Искусственный интеллект и кибербезопасность

Д.Е. Намиот, Е.А. Ильюшин, И.В. Чижов

***Аннотация*—В этой статье мы рассматриваем связь систем искусственного интеллекта и кибербезопасности. В современной трактовке, системы искусственного интеллекта – это системы машинного обучения, иногда это еще более сужается до искусственных нейронных сетей. Если мы говорим о все более широком проникновении машинного обучения в разные сферы применения информационных технологий, то, естественно, что должны возникать пересечения с кибербезопасностью. Но проблема в том, что такое пересечение не может быть описано какой-то одной моделью. Сочетания Искусственный интеллект и кибербезопасность имеют множество разных аспектов применения. Общим является, естественно, использование методов машинного обучения, но задачи, а также достигнутые на сегодняшний день результаты, являются совершенно разными. Например, если применение машинного обучения для обнаружения атак и вторжений показывает реальные достижения по сравнению с применявшимися ранее подходами, то атаки на сами системы машинного обучения пока полностью побеждают возможные защиты. Классификации моделей применения машинного обучения в кибербезопасности и посвящена данная статья.**

***Ключевые слова* – искусственный интеллект, машинное обучение, кибербезопасность.**

I. Введение

Искусственный интеллект на сегодняшний день переопределил то, как используются компьютеры [1]. Искусственный интеллект становится частью повседневной жизни. Как отмечается в [1], даже такие абсолютно понятные пользовательские устройства, как мобильные телефоны уже содержат чипы для искусственного интеллекта (Pixel 6 от Google, iPhone). ИИ меняет то, как компьютеры программируются и как они используются. Благодаря машинному обучению программисты больше не пишут правила. Вместо этого они создают нейронную сеть, которая сама извлекает эти правила в процессе обучения. Это принципиально другой способ мышления.

Искусственный интеллект (а на сегодняшний день – это машинное обучение) повсюду, компьютерная безопасность должна охватывать все процессы, соответственно, эти два понятия не могли не встретиться. Именно отношения искусственного интеллекта и кибербезопасности и есть тема настоящей статьи. Эти отношения разные, решения существуют абсолютно разные, и степень решения различных проблем также разная. Тема Искусственный Интеллект и кибербезопасность не может быть представлена как одно решение (или даже совокупность нескольких решений), поскольку она описывает совершенно разные задачи.

Например, учебная магистерская программа факультета ВМК МГУ имени М.В. Ломоносова, названная Искусственный интеллект в кибербезопасности [3], исходя из ее содержания, должна бы была быть названа как Кибербезопасность систем искусственного интеллекта. Это точнее отражало бы задачи анализа устойчивости систем машинного обучения, состязательные атаки и другие, рассматриваемые в программе темы. А предлог “в” в названии больше соответствует, например, использованию машинного обучения при анализе логов для определения вторжений и т.п. Во всех задачах используется машинное обучение, но задачи совершенно разные и текущее состояние дел совершенно разное. Если применение машинного обучения для анализа разного рода журналов в целях определения шаблонов, характерных для вторжения, показывает явные успехи (это машинное обучение в кибербезопасности), то относительно кибербезопасности самих систем машинного обучения есть понимание задач при отсутствии исчерпывающих решений.

Компания Микрософт [2] предложила следующее разделение темы ИИ и кибербезопасность:

- Повышение кибербезопасности с помощью ИИ (использование ИИ в кибербезопасности)
- Кибератаки с использованием ИИ (использование ИИ для усиления кибератак)
- Кибербезопасность систем ИИ (атаки на системы ИИ)
- Использование ИИ в злонамеренных информационных операциях (фейки с использованием ИИ)

Мы будем следовать этой классификации, и оставшаяся часть статьи структурирована в соответствии с этим разделением.

Статья получена 2 июля 2022.
Намиот Д.Е. – МГУ имени М.В. Ломоносова (email: dnamiot@gmail.com)
Ильюшин Е.А. – МГУ имени М.В. Ломоносова (email: john.ilyushin@gmail.com)
Чижов И.В. – МГУ имени М.В. Ломоносова (email: ichizhov@cs.msu.ru)

II. Повышение кибербезопасности с помощью ИИ

Можно сказать, что это наиболее продвинутая на сегодняшний день область. Ценность, которую привносит здесь машинное обучение, состоит в определении атак, поиске шаблонов и закономерностей, соответствующих вторжениям, быстром анализе и приоритизации угроз, анализе накопленной информации для адаптации методов обнаружения вторжения.

Первый ответ на вопрос, зачем здесь ИИ, согласно [2], заключается в слове “автоматизация”. Автор приводит американские данные Бюро статистики труда США о том, что возможности трудоустройства в сфере кибербезопасности вырастут на 33% с 2020 по 2030 год, что более чем в шесть раз превышает средний показатель по стране [4]. Вряд ли картина в других странах отличается от приведенной. При этом, согласно исследованию рынка труда в части кибербезопасности ISC, опубликованному в октябре 2021 года, во всем мире не хватает 2.72 миллиона специалистов по кибербезопасности [5]. Соответственно, альтернативы автоматизации решения задач кибербезопасности просто нет.

Задачи кибербезопасности состоят из предотвращения атак, обнаружения атак, проведения расследований, классификации и анализе угроз, а также обучения и моделирования систем кибербезопасности.

Предотвращение атак (профилактика) – это усилия по снижению количества уязвимостей содержащихся в программном обеспечение. Типичные примеры есть, например, в обзоре [6], который описывает системы машинного обучения, выполняющие поиск вредоносных приложений на Android. Собираются характеристики приложений (рис. 1), на датасетах по приложениям обучаются классификаторы.

Analysis Type	Feature Extraction Method	Features Extracted
Static	Manifest analysis	Package name, Permissions, Intents, Activities, Services, Providers
	Code analysis	API calls, Information flow, Taint tracking, Opcodes, Native code, Cleartext analysis
Dynamic	Network traffic analysis	URLs, IPs, Network Protocols, Certificates, Non-encrypted data
	Code instrumentation	Java classes, intents, network traffic
	System calls analysis	System calls
	System resources analysis	CPU, Memory, and Battery usage, Process reports, Network usage
	User interaction analysis	Buttons, Icons, Actions/Events

Рис.1. Характеристики приложений

Есть даже статистика по используемым методам классификации, где лидирует Random Forest.

Как отмечено в [2], в 2021 году Институт AV-Test [7] обнаружил более 125 миллионов новых вредоносных программ. Способность методов машинного обучения обобщать прошлые шаблоны для обнаружения новых вариантов вредоносных программ и является ключом к построению масштабируемой системы защиты.

Можно отметить, что поиск в Google Scholar работ по запросу “ML for malware detection” показывает более 20 000 статей [8].

Глубокое обучение также активно используется в этой области [9]. В этой работе описывается система, созданная по государственному гранту Китая для ключевых технологий. Интересный сравнительный анализ моделей глубокого обучения для определения вредоносных приложений есть в работе [10]. Все такие работы имеют практическое применение, например, Microsoft 365 Defender [11] также использует глубокое обучение.

Отметим, что под словом “программы” не следует понимать здесь только код. Например, в работе [12] описывается модель глубокого обучения для определения фишинговых URL. И это только один пример из множества подобных работ. В целом, фишинговые атаки довольно разнообразны и использование машинного обучения для их обнаружения описывалось еще в 2008 году [13]. Обзор современных подходов, использующих машинное обучение в борьбе с фишингом, есть, например, в свежей работе [14].

Обнаружение атак включает выявление подозрительного поведения и оповещение о нем непосредственно по мере его возникновения. Цель состоит в том, чтобы быстро реагировать на атаки, включая определение масштаба атаки, закрытие входов для атакующих и устранение уязвимостей (бэкдоров и т.п.), которые мог эксплуатировать злоумышленник.

Очевидно, что поиск, в общем случае, неизвестных шаблонов атак потенциально может приводить к большому числу ложных срабатываний (false positives) [15]. В литературе отмечается, что основная проблема при обнаружении подозрительной активности как раз и заключается в том, чтобы найти правильный баланс между обеспечением достаточного охвата за счет поиска точных предупреждений системы безопасности и количеством ложных срабатываний.

Можно выделить следующие направления, касающиеся использования машинного обучения для предупреждений об атаках [2]:

(1) расстановка приоритетов для предупреждений о потенциальных атаках [16],

(2) выявление многочисленных попыток взлома с течением времени, которые являются частью более крупных и длительных кампаний по взлому [16],

(3) обнаружение следов действий вредоносных программ, как внутри компьютера, так и в сети [17]

(4) идентификация потока вредоносного программного обеспечения, внедряемого через конкретную организацию. Это так-называемые Living off the Land (LotL) атаки – кибератаки, в которых атакующий использует легальное программное обеспечение в организации для выполнения атакующих действий [18].

(5) определение автоматизированных подходов к смягчению последствий атак, когда требуется быстрое реагирование, чтобы предотвратить распространение атаки. Например, автоматизированная система может отключать сетевое подключение и блокировать устройство, если обнаруживается последовательность предупреждений, которая, как известно, связана с действиями программы-вымогателя [19].

В качестве достаточно подробных обзоров моделей глубокого обучения, используемых для определения атак, можно привести работы [20] и [21].

Расследование и исправление (восстановление после атак) - это методы, используемые после нарушения безопасности, предназначенные для того, чтобы предоставить клиентам целостное представление о нарушениях безопасности, включая степень нарушения, список затронутых устройств и данных, информацию о распространении атаки и о причинах инцидента. Это достаточно новая область. Как примеры можно назвать работы [22, 23]. С другой стороны, большое количество атак привело к накоплению и большого количества информации о них, так что есть материал для исследований. По этой тематике есть интересная презентация DARPA по атрибутированию атак [24].

Методы искусственного интеллекта также находят применение при анализе угроз на высоком уровне. Можно привести примеры работ [25, 26], в которых представлены фреймворки по анализу информации об атаках, помогающие, например, выявлять сходство между различными кампаниями (взломами и т.п.).

III Кибератаки с использованием ИИ

В связи с атаками используется термин наступательный ИИ. В качестве примерных обзоров можно указать работы [27, 28]. Рисунок 2 из работы [28] суммирует направления атак с использованием систем машинного обучения на матрице угроз MITRE.

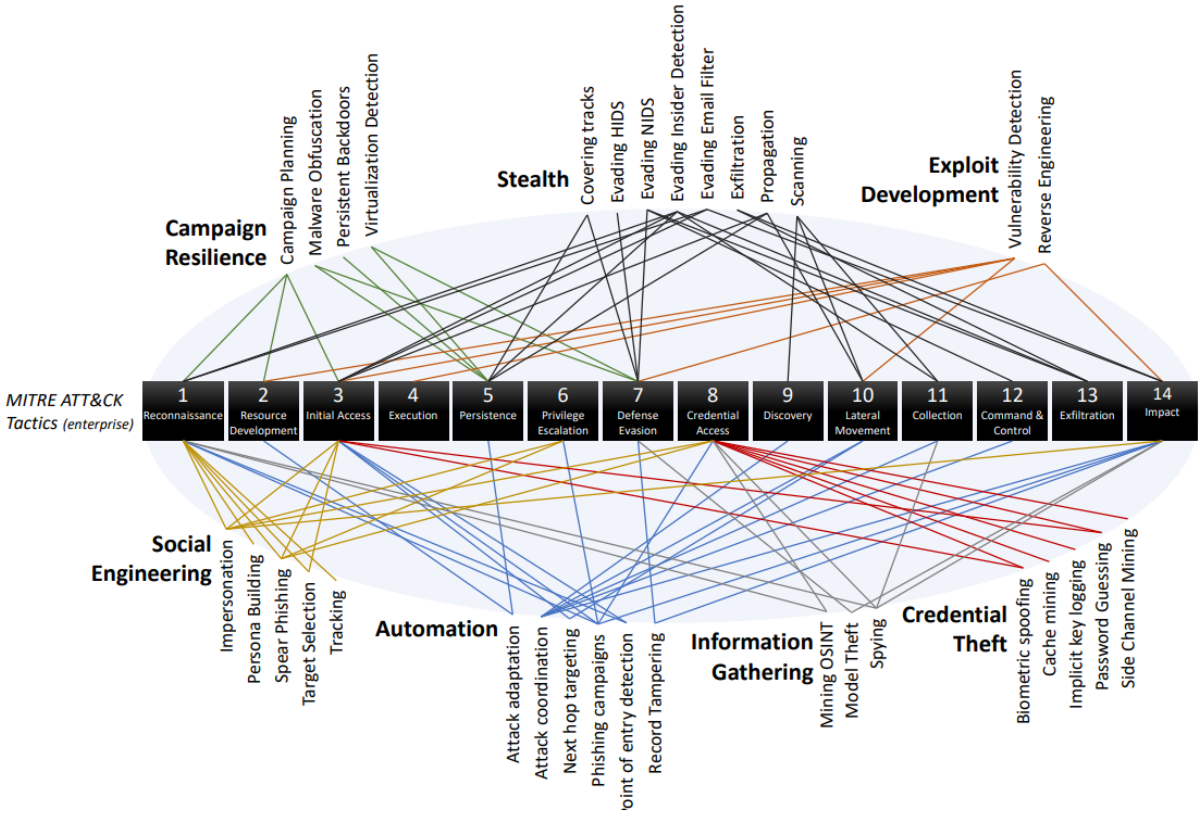

Рис. 2. Машинное обучение в кибератаках [28].

Авторы последнего обзора выделили следующие области атак с использованием ИИ.

1. Прогнозирование - сделать некоторый прогноз на

основе ранее наблюдаемых данных. Пример атаки с использованием машинного обучения - идентификация нажатий клавиш на смартфоне на основе движения (вибрации) [29, 30]. Другие приведенные примеры касались предсказания чувствительных данных для пользователей социальных сетей [31] (поиск слабого звена для атаки), поиска уязвимостей программного обеспечения [32, 33, 34].

2. Генерация – создание контента с использованием ИИ. Примеры такой генерации для наступательных целей – фальсификация медиа-данных [35], подбор паролей [36], модификацию трафика [37]. Последнее (в англоязычной литературе – traffic-space attacks) представляет собой, фактически, состязательную атаку на систему машинного обучения, которая используется для анализа трафика (определения вторжений). Цель атаки – скрыть реальное вторжение.

Дипфейки — еще один пример наступательного ИИ в этой категории. Дипфейк — это правдоподобный медиафайл. Создаются они с использованием глубокого обучения. Технология может быть использована для того, чтобы выдавать себя за жертву, имитируя ее голос или лицо при совершении фишинговой атаки [38].

3. Анализ - это задача анализа или извлечения полезной информации из данных или модели. Исследование атакуемой модели ML, с целью определения реальных факторов, влияющих, например, на классификацию. Имеется в виду использование объясняющих подходов (LIME, SHAPLEY и др.). Понимание работы атакуемой модели необходимо для создания эффективных атак или сокрытия вторжений. Если атакуемая модель недоступна, то такие эксперименты могут проводиться на ее теневой копии.

4. Поиск - это задача поиска информации или объектов для атаки по заданным критериям. Приведенные примеры – поиск (идентификация) человека по изображениям на нескольких взломанных камерах [39, 40], поиск возможных инсайдеров по семантическому анализу публикаций в социальных сетях [41], аннотирование (реферирование, суммаризация) документов при сборе данных из открытых источников (OSINT – открытая разведка) [42] (последнее есть пример автоматизации).

5. Принятие решения – это задачи разработки стратегического плана или координации операции (атаки). Примеры в ИИ — использование роевого интеллекта для управления автономной сетью ботов [43] и планирование оптимальных атак на сети [44].

В презентации [45] отмечается, что автоматизировать атаки можно и без машинного обучения, но обучение с подкреплением (reinforcement learning) имеет все шансы стать основным инструментом в осуществлении атак.

Микрософт в отчете [46] ожидает, что использование ИИ в кибератаках начнется с опытных участников, но быстро распространится на более широкую экосистему за счет повышения уровня сотрудничества и коммерциализации используемых инструментов. В частности, инструменты атакующих включают общие базовые тактики обхода защиты, как описано в атласе MITRE [47]. Одна из наиболее успешно используемых систем автоматизации в наступательном ИИ – это боты в социальных сетях [48]. Другой пример автоматизации наступательных действий представлен в работе [49] – автоматизированный тест на проникновение (penetration test), использующий обучение с подкреплением (рис. 3).

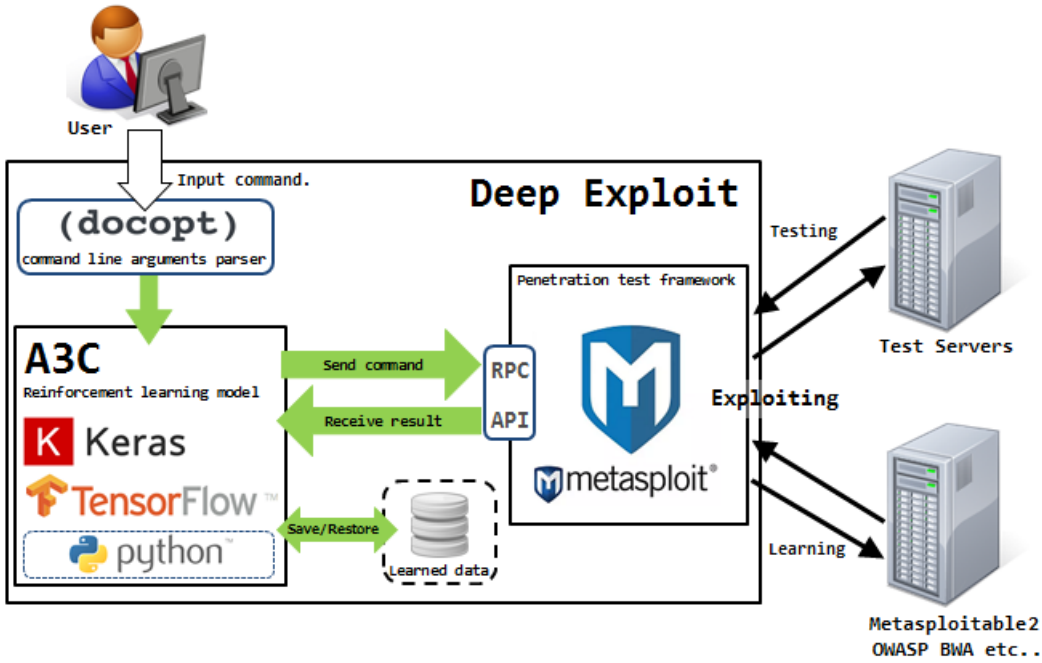

Рис. 3. Deep Exploit [49]

Машинное обучение используется для атак на биометрические системы аутентификации: подделка голоса и т.п. [50, 51].

Выше мы говорили об определении фишинговых атак с помощью машинного обучения. Но машинное обучение используется и при генерации фишинговых атак [52, 53]. Цель – обойти системы защиты, создать более привлекательный контент и побудить

пользователей кликнуть злонамеренную ссылку, установить в системе программное обеспечение и т.д. Примеры наступательных действий включают также подбор паролей [55], запутывание исходного кода программ [56], маскировку трафика [57], управление сетью ботов [58].

Наступательному ИИ посвящен отдельный воркшоп (отчет – доступен), организованный компанией Микрософт [54]. Атаки с использованием ИИ рассматриваются также в довольно подробном отчете National Security Commission on Artificial Intelligence (NSCAI) [59].

IV Атаки на системы ИИ

Это достаточно новая область для компьютерной безопасности. Атаки могут быть направлены на сами системы ИИ (фактически – на системы машинного обучения). Любая внедренная система машинного обучения есть, в конечном итоге, программа. Но проблема состоит в том, что для таких приложений традиционные методы анализа безопасности неприменимы. Проблемы с безопасностью именно таких приложений не могут быть решены традиционными методами. Конечно, скомпрометированная среда исполнения программы будет приводить к проблемам. Но это не главная беда.

Системы машинного обучения зависят от данных. На основе представленных тренировочных данных система вырабатывает некие обобщения, которые затем используются при обработке реальных (тестовых) данных. Так вот модификации данных на разных этапах конвейера машинного обучения и приводят к тому, что такие системы могут либо вовсе не работать, либо наоборот, выдавать нужные атакующему результаты. При этом специально модифицированные данные будут, вообще говоря, точно такими же, как и “чистые” данные. В общем случае, их нельзя будет различить. Более того, поскольку обучение всегда производится на некотором тренировочном наборе данных, генеральная совокупность остается, в общем случае, неизвестной. И “изменение” данных на этапе эксплуатации может случиться (и чаще всего случается) безо всяких зловредных действий. Просто потому, что так устроены сами данные. Атаками в данном случае называют именно специальное изменение данных или специальную подстановку данных, на которых система работает неверно (вообще не работает). В общем виде – это проблема устойчивости систем машинного обучения. Этой проблеме сейчас уделяется много внимания, поскольку это основное, что препятствует использованию систем ИИ в критических приложениях (авионика, ядерная безопасность и т.п.) [60, 61].

Другое название атак на системы машинного обучения – состязательные примеры [62]. Таким образом, враждебные воздействия на системы могут осуществляться в форме традиционных уязвимостей, а также с помощью новой категории: состязательных примеров.

Как примеры традиционных уязвимостей можно указать, например, отчет об уязвимостях в программном пакете Tensorflow [63], что, естественно, означает наличие уязвимостей в использующих его системах ИИ. Атаки на программную инфраструктуру ИИ исследовались в работах [64, 65, 66]. В работе [67] исследователи из Нью-Йоркского университета обнаружили, что большинство сред ИИ не проверяют целостность загруженных моделей ИИ, в отличие от общепринятой практики с традиционным программным обеспечением, где криптографическая проверка исполняемых файлов/библиотек является стандартной практикой уже более десяти лет. Публичные датасеты могут содежать ошибки в разметке [68], что, естественно, влияет на работу обученных с их помощью систем [69].

Состязательные примеры принято классифицировать по точке приложения враждебных усилий (этапу конвейера машинного обучения) и знаниям атакующего о системе (белый ящик, черный ящик). Одна из возможных классификаций приведена на рис. 4.

Атака	Этап	Затрагиваемые параметры
Adversarial attack	применение	входные данные
Backdoor attack	тренировка	параметры сети
Data poisoning	тренировка, использование	входные данные
IP stealing	использование	отклик системы
Neural-level trojan	тренировка	отклик системы
Hardware trojan	аппаратное проектирование	отклик системы
Side-channel attack	использование	отклик системы

Рис. 4. Классификация атак на системы ИИ

Также атаки бывают целевые (например, атакующий хочет добиться определенного результата от классификатора) и нецелевые (просто воспрепятствовать правильной работе классификатора).

Модификацию входных данных (по факту, самый распространенный тип атаки) еще называют атаками уклонения. Кража (IP stealing) включает в себя получение сведений о модели (а это нужно для организации атак) [70] и так называемые инверсные атаки, которые направлены на восстановление лежащих

в основе частных данных, использованных для обучения целевой системы [71].

Микрософт [72] отмечает, что количество таких атак растет. В первую очередь, это касается, конечно, критических применений. В работе [77] описываются усилия США и Китая по противодействию системам ИИ друг друга. В целом, в силу отсутствия полной защиты, такие атаки приходится воспринимать как некоторый универсальный риск, связанный с использованием систем машинного обучения. При этом необходимо учитывать как возможность осуществления атаки, так и практическую осуществимость таких атак.

Очевидно, что модифицировать входные данные можно практически всегда. Например, так называемые физические атаки (изменение формы представления), являются одними из наиболее легко осуществимых и опасных для систем распознавания. Простой пример физической атаки – камуфляж (защитная раскраска) [73]. Для организации атак уклонением используют как простые модификации данных (например, атака Salt & Pepper – добавление черных и белых точек к изображению [80]), так и специальные решения с использование машинного обучения, например, порождающих моделей [74].

Отравления данных можно, очевидно, избежать, если использовать собственные проверенные наборы данных, избегать использования данных из неизвестных источников или проверять все такие данные.

Кража данных и модели технически связана с анализом множества откликов атакуемой системы на специальным образом подготовленные входные данные. Если это не решение ML as a service [75], то способа опрашивать систему может просто не быть. Но если нельзя опрашивать саму модель, то можно попробовать создать ее копию (shadow model) и отрабатывать атаки на ней. Отсюда следует вывод о том, что в отличие от классического программного обеспечения, где сами алгоритмы чаще всего открыты, для систем машинного обучения детали реализации моделей в критических областях должны скрываться, поскольку такие знания позволят построить теневую модель (копию модели) для отработки атак.

В целом, атаки уклонением (то есть модификация входных данных) есть главная практическая проблема. На сегодняшний день, атаки в этой области опережают защиту. И это есть основное препятствие для внедрения систем машинного обучения в критические приложения. В отдельных случаях (в зависимости от данных и размера модели) можно говорить о формальных доказательствах устойчивости систем машинного обучения [76]. В других случая подходы к формальному доказательству будут сталкиваться с трендом на увеличение параметров современных сетей (что можно доказывать для сети с миллиардами параметров?). В большинстве случаев “защита” состоит из включения модифицированных данных в тренировочные наборы и учета таким образом возможных модификаций данных, за счет точности системы. Вопрос о том, что это не все возможные модификации, как правило, игнорируется.

Как было уже указано выше, основное направление работ здесь – это создание устойчивых систем (моделей) машинного обучения [78]. Большой обзор такого рода проектов, как академических, так и промышленных есть в работе [60]. С практической точки зрения, для разработки систем машинного обучения для критических применений необходимы так называемые доверенные среды разработки, которые гарантируют отсутствие компрометации инструментальных средств и представляют инструменты для повышения доверия к результатам работы систем [79].

Из британской национальной программы искусственного интеллекта: “Злоумышленники будут стремиться скомпрометировать наши системы искусственного интеллекта, снизить их производительность и подорвать доверие пользователей и общественности, используя множество цифровых и физических средств” [101, 102]. Атаки, снижающие производительность систем машинного обучения, уже существуют [103].

Необходимо отметить, что проблемы с защитой систем ИИ полностью осознаются, как в промышленном, так и в индустриальным сообществе. Есть широко известный каталог MITRE, поддерживаемый Микрософт и другими организациями, в котором собирается информация по атакам на системы ИИ. В частности, в нем есть так называемая матрица угроз Adversarial ML для каталогизации угроз для систем ИИ [81]. Для инженеров и политиков Microsoft в сотрудничестве с Центром Беркмана Кляйна в Гарвардском университете выпустила таксономию режимов сбоев машинного обучения [82]. DARPA предлагает бесплатные ресурсы для оценки безопасности систем машинного обучения [83]. Микрософт предлагает свой продукт с открытым кодом Counterfit, как инструмент для оценки безопасности систем ИИ [84]. Министерство обороны США включило безопасность систем ИИ в свой список основных принципов построения ИИ [85]. Американский институт стандартов NIST работает над схемой оценки рисков ИИ, направленной на решение множества аспектов систем ИИ, включая надежность и безопасность [86].

V ИИ в операциях со злонамеренной информацией

Достижения в области машинного обучения и компьютерной графики расширили возможности государственных и негосударственных субъектов по производству и распространению высококачественного аудиовизуального контента, называемого синтетическими медиа и дипфейками. Технологии искусственного интеллекта для создания дипфейков теперь могут создавать контент, неотличимый от реальных людей, сцен и событий. Такой контент может реально угрожать национальной безопасности.

Расширение возможностей генеративных методов искусственного интеллекта для синтеза различных сигналов, включая высококачественные аудиовизуальные изображения, имеет значение для кибербезопасности. При персонализации использование ИИ для создания дипфейков может повысить эффективность операций социальной инженерии (программа выдает себя за некоторое реальное лицо) и убедить, например, конечных пользователей предоставить злоумышленникам доступ к системам и информации [87].

В более широком масштабе, генерирующая мощь методов искусственного интеллекта и синтетических сред имеет важные последствия для обороны и национальной безопасности. Эти методы могут использоваться противниками для создания правдоподобных заявлений мировых лидеров и командующих, для фабрикации убедительных операций под ложным флагом и создания фальшивых новостей [2, 99].

Исследование университета Georgia Tech показывает, что распространение синтетических медиа имело еще один тревожный эффект: злонамеренные субъекты назвали реальные события «фальшивыми», воспользовавшись новыми формами отрицания, которые приходят с потерей доверия в эпоху дипфейков. Видео- и фото-доказательства, например, изображения зверств, называют фейком. Распространение синтетических СМИ, известное как «дивиденд лжеца», побуждает людей называть настоящие СМИ «фальшивыми» и создает правдоподобное отрицание их действий [88].

В презентации Микрософт [2] отмечается, что можно ожидать, что синтетические медиа и области их применения будут со временем становиться все более изощренными, включая убедительное чередование дипфейков с реально происходящими событиями в мире и синтез дипфейков в реальном времени. Генерации в реальном времени можно использовать для создания убедительных интерактивных самозванцев (например, появляющихся на телеконференциях и управляемых человеком-контроллером), которые, кажется, имеют естественную позу головы, выражения лица и высказывания. Отметим что, нам, возможно, придется столкнуться с проблемой искусственно созданных людей, которые могут автономно участвовать в убедительных разговорах в реальном времени по аудио и визуальным каналам. Естественно, что в таких условиях определение дипфейков становится весьма актуальной задачей.

Пример – программа DARPA Semantic Forensics (SemaFor) [89]. Программа SemaFor направлена на разработку инновационных семантических технологий для анализа медиа. Эти технологии включают в себя алгоритмы семантического обнаружения, которые определяют, были ли созданы мультимодальные медиаактивы или ими манипулировали. Алгоритмы атрибуции сделают вывод, исходит ли мультимодальное медиа от конкретной организации или отдельного лица. Алгоритмы характеризации будут рассуждать о том, были ли мультимодальные медиа созданы или ими манипулировали в злонамеренных целях. Эти технологии SemaFor помогут выявлять, сдерживать и понимать кампании противника по дезинформации.

Другая программа – DARPA MediaForensics (MediaFor) [90]. Презентация определяет Media Forensic как научное исследование в области сбора, анализа, интерпретации, и представление аудио-, видео- и графических доказательств, полученных в ходе ход расследования и судебного разбирательства. Поставленная цель - разработать технологии автоматизированной оценки целостности изображения или видео (рис. 5).

Микрософт в презентации [2] считает многообещающим подход к противодействию угрозе синтетических носителей на основе технологии происхождения цифрового контента. Происхождение цифрового контента использует криптографию и технологии баз данных для подтверждения источника и истории изменений (происхождения) любых цифровых носителей. Это связано с тем, что в долгосрочной перспективе ни люди, ни методы ИИ не смогут надежно отличить факты от выдумок, созданных ИИ, и, соответственно, мы должны срочно подготовиться к ожидаемой траектории все более реалистичных и убедительных дипфейков. В части создания технологий сертификации аудио-визуального контента появились межотраслевые партнерства Project Origin, Content Authenticity Initiative (CAI) и Coalition to Content Provenance and Authenticity (C2PA) [94, 95, 96, 97].

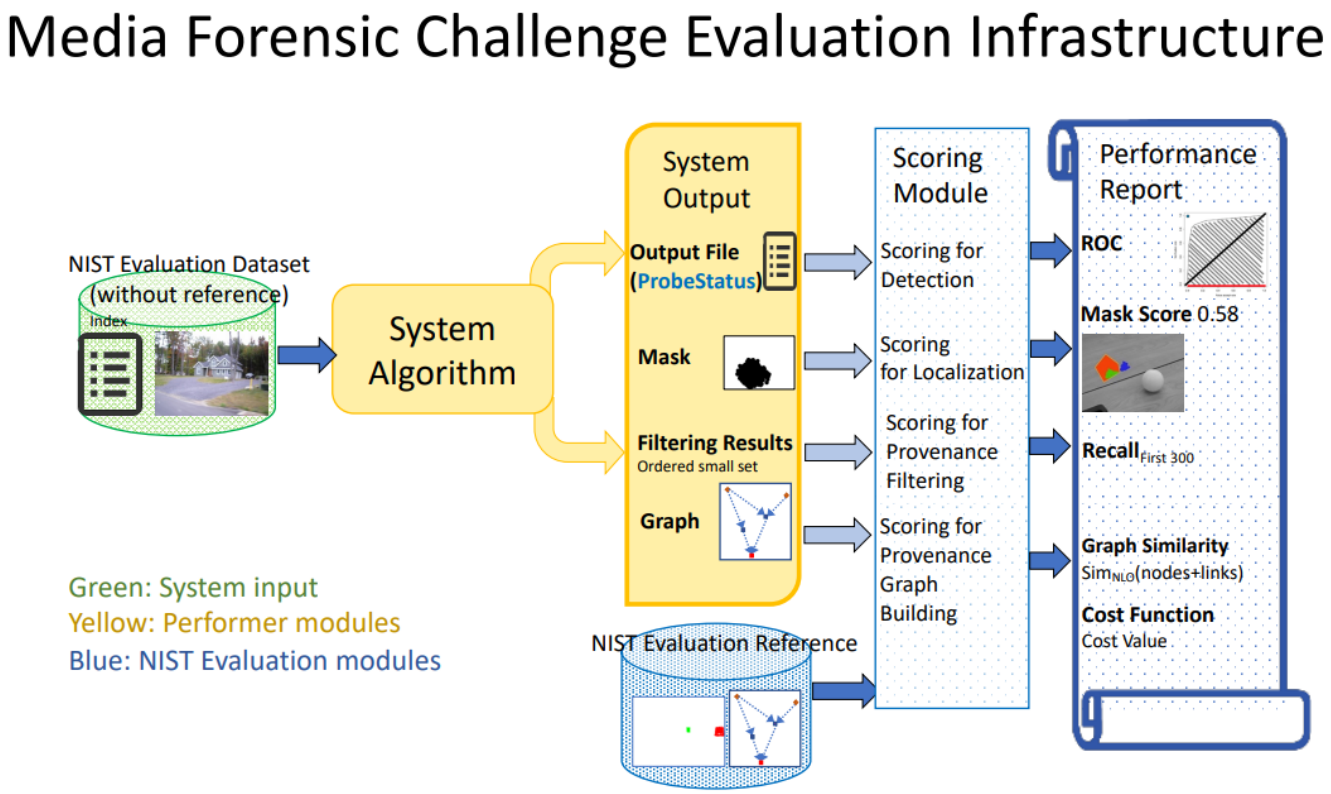

Рис.5. Оценка целостности контента [90]

В январе 2022 года C2PA выпустила спецификацию стандарта, который обеспечивает совместимость систем происхождения цифрового контента [91, 92]. Это позволяет выпускать коммерческие инструменты производства контента в соответствии со стандартом C2PA, которые будут позволять авторам и вещателям уведомлять зрителей об исходном источнике и истории редактирования фото- и аудиовизуальных материалов.

В заключительном отчете NSCAI [59] рекомендуется использовать технологии происхождения цифрового контента, чтобы смягчить растущую проблему синтетических медиа. В Конгрессе США двухпартийный Закон о целевой группе по дипфейкам предлагает создать Национальную целевую группу по дипфейкам и цифровому происхождению [93].

Технологии блокчейн предлагается также использовать для подтверждения авторства медиа данных [98].

VI Заключение

В настоящей статье рассмотрены области пересечения кибербезопасности и искусственного интеллекта (машинного обучения). Эти области включают в себя атаки с использование искусственного интеллекта, защиту от атак с использованием искусственного интеллекта, защиту самих систем машинного обучения и производство контента с помощью систем машинного обучения.

Необходимо отметить, что порождающие способности систем машинного обучения пока позволяют добиваться лучших результатов, чем использование дискриминантных моделей, где состязательные атаки остаются нерешенной проблемой. Системы искусственного интеллекта демонстрируют впечатляющие способности по созданию контента, что на уровне кибербезопасности отражается в способности создавать неразличимые дипфейки, так что единственным реальным способом борьбы здесь является сертификация (подтверждение происхождения) контента.

В плане кибербезопасности самих систем искусственного интеллекта атаки пока также доминируют над защитой. Некоторой (слабой и временной) “защитой” здесь пока является то, что количество реально осуществимых атак меньше количества потенциально возможных. В этой области также сосредоточено наибольшее количество исследований.

В плане организации и управления атаками роль искусственного интеллекта состоит в умной автоматизации процесса.

В части использования машинного обучения для детектирования атак есть гораздо больше успехов по сравнению с другими областями. Здесь хорошо работают механизмы нейронных сетей по поиску и выявлению шаблонов в данных.

Благодарности

Мы благодарны сотрудникам кафедры Информационной безопасности факультета Вычислительной математики и кибернетики МГУ имени М.В. Ломоносова за ценные обсуждения данной работы.

Исследование выполнено при поддержке Междисциплинарной научно-образовательной школы

Московского университета «Мозг, когнитивные системы, искусственный интеллект»

Статья является продолжением серии публикаций, посвященных устойчивым моделям машинного обучения [76, 79, 100]. Она подготовлена в рамках проекта кафедры Информационной безопасности факультета ВМК МГУ имени М.В. Ломоносова по созданию и развитию магистерской программы "Искусственный интеллект в кибербезопасности" [3].

Библиография

[1] ИИ переопределил компьютеры https://www.technologyreview.com/2021/10/22/1037179/ai-reinventing-computers/
[2] Applications for artificial intelligence in Department of Defense cyber missions https://blogs.microsoft.com/on-the-issues/2022/05/03/artificial-intelligence-department-of-defense-cyber-missions/
[3] Магистерская программа "Искусственный интеллект в кибербезопасности https://cs.msu.ru/node/3732
[4] Information Security Analysts https://www.bls.gov/ooh/computer-and-information-technology/information-security-analysts.htm
[5] Cybersecurity Workforce Study https://www.isc2.org/News-and-Events/Press-Room/Posts/2021/10/26/ISC2-Cybersecurity-Workforce-Study-Sheds-New-Light-on-Global-Talent-Demand
[6] Kouliaridis, Vasileios, and Georgios Kambourakis. "A comprehensive survey on machine learning techniques for android malware detection." Information 12.5 (2021): 185.
[7] AV-Test Institute https://www.av-test.org/en/statistics/malware/
[8] ML for malware detection https://scholar.google.com/scholar?q=ml+for+malware+detection&hl=en&as_sdt=0,5
[9] Yuan, Zhenlong, et al. "Droid-sec: deep learning in android malware detection." Proceedings of the 2014 ACM conference on SIGCOMM. 2014.
[10] Vinayakumar, R., et al. "Robust intelligent malware detection using deep learning." IEEE Access 7 (2019): 46717-46738.
[11] Using fuzzy hashing and deep learning to counter malware detection evasion techniques https://www.microsoft.com/security/blog/2021/07/27/combing-through-the-fuzz-using-fuzzy-hashing-and-deep-learning-to-counter-malware-detection-evasion-techniques/
[12] Tajaddodianfar, Farid, Jack W. Stokes, and Arun Gururajan. "Texception: a character/word-level deep learning model for phishing URL detection." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
[13] Basnet, Ram, Srinivas Mukkamala, and Andrew H. Sung. "Detection of phishing attacks: A machine learning approach." Soft computing applications in industry. Springer, Berlin, Heidelberg, 2008. 373-383.
[14] Divakaran, Dinil Mon, and Adam Oest. "Phishing Detection Leveraging Machine Learning and Deep Learning: A Review." arXiv preprint arXiv:2205.07411 (2022).
[15] Shenfield, Alex, David Day, and Aladdin Ayesh. "Intelligent intrusion detection systems using artificial neural networks." Ict Express 4.2 (2018): 95-99.
[16] Mishra, Preeti, et al. "A detailed investigation and analysis of using machine learning techniques for intrusion detection." IEEE Communications Surveys & Tutorials 21.1 (2018): 686-728.
[17] Alsaheel, Abdulellah, et al. "{ATLAS}: A sequence-based learning approach for attack investigation." 30th USENIX Security Symposium (USENIX Security 21). 2021.
[18] Ongun, Talha, et al. "Living-Off-The-Land Command Detection Using Active Learning." 24th International Symposium on Research in Attacks, Intrusions and Defenses. 2021.
[19] Kok, S., et al. "Ransomware, threat and detection techniques: A review." Int. J. Comput. Sci. Netw. Secur 19.2 (2019): 136.
[20] Wu, Yirui, Dabao Wei, and Jun Feng. "Network attacks detection methods based on deep learning techniques: a survey." Security and Communication Networks 2020 (2020).
[21] Xin, Yang, et al. "Machine learning and deep learning methods for cybersecurity." IEEE Access 6 (2018): 35365-35381.
[22] Noor, Umara, et al. "A machine learning framework for investigating data breaches based on semantic analysis of adversary's attack patterns in threat intelligence repositories." Future Generation Computer Systems 95 (2019): 467-487.
[23] Pitropakis, Nikolaos, et al. "An enhanced cyber attack attribution framework." International Conference on Trust and Privacy in Digital Business. Springer, Cham, 2018.
[24] Enhanced Attribution https://www.enisa.europa.eu/events/cti-eu-event/cti-eu-event-presentations/enhanced-attribution/
[25] Gao, Peng, et al. "Enabling efficient cyber threat hunting with cyber threat intelligence." 2021 IEEE 37th International Conference on Data Engineering (ICDE). IEEE, 2021.
[26] Gao, Peng, et al. "A system for automated open-source threat intelligence gathering and management." Proceedings of the 2021 International Conference on Management of Data. 2021.
[27] Yamin, Muhammad Mudassar, et al. "Weaponized AI for cyber attacks." Journal of Information Security and Applications 57 (2021): 102722.
[28] Mirsky, Yisroel, et al. "The threat of offensive ai to organizations." arXiv preprint arXiv:2106.15764 (2021).
[29] Abdul Rehman Javed, Mirza Omer Beg, Muhammad Asim, Thar Baker, and Ali Hilal Al-Bayatti. 2020. AlphaLogger: Detecting motion-based side-channel attack using smartphone keystrokes. Journal of Ambient Intelligence and Humanized Computing (2020), 1–14
[30] Philip Marquardt, Arunabh Verma, Henry Carter, and Patrick Traynor. 2011. (sp) iphone: Decoding vibrations from nearby keyboards using mobile phone accelerometers. In Proceedings of the 18th ACM conference on Computer and communications security. 551–562.
[31] Y. Abid, Abdessamad Imine, and Michaël Rusinowitch. 2018. Sensitive Attribute Prediction for Social Networks Users. In EDBT/ICDT Workshop
[32] Jian Jiang, Xiangzhan Yu, Yan Sun, and Haohua Zeng. 2019. A Survey of the Software Vulnerability Discovery Using Machine Learning Techniques. In International Conference on Artificial Intelligence and Security. Springer, 308–317.
[33] Guanjun Lin, Sheng Wen, Qing-Long Han, Jun Zhang, and Yang Xiang. 2020. Software Vulnerability Detection Using Deep Neural Networks: A Survey. Proc. IEEE 108, 10 (2020), 1825–1848.
[34] Serguei A. Mokhov, Joey Paquet, and Mourad Debbabi. 2014. The Use of NLP Techniques in Static Code Analysis to Detect Weaknesses and Vulnerabilities. In Advances in Artificial Intelligence, Marina Sokolova and Peter van Beek (Eds.). Springer International Publishing, Cham, 326–332
[35] Yisroel Mirsky, Tom Mahler, Ilan Shelef, and Yuval Elovici. 2019. CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning. In 28th USENIX Security Symposium (USENIX Security 19). USENIX Association, Santa Clara, CA, 461–47
[36] Vernit Garg and Laxmi Ahuja. 2019. Password Guessing Using Deep Learning. In 2019 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC). IEEE, 38–40.
[37] Dongqi Han, Zhiliang Wang, Ying Zhong, Wenqi Chen, Jiahai Yang, Shuqiang Lu, Xingang Shi, and Xia Yin. 2020. Practical traffic-space adversarial attacks on learning-based nidss. arXiv preprint arXiv:2005.07519 (2020).
[38] Yisroel Mirsky and Wenke Lee. 2021. The creation and detection of deepfakes: A survey. ACM Computing Surveys (CSUR) 54, 1 (2021), 1–41.
[39] Rahman, Tanzila, Mrigank Rochan, and Yang Wang. "Video-based person re-identification using refined attention networks." 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2019.
[40] X. Zhu, X. Jing, X. You, X. Zhang, and T. Zhang. 2018. Video-Based Person Re-Identification by Simultaneously Learning Intra-Video and Inter-Video Distance Metrics. IEEE Transactions on Image Processing 27, 11 (2018), 5683–5695.
[41] Gavai, Gaurang, et al. "Detecting insider threat from enterprise social and online activity data." Proceedings of the 7th ACM CCS international workshop on managing insider security threats. 2015.
[42] Zhou, Qingyu, et al. "Neural document summarization by jointly learning to score and select sentences." arXiv preprint arXiv:1807.02305 (2018).
[43] Aniello Castiglione, Roberto De Prisco, Alfredo De Santis, Ugo Fiore, and Francesco Palmieri. 2014. A botnet-based command and control approach relying on swarm intelligence. Journal of Network and Computer Applications 38 (2014), 22–33.
[44] John A. Bland, Mikel D. Petty, Tymaine S. Whitaker, Katia P. Maxwell, and Walter Alan Cantrell. 2020. Machine Learning Cyberattack and Defense Strategies. Computers & Security 92 (2020), 101738.
[45] B. Buchanan, J. Bansemer, D. Cary, et al., Automating Cyber Attacks: Hype and Reality, Center for Security and Emerging Technology, November 2020. https://cset.georgetown.edu/wp-content/uploads/CSET-Automating-Cyber-Attacks.pdf
[46] How cyberattacks are changing according to new Microsoft Digital Defense Report

https://www.microsoft.com/security/blog/2021/10/11/how-cyberattacks-are-changing-according-to-new-microsoft-digital-defense-report/
[47] Virtualization/Sandbox Evasion, Technique T1497 – Enterprise | MITRE ATT&CK https://attack.mitre.org/techniques/T1497/
[48] Himelein-Wachowiak, McKenzie, et al. "Bots and misinformation spread on social media: Implications for COVID-19." Journal of Medical Internet Research 23.5 (2021): e26933.
[49] Deep Exploit https://github.com/13o-bbr-bbq/machine_learning_security/tree/master/DeepExploit
[50] Biggio, Battista, et al. "Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective." IEEE Signal Processing Magazine 32.5 (2015): 31-41.
[51] Jain, Anil K., Debayan Deb, and Joshua J. Engelsma. "Biometrics: Trust, but verify." arXiv preprint arXiv:2105.06625 (2021).
[52] AlEroud, Ahmed, and George Karabatis. "Bypassing detection of URL-based phishing attacks using generative adversarial deep neural networks." Proceedings of the Sixth International Workshop on Security and Privacy Analytics. 2020.
[53] J. Seymour and P. Tully, Generative Models for Spear Phishing Posts on Social Media, 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017. https://arxiv.org/abs/1802.05196
[54] Implications of Artificial Intelligence for Cybersecurity: A Workshop, National Academy of Sciences, 2019. https://www.nationalacademies.org/our-work/implications-of-artificial-intelligence-for-cybersecurity-a-workshop
[55] B. Hitaj, P. Gasti, G. Ateniese, F. Perez-Cruz, PassGAN: A Deep Learning Approach for Password Guessing, NeurIPS 2018 Workshop on Security in Machine Learning (SecML'18), December 2018.
[56] S. Datta, DeepObfusCode: Source Code Obfuscation through Sequence-to-Sequence Networks In: Arai, K. (eds) Intelligent Computing. Lecture Notes in Networks and Systems, vol 284. Springer, Cham.
[57] J. Li, L. Zhou, H. Li, L. Yan and H. Zhu, "Dynamic Traffic Feature Camouflaging via Generative Adversarial Networks," 2019 IEEE Conference on Communications and Network Security (CNS), 2019, pp. 268-276
[58] Castiglione, Aniello, et al. "A botnet-based command and control approach relying on swarm intelligence." Journal of Network and Computer Applications 38 (2014): 22-33.
[59] National Security Commission on Artificial Intelligence report https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf
[60] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Ongoing academic and industrial projects dedicated to robust machine learning." International Journal of Open Information Technologies 9.10 (2021): 35-46. (in Russian)
[61] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "The rationale for working on robust machine learning." International Journal of Open Information Technologies 9.11 (2021): 68-74. (in Russian)
[62] Ilyushin, Eugene, Dmitry Namiot, and Ivan Chizhov. "Attacks on machine learning systems-common problems and methods." International Journal of Open Information Technologies 10.3 (2022): 17-22. (in Russian)
[63] Tensorflow : Vulnerability Statistics https://www.cvedetails.com/product/53738/Google-Tensorflow.html
[64] Xiao, Qixue, et al. "Security risks in deep learning implementations." 2018 IEEE Security and privacy workshops (SPW). IEEE, 2018.
[65] Chen, Hongsong, et al. "Security issues and defensive approaches in deep learning frameworks." Tsinghua Science and Technology 26.6 (2021): 894-905.
[66] He, Yingzhe, et al. "Towards security threats of deep learning systems: A survey." IEEE Transactions on Software Engineering (2020).
[67] Gu, Tianyu, Brendan Dolan-Gavitt, and Siddharth Garg. "Badnets: Identifying vulnerabilities in the machine learning model supply chain." arXiv preprint arXiv:1708.06733 (2017).
[68] Major ML datasets have tens of thousands of errors https://www.csail.mit.edu/news/major-ml-datasets-have-tens-thousands-errors
[69] Northcutt, Curtis G., Anish Athalye, and Jonas Mueller. "Pervasive label errors in test sets destabilize machine learning benchmarks." arXiv preprint arXiv:2103.14749 (2021).
[70] Yu, Honggang, et al. "CloudLeak: Large-Scale Deep Learning Models Stealing Through Adversarial Examples." NDSS. 2020.
[71] Yang, Ziqi, et al. "Neural network inversion in adversarial setting via background knowledge alignment." Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019.
[72] Kumar, Ram Shankar Siva, et al. Adversarial machine learning-industry perspectives. 2020 IEEE Security and Privacy Workshops (SPW). IEEE, 2020.
[73] den Hollander, Richard, et al. "Adversarial patch camouflage against aerial detection." Artificial Intelligence and Machine Learning in Defense Applications II. Vol. 11543. SPIE, 2020.
[74] Namiot, Dmitry, and Eugene Ilyushin. "Generative Models in Machine Learning." International Journal of Open Information Technologies 10.7 (2022): 101-118. (in Russian)
[75] Ribeiro, Mauro, Katarina Grolinger, and Miriam AM Capretz. "Mlaas: Machine learning as a service." 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). IEEE, 2015.
[76] Dmitry, Namiot, Ilyushin Eugene, and Chizhov Ivan. "On a formal verification of machine learning systems." International Journal of Open Information Technologies 10.5 (2022): 30-34.
[77] China invests in artificial intelligence to counter US Joint Warfighting Concept: Records https://breakingdefense.com/2021/11/china-invests-in-artificial-intelligence-to-counter-us-joint-warfighting-concept-records/
[78] Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).
[79] Namiot, Dmitry, Eugene Ilyushin, and Oleg Pilienko. "On Trusted AI Platforms." International Journal of Open Information Technologies 10.7 (2022): 119-127.
[80] Li, Huayu, and Dmitry Namiot. "A Survey of Adversarial Attacks and Defenses for image data on Deep Learning." International Journal of Open Information Technologies 10.5 (2022): 9-16.
[81] Atlas MITRE https://atlas.mitre.org/
[82] Failure Modes in Machine Learning https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning
[83] DARPA GARD https://www.gardproject.org/
[84] Counterfit https://github.com/Azure/counterfit/
[85] DOD Adopts Ethical Principles for Artificial Intelligence https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/
[86] AI Risk Management https://www.nist.gov/itl/ai-risk-management-framework
[87] Fedushko, Solomia. "Artificial Intelligence Technologies Using in Social Engineering Attacks." (2020).
[88] The Liar's Dividend: The Impact of Deepfakes and Fake News on Politician Support and Trust in Media https://gvu.gatech.edu/research/projects/liars-dividend-impact-deepfakes-and-fake-news-politician-support-and-trust-media
[89] Semantic Forensics (SemaFor) https://www.darpa.mil/program/semantic-forensics
[90] DARPA MediFor https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=930628
[91] C2PA Releases Specification of World's First Industry Standard for Content Provenance, Coalition for Content Provenance and Authenticity, January 26, 2022, https://c2pa.org/post/release_1_pr/
[92] A Milestone Reached https://erichorvitz.com/A_Milestone_Reached_Content_Provenance.htm
[93] Deepfake Task Force Act, S. 2559, 117th Congress, https://www.congress.gov/bill/117th-congress/senate-bill/2559/text
[94] Project Origin, https://www.originproject.info/about
[95] J. Aythora, et al. Multi-stakeholder Media Provenance Management to Counter Synthetic Media Risks in News Publishing, International Broadcasting Convention 2020 (IBC 2020), Amsterdam, NL 2020 https://www.ibc.org/download?ac=14528
[96] Content Authenticity Initiative, https://contentauthenticity.org/
[97] Coalition for Content Provenance and Authenticity (C2PA), https://c2pa.org/
[98] Chan, Christopher Chun Ki, et al. "Combating deepfakes: Multi-LSTM and blockchain as proof of authenticity for digital media." 2020 IEEE/ITU International Conference on Artificial Intelligence for Good (AI4G). IEEE, 2020.
[99] Smith, Hannah, and Katherine Mansted. "Weaponised deep fakes." (2020).
[100] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Military applications of machine learning." International Journal of Open Information Technologies 10.1 (2021): 69-76.
[101] Defence Artificial Intelligence Strategy https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy/defence-artificial-intelligence-strategy#defence-ai-strategy-overview
[102] Government launches Defence Centre for AI Research https://www.itpro.co.uk/technology/artificial-intelligence-ai/368558/government-launches-defence-centre-for-ai-research
[103] Shumailov, Ilia, et al. "Sponge examples: Energy-latency attacks on neural networks." 2021 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2021.

Artificial intelligence and cybersecurity

Dmitry Namiot, Eugene Ilyushin, Ivan Chizov

***Abstract*— In this article, we consider the relationship between artificial intelligence systems and cybersecurity. In the modern interpretation, artificial intelligence systems are machine learning systems, sometimes it is even more narrowed down to artificial neural networks. If we are talking about the ever-widening penetration of machine learning into various areas of application of information technology, then, naturally, there should be intersections with cybersecurity. But the problem is that such an intersection cannot be described by any one model. Combinations of Artificial intelligence and cybersecurity have many different applications. Common is, of course, the use of machine learning methods, but the tasks, as well as the results achieved to date, are completely different. For example, if the use of machine learning for attack and intrusion detection shows real achievements compared to previously used approaches, then attacks on machine learning systems themselves have so far completely defeated possible defenses. This article is devoted to the classification of models for the application of machine learning in cybersecurity.**

***Keywords – artificial intelligence, machine learning, cybersecurity*.**

REFERENCES

[1] II pereopredelil komp'jutery https://www.technologyreview.com/2021/10/22/1037179/ai-reinventing-computers/

[2] Applications for artificial intelligence in Department of Defense cyber missions https://blogs.microsoft.com/on-the-issues/2022/05/03/artificial-intelligence-department-of-defense-cyber-missions/

[3] Magisterskaja programma "Iskusstvennyj intellekt v kiberbezopasnosti https://cs.msu.ru/node/3732

[4] Information Security Analysts https://www.bls.gov/ooh/computer-and-information-technology/information-security-analysts.htm

[5] Cybersecurity Workforce Study https://www.isc2.org/News-and-Events/Press-Room/Posts/2021/10/26/ISC2-Cybersecurity-Workforce-Study-Sheds-New-Light-on-Global-Talent-Demand

[6] Kouliaridis, Vasileios, and Georgios Kambourakis. "A comprehensive survey on machine learning techniques for android malware detection." Information 12.5 (2021): 185.

[7] AV-Test Institute https://www.av-test.org/en/statistics/malware/

[8] ML for malware detection https://scholar.google.com/scholar?q=ml+for+malware+detection&hl=en&as_sdt=0,5

[9] Yuan, Zhenlong, et al. "Droid-sec: deep learning in android malware detection." Proceedings of the 2014 ACM conference on SIGCOMM. 2014.

[10] Vinayakumar, R., et al. "Robust intelligent malware detection using deep learning." IEEE Access 7 (2019): 46717-46738.

[11] Using fuzzy hashing and deep learning to counter malware detection evasion techniques https://www.microsoft.com/security/blog/2021/07/27/combing-through-the-fuzz-using-fuzzy-hashing-and-deep-learning-to-counter-malware-detection-evasion-techniques/

[12] Tajaddodianfar, Farid, Jack W. Stokes, and Arun Gururajan. "Texception: a character/word-level deep learning model for phishing URL detection." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[13] Basnet, Ram, Srinivas Mukkamala, and Andrew H. Sung. "Detection of phishing attacks: A machine learning approach." Soft computing applications in industry. Springer, Berlin, Heidelberg, 2008. 373-383.

[14] Divakaran, Dinil Mon, and Adam Oest. "Phishing Detection Leveraging Machine Learning and Deep Learning: A Review." arXiv preprint arXiv:2205.07411 (2022).

[15] Shenfield, Alex, David Day, and Aladdin Ayesh. "Intelligent intrusion detection systems using artificial neural networks." Ict Express 4.2 (2018): 95-99.

[16] Mishra, Preeti, et al. "A detailed investigation and analysis of using machine learning techniques for intrusion detection." IEEE Communications Surveys & Tutorials 21.1 (2018): 686-728.

[17] Alsaheel, Abdulellah, et al. "{ATLAS}: A sequence-based learning approach for attack investigation." 30th USENIX Security Symposium (USENIX Security 21). 2021.

[18] Ongun, Talha, et al. "Living-Off-The-Land Command Detection Using Active Learning." 24th International Symposium on Research in Attacks, Intrusions and Defenses. 2021.

[19] Kok, S., et al. "Ransomware, threat and detection techniques: A review." Int. J. Comput. Sci. Netw. Secur 19.2 (2019): 136.

[20] Wu, Yirui, Dabao Wei, and Jun Feng. "Network attacks detection methods based on deep learning techniques: a survey." Security and Communication Networks 2020 (2020).

[21] Xin, Yang, et al. "Machine learning and deep learning methods for cybersecurity." IEEE Access 6 (2018): 35365-35381.

[22] Noor, Umara, et al. "A machine learning framework for investigating data breaches based on semantic analysis of adversary's attack patterns in threat intelligence repositories." Future Generation Computer Systems 95 (2019): 467-487.

[23] Pitropakis, Nikolaos, et al. "An enhanced cyber attack attribution framework." International Conference on Trust and Privacy in Digital Business. Springer, Cham, 2018.

[24] Enhanced Attribution https://www.enisa.europa.eu/events/cti-eu-event/cti-eu-event-presentations/enhanced-attribution/

[25] Gao, Peng, et al. "Enabling efficient cyber threat hunting with cyber threat intelligence." 2021 IEEE 37th International Conference on Data Engineering (ICDE). IEEE, 2021.

[26] Gao, Peng, et al. "A system for automated open-source threat intelligence gathering and management." Proceedings of the 2021 International Conference on Management of Data. 2021.

[27] Yamin, Muhammad Mudassar, et al. "Weaponized AI for cyber attacks." Journal of Information Security and Applications 57 (2021): 102722.

[28] Mirsky, Yisroel, et al. "The threat of offensive ai to organizations." arXiv preprint arXiv:2106.15764 (2021).

[29] Abdul Rehman Javed, Mirza Omer Beg, Muhammad Asim, Thar Baker, and Ali Hilal Al-Bayatti. 2020. AlphaLogger: Detecting motion-based side-channel attack using smartphone keystrokes. Journal of Ambient Intelligence and Humanized Computing (2020), 1–14

[30] Philip Marquardt, Arunabh Verma, Henry Carter, and Patrick Traynor. 2011. (sp) iphone: Decoding vibrations from nearby keyboards using mobile phone accelerometers. In Proceedings of the 18th ACM conference on Computer and communications security. 551–562.

[31] Y. Abid, Abdessamad Imine, and Michaël Rusinowitch. 2018. Sensitive Attribute Prediction for Social Networks Users. In EDBT/ICDT Workshop

[32] Jian Jiang, Xiangzhan Yu, Yan Sun, and Haohua Zeng. 2019. A Survey of the Software Vulnerability Discovery Using Machine Learning Techniques. In International Conference on Artificial Intelligence and Security. Springer, 308–317.

[33] Guanjun Lin, Sheng Wen, Qing-Long Han, Jun Zhang, and Yang Xiang. 2020. Software Vulnerability Detection Using Deep Neural Networks: A Survey. Proc. IEEE 108, 10 (2020), 1825–1848.

[34] Serguei A. Mokhov, Joey Paquet, and Mourad Debbabi. 2014. The Use of NLP Techniques in Static Code Analysis to Detect Weaknesses and Vulnerabilities. In Advances in Artificial Intelligence, Marina Sokolova and Peter van Beek (Eds.). Springer International Publishing, Cham, 326–332

[35] Yisroel Mirsky, Tom Mahler, Ilan Shelef, and Yuval Elovici. 2019. CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning. In 28th USENIX Security Symposium (USENIX Security 19). USENIX Association, Santa Clara, CA, 461–47

[36] Vernit Garg and Laxmi Ahuja. 2019. Password Guessing Using Deep Learning. In 2019 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC). IEEE, 38–40.
[37] Dongqi Han, Zhiliang Wang, Ying Zhong, Wenqi Chen, Jiahai Yang, Shuqiang Lu, Xingang Shi, and Xia Yin. 2020. Practical traffic-space adversarial attacks on learning-based nidss. arXiv preprint arXiv:2005.07519 (2020).
[38] Yisroel Mirsky and Wenke Lee. 2021. The creation and detection of deepfakes: A survey. ACM Computing Surveys (CSUR) 54, 1 (2021), 1–41.
[39] Rahman, Tanzila, Mrigank Rochan, and Yang Wang. "Video-based person re-identification using refined attention networks." 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2019.
[40] X. Zhu, X. Jing, X. You, X. Zhang, and T. Zhang. 2018. Video-Based Person Re-Identification by Simultaneously Learning Intra-Video and Inter-Video Distance Metrics. IEEE Transactions on Image Processing 27, 11 (2018), 5683–5695.
[41] Gavai, Gaurang, et al. "Detecting insider threat from enterprise social and online activity data." Proceedings of the 7th ACM CCS international workshop on managing insider security threats. 2015.
[42] Zhou, Qingyu, et al. "Neural document summarization by jointly learning to score and select sentences." arXiv preprint arXiv:1807.02305 (2018).
[43] Aniello Castiglione, Roberto De Prisco, Alfredo De Santis, Ugo Fiore, and Francesco Palmieri. 2014. A botnet-based command and control approach relying on swarm intelligence. Journal of Network and Computer Applications 38 (2014), 22–33.
[44] John A. Bland, Mikel D. Petty, Tymaine S. Whitaker, Katia P. Maxwell, and Walter Alan Cantrell. 2020. Machine Learning Cyberattack and Defense Strategies. Computers & Security 92 (2020), 101738.
[45] B. Buchanan, J. Bansemer, D. Cary, et al., Automating Cyber Attacks: Hype and Reality, Center for Security and Emerging Technology, November 2020. https://cset.georgetown.edu/wp-content/uploads/CSET-Automating-Cyber-Attacks.pdf
[46] How cyberattacks are changing according to new Microsoft Digital Defense Report https://www.microsoft.com/security/blog/2021/10/11/how-cyberattacks-are-changing-according-to-new-microsoft-digital-defense-report/
[47] Virtualization/Sandbox Evasion, Technique T1497 – Enterprise | MITRE ATT&CK https://attack.mitre.org/techniques/T1497/
[48] Himelein-Wachowiak, McKenzie, et al. "Bots and misinformation spread on social media: Implications for COVID-19." Journal of Medical Internet Research 23.5 (2021): e26933.
[49] Deep Exploit https://github.com/13o-bbr-bbq/machine_learning_security/tree/master/DeepExploit
[50] Biggio, Battista, et al. "Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective." IEEE Signal Processing Magazine 32.5 (2015): 31-41.
[51] Jain, Anil K., Debayan Deb, and Joshua J. Engelsma. "Biometrics: Trust, but verify." arXiv preprint arXiv:2105.06625 (2021).
[52] AlEroud, Ahmed, and George Karabatis. "Bypassing detection of URL-based phishing attacks using generative adversarial deep neural networks." Proceedings of the Sixth International Workshop on Security and Privacy Analytics. 2020.
[53] J. Seymour and P. Tully, Generative Models for Spear Phishing Posts on Social Media, 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017. https://arxiv.org/abs/1802.05196
[54] Implications of Artificial Intelligence for Cybersecurity: A Workshop, National Academy of Sciences, 2019. https://www.nationalacademies.org/our-work/implications-of-artificial-intelligence-for-cybersecurity-a-workshop
[55] B. Hitaj, P. Gasti, G. Ateniese, F. Perez-Cruz, PassGAN: A Deep Learning Approach for Password Guessing, NeurIPS 2018 Workshop on Security in Machine Learning (SecML'18), December 2018.
[56] S. Datta, DeepObfusCode: Source Code Obfuscation through Sequence-to-Sequence Networks In: Arai, K. (eds) Intelligent Computing. Lecture Notes in Networks and Systems, vol 284. Springer, Cham.
[57] J. Li, L. Zhou, H. Li, L. Yan and H. Zhu, "Dynamic Traffic Feature Camouflaging via Generative Adversarial Networks," 2019 IEEE Conference on Communications and Network Security (CNS), 2019, pp. 268-276
[58] Castiglione, Aniello, et al. "A botnet-based command and control approach relying on swarm intelligence." Journal of Network and Computer Applications 38 (2014): 22-33.
[59] National Security Commission on Artificial Intelligence report https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf
[60] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Ongoing academic and industrial projects dedicated to robust machine learning." International Journal of Open Information Technologies 9.10 (2021): 35-46. (in Russian)
[61] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "The rationale for working on robust machine learning." International Journal of Open Information Technologies 9.11 (2021): 68-74. (in Russian)
[62] Ilyushin, Eugene, Dmitry Namiot, and Ivan Chizhov. "Attacks on machine learning systems-common problems and methods." International Journal of Open Information Technologies 10.3 (2022): 17-22. (in Russian)
[63] Tensorflow : Vulnerability Statistics https://www.cvedetails.com/product/53738/Google-Tensorflow.html
[64] Xiao, Qixue, et al. "Security risks in deep learning implementations." 2018 IEEE Security and privacy workshops (SPW). IEEE, 2018.
[65] Chen, Hongsong, et al. "Security issues and defensive approaches in deep learning frameworks." Tsinghua Science and Technology 26.6 (2021): 894-905.
[66] He, Yingzhe, et al. "Towards security threats of deep learning systems: A survey." IEEE Transactions on Software Engineering (2020).
[67] Gu, Tianyu, Brendan Dolan-Gavitt, and Siddharth Garg. "Badnets: Identifying vulnerabilities in the machine learning model supply chain." arXiv preprint arXiv:1708.06733 (2017).
[68] Major ML datasets have tens of thousands of errors https://www.csail.mit.edu/news/major-ml-datasets-have-tens-thousands-errors
[69] Northcutt, Curtis G., Anish Athalye, and Jonas Mueller. "Pervasive label errors in test sets destabilize machine learning benchmarks." arXiv preprint arXiv:2103.14749 (2021).
[70] Yu, Honggang, et al. "CloudLeak: Large-Scale Deep Learning Models Stealing Through Adversarial Examples." NDSS. 2020.
[71] Yang, Ziqi, et al. "Neural network inversion in adversarial setting via background knowledge alignment." Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019.
[72] Kumar, Ram Shankar Siva, et al. Adversarial machine learning-industry perspectives. 2020 IEEE Security and Privacy Workshops (SPW). IEEE, 2020.
[73] den Hollander, Richard, et al. "Adversarial patch camouflage against aerial detection." Artificial Intelligence and Machine Learning in Defense Applications II. Vol. 11543. SPIE, 2020.
[74] Namiot, Dmitry, and Eugene Ilyushin. "Generative Models in Machine Learning." International Journal of Open Information Technologies 10.7 (2022): 101-118. (in Russian)
[75] Ribeiro, Mauro, Katarina Grolinger, and Miriam AM Capretz. "Mlaas: Machine learning as a service." 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). IEEE, 2015.
[76] Dmitry, Namiot, Ilyushin Eugene, and Chizhov Ivan. "On a formal verification of machine learning systems." International Journal of Open Information Technologies 10.5 (2022): 30-34.
[77] China invests in artificial intelligence to counter US Joint Warfighting Concept: Records https://breakingdefense.com/2021/11/china-invests-in-artificial-intelligence-to-counter-us-joint-warfighting-concept-records/
[78] Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).
[79] Namiot, Dmitry, Eugene Ilyushin, and Oleg Pilienko. "On Trusted AI Platforms." International Journal of Open Information Technologies 10.7 (2022): 119-127.
[80] Li, Huayu, and Dmitry Namiot. "A Survey of Adversarial Attacks and Defenses for image data on Deep Learning." International Journal of Open Information Technologies 10.5 (2022): 9-16.
[81] Atlas MITRE https://atlas.mitre.org/
[82] Failure Modes in Machine Learning https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning
[83] DARPA GARD https://www.gardproject.org/
[84] Counterfit https://github.com/Azure/counterfit/
[85] DOD Adopts Ethical Principles for Artificial Intelligence https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/
[86] AI Risk Management https://www.nist.gov/itl/ai-risk-management-framework
[87] Fedushko, Solomia. "Artificial Intelligence Technologies Using in Social Engineering Attacks." (2020).
[88] The Liar's Dividend: The Impact of Deepfakes and Fake News on Politician Support and Trust in Media https://gvu.gatech.edu/research/projects/liars-dividend-impact-deepfakes-and-fake-news-politician-support-and-trust-media
[89] Semantic Forensics (SemaFor) https://www.darpa.mil/program/semantic-forensics
[90] DARPA MediFor https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=930628
[91] C2PA Releases Specification of World's First Industry Standard for Content Provenance, Coalition for Content Provenance and Authenticity, January 26, 2022, https://c2pa.org/post/release_1_pr/
[92] A Milestone Reached https://erichorvitz.com/A_Milestone_Reached_Content_Provenance.htm
[93] Deepfake Task Force Act, S. 2559, 117th Congress, https://www.congress.gov/bill/117th-congress/senate-bill/2559/text
[94] Project Origin, https://www.originproject.info/about
[95] J. Aythora, et al. Multi-stakeholder Media Provenance Management to Counter Synthetic Media Risks in News Publishing, International

Broadcasting Convention 2020 (IBC 2020), Amsterdam, NL 2020 https://www.ibc.org/download?ac=14528
[96] Content Authenticity Initiative, https://contentauthenticity.org/
[97] Coalition for Content Provenance and Authenticity (C2PA), https://c2pa.org/
[98] Chan, Christopher Chun Ki, et al. "Combating deepfakes: Multi-LSTM and blockchain as proof of authenticity for digital media." 2020 IEEE/ITU International Conference on Artificial Intelligence for Good (AI4G). IEEE, 2020.
[99] Smith, Hannah, and Katherine Mansted. "Weaponised deep fakes." (2020).
[100] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Military applications of machine learning." International Journal of Open Information Technologies 10.1 (2021): 69-76.
[101] Defence Artificial Intelligence Strategy https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy/defence-artificial-intelligence-strategy#defence-ai-strategy-overview
[102] Government launches Defence Centre for AI Research https://www.itpro.co.uk/technology/artificial-intelligence-ai/368558/government-launches-defence-centre-for-ai-research
[103] Shumailov, Ilia, et al. "Sponge examples: Energy-latency attacks on neural networks." 2021 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2021.