# 1

### **Анализ уязвимостей и угроз, связанных с применением методов машинного обучения в области кибербезопасности.**

В современном цифровом мире, где технологии машинного обучения становятся все более распространенными и востребованными, вопросы кибербезопасности приобретают особую важность. Применение методов машинного обучения открывает новые возможности для обнаружения и предотвращения киберугроз, однако это с собой несет и ряд уязвимостей. Анализ этих рисков является ключевым аспектом обеспечения безопасности информационных систем и данных. Важной задачей является изучение и анализ уязвимостей, связанных с использованием методов машинного обучения в киберпространстве, а также в выявлении возможных угроз и способов их предотвращения. Понимание вызовов, связанных с применением методов ML, позволит разработать эффективные стратегии защиты и повысить уровень безопасности в цифровом пространстве.

Анализируя доступные источники, а также рассматривая работу систем на базе машинного обучения можно выделить и разобрать следующий список угроз.

**Состязательные атаки на модели машинного обучения:**

Состязательные атаки представляют собой специальные воздействия на элементы конвейера , направленные на манипуляцию поведением моделей. Целью таких атак является создание обученной машины, где атакующий ищет набор незаметных изменений во входных данных, чтобы модель неправильно классифицировала их. Типичный пример состязательной атаки: изменение некоторых пикселей в изображении перед загрузкой, чтобы система распознавания изображений не могла классифицировать результат. Эти незаметные для человека изменения, называемые состязательными примерами, могут обмануть систему машинного обучения. Состязательные атаки могут быть направлены на различные типы данных, включая изображения, тексты, аудио-данные и временные ряды. Они могут быть использованы для обмана систем распознавания лиц, спам-фильтров, а также для других целей, где необходимо изменить результаты классификации. Существует несколько методов проведения состязательных атак, таких как FGSM (Fast Gradient Sign Method), DeepFool и JSMA (Jacobian-based Saliency Map Attack). Они различаются по эффективности создания состязательных примеров, количеству возмущений и вычислительным ресурсам. Для защиты от состязательных атак применяются методы состязательного машинного обучения, которые помогают выявлять и устранять аномалии в моделях искусственного интеллекта. Однако злоумышленники также могут использовать эти методы в своих целях, например, для обхода систем безопасности.

**Предвзятость и необъяснимость моделей ИИ:**

Одной из ключевых угроз, связанных с применением методов машинного обучения, является предвзятость и необъяснимость систем. Предвзятость может возникать из-за особенностей входных данных, на которых обучается модель, или быть искусственно внесенной в процессе разработки. Например, система распознавания лиц может неверно идентифицировать темнокожих пользователей, если в обучающих данных недостаточно представлены лица разных рас. Проблема усугубляется тем, что современные сложные модели ИИ работают по принципу "черного ящика" - их внутренние механизмы принятия решений становятся непостижимыми для человека. Это затрудняет диагностику предвзятости и исправление ошибок в работе таких систем. Кроме того, предвзятость может быть направленно внесена злоумышленниками для манипуляции поведением модели. Пример: Алгоритм оценки нуждаемости пациентов в медицинской помощи считал чернокожего пациента менее нуждающимся, чем белого, даже если у первого было больше объективных причин для получения помощи. Причина - в данных для обучения алгоритма были заложены расовые предубеждения, существующие в обществе. Но в коде самого алгоритма не было прямого указания отдавать предпочтение белым пациентам. Для противодействия угрозе предвзятости и необъяснимости моделей ИИ необходимо:

* Тщательно анализировать данные для обучения на наличие предвзятости
* Применять методы объяснимого искусственного интеллекта (XAI), позволяющие понять логику принятия решений моделью
* Разрабатывать модели с соблюдением принципов прозрачности и интерпретируемости
* Только комплексный подход к обеспечению безопасности и прозрачности моделей машинного обучения позволит минимизировать угрозы, связанные с их предвзятостью и необъяснимостью.

**Недостаток данных и ложные срабатывания:**

Недостаточное количество или некачественные обучающие данные могут приводить к тому, что модель будет неправильно классифицировать события и ошибочно срабатывать на безопасные ситуации. Проблема недостатка данных особенно актуальна для редких или новых типов атак, когда в распоряжении разработчиков моделей машинного обучения нет достаточного количества примеров для обучения. Это может привести к тому, что модель будет плохо обобщать и ошибаться при классификации подобных атак в реальных условиях. Система обнаружения вторжений может пропускать новые типы вредоносных действий, если они не были представлены в обучающих данных. Помимо количества, важное значение имеет качество обучающих данных. Если в них присутствуют ошибки, неточности или шумы, модель машинного обучения будет воспринимать их как нормальные ситуации и ложно классифицировать реальные угрозы. Например, система обнаружения аномалий в сетевом трафике может ошибочно считать вредоносные действия нормальным трафиком, если в обучающих данных присутствовали зашумленные или некорректно размеченные примеры. Для противодействия угрозе ложных срабатываний необходимо:

* Применять методы контроля качества данных и мониторинга работы моделей
* Использовать ансамблевые методы, которые более устойчивы к недостатку данных
* Проводить регулярное тестирование моделей на новых типах данных
* **Время реагирования на угрозу:**

Известно, что современные кибератаки очень быстро переходят от эксплуатации уязвимости к развертыванию вредоносных действий. Раньше, прежде чем начать атаку, злоумышленникам приходилось вручную проверять все слабые места и обходными путями выводить из строя системы безопасности – иногда этот процесс мог занимать недели. Однако, с развитием технологий, время, необходимое для развертывания атаки, значительно сократилось. Теперь злоумышленники могут автоматизировать процесс поиска уязвимостей и запуска вредоносного кода. Это ставит под угрозу эффективность традиционных систем обнаружения вторжений и реагирования на инциденты, основанных на правилах. Такие системы часто не успевают обновлять свои сигнатуры и правила быстрее, чем развиваются атаки. Модели машинного обучения, в свою очередь, также сталкиваются с проблемой времени реагирования. Для того, чтобы обучить ее распознавать новые типы атак, требуется время на сбор данных, разметку и переобучение. Кроме того, развертывание обновленной модели в рабочую среду также занимает время. Это создает окно уязвимости, которым могут воспользоваться злоумышленники. Для минимизации данной угрозы необходимо:

* Также использовать ансамблевые методы которые более устойчивы к новым типам атак
* Применять методы трансферного обучения для ускорения адаптации моделей к новым данным
* Автоматизировать процесс переобучения и развертывания моделей для сокращения времени реагирования
* Использовать методы обнаружения аномалий для выявления неизвестных угроз

**Противоречие с законами о конфиденциальности данных:**

Для эффективного обучения моделей машинного обучения требуются большие объемы разнообразных данных. Однако сбор и использование персональных данных граждан регулируется строгими законодательными нормами. Законы о конфиденциальности устанавливают ряд требований к обработке персональных данных, включая получение согласия субъекта, ограничение целей использования, обеспечение безопасности и другие. Несоблюдение этих требований может повлечь за собой серьезные штрафы и репутационные потери для организаций. В то же время, для достижения высокой точности моделей машинного обучения часто требуется использовать данные, полученные без явного согласия или в обход ограничений. Условно, система обнаружения мошенничества в финансовых транзакциях может значительно повысить свою эффективность, если будет использовать данные о транзакциях клиентов из разных банков. Однако передача и объединение таких данных между банками может противоречить законам о банковской тайне и защите персональных данных. Для минимизации данной уязвимости необходимо:

* анализировать соответствие используемых данных требованиям законодательства о конфиденциальности
* Применять методы обезличивания и анонимизации данных для снижения рисков
* Получать явное согласие субъектов на использование их персональных данных в целях обучения моделей
* Ограничивать доступ к персональным данным в процессе разработки и эксплуатации систем

**Угрозы взлома и нарушения конфиденциальности:**

Инверсия модели представляет собой метод атаки на системы машинного обучения, который может привести к восстановлению информации об обучающих данных и нарушению конфиденциальности. Этот вид атаки основан на обратном инжиниринге обучения модели, позволяя злоумышленнику извлекать подробности об обучающих данных, таких как изображения лиц или текстовые сообщения. Процесс инверсии модели начинается с обучения инверсионной модели, которая затем используется для восстановления информации об обучающих данных, проходивших через модель машинного обучения. Злоумышленник может начать с выходных данных модели, постепенно раскрывая внутреннюю структуру и веса, чтобы восстановить оригинальные входные данные. Это позволяет злоумышленнику получить доступ к конфиденциальным данным, использованным для обучения модели, что может привести к серьезным нарушениям конфиденциальности и утечкам информации. Это лишь пример одного из типов атак на систему.

**Необходимость обеспечения объяснимости системных решений ИИ:**

Модели ИИ, особенно основанные на глубоком обучении, часто работают по принципу "черного ящика", когда их внутренняя логика становится непрозрачной для человека. Это создает серьезные проблемы в области кибербезопасности, где требуется объяснение и обоснование решений систем ИИ. Без понимания логики работы таких систем, специалистам сложно выявлять ошибки, манипуляции и атаки, которые могут быть в них встроены. Кроме того, непрозрачность ИИ снижает доверие к таким системам со стороны пользователей и регуляторов. Для решения этой проблемы необходимо обеспечить объяснимость системных решений. Применяются специальные методы, такие как визуализация признаков, генерация контрфактических примеров и использование интерпретируемых моделей. Эти подходы позволяют "заглянуть внутрь" черного ящика нейронных сетей и понять логику их работы. Обеспечение объяснимости является важным условием для успешного внедрения. Только объяснимые системы ИИ способны завоевать доверие специалистов и обеспечить прозрачность их функционирования. Недостаточная объяснимость решений представляет собой серьезную уязвимость, которую необходимо устранять для обеспечения безопасности информационных систем.

# 2

**Реальная обстановка ML-систем в сфере информационной безопасности.**

Внедрение систем машинного обучения несет ряд важных преимуществ:

* алгоритмы способны анализировать большие объемы данных и выявлять сложные, ранее неизвестные атаки, основываясь на поведенческих моделях. Это позволяет быстрее реагировать на инциденты и предотвращать ущерб.
* ML позволяет автоматизировать многие рутинные процессы в сфере кибербезопасности, такие как анализ журналов событий, обнаружение аномалий, классификация вредоносных файлов. Это высвобождает время специалистов для решения более сложных задач.
* способность выявлять тонкие закономерности и связи в данных, недоступные человеческому анализу. Это позволяет более точно определять риски и принимать обоснованные решения по защите.
* ML-системы могут самообучаться на основе новых данных, постоянно совершенствуя свои модели выявления угроз. Это критически важно в условиях быстро меняющегося ландшафта киберугроз.
* способность обрабатывать огромные объемы данных и обслуживать тысячи пользователей, сохраняя высокую производительность. Это позволяет внедрять их в крупных распределенных инфраструктурах.

Однако чтобы действительно систему можно было внедрить и использовать нужно будет решить все ранее описанные проблемы. Именно потому что присутствует необходимость в разрешении или как минимум минимизации каждой уязвимости вы навряд ли сможете найти кейс реального и полноценного использования ИИ в сфере информационной безопасности, зачастую это будут редкие случаи и то применяемые по большей мере для частных автоматизаций тех или иных процессов. внедрение ML в сферу информационной безопасности открывает новые возможности для повышения эффективности защиты от киберугроз. Однако важно помнить, что система не заменяет человеческий анализ, а дополняет его, позволяя специалистам сосредоточиться на решении наиболее сложных задач

# Список литературы

Исследование атак на модели машинного обучения в сетях 5G на основе генеративно-состязательных сетей / Д. И. Парфенов, И. П. Болодурина, Л. В. Легашев [и др.] // Прикаспийский журнал: управление и высокие технологии. – 2023. – № 1(61). – С. 89-96. – EDN VNNVJT.

Намиот, Д. Е. О работе AI Red Team / Д. Е. Намиот, Е. В. Зубарева // International Journal of Open Information Technologies. – 2023. – Т. 11, № 10. – С. 130-139. – EDN NRSLMA.

Осман, С. Ш. О. Перспективы искусственного интеллекта в системах кибербезопасности / С. Ш. О. Осман // Наукосфера. – 2023. – № 9-1. – С. 213-217. – EDN DRYSZN.

Сычев, Д. И. Искусственный интеллект и кибербезопасность: будущие тенденции и вызовы / Д. И. Сычев // Международный журнал информационных технологий и энергоэффективности. – 2023. – Т. 8, № 5-2(31). – С. 9-14. – EDN FBZSBS.

Ламонина, Л. В. К вопросу о применении искусственного интеллекта в обеспечении информационной безопасности / Л. В. Ламонина, О. Б. Смирнова // Электронный научно-методический журнал Омского ГАУ. – 2023. – № 3(34). – EDN MARYTD.