Падарян Вартан
Заведующий лабораторией ИСП РАН
vartan@ispras.ru
22 февраля 2025 г.
Цифровая устойчивость промышленных систем

Безопасная разработка технологий ИИ

ИСП РАН

Доверенный искусственный интеллект (ИИ): мировой тренд

ИСП РАН

Путь к ИИ, которому мы сможем доверять

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)

Публикационная активность к 2025 году:
3000+ научных статей

Число проектов на GitHub:
2000+

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!**

Всё развивается так же, как и в случае ПО без ИИ

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 / 2024, 6 уровней доверия СЗИ, ГОСТ Р 71206-2024, ГОСТ Р 71207-2024**
общие требования, безопасный компилятор и статический анализ

Евросоюз: The Cybersecurity Act (EU 881/2019), система сертификации ПО, сервисов и процессов

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

- **Выяснилось, что недостаточно использовать классические методы защиты (по периметру, проверка доступа, антивирусы)**
- **Возникла необходимость в разработке новых моделей, методов и технологий в области анализа и трансформации программ, чтобы**

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

Доверенный ИИ: уже не только регуляторика!

ИСП РАН

https://compl-ai.org/

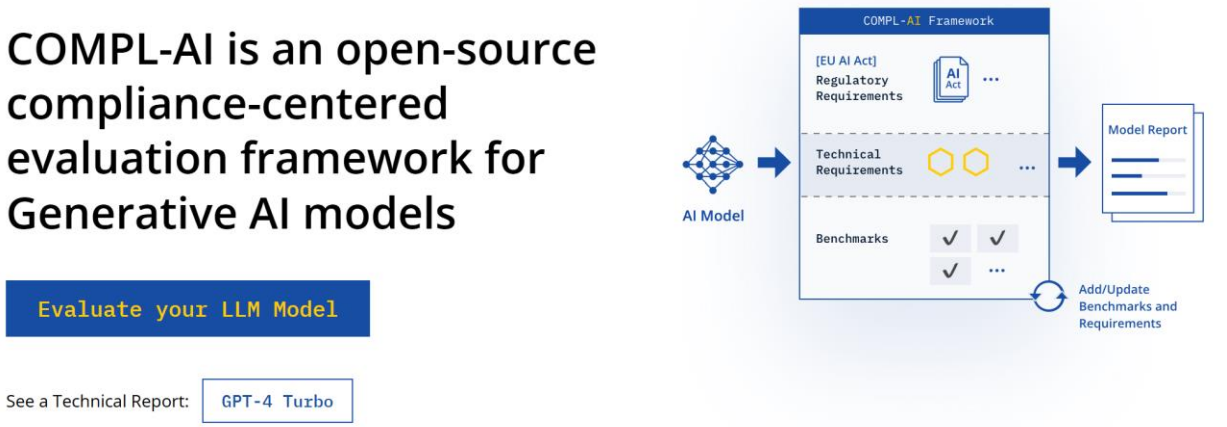
Фреймворк для оценки больших языковых моделей на соответствие EU AI Act – главному европейскому закону об искусственном интеллекте

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 ИИ как первый шаг на пути воплощения требований ЕС к искусственному интеллекту в технические требования, помогая поставщикам моделей ИИ внедрять EU AI Act».

Томас Ренье, спикер Еврокомиссии

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

Доверенный ИИ: комплексное понимание

ИСП РАН

ДОВЕРИЕ К ФУНКЦИОНИРОВАНИЮ
(доверие со стороны кибербезопасности)

в обычных условиях

- ❑ Доверие к процессу разработки
- ❑ Оценка качества данных для обучения (в т.ч. согласованности)
- ❑ Предсказуемость качества работы технологий ИИ
- ❑ Устойчивость и стабильность работы, борьба с устареванием моделей
- ❑ Интерпретируемость результатов

в условиях противодействия

- ❑ Устойчивость к состязательным атакам
- ❑ Устойчивость к краже моделей и данных
- ❑ Обнаружение закладок и вредоносного кода:
  - в библиотеках и фреймворках
  - в данных для обучения
  - в моделях

ЭТИЧЕСКИЕ АСПЕКТЫ
(доверие с социогуманитарной стороны)

- ❑ Борьба со склонностью к предвзятым оценкам
- ❑ Обеспечение приватности и работа с персональными данными
- ❑ Предотвращение деградации качества работы пользователей из-за чрезмерного доверия к технологиям ИИ
- ❑ Обнаружение манипуляций в измерениях
- ❑ …

Доверенный ИИ в РФ: обзор инициатив

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 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

2024

ФСТЭК России ведёт активные работы по стандартизации безопасности ПО, реализующего технологии ИИ

Доверенный ИИ в РФ: направления работ ФСТЭК России

ИСП РАН

Направления стандартизации безопасности ПО, реализующего технологии ИИ

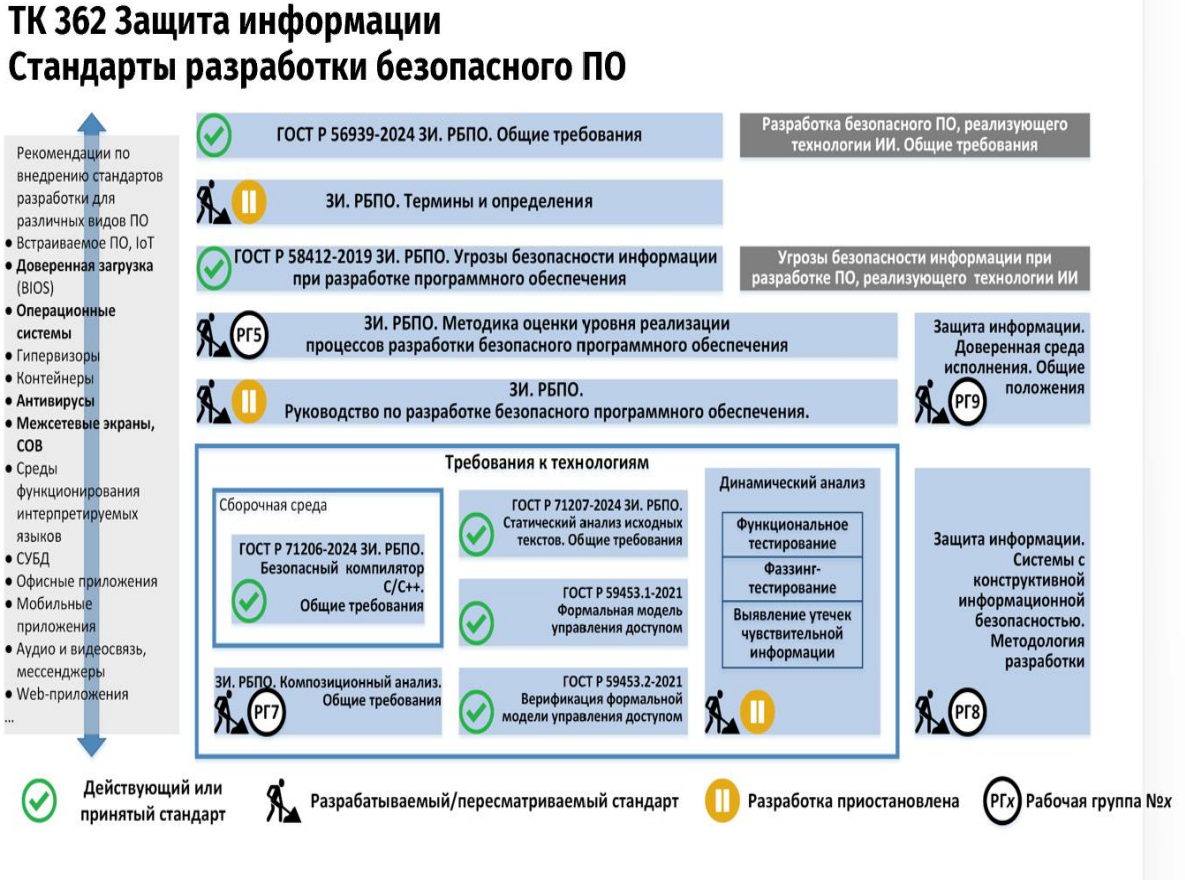
Угрозы безопасности информации при разработке ПО, реализующего технологии ИИ

- ❑ **Актуализация и расширение перечня угроз УБИ.218-222**
- ❑ **Способы реализации угроз безопасности**

Разработка безопасного ПО, реализующего технологии ИИ

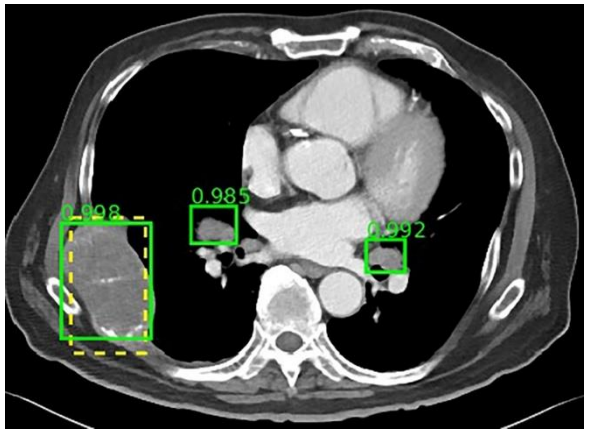
❑**Дополнительные требования в отношении процессов РБПО ГОСТ Р 56939-2024** (формирование и предъявление требований безопасности к ПО; разработка, уточнение и анализ архитектуры ПО и др.)

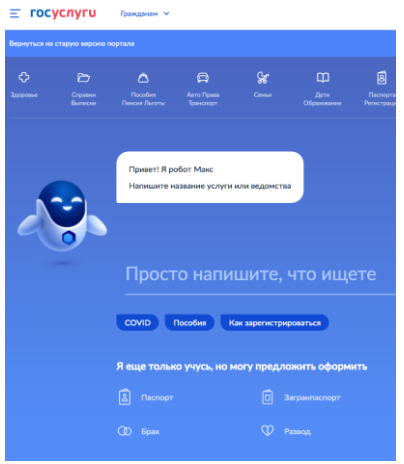
❑**Дополнительные процессы РБПО, возникающие при реализации технологий ИИ** (управление обучающими наборами данных, обучение и дообучение моделей, испытания моделей и др.)



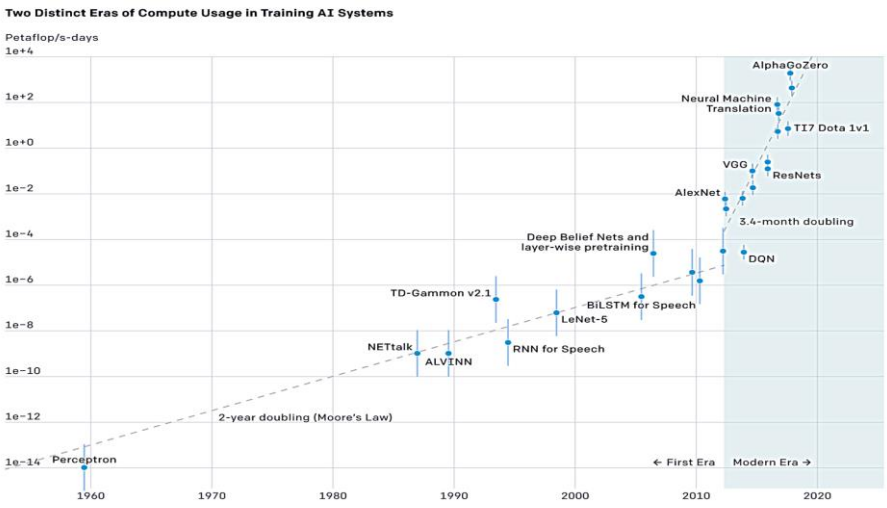
Предпосылки создания Исследовательского центра доверенного ИИ «первой волны»

ИСП РАН





- ❑ **Рост числа внедрений технологий**
- ❑ **Снижение требований к квалификации разработчиков**
- ❑ **Открытый доступ к знаниям и реализациям передовых алгоритмов и методик (научные публикации и свободное ПО)**
- ❑ **Стандартные интерфейсы программных библиотек**
- ❑ **Открытые фреймворки, реализующие системный уровень (взаимодействие с аппаратурой, эффективная программная реализация математического аппарата – линейная алгебра, оптимизация)**



Исследовательский центр доверенного ИИ: задачи и результаты в рамках «первой волны»

ИСП РАН

Поставленные 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

Результаты синхронизируются с ФСТЭК России и используются при подготовке ГОСТов

Публикационная активность Центра к 2025: 70+ статей по темам доверенного ИИ (A*/Q1)

ДОВЕРЕННЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

KASPERSKY lab · IPC InterProCom · ТЕХНОПРОМ · ЕС-ЛИЗИНГ

Разработанные продукты

I. Платформа доверенного ИИ:

- **Доверенные фреймворки машинного обучения**
- **Среда анализа фреймворков и библиотек**
- **Отчуждаемые инструменты:**
  - 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

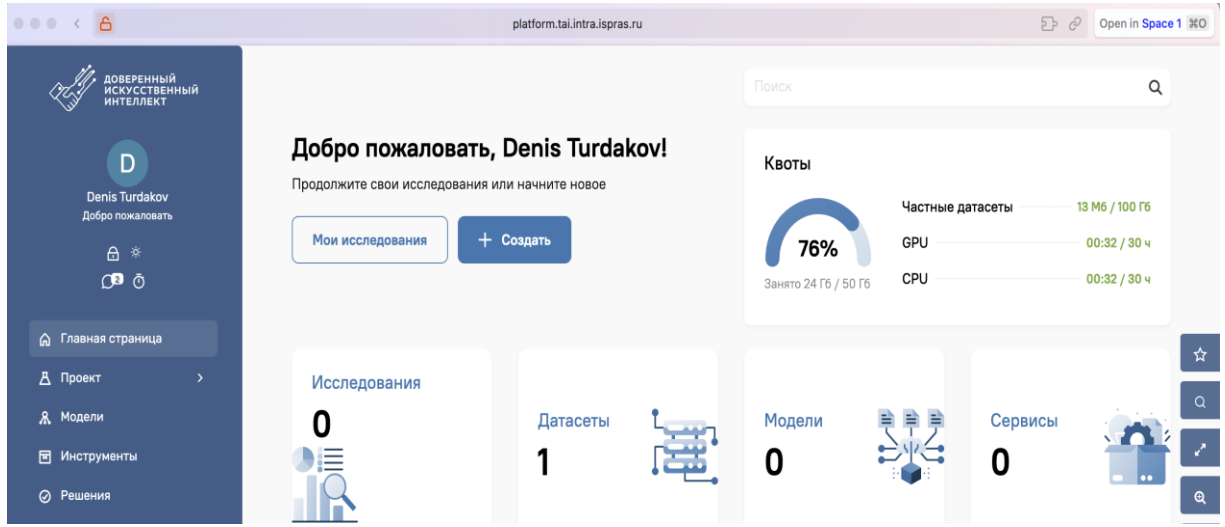
II. Доверенная версия платформы Talisman

Результат I, Платформа доверенного ИИ: общее описание

ИСП РАН

Ядро – собственное решение класса MLOps/MLSecOps

- Аналоги ClearML, MLFlow, KubeFlow
- Функциональные преимущества перед аналогами:
  - интеграция в жизненный цикл новых и существующих систем ИИ
  - быстрое подключение новых инструментов анализа и мониторинга моделей и датасетов
- Преимущества обеспечиваются за счет нового способа описания моделей машинного обучения, разработанного в Центре
- Ликвидация известных проблем с безопасностью на уровне кода (TrustFlow, TrustTorch, отказ от pickle)



Платформа обеспечивает инструментальное сопровождение созданных Центром **методик разработки доверенных интеллектуальных систем**

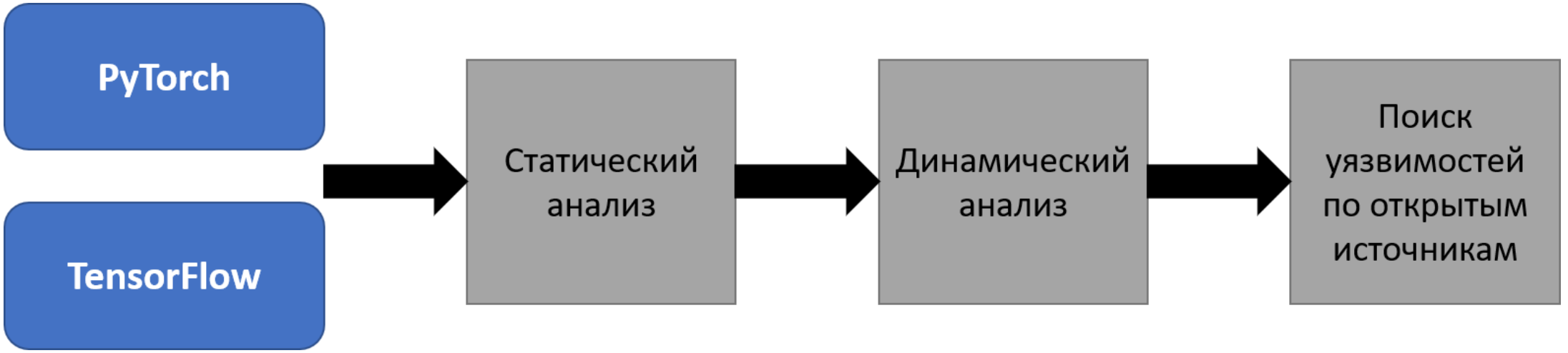
! Планы в рамках «третьей волны»:
дальнейшее развитие технологий в связи с новыми угрозами;
передача технологий в прикладные отрасли

Анализ фреймворков машинного обучения

ИСП РАН

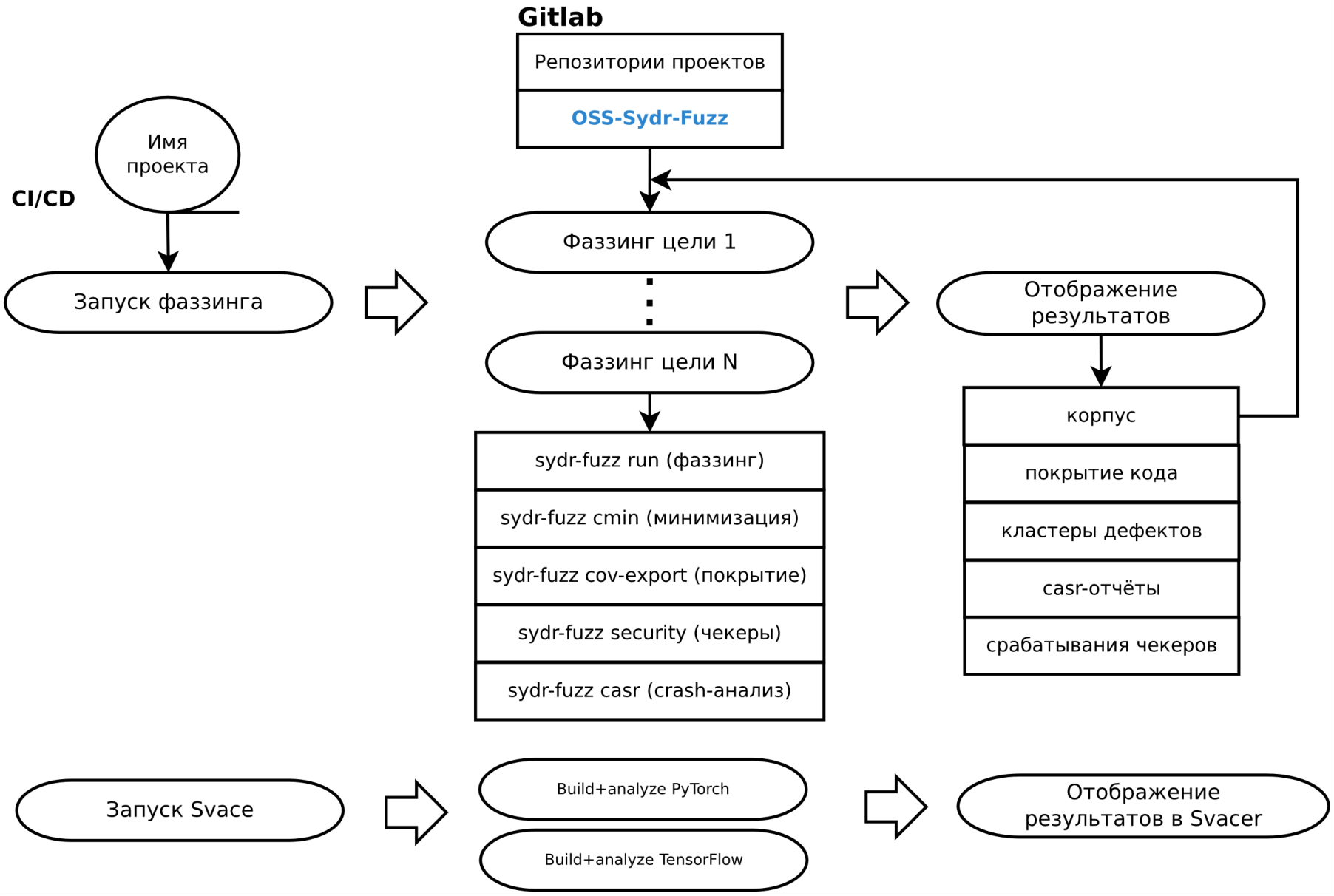
ML/DL фреймворки – фундаментальные библиотеки для разработки продуктов, использующих технологии искусственного интеллекта

Цель: обезопасить использование ИИ-фреймворком путем внедрения для них SDL с использованием собственных средств, инфраструктуры и инструментов (Svace/Sydr/…)



Система непрерывного анализа ИИ-фреймворков

ИСП РАН



Найденные ошибки в ИИ-фреймворках

ИСП РАН

Проект	Ошибки (Sydr|Svace)	Исправлено	Принято в апстрим
TensorFlow	36 (5|31)	36	30
PyTorch	52 (35|17)	52	47

При поиске и исправлении ошибок также учитывались зависимости проектов: mkl-dnn, XNNPACK, llvm-project, miniz, opencv, FFmpeg, torchvision и др.

Подробности о найденных ошибках:
github.com/ispras/oss-sydr-fuzz/blob/master/TROPHIES.md

Результат I, Платформа доверенного ИИ: состав

ИСП РАН

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

Фреймворки TrustFlow* и TrustTorch*

*Версии TensorFlow и PyTorch, прошедшие через методики разработки безопасного ПО (РБПО)

- ❑ Переданы для тестовой эксплуатации гос. заказчику, TrustTorch прошел государственные испытания в составе изделия

Среда анализа фреймворков и библиотек машинного обучения

- ❑ Автоматизация методик РБПО, учитывающая специфику фреймворков машинного обучения
- ❑ Может быть использована для проверки других фреймворков

Результат II, доверенная версия платформы Talisman: общее описание

ИСП РАН

Low-code **платформа для построения интеллектуальных информационно-аналитических систем**. Обеспечивает значительное повышение полноты и качества информации, используемой для принятия управленческих решений:

- **Объединяет инструменты для автоматизации типовых задач обработки данных, позволяющие извлекать информацию из 100+ естественных языков**
- **Использует современные глубокие нейронные сети**
- **Объединяет 50+ моделей машинного обучения (анализ текстов, изображений, графов, таблиц)**
- **Собирает данные из Интернета и корпоративных хранилищ**

! Разработка в рамках «первой волны»

- **Интеграция с Платформой ДИИ**
- **В разработке решений на основе Talisman применяются технологии и методики Центра** (доверенные фреймворки, анализ используемых датасетов, анализ моделей машинного обучения)

БИЗНЕСУ

- Анализ документации
- Проведение конкурентной разведки
- Оптимизация управления персоналом
- Объективная оценка эффективности деятельности

ГОССТРУКТУРАМ

- Создание систем мониторинга угроз информационной безопасности и прогнозирования компьютерных атак,
- Создание систем технологической разведки и ведения базы знаний по предметной области заказчика

ВУЗАМ

Масштабирование опыта с МГИМО МИД России, где создана система интеллектуального анализа данных в области международных отношений на базе Talisman (в том числе для обучения специалистов)

Консорциум исследований безопасности технологий искусственного интеллекта

ИСП РАН

- РГ 1 Разработка нормативно-правовых актов – Positive Technologies
- РГ 2 Тестирование технологий искусственного интеллекта – НТЦ ЦК
- РГ 3 Создание и развитие реестра доверенных решений искусственного интеллекта – Газинформсервис
- РГ 4 Разработка безопасных технологий искусственного интеллекта – ИСП РАН

Участники (16 декабря 2024):

АНО «НТЦ ЦК»

АО «Позитив Текнолоджиз»

Ассоциация ФинТех

Гарда

Минцифры России

НГТУ

Ожидается вступление новых участников консорциума

РГ 4 Разработка безопасных технологий искусственного интеллекта

ИСП РАН

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

Ожидаемые результаты

ИСП РАН

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

Спасибо!

ИСП РАН