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 исследований в сфере информационной безопасности.

С точки зрения развития сферы защиты информации наиболее важным результатом является возможность производить определение наиболее опасных последовательностей атакующих воздействий на системы искусственного интеллекта, что повышает качество аудита информационной безопасности в части составления сценариев атак, что на современном этапе развития интеллектуальных технологий является критично важным.

Актуальность темы диссертационного исследования обусловлена необходимостью разработки моделей определения наиболее опасных последовательностей атакующих воздействий на системы ИИ и их подсистемы для улучшения качества аудита ИБ, учитывая высокую динамику развития технологий нападения на системы ИИ и увеличение количества инцидентов безопасности, связанных с компрометацией вычислительных моделей ИИ.

Основной целью диссертационной работы является повышение степени полноты описания атак на системы искусственного интеллекта при аудите их информационной безопасности.

В работе для достижения этой цели последовательно решаются следующие задачи:

1. Произведение анализа существующих подходов, инструментов поиска и оценки событий безопасности в информационных системах с элементами ИИ при неоднозначно интерпретируемых входных данных о потенциально опасных, деструктивных воздействиях на ресурсы информационных систем.

2. Определение возможности использования доступных наборов данных для разработки модели анализа атак на системы ИИ и определение параметров используемых при моделировании данных.

3. Разработка модели построения и анализа атак на системы ИИ при определении мер противодействия атакам.

4. Разработка методических рекомендаций и алгоритма, позволяющих моделировать атакующие воздействия в процессе аудита ИБ подсистем

искусственного интеллекта (ПИИ), изучать их динамику, использовать для составления сценарии атак.

5. Разработка архитектуры программного решения, позволяющего автоматизировать процессы моделирования и проведения оценки работоспособности моделей.

Дополнительно рассматриваются следующие проблемы:

– определение функции вознаграждения в моделях марковских процессов принятия решений;

– определение доступных злоумышленнику действий с учетом интерпретации действий посредством использования баз знаний MITRE и ФСТЭК России;

– уровни детализации описания действий злоумышленника и, соответственно, предлагаемых моделей;

– использование полученных математических моделей представления данных, алгоритмов и комплексов программ для решения прикладных задач.

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

Во введении определены 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 диссертации проведен анализ целей, задач и возможностей способов моделирования атакующих воздействий с использованием марковских процессов принятия решений. Рассмотрены особенности проведения аудита систем ИИ с учетом известных способов описания атакующих воздействий (MITRE ATLAS и Методики оценки угроз безопасности информации ФСТЭК России). Также она содержит классическую постановку задачи моделирования последовательностей атакующих воздействий. Информативной является приведенная сравнительная таблица методов, позволившая сделать общий вывод об ограниченности применения каждого метода в отдельности.

Во второй главе определяется специфика применения методов моделирования атак на ПИИ с использованием марковских процессов принятия решений. Приводятся функция ценности, вознаграждения, определяются математические основы формирования последовательности атакующих воздействий как графа атаки. Выделяются два представления о переходах в марковских процессах при моделировании атак: моделирование режиме в on-line, когда учитывается возможность возврата злоумышленника в уже полученные состояния атаки, и моделирование в off-line без имитации реального поведения злоумышленника.

Определяется специфика формирования функции вознаграждения, производится определение ограничений при моделировании.

В третьей главе приводятся модели формирования последовательностей атакующих воздействий, сформированные на основе марковских процессов принятия решений. В основу моделей положен метод итераций по значениям для оценки функций полезности состояний. При моделировании вводятся уровни абстракции, и, соответственно, формируются три типа моделей, каждая из которых имеет два режима применения:on-line и off-line:

- типовая (общая) модель на основе марковских процессов принятия решений (МППР) для общего анализа специфики атаки в режимах аудита on-line и off-line на системы с элементами искусственного интеллекта без конкретизации действий по тактикам;
- упрощенные модели с использованием действий и состояний из классификации тактик MITRE ATLAS в режимах аудита on-line и off-time, в которых тактики объединены на основе сходства функционального назначения техник в несколько классов, а также учтены категории действий (совокупности техник), выделенные на основе необходимости обязательного взаимодействия во время атак;
- **подробные (полные) модели** с использованием действий и состояний из классификации тактик MITRE ATLAS,**не предполагающие ограничений при переходах между состояниями.**

Для каждого типа моделей приводятся ограничения и рекомендации по использованию. Важным результатом является формальное описание последовательности атакующих воздействий, позволяющее создать обоснованный математически сценарий атаки как с применением вычислительных средств, так и без них. Приоритетным в этом отношении являются модели МППР на основе ValueIteration. Также к важному результату можно отнести разработанные методики и алгоритмы, улучшающие процесс аудита систем ИИ.

В четвертой главе приводится экспериментальная оценка разработанных моделей. Полнота при моделировании атак на ПИИ с использованием МППР в данном случае представляется как способность модели учитывать все состояния, возникающие при воздействии атакующих действий на систему ИИ, и переходы между ними. При испытании моделей полнота описания набора действий определяется с учетом: всех состояний системы (максимально полное описание состояний достигается при использовании тактик MITRE ATLAS); типов действий (нелегальное взаимодействие ((D), легальное взаимодействие (C)), также все возможные действия атакующего с учетом методики описания сценария атаки (действия как тактики из MITRE ATLAS и Методики оценки угроз безопасности информации ФСТЭК России) при использовании полносвязанного графа атаки); 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 искусственного интеллекта.

Основные результаты и выводы работы достаточно полно и всесторонне обоснованы. В целом работа производит хорошее впечатление, ее результаты можно трактовать как новое решение научно-технической задачи, имеющей существенное значение для развития методов проведения аудита безопасности ИИ и математического моделирования атакующих воздействий на ИИ.

Автор грамотно подошел к построению новых математических алгоритмов и численных методов, успешно реализовал соответствующие алгоритмы в виде комплексов программ.

Работа прошла соответствующую апробацию, основные результаты отражены 17-ти публикациях, в том числе, в 5-ти статьях, опубликованных в ведущих рецензируемых журналах, входящих в перечень ВАК, в материалах четырех международных конференциях. Получено 7 свидетельств о государственной регистрации программ для ЭВМ.

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

Диссертация соответствует научной 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. С применением предложенного аппарата МППР появится возможность с помощью математических методов обосновать построение сценария атаки как последовательности событий безопасности, прогнозировать появление данных событий с учетом специфики подсистемы ИИ, учитывая при этом руководящие документы и базы знаний, используемые при составлении моделей угроз безопасности ИС. Это позволит улучшить и контролировать совместную работу смежных систем контроля и поиска событий безопасности (за счет введения методов анализа на основе МППР), что положительно повлияет на качество управления информационной безопасностью организаций и оперативность реагирования на угрозы.

2. Предлагается использовать совмещение нескольких методов анализа признаков событий безопасности, связанных с компрометацией ПИИ, ориентируясь на вычислительные методы, основанные на МППР. Комплекс марковских моделей, с учетом формализации типовых атакующих последовательностей (техник и тактик), позволит выявлять и учитывать ранее необнаруженные этапы развития вредоносного воздействия на информационные системы с ПИИ. При анализе учитываются: топология сети предприятия, архитектурные особенности подсистем ИИ, методики описания действий атакующего, применяемые при аудите, и другие факторы.

3. Разработаны модели и методика определения атакующих последовательностей для сценариев атак на ПИИ, формируемых в процессе аудита ИБ ПИИ.

Обоснованность и достоверность научных выводов, представленных в диссертации, достигаются благодаря тщательному анализу современных исследований в этой области. Это подтверждается тем, что результаты, полученные с помощью компьютерной реализации, согласуются с теоретическими положениями.

Практическая значимость результатов диссертационной работы состоит в возможности использования разработанных моделей, методики, алгоритма для решения конкретных прикладных задач:

– построение последовательности атакующих воздействий с целью формирования сценария атак на системы ИИ;

– улучшить процедуры аудита посредством внедрения и использования предложенных моделей, применять разработанные модели и методику к формированию модели угроз для систем ИИ.

Достоверность результатов обосновывается строгими математическими определениями и доказательствами.

Замечания по диссертационной работе:

1. Доказательства ряда утверждений излагаются излишне подробно, что увеличивает объем диссертационной работы и, соответственно, затрудняет чтение.

2. В пункте 2.4 модель нарушителя чрезвычайно сжата, что требует для понимания специфики формирования и влияния модели нарушителя на процесс моделирования атак использовать материал следующих глав.

3. Моделирование в главе 3 учитывается от внешнего нарушителя, специфика моделирования внутреннего нарушителя учитывается, однако не выделяется в отдельное описание.

4. Стиль изложения материала диссертационного исследования сложен для понимания сути работы.

Заключение

Высказанные выше 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. Полученные автором результаты являются достаточно 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 научной специальности **2.3.6. – «Методы и системы защиты информации, информационная безопасность».**

Официальный оппонент:

Заведующий кафедрой
«Системы информационной безопасности»
ФГБОУ ВО "Брянский государственный технический университет",
доктор технических наук, доцент

[redacted] Михаил Юрьевич Рытов

«28» 10 2025 года

241035, г. Брянск, бульвар 50-лет Октября, д.7
Федеральное бюджетное государственное
образовательное учреждение высшего образования
«Брянский государственный технический университет» (ФГБОУ ВО «БГТУ»)
Тел.: (4832)51-13-77, 89103300237
e-mail: rmy@tu-bryansk.ru