МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН)

14-я линия, д. 39, г. Санкт-Петербург, 199178
Тел.: (812) 328-33-11, факс: (812) 328-44-50,
e-mail: info@spcras.ru, web: http://www.spcras.ru
ОКПО 04683303, ОГРН 1027800514411,
ИНН/КПП 7801003920/780101001

УТВЕРЖДАЮ

Директор СПб ФИЦ РАН
доктор технических наук,
профессор РАН

[redacted] Ронжин А.Л.

« 21 » ноября 2025 года

«21» 11 2025 № 60/01-01- 1048

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

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

Актуальность темы исследования

Диссертационная работа Владислава Владимировича Подтопельного посвящена созданию моделей, позволяющих определять наиболее выгодные для злоумышленника последовательности атакующих воздействий на системы и подсистемы искусственного интеллекта при аудите их информационной безопасности. Под последовательностью понимается порядок действий злоумышленника и состояний атаки или атакуемого объекта, приводящий к достижению цели атакующего и компрометации системы ИИ. Основным направлением исследований автора является разработка моделей формирования таких последовательностей на основе марковских процессов принятия решений (МППР) с использованием баз знаний MITRE ATLAS и методических подходов ФСТЭК России.

Интеграция систем искусственного интеллекта в современные корпоративные информационные системы, а также в объекты критической информационной инфраструктуры сопровождается появлением новых категорий угроз. Рост числа инцидентов, связанных с атаками на модели машинного обучения, искажением обучающих данных, эксплуатацией уязвимостей алгоритмов и инфраструктурных компонентов подтверждает высокую актуальность проблемы. К характерным примерам относятся атаки «белого», «серого» и «черного» ящика, атаки отравления данных, манипуляции входными выборками и воздействия, направленные на изменение логики работы вычислительных моделей.

Специфика подсистем искусственного интеллекта заключается в высокой сложности их архитектур, зависимости от качества данных и отсутствии устойчивых механизмов защиты от целенаправленных вредоносных воздействий. Эта сложность формирует дополнительные вызовы для аудиторов ИБ: традиционные методы анализа угроз и моделирования атак не учитывают особенности функционирования ИИ и не позволяют описать реальные сценарии компрометации с достаточной полнотой. В условиях расширяющегося применения интеллектуальных технологий в производственных, коммерческих и государственных информационных системах, а также растущих требований регуляторов, возрастает необходимость разработки методик, учитывающих специфику логики работы ИИ, особенности их уязвимостей и динамику атак.

В этой связи создание формализованных моделей определения последовательностей атакующих воздействий на подсистемы ИИ, основанных на МППР и опирающихся на современную базу знаний MITRE ATLAS, является актуальной научной задачей. Решение этой задачи позволяет обеспечить более полное и достоверное моделирование поведения злоумышленника, повысить качество аудита информационной безопасности и повысить защищенность систем искусственного интеллекта за счет улучшения процедур анализа угроз и формирования обоснованных сценариев атак.

Оценка структуры и содержания работы

Диссертация состоит из введения, четырех глав, заключения, списка литературы из 136 источников. Объем диссертации составляет 250 страниц.

Во введении изложена история вопроса, поставлены основные задачи, решаемые автором, и сводка результатов по этим задачам. Приведены понятия и обозначения, используемые в диссертации, обсуждается общая мотивировка решаемых задач, проведен обзор литературы по теме диссертации, сформулированы достигнутые результаты и краткое содержание работы.

В первой главе диссертации проведен анализ целей, задач и возможностей способов моделирования атакующих воздействий с использованием марковских моделей принятия решений. Рассмотрены особенности проведения аудита систем ИИ с учетом известных способов описания атакующих воздействий (MITRE ATLAS и Методики ФСТЭК). 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.

Важным результатом является обоснование и формирования функции вознаграждения, а также адаптация множеств состояний МППР и переходов между ними к задачам определения наилучшей последовательности атакующих воздействий на системы ИИ, а также внедрения в адаптированную модель тактик и техник баз знаний MITRE ATLAS и ФСТЭК.

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 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 по необходимости обязательного взаимодействия во время атак;

- **подробные (полные) модели,** с использованием действий и состояний из классификации тактик MITRE ATLAS **не предполагающее 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 Valie Iteration. 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 ФСТЭК) 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 учете специфики уязвимостей и адресации узла, которые также учитываются при аудите систем искусственного интеллекта.

Следует подчеркнуть, что полученные в диссертации результаты представлены последовательно и изложены логично. Диссертация и автореферат оформлены в соответствии с принятыми для научных квалификационных работ нормами и требованиями

Область исследования диссертации соответствует пунктам паспорта научной специальности 2.3.6 «Методы и системы защиты информации, информационная безопасность»:

п. 3 «Методы, модели и средства выявления, идентификации, классификации и анализа угроз нарушения информационной безопасности объектов различного вида и класса»;

п. 6 «Методы, модели и средства мониторинга, предупреждения, обнаружения и противодействия нарушениям и компьютерным атакам в компьютерных сетях»;

п. 9 «Модели противодействия угрозам нарушения информационной безопасности для любого вида информационных систем, позволяющие получать оценки показателей информационной безопасности».

п. 15 «Принципы и решения (технические, математические, организационные и др.) по созданию новых и совершенствованию

существующих средств защиты информации и обеспечения информационной безопасности».

п. 16. «Модели, методы и средства обеспечения аудита и мониторинга состояния объекта, находящегося под воздействием угроз нарушения его информационной безопасности, и расследования инцидентов информационной безопасности в автоматизированных информационных системах»

Оформление диссертации соответствует ГОСТ Р 7.0.11-2011. Автореферат диссертации выполнен с соблюдением установленных требований, полностью отражает ее содержание, полученные в ней практические и теоретические результаты и выводы.

Новизна полученных результатов

1. Модели определения последовательности и анализа атакующих воздействий на ПИИ, отличающиеся использованием марковских процессов принятия решений для формализации сценариев атак как последовательности событий безопасности с учетом специфики подсистем ИИ, а также интеграцией тактик и техник MITRE ATLAS. Применение таких моделей обеспечивает улучшение взаимодействия смежных систем контроля и поиска событий безопасности и повышает качество аудита ИБ благодаря введению математически обоснованных процедур анализа.

2. Алгоритм определения последовательности действий и состояний при атаке на ПИИ, отличающийся совмещением методов анализа признаков событий безопасности, связанных с компрометацией ПИИ, и вычислительных процедур на основе МППР. Алгоритм учитывает топологию сети предприятия, архитектурные особенности подсистем ИИ, используемые методики описания действий атакующего и формализованные атакующие последовательности, что позволяет выявлять ранее неочевидные этапы развития атаки.

3. Методика определения последовательностей атакующих воздействий на ПИИ, отличающаяся интеграцией разработанных моделей и алгоритмов в единую процедуру формирования сценариев атак, ориентированную на применение при аудите ИБ ПИИ. Методика поддерживает многослойное описание состояний и действий атакующего, обеспечивает уточнение

сценариев на разных уровнях детализации и формирует последовательности, отражающие реальные механизмы компрометации систем ИИ.

4. Архитектура и программные компоненты системы построения последовательности атакующих воздействий на ПИИ на основе разработанных моделей МППР, отличающиеся практической реализацией методов, моделей и алгоритмов, заявленных в новизне работы, включая автоматизацию формирования атакующих последовательностей, анализ состояний и событий безопасности, а также использование вычислительного аппарата МППР для оценки оптимальных стратегий злоумышленника. Архитектура обеспечивает применение результатов диссертации в реальных задачах аудита ИБ ПИИ.

Степень достоверности результатов исследования

Достоверность научных положений, результатов и выводов подтверждается корректной постановкой задач и выбором методов исследования, обсуждением полученных результатов на научных конференциях, их апробацией в процессе проведения экспериментов с использованием разработанных программных средств, практическим применением разработанных предложенных моделей, методики и алгоритма определения последовательностей атакующих воздействий при решении прикладных задач аудита информационной безопасности подсистем искусственного интеллекта. Подтопельный В.В. грамотно подошел к построению новых математических алгоритмов и численных методов, успешно реализовал соответствующие алгоритмы в виде комплексов программ. Научные положения, выводы и результаты диссертационной работы корректны и научно обоснованы.

Публикации. По материалам исследования опубликовано 17 работ, в том числе 5 статей в научных изданиях из Перечня рецензируемых научных изданий, рекомендованных ВАК, и материалы 4 международных конференций. Получено 7 свидетельств о государственной регистрации программ для ЭВМ.

Теоретическая и практическая значимость результатов, полученных автором диссертации

Теоретическая значимость предложенных результатов заключается в

развитии методологической базы моделирования атак на подсистемы искусственного интеллекта за счет применения марковских процессов принятия решений. Предложенные модели, методика и алгоритм формируют научно обоснованный аппарат для определения последовательностей атакующих воздействий на системы ИИ, обеспечивая более полное и структурированное описание сценариев атак с учетом их динамики, тактик и техник, отраженных в современных базах знаний.

Практическая значимость и реализация результатов работы состоит в возможности применения разработанных моделей, методики и алгоритма при решении прикладных задач аудита информационной безопасности систем искусственного интеллекта. Полученные результаты позволяют формировать последовательности атакующих воздействий и выстраивать сценарии атак на системы ИИ, повышать эффективность процедур аудита за счет использования предложенных моделей, а также применять разработанный аппарат при формировании моделей угроз.

Рекомендации по использованию результатов и выводов диссертации

Материалы диссертационной работы могут быть использованы в организациях, занимающихся анализом и обеспечением безопасности систем ИИ, а также интегрированы в образовательные программы подготовки специалистов в сфере информационной безопасности. Также они могут представлять интерес для научных и образовательных учреждений, в которых ведутся исследования в области информационной безопасности. К их числу относятся Московский государственный университет имени М.В. Ломоносова, Казанский (Приволжский) федеральный университет, Уральский федеральный университет, Томский государственный национальный исследовательский университет, Московский педагогический государственный университет. Кроме того, полученные результаты могут служить материалом для различных университетских курсов и спецкурсов по ИБ и ИИ. Кроме того, интерес к результатам могут проявить и коммерческие компании, специализирующиеся на разработке и анализе защищённых цифровых платформ, например ООО

«Газинформсервис», АО «Позитив Текнолоджиз», а также другие российские организации, работающие в области информационной безопасности.

Замечания по диссертационной работе

1. В ряде фрагментов главы, посвященной подробному построению моделей и алгоритмов (разделы 3.2, 3.3, 3.4), материал изложен чрезмерно детально. Было бы целесообразно предварять их более компактным обобщенным описанием модели, что позволило бы уменьшить громоздкость последующего изложения.

2. Хотя в диссертации активно используется база знаний MITRE ATLAS, специфика атак на вычислительные модели искусственного интеллекта представлена менее полно, чем инфраструктурные и сетевые аспекты. Проблематика атак на данные, модели и механизмы обучения освещена преимущественно через тактики ATLAS с использованием стандарта для оценки степени опасности уязвимостей CVSS, однако глубокая связь между внутренними уязвимостями моделей ИИ и построением марковских стратегий остается раскрытой не в полной мере.

3. Включение в качестве приложений исходных кодов разработанных программных компонентов, повысило бы уровень воспроизводимости результатов и упростило бы практическое использование предложенного программного решения.

4. Работа обоснованно опирается на базу MITRE ATLAS и частично на методические документы ФСТЭК, однако современные международные источники по угрозам для ИИ освещены слабо. Это несколько сужает контекст исследования, особенно с учетом активного развития международных стандартов и таксономий угроз в области искусственного интеллекта.

5. Оформление отдельных элементов диссертационной работы не в полной мере соответствует ГОСТ Р 7.0.11–2011, а оформление библиографических записей не полностью соответствует ГОСТ 7.1 и ГОСТ 7.80.

Приведенные замечания в целом не снижают общей положительной оценки работы и не ставят под сомнение основные научные и практические

результаты диссертационной работы.

Заключение

Диссертация Подтопельного Владислава Владимировича на соискание ученой степени кандидата технических наук является законченной научно-квалификационной работой, в которой на основании выполненных автором исследований разработаны модели, методика и алгоритм определения последовательностей атакующих воздействий на подсистемы искусственного интеллекта, а также архитектура программного решения для автоматизации моделирования атак. Представленные результаты обеспечивают повышение полноты и обоснованности сценариев атак, применяемых при аудите информационной безопасности систем ИИ, и могут быть использованы в практических задачах анализа защищенности.

Диссертация соответствует требованиям п. 9-14 Положения о порядке присуждения ученых степеней, утвержденного Постановлением Правительства Российской Федерации № 842 от 24.09.2013 г. (с последующими изменениями), а ее автор – Подтопельный Владислав Владимирович - заслуживает присуждения ему ученой степени кандидата технических наук по специальности 2.3.6. Методы и системы защиты информации, информационная безопасность.

Диссертационная работа Подтопельного В.В. и отзыв на нее обсуждены на заседании лаборатории проблем компьютерной безопасности Федерального государственного бюджетного учреждения науки "Санкт-Петербургский Федеральный исследовательский центр Российской академии наук". Протокол заседания № 10 от 13 октября 2025 года.

Главный научный сотрудник лаборатории Проблем компьютерной безопасности
д.т.н, профессор [redacted] Саенко Игорь Борисович

старший научный сотрудник лаборатории Проблем компьютерной безопасности
к.т.н., доцент [redacted] Новикова Евгения Сергеевна

Наши реквизиты: Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН); юр.адрес: 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178; тел.: +7 (812) 328 33 11, факс: +7 (812) 328 44 50, электронная почта: info@spcras.ru.

«Я, Ронжин Андрей Леонидович, директор СПб ФИЦ РАН даю согласие на включение своих персональных данных в документы, связанные с работой диссертационного совета и их дальнейшую обработку»

доктор технических наук, профессор РАН [redacted] Ронжин Андрей Леонидович

«Я, Саенко Игорь Борисович, главный научный сотрудник лаборатории «Проблем компьютерной безопасности» СПб ФИЦ РАН даю согласие на включение своих персональных данных в документы, связанные с работой диссертационного совета и их дальнейшую обработку»

доктор технических наук, профессор [redacted] Саенко Игорь Борисович

«Я, Новикова Евгения Сергеевна, старший научный сотрудник лаборатории «Проблем компьютерной безопасности» СПб ФИЦ РАН даю согласие на включение своих персональных данных в документы, связанные с работой диссертационного совета и их дальнейшую обработку»

кандидат технических наук, доцент [redacted] Новикова Евгения Сергеевна