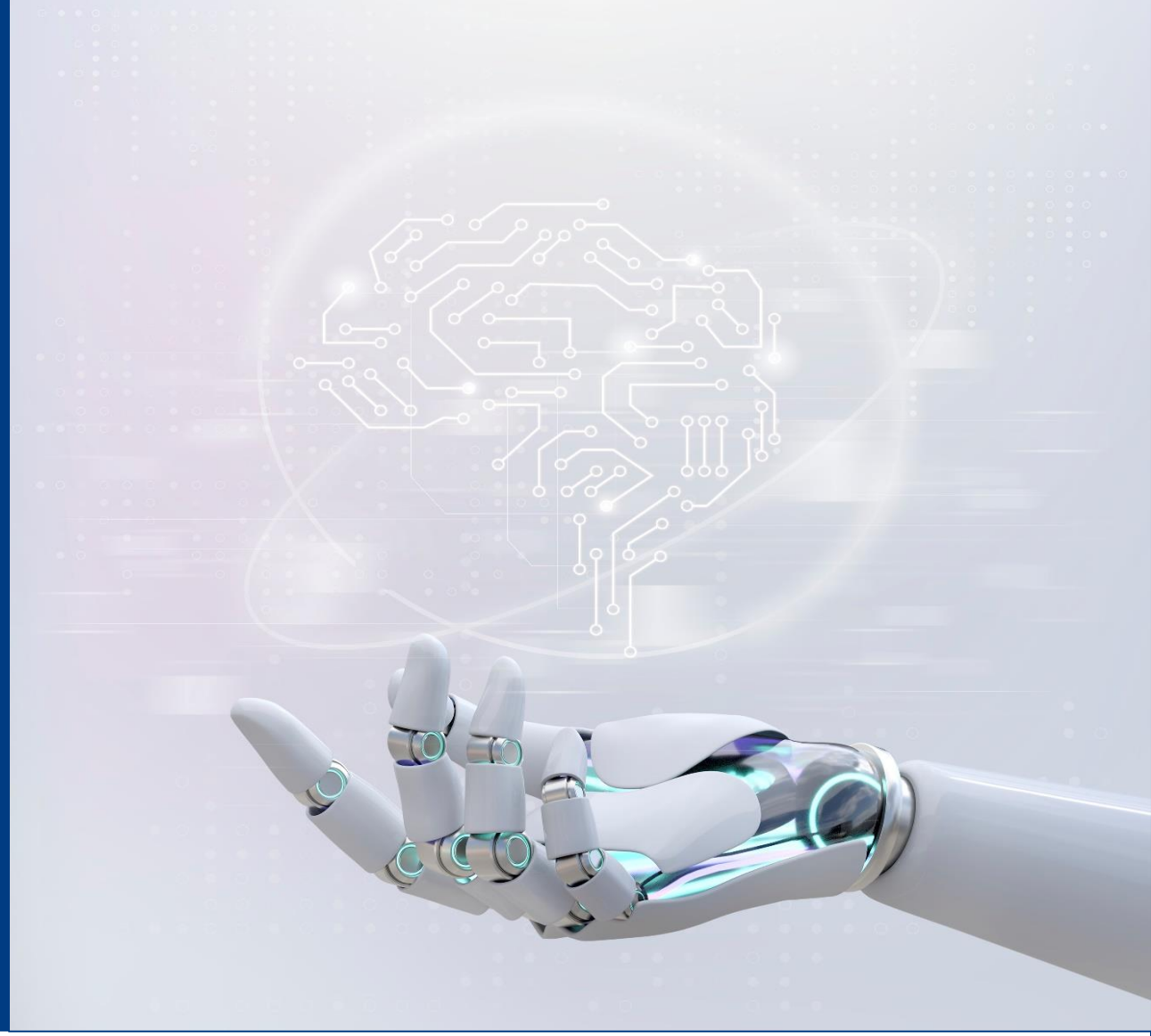
Системное программирование и технологии создания доверенных систем (в том числе с искусственным интеллектом)

Арутюн Аветисян
директор ИСП РАН
академик РАН
arut@ispras.ru
30 января 2025 года



ИСП РАН

Объединенный институт ядерных исследований
НАУКА СБЛИЖАЕТ НАРОДЫ

Сложность современного ПО

ИСП РАН

Сейчас программы:
Быстро растут
Усложняются
Не бывают изолированными

В то же время программам необходимы:

Эффективность

Продуктивность

Доверенность*

**Доверенность в данном контексте = безопасность*

Развитие ИТ на основе открытого кода

ИСП РАН

Ускоренное развитие ИТ происходит в модели коллаборативной экономики

- **Компании используют в качестве основы для своих продуктов проекты с открытым исходным кодом, в которые сами же вкладываются совместно**
- **В одиночку этот процесс осилить невозможно из-за сложности современного ПО**

GitHub

SOURCEFORGE

GitLab

Bitbucket

Немного статистики

GitHub-2024
518 млн проектов
(+25% за год)

ОС, фреймворки
PyTorch 7 млн строк кода
TensorFlow 10 млн строк кода
Debian >2 млрд строк кода

В январе 2025 года исходный код ядра Linux превысил 40 миллионов строк: это в 2 раза больше, чем 10 лет назад!

Эффективность и продуктивность, пример I: компиляторы и наборы инструментов

ИСП РАН

Компьютерные оптимизации и наборы инструментов

- **iOS**: Objective-C/LLVM – Swift/LLVM (2014)
- **Android**: Java/Dalvik (2010), Java/Android Runtime (ART) (2014), Jack (2016), **Kotlin** (2019)
- **Tizen**: C++, JavaScript/WebKit & V8 (2012), C# / Roslyn (2016)
- **Десктопы и серверы**: Оптимизации всей программы / GCC, LLVM, двоичная оптимизация / **Bolt**, формат и оптимизатор WebAssembly, языки Go, **Rust**, Dart (фреймворк Flutter)
- **Многоядерные процессоры, GPU-ускорители**: стандарты OpenCL, OpenMP / GCC, LLVM
- **Опыт ИСП РАН**: 5 официальных ревьюеров GCC, OpenMP для GPU/CUDA в GCC, опережающая компиляция для JavaScript, оптимизация размера дистрибутива ОС Tizen (уменьшение на ~20%)...

Сотни миллионов строк кода – справиться в одиночку невозможно:

- Закрытые компиляторы Intel, IBM, Microsoft проиграли конкуренцию (используют LLVM/Clang)
- «Эльбрус»: собственный компилятор GCC, затраты на поддержку своей отдельной кодогенерации, трата ресурсов и отставание от основной версии
- **Опыт ИСП РАН**: исследования по поддержке стандарта OpenCL для FPGA (2012 г.)
  Спустя 3-4 года это стало мейнстримом для основных производителей Xilinx / Altera

Эффективность и продуктивность, пример II: классическая виртуализация и облачные технологии

ИСП РАН

Среди коммерческих решений преобладает VMware, в меньшем объеме Citrix

Закрытые облачные провайдеры (Amazon, Microsoft, Alibaba...) используют открытые компоненты (Openstack, Xen, Linux KVM)

Контейнеризация часто используется поверх ВМ даже в коммерческих облаках и является дополнением

Открытые облачные платформы – Openstack, OpenNebula и Eucalyptus:

- Openstack выиграл конкуренцию – 40+ модулей IaaS + PaaS, 10+ млн. строк кода против 0.5 млн. строк / только IaaS у OpenNebula и Eucalyptus
- Тем не менее, сначала OpenNebula был самым продвинутым проектом и развернут в CERN
- Причина – **отсутствие сообщества**! OpenNebula и Eucalyptus развиваются отдельными командами, тогда как Openstack взял курс на строительство сообщества, сейчас – 400+ разработчиков
- **Разработки в России**: SberCloud/cloud.ru (Сбербанк), VK cloud, Тионикс и Рустэк (куплены РТК), AccentOS, CloudX, T1, Asperitas (ИСП РАН)

Опыт ИСП РАН: среда Asperitas (реестр отечественного ПО №5921) на базе Openstack

- Минимальная служебная операционная система без пользовательских функций
- IaaS на базе Openstack, PaaS, контейнеризация
- Рабочие столы в веб-браузере по запросу на базе Linux контейнеров (podman/docker)
- Полностью изолированная установка из собственных источников без доступа к сети Интернет
- В 2024 получила от ФСТЭК России сертификат соответствия требованиям доверия и требованиям к средствам виртуализации по 4 уровню

Asperitas как часть инфраструктуры «мегасайенс» и основа коммерческой платформы ACloud

2023: Договор о стратегическом сотрудничестве с российским оператором ИТ-инфраструктуры «Системные решения»

Ключевое направление – разработка и интеграция коммерческой платформы Acloud (выпущена в 2024)

2024: Консорциум с Курчатовским институтом и Объединенным институтом ядерных исследований для ИТ-обеспечения исследовательской инфраструктуры класса «мегасайенс»

Он будет использовать Национальную исследовательскую компьютерную сеть, к которой подключатся все научные организации России

Доверенность (безопасность): развитие стандартов и требований

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 (**6 уровней доверия**)

ГОСТ Р 71206-2024 и ГОСТ Р 71207-2024 по безопасному компилятору и статическому анализу

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

- **Недостаточно использовать классические методы защиты (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

Ошибки в ПО – основная причина уязвимости систем

ИСП РАН

Принципиальное наличие уязвимостей* в ПО и аппаратуре:

функциональные, архитектурные, программного кода/микрокода.

***Границы между ошибками программиста, закладками, НДВ размыты**

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 ПО и программы для взлома
- Постоянные угрозы

Рост ресурсов и сложности атак

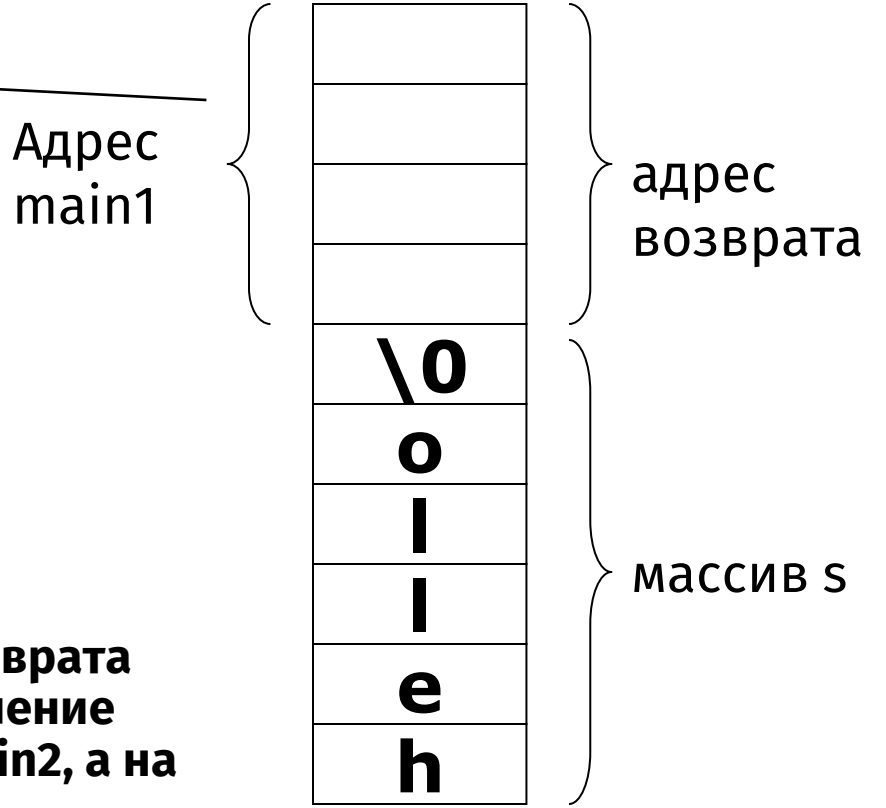
Пример I: переполнение буфера (уровень исходного кода)

ИСП РАН

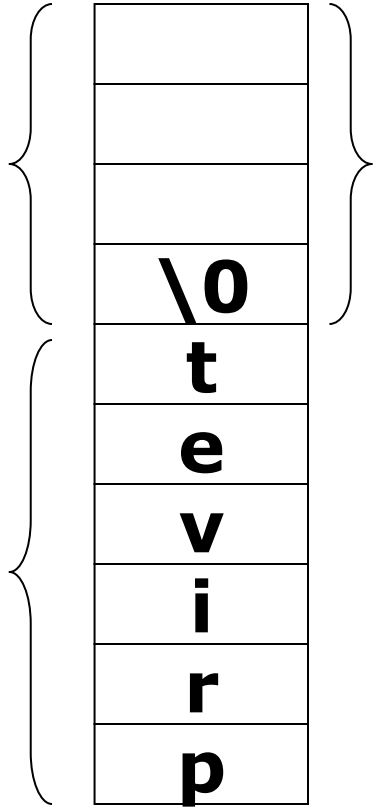
```
f(char * p)
{    char s[6];
     strcpy(s,p);
}
main1 ()
{    f("hello");
}
main2 ()
{    f("privet");
}
```

В случае main2 адрес возврата перезапишется, и управление будет передано не на main2, а на другой участок кода

Стек после выполнения функции f, вызванной из main1

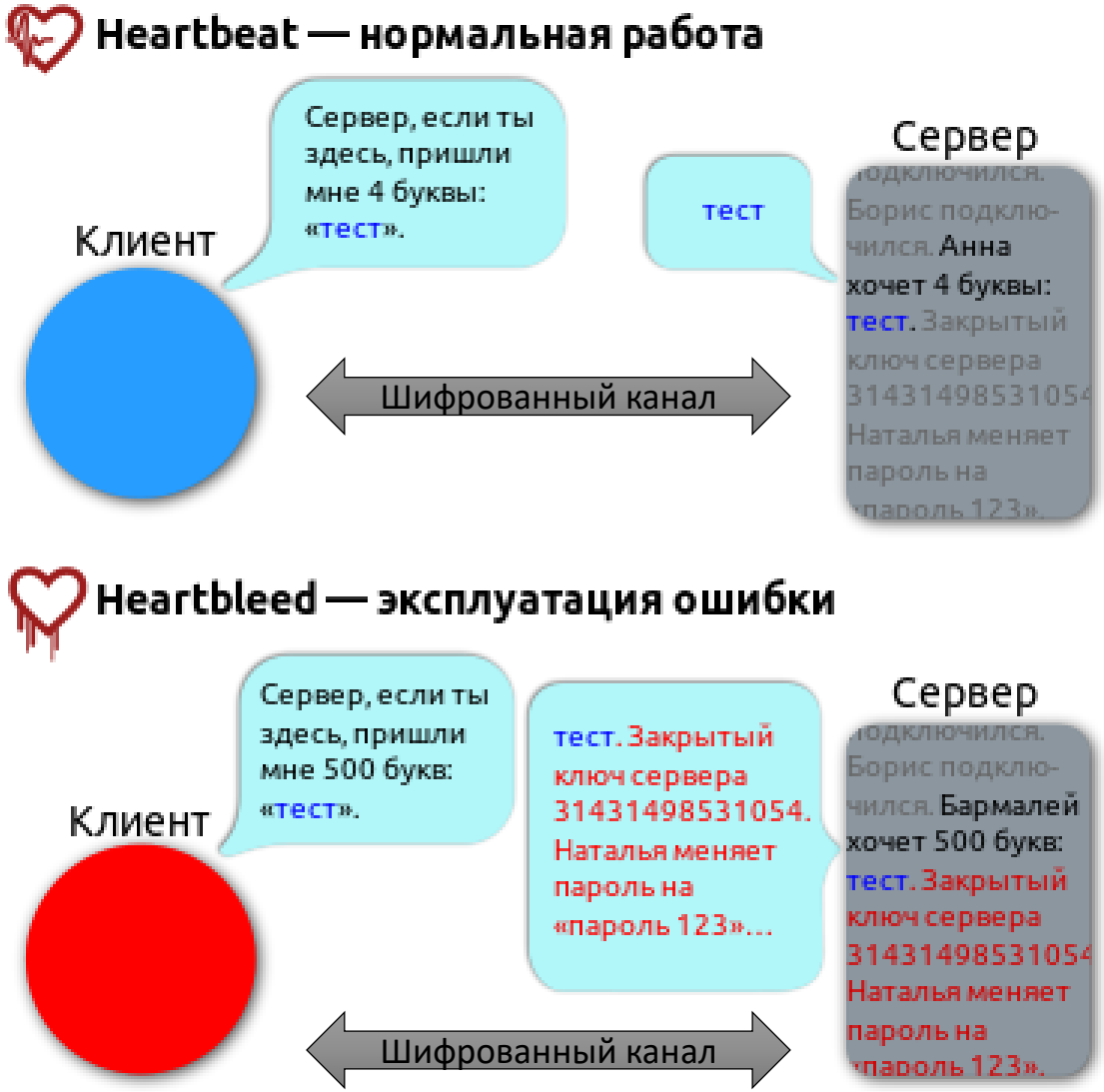


Стек после выполнения функции f, вызванной из main2



На место адреса main2 записали лишний байт

Реализация:
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

Пример II:
небезопасная оптимизация (среда компиляции и сборки)

ИСП РАН

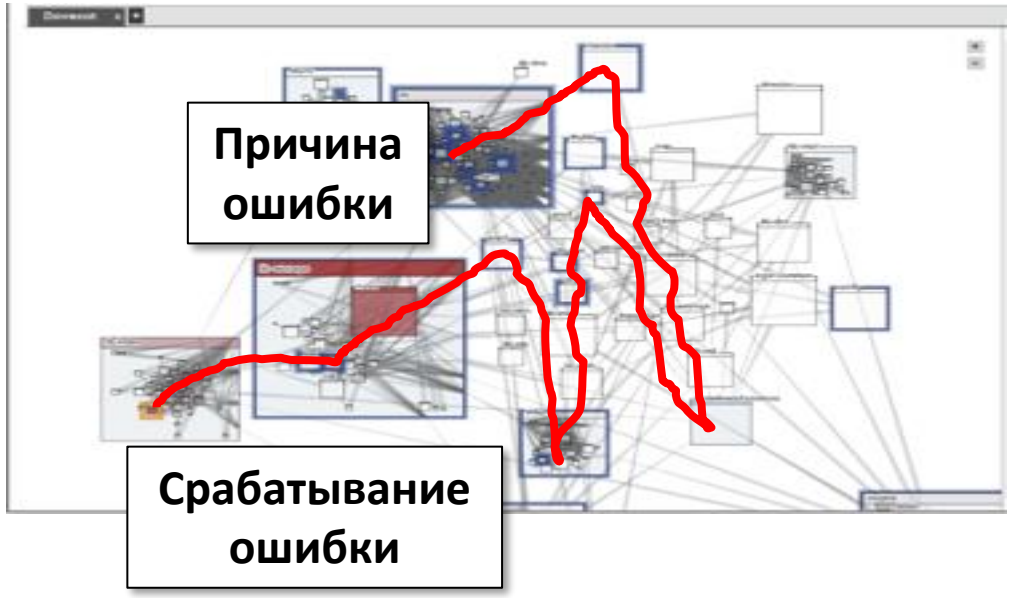
```
bool auth() {
  char buf[N];
  bool res;

  read_password(buf, N);
  res = check_password(buf);

  memset(buf, 0, N);
  return res;
}
```

Компилятор удаляет обнуление буфера с паролем, т.к. с его точки зрения после обнуления буфер не используется. При этом пароль останется на стеке

Пример III:
слабость кодирования обработки входных данных

ИСП РАН

Типы ошибки: слабость кодирования обработки входных данных, переполнение буфера



Прежде чем достичь места реализации ошибки, введённые извне данные «проходят» по многим функциям разных модулей

Модуль с функцией считывания файла-архива

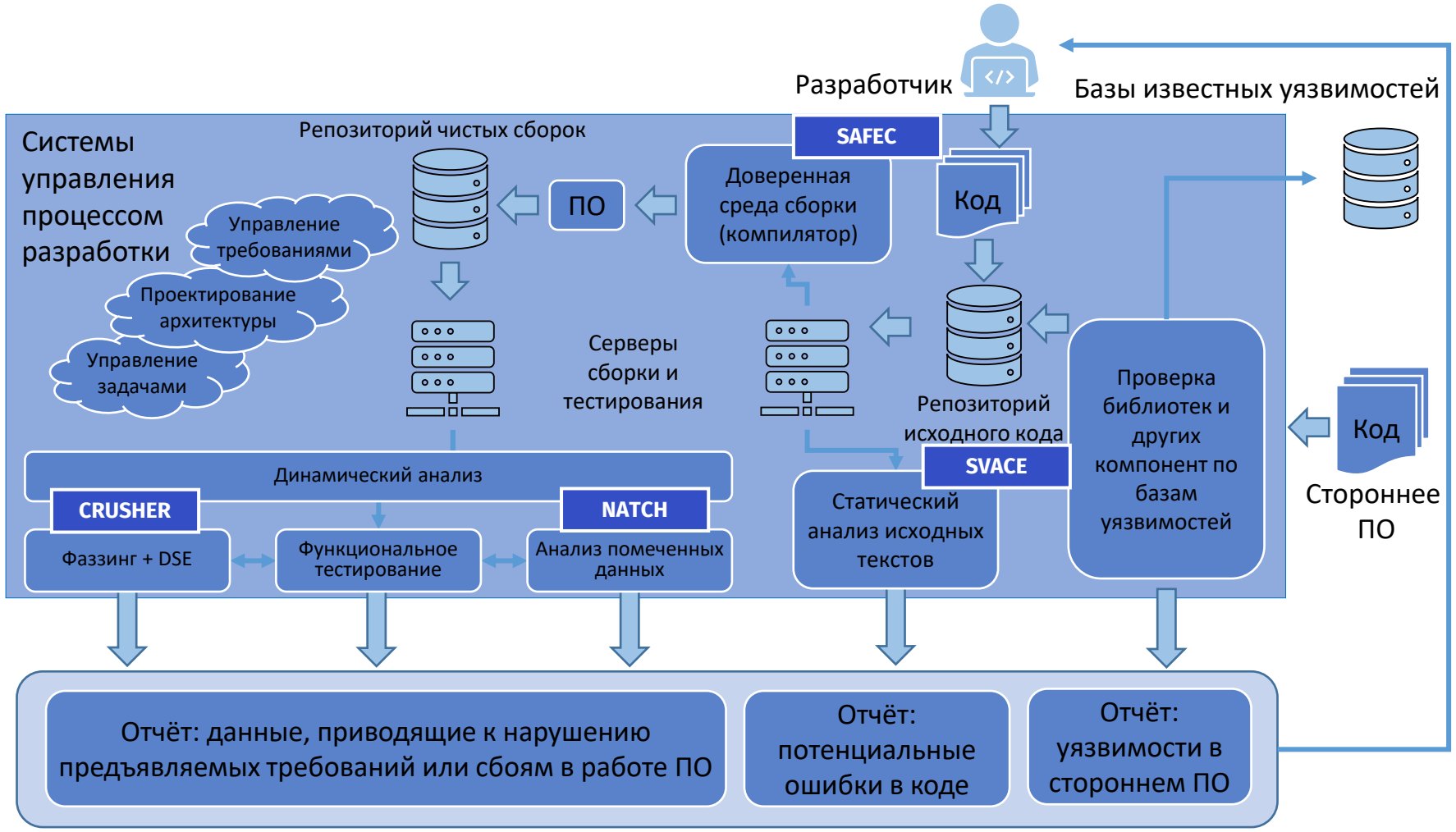
```
[tainted] Call of read
95              if(read(fd, file_hdr, SIZEOF_NEWLHD) != SIZEOF_NEWLHD) {
96                  free(file_hdr);
97                  return NULL;
98              }
99              file_hdr->flags = unrar_endian_convert_16(file_hdr->flags);
100             file_hdr->head_size = unrar_endian_convert_16(file_hdr->head_size);
101             file_hdr->pack_size = unrar_endian_convert_32(file_hdr->pack_size);
102             file_hdr->unpack_size = unrar_endian_convert_32(file_hdr->unpack_size);
103             file_hdr->file_crc = unrar_endian_convert_32(file_hdr->file_crc);
Composite 'file_hdr' taints element 'file_hdr->name_size'
104             file_hdr->name_size = unrar_endian_convert_16(file_hdr->name_size);
116             return file_hdr;
```

Функция в другом модуле. Ранее считанные извне данные определяют размер копируемой памяти

```
1484            /* Enter response type, length and copy payload */
1485            *bp++ = TLS1_HB_RESPONSE;
1486            s2n(payload, bp);
Tainted data from /home/shimnik/openssl/ssl/s3_pkt.c+239 reached a sink.
8. [SINK] *(s->s3->rrec.data + @ ) reaches the sink
1487            memcpy(bp, pl, payload);
1488            bp += payload;
1489            /* Random paddi
1490            RAND_pseudo_byt
1491
1492            r = dtls1_write                    3EAT, buffer, 3 + payload + paddi
```

Go to declaration
Go to definition
Show usage

Разработка безопасного ПО: типовой конвейер



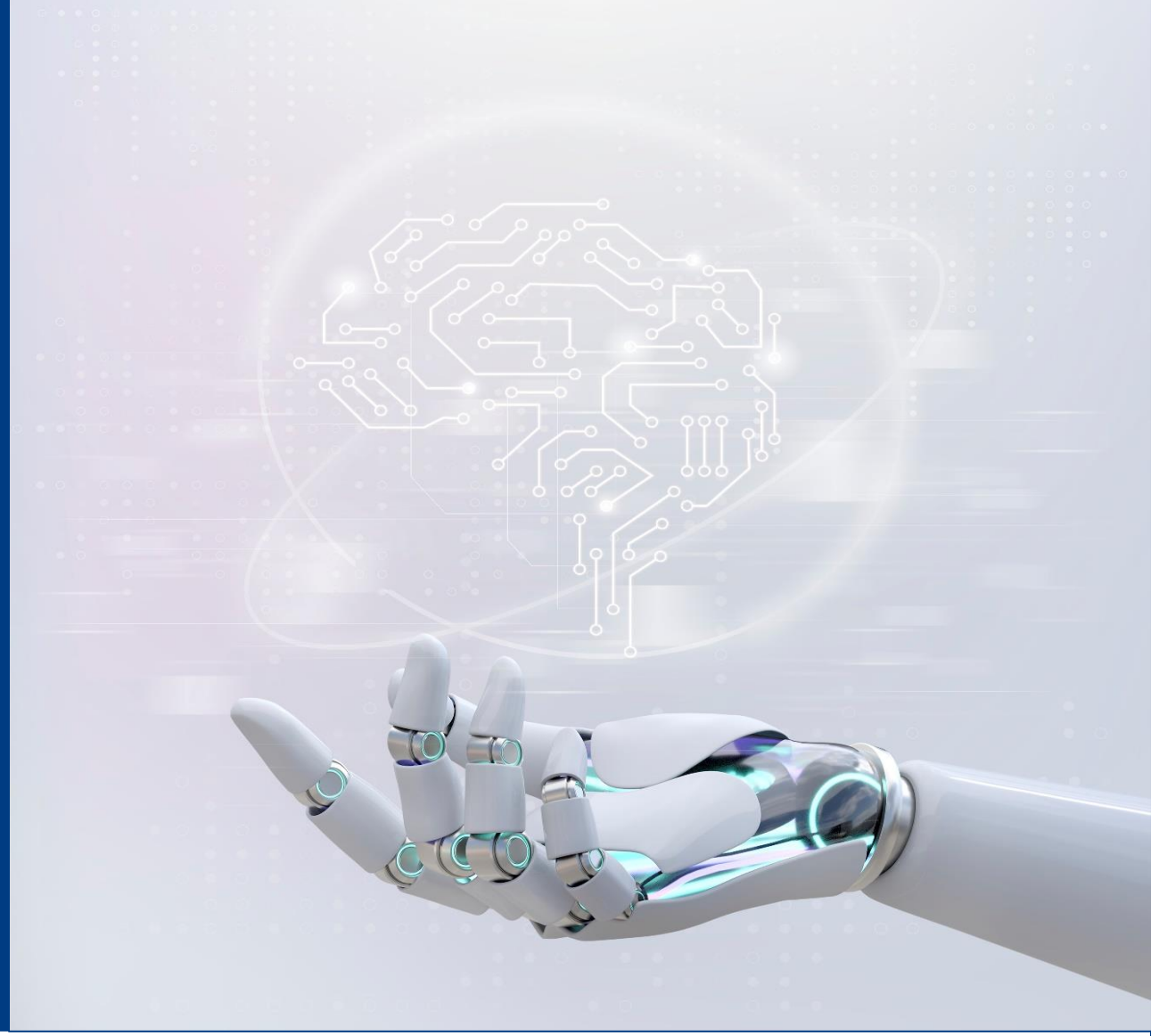
РБПО специализируется в каждой индустрии (авиация, автомобили, космос, госсектор) под нормативные документы и специфику ПО отрасли.

Технологический стек ИСП РАН развивается с 2002 г. на базе результатов фундаментальных научных исследований* по контрактам с зарубежными и российскими компаниями: «Лаборатория Касперского», ГК Astra Linux, СберТех и др.

200+ компаний используют наши технологии

***2021: новая специальность ВАК «Кибербезопасность» утверждена Приказом Минобрнауки №118**

Новый вызов: искусственный интеллект



Искусственный интеллект (ИИ) внедряется повсюду

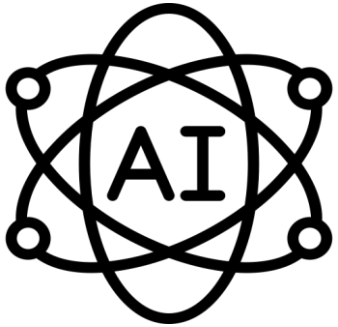
«Искусственный интеллект – комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые с результатами интеллектуальной деятельности человека или превосходящие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

Национальная стратегия развития искусственного интеллекта на период до 2030 года

Большие языковые модели и приложения
Системы машинного перевода, сервисы автоматической транскрибации, умные помощники в смартфонах и т.д.

Медицина
Компьютерная диагностика, подбор лечения. Фитнес-браслеты, глюкометры и другие устройства

Транспорт
Беспилотные автомобили



Системы безопасности
Распознавание лиц с помощью компьютерного зрения

Финансы
Обнаружение мошенничества и отмывания денег, кредитный скоринг, голосовые помощники и чат-боты

Исследование космоса
Автономная космическая навигация (роботы на Марсе)

Торговля
Рекомендации в ритейле, роботизация складского бизнеса

Промышленность
Роботизация производства

И будет внедряться дальше!

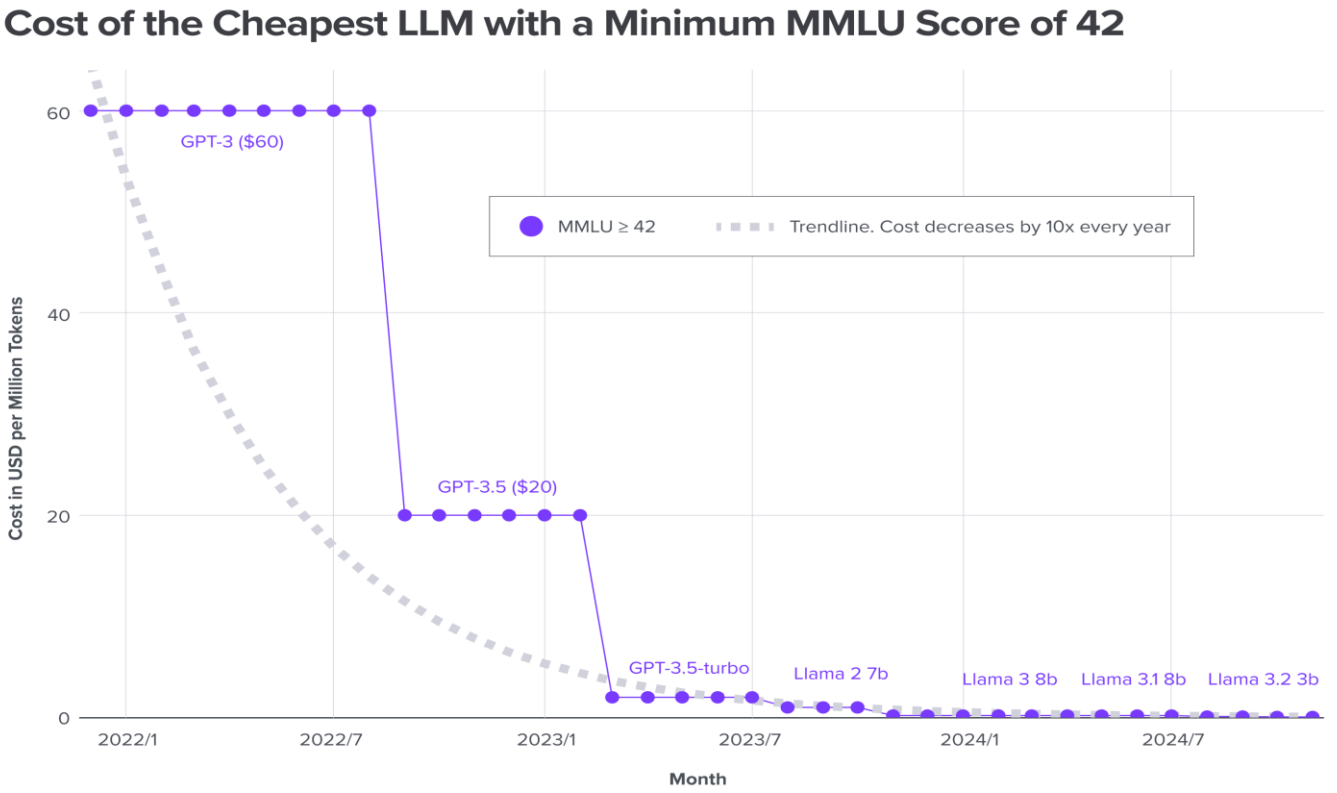
ИСП РАН

ИИ уже 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/

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
- **Представляет собой технологию без субъектности**

СИЛЬНЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ* (КОГДА?)
Strong AI, General AI

Методы: ?

- Делает интеллектуальные выводы
- Решает задачи на уровне человека
- Использует стратегии, планирует действия
- Функционирует в условия неопределенности
- Общается на естественном языке
- Способен к абстрактному мышлению
- **Пока не существует**

ИИ – разум ли это?
Иммануил Кант («Критика чистого разума», 1781): «разум есть способность, дающая нам принципы априорного знания».
Априорное знание – полученное до опыта. Апостериорное знание - эмпирическое знание, возможное только посредством опыта

Автор: Chris Noessel

2022: Холдинг Alphabet (Google) уволил старшего инженера-программиста Блейка Лемуана, который заявил, что у языкового чат-бота LaMDA есть собственный разум

***Автор терминов – американский философ Джон Сёрл. Впервые упомянуты в его статье «Разум мозга – компьютерная программа?» (1990)**

https://psychosearch.ru/teoriya/psikhika/338-searle-john-razum-mozga-kompyuternaya-programma

Здесь тоже важен открытый код!

ИСП РАН

2022: ChatGPT американской компании OpenAI (закрытое решение)

2023: Mistral 7B французского стартапа Mistral AI (открытое решение). Модель была выпущена под лицензией Apache 2.0, все могли ее использовать, изменять и распространять – с условием указания авторства изначального продукта

2023: LLaMa американской компании Meta AI (открытое решение)

2025: DeepSeek R1 китайской компании DeepSeek (открытое решение): значительно дешевле ChatGPT, нет региональных ограничений

Через несколько дней после релиза DeepSeek обогнал ChatGPT по числу загрузок в магазине приложений Apple. Успех DeepSeek привел к рекордному падению акций Nvidia на 17%

30 ноября 2022: появление генеративного ИИ (ChatGPT от OpenAI)

Открытые решения позволяют быстрее и дешевле развиваться, поэтому здесь этот тренд тоже победит

Эффективность и продуктивность: средства разработки искусственного интеллекта

ИСП РАН

Основные открытые фреймворки – PyTorch и TensorFlow:

- Разрабатываются под руководством крупной компании (Meta AI, Google) крупными сообществами
- Активная разработка – десятки изменений кода в день
- Написаны на языках C++ (вычислительные ядра) и Python (прикладные интерфейсы)
- Скорость работы достигается за счет:
  - ✓ поддержки современных процессоров x86-64 и ARM,
  - ✓ низкоуровневых оптимизаций под конкретные архитектуры,
  - ✓ специализированных математических библиотек (Intel MKL),
  - ✓ графических ускорителей (NVIDIA CUDA, AMD ROCm, macOS GPU…)
- Поддержка произвольных ускорителей для инференса моделей – через открытые стандарты OpenCL, ONNX, Apache TVM с адаптацией под конкретную архитектуру

CatBoost – открытый фреймворк компании Яндекс:

- Несмотря на открытость, сообщество в полной мере построить не удалось (в 10 раз меньше разработчиков по сравнению с PyTorch)

Пример отечественного ПО: фреймворк NEUROMATRIX® DEEP LEARNING (НТЦ «Модуль») позволяет конвертировать предобученные ONNX-модели в собственный формат nm8 для инференса на собственных ускорителях

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 их 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 зависит 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

В СЛУЧАЕ ИИ КИБЕРБЕЗОПАСНОСТЬ – ТОЛЬКО ЧАСТЬ ДОВЕРЕННОСТИ

«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...»

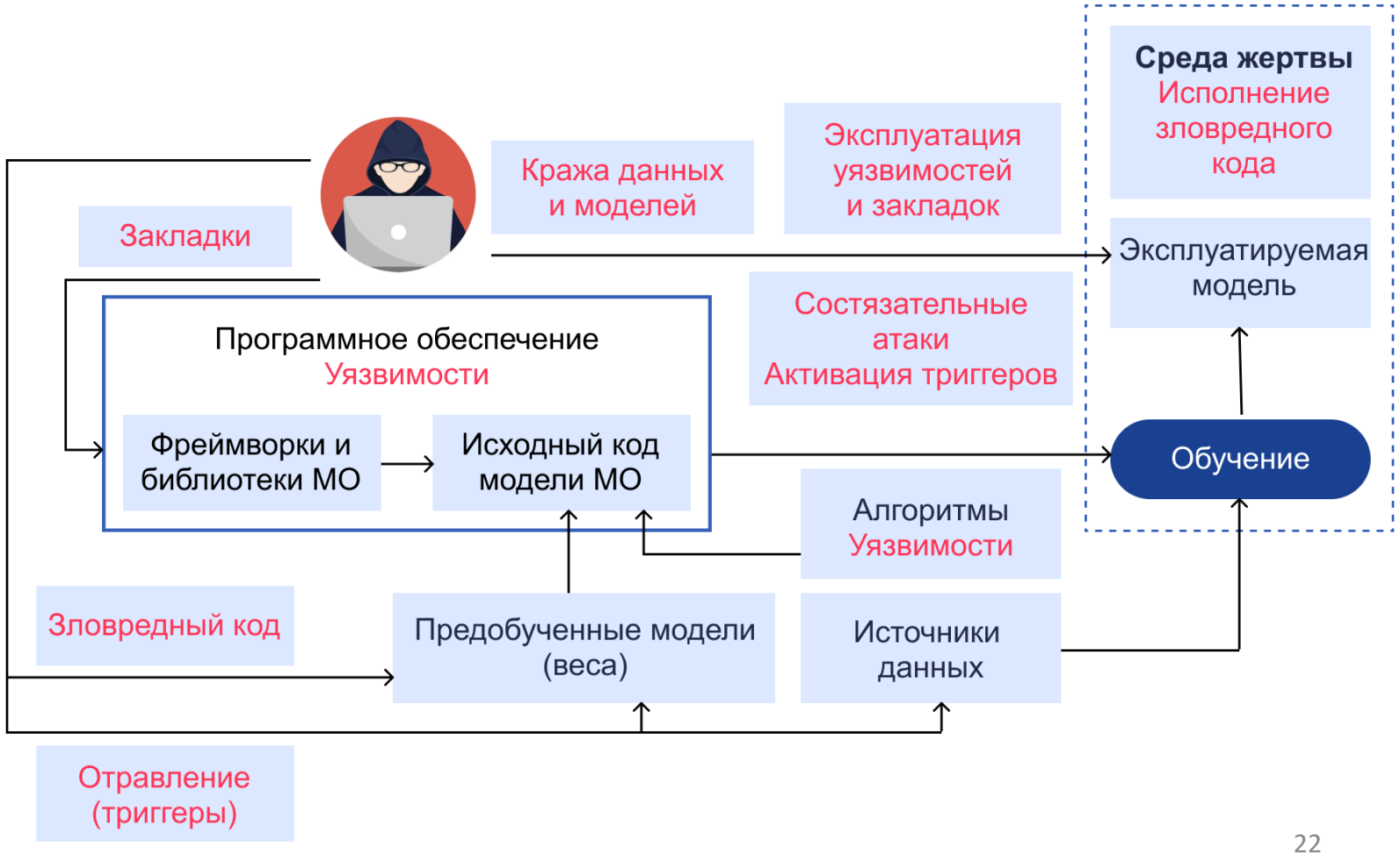
Отчёт OOH Governing AI for Humanity 2024

Вначале про кибербезопасность: источники угроз в ИИ

ИСП РАН

Особенность систем с ИИ: основная информация содержится не в программном коде, а в данных!
Поэтому традиционные инструменты разработки безопасного ПО малоприменимы к ИИ

- **Исходный код инфраструктур машинного обучения** (уязвимости, закладки)
- **Данные** (отравление данных, кража из облачных сред)
- **Алгоритмы** (предобученные модели с закладками или вредоносным ПО)



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

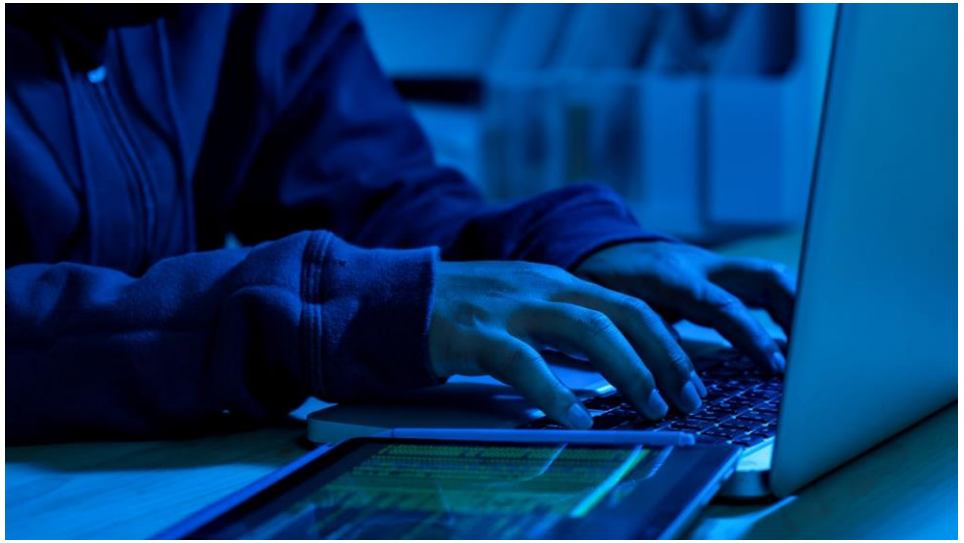
С помощью фейковой видеоконференции преступники вынудили сотрудника транснациональной корпорации перевести им $25,6 млн.

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

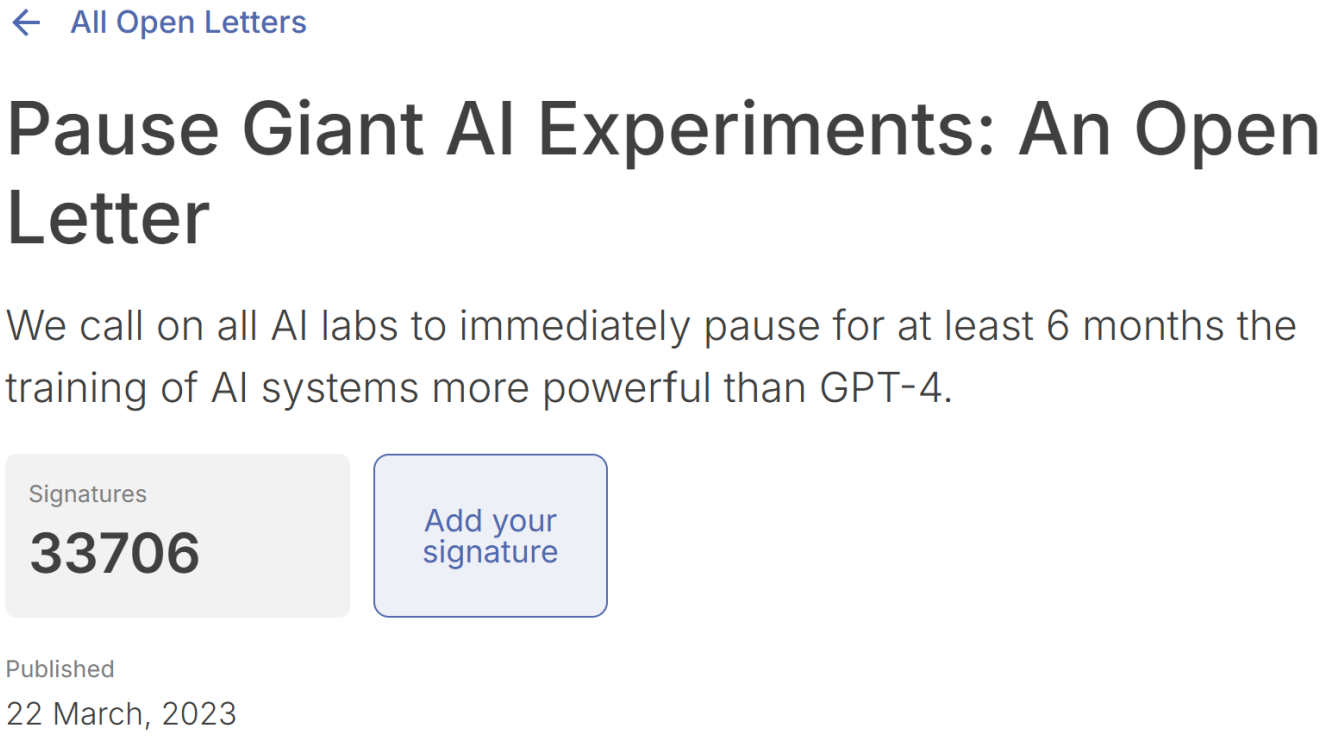
… и это далеко не все угрозы!

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/

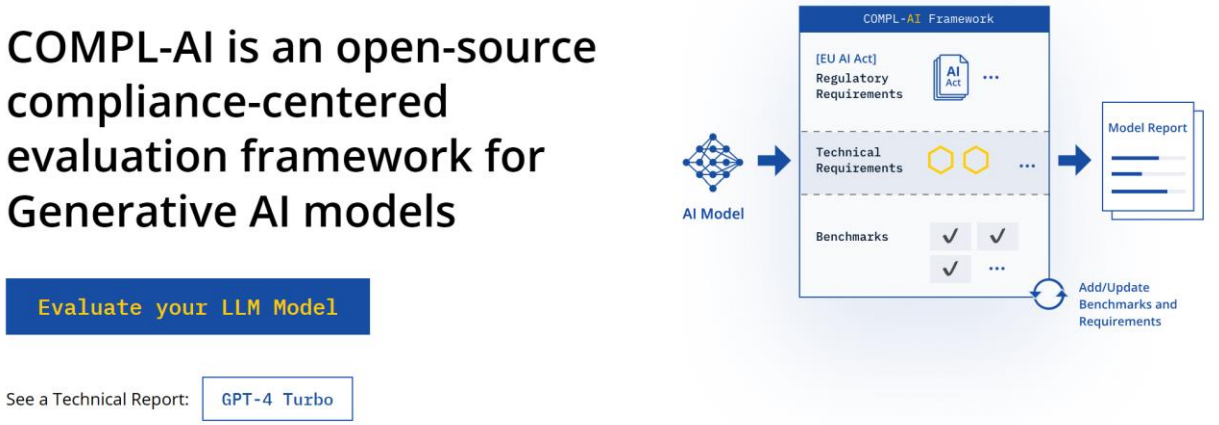
Фреймворк для оценки больших языковых моделей на соответствие EU AI Act – главному европейскому закону об искусственном интеллекте

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

Декларация об ответственной разработке и использовании сервисов на основе генеративного ИИ

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

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

Доверенный ИИ: этапы разработки, методы и технологии

ИСП РАН

Эксплуатация моделей

- Методы защиты моделей от атак на этапе эксплуатации
- Методы обнаружения дрейфа данных
- Инструменты оценки устойчивости обученных моделей к атакам
- Инструменты маркирования сгенерированного контента

Обучение моделей

- Методы объяснения моделей
- Инструменты для повышения доверия к предобученным моделям
- Методы федеративного обучения
- Методы выявления предвзятости моделей

Датасеты

- Инструменты проверки наличия аномалий в наборах данных

Фреймворки

- Доверенные фреймворки и библиотеки машинного обучения

Всё это требуется для создания доверенных систем с использованием ИИ

В ИСП РАН ведутся исследования и разработки методов и технологий, необходимых на всех этих этапах

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 2024 году: 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)

Программные инструменты для борьбы с угрозами разработаны в Центре в 2021-2024

ИСП РАН

Инструменты предназначены для:

- **тестирования моделей машинного обучения на устойчивость к состязательным атакам**
  для задач классификации/сегментации изображений, детекции объектов, распознавание речи, классификации текстов, оценки качества изображений/видео
- **защиты моделей машинного обучения от состязательных атак**
- **защиты от копирования обученных моделей машинного обучения**
- **защиты от извлечения обучающих данных из обученных моделей**
- **выявления и устранения закладок и зловредного кода в предобученных моделях машинного обучения**
- **объяснения моделей**
- **обнаружения обнаружения аномалий и дрейфа данных**
- **выявления предвзятости моделей**

...и нужны не только для обеспечения кибербезопасности, но и для борьбы с социогуманитарными угрозами!

Примеры инструментов

Инструмент защиты моделей от состязательных атак

Для задач классификации/сегментации изображений, детекции объектов, распознавания речи, классификации текстов, оценки качества изображений/видео

- ✓ Алгоритмы состязательного обучения
- ✓ Улучшенные алгоритмы состязательного обучения (Fast adversarial training, TRADES, A2T)
- ✓ Алгоритмы обеспечения сертифицированной устойчивости (Smooth Adversarial Training, DensePure, CC-Cert, Certified Robustness to Adversarial Word Substitutions)
- ✓ Алгоритмы аугментации изображений для задачи классификации (Pad & Crop, CutOut, CutMix, MixUp)
- ✓ Алгоритмы расширения обучающего набора (Denoising Diffusion Probabilistic Model, с применением подхода самообучения)
- ✓ Алгоритмы, основанные на модификации архитектуры нейронной сети (Robust Principles: Architectural Design Principles for Adversarially Robust CNNs)

ПО для выявления и устранения закладок и зловредного кода в предобученных моделях машинного обучения

- ✓ Алгоритмы встраивания закладок
- ✓ Алгоритмы встраивания зловредного кода
- ✓ Инструменты обнаружения атак с внедрением закладок
- ✓ Средства устранения имеющихся закладок в обученных нейросетевых моделях
- ✓ Средства защиты от эксплуатации зловредного кода в составе обученных моделей: (алгоритмы защиты на основе дистилляции знаний и квантования обученных параметров нейронной сети)
- ✓ Изолированное окружение для мониторинга активности приложения при работе с моделью ИИ, с возможностью сбора логов активности приложения и поиском известных IoC

Атаки через закладки

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

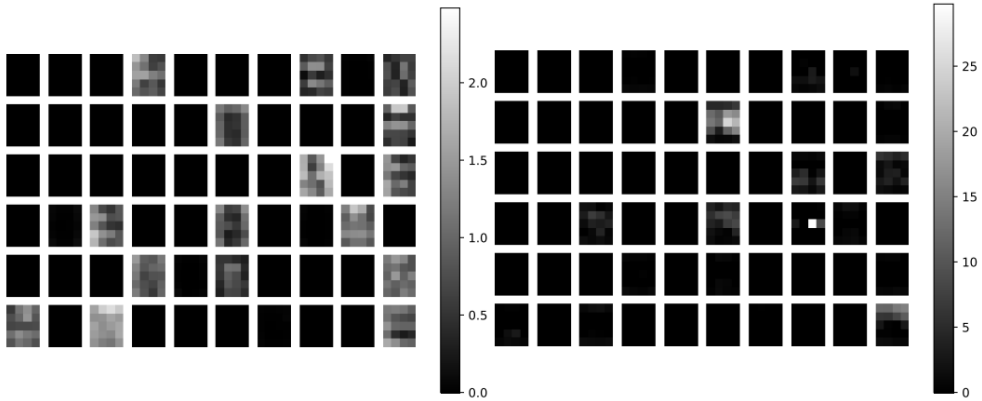
Свойства **успешных** атак через закладки:

- 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)

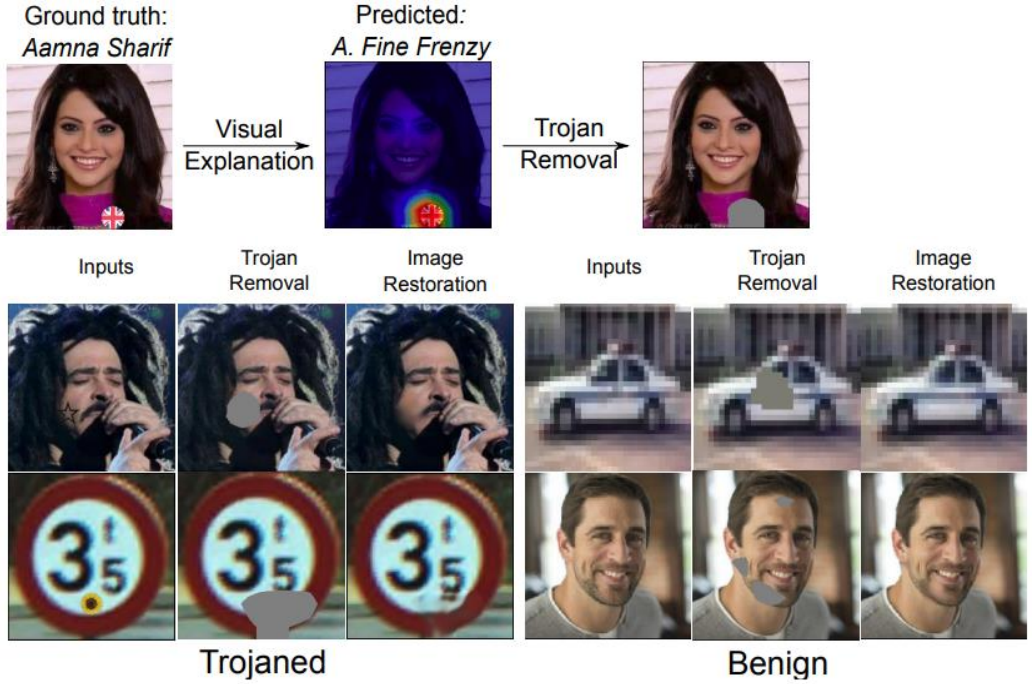
Наш результат:

- Новая атака через закладки на глубокие классификаторы текста через переупорядочивания слов
- 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



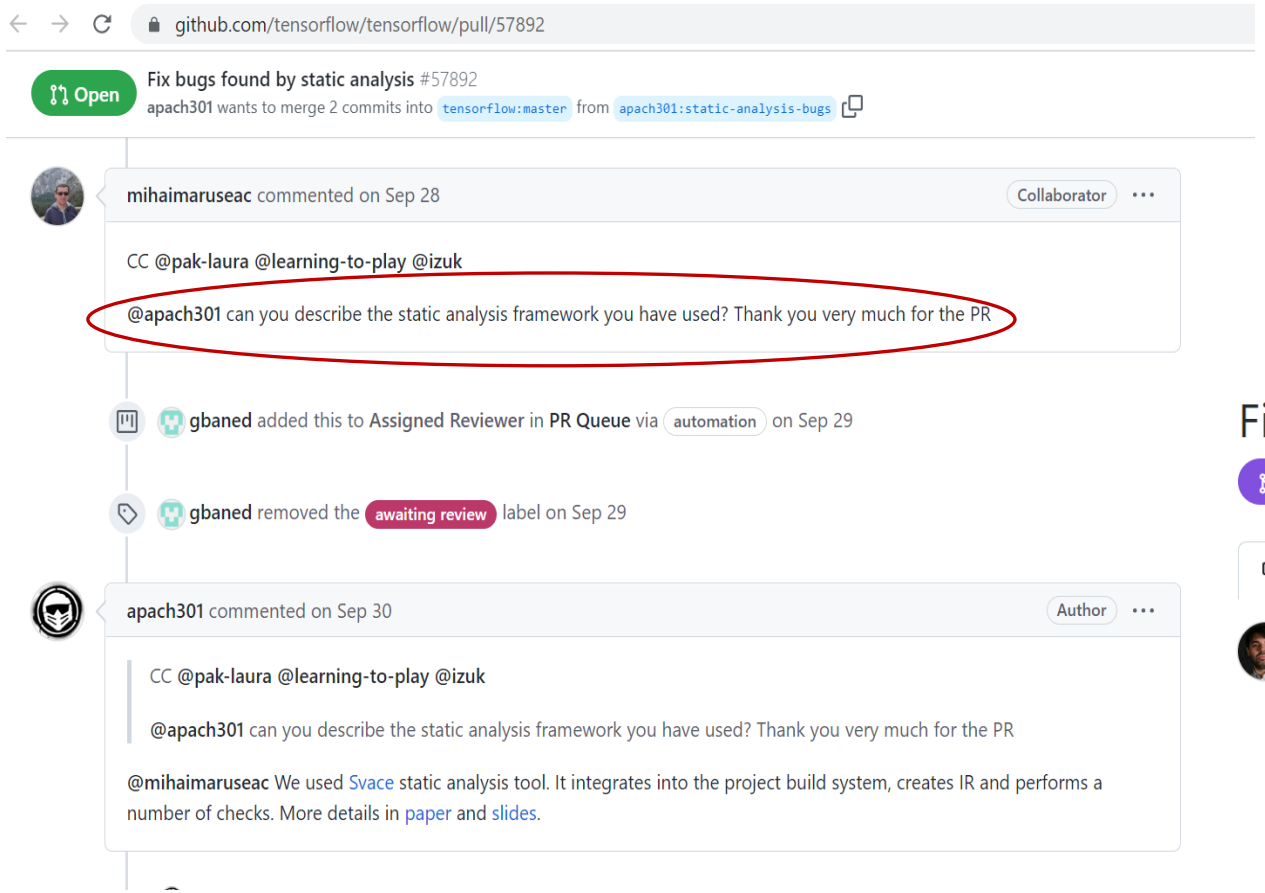
...и другие подходы

Доверенные фреймворки машинного обучения TrustFlow* and TrustTorch*

ИСП РАН

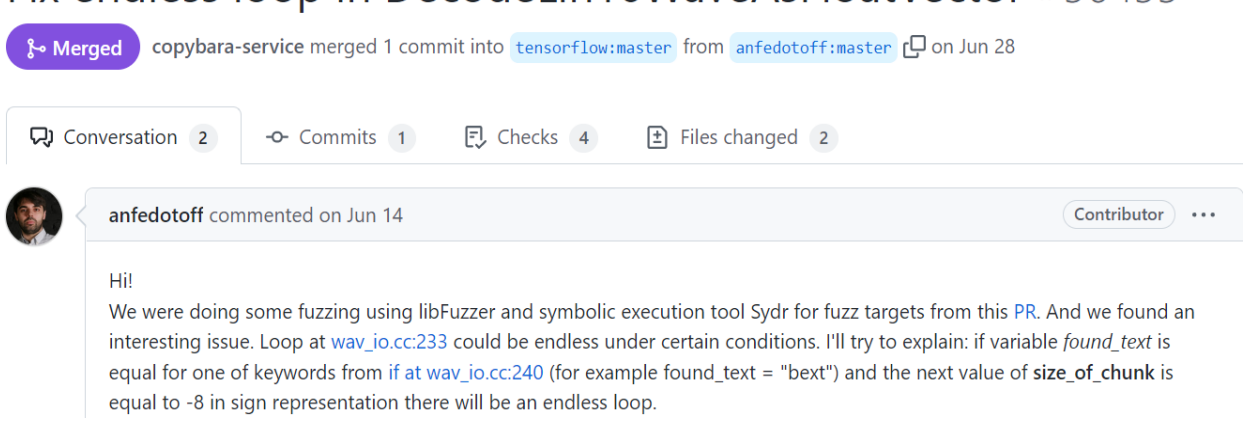
**Доверенные версии PyTorch и TensorFlow*

Проанализированы инструментами ИСП РАН Svace (статический анализ) и Sydr (гибридный фаззинг)



80+

патчей принято в PyTorch и TensorFlow



Фреймворки TrustFlow и TrustTorch постоянно обновляются; они включены в Платформу доверенного ИИ для разработки соответствующих систем

Другие направления: федеративное обучение для обеспечения безопасности

Молодёжная лаборатория по федеративному обучению (при поддержке Минобрнауки России)

Научная тема:
Разработка алгоритмов федеративного обучения: безопасное и эффективное обучение больших моделей с приложениями к медицинским задачам (2024-2026)

Проект направлен на на решение ключевых вопросов федеративного обучения не только с теоретической точки зрения, но и с точки зрения практического применения полученных результатов к медицинским задачам

Некоторые направления проекта:

- способы защиты приватности
- атаки на приватность обучения
- атаки на качество обучения

ИСП РАН + Яндекс + Сеченовский университет

2024:

Впервые в России проведён эксперимент по обучению модели классификации 12-канальных ЭКГ с использованием алгоритмов федеративного машинного обучения.

Яндекс, ИСП РАН и Сеченовский Университет, используя федеративный подход, создали нейросеть, которая по данным электрокардиограмм выявляет фибрилляцию предсердий — одну из наиболее распространенных патологий сердца. Технология делает это с высокими показателями чувствительности и специфичности.

https://habr.com/ru/news/849144/

Другие направления: 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**

Другие направления: 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**

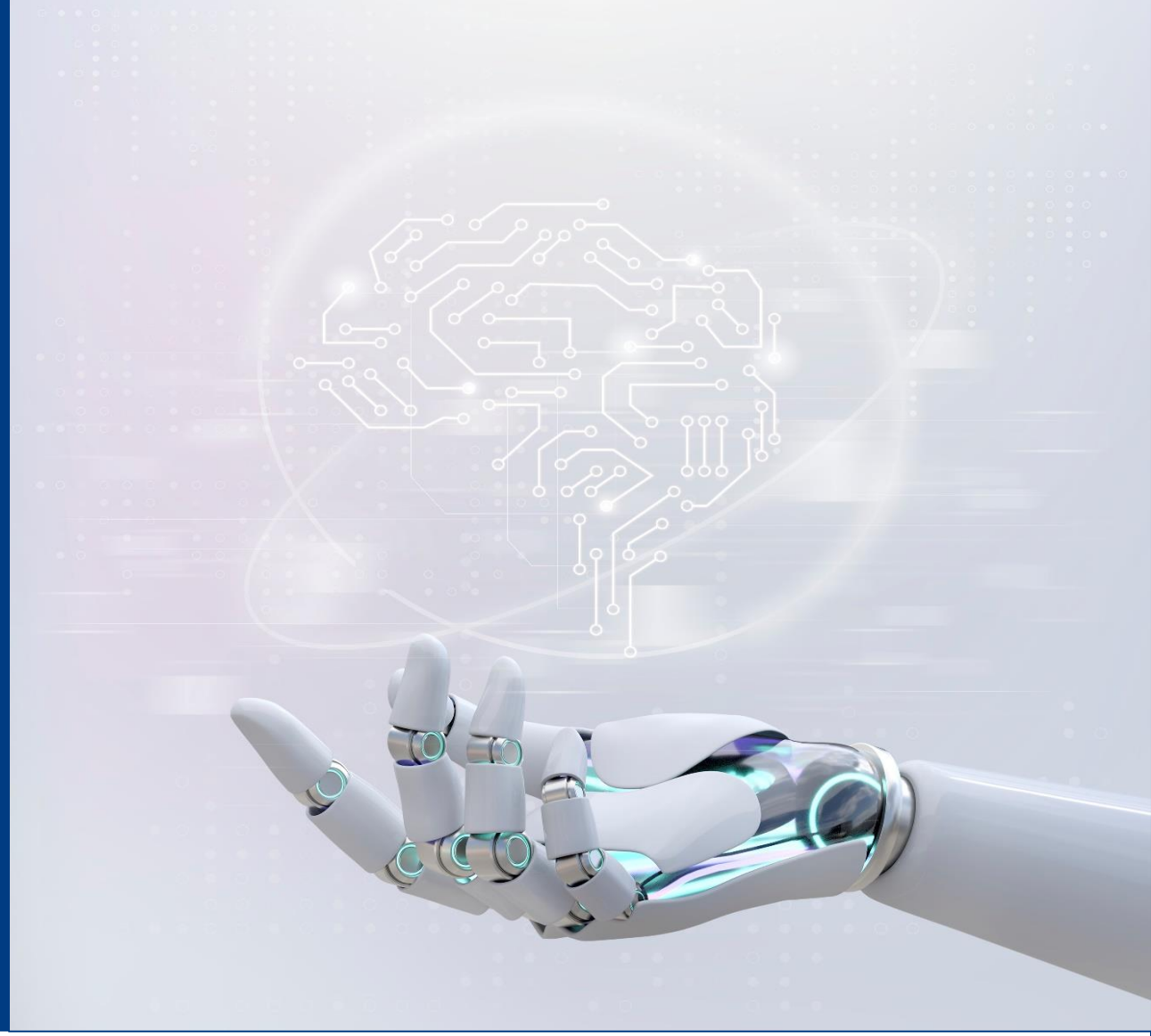
В числе задач:
- анализ существующих методов внедрения цифровых водяных знаков в контент, синтезированный генеративными моделями
- исследование особенностей применения определенного идентификатора в качестве backdoor (управляющего сигнала) в задачах генерации контента
- исследование особенностей внедрения 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**

Как развиваться дальше?
Только совместными усилиями!



Предлагаемая модель 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~~

Пример реализации модели

Центр исследований безопасности системного ПО: достижения

ИСП РАН

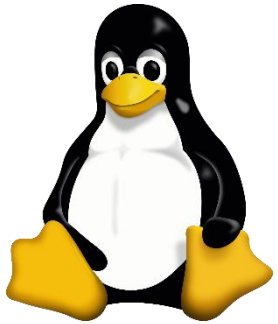
Создан на базе ИСП РАН по инициативе ФСТЭК России

Исследования ведут >70 компаний и вузов

Результаты:

- ведётся сопровождение собственых веток ядра **Linux 5.10 и 6.1**;
- выявлены **>30 критических уязвимостей** в ядре Linux;
- **>500 исправлений** уже приняты в основную ветку ядра;
- **>100** исправлений приняты в основные ветки компонентов OpenSSL, Qemu, libvirt, CPython, Lua, .NET6 Runtime;
- ведутся доработки ядра, нацеленные на повышение его безопасности;
- и главное: возникло растущее сообщество из **>80 специалистов**, которые проводят работы (решение проблемы кадрового голода)

! Создана масштабируемая экосистема, которая обеспечивает генерацию кадров и технологий и как следствие – технологическую независимость



QEMU

libvirt VIRTUALIZATION API

OpenSSL Cryptography and SSL/TLS Toolkit

Наш результат: единая сервисная платформа на основе стека технологий ИСП РАН

ИСП РАН

Физика высоких энергий | Генетика | Почвоведение | Медицина | Материаловедение

- ✓ **Научная основа**
- ✓ **Открытый код**
- ✓ **Взаимодействие с индустрией**

Прикладные сервисы

- Группа сервисов обработки данных в области физики высоких энергий
- Группа сервисов анализа данных и моделирования почв
- Группа сервисов анализа данных биомедицинского домена

- Разметка данных с учетом специфики решаемых задач
- Полный жизненный цикл нейросетевых моделей
- Представление знаний
- Обработка текстов

Инструменты сбора, хранения и анализа больших данных

- TALISMAN (ИСП РАН)
- Репозиторий доверенного ПО и наборов данных (ИСП РАН)
- Конвейер разработки доверенного ПО (ИСП РАН)
- Платформа создания ИИ моделей в медицине (НЦМУ совместно с Сеченовским Университетом) (ИСП РАН)

- Инструмент работы с большими данными
- Интеграция с внешними и внутренними «поставщиками» данных
- Технологии AI, ML, DL
- AI Ready данные

Облачная среда

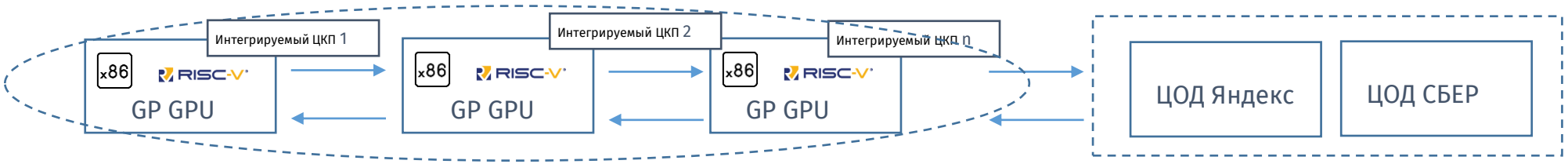
ASPERITAS (сертификат ФСТЭК России) (ИСП РАН)

Облачная платформа: высокий уровень адаптации инфраструктуры под конкретные классы задач, , заданный уровень доверия в соответствии с требованиями ФСТЭК России

FANLIGHT (в Едином реестре российского ПО). Платформа коллаборативной работы в рамках тематических Web-лабораторий

- Миграция legacy-систем в облако
- Инфраструктура как сервис на базе Openstack (IaaS)
- Платформенные сервисы по запросу (PaaS)
- Микросервисная архитектура и контейнеризация

Сеть ЦОД



Спасибо за внимание!