



VII МЕЖДУНАРОДНЫЙ НАУЧНЫЙ ФОРУМ

ШАГ В БУДУЩЕЕ:

ГЛОБАЛЬНЫЙ ФОРСАЙТ,
ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ
И СТРАТЕГИЧЕСКОЕ ЛИДЕРСТВО

Искусственный интеллект: актуальные задачи и модель развития

Аветисян Арутюн Ишханович

Директор Института системного программирования им. В.П. Иванникова Российской академии наук (ИСП РАН)

Искусственный интеллект: направления развития

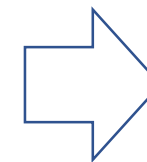
10 ключевых направлений

определены по итогам стратегической форсайт-сессии по поисковым исследованиям в сфере ИИ в 2024 г.

- Архитектуры, алгоритмы машинного обучения, оптимизация и математика
- Вычисления для ИИ
- Данные для ИИ
- Фундаментальные и генеративные модели
- Безопасность, доверие и объяснимость
- ИИ для узких задач
- Управление, принятие решений и агентные/мультиагентные системы
- Элементы «сильного» ИИ
- Взаимодействие человека и ИИ
- Социогуманитарные и экономические аспекты



- ✓ Цифровая трансформация всех отраслей экономики
- ✓ Повышение производительности труда
- ✓ Значительное снижение стоимости ИИ



Дальнейшее внедрение ИИ неизбежно!



Для качественного внедрения необходимы:

- I. Безопасность, доверие и объяснимость
- II. Модели долгосрочного развития

I. Безопасность, доверие и объяснимость: примеры проблем

Аварии с участием беспилотных автомобилей

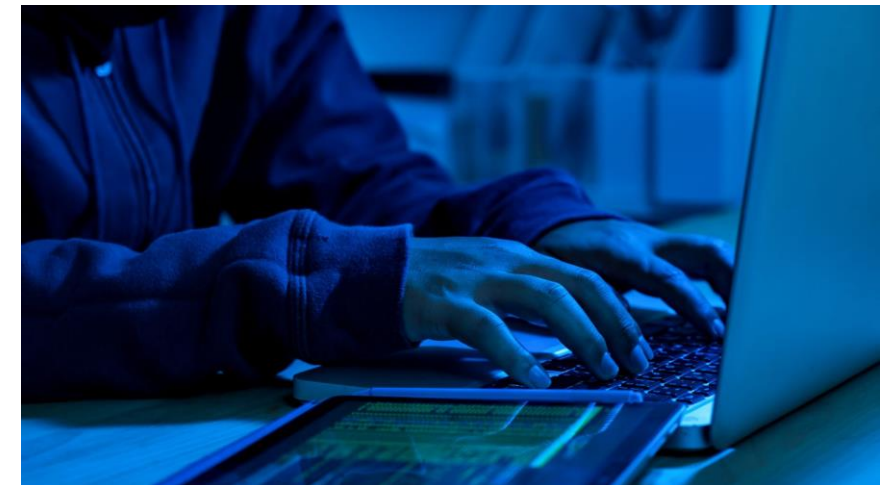
В 2023 в США произошло ДТП – беспилотный автомобиль Cruise сбил человека и протащил его 6 метров. В 2024 General Motors объявила о прекращении разработки роботизированных такси Cruise.

Мошенничества с дипфейками

В 2024 в Гонконге преступники выманили у сотрудника компании \$25,6 млн с помощью фэйковой видеоконференции. В том же году в Китае был вовремя заморожен аналогичный перевод в размере \$258 тысяч. По подсчёту аналитиков Сбербанка, за 8 месяцев 2024 количество преступных схем с использованием дипфейков в России выросло 30 раз!

Манипуляции

По данным ВОЗ, каждый восьмой человек в мире живёт с психическим расстройством (это больше миллиарда!). Каждый год более 720 тысяч человек совершают суицид. В 2023 в Бельгии мужчина покончил жизнь самоубийством после шести недель общения с чат-ботом на основе генеративного ИИ.



И многие другие!

I. Безопасность, доверие и объяснимость: развитие регуляторики

Регуляторика как ответ на технологические вызовы развивается так же, как ранее для обычного ПО (появление инструментов и методик безопасной разработки и т.п.).

Особенность ИИ – датацентричность!

Публикационная активность по теме доверенного ИИ к 2025 году: **3000+** научных статей
Число проектов на **GitHub: 2000+**

Рост числа инициатив и поиск баланса между регуляторикой и развитием технологий

2019

- Национальная стратегия развития ИИ до 2030 года (Россия)

2021

- Кодекс этики в сфере ИИ (сейчас объединяет **850** подписантов, в том числе **42** зарубежных из **24** стран), Россия
- **Федеральный проект «Искусственный интеллект»** (поддержка **6** центров «первой волны»), Россия
- **ГОСТ Р 59525-2021 «Интеллектуальные методы обработки медицинских данных»**, Россия

С 2021 – активное развитие тематики доверенного ИИ

2022

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

2023

- **Executive Order on Safe, Secure, and Trustworthy AI** (США), аннулирован в **2025**
- **NIST Trustworthy & Responsible AI Resource Center (AIRC)** (США)
- **Hiroshima AI Process (G7)**
- **ENSIA**, методика (Евросоюз)
- **Временные регуляторные документы про генеративный ИИ** о необходимости пометок контента (Китай)

2024

- Резолюция Генассамблеи ООН по безопасным системам ИИ
- США и Великобритания заключили **договор о безопасности в сфере ИИ** (первый в этой сфере)
- **EU AI Act** (некоторые технологии ИИ предлагается запретить, а сгенерированный контент – обязательно маркировать). В его рамках: проект **AI Code of Practice**.
- **European AI Office** – для координации работ с ИИ
- **Национальная стратегия развития ИИ до 2030 года, новая редакция** (Россия)
- **Консорциум для исследований безопасности технологий искусственного интеллекта:** НТЦ ЦК, Академия криптографии, ИСП РАН и другие (Россия)

И многое другое!

I. Безопасность, доверие и объяснимость: анализ не только ПО, но и моделей и данных



Исследовательский центр доверенного ИИ ИСП РАН

датацентричность!

Анализ данных

- ПО для обнаружения аномалий и дрейфа данных (методы обнаружения аномалий в задаче генерации текста с помощью больших языковых моделей; для мультимодальных моделей; в задаче семантической сегментации и др.)

+ цифровые водяные знаки

разработка системы **DocMarking** и совместный проект с МИАН, включающий создание технологий для различения естественных и синтезированных данных

Доверенные фреймворки

Доверенные версии фреймворков **PyTorch** и **TensorFlow**

- **TrustTorch** рекомендован для принятия на снабжение ВС РФ и организации промышленного производства
- **TrustFlow** введен в эксплуатацию в продукте АО «Лаборатория Касперского» «**Kaspersky Machine Learning for Anomaly Detection**» v. 3.0

+ федеративное обучение

в 2024 вместе с Яндексом и Сеченовским университетом впервые применили федеративное обучение для решения задач медицины; комплекс соответствующих технологий признан одним из важнейших результатов РАН

датацентричность!

Анализ моделей машинного обучения

- ПО для тестирования моделей на устойчивость к состязательным атакам и защиты от таких атак
- для защиты от копирования обученных моделей
- для защиты от извлечения обучающих данных
- для выявления и устранения закладок и зловредного кода в предобученных моделях
- для объяснения моделей
- для выявления предвзятости

II. Модели долгосрочного развития: наш опыт на основе открытого кода



II. Модели долгосрочного развития: что мы предлагаем для ИИ?

Ограничения доступа к технологиям
быть не должно!



❑ Предлагаем объединиться для совместной работы, в том числе по созданию средств разработки технологий ИИ

❑ Приглашаем к сотрудничеству в современной модели коллаборативной экономики



VII МЕЖДУНАРОДНЫЙ НАУЧНЫЙ ФОРУМ

ШАГ В БУДУЩЕЕ:

ГЛОБАЛЬНЫЙ ФОРСАЙТ,
ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ
И СТРАТЕГИЧЕСКОЕ ЛИДЕРСТВО

Спасибо за внимание!