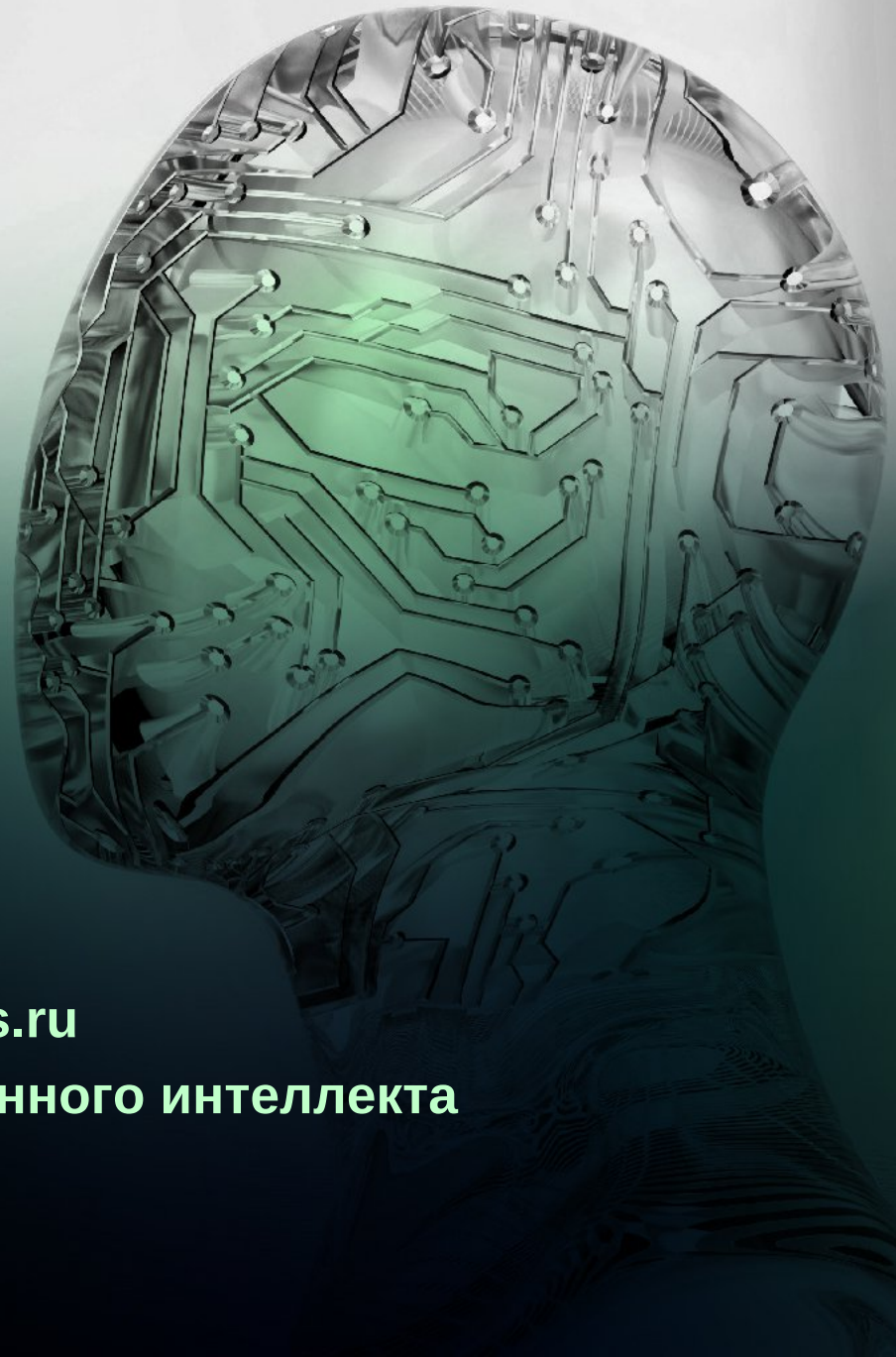
КОНСОРЦИУМ
ИССЛЕДОВАНИЙ
БЕЗОПАСНОСТИ
ТЕХНОЛОГИЙ
ИСКУССТВЕННОГО
ИНТЕЛЛЕКТА



Безопасная разработка технологий ИИ

Падарян Вартан Андроникович vartan@ispras.ru

III Форум Технологии доверенного искусственного интеллекта

20 мая 2025

РАБОЧАЯ ГРУППА №4
Общая информация

Председатель:

Падарян В.А. ИСП РАН

Участники:

ИСП РАН

АК РФ

АНО «НТЦ ЦК»

АО «Позитив Текнолоджиз»

Ассоциация ФинТех

ГК «Гарда»

Минцифры России

НГТУ НЭТИ

ГК «Свордфиш Секьюрити»

ООО «Газинфосервис»

АО «ИнфоТеКС»

ООО «АйТи Бастион»

НИУ ВШЭ

ИТ-кампус Неймарк

ООО «Код Безопасности»

РТУ МИРЭА

СПб ФИЦ РАН

АПКИТ

InfoWatch

РГГУ

ПАО ВТБ

НИЯУ МИФИ

ООО «Солар Секьюрити»

Вызовы

безопасности и доверия ИИ

Угрозы на всех этапах жизненного цикла

Эксплуатация моделей

- Атаки, эксплуатируют особенности ИИ
- Принципиально отличаются от атак на классическое ПО (через данные, а не код)
- Сложность обнаружения атаки

Обучение моделей

- Высокая стоимость ошибки (стоимость цикла обучения)
- Закладки в наборах данных

Датасеты

- Уязвимости в распространенном ПО (фреймворки, библиотеки)

Фреймворки

- Невозможно обеспечить доверие к технологиям ИИ только за счет НПА
- Нужны инструменты для разработки доверенных систем ИИ
- Должны быть определены требования к инструментам и методикам
  - Поддержка лучших практик разработки, принятых в области ИИ (MLOps)
  - Обеспечение безопасности (MLSecOps)
  - Увеличение скорости промышленной разработки, а не новые барьеры
  - Повышение эффективности использования аппаратуры
  - …

ЦЕЛИ И ЗАДАЧИ РГ№4

Цели

Выработка согласованных требований к процессам разработки безопасных технологий ИИ

Определение состава мер и средств, обеспечивающих выполнение этих требований

ЗАДАЧИ

Организационные, в рамках консорциума

- Синхронизация по перечню угроз безопасности ИИ, разработка предложений в части НПА (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

РАБОЧАЯ ГРУППА №4

План работ

Февраль	Стабилизация состав РГ, определение ролей и вклада участников
Март	Сбор подходов от участников, изучение мировых практик
Апрель	Сведение лучших практик разработки безопасных технологий ИИ, подготовка первой редакции.
Июнь	Определение состава инструментальных средств (с учетом доступности open source инструментов) , оценка требуемых для их работы вычислительных ресурсов
Июль	Выбор ПО для пилотирования, определение облика MVP - конвейера разработки безопасного ПО, реализующего технологии ИИ
Август	Пилотирование MVP MLSecOps-платформы на площадке ИСП РАН. Создание образа методологических материалов - политик, регламентов, рекомендаций, руководств для использования платформы.
Октябрь	Доработка MLSecOps по результатам пилотирования. Доработка методологических материалов с описанием работы платформы и ее функций.
Ноябрь	Разработка первой редакции методических указаний по разработке безопасного ПО, реализующего технологии ИИ . Разработка программы обучения DevOps инженеров эксплуатации MLSecOps.
Декабрь	Развертывание конвейера MLSecOps для тестирования безопасности технологий ИИ в составе испытательных стендов РГ-2, обучение операторов испытательных стендов.

СВОДКА ЛУЧШИХ ПРАКТИК
разработки безопасных технологий ИИ

Open source инструменты постепенно заменяем доверенными версиями, аналогично PyTorch и TensorFlow

Подготовка обучающих данных

Получение обучающих данных из доверенных источников, в перспективе – из реестра доверенного ИИ

Безопасный формат моделей, например **SafeTensors**

Использование доверенных фреймворков

Выявление конфиденциальных сведений, **Microsoft Presidio**

Выявление аномалий данных, в ИСП РАН идет разработка инструмента на базе **Anomalib** и **PyOD**

Выявление признаков отравления и закладок, в ИСП РАН идет разработка инструмента

Устранение несбалансированности обучающих данных, **Cleanlab** и **Fastdup**

Устранение предвзятости обучающих данных, в ИСП РАН идет разработка инструмента

Защита от состязательных атак, в ИСП РАН идет разработка инструмента

СВОДКА ЛУЧШИХ ПРАКТИК
разработки безопасных технологий ИИ

Обучение модели

Обучение только на доверенной инфраструктуре

Дифференциально-приватное обучение, **Opacus** (только для PyTorch/TrustTorch)

Федеративное обучение, в ИСП РАН идет разработка инструмента,
на текущий момент применяем **PySyft** либо **Flower**

Удаление незначимых частей модели (сжатие), в ИСП РАН идет разработка инструмента

СВОДКА ЛУЧШИХ ПРАКТИК
разработки безопасных технологий ИИ

Тестирование обученной модели (задача вне области ответственности РГ №4)

Устойчивость к состязательным атакам (для классификаторов), в ИСП РАН идет разработка инструмента, временно применяем **RobustBench**

Устойчивость к sponge examples, в ИСП РАН идет разработка инструмента

Устойчивость к промптам-эксплойтам (для языковых моделей), **Garak** и **JailbreakBench**

Возможности несанкционированного извлечения из модели конфиденциальной информации, **Guardrails**, **Microsoft Presidio**, **ML Privacy Meter**, **Model Inversion Attack Toolbox**, **AIJack**

Проверка этичности (для языковых моделей), **DeepEval**

Оценка предвзятости, **Fairlearn** и **AIF 360**

Оценка интерпретируемости предлагаемых моделью решений, в ИСП РАН идет разработка инструмента, временно применяем **Captum**, **SHAP**, **InterpretML**

Общее тестирование, **ModelScan** и **Watchtower**, ручной пентестинг модели с учетом ее целевого назначения и актуальной модели нарушителя

СВОДКА ЛУЧШИХ ПРАКТИК
разработки безопасных технологий ИИ

Эксплуатация

Защита от несанкционированного извлечения из модели конфиденциальной информации, **Guardrails** и **Microsoft Presidio**

Защита от несанкционированной дистилляции знаний из модели, в ИСП РАН идет разработка инструмента на базе **AgentOps** и **OpenLLMetry**

Защита от состязательных атак, в ИСП РАН идет разработка инструмента

Защита от sponge examples, в ИСП РАН идет разработка инструмента

Защита от промптов-эксплойтов (для языковых моделей), **Guardrails**

Ведение подробных журналов, **AgentOps** и **OpenLLMetry**

Генерация водяных знаков (для генеративных моделей), в ИСП РАН идет разработка инструмента

Выявление аномалий данных и дрейфа, в ИСП РАН идет разработка инструмента на базе **Anomalib** и **PyOD**

Интерпретация предлагаемых моделью решений, в ИСП РАН идет разработка инструмента, **Captum** и **SHAP**

РАБОЧАЯ ГРУППА №4

Ожидаемые результаты

Функционирующий конвейер MLSecOps для тестирования безопасности технологий ИИ в интересах Реестра решений доверенного ИИ

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

Регулирование ИИ

МИРОВОЙ ОПЫТ

Постоянный рост числа инициатив

2020

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

2022

- **AI Bill of Rights** (США)
- **ГОСТ Р 59276—2020**

2023

- **Executive Order on Safe, Secure, and Trustworthy AI** (США)
- **NIST AIRC**, исследовательский центр (США)
- **Hiroshima AI Process** (G7)
- **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
- **Обновленная Национальная стратегия развития ИИ в России до 2030 года**

ЗАДЕЛ ИСП РАН

Исследовательский центр доверенного искусственного интеллекта

Платформа для разработки доверенных технологий ИИ

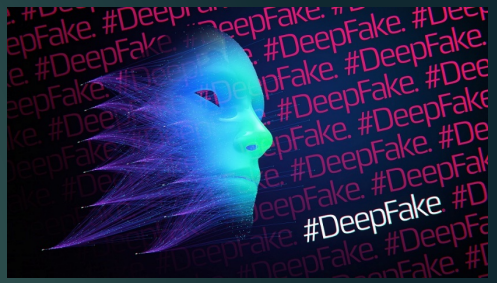
- Ядро платформы – собственное решение класса MLOps/MLSecOps
- Доверенные ML-фреймворки PyTorch и TensorFlow
- Инфраструктура и отчуждаемая методология по созданию новых доверенных фреймворков и библиотек
- Быстрое подключение новых инструментов анализа и мониторинга моделей и датасетов
- Интеграция в жизненный цикл новых и существующих систем ИИ
- Платформа обеспечивает инструментальное сопровождение созданных Центром методологий разработки доверенных систем, использующих ИИ

Отчуждаемые инструменты

- Тестирования и защиты моделей машинного обучения от состязательных атак на этапе эксплуатации
- Выявления и устранения предвзятости моделей машинного обучения
- Интерпретации моделей машинного обучения
- Обнаружения аномалий и дрейфа в наборах данных
- Выявления и устранения закладок и вредоносного кода в предобученных моделях машинного обучения
- Защиты от копирования обученных моделей машинного обучения и защиты от извлечения обучающих данных из обученных моделей

ЗАДЕЛ ИСП РАН

Исследовательский центр доверенного искусственного интеллекта



Решение гуманитарных проблем

Борьба с Deepfake и кражами моделей

Цифровые водяные знаки*

- Маркирование сгенерированного контента
- dataset watermarking позволяет установить или опровергнуть факт обучения нейронной сети на заданном наборе данных

** OpenAI, Alphabet, Meta Platforms, Anthropic, Inflection, Amazon, Microsoft взяли добровольные обязательства по встраиванию водяных знаков в ИИ-контент*

- ИСП РАН совместно с МИАН ведет работу «Безопасность данных в вопросах источников происхождения, конфиденциальности, распределенного хранения и обработки, в том числе, для задач машинного обучения» (2024-2026)
- В ИСП РАН развивается система внедрения цифровых водяных знаков DocMarking – для противодействия анонимности при утечках документов