ОТЗЫВ

официального оппонента на диссертационную работу Подтопельного Владислава Владимировича «Модели и методика определения последовательностей атакующих воздействий на системы искусственного интеллекта при аудите информационной безопасности», представленную на соискание учёной степени кандидата технических наук по научной специальности 2.3.6. – Методы и системы защиты информации, информационная безопасность

Диссертационная работа В. В. Подтопельного «Модели и методика определения последовательностей атакующих воздействий на системы искусственного интеллекта при аудите информационной безопасности» содержит оригинальные результаты исследования в трех областях:

- область аудита информационной безопасности (ИБ) систем искусственного интеллекта (ИИ);
- область моделирования (определение последовательности атакующих воздействий на системы ИИ) с использованием методов марковских процессов принятия решений (МППР);
- область разработки программного обеспечения (для автоматизации расчётов).

В условиях активного развития технологий и увеличения количества инцидентов, связанных с компрометацией систем ИИ, разработка методов определения наиболее опасных последовательностей атак становится критически важной задачей. Целью диссертационной работы является повышение степени полноты описания атак на системы ИИ при аудите их ИБ.

Актуальность темы диссертационного исследования связана с необходимостью разработки моделей определения наиболее опасных последовательностей атакующих воздействий на системы ИИ и их подсистем для улучшения качества аудита ИБ.

В работе для достижения этой цели последовательно решаются следующие задачи:

1. Анализ современных подходов, механизмов выявления и оценки инцидентов безопасности в системах ИИ.
2. Определение особенностей применения существующих баз знаний для разработки модели анализа атак на системы ИИ, определение специфики используемых при моделировании данных.
3. Разработка модели построения и анализа атак на системы ИИ.
4. Разработка методики и алгоритма моделирования атакующих воздействий в процессе аудита ИБ подсистем ИИ для последующего составления сценариев атак.
5. Разработка программного решения для автоматизации процессов моделирования и проведения оценки работоспособности моделей.

Кроме того, в рамках настоящего исследования рассматриваются следующие сопутствующие проблемы:

1. Формализация функции вознаграждения в контексте МППР, что является фундаментальным аспектом для оптимизации поведения системы в условиях неопределенности и динамически изменяющихся условий.
2. Идентификация и классификация потенциально доступных злоумышленнику действий с учетом их интерпретации на основе баз знаний MITRE и

ФСТЭК. Это позволяет провести всесторонний анализ угроз и разработать адекватные меры защиты.

3. Определение уровней детализации описания действий злоумышленника, что напрямую влияет на точность и эффективность предлагаемых моделей. Важным аспектом является баланс между детализацией и генерализацией, позволяющий избежать избыточности и недостаточности информации.

4. Применение разработанных математических моделей, алгоритмов и программных комплексов для решения прикладных задач в области кибербезопасности и управления рисками. Это включает в себя моделирование сценариев атак, прогнозирование поведения злоумышленников и оптимизацию защитных механизмов.

Общая характеристика работы. Диссертационная работа содержит введение, 4 главы, заключение, список литературы и приложения. Диссертационная работа оформлена на 250 страницах, включает 50 рисунков и 20 таблиц.

Во введении указаны объект и предмет исследования, цели исследования, обоснована актуальность темы исследования, сформулированы задачи и методы их решения, приведены основные положения, выносимые на защиту, дается краткое содержание диссертации по главам.

В первой главе работы представлен анализ целей, задач и возможностей методов моделирования атак с использованием МППР. В этой главе также рассматриваются особенности аудита систем ИИ с учётом известных методов описания атак (MITRE ATLAS и Методики ФСТЭК). Кроме того, в разделе представлена классическая постановка задачи моделирования последовательностей атак. В работе приводится сравнительная таблица методов, которая позволяет сделать общий вывод об ограниченности применения каждого метода в отдельности.

Во второй главе работы рассматривается специфика применения методов моделирования атак на системы ИИ с использованием МППР.

Вводится понятие функции ценности и вознаграждения, а также определяются математические основы формирования последовательности атакующих воздействий в виде графа атак.

В процессе моделирования атак выделяются два подхода к представлению переходов в МППР: моделирование в режиме реального времени, когда учитывается возможность возврата злоумышленника в уже достигнутые состояния атаки (on-line режим), и моделирование в режиме без имитации реального поведения злоумышленника (off-line режим).

Важным результатом является обоснование и формирование функции вознаграждения, а также адаптация множеств состояний и переходов МППР к задачам определения оптимальной последовательности атакующих воздействий на системы ИИ. Кроме того, в адаптированную модель интегрируются методы и техники баз знаний MITRE и ФСТЭК.

Определяются ограничения при моделировании, а также производится адаптация множеств состояний и переходов МППР к задачам определения оптимальной последовательности атакующих воздействий на системы ИИ.

В третьей главе работы рассматриваются модели формирования последовательностей атак, основанные на МППР.

В основе этих моделей лежит метод итераций по значениям для оценки полезности состояний. При моделировании используются уровни абстракции, что

позволяет создать три типа моделей. Каждая из них имеет два режима применения: on-line и off-line.

1. Типовая модель на основе МППР. Она предназначена для общего анализа специфики атаки в режимах аудита on-line и off-line для систем с элементами ИИ. В этой модели не конкретизируются действия по тактикам.

2. Упрощённые модели. Они используют действия и состояния из классификации тактик MITRE ATLAS в режимах аудита on-line и off-line. В этих моделях тактики объединены в классы на основе функционального назначения техник. Также учитываются категории действий (совокупности техник), которые выделены на основе необходимости обязательного взаимодействия во время атак.

3. Подробные модели. Они используют действия и состояния из классификации тактик MITRE ATLAS. В этих моделях не предусмотрены ограничения при переходах между состояниями.

Для каждого типа моделей приведены ограничения и рекомендации по использованию.

Результатом работы является формальное описание последовательности атакующих воздействий, которое позволяет создать математически обоснованный сценарий атаки как с применением вычислительных средств, так и без них.

Особое внимание уделяется моделям МППР на основе Value Iteration. Были разработаны методики и алгоритмы, которые улучшают процесс аудита систем ИИ.

В четвёртой главе проводится практическая проверка разработанных моделей.

В контексте моделирования атак на системы ИИ с применением методов машинного обучения полнота модели определяется её способностью учитывать все возможные состояния системы и переходы между ними при воздействии атакующих действий.

При испытании моделей полнота описания набора действий определяется с учётом: всех состояний системы (максимально полное описание состояний достигается при использовании тактик MITRE ATLAS); типов действий (нелегальное взаимодействие (D), легальное взаимодействие (C)); всех возможных действий атакующего с учётом методики описания сценария атаки (действия как тактики из MITRE ATLAS и Методики ФСТЭК) при использовании полносвязанного графа атак; обратных действий атакующего (возвращение злоумышленника в предыдущее состояние атаки (действия сброса (R)) в режимах on-line); перебора всех вариантов последовательностей при учёте специфики уязвимостей и адресации узла, которые также учитываются при аудите систем ИИ.

Результаты диссертационного исследования обоснованы и соответствуют поставленным задачам. Совокупность приведенных в работе результатов можно рассматривать как инновационное решение научно-технической проблемы, которое имеет большое значение для развития сферы защиты информации, в частности, аудита безопасности ИИ.

Автор продемонстрировал компетентный подход к разработке новых математических алгоритмов и численных методов, а также успешно воплотил их в виде программных комплексов.

Работа прошла соответствующую апробацию, основные результаты отражены 17 публикациях, в том числе, в пяти статьях, опубликованных в ведущих рецензируемых журналах, входящих в перечень ВАК, в материалах четырех

международных конференций. Получено семь свидетельств о государственной регистрации программ для ЭВМ.

Научные положения, выводы и результаты диссертационной работы корректны и научно обоснованы.

Диссертация соответствует специальности 2.3.6 «Методы и системы защиты информации, информационная безопасность».

Научная новизна результатов исследования заключается в следующем (все результаты, выносимые на защиту, являются новыми):

1. Разработанный аппарат МППР позволяет математически обосновать создание сценариев атак, представляющих собой последовательность атакующих воздействий и состояний безопасности системы ИИ. Это дает возможность прогнозировать их возникновение с учетом специфики подсистем ИИ. При этом учитываются руководящие документы и базы знаний, используемые при моделировании угроз ИБ. Такой подход способствует улучшению и контролю взаимодействия смежных систем контроля и обнаружения событий безопасности. Внедрение методов анализа на основе МППР положительно влияет на качество управления ИБ организаций и оперативность реагирования на угрозы.

2. Разработанный алгоритм марковских моделей, учитывающий формализацию типовых атакующих последовательностей (техник и тактик). Алгоритм позволяет выявлять и анализировать ранее не обнаруживаемые этапы вредоносного воздействия на информационные системы с подсистемами ИИ. В процессе анализа атак учитываются топология сети предприятия, архитектурные особенности подсистем ИИ, методики описания действий атакующего, применяемые в ходе аудита ИБ, а также другие значимые факторы.

3. Разработаны модели и методика для определения атакующих последовательностей, используемых в сценариях атак на системы ИИ, в том числе системы, выступающие в роли подсистем ИИ.

Обоснованность и достоверность представленных в диссертации научных выводов подтверждаются тщательным анализом современных исследований в данной области. Результаты, полученные с помощью компьютерной реализации теоретических положений, согласуются с проведенными исследованиями, что свидетельствует о корректности и надежности предложенных подходов.

Практическая ценность результатов диссертационного исследования заключается в том, что разработанные модели и методика могут быть использованы для решения практических задач, то есть они позволяют:

– разработать сценарии атак, опираясь на принципы построения марковских моделей;

– оптимизировать процедуры аудита за счёт внедрения и применения предложенных моделей;

– использовать разработанный подход к созданию модели угроз для систем ИИ.

Достоверность результатов подтверждается строгими математическими обоснованиями и доказательствами.

Замечания по диссертационной работе:

1. Во второй главе описываются виды доступа, которые используются при определении параметров моделей второго типа в третьей главе, при этом для моделей третьего типа упоминается только возможность их применения.

2. Для моделей второго типа указывается применимость к описанию атак на инфраструктуру объекта, подробно не объяснено, почему приоритет отдается анализу и построению атак на логические компоненты систем ИИ.

3. Методика моделирования предусматривает применение как ручного, так и автоматизированного способа определения атакующих последовательностей. Автором также предлагается использование разработанной системы, реализующей автоматизированный способ определения атакующих последовательностей. Однако подробно не описывается, в каких случаях допустимо определять искомые последовательности без средств автоматизации.

Заключение

Приведенные замечания не снижают высокого уровня проведенной соискателем работы, из чего следует, что диссертационная работа В. В. Подтопельного «Модели и методика определения последовательностей атакующих воздействий на системы искусственного интеллекта при аудите информационной безопасности» является завершенной научно-квалификационной работой. Результаты, полученные автором, являются новыми, обоснованными и достоверными. Автореферат полностью соответствует содержанию диссертации.

Работа отвечает требованиям Положения ВАК о порядке присуждения ученых степеней, а ее автор, **Подтопельный Владислав Владимирович**, заслуживает присуждения ученой степени кандидата технических наук по специальности **2.3.6 «Методы и системы защиты информации, информационная безопасность».**

Официальный оппонент.
к.т.н., доцент, программист
АО «Клаудран»

«20» НОЯБРЯ 2025 г. [redacted] А.А. Браницкий

адрес: Россия, 195197, город Санкт-Петербург, Полюстровский пр-кт, д. 28 литера Д, помещ. 2-н
e-mail:
тел.:

Подпись А.А. Браницкого заверяю: [redacted]

АКЦИОНЕРНОЕ ОБЩЕСТВО «КЛАУДРАН» ИНН 7804718379 «КлаудРан» ОГРН 1257800048790 САНКТ-ПЕТЕРБУРГ