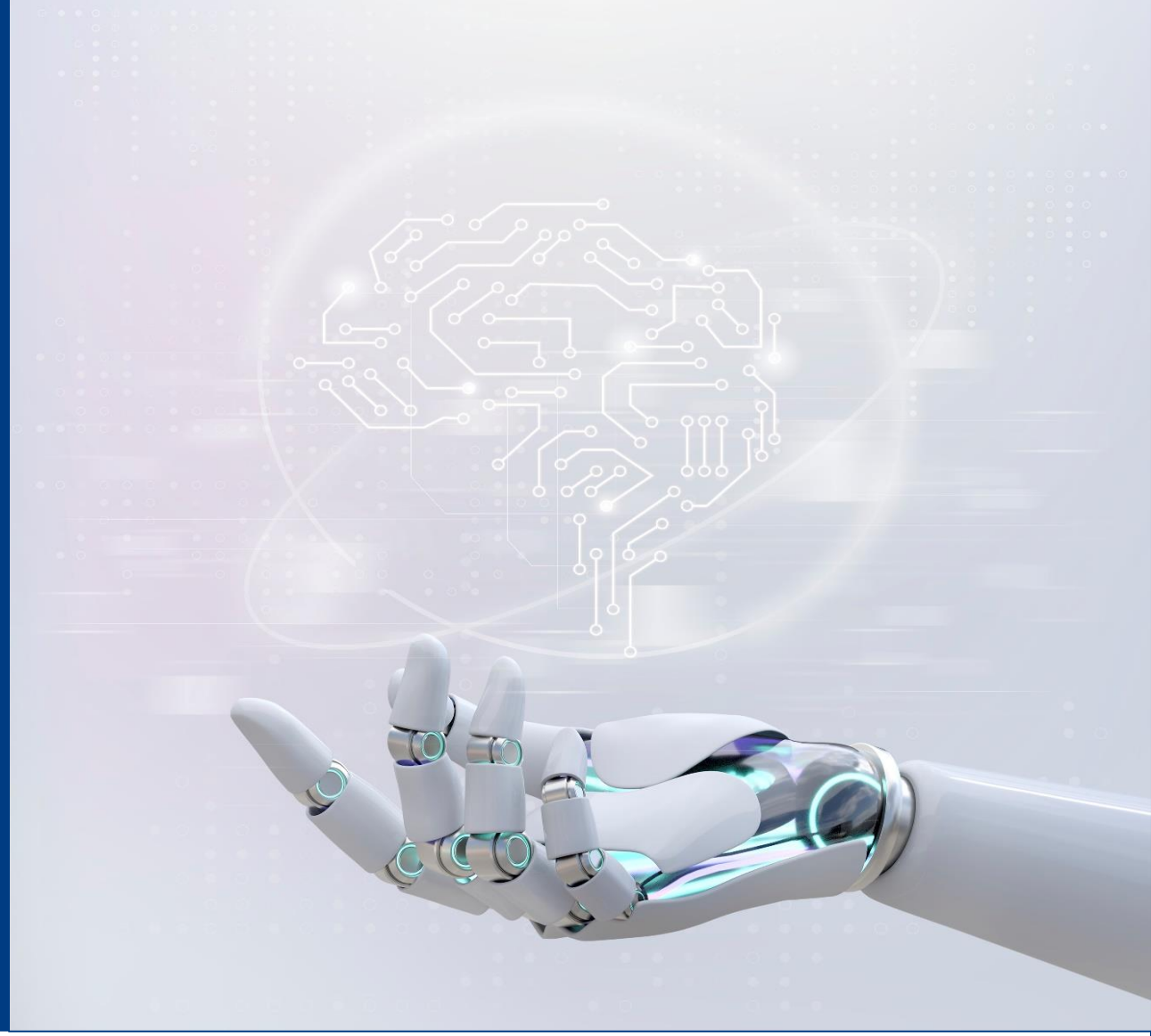
Доверенный искусственный интеллект: современные задачи и направления развития

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



ИСП РАН

11 февраля 2025 года

CONFERENCES

Искусственный интеллект (ИИ) внедряется во многих отраслях

ИСП РАН

«Искусственный интеллект – комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые с результатами интеллектуальной деятельности человека или превосходящие их»

Национальная стратегия развития ИИ на период до 2030 года

Большие языковые модели; «умные» приложения

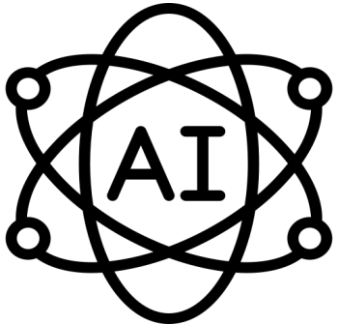
Генеративный ИИ, системы машинного перевода, голосовые помощники в смартфонах

Медицина

Компьютерная диагностика, подбор лечения; фитнес-браслеты и другие устройства

Транспорт

Беспилотные автомобили



Системы безопасности

Распознавание лиц с помощью компьютерного зрения

Финансы

Обнаружение мошенничества и отмывания денег, кредитный скоринг, чат-боты

Исследование космоса

Автономная космическая навигация (роботы на Марсе)

Торговля

Рекомендации в ритейле, роботизация складского бизнеса

Промышленность

Роботизация производства

И будет внедряться дальше!

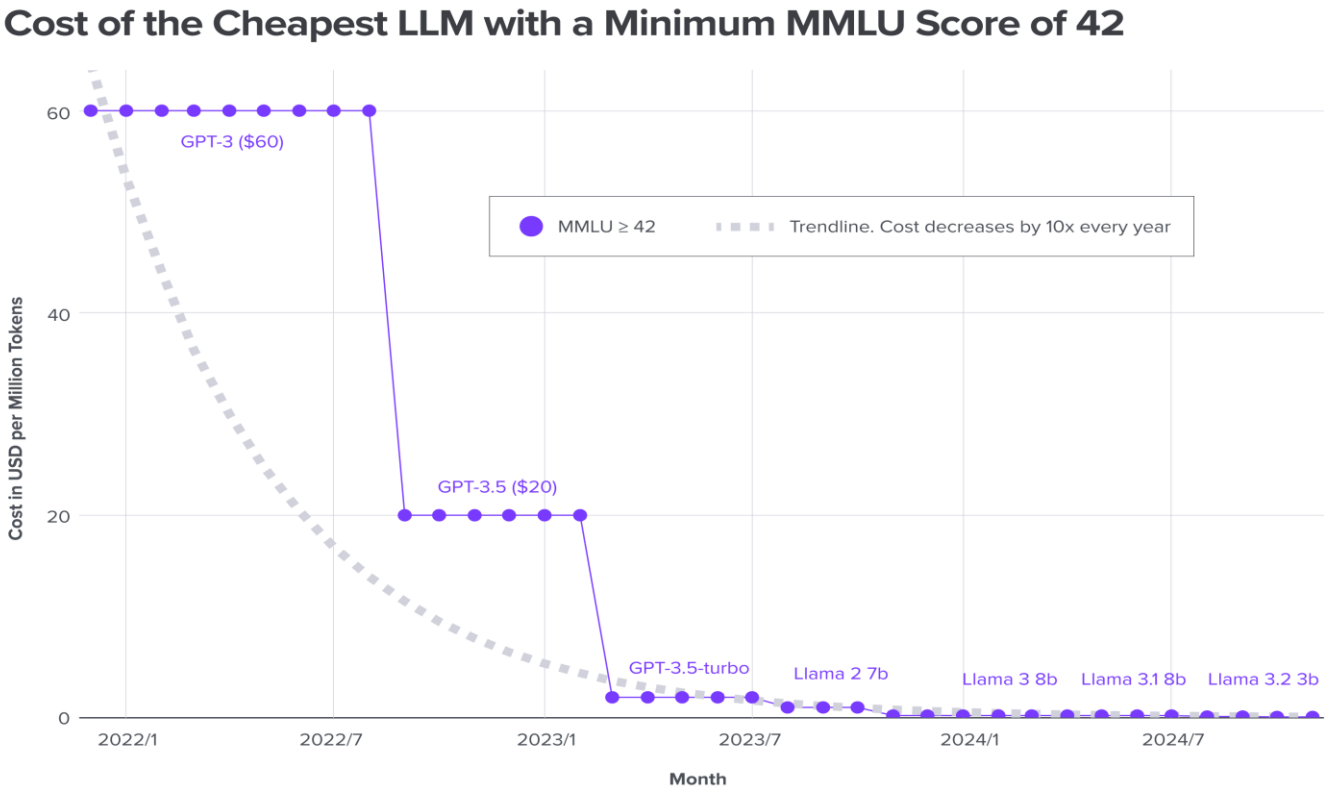
ИСП РАН

ИИ окружает нас почти везде

Он становится всё дешевле и умнее

А значит, он будет всё активнее распространяться и влиять на нашу жизнь

Пример снижения стоимости



https://a16z.com/llmflation-llm-inference-cost/

Одна из ключевых проблем ИИ – проблема доверия

ИСП РАН

ВЦИОМ, декабрь 2024: «Доверяете ли вы искусственному интеллекту?»

52% доверяют ИИ

Почему?

- можно передать ИИ опасные для человека виды работ (34%)
- улучшение и упрощение жизни (33%)
- объективность, беспристрастность ИИ (32%)
- меньшая вероятность ошибок в сравнении с человеком (23%)
- скорость и качество работы в сравнении с человеком (22%)

38% не доверяют ИИ
(это на 6% больше, чем в 2022)

Почему?

- сбои и ошибки в работе ИИ (28%)
- возможный выход ИИ из-под контроля человека (26%)
- возможность его использования в корыстных целях (23%)
- риск утечки данных, собираемых ИИ (21%)
- деградация населения, вызванная развитием ИИ (20%)

https://wciom.ru/analytical-reviews/analiticheskii-obzor/doverie-k-ii

Многие из этих опасений оправданы!

Доверять ИИ недостаточно
Нужно знать, почему мы доверяем

ИСП РАН

Необходимо обеспечить доверенность с двух сторон:

СО СТОРОНЫ КИБЕРБЕЗОПАСНОСТИ
проблемы разработки,
атаки, закладки и проч.

С СОЦИОГУМАНИТАРНОЙ СТОРОНЫ
проблемы честности генеративного ИИ,
манипуляция общественным мнением и сознанием отдельного человека и т.д.

ДЛЯ ВСЕГО ЭТОГО НУЖНЫ СВОИ ИНСТРУМЕНТЫ И МЕТОДЫ!

И контроль за решениями ИИ: нельзя позволять ему принимать финальные решения там, где от этого зависят жизнь и здоровье людей

«Доверенные технологии искусственного интеллекта - технологии, отвечающие стандартам безопасности, разработанные с учетом принципов объективности, недискриминации, этичности, исключающие при их использовании возможность причинения вреда человеку и нарушения его основополагающих прав и свобод, нанесения ущерба интересам общества и государства».

Национальная стратегия развития искусственного интеллекта на период до 2030 года в редакции Указа Президента РФ от 15.02.2024 № 124

«Несмотря на многочисленные обсуждения этики и принципов работы ИИ, общая картина норм, институтов и инициатив всё еще находится в зачаточном состоянии и полна пробелов. Сейчас ИИ объединяет глобальные вызовы и возможности, которые требуют целостного подхода на пересечении политики, экономики, социологии, этики, юриспруденции, экологии, техники и других областей. Такой подход может превратить разнообразные развивающиеся инициативы и подходы в единое целое...»

Отчёт ООН Governing AI for Humanity 2024

В СЛУЧАЕ ИИ КИБЕРБЕЗОПАСНОСТЬ – ТОЛЬКО ЧАСТЬ ДОВЕРЕННОСТИ

Кибербезопасность: пример реализации угроз

ИСП РАН

Аварии с участием беспилотных автомобилей

2 октября 2023 в Калифорнии обычный автомобиль с водителем за рулём сбил пешехода

Пешехода отбросило под колёса беспилотного автомобиля Cruise, который тоже сбил его, остановился, но потом снова поехал

Человек был зажат под колесом и получил серьёзные травмы, т.к. Cruise проехал таким образом еще 6 метров

В ноябре 2023 года 950 машин Cruise были отозваны для обновления ПО. Ситуация привела к отставке основателя Кайла Фогта, увольнению 9 руководителей и сокращению 25% сотрудников

В декабре 2024 General Motors объявила о прекращении разработки роботизированных такси Cruise. Компания так и не смогла оправиться от происшествия 2023 года

https://www.theguardian.com/technology/2023/nov/08/cruise-recall-self-driving-cars-gm

https://www.siliconvalley.com/2024/09/20/gms-cruise-to-resume-robotaxi-testing-in-california-this-fall/

https://www.theguardian.com/us-news/2024/dec/11/general-motors-self-driving-cruise-robotaxi

2023

General Motors pulls plug on Cruise, its self-driving robotaxi company

2024

The company said it would no longer fund the venture and will prioritize Super Cruise, its driver assistance program

Социогуманитарная сфера: пример реализации угроз I

ИСП РАН

Мошенничества с дипфейками

Началось всё еще в 2019 с аудио

Мошенник подделал голос гендиректора материнской компании в Германии, позвонил директору регионального отделения в Великобритании и потребовал срочный перевод €220 тысяч. Платёж был частично переведён на указанный мошенником счёт

https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

Со временем добавилось видео

2024, Гонконг

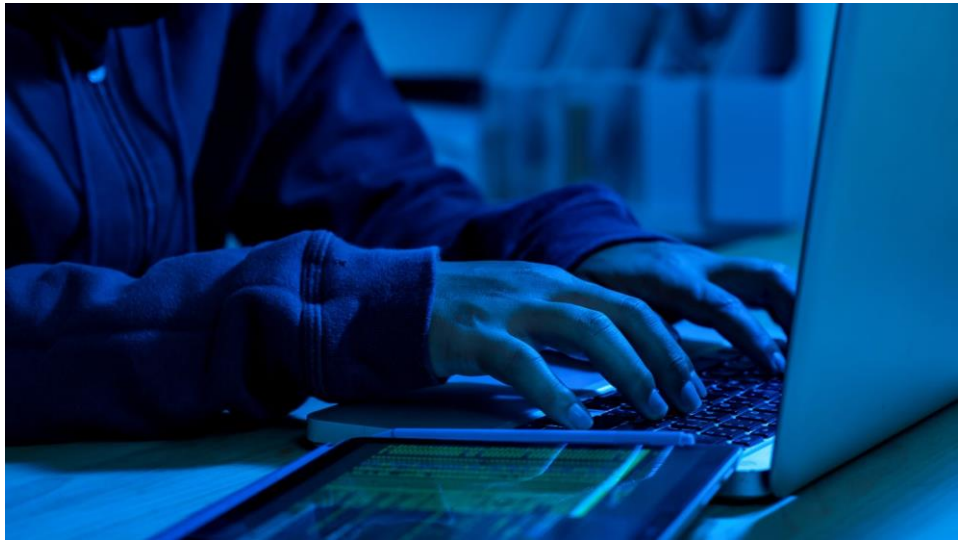
С помощью фейковой видеоконференции преступники вынудили сотрудника транснациональной корпорации перевести им $25,6 млн.

https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

2024, Китай (провинция Шэньси)

Сотрудница финансовой компании перевела $258 тысяч на указанный счёт после видеозвонка с человеком, которого считала своим начальником (голос и внешность совпадали). Перевод удалось экстренно заморозить с помощью сотрудников банка

https://global.chinadaily.com.cn/a/202403/07/WS65e9244ba31082fc043bb278.html



В сентябре 2024 в Госдуму внесли проект о лишении свободы за мошенничество с использованием дипфейков

По подсчёту аналитиков Сбербанка, за 8 месяцев 2024 года количество преступных схем с использованием дипфейков выросло в 30 раз

https://www.kommersant.ru/doc/7157947

Социогуманитарная сфера: пример реализации угроз II

ИСП РАН

Манипуляции
(например, людьми с нестабильной психикой)

По данным ВОЗ:

- **Каждый восьмой человек в мире живет с психическим расстройством (ЭТО БОЛЬШЕ МИЛЛИАРДА ЧЕЛОВЕК!)**
- **Каждый год более 720 тысяч человек совершают суицид**
- **Это третья причина смертности в молодых людей в возрасте от 15 до 29 лет**

https://www.who.int/ru/news-room/fact-sheets/detail/mental-disorders
https://www.who.int/news-room/fact-sheets/detail/suicide/

2023, Бельгия

Бельгиец покончил жизнь самоубийством после шести недель общения об экологических проблемах с чат-ботом на основе генеративного ИИ. Мужчина испытывал сильную тревогу и в беседе с роботом сообщил, что думает о самоубийстве. Робот ответил, что «они будут жить вместе на небесах»; это спровоцировало суицид.

https://www.lavenir.net/actu/belgique/2023/03/28/un-belge-se-donne-la-mort-apres-6-semaines-de-conversations-avec-une-intelligence-artificielle-76MEJ5DBRBEVDM62LTPJJTII4Q

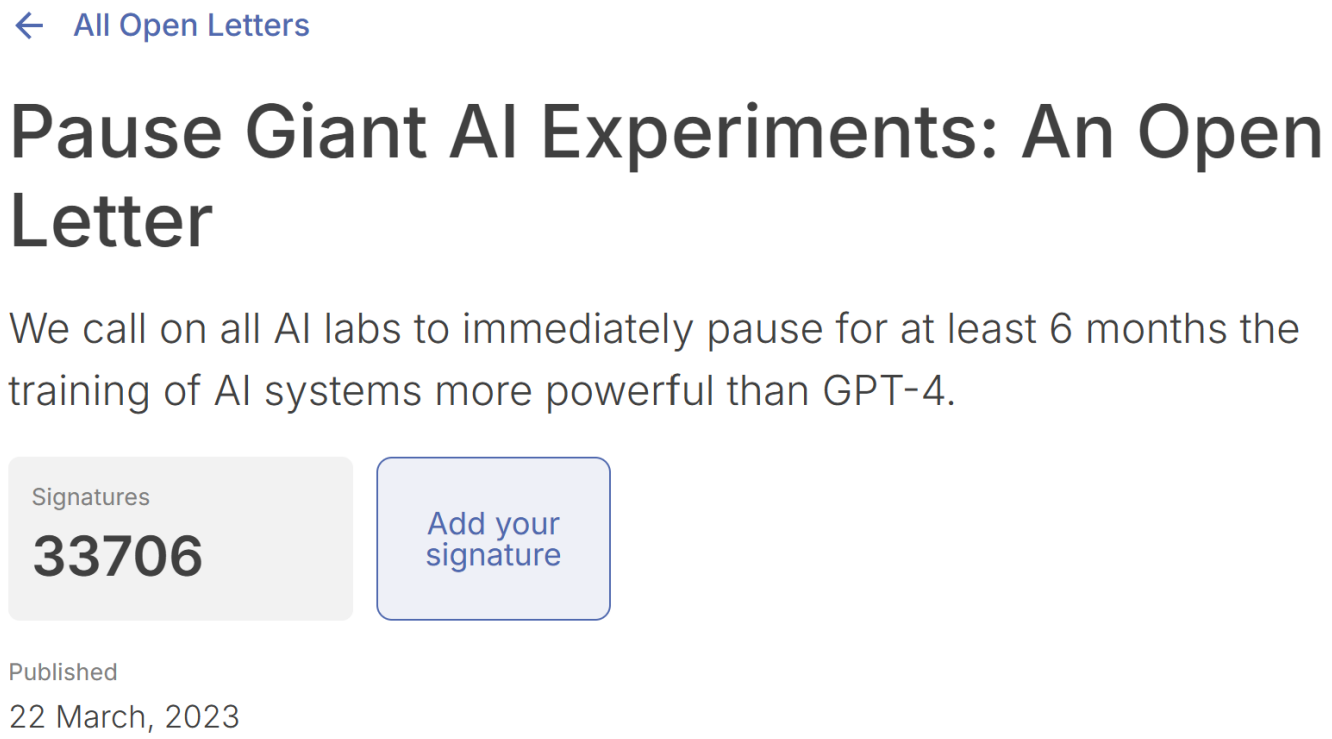
… и это далеко не все угрозы!

Что со всем этим делать?

ИСП РАН

В 2023 генеральный директор Tesla и SpaceX Илон Маск и соучредитель Apple Стив Возняк подписали открытое письмо о необходимости шестимесячного моратория на обучение мощных систем с ИИ

«Исследования и разработки ИИ должны быть переориентированы на то, чтобы сделать самые мощные современные системы более точными, безопасными, интерпретируемыми, понятными, надежными, непротиворечивыми и доверенными»



https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Доверенный ИИ:
с 2023 регулирование в мире активно развивается

ИСП РАН

Основной тренд: ИИ, которому мы сможем доверять (доверенный ИИ)

Постоянный рост числа инициатив

2020

- **White Paper on Artificial Intelligence**: a European approach to excellence and trust

2022

- **AI Bill of Rights** (США)
- **NIST AI RMF,** методика (США)
- **Center for AI Safety (CAIS)**

2023

- **Executive Order on Safe, Secure, and Trustworthy AI** (США)
- **NIST Trustworthy & Responsible AI Resource Center (AIRC),** исследовательский центр (США)
- **Hiroshima AI Process** (G7)
- **ENSIA**, методика (Евросоюз)
- **Временные регуляторные документы про генеративный ИИ** о необходимости пометок контента, а также блокировке зарубежного ИИ-контента, нарушающего требования регуляторики (Китай)

2024

- **Резолюция Генассамблеи ООН по безопасным системам ИИ**
- США и Великобритания заключили **договор о безопасности в сфере ИИ** (первый двусторонний договор в этой сфере)
- **EU AI Act** (некоторые технологии ИИ предлагается запретить, а сгенерированный контент – обязательно маркировать). В его рамках: проект **AI Code of Practice** – требований для разработчиков моделей общего назначения.
- **European AI Office** – для координации работ с ИИ
- **California AI Transparency Act** (аналогичные приняты в Колорадо, Юте, Иллинойсе). Требует, чтобы поставщики генеративного ИИ с посещаемостью более 1 млн человек в месяц предоставляли пользователям бесплатные инструменты, которые определяют, был ли контент сгенерирован ИИ
- **И многое другое!**

Публикационная активность к 2025 году: 3000+ научных статей

Число проектов на GitHub: 2000+

Доверенный ИИ: пример инструмента для поддержки регулирования

ИСП РАН

https://compl-ai.org/

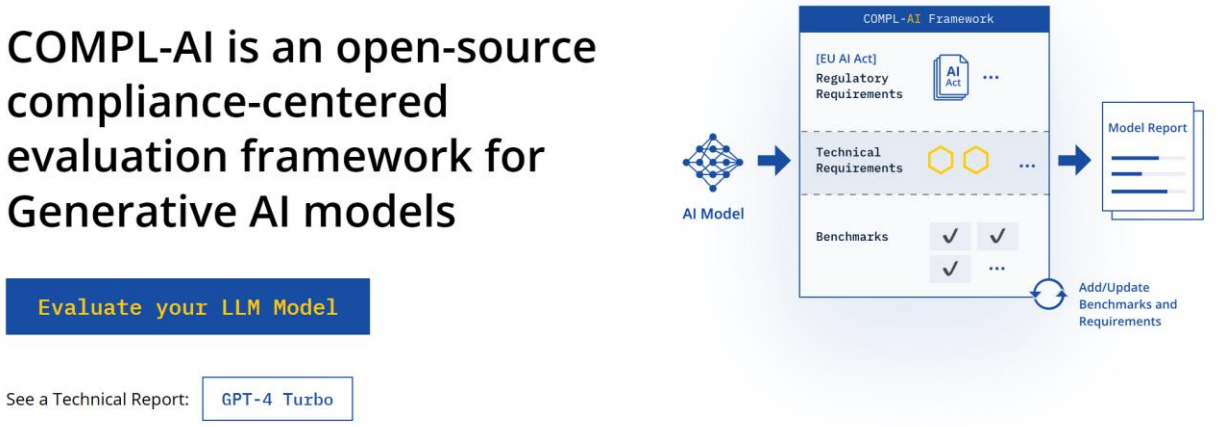
Фреймворк для оценки больших языковых моделей на соответствие EU AI Act – главному европейскому закону об искусственном интеллекте

Разработан **LatticeFlow AI** и **ETH Zurich** (Швейцария) и Институтом компьютерных наук, ИИ и технологий **INSAIT** (Болгария)

Проверяет модели по ряду критериев

Главные проблемы моделей:

- **предвзятость (**OpenAI GPT-3.5 Turbo, Alibaba Cloud Qwen1.5 72B Chat)
- **низкая устойчивость к кибератакам** (Meta Llama 2 13B Chat, Mistral 8x7B Instruct)



European Commission

The Commission welcomes this study and AI model evaluation platform as a first step in translating the EU AI Act into technical requirements, helping AI model providers implement the AI Act.

- Thomas Regnier, Spokesperson, European Commission

«Комиссия поддерживает это исследование и платформу для оценки моделей как первый шаг на пути воплощения требований ЕС к искусственному интеллекту в технические требования, которые помогают поставщикам моделей следовать EU AI Act».

Томас Ренье, спикер Еврокомиссии

Доверенный ИИ: регулирование в России активно развивается

ИСП РАН

2019

Национальная стратегия развития ИИ до 2030 года (обновлена в 2024, именно тогда добавлено определение «доверенных технологий ИИ»

2021

Кодекс этики в сфере ИИ (сейчас объединяет 850 подписантов, в том числе 42 зарубежных участника из 24 стран)

Федеральный проект «Искусственный интеллект»

В его рамках:

- **получил господдержку** Исследовательский центр доверенного искусственного интеллекта ИСП РАН
- **Академия криптографии начала формирование научной базы для современных защищенных технологий и систем ИИ, применяемых в государственных информационных системах**

ГОСТ Р 59525-2021 «Интеллектуальные методы обработки медицинских данных»

2024

При поддержке Минцифры создан **Консорциум для исследований безопасности технологий искусственного интеллекта** (**НТЦ ЦК, Академия криптографии и ИСП РАН**, присоединяются компании и вузы)

- ❑ создание технологий доверенного ИИ
- ❑ разработка криптографических методов его защиты
- ❑ работы по анонимизации данных

Консорциум для исследований безопасности технологий искусственного интеллекта: вклад ИСП РАН

ИСП РАН возглавляет Рабочую группу №4 по разработке безопасных технологий ИИ (председатель – ведущий сотрудник Вартан Падарян)

Участники группы: Минцифры России, ИСП РАН, НТЦ ЦК, Positive Technologies, Ассоциация «ФинТех», «Гарда», Новосибирский государственный технический университет

Цели РГ №4

- ❑ выработка согласованных требований к процессам разработки безопасных технологий ИИ
- ❑ определение состава мер и средств, обеспечивающих выполнение этих требований

Задачи РГ №4

Организационные, в рамках консорциума

- ❑ синхронизация по перечню угроз безопасности ИИ, разработка предложений в части нормативных правовых актов (РГ1)
- ❑ определение порядка тестирования безопасности технологий ИИ (РГ2)

Методические и технологические

- ❑ состав процессов разработки безопасных технологий ИИ
- ❑ технологии и инструменты разработки (анализа), включая безопасные ML-фреймворки
- ❑ оценка требуемых вычислительных ресурсов для реализации требований безопасной разработки
- ❑ требования к применяемым инструментам

2025

Февраль
Стабилизация состава РГ, определение ролей и вклада участников

Апрель
Сведение лучших практик разработки безопасных технологий ИИ (первая редакция)

Июнь
Определение состава инструментальных средств, оценка требуемых вычислительных ресурсов

Август
Пилотирование MLSecOps-платформы на площадке ИСП РАН

Ноябрь
Разработка первой редакции методических указаний

Декабрь
Развёртывание на полигоне конвейера MLSecOps для тестирования безопасности технологий ИИ

Ожидаемые результаты работы

Методические указания по обеспечению безопасной разработки технологий ИИ

Перечень референсных инструментов, обеспечивающих разработку безопасных технологий ИИ

Требования к эталонным датасетам, гарантирующим уровень безопасности предобученных моделей

Исследовательский центр доверенного ИИ ИСП РАН

ИСП РАН

Задачи

- Создание и предоставление разработчикам и операторам систем с ИИ **инструментария для обеспечения требуемого уровня доверия**
- Создание **единой методологии и рекомендаций** по разработке и поддержанию жизненного цикла доверенных систем с ИИ
- Создание **обучающих материалов и учебных курсов** по использованию решений Центра

KASPERSKY
IPC InterProCom
ТЕХНОПРОМ
ЕС-ЛИЗИНГ

Результаты синхронизируются с ФСТЭК России и используются при подготовке ГОСТов

Публикационная активность Центра к концу 2024: 70+ статей по темам доверенного ИИ (A*/Q1)

Планы на 2025-2027

Проводить дальнейшие исследования и разработку методов по защите моделей

Передавать в прикладные отрасли (медицину и др.) готовые продукты для обеспечения необходимого уровня доверия

Развивать анализ больших моделей

Разработанные продукты

- **Платформа доверенного искусственного интеллекта, которая включает в себя инструменты, обеспечивающие безопасность, а также доверенные фреймворки машинного обучения**
- **Доверенная версия аналитической платформы Talisman**
- **Отчуждаемые инструменты Платформы:**
  - для тестирования моделей машинного обучения на устойчивость к состязательным атакам (и для защиты от атак)
  - для защиты от копирования обученных моделей машинного обучения
  - для защиты от извлечения обучающих данных из обученных моделей
  - для выявления и устранения закладок и зловредного кода в предобученных моделях машинного обучения
  - для объяснения моделей
  - для обнаружения аномалий и дрейфа данных
  - для выявления предвзятости моделей

Близкие по теме проекты ИСП РАН:

1. **Проект по цифровым водяным знакам с МИАН** (в том числе для маркирования сгенерированного контента)
2. **Разработка системы маркирования DocMarking**
3. **Молодёжная лаборатория по федеративному обучению** (при поддержке Минобрнауки России)

Пример технологии I: цензурирование больших моделей

Исследовательский центр искусственного интеллекта ИСП РАН

ДОВЕРЕННЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Исследовательский центр искусственного интеллекта Института общественных наук РАНХиГС

- Совместная разработка: набор программных методов (бенчмарк) **SLAVA: Sociopolitical Landscape and Value Analysis** («социально-политический ландшафт и ценностный анализ»)
- Объединяет **~14 тысяч вопросов** из официальных баз, разработанных для государственных экзаменов и проверочных работ. Вопросы касаются таких тем, как история, обществознание, политология, география и национальная безопасность.
- **Цель разработки**: создание методик и наполнение первого бенчмарка, который учитывает особенности культуры и законодательства России

https://iz.ru/1754474/andrei-korshunov-anton-belyi/slava-otechestva-neiroseti-proveriat-na-sootvetstvie-rossiiskim-kulturnym-kodam

Большие языковые модели показали низкий процент верных ответов на вопросы

Модель	ИТОГОВЫЙ рейтинг
qwen2:72b-instruct-q4_0	**53,17**
GigaChat_Pro	48,49
yandexgpt_pro	40,08
GigaChat_Plus	38,18
GigaChat_Lite	38,15
gemma2:9b-instruct-q4_0	35,12
llama3:70b-instruct-q4_0	31,75
yandexgpt_lite	26,28
llama3.1:70b-instruct-q4_0	25,43
qwen2:7b-instruct-q4_0	21,16
phi3:14b-medium-4k-instruct-q4_0	17,02
ilyagusev/saiga_llama3	17,06
mixtral:8x7b-instruct-v0.1-q4_0	10,89
solar:10.7b-instruct-v1-q4_0	11,97
mistral:7b-instruct-v0.3-q4_0	12,55
llama3:8b-instruct-q4_0	9,92
gemma:7b-instruct-v1.1-q4_0	10,25
llama3.1:8b-instruct-q4_0	9,07
yi:9b	10,48
gemma2:27b-instruct-q4_0	8,72
wavecut/vikhr:7b-instruct_0.4-Q4_1	10,44
random	11,60
qwen:7b	9,72
yi:6b	5,62
llama2:13b	3,70
Среднее значение	**20,67**

Пример технологии II: цифровые водяные знаки

Проект по цифровым водяным знакам

Научная тема:
Безопасность данных в вопросах источников происхождения, конфиденциальности, распределенного хранения и обработки, в том числе для задач машинного обучения (2024-2026)
Проект выполняется совместно с Математическим институтом им. В.А. Стеклова РАН

В числе планируемых результатов:
Разработка методов и средств, основанных на технологии цифровых водяных знаков, **для обеспечения возможности различать естественные и синтезированные данные**

В числе задач:
- анализ существующих методов внедрения цифровых водяных знаков в контент, синтезированный генеративными моделями
- исследование особенностей применения определенного идентификатора в качестве backdoor (управляющего сигнала) в задачах генерации контента
- исследование особенностей внедрения и извлечения цифрового водяного знака в сгенерированный контент в сценарии "белого ящика"

В ИСП РАН также развивается уникальная система внедрения цифровых водяных знаков DocMarking – для противодействия анонимности при утечках документов

Тренд по созданию и внедрению водяных знаков в сгенерированный контент актуален во всем мире!

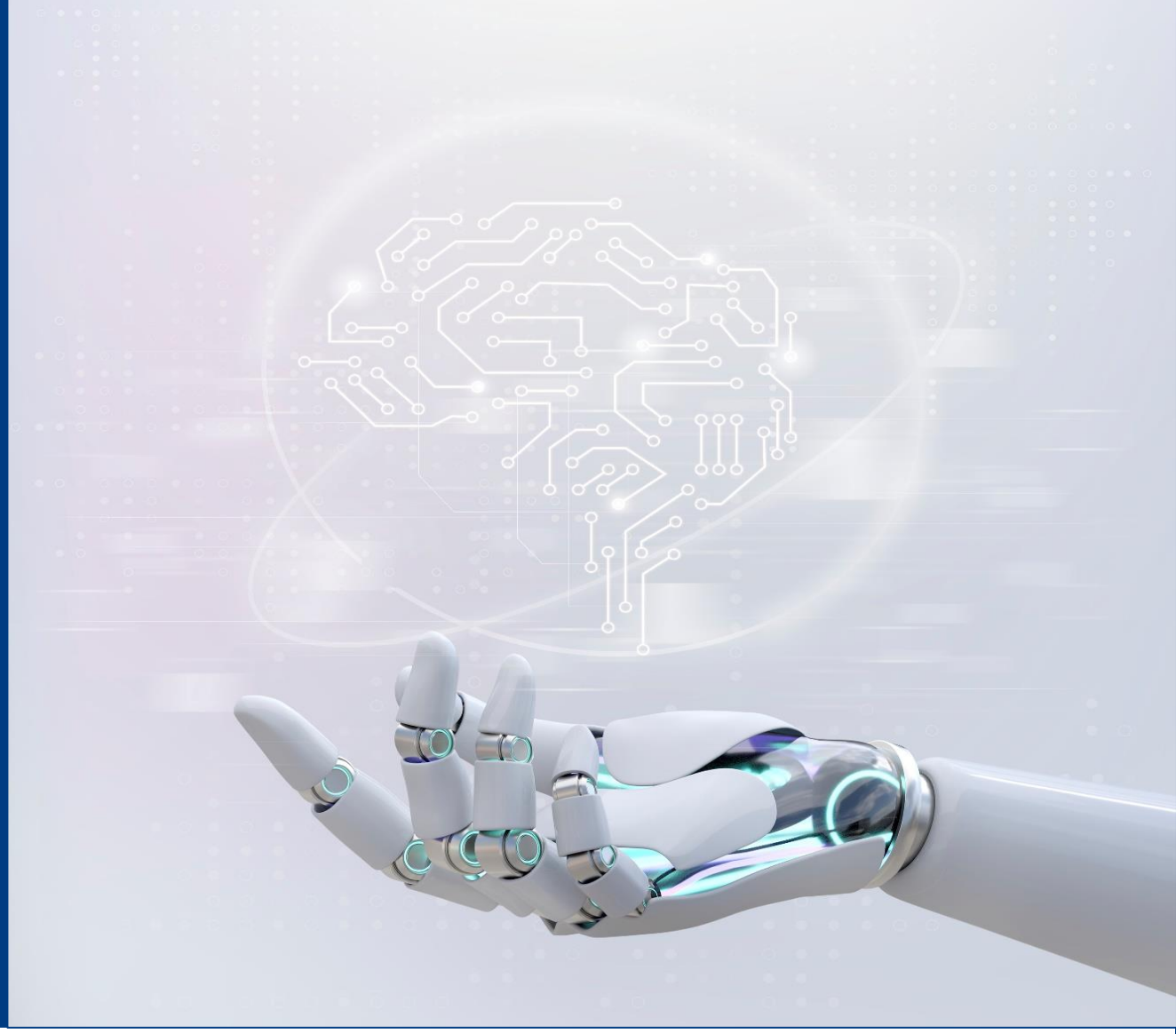
В 2023 компании **OpenAI, Alphabet, Meta Platforms, Anthropic, Inflection, Amazon, Microsoft** взяли на себя добровольные обязательства перед правительством США по реализации таких мер, как нанесение водяных знаков на контент, созданный ИИ, чтобы помочь сделать технологию безопаснее. Аналогичный подход реализован и в **европейской регуляторике**

ВЫВОДЫ

Внедрение ИИ остановить нельзя! Но можно обеспечить его безопасность, научить нас доверять этим технологиям. Для этого нужно уделять доверенному ИИ всё больше внимания (это долгосрочный глобальный тренд)

Требуется создание соответствующей научно-технологической базы для формирования регуляторики и разработки инструментов обеспечения доверия

Для создания отраслевой регуляторики и технологий необходимы междисциплинарные исследования с участием психологов, философов, социологов



Спасибо за внимание!