Российская неделя высоких технологий РНВТ

Эшелон комплексная безопасность

Актуальные вопросы оценки соответствия интеллектуальных СЗИ (постановка задачи)

РНВТ-2022, Тематическая секция «Искусственный интеллект – новая основа безопасности критической инфраструктуры государства», Москва, ЦВК «Экспоцентр», 27 апреля 2022 г.

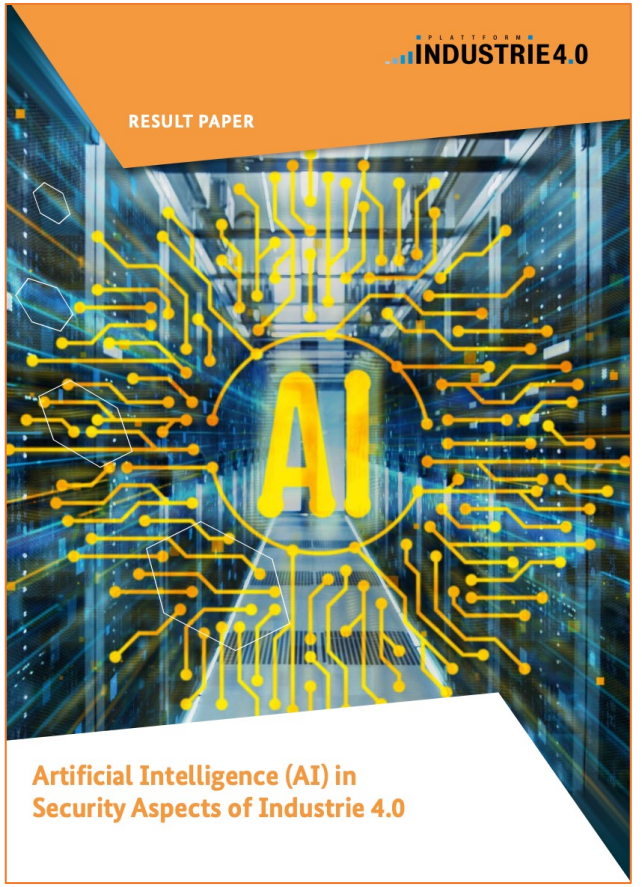
Алексей Марков,
д.т.н., CISSP, CEH

План

1. Актуальность
2. Угрозы и контрмеры ИБ в области ИИ
3. Оценка соответствия и системы сертификации
4. Выводы и предложения

Область определения

- Позиция ISO/IEC: AI == ML
- Интеллектуальные СЗИ – это СЗИ, предназначенные для решения различных задач обеспечения ИБ, создание которых в обязательном порядке связано с использованием специальным образом подготовленных НД, описывающих представительную совокупность примеров решения этих задач (по ТК-164)



Актуальность

Gartner Top 10 Strategic Technology Trends

AI Security

Конвергенция трендов: кибербезопасность …

Статистика и прогнозы

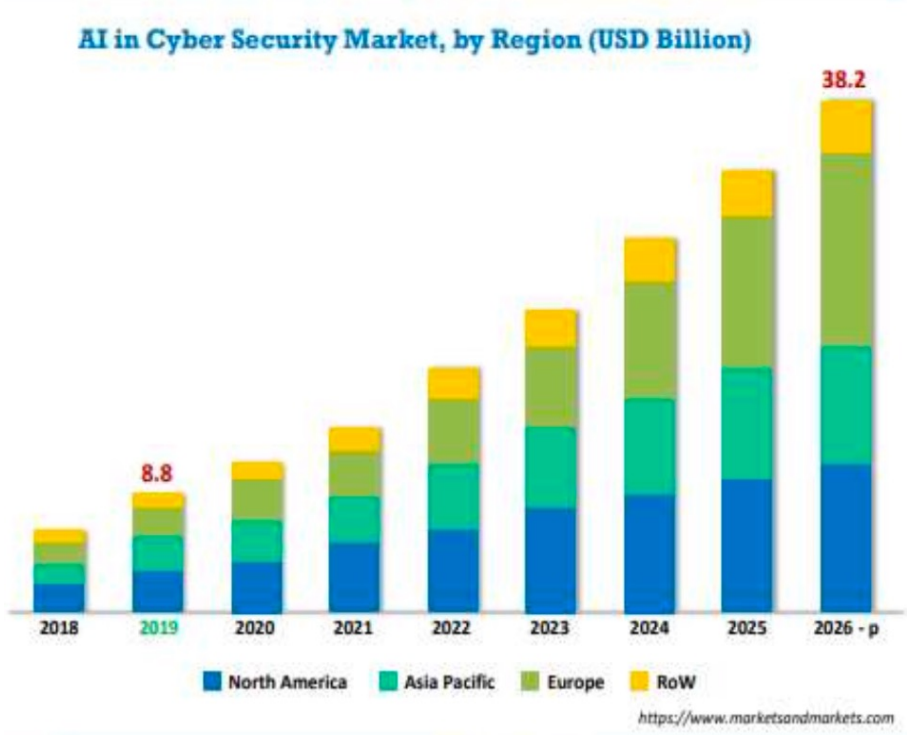
- MSM на 2026 г. – Рынок ML & Cybersecurity - 2.83 трлн. ₽ >> в 135 раз..
- Gartner 2022 г. - 30% кибератак с использованием ИИ будут составлять состязательные атаки
- Splunk 2022 г.: 63% организаций используют ML для аналитики по ИБ
- Cylance 2022: 62% хакеров планируют использовать ИИ
- Экспоненциальный рост исследований в вузах (лидеры 4 Пром. рев.)
- …

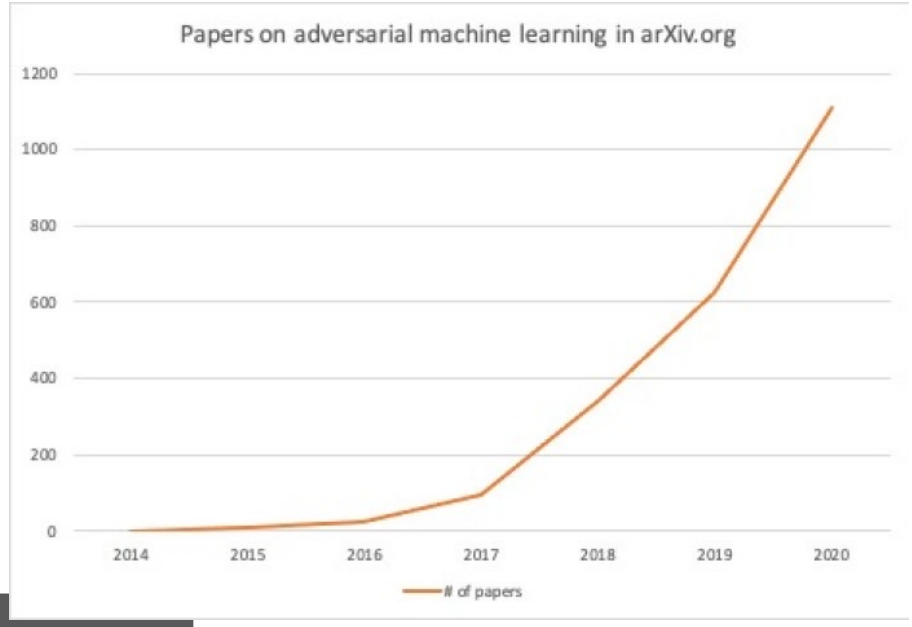
Инциденты ИБ

- Переученные боты, ошибки идентификации

Регуляторы

- Проработка тематических стандартов MITRE, NIST, ISO/IEC…
- Система сертификации ИИ по эгидой TÜV

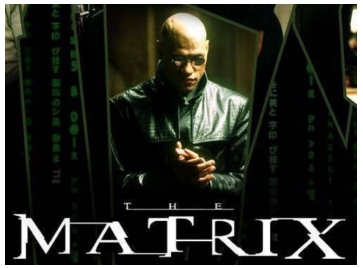




Угрозы и контрмеры ИБ в области ИИ

- Новые классы атак (состязательные атаки)
- Эволюция угроз (интеллектуальные атаки)
- Эволюция СЗИ с элементами ИИ (СЗИ NG)
  - Эволюция САЗ и технологий разработки БПО с элементами ИИ

1. Атаки на системы ИИ



- Атаки на данные (искажение разметки, «черный лебедь»)
- Атаки на алгоритмы (белый ящик, черный ящик, ндв - заимствованные компоненты, бюллетени и БД и т.д.)
- Атака на наличие конфиденциальных данных
- Атаки на среду и др.

MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems

Reconnaissance	Resource Development	Initial Access	ML Model Access	Execution	Persistence	Defense Evasion	Discovery	Collection	ML Attack Staging
5 techniques	7 techniques	2 techniques	4 techniques	1 technique	2 techniques	1 technique	3 techniques	2 techniques	4 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution	Poison Training Data	Evade ML Model	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities	Valid Accounts	ML-Enabled Product or Service		Backdoor ML Model		Discover ML Model Family	Data from Information Repositories	Backdoor ML Model
Search Victim-Owned Websites	Develop Adversarial ML Attack Capabilities		Physical Environment Access				Discover ML Artifacts		Verify Attack
Search Application Repositories	Acquire Infrastructure		Full ML Model Access						Craft Adversarial Data
Active Scanning	Publish Poisoned Datasets								
	Poison Training Data								
	Establish Accounts								

Full ML Model Access

Summary

Adversaries may gain full "white-box" access to a machine learning model. This means the adversary has complete knowledge of the model architecture, its parameters, and class ontology. They may exfiltrate the model to Craft Adversarial Data and Verify Attack in an offline where it is hard to detect their behavior.

ID: AML.T0044

Tactics: ML Model Access

Number of case studies: 1

Case study examples

Backdoor Attack on Deep Learning Models in Mobile Apps

This provided the researches with full access to the ML model, albeit in compiled, binary form.

2. Атаки с использованием ИИ

- Сбор (фильтрация) информации о жертве
- Имперсонация (SET, фишинг и т.д.)
- Несанкционированный доступ (обход CAPTCHA и подбор паролей)
- Проведение атак (VA, Malware, DDoS, crowdturfing/фейки)
- Автоматизация APT-атак (HiveNet – управление ботнетами)

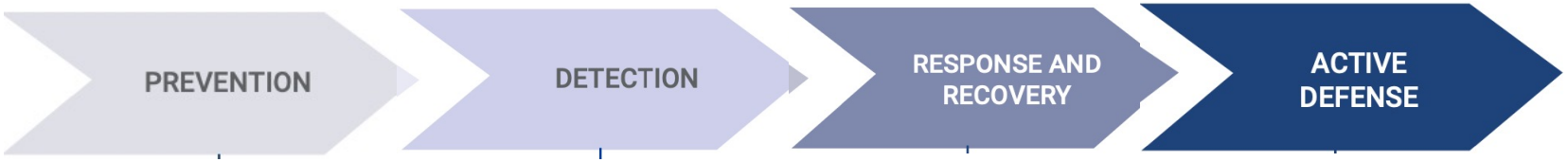
Metric \ Model	LSTM	Markov Chain
Training Speed	Days	Seconds
Accuracy	High	Medium
Availability	Public	Public
Size	Large	Small

DARPA: CTF-соревнования Cyber Grand Challenge 2016
Участник – система ИИ **Mayhem**

3. СЗИ NG. Пример классификации по CBI

- Идентификация, аутентификация, борьба с мошенничеством
- Риск-менеджмент
- Ловушки, Honeypot
- Автоматизация
- Проактивная защита, NDR, SIEM/IPS
- Поведенческий анализ, UEBA
- Mobile/IoT-безопасность
- Безопасность приложений

3. СЗИ NG. Пример классификации по CSET



Предотвращение | Выявление | Реагирование | Проактивная защита

CYBERSECURITY TASK	UNDERLYING TECHNOLOGY OF NEW AI APPLICATIONS
Fuzzing	Deep Learning
Pentesting	Reinforcement Learning
Bug Triage and Classification	NLP, Traditional ML Methods
Vulnerability Severity Assessment	NLP, Traditional ML Methods

CYBERSECURITY TASK	UNDERLYING TECHNOLOGY OF NEW AI APPLICATIONS
Accurate Detection	Deep Learning
Alert Prioritization	Deep Learning
Adversarial Hardening of Detection Systems	GANs

CYBERSECURITY TASK	UNDERLYING TECHNOLOGY OF NEW AI APPLICATIONS
Adversary Engagement	Reinforcement Learning
Moving Target Defense	Reinforcement Learning

CYBERSECURITY TASK	UNDERLYING TECHNOLOGY OF NEW AI APPLICATIONS
Deceptive Document Generation	NLP, GANs
Dynamic Honeypotting	Reinforcement Learning
Automated Phishing Response	NLP, Reinforcement Learning
Dark Web Threat Intelligence	NLP
Attack Clustering for Attribution	NLP, Traditional ML Methods
Code De-Anonymization	NLP

Какие могут быть варианты оценки соответствия интеллектуальных СЗИ?

Оценка соответствия и системы сертификации СЗИ по требованиям БИ

- Сертификация СЗИ
- Аттестация ОИ

- Обязательная сертификация
  - Минобороны России
  - ФСБ России
  - ФСТЭК России

- Добровольная сертификация
  - Банки
  - другие

- Сертификация
  - СЗИ
  - СМИБ

Примеры вариантов требований

ФСТЭК России

- По классу защищенности от НСД (РД Гостехкомиссии)
- По классу защищенности от НСД (ПЗ/ЗБ ФСТЭК России)
- По требованиям доверия 76 Приказ ФСТЭК (уязвимости, БПО)

Банк России

- **По ОУД (ОУД4/4+) по ГОСТ Р ИСО/МЭК 15408-3-2013**
- По Профилю защиты (ЦБ РФ).
- По требованиям ГОСТ Р 56939-2016

Минобороны России

- На соответствие РДВ
- По контролю НДВ
- По РД АС, РД СВТ
- По ПЗ
- КИКТ и др.

ФСБ России

СТБИ

- ГОСТ Р 56939-2016
- **ГОСТ Р ИСО/МЭК 15408-3-2012/2013**
- **ГОСТ Р ИСО/МЭК 27001-2006**
  ГОСТ Р 57580.1-2017

Профильные технические комитеты

- ТК-164 «Искусственный интеллект»
- ТК-362 «Защита информации»
- ТК-26 «Криптографическая защита информации»
- ТК-22 «Информационные технологии»
- ТК-122 «Стандарты финансовых операций»
- другие

- ISO/IEC JTC 1/SC 42 Artificial intelligence
- ISO/IEC JTC 1/SC 27 Information security, cybersecurity and privacy protection
- другие

Примеры стандартов в области ИИ, релевантных ИБ

- **По линии ISO/IEC JTC 42 – 11 стандартов, в разработке 26**
- ISO/IEC TR 24028:2020 IT— Artificial intelligence — Overview of trustworthiness in artificial intelligence
- ISO/IEC AWI TR 5469 Artificial intelligence — Functional safety and AI systems
- ISO/IEC DIS 23894 IT— Artificial intelligence — Risk management
- ISO/IEC AWI TS 29119-11 IT — Artificial intelligence — Testing for AI systems — Part 11:
- …
- **По линии ISO/IEC JTC 27**
- ISO/IEC AWI 27562 Security and privacy in artificial intelligence use cases
- ISO/IEC DTR 27563 Impact of security and privacy in artificial intelligence
- ….

- **gost.ru – 11 стандартов**

РЕЗУЛЬТАТЫ ПОИСКА

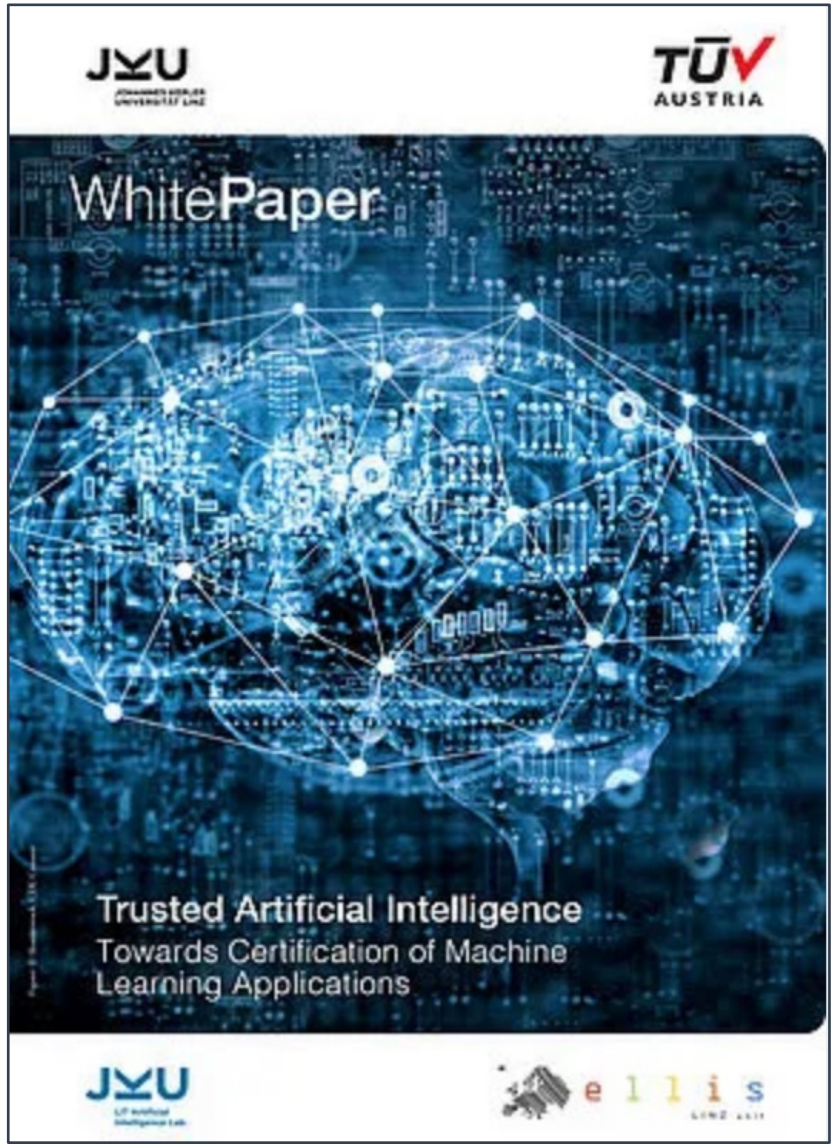
Найдено: 11 документов

Обозначение ГОСТ	Наименование
ГОСТ Р 59276-2020	Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения
ГОСТ Р 59277-2020	Системы искусственного интеллекта. Классификация систем искусственного интеллекта
ГОСТ Р 59900-2021	Системы искусственного интеллекта. Типовые требования к контрольным выборкам исходных данных для испытания систем искусственного интеллекта в образовании
ГОСТ Р 59920-2021	Системы искусственного интеллекта. Системы искусственного интеллекта в сельском хозяйстве. Требования к обеспечению характеристик эксплуатационной безопасности систем автоматизированного управления движением сельскохозяйственной техники
ГОСТ Р 59921.2-2021	Системы искусственного интеллекта в клинической медицине. Часть 2. Программа и методика технических испытаний
ГОСТ Р 59921.3-2021	Системы искусственного интеллекта в клинической медицине. Часть 3. Управление изменениями в системах искусственного интеллекта с непрерывным обучением
ГОСТ Р 59921.4-2021	Системы искусственного интеллекта в клинической медицине. Часть 4. Оценка и контроль эксплуатационных параметров
ГОСТ Р 59921.5-2022	Системы искусственного интеллекта в клинической медицине. Часть 5. Требования к структуре и порядку применения набора данных для обучения и тестирования алгоритмов

Пример системы сертификации систем машинного обучения TÜV

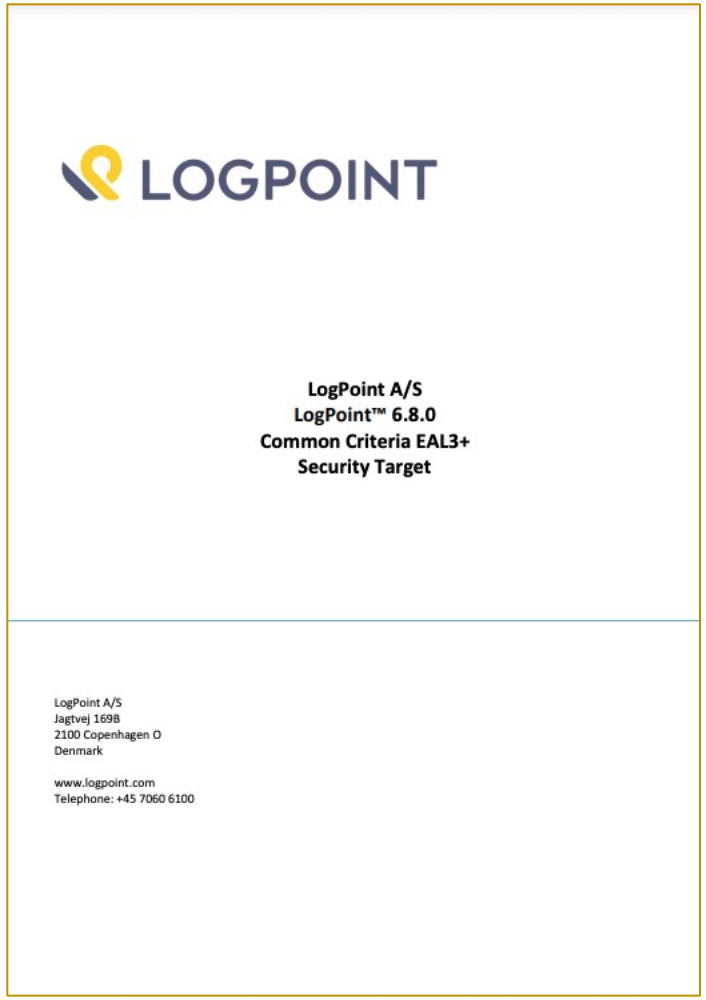
Требования:

- фактическое функционирование (РДВ)
- надежность обученной машины
- безопасность ПО
- соответствие требованиям сферы применения ПО (НДВ)
- обеспечение защиты и конфиденциальности данных
- этические аспекты взаимодействия машины и человека



Пример сертификации интеллектуального СЗИ по линии Common Criteria

- UEBA (machine learning) LogPoint SIEM


LOGPOINT

LogPoint A/S
LogPoint™ 6.8.0
Common Criteria EAL3+
Security Target

LogPoint A/S
Jagtvej 169B
2100 Copenhagen O
Denmark

www.logpoint.com
Telephone: +45 7060 6100

Выводы и предложения

- Отсутствие нормативной и методический базы в области применения технологии ИИ в области ИБ создает угрозы различного уровня вплоть до глобального
- Вопрос создания подсистемы, регулирующей ИБ интеллектуальных систем защиты, вопрос времени по объективным причинам
- Чем быстрее будут исследованы вопросы применения ИИ в области ИБ, тем быстрее страна получит стратегическое преимущество в глобальном ИПБ
- В ряде систем сертификации средств защиты информации это можно сделать различными способами
- Вопросы стандартизации могут лежать на стыке работы различных ТК
- Очевидна важность гармонизации международной нормативной базы в области ИИ и ИБ, а также разработки отечественных НПА и методических документов
- Чрезвычайно актуальным является исследование эффективности применения технологий ИИ в области оценки соответствия (аудита, анализа защищенности, сертификационных и аттестационных испытаний и др).

СПАСИБО ЗА ВНИМАНИЕ!

Марков Алексей Сергеевич
a.markov@npo-echelon.ru