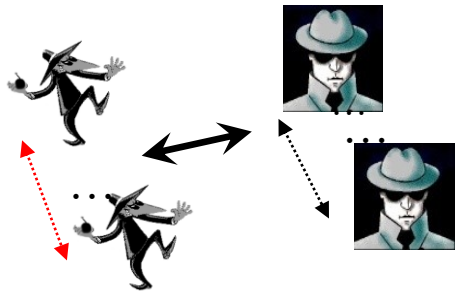
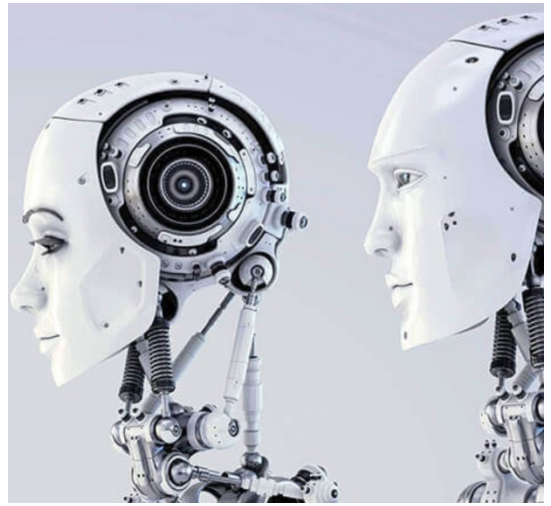
Ключевые области внимания на стыке искусственного интеллекта (ИИ) и кибербезопасности



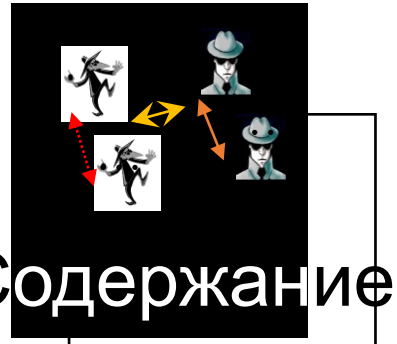
И.В. Котенко

Санкт-Петербургский Федеральный
исследовательский центр
Российской академии наук

СПб ФИЦ РАН



РусКрипто’2023 - 23 марта 2023 г.



Содержание

- Введение
- Тренды в ИИ
- Повышение кибербезопасности на основе ИИ
- Использование ИИ для кибератак кибербезопасности
- Уязвимости систем ИИ к атакам
- Использование ИИ во вредоносных информационных операциях
- Заключение

Ключевые области внимания на стыке ИИ и кибербезопасности

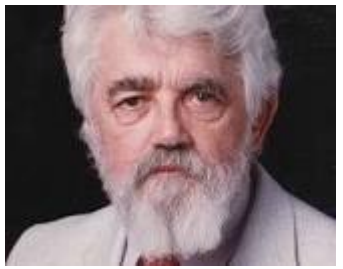
- ❑ **Повышение кибербезопасности с помощью ИИ**
- ❑ **Использование ИИ для кибератак**
- ❑ **Уязвимости систем ИИ к атакам**
- ❑ **Использование ИИ во вредоносных информационных операциях (фейки с использованием ИИ)**

[Applications for artificial intelligence in Department of Defense cyber missions. https://blogs.microsoft.com/on-the-issues/2022/05/03/artificial-intelligence-department-of-defense-cyber-missions/]
[Д.Е. Намиот, Е.А. Ильюшин, И.В. Чижов. Искусственный интеллект и кибербезопасность. 2022]

Содержание

- Введение
- Тренды в ИИ
- Повышение кибербезопасности на основе ИИ
- Использование ИИ для кибератак кибербезопасности
- Уязвимости систем ИИ к атакам
- Использование ИИ во вредоносных информационных операциях
- Заключение

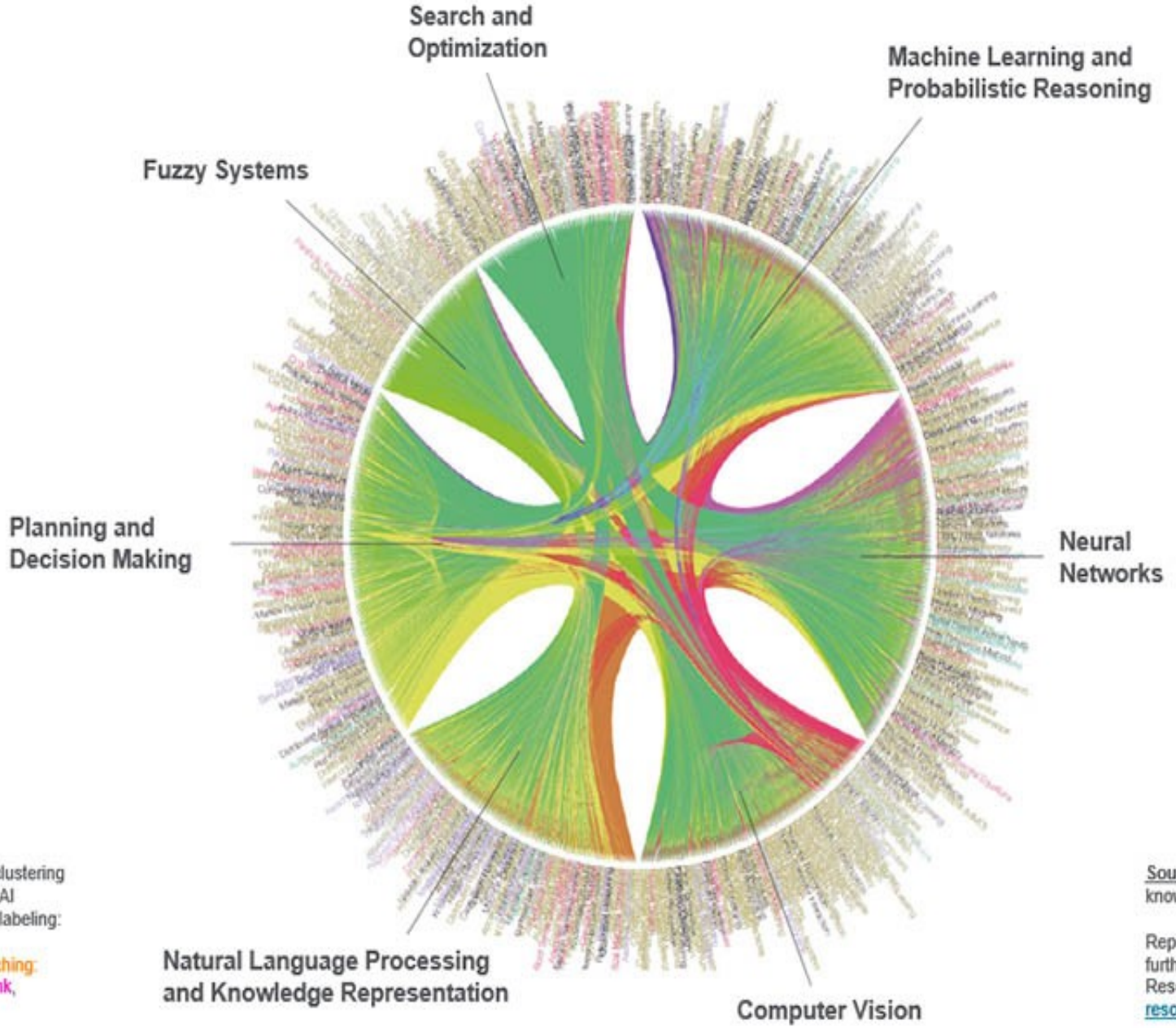
Карта искусственного интеллекта



John McCarthy

[V.Osaulenko, 2020]

Кластеры исследований ИИ



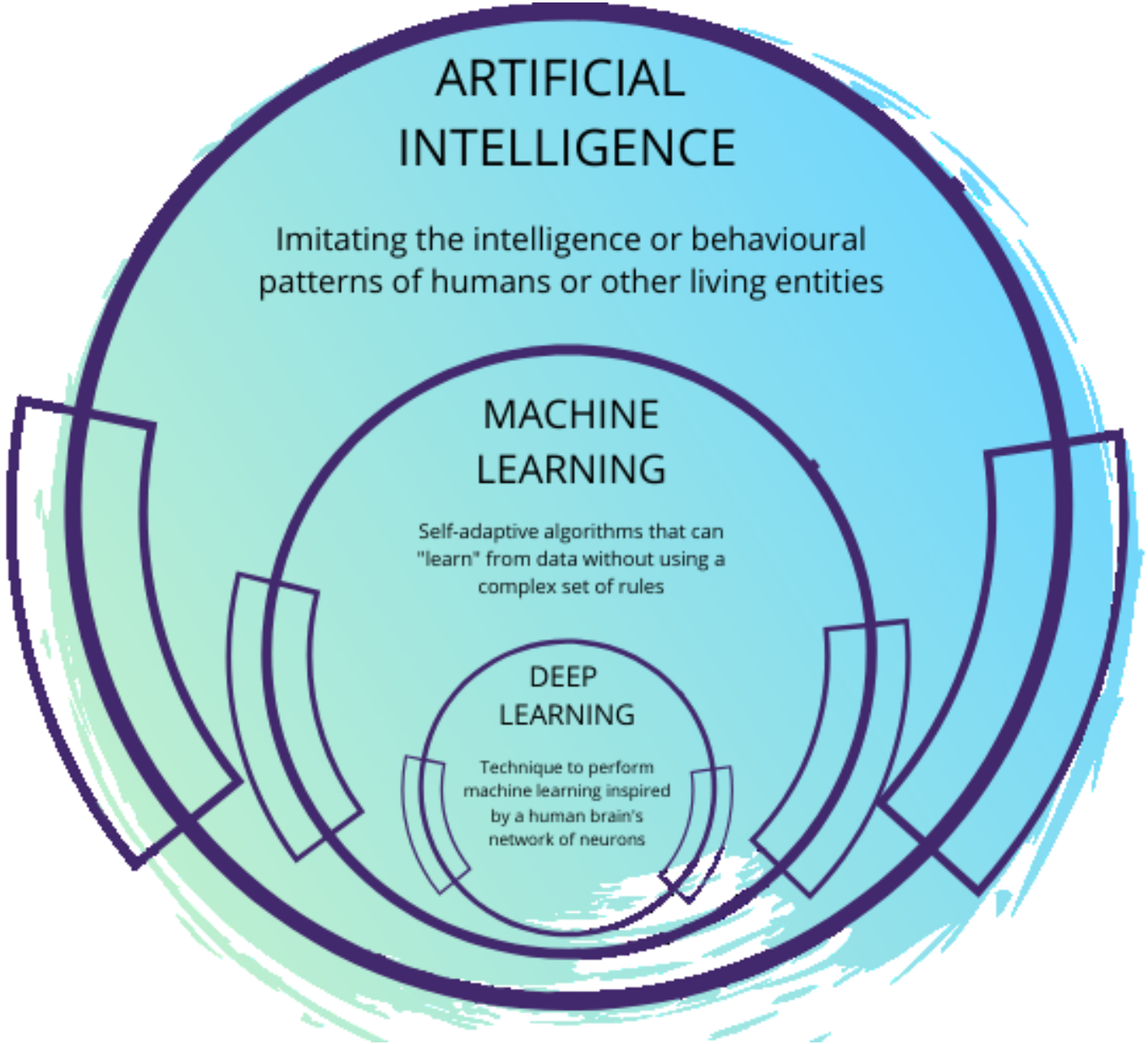
Using keyword co-occurrences with unsupervised clustering on article-level, based on keywords extracted from AI documentation of different actors and using expert labeling:

The color of the keyword represents its origin: Teaching: orange, Industry: green, Research: blue, Media: pink, multiple: black

Source: Elsevier, 2018, Artificial Intelligence: how knowledge is created, transferred, and used

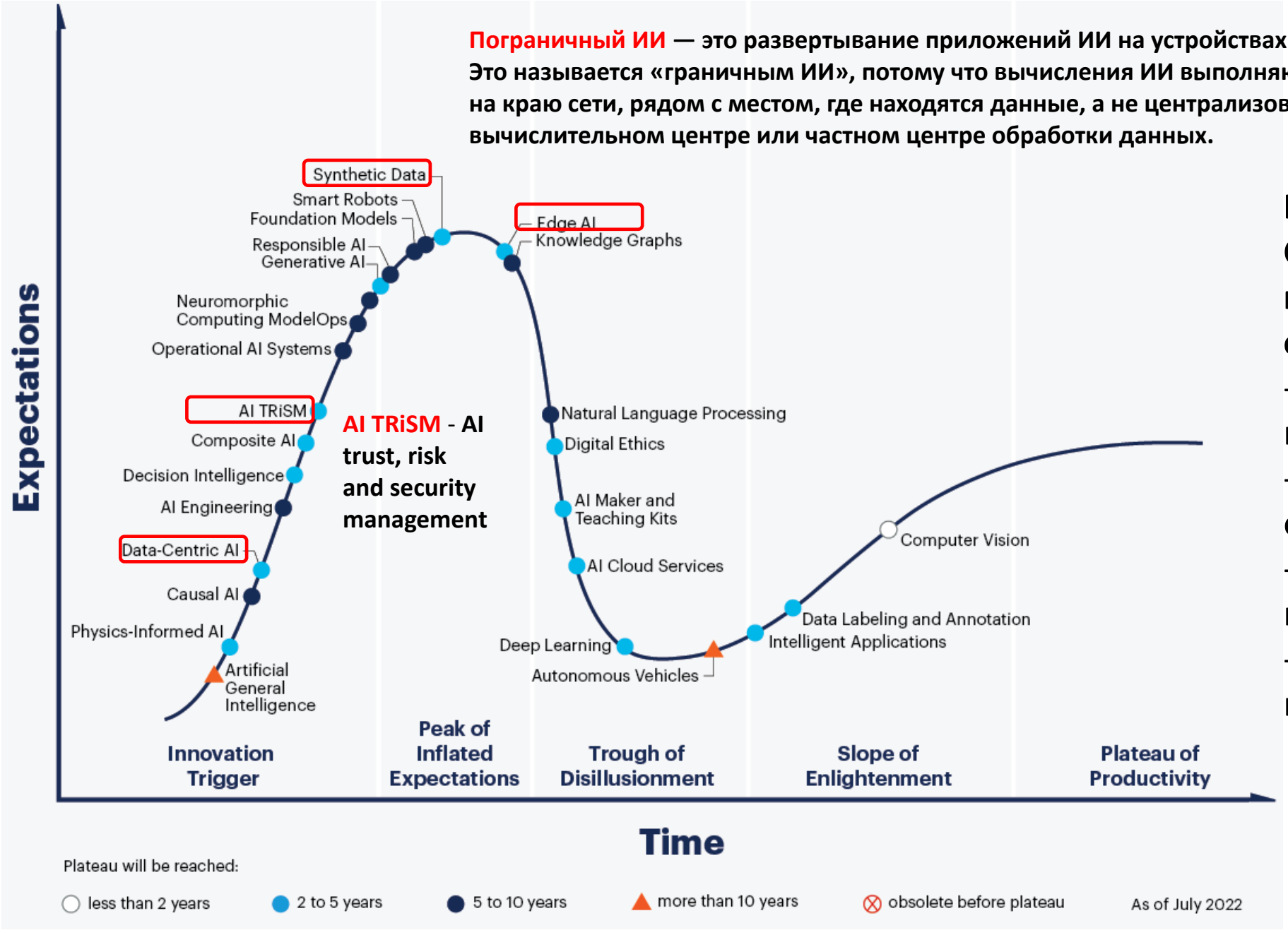
Report will be launched: Dec. 11th, 2018. Download and further information from Elsevier's Artificial Intelligence Resource Center: https://www.elsevier.com/connect/ai-resource-center

ИИ, машинное обучение и глубокое обучение



Цикл зрелости (Hype Cycle) для ИИ, 2022, Gartner

Пограничный ИИ — это развертывание приложений ИИ на устройствах по всему физическому миру. Это называется «граничным ИИ», потому что вычисления ИИ выполняются рядом с пользователем на краю сети, рядом с местом, где находятся данные, а не централизованно в облачном вычислительном центре или частном центре обработки данных.



Инновации ИИ в Hype Cycle отражают приоритеты в четырех основных категориях:

- ИИ, ориентированный на данные ИИ
- Модельно-ориентированный ИИ
- ИИ, ориентированный на приложения
- ИИ, ориентированный на человека

Содержание

- Введение
- Тренды в ИИ
- Повышение кибербезопасности на основе ИИ
- Использование ИИ для кибератак кибербезопасности
- Уязвимости систем ИИ к атакам
- Использование ИИ во вредоносных информационных операциях
- Заключение

Сферы применения ИИ в кибербезопасности

- Обнаружение и реагирование на кибератаки
- Обнаружение мошенничества в бизнес-процессах
- Управление событиями безопасности
- Защита конечных точек
  Защита приложений, управление уязвимостями
- Контроль доступа и аутентификация
- Анализ поведения пользователей и устройств
- Обнаружение вредоносных программ
- Анти-Фишинг

Мультипредставления и мультимодальная аналитика

- Ключевым недостатком существующих исследований и разработок в ИИ для кибербезопасности является **изолированное использование отдельных наборов данных**. Это часто является результатом отсутствия доступа к множеству наборов данных (в академических кругах) и/или полного понимания взаимосвязей между несколькими наборами данных.
- Чтобы решить эту проблему, будущие исследования ИИ для кибербезопасности следует направить на более **целостное использование характеристик нескольких источников данных**.
- **Перспективные подходы включают сопоставление сущностей глубокого обучения, сопоставление коротких текстов (например, на основе глубоко структурированных семантических моделей), подходы с несколькими представлениями (например, с несколькими источниками) и стратегии многозадачного обучения**.
- Использование знаний из нескольких наборов данных с помощью **трансферного обучения** (transfer learning) и/или **федеративного обучения** также следует использовать для повышения производительности задач.
- Каждая модель может быть **расширена для учета ключевых аспектов** предметной области (например, своевременности, интерпретируемости и т. д.) и повышения способности модели к обучению.
- Успешное **объединение нескольких источников данных** может привести к новым производным атрибутам, улучшенным показателям управления рисками и, в конечном счете, к целостному представлению о состоянии кибербезопасности организации.

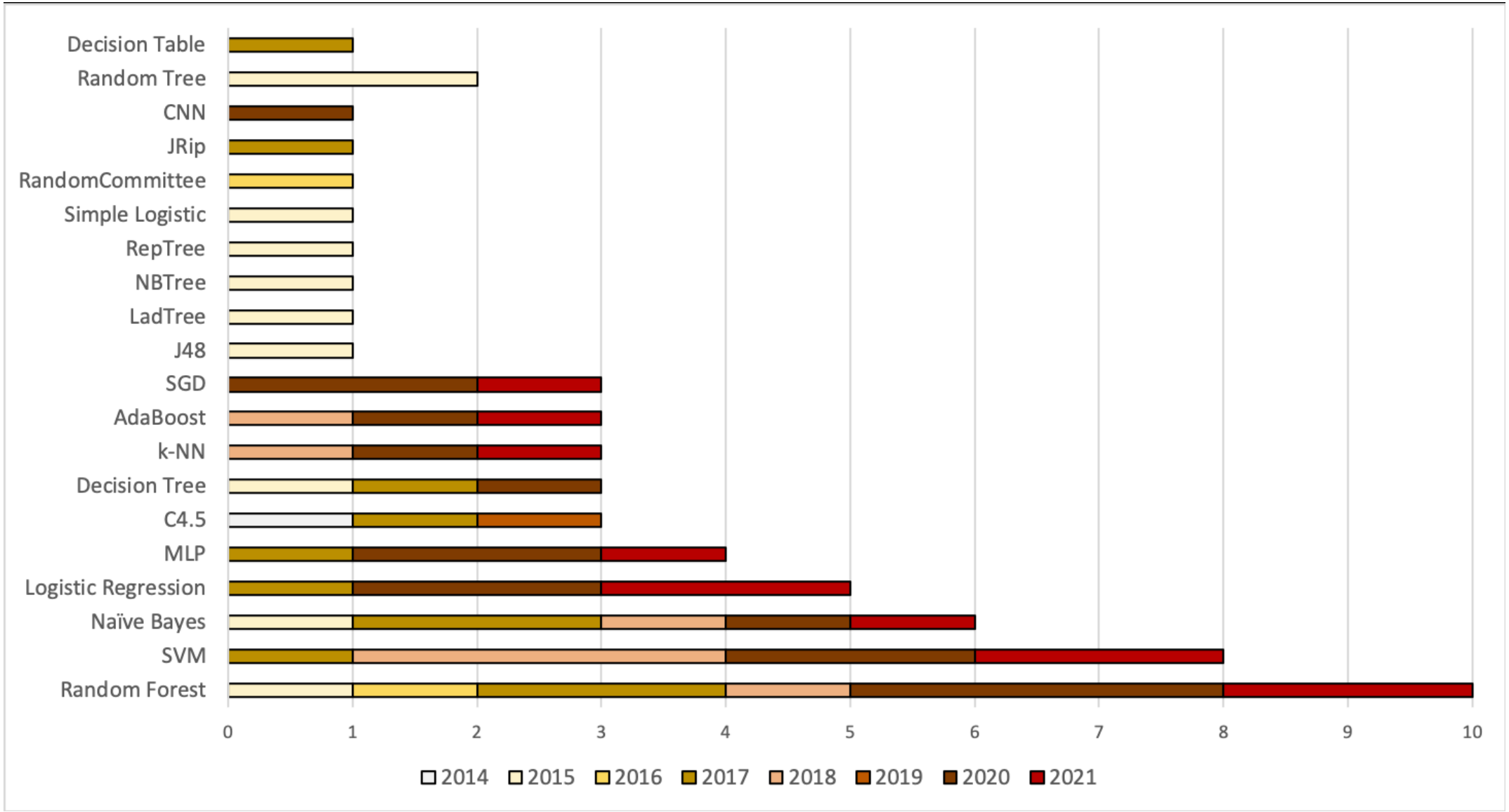
Объяснимые и интерпретируемые подходы к ИИ

- Многие современные аналитические процедуры с поддержкой ИИ, особенно основанные на глубоком обучении, часто являются **черными ящиками**, которые не объясняют, как модель достигла своего результата.
- **Отсутствие объяснимости модели** может привести к серьезным проблемам с доверием к ИИ, настройкой модели, производительностью модели и другим проблемам.
- Проблемы с существующими подходами к аналитике с поддержкой ИИ побудили многих исследователей разработать **объяснимые методы ИИ (XAI)**.
- В целом существуют две основные **категории XAI**: внутренние (intrinsic) и апостериорные (post-hoc).
- **Внутренние подходы** включают в модель такие методы, как правила принятия решений, механизмы внимания, пути рассуждений, маски и/или графы и другие. Внутренние подходы XAI работают во время обучения и выполнения модели.
- **Апостериорные методы**, такие как визуализация, контр-фактический анализ (counter-factual analysis), суррогатные модели, важность концепций, LIME и SHAP, обычно представляют собой независимые от модели подходы, которые объясняют различные компоненты модели после ее схождения.

Расширенный интеллект — интерфейсы человек-ИИ

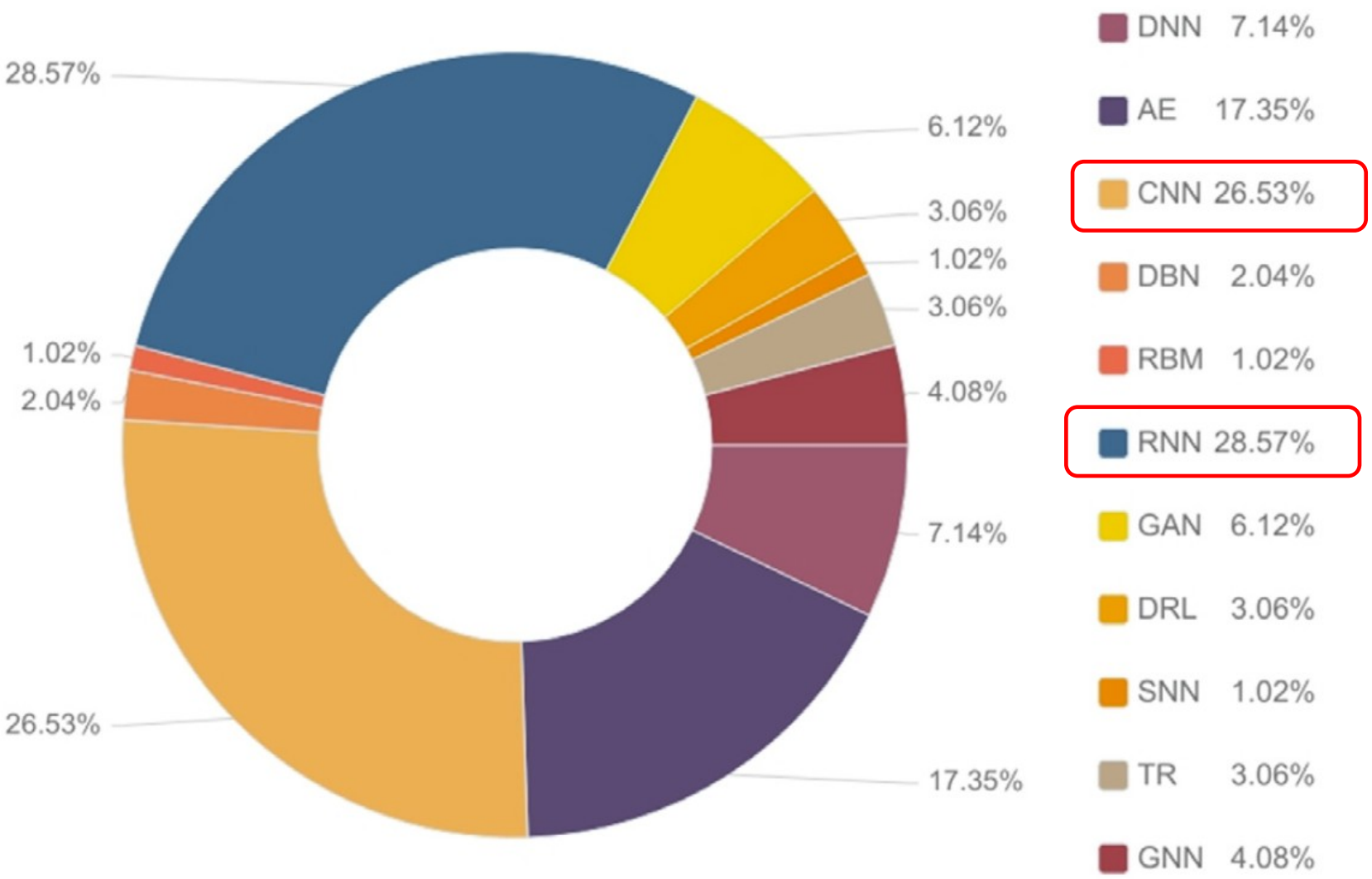
- Многие специалисты по кибербезопасности утверждают, что алгоритмы и системы на основе ИИ не должны сами принимать решения в области кибербезопасности. Скорее, подходы на основе ИИ, должны быть **тесно связаны с действиями человека** (например, аналитик безопасности является активным участником аналитического процесса), чтобы способствовать принятию эффективных решений.
- Эти подходы, также называемые **расширенным интеллектом или интерфейсами человек-ИИ**, могут привести к значительному повышению производительности по сравнению с использованием алгоритма или человека по отдельности.
- Существуют **три общих подхода к интерфейсам человека и ИИ**: (1) замещение (ИИ заменяет людей), (2) увеличение (ИИ и человек синергетически дополняют друг друга) и (3) сборка (люди и ИИ динамически собираются, чтобы сотрудничать и функционировать как единое интегрированное целое).
- **Способы взаимодействия человека и ИИ** для решения критических и фундаментальных задач кибербезопасности недостаточно исследованы. Такие исследования должны использовать **междисциплинарный подход**, особенно с упором на перспективы когнитивной науки, психологии, взаимодействия человека с компьютером и других областей.

Использование базовых моделей машинного обучения



Количество работ с использованием базовых моделей классификации для поиска вредоносных приложений на Android в год [Kouliaridis&Kambourakis, 2021]

Использование моделей DL в статьях [M. Macas et al., 2022]



DNN - Deep Neural Networks
AE - Autoencoder
CNN - Convolutional Neural Network
DBN - Deep Belief Network
RBM - Restricted Boltzmann Machine
RNN - Recurrent Neural Network
GAN - Generative Adversarial Network
DRL - Deep Reinforcement Learning
SNN - Siamese neural network
TR - Transformer
GNN - Graph Neural Network

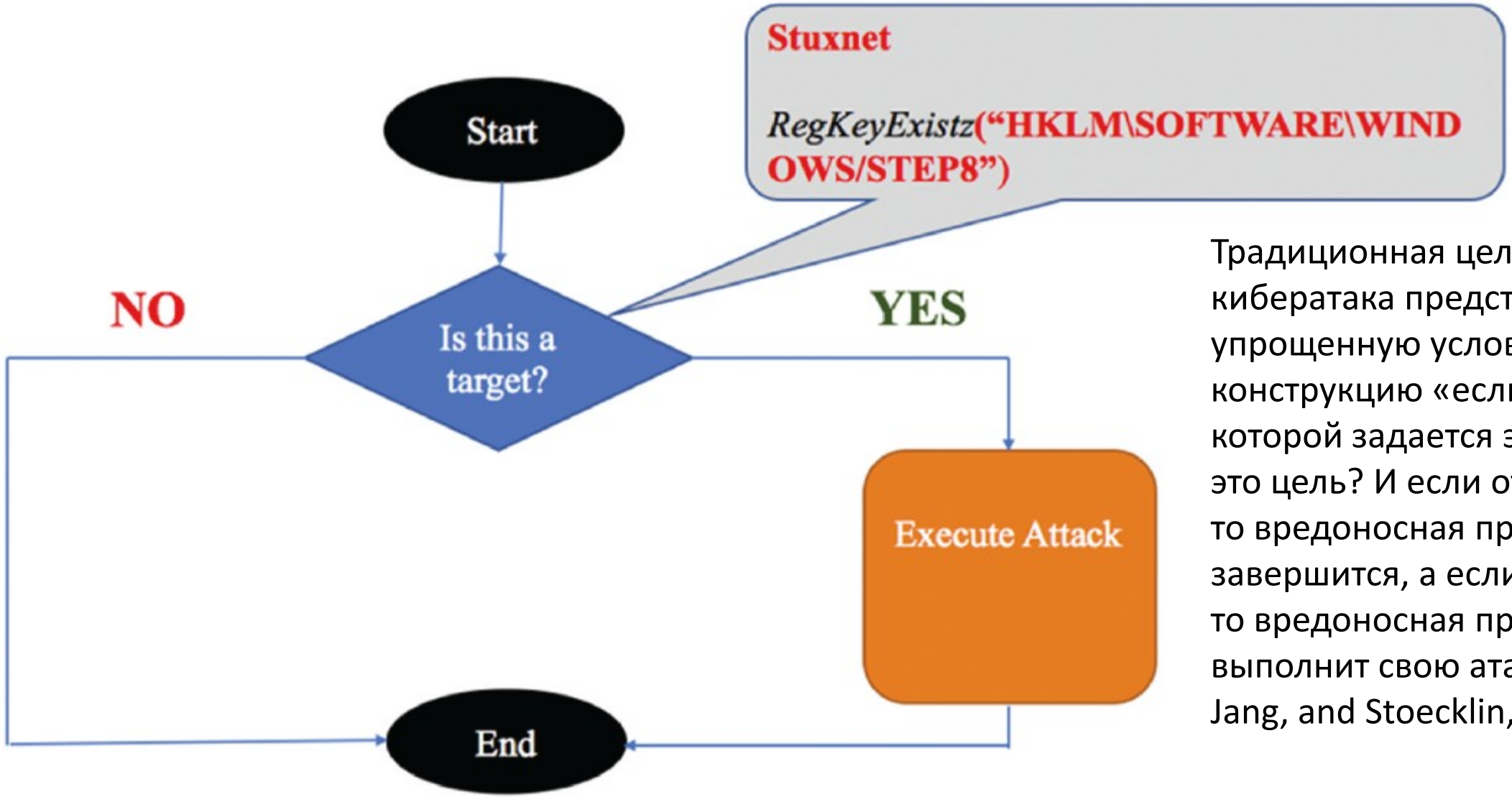
Содержание

- Введение
- Тренды в ИИ
- Повышение кибербезопасности на основе ИИ
- Использование ИИ для кибератак кибербезопасности
- Уязвимости систем ИИ к атакам
- Использование ИИ во вредоносных информационных операциях
- Заключение

Умная атака / атака, управляемая ИИ / атака с помощью ИИ

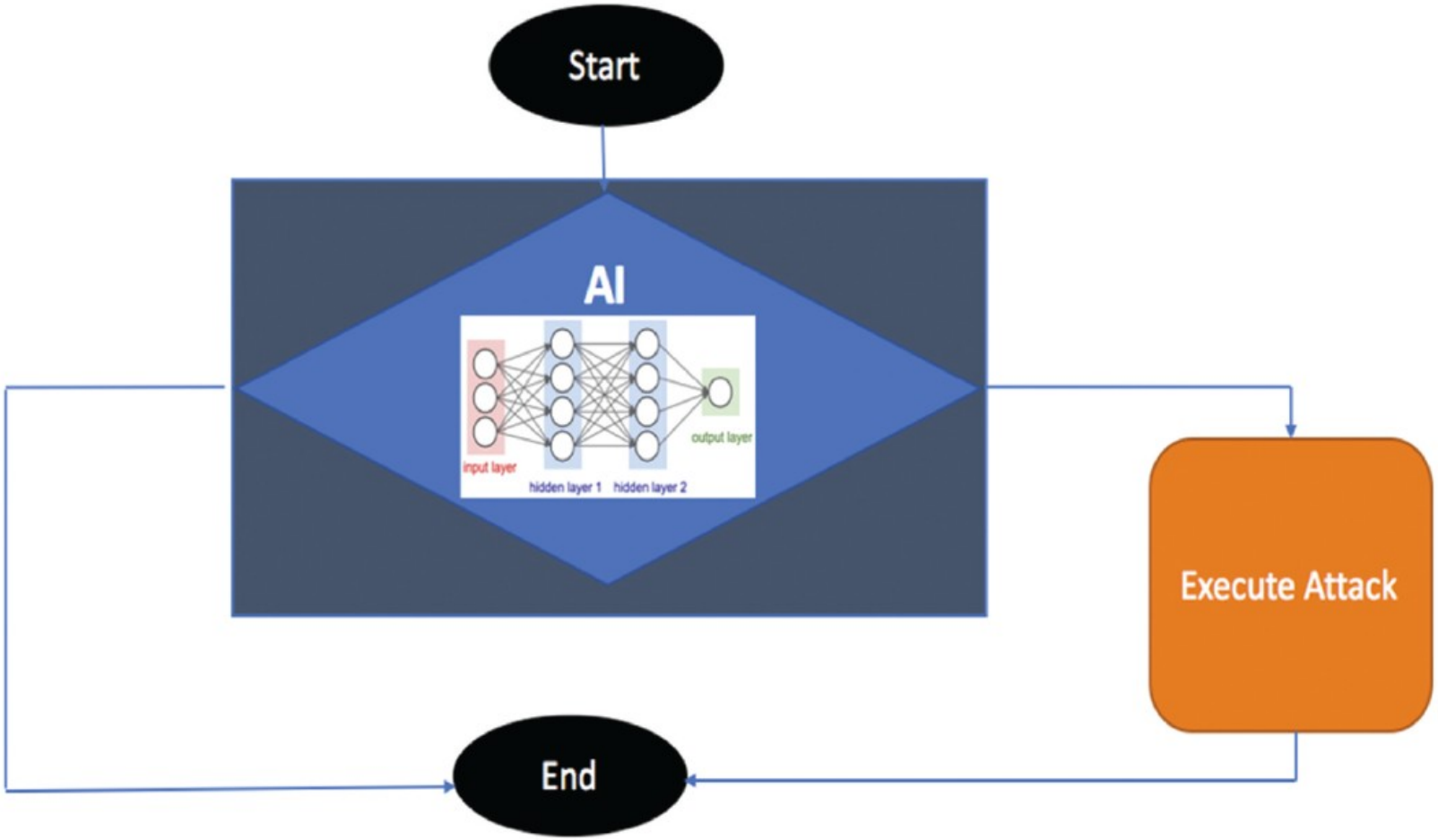
- **Умную/интеллектуальную (smart) атаку** можно определить «*как атаку с использованием ИИ, в которой злоумышленники могут использовать технологии ИИ для атаки на (интеллектуальные) компоненты внутри автономных систем. Умная атака обычно осуществляется посредством постоянного, точно нацеленного, комбинированного и многоуровневого использования нескольких зон безопасности замаскированным образом*» [J.Li et al., 2018].
- «**интеллектуальность обычно включает** в себя более сложную функциональность и более сложные взаимодействия, что, в свою очередь, увеличивает потенциальную поверхность атаки», и новые уязвимости могут быть созданы из-за охвата большего числа пользователей.
- Вредоносное использование ИИ для компрометации цифровой безопасности известно как **кибератака с использованием ИИ**, при которой киберпреступники могут обучать (программных) роботов социальной инженерии целей (атаки) с человеческим или сверхчеловеческим уровнем производительности [Brundage et al., 2018].
- **Атаки с помощью ИИ** могут **адаптироваться** к среде, которую они компрометируют, обучаясь на основе контекстных данных для эмуляции доверенных элементов киберпространства или выявления слабых мест [DarkTrace, 2020].
- Индустрия и сообщество безопасности должны понять, как ИИ применяется для кибератак и где находятся слабые места, чтобы **найти для них лучшую вакцину** [D.Patterson, 2018]

Традиционная логика принятия решения о целенаправленной кибератаке [Kirat, Jang, and Stoecklin, 2018]



Традиционная целевая кибератака представляет собой упрощенную условную конструкцию «если-то», в которой задается этот вопрос; это цель? И если ответ «Нет», то вредоносная программа завершится, а если ответ «Да», то вредоносная программа выполнит свою атаку. [Kirat, Jang, and Stoecklin, 2018].

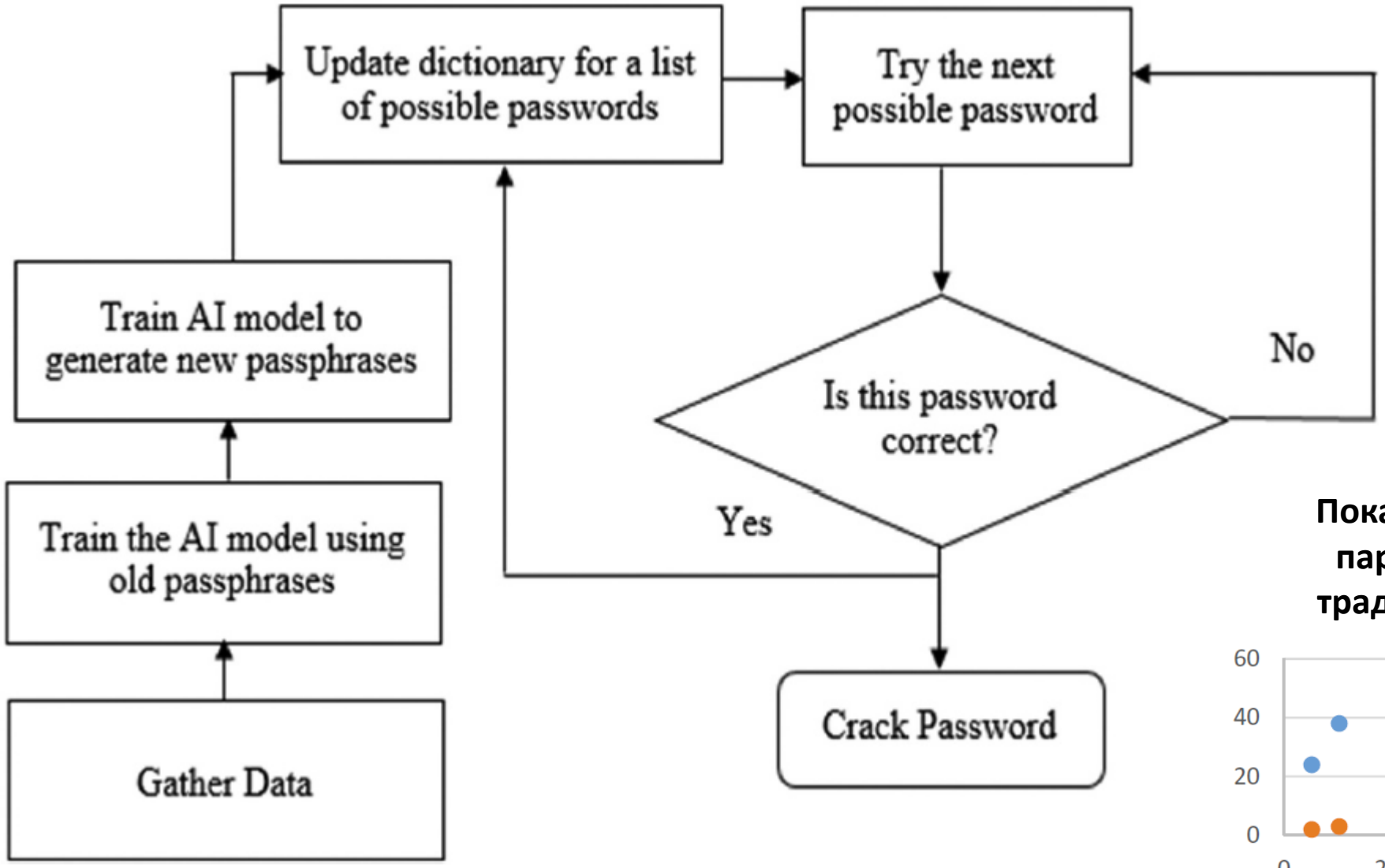
Принятие решения о кибератаке на основе DNN [Kirat, Jang, and Stoecklin, 2018]



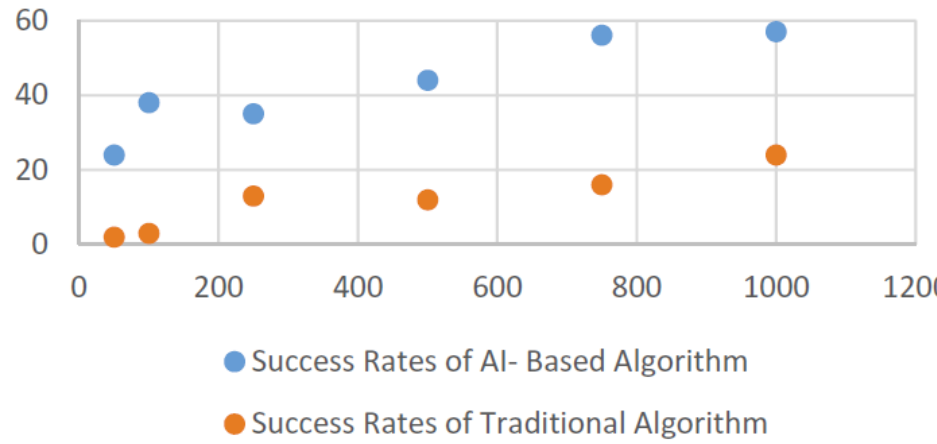
- Поскольку киберпреступники поняли, что эксперты по кибербезопасности используют **песочницы** для анализа и борьбы с этими традиционными целевыми атаками, они **трансформируют эту упрощенную форму условной конструкции «если-то» в логику принятия решений с использованием глубоких нейронных сетей (DNN).**
- С концепцией DNN злоумышленники могут решить, атаковать или нет. Проблема для защитника заключается в том, что будет крайне сложно выяснить, **что является фактическим вредоносным кодом**, а что — верной целью [Kirat, Jang, and Stoecklin, 2018].

https://www.blackhat.com/us-18/briefings/schedule/index.html#dee%20plocker%E2%80%94concealing-targeted-attacks-with-ailocksmithing-11549
https://github.com/CyberWarefare/DeepLocker

Атаки методом перебора паролей на основе ИИ [Trieu and Yang, 2018]



Показатели успешности перебора паролей с помощью ИИ против традиционной атаки грубой силы



Области использования ИИ во вредоносных программах [Thanh&Zelinka, 2019]

Содержание

- Введение
- Тренды в ИИ
- Повышение кибербезопасности на основе ИИ
- Использование ИИ для кибератак кибербезопасности
- Уязвимости систем ИИ к атакам
- Использование ИИ во вредоносных информационных операциях
- Заключение

MITRE ATLAS (Ландшафт состязательных угроз для систем ИИ)

MITRE ATLAS— это база знаний о тактике, методах и тематических исследованиях злоумышленников для систем машинного обучения (ML), основанная на реальных наблюдениях, демонстрациях от красных команд ML и групп безопасности, и состояние возможного из академических исследований.

Защитные методы машинного обучения [Q.LIU et al., 2018]

Содержание

- Введение
- Тренды в ИИ
- Повышение кибербезопасности на основе ИИ
- Использование ИИ для кибератак кибербезопасности
- Уязвимости систем ИИ к атакам
- Использование ИИ во вредоносных информационных операциях
- Заключение

Использование ИИ в вредоносных информационных операциях

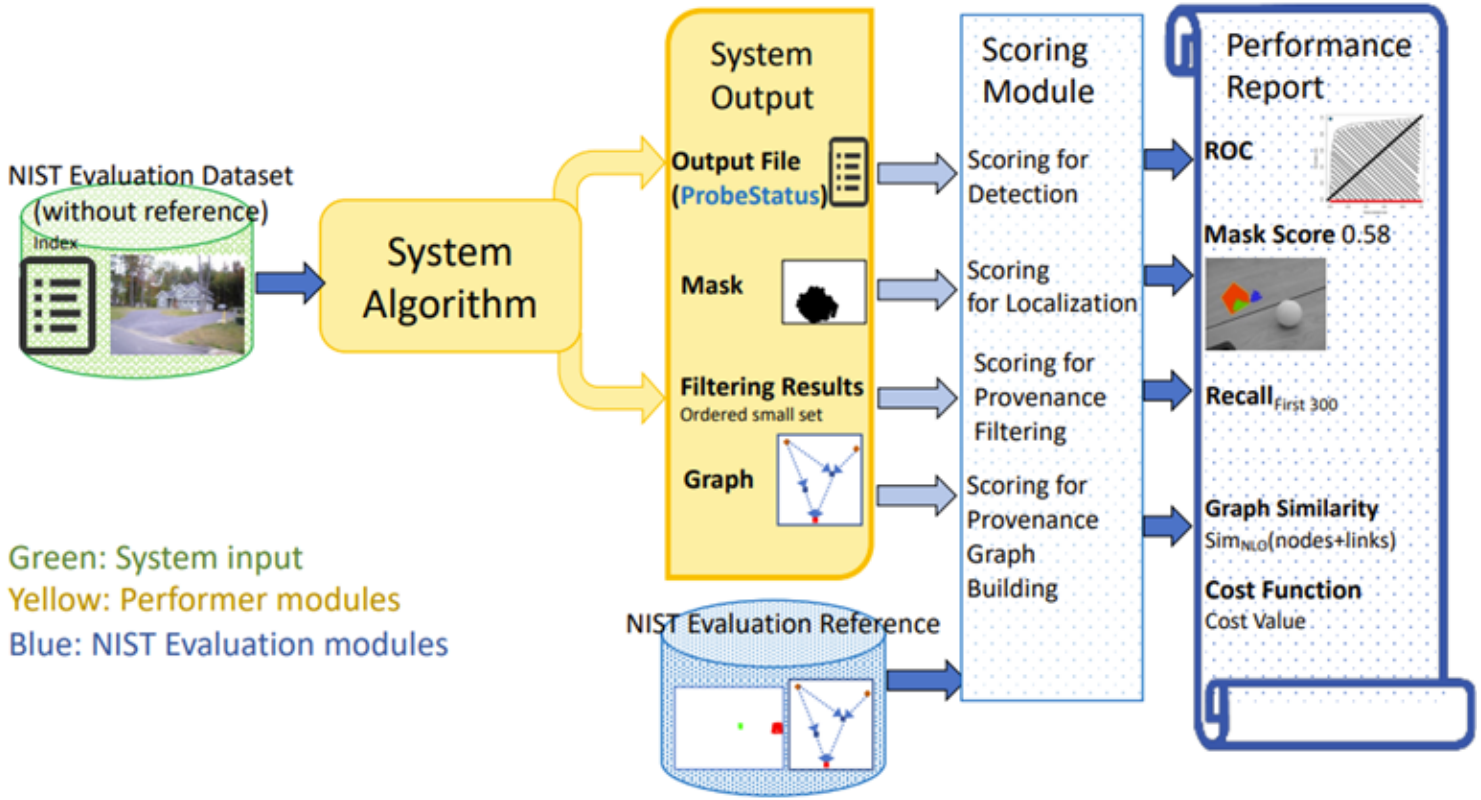
- Расширяются возможности по производству и распространению высококачественного аудиовизуального контента, называемого **синтетическими медиа и дипфейками**. Технологии ИИ для создания дипфейков теперь могут создавать контент, неотличимый от реальных людей, сцен и событий.
- ИИ для создания дипфейков может повысить эффективность **операций социальной инженерии** (программа выдает себя за некоторое реальное лицо) и убедить, например, конечных пользователей предоставить злоумышленникам доступ к системам и информации .
- Эти методы могут использоваться для создания правдоподобных заявлений мировых лидеров и командующих, для фабрикации убедительных **операций под ложным флагом и создания фальшивых новостей.**
- Злонамеренные субъекты называют реальные события «фальшивыми», видео- и фото-доказательства, например, изображения зверств, называют фейком. Распространение синтетических СМИ, известное как «**дивиденд лжеца**», побуждает людей называть настоящие СМИ «фальшивыми» и создает правдоподобное отрицание их действий
- Синтетические медиа и области их применения становятся все более изощренными, включая убедительное **чередование дипфейков с реально происходящими событиями** и синтез дипфейков в реальном времени.

[https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=930628 **]**

[Д.Е. Намиот, Е.А. Ильюшин, И.В. Чижов. Искусственный интеллект и кибербезопасность. 2022]

Примеры решений

- Программа **DARPA Semantic Forensics (SemaFor)** - направлена на разработку инновационных семантических технологий для анализа медиа.
- Программа **DARPA MediaForensics (MediaFor**) . Поставленная цель - разработать технологии автоматизированной оценки целостности изображения или видео



- **Майкрософт:** подход к противодействию угрозе синтетических носителей на основе технологии происхождения цифрового контента.

[https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=930628]

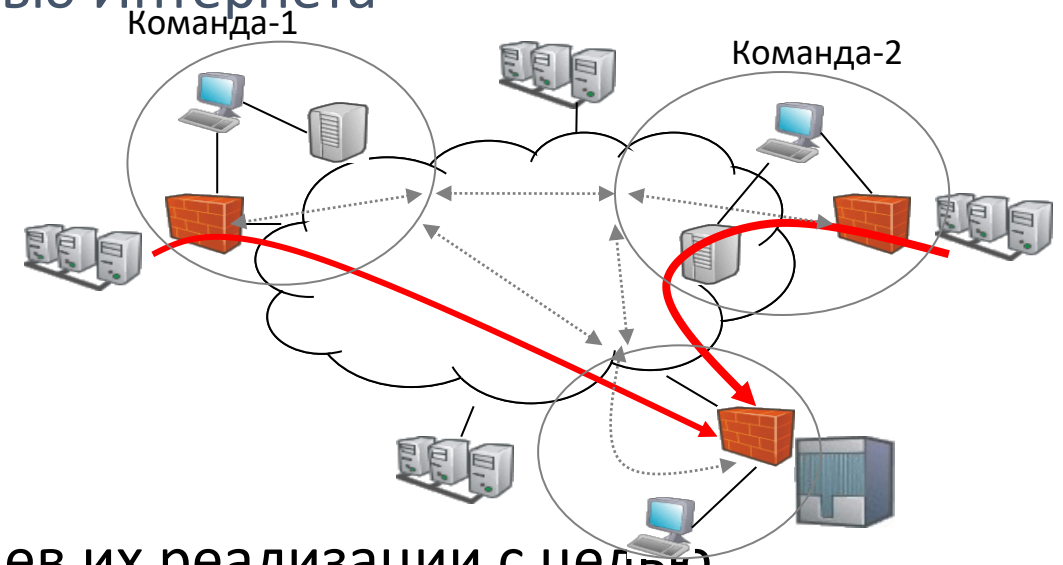
Содержание

- Вступление
- Тренды в ИИ и кибербезопасности
- Ключевые области внимания на стыке ИИ и кибербезопасности
- Использование ИИ в кибербезопасности
- Ландшафт киберугроз на основе ИИ
- Атаки на системы на основе ИИ
- Приложения кибербезопасности на основе ИИ
- Заключение

Многоагентное моделирование атак и механизмов защиты

Основные положения подхода к моделированию (1/2)

- Кибернетическое противоборство представляется в виде взаимодействия различных команд программных агентов
- Процессы происходят в среде, задаваемой моделью Интернета
- Выделяются команды агентов атаки, защиты и пользователей
- Команды взаимодействуют между собой: противоборствуют, кооперируются, адаптируются



- Команда агентов-злоумышленников эволюционирует посредством генерации новых экземпляров и типов атак, а также сценариев их реализации с целью преодоления подсистемы защиты.
- Команда агентов защиты адаптируется к действиям злоумышленников путем изменения исполняемой политики безопасности, формирования новых экземпляров механизмов и профилей защиты.

Многоагентное моделирование атак и механизмов защиты
Основные положения подхода к моделированию (2/2)

- Предлагаемый подход базируется на **комбинировании** элементов теории общих намерений, теории разделяемых планов и комбинированных подходов
- **Структура команды агентов** описывается в терминах иерархии групповых и индивидуальных ролей в различных сценариях действий
- **Спецификация иерархии планов** действий осуществляется для каждой из ролей.
- Для каждого плана описываются:
  (1) начальные условия, когда план предлагается для исполнения;
  (2) условия, при которых план прекращает исполняться;
  (3) действия, выполняемые на уровне команды, как часть общего плана
- **Назначение ролей и распределение планов** между агентами выполняется в два этапа: (1) сначала план распределяется в терминах ролей, (2) каждой из ролей ставится в соответствие агент

Многоагентное моделирование атак и механизмов защиты

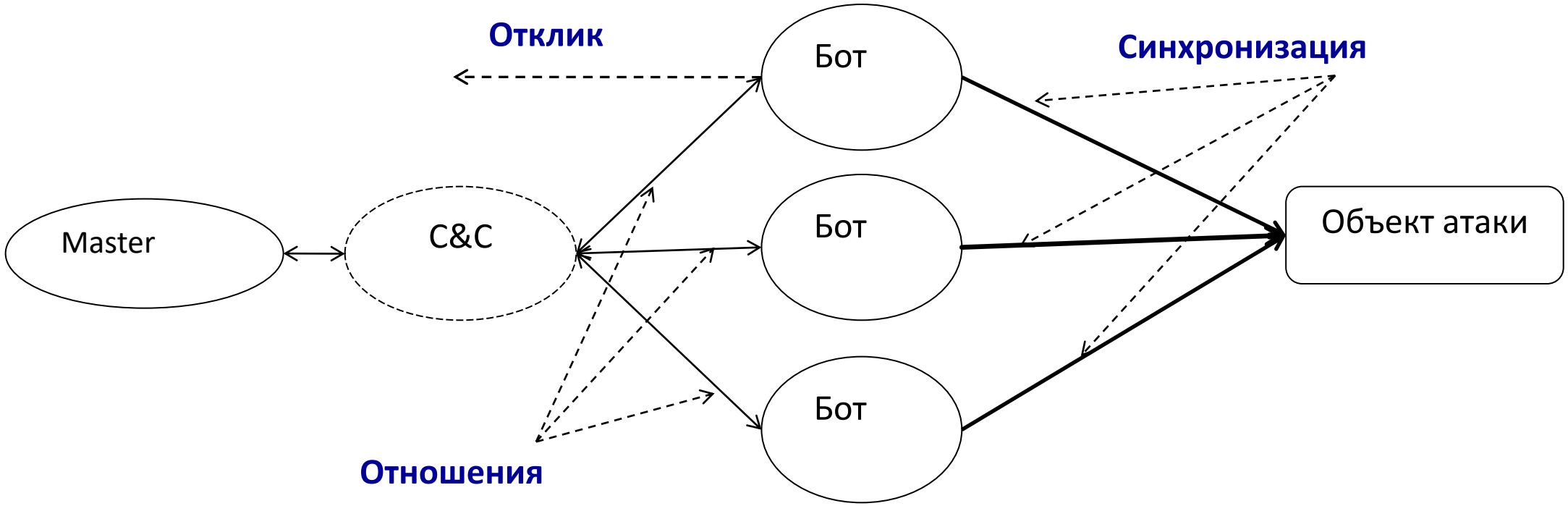
Процедуры поддержки командной работы

1. **Процедуры обеспечения согласованности действий агентов в команде** *(группе, индивидуально)* по некоторому общему плану.
2. **Процедуры мониторинга и восстановления функциональности команды** *(группы, индивидуально)* за счет переназначения “утерянных” ролей тем членам команды, которые в состоянии выполнить эту работу
3. **Процедуры обеспечения селективности коммуникаций**; основываются на расчете важности того или иного сообщения с учетом его “стоимости” и выгоды, получаемой при этом.

Kotenko I., Konovalov A., Shorov A. **Agent-based simulation of cooperative defence against botnets** // **Concurrency and Computation: Practice and Experience**, Vol. 24, Issue 6, 25 April 2012. P. 573-588.

Многоагентное моделирование атак и механизмов защиты

Метрики для обнаружения бот-сетей (1/2)



Возможная интерпретация:

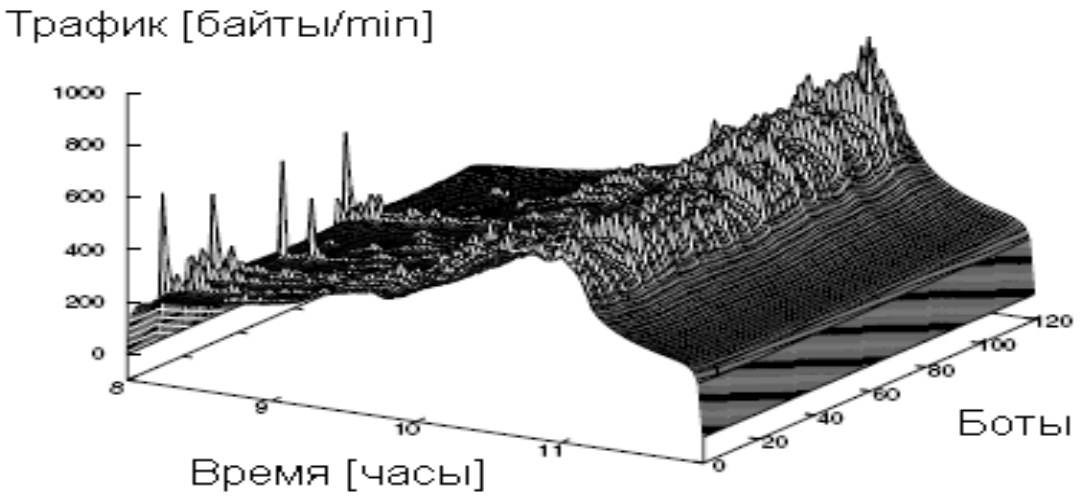
- Отношение – количество активных клиентов в канале (IRC)
- Отклик – время отклика клиентов на запрос
- Синхронизация – синхронизм трафика, посланного клиентами

[Akiyama, etc., 2007]

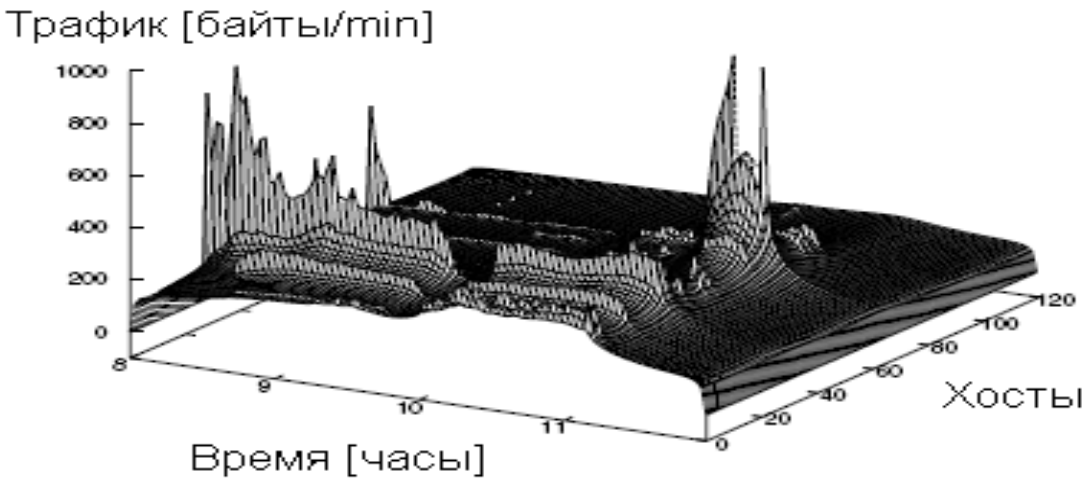
Многоагентное моделирование атак и механизмов защиты

Метрики для обнаружения бот-сетей (2/2)

Синхронизированный трафик



IRC-трафик бот-сети



IRC-трафик обычных пользователей

[Akiyama, etc., 2007]

Многоагентное моделирование атак и механизмов защиты

Критерии адаптации команд агентов (1/2)

- Субъекты взаимодействия (*S*): команды атаки и защиты.
- Общий подход к адаптации:

$$\min_{S(t)} \sum_{i=1}^{n} C_i\left(S(t), K_D(t)\right)$$

Отражает стоимости

Конфигурация системы

Показатель мощности атаки/защиты

  - Критерий адаптации команды защиты:

$$\min_{S(t)} \sum_{i=1}^{n} \left\{C_{FP}\left(S(t), K_D(t)\right) + C_{FN}\left(S(t), K_D(t)\right) + C_T\left(S(t), K_D(t)\right)\right\}$$

Процент ложных срабатываний

Процент пропусков атак

Продолжительность атаки

  - Критерий адаптации команды атаки:

$$\min_{E(t)} \sum_{i=1}^{n} \left\{C_P\left(E(t), K_A(t)\right) + C_D\left(S(t), K_A(t)\right)\right\}$$

Количество пакетов

Количество обезвреженных «демонов»

Многоагентное моделирование атак и механизмов защиты

Критерии адаптации команд агентов (2/2)

$K_D(t) = \{M_i, TK_j\}$ – конфигурация системы защиты на время t,

где M_i – метод защиты и его параметры (полученные во время обучения),

TK_j – схема кооперации (без кооперации, на уровне фильтров, сэмплеров и полная)

$K_A(t) = \{I_i, R_j\}$ – параметры атаки на время t,

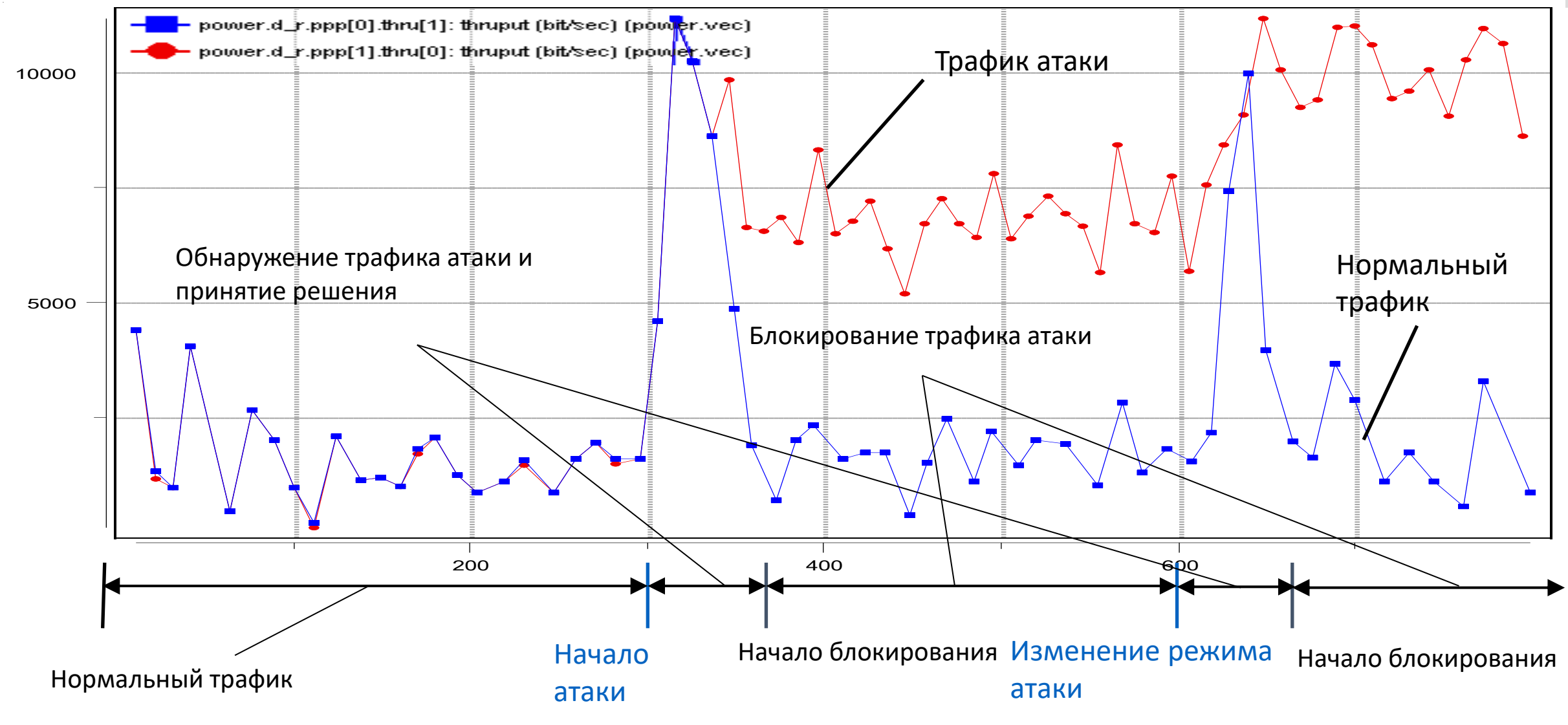
где I_i – интенсивность атаки (задается злоумышленником),

R_j – метод подмены адреса отправителя (без подмены, постоянная, случайная, случайная той же подсети)

Многоагентное моделирование атак и механизмов защиты

Принятие решения и функционирование

Графики изменения пропускной способности канала на входе в защищаемую сеть (зависимость бит/с от времени) до (красный) и после фильтра (синий)



Многоагентное моделирование атак и механизмов защиты

Параметры моделирования

- *Топология и конфигурация сети:* количество и типы хостов и каналов связи между ними, характеристики каналов и хостов.
- *Конфигурация команд атаки:* количество демонов; адрес и порт мастера для взаимодействия; порт демона для посылки пакетов атаки; адрес и порт цели атаки; время атаки; интенсивность атаки; метод подмены адреса отправителя.
- *Параметры команд защиты*: адрес защищаемого узла, адрес и порт “детектора” для взаимодействий, размер ответа на запрос и время обработки запроса сервером; схема адаптации и т.п.
- *Параметры команды пользователей*: количество пользователей; адрес и порт сервера; время начала работы; количество, запросов, интервал между запросами, размер запросов к серверу в одном соединении; интервал между соединениями.
- *Параметры кооперации агентов защиты:* схема кооперации.
- *Параметры моделирования*: продолжительность моделирования, количество экспериментов и др.
- *Параметры атак*
- *Параметры механизмов защиты*

Заключение

- В докладе **представлено современное состояние использования ИИ в кибербезопасности** в виде сопоставления методов и технологий нападения и защиты. Рассмотрены аспекты защиты от атак с использованием ИИ, атаки с использованием ИИ, защиты самих систем машинного обучения и использование ИИ в вредоносных информационных операциях.

- Проведен анализ данных, используемых для решения задач кибербезопасности, обобщены существующие области применения ИИ для кибербезопасности.

- Рассмотрен пример исследования в области моделирования интеллектуальных атак.

- Данные исследования выполняются при финансовой поддержке Гранта РНФ № 21-71-20078 в СПб ФИЦ РАН.

(25-30 сентября 2023 г., Санкт-Петербург, Россия)

- 7-я Международная научная конференция
- **«Интеллектуальные информационные технологии в технике и на производстве» (IITI’23)**

https://iiti.rgups.ru/

- Проблемы кибербезопасности в условиях становления Industry 4.0.

6 марта 2023 года -> 27 марта 2023 года
Окончание приема заявок на участие. Для подачи заявки необходимо заполнить все поля (**кроме файла публикации**) и отправить форму через вкладку New Submission на EasyChair (https://easychair.org/conferences/?conf=iiti23). Для этого необходимо быть зарегистрированным в системе.

3 апреля 2023 года. -> 17 апреля 2023 года
Предоставление полного текста публикаций на сайте EasyChair.

Контакты

Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), Санкт-Петербургский институт информатики и автоматизации Российской академии наук, Лаборатория проблем компьютерной безопасности,

Адрес: 39, 14 линия В.О., Санкт-Петербург, 199178

Телефон: +7(812)328-71-81

URL: http://comsec.spb.ru

Контакты:

ГНС, д.т.н., проф. Котенко Игорь Витальевич, ivkote@comsec.spb.ru, http://comsec.spb.ru/kotenko

СПб ФИЦ РАН

Благодарности

- Работа выполнена при финансовой поддержке гранта РНФ 21-71-20078 в СПб ФИЦ РАН.

СПИИРАН