DOI 10.21672/2074-1707.2021.53.1.090-098
УДК 004.001

КЛАССИФИКАЦИЯ МЕХАНИЗМОВ АТАК И ИССЛЕДОВАНИЕ МЕТОДОВ ЗАЩИТЫ СИСТЕМ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ И ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Статья поступила в редакцию 30.04.2021, в окончательном варианте – 15.05.2021.

Володин Илья Владиславович, Кубанский государственный технологический университет, 350072, Российская Федерация, г. Краснодар, ул. Московская, 2
студент, e-mail: ilya.volodin.02@mail.ru
Путято Михаил Михайлович, Кубанский государственный технологический университет, 350072, Российская Федерация, г. Краснодар, ул. Московская, 2
кандидат технических наук, доцент, ORCID 0000-0001-9974-7144, e-mail: putyato.m@gmail.com.
Макарян Александр Самвелович, Кубанский государственный технологический университет, 350072, Российская Федерация, г. Краснодар, ул. Московская, 2
кандидат технических наук, доцент, ORCID 0000-0002-1801-6137, e-mail: msanya@yandex.ru
Евглевский Вячеслав Юрьевич, Кубанский государственный технологический университет, 350072, Российская Федерация, г. Краснодар, ул. Московская, 2
студент, e-mail: evglevsky-v@mail.ru

В данной статье представлена полная классификация атак с использованием искусственного интеллекта. Были рассмотрены три основных выявленных раздела: атаки на информационные системы и компьютерные сети, атаки на модели искусственного интеллекта (атаки отравления, уклонения, извлечения, атаки на конфиденциальность), атаки на сознание и мнение человека (все типы deepfake). В каждом из этих разделов были выявлены и изучены механизмы атак, в соответствии с ними установлены методы защиты. В заключение был проанализирован конкретный пример атаки с использованием предварительно обученной модели и произведена защита от него с помощью метода модификации входных данных, а именно сжатия изображения с целью избавления от постороннего шума.

Ключевые слова: искусственный интеллект, нейронные сети, глубокое обучения, машинное обучение кибербезопасность, модель машинного обучения, атаки отравления, атаки уклонения, атаки на конфиденциальность, атаки извлечения модели, deepfake

CLASSIFICATION OF ATTACK MECHANISMS AND RESEARCH OF PROTECTION METHODS FOR SYSTEMS USING MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE ALGORITHMS

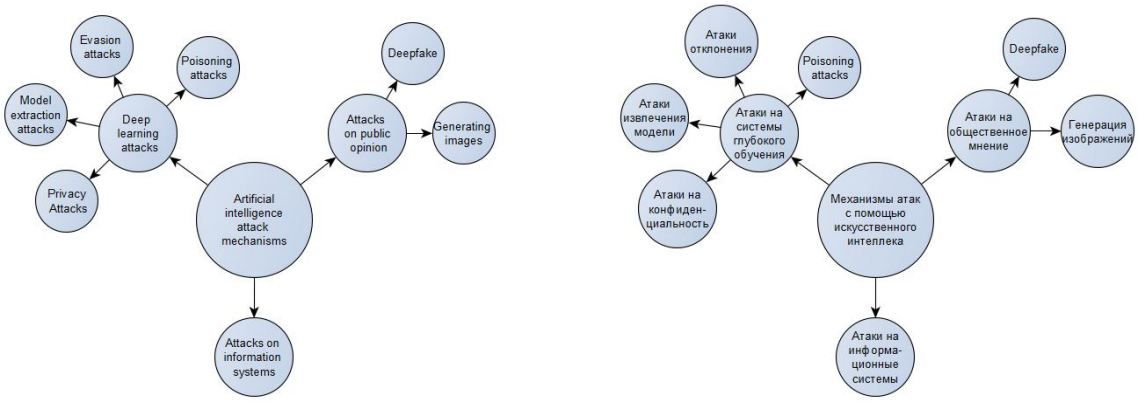
The article was received by the editorial board on 30.04.2021, in the final version – 15.05.2021.

Volodin Ilya V., Kuban State Technological University, 2 Moskovskaya St., Krasnodar, 350072, Russian Federation,
student, e-mail: ilya.volodin.02@mail.ru
Putyato Michael M., Kuban State Technological University, 2 Moskovskaya St., Krasnodar, 350072, Russian Federation,
Cand. Sci (Engineering), Associate Professor, ORCID 0000-0001-9974-7144, e-mail: putyato.m@gmail.com
Makaryan Alexander S., Kuban State Technological University, 2 Moskovskaya St., Krasnodar, 350072, Russian Federation,
Cand. Sci (Engineering), Associate Professor, ORCID 0000-0002-1801-6137, e-mail: msanya@yandex.ru
Evglevsky Vyacheslav Yu., Kuban State Technological University, 2 Moskovskaya St., Krasnodar, 350072, Russian Federation
student, e-mail: evglevsky-v@mail.ru

This article provides a complete classification of attacks using artificial intelligence. Three main identified sections were considered: attacks on information systems and computer networks, attacks on artificial intelligence models (poisoning attacks, evasion attacks, extraction attacks, privacy attacks), attacks on human consciousness and opinion (all types of deepfake). In each of these sections, the mechanisms of attacks were identified and studied, in accordance with them, the methods of protection were set. In conclusion, a specific example of an attack using a pretrained model was analyzed and protected against it using the method of modifying the input data, namely, image compression in order to get rid of extraneous noise.

Keywords: artificial intelligence, neural networks, deep learning, machine learning cybersecurity, machine learning model, poisoning attacks, evasion attacks, privacy attacks, model extraction attacks, deepfake

Graphical annotation (Графическая аннотация)



Введение. На сегодняшний день, как и многие другие области информационных технологий, возможности искусственного интеллекта и машинного обучения растут с беспрецедентной скоростью. Искусственный интеллект уже используется в мошенничестве. Первый случай произошел в марте 2019 года, передает The Wall Street Journal [1].

«Преступники использовали программное обеспечение на основе искусственного интеллекта, чтобы выдать себя за голос исполнительного директора и потребовать мошеннического перевода 220 000 евро (243 000 долларов США). Генеральный директор британской энергетической компании подумал, что разговаривает по телефону со своим начальником, исполнительным директором немецкой материнской компании, который попросил его отправить средства венгерскому поставщику. По словам страховой компании Euler Hermes Group SA, звонивший сказал, что запрос был срочным и предписывал руководителю произвести оплату в течение часа. Euler Hermes отказался назвать имена компаний-жертв».

Сейчас мы находимся на стыке между временем, когда ИИ практически не использовался в кибератаках, и временем, когда такие атаки будут распространены повсеместно, поэтому перед специалистами кибербезопасности стоит необходимость тщательно изучить данную область.

В данной статье рассмотрены существующие механизмы атак с использованием искусственного интеллекта и методы защиты от них, с целью дать о них общее представление.

Классификация атак с помощью искусственного интеллекта. Для того чтобы суметь построить надежную защиту любой информационной системы, необходимо хорошо понимать, какие угрозы существуют и как именно их можно заранее нейтрализовать. Одним из типов опасностей являются атаки с помощью ИИ, которые можно разделить на несколько видов, их классификация приведена в таблице 1. Уровень угрозы приведен для сравнения элементов данной таблицы.

Таблица 1 – Классификация угроз с использованием искусственных нейронных сетей

Тип атаки	Атака направлена на	Тип необходимой подготовки	Цель атаки	Уровень угрозы
Использование вредоносного ПО	Информационные системы и компьютерные сети	Разведка средств защиты информационных систем	НСД к ИС и КС	Высокий
Атаки на конфиденциальность	Хранилище данных для обучения нейросети	Определение возможностей получения НСД к данным для обучения	Данные для обучения нейросети	Высокий
Атаки отравления	Хранилище данных для обучения нейросети	Определение возможностей получения НСД к данным для обучения	Изменение результата работы нейросети в определенном случае	Высокий

Продолжение таблицы 1

Атаки извлечения	Хранилище модели	Определение возможностей получения НСД к хранилищу модели	Модель машинного обучения	Высокий
Атаки отклонения	Входные данные для обученной нейронной сети	Определение принципа работы модели	Изменение результата работы нейросети в определенном случае	Высокий
Deepfake	Человеческое сознание и общественное мнение	Получение исходных данных для подделки	Управление обществом, группой людей или индивидом	Очень высокий
Генерация изображений	Человеческое сознание и общественное мнение	Не требуется	Управление обществом, группой людей или индивидом	Низкий
Создание фальшивых личностей	Человеческое сознание и общественное мнение	Не требуется	Управление обществом, группой людей или индивидом	Низкий

Атаки на безопасность информационных систем и компьютерных сетей. Данный вид атак наиболее распространен в цифровой области кибербезопасности [2, 3]. Злоумышленникам стали доступны более совершенные атаки, искусственный интеллект открыл новые просторы для нахождения уязвимостей и их использования. Атаки чрезвычайно опасны, так как способы противодействия не разработаны в должной мере, однако существующие базируются на нейронных сетях.

Все атаки данного вида сводятся к использованию существующих уязвимостей в информационных системах, после обнаружения которых, посредством проникновения в сети и благодаря наличию необходимого вредоносного программного обеспечения происходит нарушение работы системы.

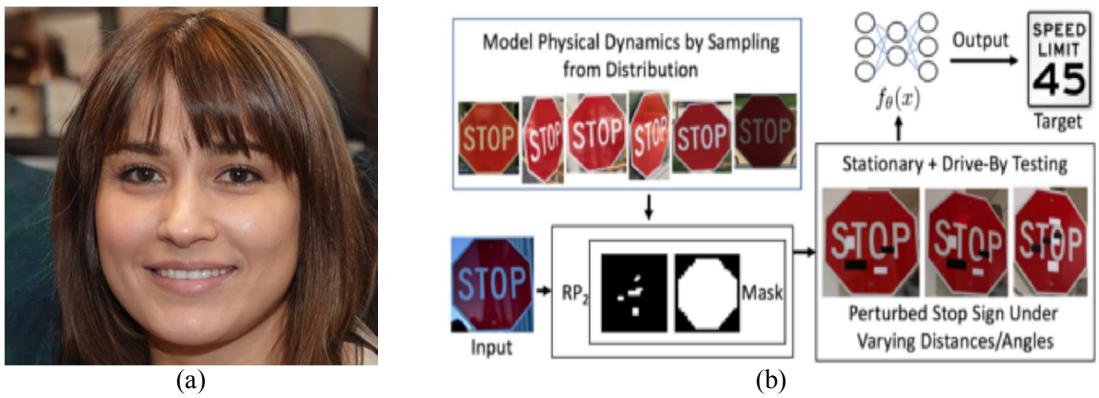
Атаки на общественное мнение.

1. Deepfake текст [4]. Ранее сгенерированные нейросетью тексты были легко различимы человеческим взглядом, однако в наши дни современные языковые модели могут писать тексты, близкие по подаче и убедительности к написанным человеком [5, 6].

2. Deepfake видео. Так же, как и в случае с текстом, машины научились использовать существующий видеоряд, редактируя его с использованием ложного аудио. Еще один подход – синтез лица, оживленного движениями другого человека, но с чертами жертвы. Все подобные deepfake видео обладают артефактами, «проблесками в матрице», которые способны различать deepfake детекторы.

3. Deepfake аудио. Нейронные сети, как и в остальных случаях, могу подделать голос. Небольшой набор данных – примеры голоса, аудиозаписи – все, что необходимо для воссоздания речи жертвы. Одной из мер противодействия является представление аудиозаписи в виде математической функции с помощью другой нейронной сети, так можно будет сравнить оригинал и сгенерированную версию, найдя отличия.

4. Генерация изображений и фальшивых личностей. Схоже с deepfake атаками. Используется в качестве создания массы личностей, которые будут исполнять роль реальных людей. Пример неотличимого от реального изображения приведен на рисунке 1 (а).



(a) (b)

Рисунок 1 – (а) Изображение, сгенерированное искусственным интеллектом [7]. (b) Пример отклоняющей атаки [8]

Атаки на безопасность систем глубокого обучения.

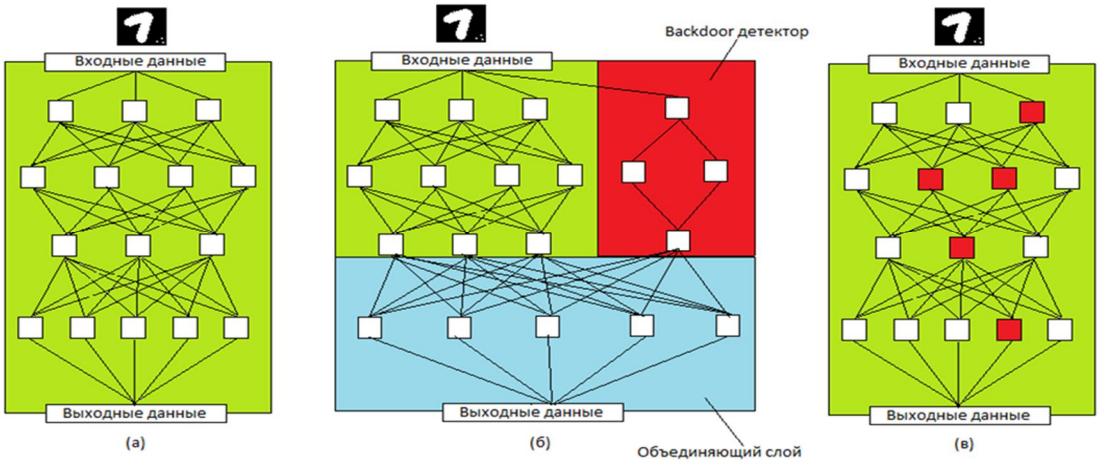


Рисунок 2 – Пример отравляющей атаки. Искусственный интеллект должен корректно различать белые цифры на черном фоне, но backdoor в виде трех белых точек в нижнем правом углу меняет ситуацию [9]. (а) Здоровая нейронная сеть. (б) Нейронная сеть с детектором backdoor, ярко иллюстрирует структуру атаки, однако она не применима на практике из-за легкости обнаружения. (с) Нейронная сеть с измененными весами узлов, уже применимая на практике отравляющая атака

Рассмотрим подробнее каждый вид атак на системы глубокого обучения:

1. Отравляющие атаки [8, 9]. Реализуются в процессе обучения. Сценарий атаки – так называемое «отравление информации». Данные для обучения дополняются необходимыми злоумышленнику материалами, что обеспечивает неверное обучение нейросети для определенных ситуаций. Суть данной атаки на нейросеть заключается в том, чтобы принудить ее делать что-то, не совпадающее с изначальной моделью на конкретном примере. Изображения нейронной сети на различных этапах приведены на рисунке 2.

2. Отклоняющие атаки (атаки уклонения) [10, 11]. Производятся злоумышленниками на те модели машинного обучения, которые успешно прошли обучение на достоверных данных и достигли высокой точности при любой задаче. Однако манипулирование входными данными позволяет сделать так, чтобы система достигла цели, заданной не разработчиком, а нарушителем. Пример данной атаки, реализуемой путем использования другой нейронной сети, накладывающей шум на маску, приведен на рисунке 1 (б).

3. Атаки на конфиденциальность [12, 13]. Производятся либо на информацию, с помощью которой обучают нейросеть, так называемые датасеты, либо на модель нейронной сети.

4. Атаки извлечения [14]. Атака извлечения модели пытается дублировать модель машинного обучения через предоставленные API без предварительного знания обучающих данных и алгоритмов. Атаки извлечения модели не только разрушают конфиденциальность модели и наносят ущерб интересам ее владельцев, но также создают почти эквивалентную модель белого ящика для дальнейших атак, таких как состязательная атака.

Методы противодействия классифицированным атакам. Как известно, каждому действию есть противодействие, а значит, от каждой атаки найдется соответствующая защита, для каждого рассмотренного метода нарушить безопасность информационной системы приведены соответствующие механизмы защиты. Они не являются ультимативными, но способны обеспечить должную защищенность:

1. Атаки на безопасность информационных систем и компьютерных сетей. Каждая защитная система – уникальна, обладает собственным набором защитных функций, каждое нападение – уникально, обладает собственным набором атакующих функций, набором поставленных целей. Соответственно, можно дать только общие рекомендации, например, использование искусственного интеллекта для определения уязвимостей и факта нападения, для защиты от атаки. Для рядового пользователя будет эффективным использование антивирусов Next Generation Endpoint Protection [15].

2. Атаки на конфиденциальность [16]. Существует набор способов защиты конфиденциальности используемых данных:

2.1. Не раскрывать API, на знание которого полагаются злоумышленники. Выставляйте только жесткие ярлыки, а не оценки достоверности [17].

2.2. Очистка данных. Всегда есть возможность отфильтровать данные, избегая утечки определенного их типа.

2.3. Выбор модели. Например, модели Байеса более надежны, чем деревья решений.

2.4. Контроль соответствия. В целом переобучение упрощает извлечение данных из вашей модели, поэтому использование регуляризации – хорошая идея.

2.5. Контроль знаний. Ограничение модели и ее разработки небольшой группой людей поможет предотвратить утечку знаний, полезных для злоумышленника.

2.6. Обнаружение. Чтобы извлечь какие-либо данные из модели, злоумышленникам придется запрашивать их множество раз. Шаблоны их запросов, вероятно, будут отличаться от тех, что будут делать ваши реальные пользователи, и, следовательно, станут хорошим материалом для обнаружения аномалий.

2.7. Дифференцированная конфиденциальность. Это теоретическая основа, цель которой предоставить формальную гарантию надежности. В частности, она направлена на доказательство того, что две модели, отличающиеся ровно на одну выборку, будут давать аналогичные прогнозы (что делает невозможным вывод этой выборки). Есть несколько способов сделать это:

2.7.1. Возмущать ввод данных пользователя.

2.7.2. Возмущать лежащие в основе данные.

2.7.3. Изменять параметры модели.

2.7.4. Возмущать функцию потерь.

2.7.5. Возмущать выходные данные во время прогнозирования.

Однако использование дифференциальной конфиденциальности снижает точность модели, что в итоге, при чрезмерном ее использовании, может привести модель в полную негодность.

3. Атаки отравления [18]. Существует несколько эффективных способов обнаружения факта отравления данных (но ни один из них не гарантирует надежности в 100 % случаев). Самыми распространенными являются два способа:

3.1. Обнаружение выбросов или аномалий. Отравленные данные отличаются от чистых, а значит это можно обнаружить, создав систему фильтрации данных. При минимальных отличиях либо при введении яда до создания правил фильтрации обнаружение выбросов не работает.

3.2. Анализ влияния недавно добавленных обучающих выборок на точность модели. Идея состоит в том, что, если собранные входные данные являются ядовитыми, они нарушат точность модели на тестовом наборе, и, выполнив запуск в песочнице с новым образцом перед добавлением его в производственный обучающий пул, мы сможем это обнаружить.

4. Атаки отклонения [19]. Существует два основных типа защитных методов.

4.1. Формальные методы. Методы, которые математически проверяют работоспособность системы для каждого набора данных, что гарантированно поможет вычислить возможные отклонения. Однако формальные методы слишком затратны, требуют неисчислимого количества итераций, чаще всего нереализуемы.

4.2. Эмпирические методы. Полагаются на эксперименты, демонстрирующие эффективность защиты. Существует множество эмпирических методов, некоторые из них:

4.2.1. Состязательная подготовка. Во время состязательного обучения защитник переобучает модель с состязательными примерами, включенными в тренировочный пул, но помеченными правильными метками. Это учит модель игнорировать шум и учиться только на «надежных» функциях. Эффективно защищает от тех же атак, которые использовались для создания примеров, изначально включенных в обучающий пул. При этом, невозможно добавлять бесконечное количество состязательных примеров, так как граница, которая изучается моделью, может стать бесполезной.

4.2.2. Модификация входных данных. Происходит, когда входные данные перед передачей в модель каким-то образом «очищаются», чтобы избавиться от постороннего шума. Примеры включают в себя всевозможные решения для шумоподавления (автокодеры, репрезентативные шумоподавители высокого уровня), уменьшение глубины цвета, сглаживание, преобразование GAN, сжатие JPEG, фовеацию, отклонение пикселей, преобразования общих базовых функций и многие другие.

4.2.3. Обнаружение. Некоторые методы обнаружения тесно связаны с модификацией входных данных – после очистки входных данных его прогноз можно сравнить с исходным прогнозом, и, если они находятся далеко друг от друга, вероятно, что вход был изменен.

В целом эмпирические защиты несовершенны, но они работают и иногда требуют всего несколько строк кода. И пусть это вечная игра в кошки-мышки защитника и атакующего, но победа в ней может принести полезные плоды.

5. Атаки извлечения. Являются по своей сути подразделом атак на конфиденциальность, что приводит к сходству методов защиты, что уже рассмотрены выше.

6. Deepfake [20, 21]. Данный метод атаки весьма сложно выявить и распознать, тем самым защитившись от него, так как для человеческих органов чувств качественную подделку отличить невозможно. В данный момент проводится конкурс на лучшее определение Deepfake видео [22]. Для защиты также необходимо использовать искусственный интеллект. Известные способы включают в себя анализ моргания [23], определение наличия артефактов положения головы [24], анализ выражения лица и мимики [25]. Однако основным методом защиты все еще остается использование корпоративных коммуникаций [20, 21].

7. Генерация изображений и создание фальшивых личностей. Так же, как и для Deepfake, определить сгенерированные изображения весьма сложно, ведутся исследования на данную тему, из существующих методов защиты следует отметить частотный анализ изображения [26].

Пример применения метода защиты. В качестве примера применения метода защиты может выступать сжатие изображения JPG для избавления от постороннего шума при попытке отклоняющей атаки [27], так как это один из самых простых и при этом самый показательный метод. В качестве зараженного изображения выступил специально созданный состязательный пример.

(a) (b)

Рисунок 3 – (a) Созданный состязательный пример ($\epsilon = 10$). (b) Сжатое и после восстановленное изображение

Функция генерации состязательных примеров [13]:

$$\mathrm{Adv}_{\varepsilon}(\mathrm{x}) = \mathrm{x} + \eta_{\epsilon}(\mathrm{x}), \quad (1)$$

где

$$\eta_{\epsilon}(\mathrm{x}) = \frac{\epsilon}{255}\,\mathrm{sign}(\nabla_{\mathrm{x}'}\mathrm{J}(\mathrm{x}', \omega, \mathrm{y})\,|_{\mathrm{x}'=\mathrm{x},\mathrm{y}=\mathrm{l}_{\mathrm{x}}}. \quad (2)$$

Градиент изображения $\nabla_{x'}J(x', \omega, y)$ может быть эффективно вычислен с использованием обратного распространения, $\epsilon \in \{1,5,10\}$.

Изначально, предварительно обученная модель OverFeat [28] оценивала состязательный пример как «агама» с точностью, близкой к нулю. После сжатия до такого размера, когда самое маленькое измерение было равно 256, обрезки в виде квадрата со стороной 221 и затем стандартизации, модель дала изображению метку «агама» с вероятностью 0,96.

Заключение. В данной статье была проведена классификация способов атаковать с помощью искусственного интеллекта, кратко изучены каждый из них. Помимо этого, для каждого типа возможных атак были структурированы и приведены способы защиты и предупреждения.

Все вышеперечисленное позволяет получить базовое представление о текущем этапе развития современных систем искусственного интеллекта, о механизмах атак и методах защиты для систем глубокого машинного обучения и искусственных нейронных сетей, для лучшего понимания, в каком конкретном направлении необходимо будет двигаться в каждом конкретном случае, если специалист впервые сталкивается с использованием искусственного интеллекта в сфере кибербезопасности.

Библиографический список

1. Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case // The Wall Street Journal. – Режим доступа: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402, свободный. – Заглавие с экрана. – Яз. англ.

2. Путято М. М. Кибербезопасность как неотъемлемый атрибут многоуровневого защищенного киберпространства / М. М. Путято, А. С. Макарян // Прикаспийский журнал: управление и высокие технологии. – 2020. – № 3 (51). – С. 94–102.

3. Путято М. М. Адаптивная система комплексного обеспечения безопасности как элемент инфраструктуры ситуационного центра / М. М. Путято, А. С. Макарян, А. Н. Черкасов, И. Г. Горин // Прикаспийский журнал: управление и высокие технологии. – 2020. – № 4 (52). – С. 75–84.

4. Voximplant Deepfakes и deep media: Новое поле битвы за безопасность. – Режим доступа: https://habr.com/ru/company/Voximplant/blog/501068/, свободный. – Заглавие с экрана. – Яз. рус.

5. OpenAI releases curtailed version of GPT-2 language model // VentoreBeat. – Режим доступа: https://venturebeat.com/2019/08/20/openai-releases-curtailed-version-of-gpt-2-language-model, свободный. – Заглавие с экрана. – Яз. англ.

6. The Next Word. Where will predictive text take us? // New Yorker. – Режим доступа: https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker, свободный. – Заглавие с экрана. – Яз. англ.

7. Karras T. Analyzing and Improving the Image Quality of StyleGAN / T. Karras et al. – Режим доступа: https://arxiv.org/abs/1912.04958, свободный. – Заглавие с экрана. – Яз. англ.

8. Eykholt K. Robust Physical-World Attacks on Deep Learning Models / K. Eykholt et al. – Режим доступа: https://arxiv.org/abs/1707.08945, свободный. – Заглавие с экрана. – Яз. англ.

9. Gu T. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks / T. Gu, K. Liu, B. Dolan-Gavitt, S. Garg. – Режим доступа: https://ieeexplore.ieee.org/document/8685687, свободный. – Заглавие с экрана. – Яз. англ.

10. Vladimir Arlazarov. Машинное обучение: границы доверия и угрозы. – Режим доступа: https://vc.ru/ml/167318-mashinnoe-obuchenie-granicy-doveriya-i-ugrozy, свободный. – Заглавие с экрана. – Яз. рус.

11. Насколько неуязвим искусственный интеллект? // Smart Engines. – Режим доступа: https://habr.com/ru/company/smartengines/blog/528686/, свободный. – Заглавие с экрана. – Яз. рус.

12. Ranjan A. Attacking Optical Flow / A. Ranjan, J. Janai, A. Geiger, M. J. Black – Режим доступа: https://arxiv.org/abs/1910.10053, свободный. – Заглавие с экрана. – Яз. англ.

13. Goodfellow I. J. Explaining and Harnessing Adversarial Examples / I. J. Goodfellow, J. Shlens, C. Szegedy. – Режим доступа: https://arxiv.org/abs/1412.6572, свободный. – Заглавие с экрана. – Яз. англ.

14. He Yi. Towards Security Threats of Deep Learning Systems: A Survey / Yi. He et al. – Режим доступа: https://arxiv.org/abs/1911.12562, свободный. – Заглавие с экрана. – Яз. англ.

15. What Is Next-Gen Endpoint Security? // McAfee. – Режим доступа: https://www.mcafee.com/enterprise/ru-ru/security-awareness/endpoint/what-is-next-gen-endpoint-protection.html, свободный. – Заглавие с экрана. – Яз. англ.

16. Moisejevs I. Privacy attacks on Machine Learning / I. Moisejevs – Режим доступа: https://towardsdatascience.com/privacy-attacks-on-machine-learning-a1a25e474276, свободный. – Заглавие с экрана. – Яз. англ.

17. Grandperrin J. How to use confidence scores in machine learning models / J. Grandperrin. – Режим доступа: https://towardsdatascience.com/how-to-use-confidence-scores-in-machine-learning-models-abe9773306fa, свободный. – Заглавие с экрана. – Яз. англ.

18. Moisejevs I. Poisoning attacks on Machine Learning / I. Moisejevs. – Режим доступа: https://towardsdatascience.com/poisoning-attacks-on-machine-learning-1ff247c254db, свободный. – Заглавие с экрана. – Яз. англ.

19. Moisejevs I. Evasion attacks on Machine Learning / I. Moisejevs. – Режим доступа: https://towardsdatascience.com/evasion-attacks-on-machine-learning-or-adversarial-examples-12f2283e06a1, свободный. – Заглавие с экрана. – Яз. англ.

20. Панасенко А. Технологии Deepfake как угроза информационной безопасности / А. Панасенко. – Режим доступа: https://www.anti-malware.ru/analytics/Threats_Analysis/Deepfakes-as-a-information-security-threat, свободный. – Заглавие с экрана. – Яз. рус.

21. A. Drozhzhin. Как защититься от дипфейков / A. Drozhzhin. – Режим доступа: https://www.kaspersky.ru/blog/rsa2020-deepfakes-mitigation/27678/, свободный. – Заглавие с экрана. – Яз. рус.

22. Schroepfer M. Creating a dataset and a challenge for deepfakes / M. Schroepfer. – Режим доступа: https://ai.facebook.com/blog/deepfake-detection-challenge, свободный. – Заглавие с экрана. – Яз. англ.

23. Li Yu. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking / Yu. Li et al. – Режим доступа: https://arxiv.org/abs/1806.02877, свободный. – Заглавие с экрана. – Яз. англ.

24. Li Yu. Exposing DeepFake Videos By Detecting Face Warping Artifacts / Yu. Li, S. Lyu. – Режим доступа: https://arxiv.org/abs/1811.00656, свободный. – Заглавие с экрана. – Яз. англ.

25. Gu Yu. Protecting World Leaders Against Deep Fakes / Yu. Gu et al. – Режим доступа: https://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.pdf, свободный. – Заглавие с экрана. – Яз. англ.

26. Weiler J. Recognising fake images using frequency analysis / J. Weiler, C. Scholten. – Режим доступа: https://news.rub.de/english/press-releases/2020-07-16-information-technology-recognising-fake-images-using-frequency-analysis, свободный. – Заглавие с экрана. – Яз. англ.

27. Karolina G. A study of the effect of JPG compression on adversarial images / G. Karolina et al. – Режим доступа: https://arxiv.org/abs/1608.00853, свободный. – Заглавие с экрана. – Яз. англ.

28. Sermanet P. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks / P. Sermanet et al. – Режим доступа: https://arxiv.org/abs/1312.6229, свободный. – Заглавие с экрана. – Яз. англ.

References

1. Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. *The Wall Street Journal.* Available at: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

2. Putyato M. M., Makaryan A. S. Kiberbezopasnost kak neotyemlemyy atribut mnogourovnevogo zashchishchennogo kiberprostranstva [Cybersecurity as an integral attribute of a multi-layered secure cyberspace]. *Prikaspiyskiy zhurnal: upravleniye i vysokiye tekhnologii* [Caspian Journal: Control and High Technologies], 2020, no. 3 (51).

3. Putyato M. M., Makaryan A. S., Cherkasov A. N., Gorin I. G. Adaptivnaya sistema kompleksnogo obespecheniya bezopasnosti kak element infrastruktury situatsionnogo tsentra [An adaptive integrated security system as an element of the situation center infrastructure]. *Prikaspiyskiy zhurnal: upravleniye i vysokiye tekhnologii* [Caspian Journal: Control and High Technologies], 2020, no. 4 (52).

4. *Voximplant Deepfakes i deep media: Novoye pole bitvy za bezopasnost* [Voximplant Deepfakes and deep media: A new battleground for security]. Available at: https://habr.com/ru/company/Voximplant/blog/501068.

5. OpenAI releases curtailed version of GPT-2 language model. *VentoreBeat*. Available at: https://openai.com/blog/openai-five-defeats-dota-2-world-champions.

6. The Next Word. Where will predictive text take us? *New Yorker*. Available at: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

7. Karras T. et al. *Analyzing and Improving the Image Quality of StyleGAN*. Available at: https://arxiv.org/abs/1912.04958.

8. Eykholt K. et al. *Robust Physical-World Attacks on Deep Learning Models*. Available at: https://arxiv.org/abs/1707.08945.

9. Gu T., Liu K., Dolan-Gavitt B., Garg S. *BadNets: Evaluating Backdooring Attacks on Deep Neural Networks.* Available at: https://ieeexplore.ieee.org/document/8685687.

10. Arlazarov V. *Machine Learning: Trust Boundaries and Threats*. Available at: https://vc.ru/ml/167318-mashinnoe-obuchenie-granicy-doveriya-i-ugrozy.

11. *How invulnerable is artificial intelligence? Smart Engines.* Available at: https://habr.com/ru/company/smartengines/blog/528686/.

12. Ranjan A., Janai J., Geiger A., Black M. J. *Attacking Optical Flow.* Available at: https://arxiv.org/abs/1910.10053.

13. Goodfellow I. J., Shlens J., Szegedy C. *Explaining and Harnessing Adversarial Examples.* Available at: https://arxiv.org/abs/1412.6572.

14. He Y. et al. *Towards Security Threats of Deep Learning Systems: A Survey.* Available at: https://arxiv.org/abs/1911.12562

15. *What Is Next-Gen Endpoint Security? McAfee.* Available at: https://www.mcafee.com/enterprise/ru-ru/security-awareness/endpoint/what-is-next-gen-endpoint-protection.html.

16. Moisejevs I. *Privacy attacks on Machine Learning.* Available at: https://towardsdatascience.com/privacy-attacks-on-machine-learning-a1a25e474276.

17. Grandperrin J. *How to use confidence scores in machine learning models.* Available at: https://towardsdatascience.com/how-to-use-confidence-scores-in-machine-learning-models-abe9773306fa.

18. Moisejevs I. *Poisoning attacks on Machine Learning.* Available at: https://towardsdatascience.com/poisoning-attacks-on-machine-learning-1ff247c254db.

19. Moisejevs I. *Evasion attacks on Machine Learning.* Available at: https://towardsdatascience.com /evasion-attacks-on-machine-learning-or-adversarial-examples-12f2283e06a1.

20. Panasenko A. *Tekhnologii Deepfake kak ugroza informatsionnoy bezopasnosti* [Deepfake technologies as a threat to information security]. Available at: malware.ru/analytics/Threats_Analysis/Deepfakes-as-a-information-security-threat.

21. Drozhzhin A. *How to protect yourself from deepfakes.* Available at: https://www.kaspersky.ru/blog/rsa2020-deepfakes-mitigation/27678/.

22. Schroepfer M. *Creating a dataset and a challenge for deepfakes.* Available at: https://ai.facebook.com/blog/deepfake-detection-challenge.

23. Li Yu. et al. *In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking.* Available at: https://arxiv.org/abs/1806.02877.

24. Li Y., Lyu S. *Exposing DeepFake Videos By Detecting Face Warping Artifacts.* Available at: https://arxiv.org/abs/1811.00656.

25. Gu Y. et al. *Protecting World Leaders Against Deep Fakes.* Available at: https://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.pdf.

26. Weiler J., Scholten C. *Recognising fake images using frequency analysis.* Available at: https://news.rub.de/english/press-releases/2020-07-16-information-technology-recognising-fake-images-using-frequency-analysis.

27. Karolina G. et al. *A study of the effect of JPG compression on adversarial images.* Available at: https://arxiv.org/abs/1608.00853.

28. Sermanet P. et al. *OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks.* Available at: https://arxiv.org/abs/1312.6229.