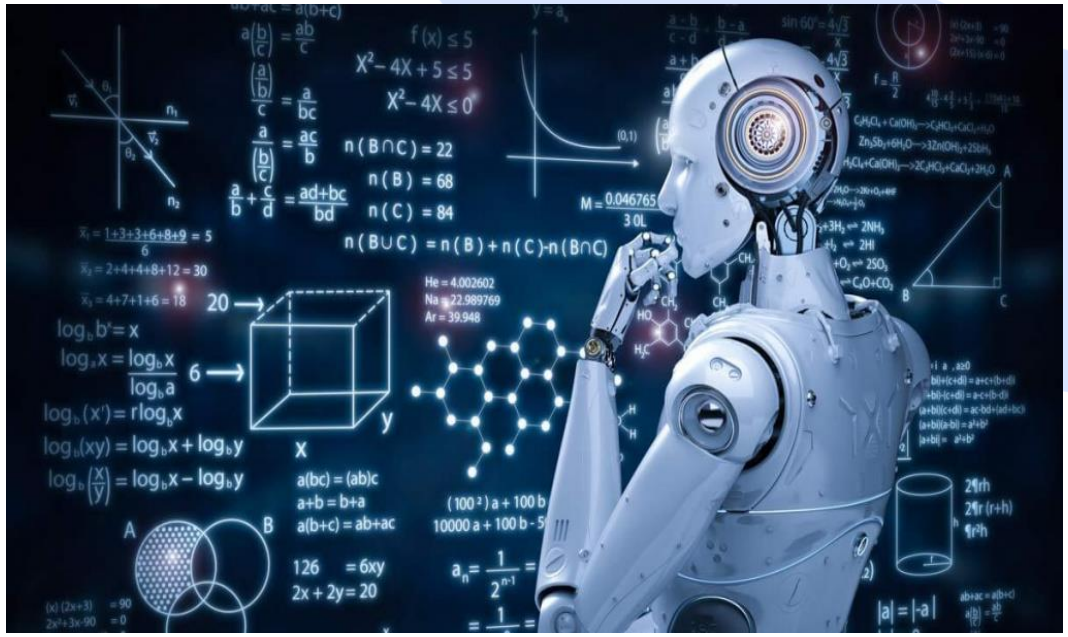
2011 – IBM Watson выиграл шоу Jeopardy! («Своя игра»)

2011 – персональный ассистент в смартфоне (Siri)

2016 – AlphaGO выиграла у профессионального игрока в Го

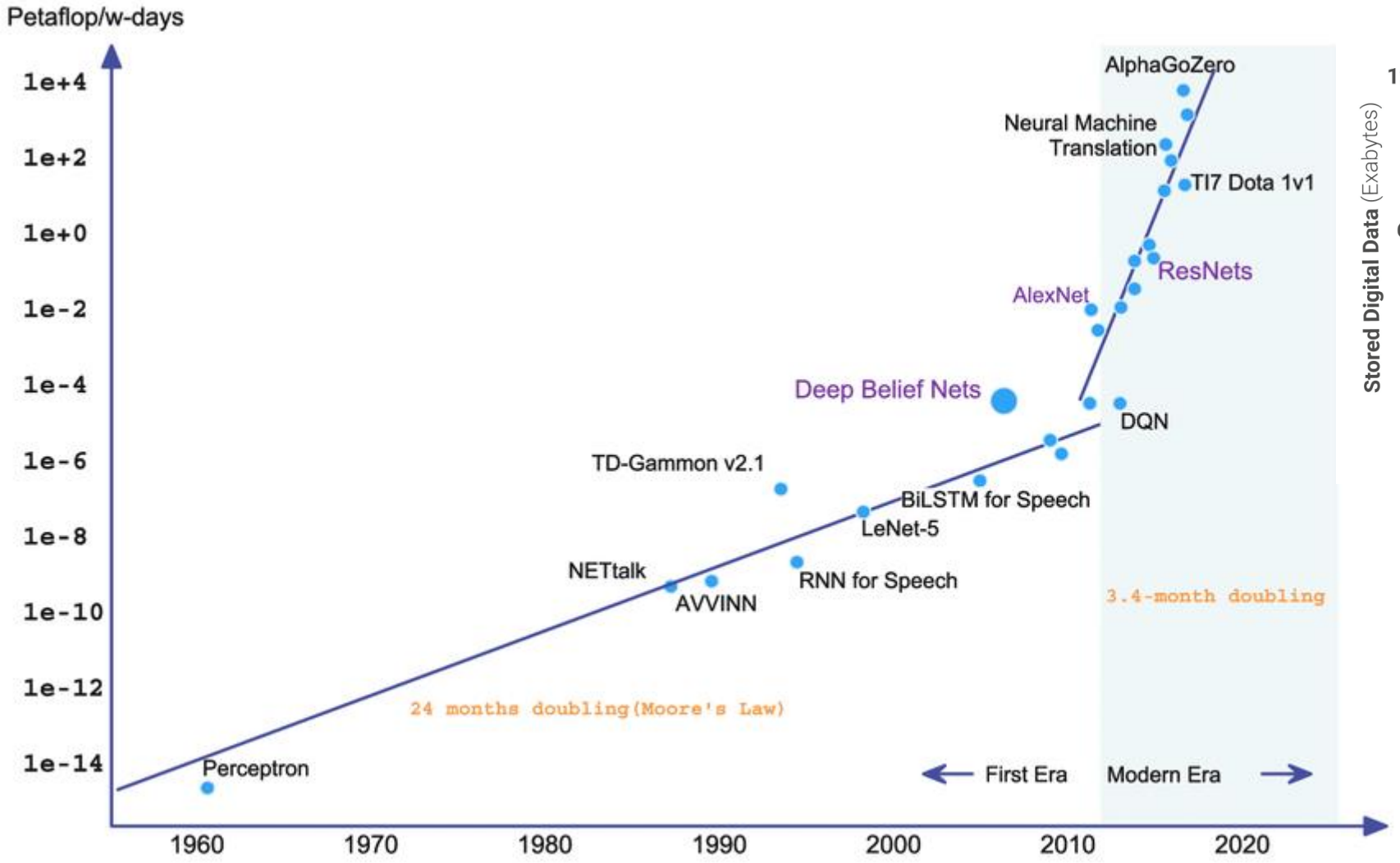
2016 – Google Translate начинает использовать нейронный машинный перевод для 8 языков

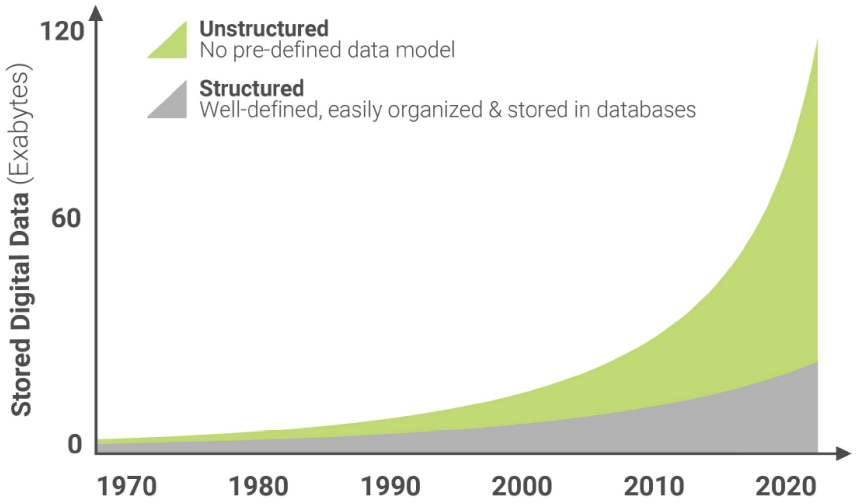
2022 – по н.в.: появление и развитие «больших моделей»: Open AI ChatGPT, YandexGPT2, RuGPT3 (Сбер) и другие.



Переход от моделирования к решению задач по аналогии

Вычислительные ресурсы и большие данные: двигатели развития ИИ





- **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

Принципиальная значимость математики в развитии ИИ:
теорема А.Н. Колмогорова

ИСП РАН

$I^n = [0,1]^n$ – n-мерный куб,
$n = 1, 2, \ldots$
$C(I^n)$ – соответствующее пространство непрерывных функций

Теорема (А.Н. Колмогоров, ДАН СССР, Т. 114, № 5 (1957) 953-956)

При любом $n = 2, 3 \ldots$ существуют такие непрерывные на $[0,1]$ функции $\psi^{p,q}(x)$, $1 \leqslant p \leqslant n$, $1 \leqslant q \leqslant 2n+1$, что каждая функция n переменных $f(x_1, \ldots, x_n) \in C(I^n)$ представима в виде

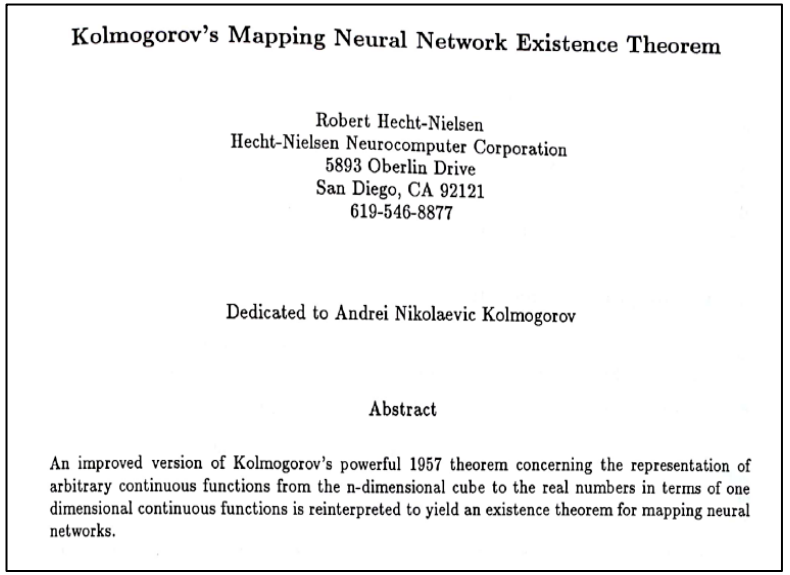
$$f(x_1, \ldots, x_n) = \sum_{q=1}^{2n+1} \chi_q \left(\sum_{p=1}^{n} \psi^{p,q}(x_p) \right),$$

где $\chi_q \in C(\mathbb{R})$, $1 \leqslant q \leqslant 2n+1$.

В Трудах первой научной конференции по нейрокомпьютингу (США, 1987) была опубликована важная статья «Теорема Колмогорова о представимости непрерывных функций нейронными сетями»

Автор - Robert Hecht-Nielsen (1947-2019),

профессор университета Калифорнии (Сан-Диего), один из пионеров практического использования нейросетей для вычислений, автор первого учебника по теме «нейрокомпьютинг», лауреат премий IEEE Neural Networks Pioneer Award и International Gabor Award of the International Neural Network Society



Kolmogorov's Mapping Neural Network Existence Theorem

Robert Hecht-Nielsen
Hecht-Nielsen Neurocomputer Corporation
5893 Oberlin Drive
San Diego, CA 92121
619-546-8877

Dedicated to Andrei Nikolaevic Kolmogorov

Abstract

An improved version of Kolmogorov's powerful 1957 theorem concerning the representation of arbitrary continuous functions from the n-dimensional cube to the real numbers in terms of one dimensional continuous functions is reinterpreted to yield an existence theorem for mapping neural networks.

«Теорема Колмогорова – поразительная, очень мощная, и большинство математиков, впервые о ней узнав, не могут поверить, что она и правда работает. Несмотря на это, ***она оказалась не слишком полезной для доказательства других важных теорем****. Как говорят математики,* ***она не нашла своего использования****. И цель моей статьи – показать, что в нейронных вычислениях это совершенно не так!»*

Принципиальная значимость математики в развитии ИИ:
теорема А.Н. Тихонова и другие исследования

ИСП РАН

ИИ – это аппроксимация функции отображения некоторого множества в другое, придуманной естественным интеллектом

Математическое обоснование существования сходимости и устойчивости таких аппроксимаций для конечномерных множеств – **теорема А.Н. Тихонова о неподвижных точках отображения на упорядоченных ограниченных множествах**

В основе успеха ChatGPT – приложение этой теоремы:

поиск на множестве из тысяч произведений реальных авторов (большие данные) + метрика близости фрагментов этих произведений к заданному

Проблема построения аппроксимации функции отображения бесконечномерного пространства в конечномерное относится к категории некорректных задач, решение которых либо отсутствует, либо множественно, либо неустойчиво.

! **Существенное снижение требований к производительности компьютера при вычислении значения аппроксимирующей функции**

Современные исследования:

Всемирный конгресс «Теория систем, алгебраическая биология, искусственный интеллект: математические основы и приложения», Москва, РАН, 2023

«Математические проблемы создания искусственных нейронных сетей и ИИ», академик РАН В.Б. Бетелин, проф. В.А. Галкин

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

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)

Регулирование: Европа, США

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

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)

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

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)

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

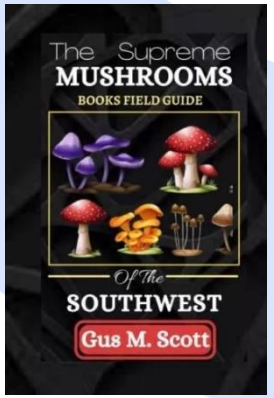
?

Пример (Reuters, 2018): в Amazon создали модель для выбора кандидатов на должности разработчиков. Однако потом выяснили, что система не оценивает кандидатов 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



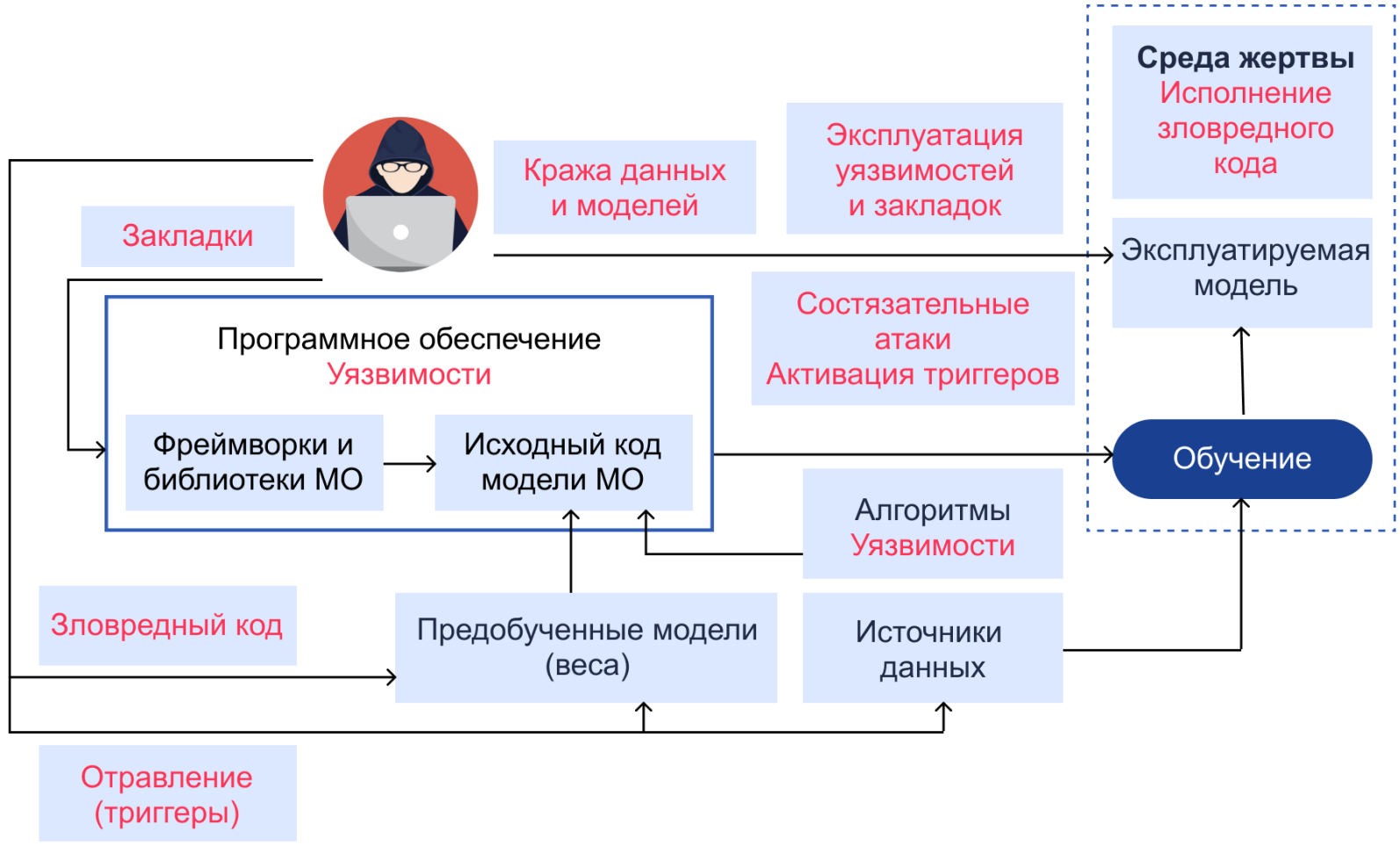
Пример (The Guardian, 2023): в продаже появились книги по сбору грибов, написанные ChatGPT. Специалисты не рекомендуют грибникам их использовать, т.к. в книгах есть ошибки.

СЛАБЫЙ ИИ

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).

Отравленные данные: вставка закладок

Целевая метка

Триггер

выбираются атакующим

$$\theta' = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[(1-\lambda)\cdot\mathcal{L}(F(x;\theta),y) + \lambda\cdot\mathcal{L}(F(x+t;\theta),y_T)]$$

Атака №1: Заставить модель присвоить заведомо ложную целевую метку при наличии триггера во входных признаках

https://arxiv.org/abs/1708.06733

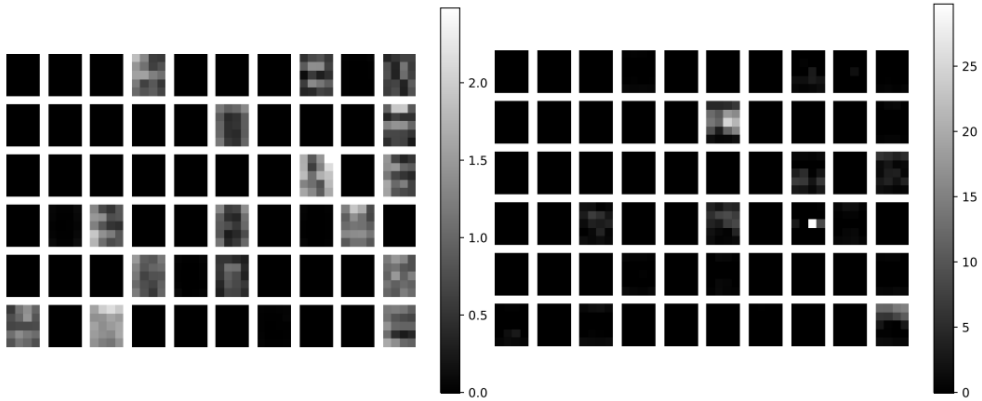
Свойства **успешных** атак через закладки:

- хорошая точность на чистых тестовых данных
- сбой модели на отравленных данных
- атака незаметна: малая доля отравленных данных, малый размер триггера

$$\mathbf{p} = \underset{\mathbf{x}}{\operatorname{argmin}} \; \|f(\mathbf{x}) - f(\mathbf{t})\|_2^2 + \beta \|\mathbf{x} - \mathbf{b}\|_2^2$$

Атака №2: Заставить модель присвоить выбранную целевую метку (*рыба*) на заданном множестве целей из другого класса (*собака*) https://arxiv.org/abs/1804.00792

Защита от атак через закладки



(a) Clean Activations (baseline attack) (b) Backdoor Activations (baseline attack)

Чистка зловредных нейронов https://arxiv.org/abs/1805.12185

- По отравленным тренировочным данным найти редко активируемые нейроны (скорее всего они кодируют триггер)

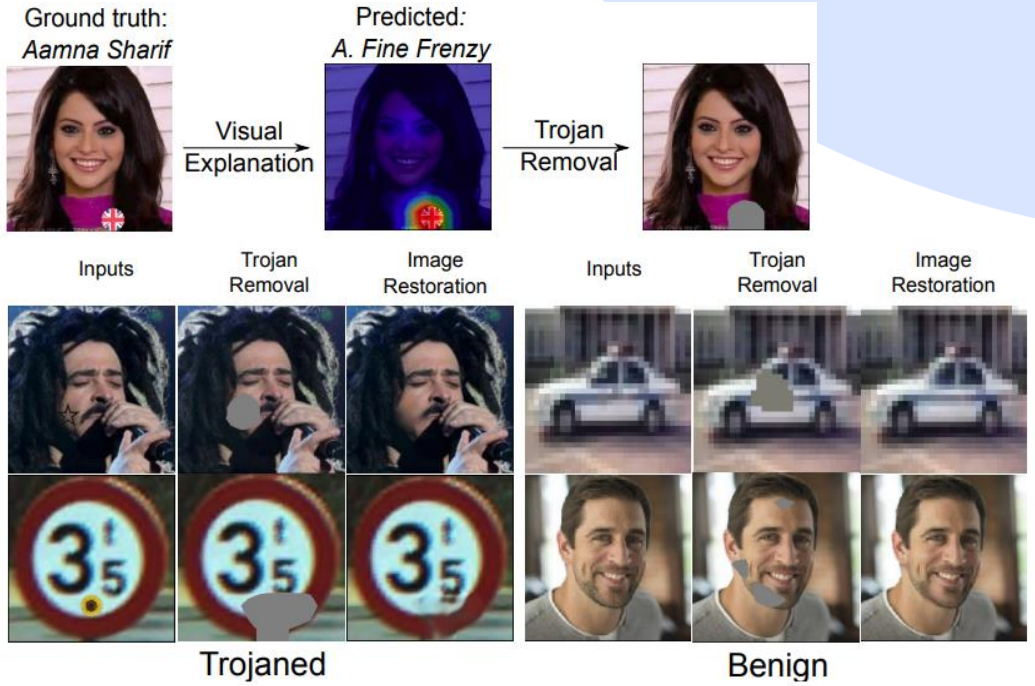
Наш результат:

- Новая атака через закладки на глубокие классификаторы текста через переупорядочивания слов (доклад на AINL 2023)
- защитные алгоритмы от незаметного отравления картинок (ведутся исследования)

Обнаружение и **удаление** триггера через объяснимый ИИ

- затем применение генеративной модели для восстановления изображения

https://arxiv.org/abs/1908.03369

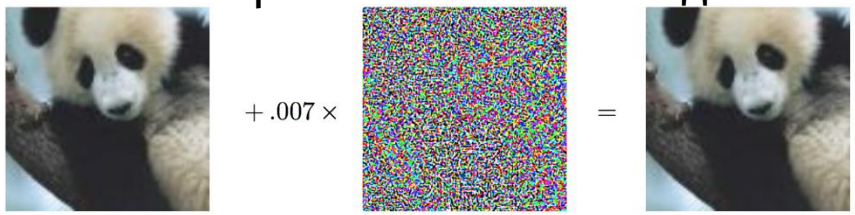


...и другие подходы

Модели машинного обучения: состязательные атаки I

ИСП РАН

Незаметная градиентная атака на классификаторы изображений: максимизация ошибки через изменение входа

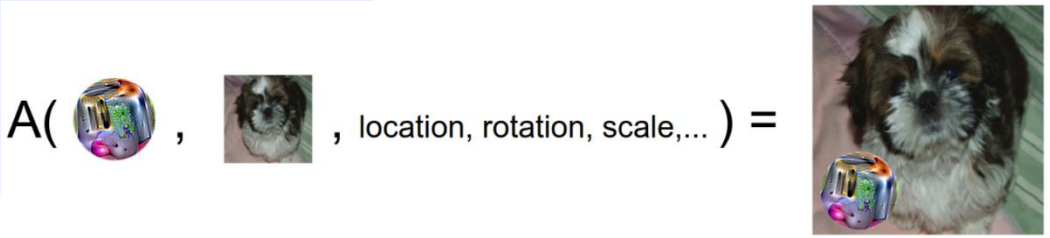


$$x' = \text{Proj}_{[0,1]^d}(x + \epsilon\text{sign}(\nabla_x J(\theta, x, y)))$$ без целевого класса

$$x' = \text{Proj}_{[0,1]^d}(x - \epsilon\text{sign}(\nabla_x J(\theta, x, y')))$$ направленная (на класс y')

$$x_{k+1} = \text{Proj}_{B_\epsilon \cap [0,1]^d}(x_k + \alpha\text{sign}(\nabla_x J(\theta, x_k, y)))$$ **итерационная** (сильная)

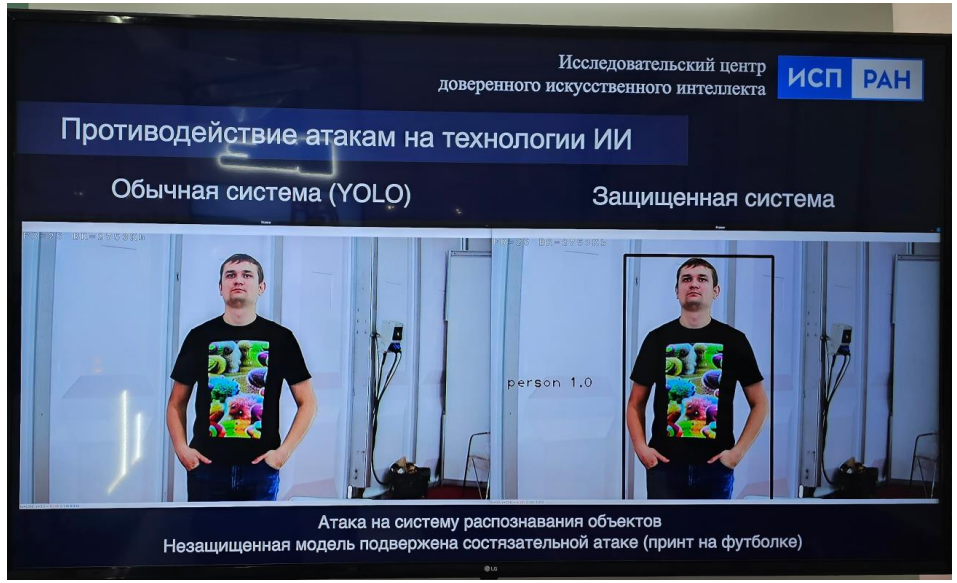
Состязательные патчи: выучить универсальный зловредный стикер



$$\hat{p} = \arg\max_p \mathbb{E}_{x\sim X, t\sim T, l\sim L}\left[\log \Pr(\hat{y}|A(p, x, l, t))\right]$$

изменения алгоритма →

Состязательные футболки
(справа реализована защита)



Модели машинного обучения: состязательные атаки II

ИСП РАН

Возможность переноса состязательных примеров на другие модели и наборы данных, в результате – возможность атак черного ящика без запросов

Как защититься?

- Сложная исследовательская задача
- Современная защита: состязательная тренировка + методы сглаживания ошибок

$$\min_{\theta} \sum_{i=1}^{N} \max_{\|\delta\|_p \leq \epsilon} \mathcal{L}(f_\theta(x_i + \delta), y_i)$$

Устойчивая точность по ImageNet (апрель 2023): 59.56% (https://robustbench.github.io/)

Наш результат:

- новая атака на определение объектов через генеративные состязательные сети и диффузионные модели (доклад на AINL 2023)
- состязательная тренировка для защиты от состязательных атак на основе оценки качества через глубокие нейронные сети (IQA, ведутся исследования)

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)
…

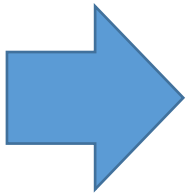
ПРОБЛЕМА
Отсутствие общей среды для прозрачного одновременного использования разных инструментов

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 контент с целью защиты авторских прав и подтверждения целостности самого документа.

Федеративное обучение: проблема I
объединение приватных данных

ИСП РАН

Классическое обучение

- Локально на своих данных
- Распределено на открытых данных

WIKIPEDIA
IMAGENET

Залежи приватных данных на пользовательских устройствах **не используются**

Решение:
федеративное обучение – объединить приватные данные в одном ИИ

Виды

Горизонтальное:
Одинаковая инф-ция про разные объекты

Единая модель МО

Вертикальное:
Разная инф-ция про одинаковые объекты

Единая модель МО

Перспективы

- Объединение **разной информации** в рамках одного ИИ
- Огромные объёмы **новых уникальных данных** для обучения ИИ
- **Персонализация:** каждому клиенту своя уникальная модель ИИ

Коммуникации: проблема II
необходимо эффективное общение

ИСП РАН

Коммуникации – **бесполезная трата времени** с точки зрения прогресса обучения ИИ

↓

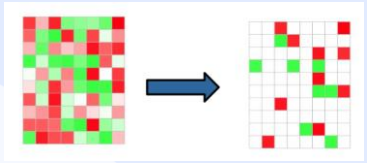
Могут занимать около **30-95%** времени обучения

↓

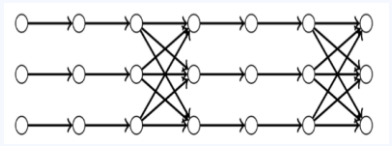
Организация коммуникации – **борьба за быстрое обучение** современного ИИ

Способы

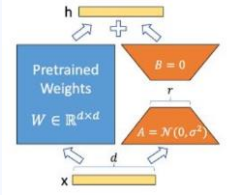
Сжатие посылок



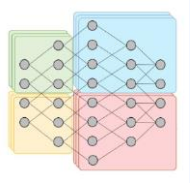
Пропуск раундов общения



Временное упрощение модели



Разделение модели на части



Уменьшение коммуникационных затрат до **15 раз**, ускорение процесса обучения до **5 раз**

Федеративное обучение: научные школы в России

ИСП РАН

Жёсткие требования к стоимости коммуникаций (т.е. к объему передаваемых данных) и конфиденциальности информации клиентов – две основные проблемы в федеративном обучении.

В этих вопросах, а также при оптимальной квантизации сигналов активно используются результаты из многомерной геометрии, полученные акад. Б.С. Кашиным еще в 70-х годах прошлого века, в частности, так называемое **«Kashin representation»**.

См., например,

W.N. Chen, P. Kairouz, A. Ozgun «Breaking the Communication – Privasy – Acuracy Trilemma»,

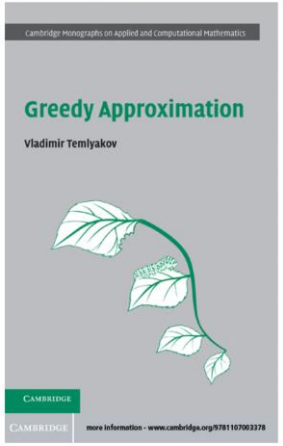
IEEE Transactions on Information Theory, v. 69, № 2, 1261-1281 (2022)

M. Safaryan, E. Shulgin, P. Richtarik «Uncertainty principle for communication compression in distributed and federated learning and the search for an optimal compressor», Information and Inference:

A Journal of the IMA, v. 11, 557-580 (2022)

Важную роль в решении многих задач машинного обучения играют так называемые **жадные алгоритмы**. Российские математики занимают в этой области передовые позиции.

Автором единственной монографии по жадным алгоритмам и жадным приближениям является заведующий недавно созданным в Математическом институте им. В.А. Стеклова РАН отделом математических основ искусственного интеллекта профессор В.Н. Темляков



Другие исследования

Beznosikov, A., Horváth, S., Richtárik, P., & Safaryan, M. (2023). On biased compression for distributed learning. Journal of Machine Learning Research, 24(276), 1-50.

Gorbunov, E., Rogozin, A., Beznosikov, A., Dvinskikh, D., & Gasnikov, A. (2022). Recent theoretical advances in decentralized distributed convex optimization. In High-Dimensional Optimization and Probability: With a View Towards Data Science (pp. 253-325). Cham: Springer International Publishing.

Metelev, D., Rogozin, A., Kovalev, D., & Gasnikov, A. (2023, July). Is consensus acceleration possible in decentralized optimization over slowly time-varying networks?. In International Conference on Machine Learning (pp. 24532-24554). PMLR.

Цифровые водяные знаки (Watermarking),
технология маркирования

ИСП РАН

Для синтезированного контента

- В 2023 семь ведущих компаний в области ИИ (OpenAI, Meta.Platforms, Alphabet и др.) обязались разработать систему цифровых водяных знаков для всех форм синтезированного контента [1]
  - Обнаружение сгенерированного текста [2], обнаружение дипфейков в аудио-/видео-потоках и изображениях [3]
  - Математическое обоснование: устойчивости ЦВЗ к преобразованиям над контентом, вероятности извлечения ЦВЗ

Для глубоких нейронных сетей и обучающих наборов данных

- Защита обученных нейросетевых моделей от анонимной кражи [4, 5]
  - Интеллектуальная собственность в сфере обученных нейронных сетей
- Защита обучающих наборов данных от анонимной кражи [6]
  - Интеллектуальная собственность в сфере наборов данных (*dataset watermarking*) – установить/опровергнуть факт обучения нейронной сети на заданном наборе данных

- **Для контента, созданного человеком (защита авторских прав и отслеживание распространения)**
  - Изображения [7], аудио- [8] и видео-потоки [9]
    - Дискретные Фурье, косинусное , вейвлет преобразования (DFT/DCT/DWT), модели архитектуры «кодер-декодер»
  - Предотвращение анонимных утечек документов – DocMarking [10]
    - Статистические особенности изображений документов, вероятностные оценки точности извлечения ЦВЗ

[1] SecurityLab.ru
[2] Can ai-generated text be reliably detected? (Sadasivan V. S. et al., 2023)
[3] Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data (Yu N. et al., 2021)
[4] Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring (Adi Y. et al., 2018)
[5] A survey of deep neural network watermarking techniques (Li Y., Wang H., Barni M., 2021)
[6] Open-sourced Dataset Protection via Backdoor Watermarking (Li Y. et al., 2020)
[7] Reading watermarks with a camera phone from printed images (Pramila, 2018) [VPN]
[8] Hybrid blind audio watermarking for proprietary protection, tamper proofing, and self-recovery (Hu H. T., Lee T. T., 2019)
[9] Implementation-Free Forensic Watermarking for Adaptive Streaming with A/B Watermarking (Mareen H., Van Wallendael G., Lambert P., 2022)
[10] DocMarking: Real-Time Screen-Cam Robust Document Image Watermarking (Yakushev A. et al., 2023)

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

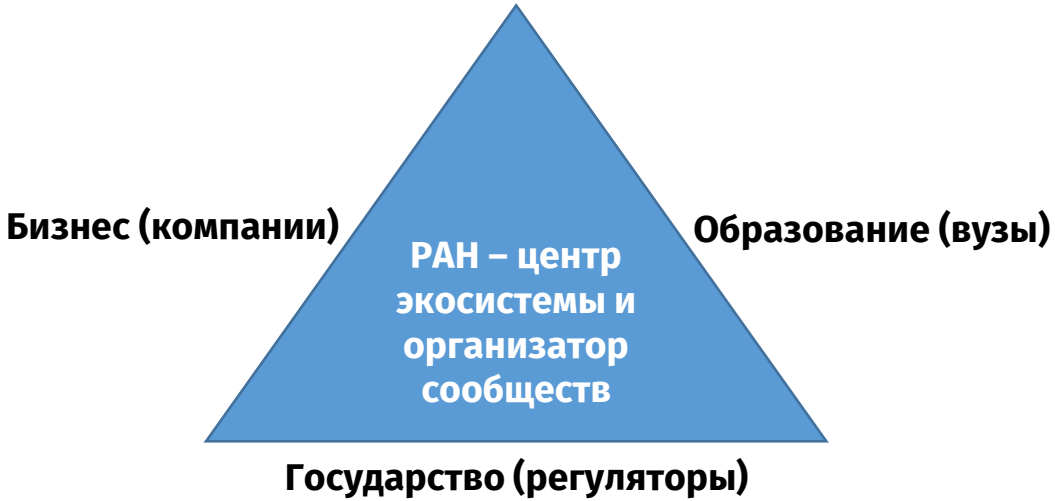
Проблемы

~~Технологические риски~~

~~Кадровые риски~~

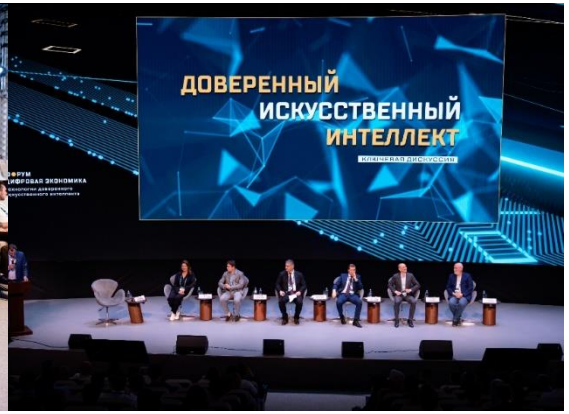
~~Политические риски~~

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!

300 лет
Российская Академия Наук

ИСП РАН