EUROPEAN UNION AGENCY
FOR CYBERSECURITY

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И КИБЕРБЕЗОПАСНОСТЬ

ИССЛЕДОВАТЬ

Краткий обзор исследований и инноваций ENISA

ИЮНЬ 2023 Г.

О ЕНИСЕ

Агентство Европейского Союза по кибербезопасности, ENISA, является агентством Союза, занимающимся достижением высокого общего уровня кибербезопасности в Европе. Агентство Европейского Союза по кибербезопасности, созданное в 2004 году и усиленное Законом ЕС о кибербезопасности, вносит свой вклад в киберполитику ЕС, повышает надежность продуктов, услуг и процессов ИКТ с помощью схем сертификации кибербезопасности, сотрудничает с государствами-членами и органами ЕС и помогает Европе подготовиться к кибервызовы завтрашнего дня. Посредством обмена знаниями, наращивания потенциала и повышения осведомленности Агентство работает вместе со своими ключевыми заинтересованными сторонами над укреплением доверия к взаимосвязанной экономике, повышением устойчивости инфраструктуры Союза и, в конечном счете, обеспечением цифровой безопасности европейского общества и граждан. Более подробную информацию об ENISA и ее работе можно найти здесь:www.enisa.europa.eu.

КОНТАКТ

Для связи с авторами используйтеrit@enisa.europa.eu.
Для запросов СМИ об этой газете, пожалуйста, используйтеpress@enisa.europa.eu.

РЕДАКТОРЫ

Корина Паску (ENISA), Марко Баррос Лоренсу (ENISA)

АВТОРЫ

д-р Ставрос НТАЛАМПИРАС, Миланский университет, I; Доктор Джанлука МИСУРАКА, соучредитель и вице-президент Inspiring Futures, ES; Доктор Пьер Россель, президент Inspiring Futures CH

ISBN: 978-92-9204-637-8, DOI: 10.2824/808362

ОГЛАВЛЕНИЕ

УПРАВЛЯЮЩЕЕ РЕЗЮМЕ

Искусственный интеллект (ИИ) — это типичная технология двойного назначения, в которой злоумышленники и новаторы постоянно пытаются улучшить работу друг друга. Это обычная ситуация с технологиями, используемыми для подготовки стратегической разведки и поддержки принятия решений в критических областях. Злоумышленники учатся тому, как сделать свои атаки более эффективными, используя эту технологию для поиска и использования уязвимостей в системах ИКТ.

Сделав еще один шаг в прояснении этого первоначального утверждения: с помощью ИИ злоумышленники могут внедрять новые возможности, которые могут продлить или даже расширить практики киберугроз, которые уже существуют в течение длительного времени. Благодаря ИИ эти возможности постепенно становятся автоматизированными, и их труднее обнаружить. В данном исследовании рассматриваются некоторые из этих возможностей с исследовательской точки зрения.

В этом исследовании были рассмотрены два аспекта ИИ (категоризация объяснена в разделе 4): (a) обеспечение безопасного и надежного ИИ и предотвращение его злонамеренного использования («ИИ как криминальная услуга» или «ИИ для нанесения вреда ') и (b) использование ИИ в кибербезопасности («варианты использования ИИ» или «ИИ для защиты»).

Случаи использования ИИ в кибербезопасности многочисленны и расширяются. Их исчерпывающий перечень выходит за рамки данного исследования, так как исследования в этой области постоянно развиваются. Тем не менее, мы приводим примеры некоторых из этих вариантов использования в отчете, чтобы лучше объяснить текущие исследования в области этой технологии и изучить области, в которых необходимы дальнейшие исследования.

The **цель этого исследования - определить потребности в исследованиях ИИ для кибербезопасности и обеспечения безопасности ИИ,**в рамках работы ENISA по выполнению своего мандата в соответствии со статьей 11 Закона о кибербезопасности[1]. Этот отчет является одним из результатов этой задачи. В нем мы представляем результаты работы, проведенной в 2021 году[2]и впоследствии подтвержден в 2022 и 2023 годах заинтересованными сторонами, экспертами и членами сообщества, такими как ENISA AHWG по искусственному интеллекту.[3]. ENISA внесет свой вклад путем определения пяти ключевых потребностей в исследованиях, которые будут переданы и обсуждены с заинтересованными сторонами в качестве предложений по будущей политике и инициативам по финансированию на уровне ЕС и государств-членов.

В этом отчете не представлена приоритизация потребностей в исследованиях. ENISA ежегодно проводит расстановку приоритетов с учетом общего состояния исследований и инноваций в области кибербезопасности в ЕС, политики и инициатив по финансированию исследований и инноваций в области кибербезопасности в Союзе, а также технического анализа по конкретным темам и технологиям. Приоритеты на 2022 год можно найти в кратком отчете ENISA об исследованиях и инновациях.

Кроме того, в 2022 году ENISA провела исследование, в котором рассмотрела работу 44 исследовательских проектов, программ и инициатив в области кибербезопасности и ИИ, которые были наиболее

[1] https://digital-strategy.ec.europa.eu/en/policies/cybersecurity-act , последний доступ январь 2023 г.
[2] Соображения в этом исследовании являются результатом обзора литературы, в том числе предыдущей работы ENISA по ИИ, например, «Защита алгоритмов машинного обучения»: https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms
[3] Специальная рабочая группа по кибербезопасности искусственного интеллекта — ENISA (europa.eu) .

часть финансируется рамочными программами ЕС на период с 2014 по 2027 год. Важность этого перечня связана с особой ролью, которую ИИ играет в области исследований в области кибербезопасности, учитывая постоянное и усиливающееся взаимодействие с другими семействами технологий. Фундаментальный вопрос, лежащий в основе этого исследования, заключался в том, позволили ли Европе инвестиции в исследования и разработки в области кибербезопасности в области ИИ добиться прогресса в этой области, особенно те, которые поддерживаются фондами ЕС. Результаты этого исследования также можно найти в Кратком отчете об исследованиях и инновациях ENISA за 2022 год.

Хотя мы признаем огромный потенциал ИИ для инноваций в области кибербезопасности и множество требований, необходимых для повышения его безопасности, мы также признаем, что еще предстоит проделать большую работу, чтобы полностью раскрыть и описать эти требования. Этот отчет является лишь начальной оценкой того, где мы находимся и где нам нужно искать дальше в этих двух важных аспектах этой технологии.

Кроме того, согласно результатам исследования ENISA по финансируемым ЕС исследовательским проектам в области кибербезопасности и ИИ, упомянутым ранее, большинство рассмотренных проектов были сосредоточены на методах машинного обучения. Это можно интерпретировать двояко: как знак того, что рынок таких решений особенно ценит потенциальные преимущества машинного обучения по сравнению с другими областями ИИ, или что по какой-то причине исследования и разработки в других областях ИИ не проводятся. должным образом рассматриваются государственными спонсорами, несмотря на их признанный потенциал. В этом исследовании мы также подчеркиваем необходимость дальнейшего изучения использования машинного обучения в кибербезопасности, а также изучения других концепций ИИ.

ENISA выполнила шаги, описанные в следующем списке, чтобы определить потребности в исследованиях, представленные в главе 7.2 этого отчета.

- Выявление из существующих исследовательских работ функций и вариантов использования, в которых ИИ используется для поддержки мероприятий по кибербезопасности, представленных в главе 3.
- Выявление из существующих исследовательских работ областей, в которых кибербезопасность необходима для защиты ИИ, представлено в главе 4.
- Обзор вариантов использования ИИ, представленный в главе 5.
- Анализ открытых вопросов, проблем и пробелов, представленных в главе 6.
- Выявление областей, в которых требуются дополнительные знания.

Эти шаги были выполнены экспертами, которые внесли свой вклад в этот отчет, в основном посредством кабинетных исследований, а результаты были подтверждены членами сообщества R&I.

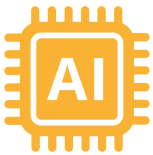
ENISA готовит эти исследования с целью использования их в качестве инструмента для разработки рекомендаций по исследованиям и разработкам в области кибербезопасности и представления их заинтересованным сторонам. Эти заинтересованные стороны являются основной целевой аудиторией настоящего отчета и включают членов более широкого научно-исследовательского сообщества (ученых, исследователей и новаторов), представителей промышленности, Европейской комиссии (ЕК), Европейского центра компетенции в области кибербезопасности (ECCC) и национальных координационных центров (НКЦ).

5 ОСНОВНЫХ ПОТРЕБНОСТЕЙ В ИССЛЕДОВАНИЯХ ДЛЯ ИИ И КИБЕРБЕЗОПАСНОСТИ

1

Test beds to study and optimise the performance of ML-based tools and technologies used for cybersecurity

2



Incentivise the development of penetration testing tools based on AI and ML to find and exploit security vulnerabilities to assess attacker behaviours

3

Development of standardised frameworks assessing the preservation of privacy and the confidentiality of information flows as well as of the designed systems

4

Development of training in AI for practitioners using real-world scenarios

5

Establishing an observatory for AI and cybersecurity threats

Примечание. Дополнительную информацию об этих приоритетах можно найти в главе 7 настоящего отчета.

ОПРЕДЕЛЕНИЕ ТЕРМИНОВ И СОКРАЩЕНИЙ

В следующем списке описаны термины, используемые в этом документе.

Искусственный Интеллект (ИИ)	Общепринятого определения ИИ не существует.[4]. Хотя общее определение отсутствует, можно отметить ряд общих черт (см. JRC[5]) в проанализированных определениях, которые можно рассматривать как основные черты ИИ: (i) восприятие окружающей среды, включая учет сложности реального мира; (ii) обработка информации (сбор и интерпретация входных данных (в виде данных); (iii) принятие решений (включая рассуждение и обучение): выполнение действий, выполнение задач (включая адаптацию и реакцию на изменения в окружающей среде) с определенной уровень автономии (iv) достижение конкретных целей.
Искусственный Интеллект системы	Системы ИИ — это программное обеспечение (которое разработано с использованием подходов машинного обучения и подходов, основанных на логике и знаниях).[6]). Кроме того, они могут для данного набора целей, определенных человеком, генерировать выходные данные, такие как контент, прогнозы и рекомендации или решения, влияющие на среду, с которой они взаимодействуют. Системы ИИ также могут включать в себя аппаратные системы, разработанные людьми, которые, имея сложную цель, действуют в физическом или цифровом измерении, воспринимая свою среду посредством сбора данных, интерпретируя собранные структурированные или неструктурированные данные, рассуждая о знаниях или обрабатывая полученную информацию. на основе этих данных и принятия решения о наилучших действиях, которые необходимо предпринять для достижения поставленной цели.[7] [8].
Искусственный нейрон сети (ANN)	Искусственные нейронные сети (ИНС), обычно называемые просто нейронными сетями (НС) или нейронными сетями, представляют собой вычислительные системы, основанные на наборе связанных единиц или узлов, называемых искусственными нейронами, которые приблизительно моделируют нейроны в биологическом мозге.
Кибер-физический системы (СУЗ)	Киберфизические системы (CPS) — это интеграция вычислений, связи и управления, которые обеспечивают желаемую производительность физических процессов.
Древо решений (ДТ)	Обучение с помощью дерева решений — это форма контролируемого машинного обучения.

[4]Европейская комиссия. Объединенный исследовательский центр. AI Watch: определение искусственного интеллекта: к рабочему определению и таксономии искусственного интеллекта. Publications Office, 2020. doi: 10.2760/382730. URL https://data.europa.eu/doi/10.2760/382730. Обновление этого технического отчета JRC в 2021 г.https://aiwatch.ec.europa.eu/document/download/e90645f1-662e-470d-9af9-848010260b1f_en предоставил качественный анализ еще 37 политических и институциональных отчетов в области ИИ, 23 соответствующих исследовательских публикаций и 3 рыночных отчетов с момента появления ИИ в 1955 году до 2021 года.

[5]То же, что и 4

[6]Предложение Комиссии по Регламенту ЕС и Общему подходу Совета к проекту Закона об искусственном интеллекте, декабрь 2022 г., https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf. Первоначальное определение в Предложении Комиссии было сужено Советом, чтобы отличить ИИ от более классических программных систем.

[7]То же, что и 4. ETSI определяет ИИ (систему) как: «Искусственный интеллект — это способность системы обрабатывать представления, как явные, так и неявные, и процедуры для выполнения задач, которые считались бы интеллектуальными, если бы их выполнял человек».

[8]Юридическое определение ИИ в проекте Регламента ЕС находится в стадии разработки в парламенте ЕС.

Глубокое обучение (дл)	Глубокое обучение[9] is part of a broader family of machine learning methods based on artificial neural networks (ANNs[10]).
Ансамбль методы	Методы, направленные на повышение точности результатов в моделях путем объединения нескольких моделей вместо использования одной модели.
Скрытый Марков Модель (ХММ)	Скрытая марковская модель (HMM) — это статистическая модель, которая также используется в машинном обучении. Его можно использовать для описания эволюции наблюдаемых событий, которые зависят от внутренних факторов, не наблюдаемых напрямую. Скрытые марковские модели (HMM) первоначально возникли в области распознавания речи. В последние годы они вызывают растущий интерес и к области компьютерного зрения.
К-означает кластеризация	Кластеризация K-средних — один из самых простых и популярных алгоритмов машинного обучения без учителя.
Машинное обучение (мл)	Машинное обучение — это подмножество ИИ, которое в основном использует расширенную статистику для создания платформ с возможностью учиться на доступных данных, выявлять закономерности и делать прогнозы, не требуя вмешательства человека.[11].
Наивный Байес классификатор (НБ)	Naive Bayes — популярный алгоритм машинного обучения с учителем.
Армирование обучение (РЛ)	Обучение с подкреплением (RL) — это область машинного обучения, связанная с тем, как интеллектуальные агенты выполняют действия в окружающей среде, чтобы максимизировать понятие кумулятивного вознаграждения. Обучение с подкреплением — это одна из трех основных парадигм машинного обучения, наряду с обучением с учителем и обучением без учителя.
Безопасность по-дизайн	Концепция разработки программного обеспечения и дизайна продукта, учитывающая соображения безопасности на ранних этапах разработки продукта.
Контролируемое машинное обучение	Обучение с учителем — это подкатегория машинного обучения, определяемая использованием помеченных наборов данных для обучения алгоритмов для точной классификации данных или прогнозирования результатов.
Вектор поддержки Машина (СВМ)	Алгоритм машины опорных векторов (SVM) — это алгоритм обучения с учителем, используемый для классификации наборов обучающих данных.
Неконтролируемое отмывание денег	Одна из трех основных парадигм машинного обучения, наряду с обучением с подкреплением и обучением с учителем, связанная с процессом вывода скрытых шаблонов из исторических данных.[12].

[9] Например, Лекун, Янн; Бенжио, Йошуа; Хинтон, Джеффри (2015). Глубокое обучение. Природа. 521 (7553): 436–444. Бибкод: 2015 Природа 521.436L. DOI: 10.1038/nature14539

[10] Например, Хардести, Ларри (14 апреля 2017 г.). Объяснение: Нейронные сети. Офис новостей Массачусетского технологического института. Проверено 2 июня 2022 г.

[11] Дипанкар Дасгупта, Захид Ахтар и Саджиб Сен. Машинное обучение в кибербезопасности: всесторонний обзор. Журнал оборонного моделирования и моделирования: приложения, методология, технология, стр. 154851292095127, сентябрь 2020 г. doi: 10.1177/1548512920951275. URL-адрес https://doi.org/10.1177/1548512920951275

[12] Хинтон, Джеффри; Сейновски, Терренс (1999). Неконтролируемое обучение: основы нейронных вычислений. Массачусетский технологический институт Пресс. ISBN 978-0262581684.

ОСНОВНЫЕ КОНЦЕПЦИИ И ФУНКЦИИ ИИ

Машинное обучение на сегодняшний день является самой популярной областью в области ИИ. Он используется в кибербезопасности различными способами. В таблице 1 ниже показано использование методов ИИ в функциях кибербезопасности.

Машинное обучение включает в себя разработку алгоритмов и статистических моделей, которые позволяют компьютерным системам учиться на собственном опыте и совершенствоваться без необходимости явного программирования. В этой главе мы разделяем существующие методы машинного обучения на две отдельные группы: традиционные инструменты и методы, основанные на нейронных сетях. Этот тип категоризации широко используется в литературе, чтобы показать преимущества и недостатки каждого инструмента.

Существуют и другие способы категоризации, в зависимости от использования информации (контролируемый или неконтролируемый), области применения (классификация, регрессия и кластеризация), глубины архитектуры (мелкая или глубокая) и т. д.

Следует также упомянуть еще одну школу мысли, а именно обучение с подкреплением (RL), гибридный подход, целью которого является изучение среды с помощью агента на основе проб и ошибок.

Таблица 1:Методы ИИ в функциях кибербезопасности (источник: авторы)

Функция безопасности/ИИ	ДТ	SVM	NB	K-означает	ХМ	ГАЗ	АННА	Си-Эн-Эн	РНН	Кодировщики	SNN
Обнаружения вторжений	Икс	Икс	Икс	Икс	Икс	Икс	Икс	Икс	Икс		Икс
Обнаружение вредоносных программ	Икс	Икс	Икс	Икс				Икс	Икс		
Уязвимость оценка	Икс										
Фильтрация спама			Икс								
Обнаружение аномалий					Икс					Икс	
Классификация вредоносных программ						Икс	Икс				Икс
Обнаружение фишинга							Икс				
Анализ трафика								Икс	Икс		
Сжатие данных										Икс	
Извлечение признаков										Икс	

1.1ТРАДИЦИОННЫЙ МЛ

Традиционные решения на основе ML включают кластеризацию DT, SVM и K-средних, которые широко используются в различных задачах кибербезопасности, таких как обнаружение спама.[13],

[13]Саумья Гоял, Р.К. Чаухан и Шабнам Парвин. Обнаружение спама с помощью KNN и механизмов дерева решений в социальных сетях. В 2016 г. Четвертая международная конференция по параллельным, распределенным и грид-вычислениям (PDGC), стр. 522–526, 2016 г. doi:10.1109/PDGC.2016.7913250.

вторжения[14] и вредоносное ПО[15], или в моделировании киберфизических систем[16]. Они будут подробно описаны в следующих разделах настоящего отчета.

1.1.1 Деревья решений (DT)

DT широко используются для обнаружения спама и вторжений.[17] из-за их способности идентифицировать правила и шаблоны в данных сетевого трафика и активности системы. DT реализует ряд правил, полученных из доступных размеченных данных, организованных в древовидную структуру.[18]. Различные методы ML, такие как DT, использовались для **обнаруживать кибератаки**. Поскольку DT полагаются на обучающие данные из прошлых инцидентов и происшествий, большинство из них не могут обнаружить новые типы, которые не являются частью набора данных.

Пространство для возможных деревьев решений экспоненциально велико, что приводит к «жадным подходам».[19] которые часто не могут найти лучшее дерево. DT не учитывают взаимодействия между атрибутами, и каждая граница решения включает только один атрибут. Особое внимание необходимо, чтобы избежать чрезмерной или недостаточной подгонки[20] (например, предварительная обрезка, постобрезка и т. д.), где сосредоточено большинство исследований[21].

В целом, DT недороги в построении, быстро классифицируют неизвестные записи, легко интерпретируются для деревьев небольшого размера, устойчивы к шуму (особенно когда используются методы, позволяющие избежать переобучения) и могут легко справляться с избыточностью.

1.1.2 Машины опорных векторов (SVM)

SVM — это тип алгоритма машинного обучения, который можно использовать для классификации или регрессионного анализа. Это один из самых известных алгоритмов для приложений кибербезопасности, поскольку он подходит для решения как **задачи обнаружения аномалий и распознавания образов** (спам, вредоносное ПО и обнаружение вторжений[22]). SVM известны своей устойчивостью к шуму.

---

[14] С. Кришнавени, Палани Виннешвар, С. Кишор, Б. Джоти и С. Сивамохан. Система обнаружения вторжений на основе аномалий с использованием метода опорных векторов. В Достижениях в области интеллектуальных систем и вычислений, страницы 723–731. Springer Singapore, 2020. doi:10.1007/978-981-15-0199-9_62. URL-адрес https://doi.org/10.1007/978-981-15-0199-9_62

[15] Басир Печаз, Маджид Вафайе Джахан и Мехрдад Джалали. Обнаружение вредоносного ПО с использованием скрытой марковской модели на основе метода выбора признаков марковского покрытия. В 2015 г. Международный конгресс по технологиям, коммуникации и знаниям (ICTCK), стр. 558–563, 2015 г. doi:10.1109/ICTCK.2015.7582729.

[16] Чезаре Алиппи, Ставрос Нталампирас и Мануэль Ровери. Идентификация неисправностей датчиков без использования моделей в режиме онлайн и изучение словаря в киберфизических системах. В 2016 г. Международная объединенная конференция по нейронным сетям (IJCNN), стр. 756–762, 2016 г. doi: 10.1109/IJCNN.2016.7727276.

[17] Б.К. Нирупама; М. Ниранджанамурти, Обнаружение сетевых вторжений с использованием дерева решений и случайного леса. В 2022 г. Международная конференция по достижениям в области вычислительной техники, связи и прикладной информатики (ACCAI), DOI: 10.1109/ACCAI53970.2022.9752578. Маниш Кумар, М. Ханумантаппа и ТВ Суреш Кумар. Система обнаружения вторжений с использованием алгоритма дерева решений. В 2012 г. 14-я Международная конференция IEEE по коммуникационным технологиям, страницы 629–634, 2012 г. DOI: 10.1109 / ICCT.2012.6511281.

[18] Виктор Х. Гарсия, Рауль Монрой и Марисела Кинтана. Обнаружение веб-атак с использованием ID3. В «Профессиональной практике в области искусственного интеллекта», страницы 323–332. Springer US, 2006. doi: 10.1007/978-0-387-34749-3_34. URL-адрес https://doi.org/10.1007/978-0-387-34749-3_34 . А также Шон Т. Миллер и Кертис Басби-Эрл. Многоперспективное машинное обучение метода ансамбля классификаторов для обнаружения вторжений. В материалах Международной конференции по машинному обучению и программным вычислениям 2017 г. - ICMLSC '17. ACM Press, 2017. doi: 10.1145/3036290.3036303. URL-адрес https://doi.org/10.1145/3036290 3036303.

[19] Подходы, основанные на эвристике, ведущие к локально оптимальному решению.

[20] Переоснащение в основном происходит, когда сложность модели выше, чем сложность данных. это означает, что модель уже зафиксировала общие закономерности, а также зафиксировала шумы. Недообучение происходит, когда сложность модели ниже сложности данных. Это означает, что эта модель не может фиксировать данные даже об общих шаблонах (сигналах). Например. https://medium.com/geekculture/what-is-overfitting-and-underfitting-in-machine-learning-8907eea8a6c4

[21] Богумил Каминьски, Михал Якубчик и Пшемыслав Шуфель. Структура для анализа чувствительности деревьев решений. Central European Journal of Operations Research, 26(1):135–159, май 2017 г. DOI:10.1007/s10100-017-0479-6. URL-адрес https://doi.org/10.1007/s10100-017-0479-6 .

[22] Байгалтуги Санджаа и Эрдэнэбат Чулуун. Обнаружение вредоносного ПО с помощью линейного SVM. In Ifost, том 2, страницы 136–138, 2013 г. doi:10.1109/IFOST.2013.6616872; Мин Ян, Синшу Чен, Юнган Луо и Ханг Чжан. Модель обнаружения вредоносных программ для Android на основе DT-SVM. Сети безопасности и связи, 2020:1–11, декабрь 2020 г. DOI: 10.1155/2020/8841233. URL-адрес https://doi.org/10.1155/2020/8841233 ; Кинан Ганем, Франсиско Х. Апарисио-Наварро, Константинос Г. Кириакопулос, Сангарапиллай Ламботаран и Джонатон А. Чемберс. Машина вектора поддержки для обнаружения сетевых вторжений и кибератак. В 2017 году Sensor Signal Processing for Defense Conference (SSPD), страницы 1–5, 2017. DOI: 10.1109/SSPD.2017.8233268.

SVM трудно интерпретировать, а это означает, что может быть трудно понять, как алгоритм пришел к своему решению (модель черного ящика). Кроме того, SVM имеют ограниченную масштабируемость и сильно зависят от выбора ядра. Другие проблемы, с которыми сталкиваются SVM, включают чувствительность к выбросам в данных, которые могут оказать существенное влияние на местоположение и ориентацию границы решения, а также трудности с классификацией набора данных в несколько классов, когда некоторые методы, такие как один против одного или один против всех может потребовать значительных вычислительных ресурсов и времени.

1.1.3 Наивный байесовский классификатор (NB)

NB — это универсальный и эффективный алгоритм машинного обучения, который часто используется в кибербезопасности. Его можно использовать для решения задач классификации в задачах кибербезопасности путем принятия статистической теории, в частности теоремы Байеса, для расчета вероятности класса, когда все функции заданы в качестве входных данных.[23].

Самым большим преимуществом NB является то, что он может работать с очень небольшими наборами данных. Это один из самых популярных алгоритмов для**фильтрация спама**[24]**, обнаружение вредоносных программ и вторжений**. Кроме того, он относительно прост в реализации и часто используется в качестве классификатора.

NB может эффективно работать даже в средах с плохими данными. Если набор данных недоступен, его все равно можно использовать в качестве алгоритма классификации. Кроме того, это надежный метод для изолированных шумовых точек.[25]. Однако NB очень склонен к переоснащению.

1.1.4 Кластеризация K-средних (кластеризация)

K-means — это популярный неконтролируемый тип алгоритма машинного обучения, используемый для кластеризации точек данных в группы на основе сходства. Кластеризация считается важной концепцией, помогающей найти структуру или закономерность в наборе неизвестных данных. Алгоритмы кластеризации, такие как K-средние, предназначены для обработки данных и обнаружения кластеров (точек данных, которые можно сгруппировать), когда они присутствуют в наборе данных. Такие кластеры можно использовать для извлечения полезной информации и потенциально для помощи в**выявление вторжений, кибератак и вредоносных программ**[26].

Известным ограничением K-средних является то, что они предполагают, что все кластеры имеют одинаковые размеры и дисперсии.[27]. Другое ограничение состоит в том, что алгоритм ограничен линейными границами данных.

1.1.5 Скрытая марковская модель (HMM)

HMM работает с распределением вероятностей по последовательностям наблюдений. HMM обычно используется в статистическом распознавании образов, где временная структура

[23] Сауараб Мукерджи и Нилам Шарма. Обнаружение вторжений с использованием наивного байесовского классификатора с уменьшением количества признаков. Procedia Technology, 4:119–128, 2012. DOI: 10.1016/j.protcy.2012.05.017. URL-адрес https://doi.org/10.1016/j.protcy.2012.05.017

[24] А. Сумитра, А. Ашифа, С. Харини и Н. Кумаресан, Наивный байесовский алгоритм, основанный на вероятности, для классификации спама по электронной почте. В 2022 г. Международная конференция по компьютерным коммуникациям и информатике (ICCCI), DOI: 10.1109/ICCCI54379.2022.9740792

[25] «Изолированная шумовая точка» имеет характеристики или значения, которые сильно отличаются от большинства точек. Поскольку таких точек по определению очень мало, их значения играют очень малую роль в условной вероятности по всем точкам.

[26] Анджли Чанана, Сурджит Сингх и К.К. Паливал. Обнаружение вредоносного ПО с использованием k-средних, оптимизированных для ga, и хм. Международная конференция по вычислительной технике, связи и автоматизации (ICCCA) 2017 г., стр. 355–362, 2017 г. DOI: 10.1109/CCAA.2017.8229842.

[27] https://hackr.io/blog/k-means-clustering, последний доступ март 2022.

Особенно важно[28]для классификации. Это мощный инструмент для обнаружения слабых сигналов. К сожалению, обучающие данные должны очень хорошо представлять проблему и быть высокого качества, чтобы оптимально принять решение и изучить количество параметров HMM. HMM можно использовать в области кибербезопасности для решения нескольких задач, а именно:**обнаружения вторжений**[29].

1.1.6Генетические алгоритмы (ГА)

GA — это эвристический алгоритм поиска, используемый для решения задач поиска и оптимизации. Этот алгоритм является подмножеством эволюционных алгоритмов[30]используется в вычислениях. ГА использует концепцию генетики и естественного отбора для решения проблем.[31].

Решения на основе GA обычно используются в задачах оптимизации и поиска. Системы на основе GA использовались в различных приложениях кибербезопасности, в том числе**обнаружение спама и вторжений**[32] [33].

Одной из многообещающих областей исследований является использование биовычислений в оборонных целях, где методы предотвращения и борьбы с хищниками могут быть адаптированы к приложениям кибербезопасности.[34]. Несколько подходов, основанных на искусственных иммунных системах для **обнаружение злоумышленника**[35]можно найти в литературе.

1,2НЕЙРОННЫЕ СЕТИ

1.2.1Искусственные нейронные сети (ИНС)

ИНС состоят из узлов, вдохновленных структурой человеческого мозга. По умолчанию они состоят из трех слоев, т. е. входного слоя, скрытого слоя и выходного слоя, хотя в зависимости от сложности задачи могут быть добавлены дополнительные скрытые слои. ИНС часто называют универсальными аппроксиматорами, потому что в процессе обучения выходные данные контролируются таким образом, чтобы свести к минимуму ошибку между желаемым и фактическим выходными данными.[36].

[28]Ахмед Хуссен Абдельазиз, Штеффен Зейлер и Доротея Колосса. Изучение весов динамических потоков для связанного аудиовизуального распознавания речи на основе hmm. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(5):863–876, 2015. DOI:10.1109/TASLP.2015.2409785.

[29]Е Ду, Хуэйцян Ван и Юнган Пан. HMM для обнаружения аномальных вторжений. В области вычислительной техники и информатики, страницы 692–697. Springer Berlin Heidelberg, 2004. DOI: 10.1007/978-3-540-30497-5_108. URL-адресhttps://doi.org/10.1007/978-3-540-30497-5_108

[30]https://www.techtarget.com/whatis/definition/эволюционный-алгоритм , последний доступ март 2022 г.

[31]Жорж Р. Харик, Фернандо Г. Лобо и Кумара Шастри. Изучение сцепления с помощью вероятностного моделирования в расширенном компактном генетическом алгоритме (ECGA). В «Масштабируемой оптимизации с помощью вероятностного моделирования», стр. 39–61. Springer Berlin Heidelberg, 2006. doi: 10.1007/978-3-540-34954-9_3. URL-адреshttps://doi.org/10.1007/978-3-540-34954-9_3

[32]Анас Аррам, Хишам Муса и Анзида Зайнал. Обнаружение спама с использованием гибридной искусственной нейронной сети и генетических алгоритмов. 13-я Международная конференция по проектированию и приложениям интеллектуальных систем, 2013 г., стр. 336–340, 2013 г. DOI: 10.1109/ISDA.2013.6920760. Хоссейн Гари и Хамид Хоссейнванд. Новые идентификаторы выбора признаков, основанные на генетическом алгоритме и SVM. В 2016 г. 8-й Международный симпозиум по телекоммуникациям (IST), страницы 139–144, 2016 г. DOI: 10.1109 / ISTEL.2016.7881798.

[33]Ин Чжан, Пейсон Ли и Синьхэн Ван. Обнаружение вторжений для IoT на основе улучшенного генетического алгоритма и сети глубокого доверия. IEEE Access, 7:31711–31722, 2019 г. DOI:10.1109/ACCESS.2019.2903723.

[34]Сияха Н. Мтунзи, Эльхадж Бенхелифа, Томаш Босаковский и Салим Харири. Биологический подход к кибербезопасности. В машинном обучении для компьютерной и кибербезопасности, страницы 75–104. CRC Press, февраль 2019 г. DOI: 10.1201/9780429504044-4. URL-адреshttps://doi.org/10.1201/9780429504044-4

[35]Ин Чжан, Пейсон Ли и Синьхэн Ван. Обнаружение вторжений для IoT на основе улучшенного генетического алгоритма и сети глубокого доверия. IEEE Access, 7:31711–31722, 2019 г. DOI:10.1109/ACCESS.2019.2903723

[36]Дэвид Э. Румелхарт, Джеффри Э. Хинтон и Рональд Дж. Уильямс. Обучение представлениям путем обратного распространения ошибок. Nature, 323 (6088): 533–536, октябрь 1986 г. DOI: 10.1038/323533a0. URL-адреshttps://doi.org/10.1038/323533a0

ИНС использовались во многих областях кибербезопасности, таких как**обнаружение мошенничества, вторжений, спама и вредоносных программ**[37]. В целом, многослойные ИНС склонны к переоснащению, если сеть слишком велика. В то же время построение модели может занять очень много времени, а тестирование может быть очень быстрым. Однако они чувствительны к шуму в обучающих данных и не обрабатывают отсутствующие атрибуты.

1.2.2Сверточные нейронные сети (CNN)

CNN — это типы нейронных сетей, специально разработанные для задач обработки изображений, таких как распознавание и классификация объектов. CNN используют подходы на основе глубокого обучения (DL), которые могут эффективно моделировать очень большие наборы данных. CNN используют серию сверточных слоев и слоев объединения для извлечения все более абстрактных функций из входных изображений. Сверточные слои применяют фильтры к входному изображению для выявления закономерностей и признаков, в то время как объединяющие слои выполняют выборку карт признаков, чтобы уменьшить вычислительную сложность сети. Выходные данные последнего слоя CNN затем передаются на полносвязный слой, который выполняет задачу классификации. Их успех последовал за огромным прорывом в графических процессорах со значительной мощностью обработки данных. Однако,

В кибербезопасности CNN использовались для**задачи обнаружения вторжений**[38].

1.2.3Рекуррентные нейронные сети (RNN)

RNN — это тип нейронной сети, который особенно хорошо подходит для последовательных данных, таких как временные ряды или текстовые данные. RNN предназначены для обработки входных данных переменной длины, обрабатывая один элемент за раз, а также сохраняя внутреннее состояние, которое суммирует предыдущие входные данные. Это внутреннее состояние передается от одного временного шага к другому, что позволяет сети фиксировать зависимости и закономерности, существующие во времени.

RNN обычно используются для**обнаружения вторжений**в наборах данных KDD99 (см. раздел 2.4) с высоким уровнем точности[39].

1.2.4Автоэнкодеры

Автоэнкодеры — это тип неконтролируемой технологии DNN, которая уменьшает размерность исходного входного пространства для устранения шума и ненужных функций.

Автоэнкодеры состоят из двух частей: кодировщика, который отображает входные данные в низкоразмерное представление, и декодера, который отображает закодированное представление обратно в исходное входное пространство. Во время обучения сеть учится минимизировать разницу между входными данными и реконструированным выходом, регулируя веса кодировщика и декодера.

---

[37]Прити Мишра, Виджай Варадхараджан, Удай Тупакула и Эммануэль С. Пилли. Подробное исследование и анализ использования методов машинного обучения для обнаружения вторжений. IEEE Communications Surveys Tutorials, 21(1):686–728, 2019. DOI:10.1109/COMST.2018.2847722.

[38]Диляра Гюмюшбаш, Тулай Йылдырым, Анджело Дженовезе и Фабио Скотти. Всесторонний обзор баз данных и методов глубокого обучения для систем кибербезопасности и обнаружения вторжений. Журнал IEEE Systems, страницы 1–15, 2020 г. DOI: 10.1109/JSYST.2020.2992966.

[39]То же, что и 38.

Помимо приложений сжатия, автокодировщики эффективны при обнаружении аномалий путем сравнения потерь реконструкции между известными и новыми данными и поэтому очень интересны для приложений кибербезопасности.[40]в том числе**обнаружение атак нулевого дня**[41].

1.2.5Сиамские нейронные сети (SNN)

SNN — это классификаторы подобия, которые используют отличительные признаки для обобщения неизвестных категорий в заданном распределении, например, для извлечения признаков или определения того, принадлежат ли две категории к одному и тому же классу, или для разделения данных на классы, которые модель никогда раньше не «видела». Этот тип нейронной сети можно использовать для задач классификации.

Архитектура SNN более сложна, и может потребоваться добавление дополнительных механизмов извлечения признаков ML. По сравнению с обычными нейронными сетями для обучения требуется больше времени, так как для построения точной модели требуется большое количество комбинаций обучающих выборок, необходимых для механизма обучения SNN.[42].

Сиамские нейронные сети имеют множество приложений для распознавания изображений, а также для обучения с самоконтролем (SSL).[43]. SNN могут быть эффективными для количественной оценки того, насколько похожи или различаются два входа, для облегчения задач ML, например классификации, обнаружения аномалий и т. д.

В кибербезопасности SSN применялись к таким задачам, как**обнаружение вредоносных программ**и **обнаружения вторжений**, изучая характерные представления входных данных, которые фиксируют соответствующие характеристики вредоносных программ или аномального сетевого трафика.

1.2.6Методы ансамбля

Методы ансамбля машинного обучения — это методы, которые объединяют несколько моделей машинного обучения для повышения их точности и стабильности. Методы ансамбля популярны, потому что они могут повысить точность отдельных моделей, уменьшить переоснащение и повысить надежность. Несмотря на то, что в большей части существующей литературы используются системы, основанные на одном инструменте на основе машинного обучения, существует несколько сценариев, в которых применялись ансамблевые методы.[44].

Причина использования ансамблевых моделей заключается в объединении типов моделей, демонстрирующих многообещающую производительность в различных случаях (например, типы атак, сети и т. д.). Такой

[40]Темесгуэн Мессей Кебеде, Уботи Джаней-Бунджоу, Барат Нараянан Нараянан, Анка Ралеску и Дэвид Капп. Классификация вредоносных программ с использованием автокодировщиков на основе архитектуры глубокого обучения и ее применение к набору данных Microsoft по классификации вредоносных программ (большой 2015 г.). В 2017 г. IEEE National Aerospace and Electronics Conference (NAECON), стр. 70–75, 2017 г. DOI: 10.1109/NAECON.2017.8268747

[41]Ханан Хинди, Роберт Аткинсон, Христос Тахтацис, Жан-Ноэль Колин, Итан Бейн и Ксавьер Беллекенс. Использование методов глубокого обучения для эффективного обнаружения атак нулевого дня. Electronics, 9(10):1684, октябрь 2020 г. DOI:10.3390/electronics9101684. URL https://doi.org/10.3390/electronics9101684

[42]https://medium.com/codex/vol-2a-siamese-neural-networks-6df66d33180e , последний доступ в марте 2022 года.

[43]Аттаулла Сахито, Эйбе Франк и Бернхард Пфарингер, Полууправляемое обучение с использованием сиамских сетей, Springer International Publishing, 2019 г., DOI: 10.1007/978-3-030-35288-2_47

[44]Дипанкар Дасгупта, Захид Ахтар и Саджиб Сен. Машинное обучение в кибербезопасности: всесторонний обзор. Журнал оборонного моделирования и моделирования: приложения, методология, технологии, сентябрь 2020 г. DOI: 10.1177/1548512920951275. URL-адрес https://doi.org/10.1177/1548512920951275

методы были использованы для нескольких приложений, включая**обнаружение вредоносных программ**[45], **обнаружения вторжений**[46], и т. д.

1,3АКТУАЛЬНОСТЬ ПОДХОДОВ НА ОСНОВЕ ГЛУБОКОГО ОБУЧЕНИЯ (ГО)

В последние годы была проделана огромная работа по разработке решений на основе DL для использования в приложениях кибербезопасности, включая**защита и оборона**[47]. Решения на основе машинного обучения смогли предложить превосходную производительность, которая часто превосходит традиционное машинное обучение, работающее с большими наборами данных, и в настоящее время представляют собой передовые технологии во многих областях.

Однако они имеют некоторые важные ограничения, которые следует учитывать при разработке и внедрении. Во-первых, это доступность и надежность наборов данных, т.е. потребность в больших наборах данных, содержащих данные высокого качества.[48]. Подавляющее большинство литературы посвящено улучшению современных характеристик, в то время как надежность наборов данных практически не рассматривается.

В современной литературе предлагаются критерии надежности.[49] [50]такие как: a) разнообразие атак, б) анонимность, в) доступные протоколы, г) полный захват (с полезной нагрузкой), д) полное взаимодействие, е) полная конфигурация сети, ж) полный трафик, з) набор функций, и) неоднородность (весь сетевой трафик и системные журналы), j) правильная маркировка и k) метаданные (полная документация по сбору данных).

К сожалению, существующие критерии надежности сосредоточены на обнаружении вторжений, в то время как аналогичные требования для других приложений кибербезопасности еще предстоит решить.

Вторым важным аспектом, который следует учитывать в этом конкретном контексте, является тот факт, что злоумышленники постоянно разрабатывают новые типы атак в обход существующих систем безопасности. Эта конкретная проблема относится к области обучения в нестационарной среде и обычно называется дрейфом понятий.[51].

Кроме того, изучаемая система может подвергнуться сдвигу в своих номинальных рабочих условиях (изменение во времени), когда номинальная модель нуждается в обновлении.[52]. Такой

[45]Санджай Кумар, Ари Вийникайнен и Тимо Хамалайнен. Оценка ансамблевых методов машинного обучения при обнаружении мобильных угроз. В 2017 г. 12-я Международная конференция по интернет-технологиям и защищенным транзакциям (ICITST), страницы 261–268, 2017 г. DOI: 10.23919 / ICITST.2017.8356396.

[46]Анна Магдалена Косек и Оливер Герке. Обнаружение аномалий на основе ансамблевой регрессионной модели для обнаружения киберфизических вторжений в интеллектуальные сети. На конференции IEEE по электроэнергетике и энергетике (EPEC) 2016 г., страницы 1–7, 2016 г. DOI: 10.1109/EPEC.2016.7771704.

[47]Диляра Гюмюшбаш, Тулай Йылдырым, Анджело Дженовезе и Фабио Скотти. Всесторонний обзор баз данных и методов глубокого обучения для систем кибербезопасности и обнаружения вторжений. Журнал IEEE Systems, страницы 1–15, 2020 г. DOI: 10.1109/JSYST.2020.2992966.

[48]Самира Пуянфар, Саад Садик, Илинь Ян, Хайман Тиан, Юдонг Тао, Мария Преса Рейес, Мей-Линг Шью, Шу-Чинг Чен и С.С. Айенгар. Опрос по глубокому обучению. ACM Computing Surveys, 51(5):1–36, январь 2019 г. DOI:10.1145/3234150. URL-адрес https://doi.org/10.1145/3234150

[49]Иман Шарафалдин, Араш Хабиби Лашкари и Али А. Горбани. На пути к созданию нового набора данных для обнаружения вторжений и характеристик трафика вторжений. В материалах 4-й Международной конференции по безопасности и конфиденциальности информационных систем. SCITEPRESS - Публикации по науке и технологиям, 2018. DOI: 10.5220/0006639801080116. URL-адрес https://doi.org/10.5220/0006639801080116

[50]Амирхоссейн Гариб, Иман Шарафалдин, Араш Хабиби Лашкари и Али А. Горбани. Платформа оценки для набора данных обнаружения вторжений. В 2016 г. Международная конференция по информатике и безопасности (ICISS), страницы 1–6, 2016 г. DOI: 10.1109/ICISSEC.2016.7885840.

[51]Грегори Дицлер, Мануэль Ровери, Чезаре Алиппи и Роби Поликар. Обучение в нестационарных средах: обзор. Журнал IEEE Computational Intelligence, 10(4):12–25, 2015 г. doi:10.1109/MCI.2015.2471196.

[52]Чезаре Алиппи, Ставрос Нталампирас и Мануэль Ровери. Безмодельное обнаружение и изоляция неисправностей в крупномасштабных киберфизических системах. IEEE Transactions on Emerging Topics in Computational Intelligence, 1(1):61–71, 2017. DOI:10.1109/TETCI.2016.2641452.

изменения должны быть своевременно обнаружены и правильно идентифицированы, чтобы механизмы защиты могли надежно функционировать.

Таким образом, обучение в нестационарных средах в области кибербезопасности остается открытой темой, и для эффективных и современных моделей безопасности требуются новые методы, способные обнаруживать изменения стационарности и соответствующим образом реагировать на них.

1,4 ОБЫЧНО ИСПОЛЬЗУЕМЫЕ НАБОРЫ ДАННЫХ О КИБЕРБЕЗОПАСНОСТИ

Вышеупомянутые инструменты и методологии на основе ML зависят от доступности данных, т.е. для создания таких решений необходимы наборы данных, наборы потенциально разнородных типов информации, атрибутов или функций. Анализируя доступные данные и обнаруживая существующие шаблоны, можно получить представление о номинальном состоянии, а также о кибератаках.

В таблице 2 представлены несколько наборов данных, широко используемых научно-исследовательским сообществом для разработки инструментов и методологий на основе машинного обучения для приложений кибербезопасности, таких как обнаружение вторжений, анализ вредоносных программ, моделирование трафика ботнетов или фильтрация спама. Список, представленный ниже, не является исчерпывающим[53], так как его цель — представить некоторые из наиболее часто используемых наборов данных и разнообразные сценарии их применения.

Таблица 2: Широко используемые наборы данных кибербезопасности

Набор данных	Описание
Кубок КДД 99[54]	Это, вероятно, наиболее широко используемый набор данных, содержащий 41 признак для обнаружения аномалий. Он был разработан и опубликован Агентством перспективных оборонных исследовательских проектов (DARPA). Он включает в себя полные пакетные данные и четыре категории атак, таких как DoS, R2L «удаленный-локальный», «пользователь-удаленный» (U2R) и зондирование. Он широко использовал подходы к обнаружению вторжений.
ДЕФКОН[55]	Этот набор данных включает в себя различные атаки для помощи в проведении соревнований по моделированию вторжений, проводимых ежегодно.
ГТУ-13[56]	Сюда входят 13 различных ситуаций реального трафика ботнета с учетом характеристик как обычного, так и фонового трафика.

[53] Поскольку новые наборы данных публикуются в быстром темпе, читатель может ознакомиться с исчерпывающим списком связанных наборов данных: Камран Шаукат, Сухуай Луо, Виджай Варадхараджан, Ибрагим А. Хамид и Мин Сюй. Обзор методов машинного обучения для обеспечения кибербезопасности за последнее десятилетие. IEEE Access, 8:222310–222354, 2020. DOI:10.1109/access.2020.3041951. URL-адрес https://doi.org/10.1109/access.2020.3041951 ; Диляра Гюмюшбаш, Тулай Йылдырым, Анджело Дженовезе и Фабио Скотти. Всесторонний обзор баз данных и методов глубокого обучения для систем кибербезопасности и обнаружения вторжений. Журнал IEEE Systems, страницы 1–15, 2020 г. DOI: 10.1109/JSYST.2020.2992966; и Икбал Х. Саркер, А.С.М. Кайес, Шахриар Бадша, Хамед Алкахтани, Пол Уоттерс и Алекс Нг. Наука о данных кибербезопасности: обзор с точки зрения машинного обучения. Журнал больших данных, 7 (1), июль 2020 г. DOI: 10.1186/s40537-020-00318-5. URL-адрес https://doi.org/10.1186/s40537-020-00318-5 .

[54] RP Lippmann, DJ Fried, I. Graf, JW Haines, KR Kendall, D. McClung, D. Weber, SE Webster, D. Wyschogrod, RK Cunningham и MA Zissman. Оценка систем обнаружения вторжений: оценка автономного обнаружения вторжений DARPA 1998 года. В материалах конференции и выставки DARPA по информационной живучести. DISCEX'00, том 2, страницы 12–26, том 2, 2000 г. DOI: 10.1109/DISCEX.2000.821506.

[55] Али Ширави, Хади Ширави, Махбод Таваллаи и Али А. Горбани. На пути к разработке систематического подхода к созданию эталонных наборов данных для обнаружения вторжений. Компьютеры и безопасность, 31(3):357–374, май 2012 г. DOI: 10.1016/j.cose.2011.12.012. URL-адрес https://doi.org/10.1016/j.cose.2011.12.012

[56] С. Гарсия, М. Гриль, Дж. Стиборек и А. Зунино. Эмпирическое сравнение методов обнаружения ботнетов. Компьютеры и безопасность, 45:100–123, сентябрь 2014 г. DOI:10.1016/j.cose.2014.05.011. URL-адрес https://doi.org/10.1016/j.cose.2014.05.011

Спам база5[57]	Это коллекция электронных писем с несколькими тысячами экземпляров, упрощающая классификацию электронной почты.
SMS-спам Коллекция6	Это включает в себя широкий спектр SMS-сообщений, помеченных как спам или не спам.
CICIDS20177	Он состоит из данных о трафике, записанных в Канадском институте кибербезопасности, и предоставляет полные пакетные данные и необработанные файлы PCAP. Интересно, что рассмотрено несколько типов атак.
CICAndMal20178[58]	Сюда входят заслуживающие доверия и вредоносные приложения, удобно разделенные на четыре класса, т. е. вредоносные программы, SMS-вредоносные программы, программы-вымогатели и рекламное ПО. Таким образом, он может облегчить идентификацию вредоносных приложений Android.
Проверка Android[59]	Он состоит из данных, характеризующих отношения, существующие между различными приложениями, организованными в ложных братьев и сестер, братьев и сестер, кузенов и сводных братьев и сестер.
Набор данных IoT-23[60]	Это набор данных, содержащий вредоносный и безопасный сетевой трафик IoT.

[57] Тьяго А. Алмейда, Хосе Мария Г. Идальго и Акебо Ямаками. Вклад в изучение фильтрации SMS-спама. В материалах 11-го симпозиума ACM по разработке документов - DocEng '11. ACM Press, 2011. DOI: 10.1145/2034691.2034742. URL-адрес https://doi.org/10.1145/2034691.2034742

[58] Эсра Чалик Баязит, Озгур Корай Сахингоз и Букет Доган. Обнаружение вредоносных программ в системах Android с традиционными моделями машинного обучения: опрос. В 2020 г. Международный конгресс по взаимодействию человека и компьютера, оптимизации и роботизированным приложениям (HORA), страницы 1–8, 2020 г. DOI: 10.1109/HORA49412.2020.9152840.

[59] Уго Гонсалес, Наталья Стаханова и Али А. Горбани. DroidKin: Облегченное обнаружение сходства приложений для Android. В конспектах лекций Института компьютерных наук, социальной информатики и телекоммуникаций, страницы 436–453. Springer International Publishing, 2015. DOI: 10.1007/978-3-319-23829-6_30. URL-адрес https://doi.org/10.1007/978-3-319-23829-6_30

[60] Себастьян Гарсия, Агустин Пармизано и Мария Хосе Эркиага, https://doi.org/10.5281/ZENODO.4743746 . Доступные наборы данных в https://www.stratosphereips.org/datasets-iot23

ИИ В КИБЕРБЕЗОПАСНОСТИ

В этом разделе обобщается текущее состояние дел в области основных применений «традиционных», давних и новых приложений ИИ (системы глубокого обучения), инструментов и методов в кибербезопасности, рассматривая обе стороны использования ИИ в контексте требований кибербезопасности, т. е. злонамеренных и добродетельных. Ниже приведен неполный список способов использования ИИ в кибербезопасности:

а. Киберпреступники, использующие ИИ для повышения своей эффективности;

б. Механизмы безопасности, включающие ИИ для обнаружения, идентификации и смягчения последствий компрометации;

в. Использование ИИ для использования уязвимостей в существующих инструментах и методологиях ИИ и других инструментов, например, состязательных атак.[61];

д. Использование ИИ при проектировании системы для защиты существующих инструментов и методологий ИИ и не связанных с ИИ (защита, создаваемая при проектировании системы).

В первых двух случаях ИИ используется как инструмент (злоумышленник может использовать ИИ для разработки атаки), а в последних двух случаях ИИ является фактической целью (атака может быть нацелена на систему на основе ИИ). Несмотря на то, что защитные механизмы на основе ИИ устраняют широкий спектр уязвимостей, они сами могут быть точками атаки. Злоумышленники используют ИИ не только для организации различных киберугроз, но и для атак на механизмы защиты на основе ИИ, используя существующие уязвимости. В таблице ниже показано использование методов ИИ в функциях кибербезопасности.

Таблица 3: Методы ИИ в функциях кибербезопасности (источник: авторы)

Функция безопасности\ИИ	ДТ	SVM	NB	K-означает	ХМ	ГАЗ	АННА	Си-Эн-Эн	РНН	Кодировщики	SNN
Обнаружения вторжений	Икс	Икс	Икс	Икс	Икс	Икс	Икс	Икс	Икс		Икс
Обнаружение вредоносных программ	Икс	Икс	Икс	Икс				Икс	Икс		
Уязвимость оценка	Икс										
Фильтрация спама			Икс								
Обнаружение аномалий					Икс					Икс	
Классификация вредоносных программ						Икс	Икс				Икс
Обнаружение фишинга							Икс				
Анализ трафика								Икс	Икс		
Сжатие данных										Икс	
Извлечение признаков										Икс	

[61]Например, в GAN с обучающим набором метод учится генерировать новые данные с той же статистикой, что и обучающий набор. GAN, обученный работе с фотографиями, может генерировать новые фотографии, которые, по крайней мере, внешне выглядят аутентичными для людей-наблюдателей, обладая многими реалистичными характеристиками.

1,5ПРИМЕРЫ ИСПОЛЬЗОВАНИЯ

Инструменты и методологии на основе ИИ могут использоваться для обнаружения и идентификации кибератак и смягчения их последствий. Такие инструменты могут обеспечить удовлетворительную производительность при низких затратах и в режиме реального времени. Существует широкий спектр методов и возможностей защиты, которые могут быть реализованы с помощью ИИ.[62]. Механизмы защиты на основе ИИ все чаще применяются в области кибербезопасности, например, в области безопасности сети и данных, защиты конечных точек, надежности доступа и т. д.[63].

После описания ключевых концепций и ключевых функций в главе 2 в следующих разделах приведены виды ИИ (задача, метод, метод), которые используются в каждой конкретной функции или операции кибербезопасности, например, предотвращение атак, обнаружение угроз и вторжений. , реагирование и восстановление после кибератак. Для этого мы рассмотрим их, используя концепции предотвращения, обнаружения, реагирования и восстановления.

1.5.1Профилактика

ИИ можно использовать для оценки уязвимостей в компьютерных системах и сетях. Алгоритмы машинного обучения часто используются при анализе данных из нескольких источников, таких как сканеры, журналы безопасности и системы управления исправлениями, для выявления уязвимостей и определения приоритетности усилий по исправлению.

Фаззеры на основе глубокого обучения[64]в настоящее время считаются наиболее перспективным путем обнаружения уязвимостей по сравнению с традиционным ML[65]. Обучение с подкреплением может искать уязвимости в компьютерной сети быстрее, чем традиционные инструменты пен-тестирования.

Таблица 4:Приложения ИИ для предотвращения атак

Задача	Пример методов, техник, подходов ИИ
фаззеры	DL
Реп-тестирование	Обучение с подкреплением
Оценка уязвимости	НЛП, традиционное машинное обучение

Источник: авторская адаптация по роману Мики и Эштона (2021 г.).

ML также может быть полезен при оценке риска в сети, например, для определения серьезности уязвимости. ИИ можно использовать для управления идентификацией пользователей и доступом к компьютерным системам и приложениям. Алгоритмы машинного обучения можно использовать для анализа поведения пользователей и

[62]В частности анализы Колумба, Луи. nd «Защита вашей компании при продаже ваших привилегированных учетных данных». Форбс. По состоянию на 23 августа 2021 г.https://www.forbes.com/sites/louiscolumbus/2018/08/21/protecting-yourcompany-when-your-privated-credentials-are-for-sale/ ; - Дилмегани Джем. 2021. «Аналитика безопасности: полное руководство [обновление 2021]». 20 августа 2018 г. https://research.aimultiple.com/security-analytics/ ; Капджемини. 2019. «Новое изобретение кибербезопасности с помощью искусственного интеллекта: новый рубеж цифровой безопасности». ИИ-в-Cybersecurity_Report_20190711_V06.pdf (capgemini.com) и — Джонс, Тим. 2019. IBM Developer «Рассмотрите ИИ и безопасность и изучите использование алгоритмов машинного обучения для обнаружения угроз и управления ими». (блог). 19 августа 2019 г. https://developer.ibm.com/articles/ai-и-безопасность /.

[63]Группа Капджемини. Новое изобретение кибербезопасности с помощью искусственного интеллекта: новый рубеж цифровой безопасности. Технический отчет, Исследовательский институт Capgemini, 01 2021. URLhttps://www.capgemini.com/research/reinventingcybersecurity-with-artificial-intelligence/

[64]Примером может служить программа NeuFuzz, основанная на глубоком обучении. Microsoft также изучила использование глубокого обучения для фаззеров, см., например,http://arxiv.org/abs/1711.04596

[65]Несколько команд в соревнованиях Cyber Grand Challenge, спонсируемых DARPA, пытались использовать машинное обучение для выявления уязвимостей программного обеспечения.

выявлять подозрительные действия, такие как попытки захвата учетной записи или попытки несанкционированного доступа.

1.5.2Обнаружение

Большинство «традиционных» приложений машинного обучения почти полностью попадают на этап обнаружения, то есть для обнаружения спама, обнаружения вторжений и обнаружения вредоносных программ, а также обнаружения атак. Большое количество существующих работ посвящено обнаружению спама в компьютерных сетях. Спам, рассылаемый по электронной почте, потребляет соответствующие ресурсы (например, полосу пропускания, хранилище и т. д.), что напрямую снижает пропускную способность и эффективность систем и сетей.

Еще одна проблема, которой активно занимается исследовательское сообщество, — это обнаружение вредоносных программ и вторжений.

Как правило, защитные механизмы предназначены для защиты от конкретных типов атак, таких как распределенный отказ в обслуживании (DDoS), зондирующие атаки.[66], дистанционные атаки на локальные (R2L)[67], несанкционированный доступ к локальному суперпользователю (U2R)[68], хостовые, сетевые, программы-вымогатели и т. д. Для борьбы с этими конкретными типами атак использовалось множество многообещающих решений на основе машинного обучения, включая контролируемые и неконтролируемые подходы.[69] [70]. Кроме того, для решения проблем обнаружения вторжений использовались биологические алгоритмы.[71] [72].

В области обнаружения вредоносных программ[73] [74] [75], мл[76]был использован для выбора соответствующих функций, выявляющих наличие вредоносных программ, а также методов обнаружения аномалий или отклонений.

Различные методы машинного обучения, такие как SVM и DT, также использовались для обнаружения кибератак, но большинство из них не могут обнаружить новые типы атак, то есть атаки, которые не являются частью набора данных, используемого при обучении. В этом случае решения должны аппроксимировать распределение доступных данных, чтобы можно было обнаружить выборки, не принадлежащие распределению. Для этого могут быть использованы адаптированные версии существующих традиционных (одноклассовые SVM, HMM и др.) и основанных на NN (ANN, CNN и др.) решений.

[66]При зондирующих атаках злоумышленник сканирует сеть для сбора информации о компьютерах с целью выявления уязвимостей.

[67]Известно, что злоумышленники запускают атаки удаленного доступа к локальным (R2L) для получения несанкционированного доступа к компьютерам-жертвам в сети.

[68]Атака, при которой злоумышленник использует обычную учетную запись для входа в систему жертвы и пытается получить привилегии root/администратора, используя некоторую уязвимость.

[69]Камран Шаукат, Сухуай Луо, Виджай Варадхараджан, Ибрагим А. Хамид и Мин Сюй. Обзор методов машинного обучения для обеспечения кибербезопасности за последнее десятилетие. IEEE Access, 8:222310–222354, 2020. doi:10.1109/access.2020.3041951. URL-адрес https://doi.org/10.1109/access.2020.3041951

[70]Статья Обзор методов машинного обучения для кибербезопасности за последнее десятилетие https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9277523 представляет собой опрос, в котором обсуждается выполнение различных исследовательских работ.

[71]Анас Аррам, Хишам Муса и Анзида Зайнал. Обнаружение спама с использованием гибридной искусственной нейронной сети и генетического алгоритма. В 2013 г. 13-я Международная конференция по проектированию и приложениям интеллектуальных систем, страницы 336–340, 2013 г. doi: 10.1109 / ISDA.2013.6920760

[72]Хоссейн Гари и Хамид Хоссейнванд. Новые идентификаторы выбора признаков на основе генетического алгоритма и SVM. В 2016 г. 8-й Международный симпозиум по телекоммуникациям (IST), стр. 139–144, 2016 г. doi: 10.1109 / ISTEL.2016.7881798.

[73]Хамед Хаддад Паджоу, Али Дегантанха, Рауф Хаями и Ким-Кванг Рэймонд Чу. Подход на основе глубокой рекуррентной нейронной сети для поиска вредоносных программ в Интернете вещей. Future Generation Computer Systems, 85:88–96, август 2018 г. doi:10.1016/j.future.2018.03.007. URL-адреchttps://doi.org/10.1016/j .будущее.2018.03.007

[74]Темесгуэн Мессей Кебеде, Уботи Джаней-Бунджоу, Барат Нараянан Нараянан, Анка Ралеску и Дэвид Капп. Классификация вредоносных программ с использованием автоэнкодеров на основе архитектуры глубокого обучения и ее применение к набору данных Microsoft по классификации вредоносных программ (большой 2015 г.). В 2017 г. IEEE National Aerospace and Electronics Conference (NAECON), стр. 70–75, 2017 г. doi:10.1109/NAECON.2017.8268747

[75]Эсра Чалик Баязит, Озгур Корай Сахингоз и Букет Доган. Обнаружение вредоносных программ в системах Android с традиционными моделями машинного обучения: опрос. В 2020 г. Международный конгресс по взаимодействию человека с компьютером, оптимизации и роботизированным приложениям (HORA), страницы 1–8, 2020 г. doi: 10.1109/HORA49412.2020.9152840.

[76]См. Micah and Ashton (2021) об исследованиях, посвященных использованию машинного обучения, например, методов HMM и DL.

Кроме того, новые данные должны быть включены в словарь для дальнейшего использования и ручного анализа. В таблице 4 ниже приведены возможные варианты использования методов ИИ для обнаружения угроз и вторжений.

Таблица 5:Приложения ИИ для обнаружения угроз и вторжений (источник: разработка авторов)

Задача	Примеры методов ИИ
Обнаружение спама	СВМ, ДТ
Обнаружения вторжений	Контролируемые и неконтролируемые подходы, био-алгоритмы
Обнаружение вредоносных программ	Стандартные классификаторы ML, DL
Обнаружение атаки	СВМ, ДТ

ЗАЩИТА ИИ

В этом отчете также рассматриваются существующие подходы к более безопасному ИИ, чтобы предотвратить использование ИИ для организации кибератак или для предотвращения атак на механизмы и инструменты на основе ИИ. Сами системы ИИ могут быть уязвимы для угроз из-за собственных уязвимостей или уязвимостей других взаимозависимых механизмов.

1,6ИИ БЕЗОПАСНОСТЬ

Security-by-design — это концепция разработки программного обеспечения, которая подчеркивает важность интеграции принципов безопасности на ранних этапах проектирования и разработки систем и приложений. Это включает рассмотрение рисков и уязвимостей безопасности на каждом этапе разработки, от архитектуры и проектирования до реализации и тестирования. В следующем списке приведены примеры методов обеспечения безопасности, которые могут быть применены к системам ИИ:

- Проведение оценок рисков безопасности и моделирования угроз для выявления потенциальных уязвимостей и векторов атак,
- Использование методов безопасного кодирования и сред разработки программного обеспечения для минимизации риска ошибок и уязвимостей кодирования,
- Внедрение методов безопасной обработки данных для защиты конфиденциальных данных и предотвращения утечки данных,
- Включение тестирования и проверки безопасности в процесс разработки для раннего выявления и устранения проблем безопасности,
- Обеспечение того, чтобы системы ИИ были прозрачными и объяснимыми, чтобы их поведение можно было проверить и проверить.

Концепции безопасности по дизайну, применимые конкретно к системам ИИ, включают:

- Конфиденциальность по дизайну: эта концепция подчеркивает важность учета соображений конфиденциальности и конфиденциальности данных при проектировании и разработке систем ИИ.
- Объяснимость за счет дизайна: эта концепция подчеркивает важность разработки прозрачных и объяснимых систем ИИ, чтобы люди могли понять и проверить их поведение.
- Надежность по дизайну: эта концепция подчеркивает важность разработки систем ИИ, устойчивых к атакам и ошибкам и способных продолжать функционировать даже перед лицом неожиданных входных данных или помех.
- Справедливость по дизайну: эта концепция подчеркивает важность разработки справедливых и беспристрастных систем ИИ, которые не увековечивают и не усиливают существующие в обществе предубеждения или дискриминацию.

1,7КИБЕРАТАКИ С ИСПОЛЬЗОВАНИЕМ ИИ

Поскольку технология ИИ продолжает развиваться, вполне вероятно, что в будущем мы увидим более изощренные и сложные кибератаки с использованием ИИ. Например, генеративно-состязательная сеть (GAN), класс платформ ML, может использоваться для создания «глубоких подделок» путем замены или манипулирования лицами или голосами на изображении или видео.

Алгоритмы на основе искусственного интеллекта также могут подготавливать убедительные фишинговые электронные письма.[77] ориентированы на частных лиц и организации. ИИ также можно использовать для повышения эффективности вредоносных программ.[78], улучшая его способность уклоняться от обнаружения, адаптироваться к изменяющимся условиям, нацеливаться на определенные уязвимости, распространяться и сохраняться в целевых системах. Вредоносное ПО, управляемое искусственным интеллектом, может использовать методы обучения с подкреплением, чтобы совершенствоваться и проводить еще более успешные атаки.

Злоумышленники могут использовать обучающие данные для создания «черного хода» в алгоритме ИИ. Злоумышленники также могут использовать ИИ, чтобы решить, какую уязвимость, скорее всего, стоит использовать. Это всего лишь несколько примеров кибератак с использованием ИИ, которые уже вызывают серьезную озабоченность.

1,8ЗАЩИТА МЕХАНИЗМОВ НА ОСНОВЕ ИИ

Системы ИИ могут быть уязвимы из-за собственных уязвимостей или слабых мест, вызванных другими взаимозависимыми механизмами. Атаки на механизмы на основе ИИ можно организовать по следующим категориям[79](неполный список).

- Атаки, использующие существующие уязвимости в популярных библиотеках программного обеспечения с открытым исходным кодом, например, pytorch, tensorflow и т. д.
- Атаки отравляют обучающие данные. Здесь предполагается, что злоумышленник имеет доступ к обучающим данным и может изменять их и вводить манипуляции, такие как неправильные метки, чтобы система ИИ, обученная на зараженных данных, выполняла обработку и/или прогнозирование в соответствии с интересами злоумышленника.
- Состязательные атаки, при которых обычно атакуемая система ИИ представляет собой глубокую нейронную сеть. Здесь злоумышленник вносит незначительные изменения в тестовые примеры, чтобы изменить предсказание системы ИИ целевым или нецелевым образом, то есть направить предсказание к заданному желаемому классу или к любому классу, отличному от правильного.
- Обратное проектирование обученной модели на основе общедоступных интерфейсов запросов, например, кража модели, инверсия модели и вывод о членстве.

В литературе было предложено несколько подходов для обеспечения безопасности и защиты механизмов на основе ИИ от таких злонамеренных попыток. Эти подходы включают следующее.

[77] https://www.wired.com/story/ai-phishing-emails/ , последний доступ март 2023 г.
[78] Конг Чыонг Тхань и Иван Зелинка. Опрос об искусственном интеллекте в вредоносных программах как угрозах нового поколения. МЕНДЕЛЬ, 25(2):27–34, декабрь 2019 г. doi:10.13164/mendel.2019.2.027. URL-адрес https://doi.org/10.13164/mendel.2019.2.027
[79] Вызовы кибербезопасности искусственного интеллекта ENISA, 2020 г., доступны по адресу https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges/@@download/fullReport

- Защита используемых программных пакетов и проверка достоверности обучающих данных.[80].
- Подходы к противодействию состязательным атакам[81] [82]которые, как правило, являются специальными и ориентированы на определенный тип атаки, который, как предполагается, известен априори. Это связано с размером пространства генерации состязательной атаки, которое потенциально имеет большие размеры. Таким образом, как традиционные, так и основанные на нейронных сетях подходы к машинному обучению могут использоваться в зависимости от спецификаций задачи.

[80]Д. Гюмюшбаш, Т. Йылдырым, А. Дженовезе и Ф. Скотти, Всесторонний обзор баз данных и методов глубокого обучения для систем кибербезопасности и обнаружения вторжений, в журнале IEEE Systems Journal, vol. 15, нет. 2, стр. 1717-1731, июнь 2021 г., DOI: 10.1109/JSYST.2020.2992966.

[81]То же сноска 79

[82]Юнфэй Сун, Тянь Лю, Тунцюань Вэй, Сянфэн Ван, Чжэ Тао и Минсон Чен. Fda3: федеративная защита от атак злоумышленников для облачных приложений IoT. IEEE Transactions on Industrial Informatics, возраст 1–1 год, 2020 г. DOI: 10.1109 / TII.2020.3005969.

ИЗБРАННЫЕ ПРИМЕРЫ

Были рассмотрены четыре основные области из-за их сильной взаимозависимости с ИИ и кибербезопасностью, а именно телекоммуникации следующего поколения (6G), кибербиотехнологии, Интернет вещей (IoT) и киберфизические системы (CPS). Поскольку некоторые из этих областей все еще находятся на ранней стадии развития (по крайней мере, первые две), ожидается, что ИИ будет способствовать увеличению их потенциала. Это предположение оправдано не только с точки зрения потенциала, но и с точки зрения безопасности.

Однако существующие инструменты кибербезопасности, использующие ИИ, могут оказаться недостаточными для защиты этих технологий и областей. Использование ИИ в новых контекстах необходимо оценивать и часто адаптировать, особенно когда он учится на данных, описывающих шаблоны атак в новых областях атак.[83]. Однако для этого требуется достаточное количество справочных данных для обучения моделей, которые могут быть еще недоступны из-за новизны этих технологий и областей.

1,9СЛЕДУЮЩЕЕ ПОКОЛЕНИЕ ТЕЛЕКОММУНИКАЦИЙ[84]

В этом разделе мы рассмотрим, как 5G, более 5G и 6G могут в равной степени получать выгоду и подвергаться риску от использования ИИ. Некоторые многообещающие возможности ИИ[85]для поддержки кибербезопасности 5G перечислены ниже (не исчерпывающе):

- оптимизация ресурсов и динамические арбитражи, особенно в ситуации массовой мультимобильности[86],
- улучшение управления и координации алгоритмов[87],
- улучшение «кривой обучения» в управлении проблемами кибербезопасности, в частности, при обнаружении аномалий, например, потенциально связанных с вредоносным ПО, или даже уже перечисленных шаблонов атак[88],
- помогает развивать более гибкие и автоматизированные возможности, способные реагировать на тонко изменяющиеся или угрожающие ситуации[89],
- помощь в разработке механизмов безопасности путем создания моделей доверия, безопасности устройств и гарантирования данных для обеспечения систематической безопасности всей сети 5G.

[83]Пужоль, Гай (2020). Faut-il avoir per de la 5G. Париж, Ларусс, с. 217-219.

[84]В этой главе мы оставим в стороне проблему распознавания лиц и наблюдения на основе ИИ с использованием 5G, которая сама по себе является полноценной темой, вызывающей растущие опасения и технологическую мощь.

[85]Хайдер, Номан; Байг, Мухаммад Зишан; Имран, Мухаммед. 2020. Искусственный интеллект и машинное обучение в сетевой безопасности 5G: возможности, преимущества и будущие направления исследований. arXiv:2007.04490, основанный на технических характеристиках 3GPP Group Services and Systems Aspects.

[86]Когда множеству мобильных агентов необходимо иметь почти одновременный доступ к телекоммуникационным услугам, объем передаваемых и контролируемых данных должен поддерживаться ИИ, а также метанаблюдение за тем, как это может быть подвержено атакам, с формами атак. сами являются источником обучения ИИ.

[87]См. наш комментарий к объяснению Пужоля выше (указ. соч.).

[88]Фактически, именно так ИИ может все больше и больше участвовать в защите узлов 5G и даже терминалов 5G, то есть все более эффективно использовать прошлый опыт атак (быть атакованными).

[89]Эта проблема времени реакции или управления временем сама по себе является такой проблемой (поскольку злоумышленники также склонны использовать ИИ, чтобы увидеть, как системы защищаются от атакующих зондов), что более высокая ожидаемая производительность 6G и обеспечение кибербезопасности кажутся неизбежными (см. об этом Гуртов, Андрей (2020) Архитектура сетевой безопасности и криптографические технологии, приближающиеся к постквантовой эре, в Белой книге 6G: Проблемы исследования в области доверия, безопасности и конфиденциальности, Университет Оулу, Финляндия, 6G Research Visions, № 9, 2020 г., в частности , стр. 16, где автор подчеркивает ценность ИИ для обеспечения динамичности, соответствующей требованиям 6G для кибербезопасности.)

Сеть IoT, включающая как классические средства, так и предложенные выше, для сложных ситуаций и схем анализа больших данных,

- развертывание возможностей, которые усиливают функции безопасности даже при отсутствии данных о реальных атаках, например, с использованием GAN (генеративно-состязательных сетей).

Однако ИИ может создать несколько проблем для инфраструктуры 5G. Согласно Suomalainen et al. (2020)[90], существует множество уязвимостей, и необходимы дополнительные исследования, эксперименты и инициативы по коллективному обучению, чтобы сделать 5G более безопасным. В этом документе были освещены несколько вопросов, связанных с использованием ИИ в 5G, например, возможность квалификации рисков той или иной ситуации по ее «объяснимости».

Многие из вышеперечисленных ожиданий вряд ли будут полностью реализованы до следующего поколения связи (после 5G/6G). Ожидается, что разработка 6G достигнет технологической зрелости и стандартизации к концу этого десятилетия. На данном этапе важно помнить, что эта область исследований все еще далека от стандартизации функций и спецификаций кибербезопасности. Тем не менее, ключевым компонентом архитектуры 6G, несомненно, будет использование возможностей искусственного интеллекта, как это предлагается в Белой книге 6G (Гуртов, указ. соч.).

Ожидается, что 6G будет «с поддержкой ИИ» в том смысле, что он будет полагаться на ИИ в своей основной функции, на физическом уровне, и позволит использовать широкий спектр новых приложений на основе ИИ с необходимой адаптируемостью в реальном времени и также станет более защищенным от оппортунистических атак на основе ИИ. Ключевые области архитектуры 6G будут в некоторой (высокой) степени полагаться на ИИ, например, интеллектуальный край в реальном времени для расширенных возможностей управления в реальном времени в масштабе, распределенный ИИ для децентрализованного принятия решений, интеллектуальное распределение радиочастот для динамической конфигурации радиокадры, интеллектуальное управление сетью для сквозной автоматизации управления сетью[91]. Некоторыми примерами новых возможностей на основе ИИ являются мультисенсорная дополненная реальность (XR), подключенная робототехника и автономные системы (CRAS) или беспроводное взаимодействие мозг-компьютер (BCI).[92].

1.10 ИНТЕРНЕТ ВЕЩЕЙ (IOT) И ИНТЕРНЕТ ВСЕГО (IOE)[93]

В контексте IoT аспекты сложности, скорости и эффективности продвигаются искусственным интеллектом. Следующее поколение IoT, скорее всего, будет определяться потребностями отрасли. Приведем лишь один пример: искусственный интеллект может помочь улучшить меры безопасности, проверяя наличие вторжений и аномалий и прогнозируя риск перебоев в обслуживании. Другой пример: искусственный интеллект играет важную роль.[94]в анализе поступающих данных и общесетевой аналитике.

[90] Суомалайнен Дж., Юхола А., Шахабуддин С., Мяммела А. и Ахмад И. 2020. Машинное обучение угрожает безопасности 5G. IEEE Access, 8, 190822–190842.https://doi.org/10.1109/ACCESS.2020.3031966

[91] Ван и др. 2020. «Безопасность и конфиденциальность в сетях 6G: новые области и новые вызовы». Цифровые коммуникации и сети 6 (3): 281–91.https://doi.org/10.1016/j.dcan.2020.07.003

[92] Сиривардхана и др. 2021. «Искусственный интеллект и безопасность 6G: возможности и проблемы». https://doi.org/10.1109/EuCNC/6GSummit51104.2021.9482503

[93] Известно, что системы Интернета вещей содержат ряд уязвимостей, но ИИ в этой области является лишь одним из возможных инструментов для обнаружения аномалий или извлечения уроков из опыта прошлых атак и существенно не отличается от его наиболее общего использования. Однако в текущей ситуации, поскольку на IoT приходится значительное количество инцидентов кибербезопасности, важно понимать природу его уязвимостей, независимо от того, могут ли некоторые из этих проблем быть смягчены благодаря ИИ или нет.

[94] Фаггелла, Даниэль. 2019. «Обоюдоострая роль искусственного интеллекта в кибербезопасности - с доктором Романом В. Ямпольским». Эмердж.https://emerj.com/ai-podcast-interviews/artificial-intelligences-double-edged-role-in-cyber-security-with-dr-roman-vyampolskiy/

ИИ, несомненно, является отличным набором инструментов для снижения рисков IoT, будь то исследование уязвимостей, предвидение проблем (или даже прогнозирование их с помощью возможностей самоотчетов), контроль межсетевых проблем, организация потоков трафика и общее снижение рисков.[95].

Особый источник уязвимостей лежит в так называемом Интернете всего (IoE), эволюции Интернета вещей (IoT), который благодаря архитектуре «крошечных ячеек» становится комплексной экосистемой, соединяющей миллиарды различных устройств.[96]. Эти устройства являются главной целью для злоумышленников[97]. Кроме того, эта архитектурная функция также создает риски для конфиденциальности с точки зрения сбора данных о местоположении и идентификационных данных. Меньшие, плотные и постоянно подключенные локальные сети потенциально будут включать нательные сети, дроны и датчики окружающей среды с низким уровнем безопасности, которые собирают и обмениваются очень конфиденциальной информацией, как мы увидим для IoT в целом, проблема, которую сообщество 6G будет решать. должны иметь дело с эффективно.

Проблемы с безопасностью возникают, когда Edge Intelligence (Cloud at the Edge) ML развертывает инструменты, уязвимые для отравления атак или других форм вторжения в процессе обучения. Преднамеренное введение ложных данных или манипулирование логикой данных может привести к ошибкам в интерпретации или недобросовестному поведению.[98]. Одной из теоретических мер противодействия этой угрозе являются системы защиты, которые способны имитировать и превосходить атакующего.

1.11 КИБЕРБЕЗОПАСНОСТЬ В КИБЕРФИЗИЧЕСКИХ СИСТЕМАХ (КФС)

Киберфизические системы являются важнейшим элементом сложных технических систем, таких как системы электроснабжения, сети водоснабжения, транспортные системы, роботизированные системы, умные здания и т. д., улучшая общее использование и контроль их компонентов. Однако их присутствие открыло двери для кибератак.[99] [100]. Цель таких злонамеренных действий может быть разной и обычно включает в себя кражу, повреждение или даже уничтожение информации и/или компонентов системы.[101]. Излишне говорить, что в то время, когда мы пишем эти строки, существует реальная ситуация с войной на Украине, которая постоянно включает в себя эти угрозы для критически важной инфраструктуры.[102].

Существует три метода обнаружения кибератак в CPS:

[95]Для этого см., в частности, Hodo, Elike, et al. 2016. «Анализ угроз сетей IoT с использованием системы обнаружения вторжений искусственной нейронной сети». 2016 Международный симпозиум по сетям, компьютерам и коммуникациям (ISNCC), 1–6 мая. https://doi.org/10.1109/ISNCC.2016.7746067

[96]Презентация на Black Hat USA 2022 экосистемы API, соединяющей устройства IoT/IOE с функциональными возможностями. https://i.blackhat.com/USA-22/Wednesday/US-22-Shaik-Attacks-From-a-New-Front-Door-in-4G-5G-Mobile-Networks.pdf

[97]То же, что**Ошибка! Закладка не определена.**

[98]Бензаид и Т. Талеб. 2020. ИИ для сетей 5G за пределами сетей 5G: средство защиты или нападения на кибербезопасность? Сеть IEEE, Vol. 34, № 6, с. 140–147, 2020.https://doi.org/10.1109/MNET.011.2000088

[99]Шридхар Адепу, Венката Редди Паллети, Гьянендра Мишра и Адитья Матхур. Расследование кибератак на систему водоснабжения. В конспектах лекций по информатике, стр. 274–291. Springer International Publishing, 2020. DOI: 10.1007/978-3-030-61638-0_16. URL-адреchttps://doi.org/10.1007/978-3-030-61638-0_16

[100]Петер Эдер-Нойхаузер, Таня Зеби, Йоахим Фабини и Гернот Формайр. Модели кибератак для сред интеллектуальных сетей. Устойчивая энергетика, сети и сети, 12:10–29, декабрь 2017 г. DOI:10.1016/j.segan.2017.08.002. URL-адресhttps://doi.org/10.1016/j.segan.2017.08.002

[101]Антонелло Монти и Фердинанда Пончи. Электроэнергетические системы. Интеллектуальный мониторинг, управление и безопасность систем критической инфраструктуры, стр. 31–65. Springer Berlin Heidelberg, сентябрь 2014 г. DOI: 10.1007/978-3-662-44160-2_2. URL-адрес https://doi.org/10.1007/978-3-662-44160-2_2

[102]Просто для представления: по данным New York Times от 17 ноября 2022 года, к тому времени уже было 126 кибератак на украинскую энергосистему со стороны россиян.Агентство ООН заявляет, что российские атаки на украинские энергосистемы угрожают атомным станциям - The New York Times (nytimes.com)

а) на основе сигнатур, т.е. поиск известных шаблонов вредоносной активности в потоке данных с использованием предопределенного словаря атак[103],

б) на основе аномалий, т.е. оценки характерных черт нормального поведения и последующего обнаружения отклонений, которые могут проявиться при вторжении[104],

с) на основе контрмер, т. е. адаптации задействованных сигналов (путем добавления информации, демонстрирующей подлинность), чтобы упростить задачу обнаружения вторжений.[105].

Вышеупомянутые методы можно использовать в качестве первой линии защиты, если вычислительные затраты относительно невелики.

Методы, основанные на аномалиях и подозрительных корреляциях с большими данными, должны быть в состоянии решать более сложные случаи вредоносных событий и изощренных атак и считаются перспективными, т. е. для семейств CPS, таких как интеллектуальные сети, автомобильные, промышленные и медицинские CPS, и изучаются в литература[106]). Это больше связано с идеей приемлемой уверенности в данной системе в данное время и в данном контексте, а не с целями измерения абсолютной эффективности.[107].

Для моделирования и обнаружения аномалий можно использовать различные методы машинного обучения, включая нейронные сети, схемы на основе правил, предопределенные схемы фильтрации подозрительных больших данных.[108]. Наблюдается значительный рост исследований в области решений на основе ML из-за повсеместного развития и применения алгоритмов DL/RL. Однако, несмотря на эти непрерывные улучшения, кажется, что текущее состояние алгоритмов безопасности не может полностью поспевать за развитием новых атак. Отчасти это связано с изобретательностью злоумышленников, но также и со сложностью защиты сложных систем, включающих не только инфраструктуры, но и всех людей внутри и вне их, что делает их настоящими информационными экосистемами.

1.12 КИБЕР БИОБЕЗОПАСНОСТЬ

Растущая конвергенция биотехнологии и ИИ является новым полем для эксплуатации. Первоначальная попытка проблематизировать область исследований на стыке кибербезопасности, киберфизической безопасности и биобезопасности привела к предложенному определению кибербезопасности как «понимание уязвимости к нежелательному наблюдению, вторжениям, злонамеренным и вредоносным действиям, которые могут происходить в или на стыках взаимосвязанных наук о жизни и медицине, кибер-, кибер-физических системах, цепочках поставок и инфраструктурных системах, а также разработка и внедрение

[103]Ху Чжэнбин, Ли Чжитан и У Цзюньци. Новая система обнаружения сетевых вторжений (NIDS), основанная на поиске сигнатур интеллектуального анализа данных. В Первом международном семинаре по обнаружению знаний и интеллектуальному анализу данных (WKDD 2008), страницы 10–16, 2008 г. doi: 10.1109 / WKDD.2008.48.

[104]Ян Нейзиль, Ондрей Крейбич и Радислав Смид. Распределенная система обнаружения неисправностей на базе IWSN для мониторинга состояния машин. IEEE Transactions on Industrial Informatics, 10(2):1118–1123, 2014. DOI:10.1109/TII.2013.2290432.

[105]Илинь Мо, Рохан Чабуксвар и Бруно Синополи. Обнаружение атак на целостность SCADA-систем. IEEE Transactions on Control Systems Technology, 22(4):1396–1407, 2014. DOI:10.1109/TCST.2013.2280899.

[106]Феликс О. Оловони, Данда Б. Рават и Чунмей Лю. Устойчивое машинное обучение для сетевых киберфизических систем: обзор безопасности машинного обучения для защиты машинного обучения для CPS. IEEE Communications Surveys & Tutorials, 23(1):524–552, 2021. ISSN 2373-745X. DOI: 10.1109/comst.2020.3036778. URL-адрес http://dx.doi.org/10.1109/COMST.2020.3036778

[107]См., в частности, Siau Keng and Wang Weiyu (2018). Укрепление доверия к искусственному интеллекту, машинному обучению и робототехнике. Журнал бизнес-технологий CUTTER(2) (PDF) Укрепление доверия к искусственному интеллекту, машинному обучению и робототехнике (researchgate.net)

[108]Сиддхарт Шридхар и Манимаран Говиндарасу. Обнаружение и смягчение атак на основе моделей для автоматического управления генерацией. IEEE Transactions on Smart Grid, 5(2):580–591, 2014. DOI:10.1109/TSG.2014.2298195.

меры по предотвращению, защите от, смягчению последствий, расследованию и связыванию таких угроз с безопасностью, конкурентоспособностью и устойчивостью».[109].

Важнейший механизм, внедренный с использованием ИИ в биотехнологии[110]— это способность автоматизировать сложные задачи без непосредственного контроля или использовать кибератаки для использования биоавтоматизации в злонамеренных целях.

В то же время они являются примерами «вызывающих озабоченность исследований двойного назначения» (DURC), то есть технологий, которые явно оказывают положительное влияние, открывая новые возможности, которые также могут быть использованы в злонамеренных целях (Pauwels, 2021). Как обсуждалось ранее, основной проблемой ИИ является объяснимость и производство воспроизводимых и пригодных для использования знаний (Jordan et al., 2020). Однако еще предстоит продемонстрировать с помощью реальных доказательств, что биоэволюция может создавать новые конкретные угрозы, которые не являются просто расширением существующей потенциальной поверхности атаки. Биометрические системы показывают, что речь идет скорее о распространении развертываний кибербезопасности, чем о реальном изменении парадигмы, но, конечно, кажется, что еще рано заканчивать эту дискуссию.

[109]Пеккуд, Дж., Гальегос, Дж. Э., Марч, Р., Бухгольц, В. Г., Раман, С. 2018. Кибербиобезопасность: от наивного доверия к осведомленности о рисках. Тенденции биотехнологии, 36(1), 4-7.https://doi.org/10.1016/j.tibtech.2017.10.012

[110]На самом деле это очень разнообразный ландшафт с очевидными возможностями перекрестного обогащения между областями применения и дисциплинами, например, одним из основных применений ИИ в биополе является помощь в идентификации и моделировании новых белков высокой -потенциальные фарма-ориентированные молекулы.

ИИ В КИБЕРБЕЗОПАСНОСТИ: ПРОБЕЛЫ ИССЛЕДОВАНИЙ И ПОТРЕБНОСТИ

В следующем разделе определяются пробелы в исследованиях, связанные с некоторыми проблемами и проблемами, указанными в предыдущей главе. Чтобы закрыть некоторые из этих пробелов, мы определили возможности для дальнейших исследований, которые представлены в разделе 7.2.

1.13 ОТКРЫТЫЕ ВОПРОСЫ И ПРОБЛЕМЫ

Подавляющее большинство систем ИИ разрабатываются на основе одного (или нескольких) из следующих предположений (в контексте данного исследования): (i) наличие обычно значительного объема высококачественных данных, представляющих как нормальные, так и атакуемые состояния система; (ii) наличие экспертных знаний в предметной области, на основе которых разрабатываются функции, адаптированные к рассматриваемой проблеме; (iii) стационарность во времени, т. е. распределения данных, представляющих состояния системы, не дрейфуют и не изменяются резко во времени (что, конечно, также зависит от используемого метода ИИ); (iv) знание словаря классов, включая полный набор всех состояний системы; (v) знание лежащих в основе аналитических взаимосвязей, управляющих контролируемой системой; (vi) что обученные модели не имеют каких-либо предубеждений,

In this chapter, we provide details on how AI raises specific issues and challenges, on which we identify further research opportunities in the next section.

There are several open issues and challenges that have not yet been addressed and that can be further explored by research. The following non-exhaustive list presents some of the most noteworthy open issues:

- achieving verifiable, reliable, explainable, auditable, robust and unbiased AI[111];
- quality of data sets: among the self-built limitations, there is the notion of 'trash in/ trash out' i.e. you need good quality inputs to get reasonable quality output[112], meaning not only the quality of data bearing in mind their practical algorithmic usability but also how well they represent the problem being tackled;

[111] Verifiable: there should proof that the AI-based approach acts correctly for a range of inputs; Reliable: the AI-based approach should operate as expected, even for inputs coming out of data distribution the system has not 'seen before; Explainable: the system should be structured in a way so that the operator is able to backtrack any obtained prediction/decision in terms of data, scenarios, and assumptions that led to it; Robust against adversarial attacks that can jeopardise an AI-based tool, thus any deployed systems should not be vulnerable to noise and specific inputs designed to manipulate its operation; Auditable: the operator should be able to 'open' and check the internal state of the deployed system at any point in time and especially when a prediction is carried out, and Unbiased: the system should not display unintended preference towards specific predictions, recommendations, etc.

[112] Pouyanfar et al, 2019, A Survey on Deep Learning DOI: 10.1145/3234150. Association for Computing Machinery (ACM)

- how to achieve end-to-end protection (data is particularly at risk when it is in transit[113]);
- how to achieve optimal accuracy under real-world conditions and not in a simulated environment[114];
- the need for computational complexity and 'low-latency operation' to be addressed especially when the system being monitored is of critical importance[115];
- the need to investigate whether the inferred models are valid or biased, or whether there are perceive changes in the time variance[116];
- Ensuring that the security of the protection mechanism is assessed following a standardised framework considering diverse malicious attempts, cases, figures of merit, etc. (security-by-design)[117];
- preservation of privacy e.g. training data and confidentiality of the information flowing in the system so that the characteristics of the system are not exposed indirectly and potentially classified information is not also revealed[118].

1.14 RESEARCH GAPS

The following non-exhaustive list provides the research gaps that were identified in our study:

- Construction of effective AI models with a relatively small amount of data by moving from big data to a small data environment;
- Elaboration on raw data targeting end-to-end solutions where feature engineering and the need for domain expertise (knowledge) is minimised or even eliminated;
- Incorporation of change detection and adaptation mechanisms to address non-stationarities (changes in the time variance of system states);
- Periodical assessment of the validity of the developed model(s) so as to promptly detect and address potential bias(es) which introduce additional vulnerabilities;
- Development of approaches to remove existing biases, imbalances, etc. which may degrade the performance of the model;
- Development of standardised data sets following these requirements in order to reliably reproduce and compare existing AI-based solutions;

113 Trantidou, et al, 2022, SENTINEL - Approachable, tailor-made cybersecurity and data protection for small enterprises, in PROCEEDINGS 2022 IEEE International Conference on Cyber Security and Resilience (CSR), DOI: 10.1109/CSR54599.2022.9850297.
114 Kavak et al, 2021, Simulation for cybersecurity: state of the art and future directions, DOI: 10.1093/cybsec/tyab005, Oxford University Press (OUP), Journal of Cybersecurity.
115 Zhenyu Guan, Liangxu Bian, Tao Shang, and Jianwei Liu. When machine learning meets security issues: A survey. In 2018 IEEE International Conference on Intelligence and Safety for Robotics (ISR), pages 158–165, 2018. doi:10.1109/IISR.2018.8535799. Liu et al, 2022, Complexity Measures for IoT Network Traffic, IEEE Internet of Things Journal, DOI: 10.1109/JIOT.2022.3197323.
116 Ntalampiras and Potamitis, 2022, Few-shot learning for modelling cyber physical systems in non-stationary environments, DOI: 10.1007/s00521-022-07903-0. Springer Science and Business Media (LLC), Journal Neural Computing and Applications.

117 Karie et al, 2021 A Review of Security Standards and Frameworks for IoT-Based Smart Environments, DOI: 10.1109/ACCESS.2021.3109886, IEEE
118 Domingo Ferrer and Alberto Blanco-Justicia, 2020, Privacy-Preserving Technologies, DOI: 10.1007/978-3-030-29053-5_14, Springer International Publishing, The International Library of Ethics, Law and Technology.

- Development of approaches to distinguish malicious attacks from faulty states[119];
- On understanding how the efficacy of AI-based tools and methodologies is altered in terms of both accuracy and computational complexity due to an increase in the scale of the system[120], and consequently an increase in the impact of a cyberattack;
- Modelling interdependent cyber-physical systems in order to assess the impact of vulnerabilities;
- The need for a standardised performance evaluation framework to enable reliable comparison between solutions addressing the same or similar problems;
- Provision of context awareness[121] in ML in order to boost resiliency;
- Bringing 'humans into the loop' e.g. training practitioners using real-world scenarios.

While these research gaps cover AI in general, they are particularly important for cybersecurity applications.

1.15 RESEARCH NEEDS

The following list presents the needs for further research on the use of AI or ML concepts in cybersecurity:

1. test beds to study and optimise the performance of ML-based tools and technologies used for cybersecurity,
2. development of penetration testing tools based on AI and ML to find and exploit security vulnerabilities to assess the behaviour of attackers,
3. development of standardised frameworks assessing the preservation of privacy and the confidentiality of information flows as well as the designed system,
4. development of AI training models for practitioners using real-world scenarios,
5. establishing an observatory for AI and cybersecurity threats.

The tables below present ENISA's proposals for future funding calls based on the needs identified in the list above.

Test-beds to optimise the performance of AI/ML-based tools and technologies used for cybersecurity
Type: AI for cybersecurity

[119] Yannis Soupionis, Stavros Ntalampiras, and Georgios Giannopoulos. Faults and cyber-attacks detection in critical infrastructures. In Critical Information Infrastructures Security, pages 283–289. Springer International Publishing, 2016. DOI:10.1007/978-3-319-31664-2_29. URL https://doi.org/10.1007/978-3-319-31664-2_29 .

[120] Cesare Alippi, Stavros Ntalampiras, and Manuel Roveri. Model-free fault detection and isolation in large-scale cyberphysical systems. IEEE Transactions on Emerging Topics in Computational Intelligence, 1(1):61–71, 2017. DOI:10.1109/TETCI.2016.2641452

[121] Context awareness refers to the ability of the protection mechanism to collect information from its surrounding and interconnected environment in order to adapt to potential changes and incorporate them into its operation. As such, protection quality could be boosted since previously unavailable information would be employed to learn the system model on-the-fly.

Description:

Testbeds are required to study and optimise the performance of ML-based tools and technologies used for cybersecurity.

Objectives:

1. Develop test beds to optimise the performance of AI/ML used in cybersecurity.

Entities:

- Security researchers
- Application developers

Beneficiaries:

- Providers of AI tools and solutions

Existing research:

Much of the existing research effort is focused on achieving optimal accuracy in a simulated environment, which usually does not reflect the performance achieved under real-world conditions. Computational complexity and real-time operation must be taken into account, especially when the system being monitored is of critical importance. In this direction, efforts are needed to improve and construct test environments to study and optimise the performance of ML-based cybersecurity tools and technologies.

Standardised frameworks assessing the preservation of privacy and confidentiality

Type: AI for Security

Description:

Standardised frameworks assessing the preservation of privacy and confidentiality of the information flows as well as of the designed solutions need to be developed.

Objectives:

1. Preservation of privacy, e.g. training data and confidentiality of the information flows in systems so that the characteristics of the systems are not indirectly exposed and potentially classified information is not revealed;
2. Privacy preservation and confidentiality offered by the designed solutions.

Entities:

- Security researchers
- Application developers
- GDPR-related specialists

Beneficiaries:

- Software industry

Existing research:

The preservation of privacy and confidentiality of the information flows and of the designed solutions are issues that are rarely considered.

AI/ML-based penetration testing

Type: AI for cybersecurity

Description:

AI-powered penetration testing

Objectives:

1. Using AI/ML to test a system to find security vulnerabilities that an attacker could exploit and then trying to figure out what an attacker will do.

Entities:

- Security researchers
- Application developers

Beneficiaries:

- Cybersecurity practitioners
- Cybersecurity industry

Existing research:

Threat actors can take advantage of training data by generating a backdoor. They can use AI to find the most likely vulnerability to exploit. Penetration testing can lead to finding vulnerabilities that give outsiders access to the data training models.

There are many automated tools that complement penetration testing tools. These automated solutions have some basic AI capabilities, and these capabilities are gradually increasing thanks to ongoing research and open competitions. For example, the 2016 Cyber Grand Challenge - a DARPA-sponsored competition - challenged people to build hacking bots and compete against each other. These artificially intelligent bots perform penetration tests to look for security vulnerabilities and close them before competing teams can exploit them. For example, Mayhem was able to find, fix and search for intrusions on its host system, while simultaneously finding and exploiting vulnerabilities on rival systems.

As we write this study, Generative Pre-trained Transformer software is emerging first through OpenChat GPT and then with the promises of a handful of competitors. Research

into how this software can enhance deception clearly needs to be undertaken urgently. Then, hopefully, the training of specialists can become more consistent, taking this development into account.

Training models for practitioners in real-world scenarios
Type: AI training
Description: Develop training on AI for practitioners using real-world scenarios.
Objectives: 1. Training on AI in real-world scenarios
Entities: • Security researchers • Universities
Beneficiaries: • Cybersecurity practitioners
Existing research: Cybersecurity is a never-ending task that is not only about stopping threats from intruders, but also about not wasting time and energy on false positives, which requires a 'rock-solid' belief in the AI model coupled with rapid escalation to human analysts. The best AI requires data scientists, statistics and as much human input as possible. The foundation for effective 'triage' activity against the multitude of risks and forms of attack is teaching AI when incidents occur, teaching an AI threat disposition system, training practitioners using real-world scenarios and conducting real behavioural threat analysis. This is also what the 'human in the loop'[122] interaction requires.

AI & cybersecurity threats Observatory
Type: AI for cybersecurity
Description:

[122] Jones, Tim. 2019. IBM Developer 'Take a Look at AI and Security and Explore the Use of Machine Learning Algorithms in Threat Detection and Management. (blog). 19 August 2019. https://developer.ibm.com/articles/ai-and-security/

Sharing real-time information on AI and cybersecurity threats, at software and hardware levels, as well as attackers' modus operandi is a must for Europe to function as a coherent defence arena.
Objectives: Develop an inventory of trends and threats at software and hardware levels as well as the modus operandi of attackers.
Entities: • European Cybersecurity Competence Centre
Beneficiaries: • Cybersecurity community
Existing research: Developing an observatory of threats would require developing a network of observatories across the EU and linked to like-minded countries and key partners and organisations. The European Cybersecurity Centre could be an organisation to play such role, provided that this particular 'observation-and-sharing' objective be specified.

CONCLUSIONS AND NEXT STEPS

AI is gaining attention in most quadrants of society and the economy, as it can impact people's daily lives and plays a key role in the ongoing digital transformation through its automated decision-making capabilities. AI is also seen as an important enabler of cybersecurity innovation for two main reasons: its ability to detect and respond to cyber threats and the need to secure AI-based applications.

The EU has long considered AI as a technology of strategic importance and refers to it in various policy and strategy documents. ENISA is contributing to these EU efforts with technical studies on cybersecurity and AI. For example, the cyber threat landscape for AI[123] raised awareness on the opportunities and challenges of this technology. The Agency has already published two studies on this topic and this report will be the third publication aiming to provide a research and innovation perspective of cybersecurity and AI. In preparing these studies, the Agency is supported by the R&I community and has established an *ad-hoc* working group[124] with experts and stakeholders from different fields and domains.

This study makes recommendations to address some of the challenges through research and identifies key areas to guide stakeholders driving cybersecurity research and development on AI and cybersecurity. These recommendations constitute ENISA's advice, in particular to the EC and ECCC, using its prerogative as an observer on the Governing Board and advisor to the Centre. The findings were used to produce an assessment of the current state of cybersecurity research and innovation in the EU and contribute to the analysis of research and innovation priorities for 2022, presented in a separate report.

In this context and as next steps, ENISA will:

1. present and discuss the research and innovation priorities identified in 2022 with members of the ECCC Governing Board and NCCs;
2. develop a roadmap and establish an observatory for cybersecurity R&I where AI is a key technology; and
3. continue identifying R&I needs and priorities as part of ENISA's mandate (Article 11 of the CSA).

123 ENISA. https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges , last access March 2023.
124 ENISA. https://www.enisa.europa.eu/topics/iot-and-smart-infrastructures/artificial_intelligence/ad-hoc-workinggroup/adhoc_wg_calls , last accessed March 2023.

TP-07-23-202-EN-N

ABOUT ENISA

The European Union Agency for Cybersecurity, ENISA, is the Union's agency dedicated to achieving a high common level of cybersecurity across Europe. Established in 2004 and strengthened by the EU Cybersecurity Act, the European Union Agency for Cybersecurity contributes to EU cyber policy, enhances the trustworthiness of ICT products, services and processes with cybersecurity certification schemes, cooperates with Member States and EU bodies and helps Europe prepare for the cyber challenges of tomorrow. Through knowledge sharing, capacity building and awareness raising, the Agency works together with its key stakeholders to strengthen trust in the connected economy, to boost resilience of the Union's infrastructure and, ultimately, to keep Europe's society and citizens digitally secure. More information about ENISA and its work can be found at: www.enisa.europa.eu.

ENISA
European Union Agency for Cybersecurity

Athens Office
Agamemnonos 14, Chalandri 15231, Attiki, Greece

Heraklion Office
95 Nikolaou Plastira
700 13 Vassilika Vouton, Heraklion, Greece

enisa.europa.eu

ISBN 978-92-9204-637-8
doi: 10.2824/808362