ПРОБЛЕМА ДОВЕРИЯ ТЕХНОЛОГИЙ КОМПЬЮТЕРНОГО ЗРЕНИЯ И СПОСОБЫ ЕЕ РЕШЕНИЯ

ГЛАВНЫЙ РАДИОЧАСТОТНЫЙ ЦЕНТР

ФЕДОТОВ
Александр Владимирович

Руководитель научно-технического центра ФГУП «ГРЧЦ»

2022 г.

РЕЗУЛЬТАТЫ ОПРОСА*

Доверяете ли вы СИИ?

*Результаты опроса российских респондентов, проведенного АНО «Национальные приоритеты» и ВЦИОМ

48% СКОРЕЕ «ДА»

42% СКОРЕЕ «НЕТ»

СРЕДНЕМИРОВОЙ УРОВЕНЬ ДОВЕРИЯ – 28%

ОСНОВНЫЕ ПРИЧИНЫ НЕДОВЕРИЯ СИ

18% ТЕХНОЛОГИИ ПЛОХО РАЗВИТЫ / НЕ ИЗУЧЕНЫ

15% СБОИ / ОШИБКИ

14% «ЧЕЛОВЕКА НЕ ЗАМЕНИТЬ»

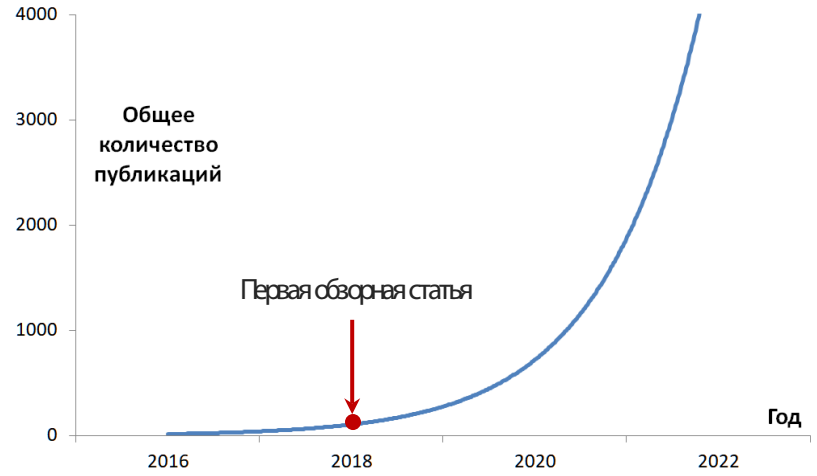
9% УГРОЗЫ БЕЗОПАСНОСТИ / ПОДВЕРЖЕНЫ АТАКАМ

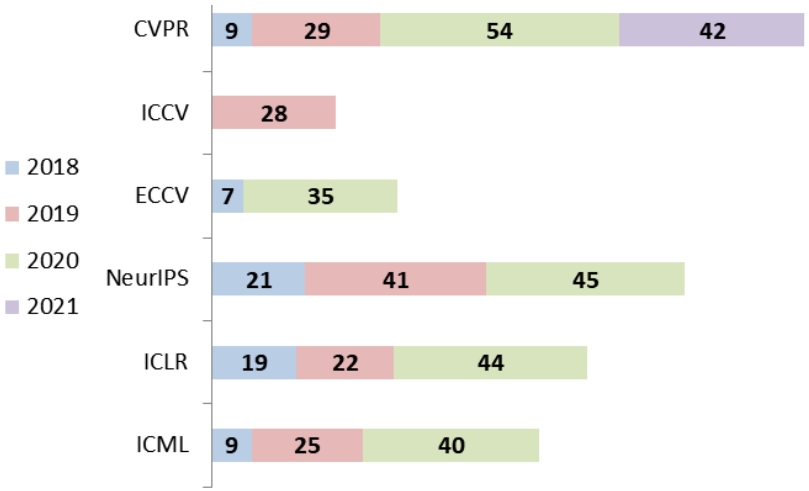
Доверие к системам искусственного интеллекта

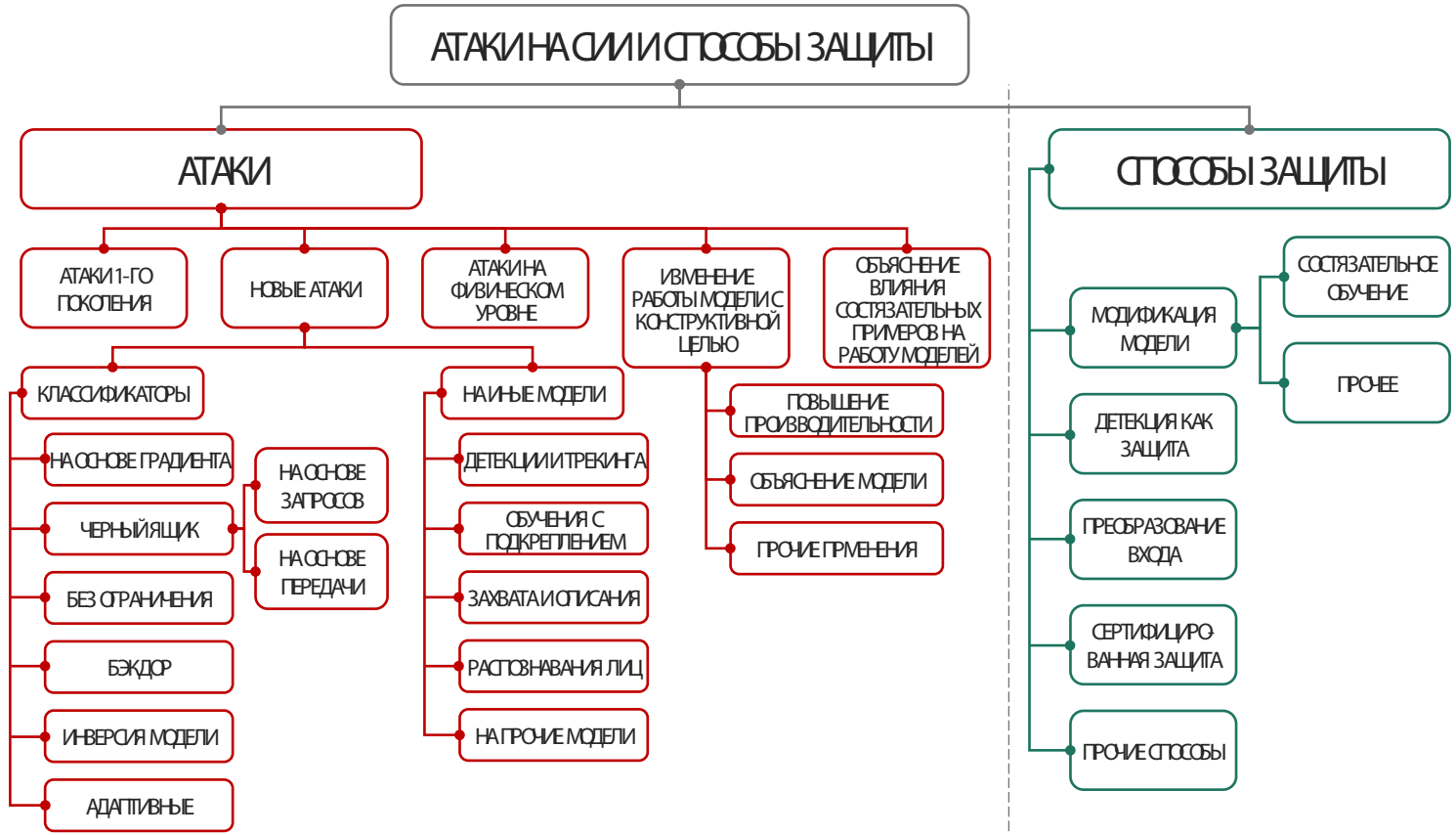
ЭЛЕМЕНТЫ ДОВЕРЕННОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Актуальность темы атак на СИИ и способов защиты от них

Количество публикаций по тематике атак на СИИ в научной литературе







Подробнее:

Риски и угрозы, возникающие вследствие атак на СИИ

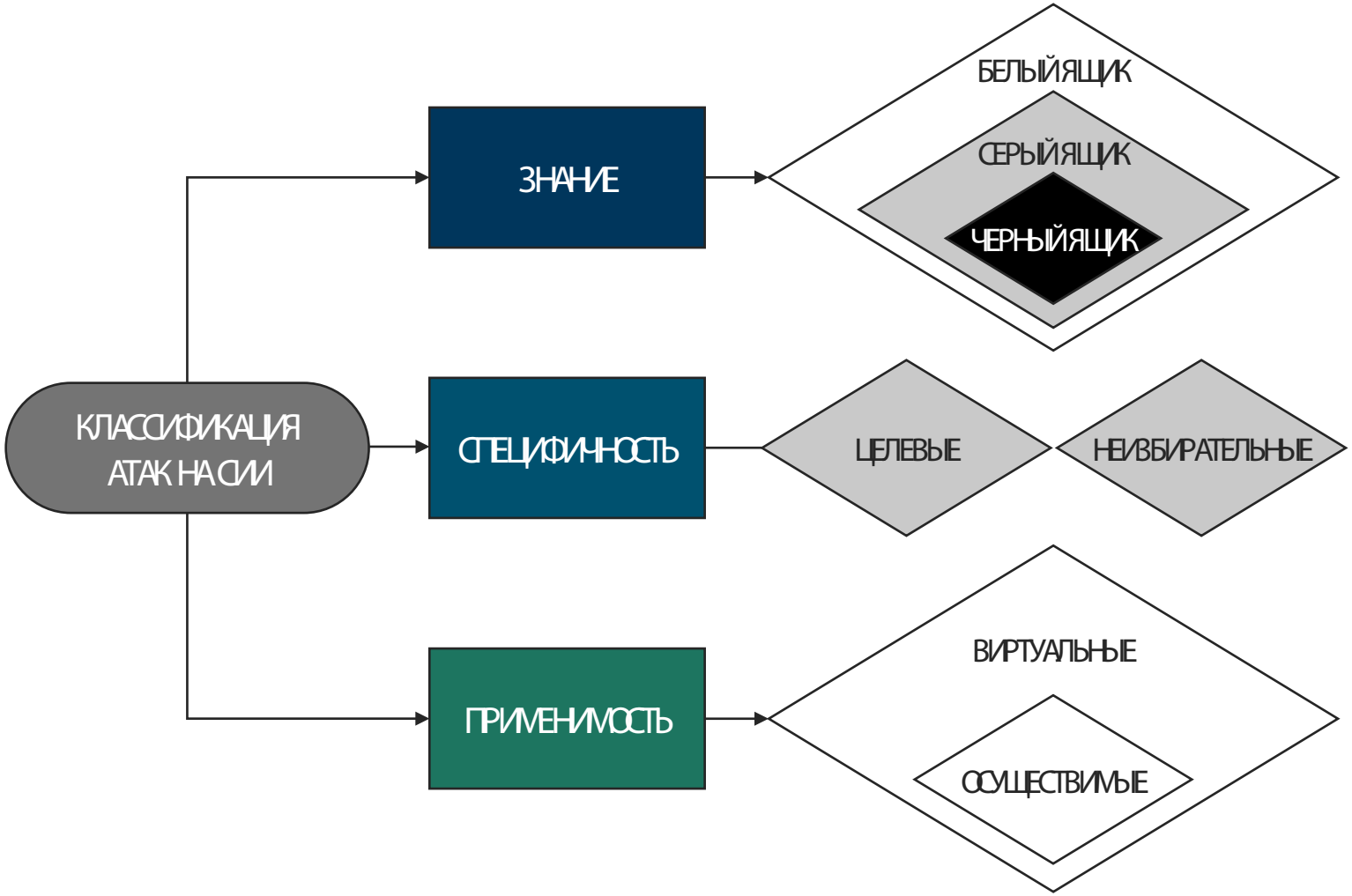
Факторы риска, заложенные в СИИ:

- НЕПОНИМАНИЕ И НЕПРЕДСКАЗУЕМОСТЬ ДЕЙСТВИЙ АЛГОРИТМОВ
- НЕДОСТАТОЧНАЯ УСТОЙЧИВОСТЬ И НАДЕЖНОСТЬ СИСТЕМ ПРИНЯТИЯ РЕШЕНИЙ

Возможные последствия и угрозы:

- ПРИНЯТИЕ СИИ И/ИЛИ РЕКОМЕНДАЦИЯ НЕВЕРНЫХ РЕШЕНИЙ ЧЕЛОВЕКУ
- МАНИПУЛЯЦИЯ ОБЩЕСТВЕННЫМ МНЕНИЕМ, В ТОМ ЧИСЛЕ С ИСПОЛЬЗОВАНИЕМ ФЕЙКОВ (СМИ, СОЦИАЛЬНЫЕ СЕТИ, РЕКОМЕНДАТЕЛЬНЫЕ СЕРВИСЫ)
- ДИСКРИМИНАЦИЯ ЛЮДЕЙ ПО ОПРЕДЕЛЕННОМУ ПРИЗНАКУ (СИСТЕМЫ СКОРИНГА, СОЦИАЛЬНОГО РЕЙТИНГА И ДР.)
- ВЫХОД СИИ ИЗ-ПОД КОНТРОЛЯ И ПРИЧИНЕНИЕ ВРЕДА ЗДОРОВЬЮ И ЖИЗНИ ЧЕЛОВЕКУ (АВТОТРАНСПОРТ, ЗДРАВООХРАНЕНИЕ, ОБОРОНА)

Таксономия атак на СИИ



Подробнее:

ТИПЫ АТАК НА СИИ

«evasion»
(атаки уклонения)

«poisoning»
(атаки отравления модели)

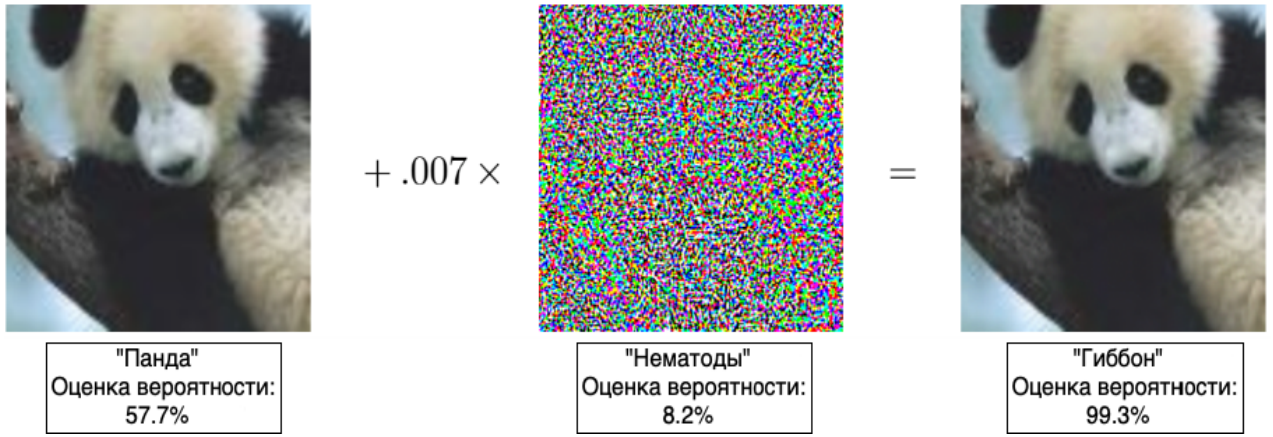
«model inversion»
(атаки инверсии модели)

«model extraction» / «data-free model extraction»
(атаки извлечения модели с использованием и без использования данных)

Атаки уклонения на модели компьютерного зрения

«evasion»
(атаки уклонения)

Атака типа «evasion» предполагает внесение «шумов» в обученные СИИ в целях нарушения корректности их работы



Подробнее:

Атаки инверсии на модели компьютерного зрения

«model inversion»
(атаки инверсии модели)

Атака типа «model inversion» предполагает вывод из модели обучающих данных и восстановление принадлежности данных или свойств данных

Изображение, которое удалось получить атакой на модель распознавания лиц методом инверсии (справа), и исходное изображение из обучающего датасета

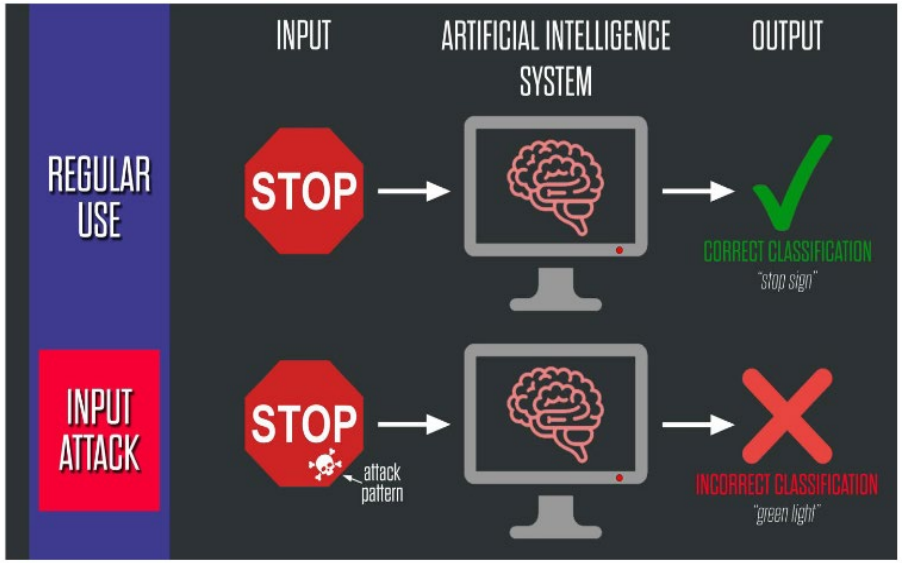
Подробнее:

«poisoning»
(атаки отравления модели)

Атаки отравления модели компьютерного зрения

Атака типа poisoning предполагает искажение обучающих данных или искажение работы алгоритмов ИИ

Подробнее:



Уровни доступа злоумышленника к целевой модели

«БЕЛЫЙ ЯЩИК»

ПОЛНОЕ ВЛАДЕНИЕ ИНСАЙДЕРСКОЙ ИНФОРМАЦИЕЙ О МОДЕЛИ, ВЕСАХ, СКОРОСТИ, МЕТОДАХ ОБУЧЕНИЯ

«СЕРЫЙ / ЧЕРНЫЙ ЯЩИК С ОЦЕНКОЙ»

ДОСТУП К НЕКОТОРЫМ ДАННЫМ МОДЕЛИ

«ЧЕРНЫЙ ЯЩИК»

КОРРЕКТИРОВКА ВРЕДОНОСНЫХ ВХОДНЫХ ДАННЫХ НА ОСНОВАНИИ РЕЗУЛЬТАТОВ, ГЕНЕРИРУЕМЫХ МОДЕЛЬЮ

«ОГРАНИЧЕННЫЙ ЧЕРНЫЙ ЯЩИК»

ДОСТУП ТОЛЬКО К ВЫХОДНЫМ ДАННЫМ МОДЕЛИ

УРОВНИ ДОСТУПА ЗЛОУМЫШЛЕННИКА К ЦЕЛЕВОЙ МОДЕЛИ

Методы атак на СИИ

- → «ADVERSARIAL» (СОСТЯЗАТЕЛЬНЫЕ АТАКИ ВЛИЯНИЯ НА КЛАССИФИКАТОР МОДЕЛИ)
- → ПИКСЕЛЬНЫЕ АТАКИ
- → ПРОЕКТИВНЫЕ ИСКАЖЕНИЯ / АФФИННЫЕ ИСКАЖЕНИЯ
- → АТАКИ НА ОПТИЧЕСКИЙ ПОТОК
- → ADVERSARIAL PATCH («ВРЕДОНОСНЫЕ ЗАПЛАТКИ»)
- → АТАКИ С ИСПОЛЬЗОВАНИЕМ ТРИГГЕРОВ
- → АУДИОИСКАЖЕНИЯ И ИМИТАЦИЯ РЕЧИ ДРУГОГО ЧЕЛОВЕКА
- → ENERGY DRAW («ЭНЕРГОАТАКА»)

Технические методы защиты

Модификация модели

Состязательные атаки с искажениями — минимизируют ошибки классиф

Защитная дистилляция — состязательный метод обучения, при котором используется для обучения меньшей модели, которая демонст выходную поверхность.

Ансамблевое состязательное обучение — классификаторов обуча вместе и объединяются для повышения надежности.

Прочее: выработка устойчивости модели при обучении на станда с квантованием.

Преобразование входа

Маскировка градиента — препятствует использование злоумышленником полезного градиента.

Регуляризация входа — используется, чтобы избежать больших градиентов на входе сети, которые делают сети уязвимыми.

Сжатие признаков — объединение выборок, которые соответствуют множеству различных векторов признаков в исходном пространстве, в одну выборку, что уменьшает пространство поиска, доступное злоумышленнику.

Достоверные способы защиты

Стандартные методы, гарантирующие устойчивость целевой модели к известным атакам.

Прочие способы

Комбинирование нескольких стратегий защиты, защита от узконаправленных атак.

Организационные методы защиты

- НЕ ЗАБЫВАТЬ ПРО КЛАССИЧЕСКИЕ ПОДХОДЫ К ИБ И НЕ ПРЕНЕБРЕГАТЬ РАЗРАБОТКОЙ МОДЕЛИ УГРОЗ И МОДЕЛИ НАРУШИТЕЛЯ ДЛЯ СИСТЕМ, ИСПОЛЬЗУЮЩИХ ИИ
- СОЗДАТЬ ПЛОЩАДКУ ДЛЯ ОБМЕНА ЛУЧШИМИ ПРАКТИКАМИ В ОБЛАСТИ ОБЕСПЕЧЕНИЯ ДОВЕРИЯ И БЕЗОПАСНОСТИ СИСТЕМ, ИСПОЛЬЗУЮЩИХ ИИ
- СОЗДАТЬ ОТЕЧЕСТВЕННУЮ ПЛАТФОРМУ ДЛЯ РАЗРАБОТКИ И РАЗВИТИЯ ТЕХНОЛОГИЙ ИИ (ОС, ФРЕЙМВОРКИ, БИБЛИОТЕКИ, ОБУЧЕНИЕ И ТЕСТИРОВАНИЕ МОДЕЛЕЙ И ПР.)
- ПРОДОЛЖАТЬ РАЗВИТИЕ ПРОЦЕССОВ СТАНДАРТИЗАЦИИ ИИ И НАУЧНЫХ ИССЛЕДОВАНИЙ В ДАННОЙ ОБЛАСТИ

Предлагаемая последовательность
по расшивке проблематики доверенного и безопасного ИИ

2022

СОЗДАТЬ НА БАЗЕ ФГУП «ГРЧЦ» ЭКСПЕРИМЕНТАЛЬНЫЙ СТЕНД

ДЛЯ АПРОБАЦИИ ОСНОВНЫХ ТИПОВ И МЕТОДОВ АТАК НА СИИ, А ТАКЖЕ СПОСОБОВ ПРОТИВОДЕЙСТВИЯ ИМ

ПЕРВЫЙ ЭТАП

2023

РАЗРАБОТАТЬ ПРОЕКТ СТАНДАРТА ПРЕДПРИЯТИЯ,

ЧТОБЫ В ПИЛОТНОМ РЕЖИМЕ РЕШИТЬ ПЕРВООЧЕРЕДНЫЕ ВОПРОСЫ ОБЕСПЕЧЕНИЯ ДОВЕРИЯ И БЕЗОПАСНОСТИ ПРОФИЛЬНЫХ ДЛЯ ОТРАСЛИ СИИ

ВТОРОЙ ЭТАП

2024

УЧАСТВОВАТЬ В РАЗРАБОТКЕ НАЦИОНАЛЬНОГО СТАНДАРТА ДОВЕРИЯ И БЕЗОПАСНОСТИ СИИ

В РАМКАХ МЕХАНИЗМА ТК-164 «ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ»

СЛЕДУЮЩИЙ ЭТАП (ИЛИ ПАРАЛЛЕЛЬНО С СТП)

Спасибо за внимание!