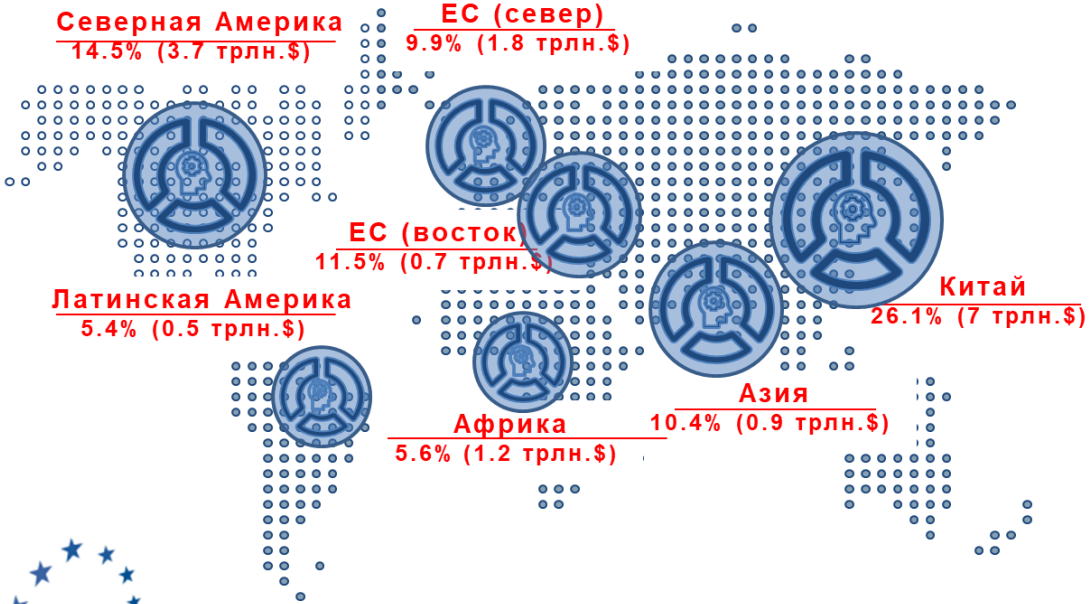
# Статический анализ фреймворков машинного обучения

Докторант кафедры
«Систем сбора и обработки информации»
Военно-космической академии имени А.Ф.Можайского
к.т.н. **Менисов Артем Бакытжанович**

**Захаров Олег Олегович**

г. Санкт-Петербург, 2023

# Эффект применения ТИИ (ВВП)




Все регионы получают экономический прирост от применения ТИИ. Прирост ВВП до 2030 г. ожидается до 58%

Самое большое развитие ТИИ – Северная Америка и Китай – 10.7 трлн.$

Ожидаемый прирост производительности в разных сферах – до 55%

enisa EUROPEAN UNION AGENCY FOR CYBERSECURITY

MITRE | ATLAS™

CUJOAI CYBERSECURITY FOR NETWORK OPERATORS

Alipay BOSCH CAICT 中国信息通信研究院 Deutsche Telekom Microsoft IBM NEC VISA ORACLE

*__Источник__: http://www.pwc.com/AI

# Классификация применения технологий искусственного интеллекта


**Источник: ИСИЭЗ НИУ ВШЭ.*

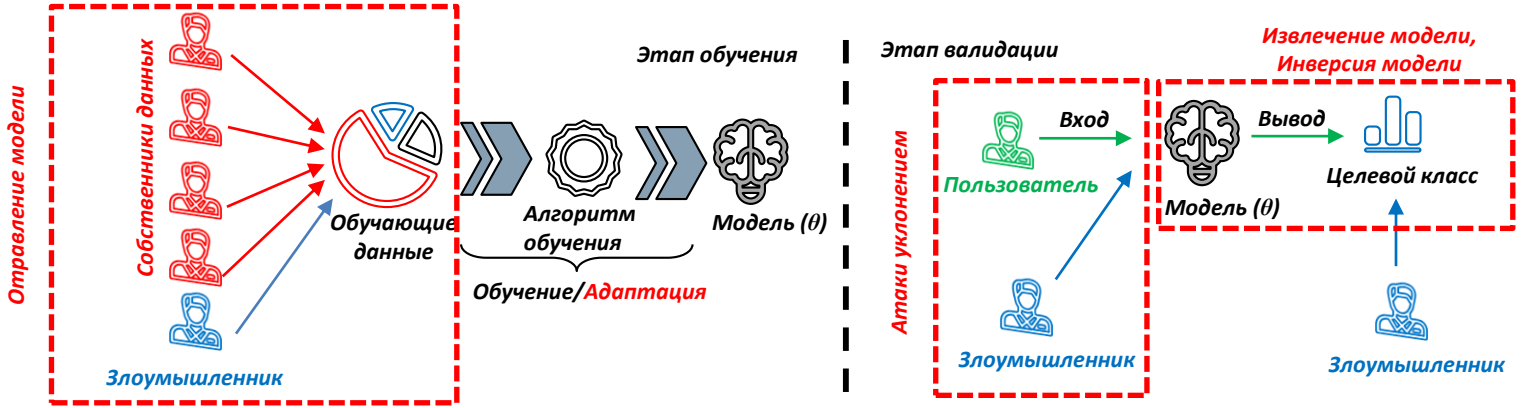
# Угрозы безопасности информации ФСТЭК России

| | |
|---|---|
| УБИ. 218 | Угроза раскрытия информации о модели машинного обучения |
| УБИ. 219 | Угроза хищения обучающих данных |
| УБИ. 220 | Угроза нарушения функционирования («обхода») средств, реализующих технологии искусственного интеллекта |
| УБИ. 221 | Угроза модификации модели машинного обучения путем искажения («отравления») обучающих данных |
| УБИ. 222 | Угроза подмены модели машинного обучения |

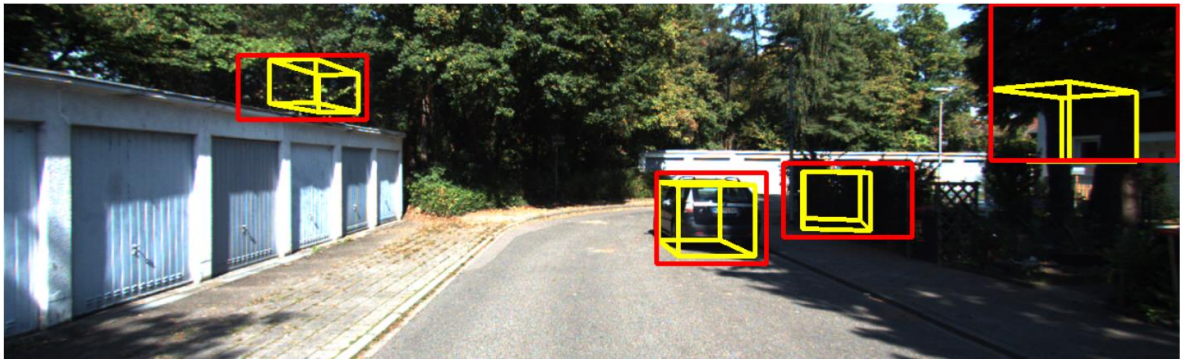
* ***Источник***: *https://bdu.fstec.ru/threat (последние обновления – декабрь 2020 г.)*

# Проблемы обеспечения защищенности ТИИ

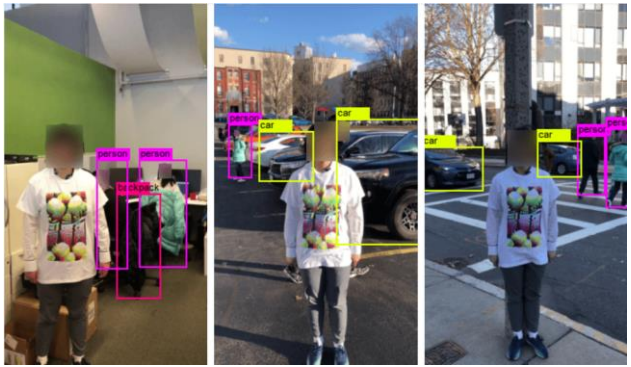
*Возможные объекты воздействия*




## Атаки на датчики беспилотных автомобилей (камеры и лидара)



## Атаки на системы распознавания



# Проблемы обеспечения защищенности ТИИ

*Актуальные последствия атак на ТИИ*

## Целостность

**Атака уклонения**
Злоумышленник изменяет запрос, чтобы получить соответствующий ответ.

**Backdoor**
Модель с бэкдором вызовет целенаправленную неправильную классификацию или ухудшит точность модели для входных данных, которые содержат триггером.

**Отравление модели**
Цель злоумышленника — испортить модель машины.

**Атака на конвейеры МО**
Атака на известные модели или наборы данных.

**Перепрограммирование ИНС**
С помощью специально созданного запроса от злоумышленника СИИ могут быть перепрограммированы на другую задачу.

## Конфиденциальность

**Использование программных зависимостей СИИ**
Злоумышленник использует традиционные уязвимости программного обеспечения.

**Инверсия модели**
Параметры, используемые в моделях МО, могут быть восстановлены.

**Кража модели**
Злоумышленники воссоздают базовую модель с близким функционалом.

**Восстановление обучающих данных**
Злоумышленник может восстановить часть информации.

## Доступность

**Перепрограммирование ИНС**
С помощью специально созданного запроса от злоумышленника СИИ могут быть перепрограммированы на задачу, которая отличается от первоначального замысла.

# Результаты статического анализа

| | CNTK 2.7 | Dlib 19.24 | Keras 2.10.0 | MXNet 1.9.1 | PyTorch 1.12.1 | Sklearn 1.1.2 | Tensorflow 2.10 |
|---|---|---|---|---|---|---|---|
| **Всего** | 57 | 185 | 54 | 1580 | 2859 | 867 | 191 |
| **Дефекты высокого уровня опасности (high)** | 12 | 52 | 0 | 465 | 1468 | 93 | 43 |
| **Дефекты среднего уровня опасности (medium)** | 41 | 112 | 16 | 918 | 1150 | 323 | 140 |
| **Дефекты низкого уровня опасности (low)** | 4 | 21 | 38 | 197 | 241 | 451 | 8 |
| **Активность (за последние 5 лет)** | 0 | 0 | 3 | 2 | 0 | 1 (1) | 323 |
| **Наиболее уязвимые компоненты** | OpenCV MKLML | libpng, zlib | | oneDNN (11 дефектов), GoogleTest (1 дефект), CUDA SDK (1 дефект) | oneDNN, GoogleTest, Tensorpipe, NCCL, Google Protobuf, onnx, ATen, SLEEF, ideep, Google Benchmark, FBGEMM | | NVIDIA NCCL |
| **The CWE Top 25 2022** | 1 | 5 | 3 | 5 | 5 | 2 | 3 |

* Применяемый статический анализатор - Coverity (Synopsys)

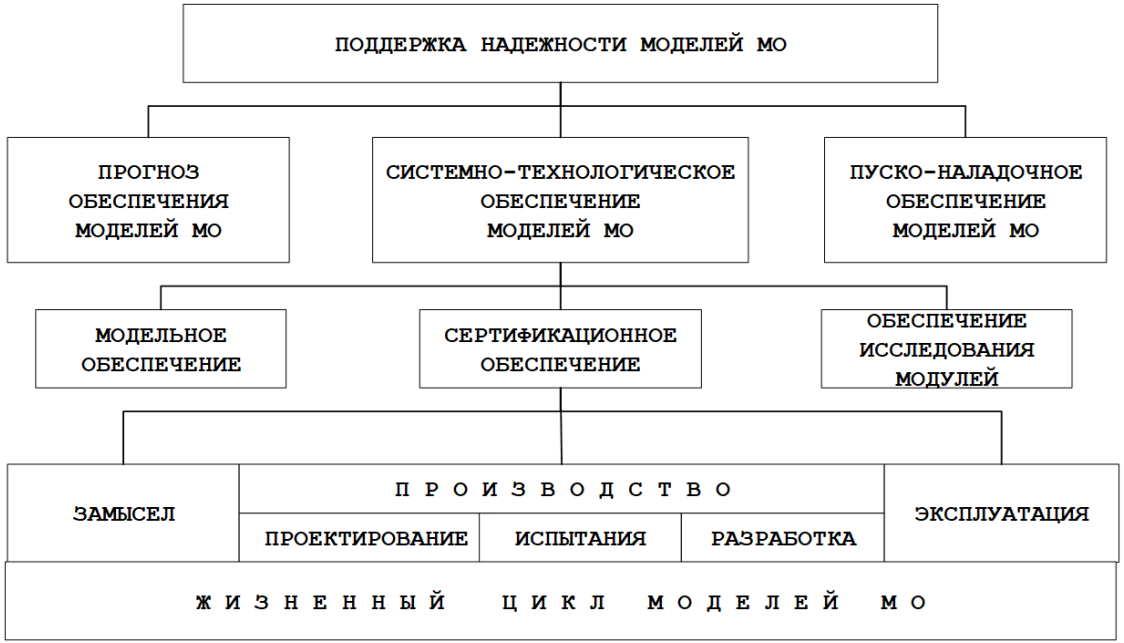
# Проблемы развития ТИИ

# Выводы

1. Несмотря на широкое распространение ТИИ и наличие большого числа открытых фреймворков машинного обучения, обеспечивающих должный уровень эффективности и производительности, создание систем доверенного ИИ является долгосрочным вызовом.

2. На текущий момент все еще не существует технической, методологической и организационной инфраструктуры для обучения разработки высоконадежных, доверенных и одновременно эффективных систем, использующих ТИИ.

# Предложения




**Необходимость исследования и разработки**

- инструментария тестирования надежности моделей машинного обучения
- инструментария интерпретируемости моделей машинного обучения
- детально регламентированного технологического процесса поддержки корректности функционирования моделей машинного обучения
- сопровождения восстановления работоспособности систем искусственного интеллекта после проведения компьютерных атак

# СПАСИБО ЗА ВНИМАНИЕ!