Федеральное государственное бюджетное образовательное учреждение
высшего образования
Калининградский государственный технический университет
(КГТУ)

На правах рукописи

ПОДТОПЕЛЬНЫЙ
Владислав Владимирович

МОДЕЛИ И МЕТОДИКА ОПРЕДЕЛЕНИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ АТАКУЮЩИХ ВОЗДЕЙСТВИЙ НА СИСТЕМЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ПРИ АУДИТЕ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Специальность 2.3.6 — Методы и системы защиты информации,
информационная безопасность

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель:
к.т.н, доц. Ветров И. А.

Калининград – 2025

ОГЛАВЛЕНИЕ

Введение 4

Глава 1. Анализ специфики информационной безопасности систем искусственного интеллекта 14

1.1 Специфика аудита безопасности систем искусственного интеллекта 14

1.2 Защищаемые информационные ресурсы и существующие угрозы 19

1.3 Анализ научных подходов к определению атакующих воздействий 27

1.4 Особенности построения ПИИ и определение специфики атакующих воздействий 35

1.5 Постановка задачи исследования 39

Выводы к главе 1 42

Глава 2. Специфика применения методов моделирования 44

2.1 Специфика рассмотрения атакующих воздействий на ПИИ в приложении к МППР 44

2.2 Определение параметров модели атакующих воздействий на ПИИ 49

2.3 Определение особенностей параметров модели 61

2.4 Определение ограничений при моделировании атак 93

Выводы к главе 2 110

Глава 3. Модели определения последовательностей атакующих воздействий на системы искусственного интеллекта как подсистемы ИС 112

3.1 Определение правил моделирования 112

3.2 Общие модели на основе МППР для анализа специфики атаки в режиме on-line и off-line 118

3.3 Модели на основе упрощенной классификации тактик 128

3.4 Модель с учетом всех тактик методики описания сценария атаки 155

3.5 Общая методическая последовательность использования моделей 167

Выводы к главе 3 176

Глава 4. Экспериментальная оценка разработанных моделей 178

4.1. Сравнение предлагаемого моделирования с альтернативными 178

4.2 Описание тестового примера и экспериментальная оценка определения путей атаки с использованием моделей МППР 189

4.3 Экспериментальная оценка применительности моделирования при использовании Q – обучения 199
4.4 Сравнительный анализ с альтернативными решениями 206
4.5 Особенности программного решения 213
Выводы к главе 4 222
Заключение 223
Список использованной литературы и электронных ресурсов 225
ПРИЛОЖЕНИЕ А 242
ПРИЛОЖЕНИЕ Б 246

Введение

Актуальность темы диссертации. Одним из наиболее распространенных методов, используемых злоумышленниками для осуществления атак на подсистемы искусственного интеллекта (ПИИ) на основе машинного обучения (в том числе на основе нейронных сетей) информационных систем является использование ложных данных при их вводе в вычислительную модель или нарушение логики работы ПИИ. Подсистемы ИИ, при получении данных, определяют их принадлежность к определённой группе информации (нейтральные, заслуживающие доверия, ложные) с учетом принятой классификации, и реагируют соответствующим образом. Цель подобных атак заключается в изменении параметров (диапазонов значений) механизма классификации в соответствии с требованиями злоумышленника без потери доверия к нему, как к источнику данных. Для выявления и блокирования атак подобного типа важно выстраивать оптимальную стратегию защиты с учетом знаний о возможных действиях (сценариях атаки) со стороны нападающего на ПИИ.

Можно выделить следующие проблемы при использовании машинного обучения, которые могут повлиять на безопасность обработки данных в ИС:

1. Необходимость обеспечения чистоты и достоверности данных. Требуется обеспечить качество данных.
2. Необходимость обеспечения корректного отбора признаков.
3. Потребность в значительных вычислительных ресурсах и времени, особенно при работе с большими объемами данных.
4. Учет опыта специалистов в сфере ИИ для корректной настройки и улучшения моделей.
5. Необходимость обеспечения баланса скорости и качества обработки данных.

Приведенные проблемы позволяют реализовать несколько основных типовых наборов атак на системы ИИ:

1. Атаки «белого ящика» подразумевают полный доступ к модели машинного обучения, включая ее архитектуру, параметры и данные.

2. Атаки типа «черный ящик» подразумевают известность только входных и выходных данных модели. Однако с помощью различных методов, таких как внедрение шума в данные или анализ выходных данных, атакующий может искажать выводы модели или даже извлекать некоторую информацию о ее внутреннем устройстве.

3. Атаки типа «серый ящик» основываются на частичной известности злоумышленнику используемой модели ИИ.

4. Атаки отравляющие данные. Этот тип атаки заключается во внесении изменений в обучающие данные модели.

5. Атаки, использующие уязвимости программной и аппаратной среды.

Следует отметить, что различные вычислительные модели в разной степени уязвимы к указанным атакам. В целом, уровень уязвимости зависит от двух факторов: известности и распространённости моделей (они могут быть типовыми и индивидуальными), и распространенности и доступности обучающих данных.

Не все из существующих методов моделирования одинаково применимы к задачам анализа атак подобного рода, поскольку уязвимости вычислительных моделей ИИ, а также их подсистем, осуществляющих сбор и обработку данных ИИ, достаточно специфичны. Атаки могут реализовываться на основе эксплуатации заданной неточности работы моделей ИИ, на основе манипуляции с исходными обучающими выборками и т.п. Таким образом, при аудите информационной безопасности возникает необходимость поиска возможных последовательностей атакующих воздействий, которые включают в свой состав эксплуатацию специфических особенностей (уязвимостей) новых интеллектуальных технологий, встраиваемых в современные информационные системы (ИС).

Цель диссертационного исследования заключается в повышении степени полноты описания атак на системы искусственного интеллекта при аудите их информационной безопасности.

Решаемая научно-техническая задача: разработка моделей и методики определения последовательностей атакующих воздействий на системы искусственного интеллекта для повышения качества процессов аудита ИБ ПИИ.

Решаемая научно-техническая задача: разработка моделей и методики определения атакующих последовательностей (вектора атаки) на системы (вычислительные модели) искусственного интеллекта для повышения качества процессов аудита ИБ ПИИ.

Научная задача заключается в разработке модели и методик построения и анализа атак на системы ИИ на основе методов марковских процессов принятия решений с учетом современной специфики формирования сценариев атак и рекомендаций в области ИБ при аудите информационных систем. 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. В контексте атак на ИИ, это означает, что модель должна включать все возможные стратегии атаки и реакции системы на эти атаки.

Для достижения данной цели в диссертационной работе поставлены и решены следующие частные задачи:

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

2. Определены возможности использования доступных наборов данных для разработки модели анализа атак на системы ИИ и определены параметры, используемые при моделировании.

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

4. Разработаны методические рекомендации и алгоритм, позволяющие моделировать атакующие воздействия в процессе аудита ИБ ПИИ, изучать их динамику, использовать для составления сценарии атак.

5. Разработана архитектура программного решения, позволяющая автоматизировать процессы моделирования и провести оценку работоспособности моделей.

Объектом исследования являются атаки, направленные на эксплуатацию уязвимостей моделей и архитектур подсистем искусственного интеллекта в контексте общей архитектуры корпоративной информационной системы, процессы построения и анализа атакующих последовательностей для повышения качества аудита защищенности систем искусственного интеллекта.

Предметом исследования выступают вычислительные модели искусственного интеллекта, методики и алгоритмы моделирования атак на подсистемы ИИ.

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

1. С применением предложенного аппарата марковских процессов принятия решений (МППР) появится возможность с помощью математических методов обосновать построение вектора атаки как последовательности событий безопасности, прогнозировать появление данных событий с учетом специфики подсистемы ИИ и с учетом руководящих документов и баз знаний, используемых 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, оперативное реагирование 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.

Основные теоретические положения диссертации были успешно апробированы на различных научных конференциях, как всероссийского, так и международного уровня. Кроме того, важнейшие результаты работы были опубликованы в ведущих рецензируемых научных изданиях.

Теоретическая и практическая значимость результатов исследования. Разработанные модели представляют собой научно-методическую основу для выявления и обоснования формируемых в процессе аудита ИБ ИИ сценариев атак, учет которых позволит повысить качество защиты ИИ. Практическая реализация позволяет построить прогноз проявления событий безопасности, основываясь на принципах построения марковских моделей. При этом повышается точность и полнота построения вектора атаки, что позволяет на практике эффективно применять разработанный подход к формированию модели угроз.

Методология и методы исследования. Используемые в диссертации в качестве математических положений применены МППР для построения вектора атак на ПИИ, методы машинного обучения как вспомогательные элементы для регистрации и анализа событий с применением аналитико-статистических методов.

Апробация результатов. Научные результаты, полученные в диссертации, внедрены в научно-исследовательскую работу, образовательный процесс и практику деятельности в ФГБОУ ВО «Калининградский государственный технический университет» (г. Калининград), ФГАОУ ВО «Балтийский федеральный университет им. И. Канта» (г. Калининград), ООО «Центр защиты информации» (г. Калининград), ОКБ Пеленг (г. Екатеринбург).

Основные положения и результаты докладывались и обсуждались на следующих конференциях:

1. Международной научно-практической конференции VIII Международного Балтийского морского форума (Калининград, 2020 г).

2. III Международной научной конференции «Экосистемы без границ - 2022» IX Международного Балтийского морского форума (Калининград, 2021 г).

3. X Национальной научной конференции с международным участием «Морская техника и технологии. Безопасность морской индустрии» в рамках X Международного Балтийского морского форума (Калининград, 2022 г).

4. XVIII Всероссийской научно-практической конференции «Информационная безопасность цифровой экономики» в рамках форума информационной безопасности «Сибирь-Дальний Восток-2022» (Хабаровск, 2022).

5. XI Национальной научной конференции с международным участием «Морская техника и технологии. Безопасность морской индустрии», в рамках XI Балтийского морского форума (Калининград, 2023 г).

6. XIX Всероссийской научно-практической конференции «Информационная безопасность цифровой экономики» в рамках форума информационной безопасности «Сибирь-Дальний Восток-2023» (Улан-Удэ, 2023).

7. Всероссийской научно-технической конференции «Актуальные проблемы радиоэлектроники и телекоммуникаций» (Самара, 2024).

8. V Всероссийской научно-практической конференции «Социотехнические и гуманитарные аспекты информационной безопасности» (Пятигорск, 2024).

Публикации. Основные результаты диссертации изложены в 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 семь свидетельств о государственной регистрации программ для ЭВМ.

Личный вклад соискателя. Все выносимые на защиту результаты получены лично автором. Лично разработаны модели атакующих воздействий применительно к ПИИ, определены обоснования для повышения ИБ ПИИ. Существенно развит метод обоснования мероприятий ИБ, исходя из динамики изменения атакующих последовательностей.

Структура и объем работы. Диссертационная работа изложена на 250 машинописных страницах, включает 4 главы, 50 рисунок, 20 таблиц и список литературы (136 наименований).

В первой главе диссертации проведен анализ целей, задач и возможностей способов моделирования атакующих воздействий с использованием марковских моделей принятия решений. Рассмотрены особенности аудита систем ИИ с учетом известных способов описания атак (MITRE ATLAS и Методики ФСТЭК). Проанализированы современные способы атак на системы ИИ, а также ПИИ, определена их специфика. Выделены проблемы при использовании машинного обучения: обеспечение чистоты и достоверности данных; корректный отбор признаков; допустимые отклонения в классификации данных. Исследуются атаки «белого ящика» (полный доступ к ПИИ), «черного ящика» (известны только входные и выходные данные модели), «серого ящика» (данные частично известны злоумышленнику), отравляющие данные, использующие уязвимости программной и аппаратной среды. Опасность уязвимости зависит от известности и распространенности моделей, обучающих данных и уязвимостей компонентов ПИИ.

На основе анализа методов моделирования атак предложены марковские процессы принятия решений (МППР) как метод, позволяющий формировать сценарии атаки на ПИИ. Сформулированы задачи исследования: разработка моделей атак на системы ИИ, алгоритмов построения и модификации последовательности атакующих воздействий, методики применения моделирования атак, архитектуры системы определения и анализа атак.

Во второй главе определяется специфика применения методов моделирования атак на ПИИ с использованием марковских процессов принятия решений. Основным параметром в описании вектора состояний является функция ценности, тип и вероятность перехода некого компрометированного состояния в новое в соответствии с последовательностью классифицируемых в методиках описания атак тактик. На основе параметров вероятностей переходов и начальных состояний формируется граф состояний атаки. Для этого необходимо ассоциировать тактики и состояния, которые будут являться вершинами графа, описывающего марковские

процессы. По требованиям методики предполагается их последовательное соединение. Для атакующих воздействий вводятся правила и ограничения. Определяется два представления о переходах в марковских процессах при моделировании атак:

1. Передвижения злоумышленника с помощью действий по состояниям (актуализациям уязвимостей) рассматриваются в режиме имитации развертывания атаки без априорного предположения о логичности действий злоумышленника (имитация в реальном времени (on-line)).

2. Второй способ описания переходов предполагается использовать при плановом аудите, в режиме off-line. Методика позволяет определить более эффективный путь атаки, исходя из того, что при построении сценария атаки не повторяются уже прошедшие этапы в силу того, что возвратные состояния отдаляют злоумышленника от целевого состояния или объекта системы.

Формируются последовательности атакующих воздействий в виде набора вершин графа на основе входных данных, описывающих состояние системы ИИ. Состояния и действия злоумышленника при атаке (с учетом уровня их детального описания) приводятся в соответствии с тактиками методологии MITRE ATLAS (далее MITRE), а их осуществление соответствует наступлению состояний, фиксирующих успех действия на некотором этапе атаки.

Осуществляется перевод оценок точности нейросетей, таких как AUC, точность (Accuracy) и полнота (Recall), в метрики CVSS с учетом того, что CVSS предназначен для оценки уязвимостей и их влияния на безопасность, тогда как AUC и другие метрики относятся к производительности моделей классификации. Определяется специфика формирования функции вознаграждения.

В третьей главе приводятся модели, построенные с учетом специфики марковских процессов принятия решений. Модель нарушителя (как и доступные ему способы взаимодействия с атакуемой системой) определяется в соответствии с представляемыми моделями атакующих воздействий, в которых учитывается известность о компонентах ПИИ и специфика доступности компонентов ПИИ, местоположение нарушителя относительно ПИИ.

При моделировании используется, как основной, метод итераций по значениям для оценки функций полезности состояний. Каждая функция полезности обновляется на основе текущих значений и максимизации ожидаемого вознаграждения с учетом вероятностей переходов. Методы фокусируются на оценке ценности состояний или пар «состояние-действие». Они вычисляют ожидаемую полезность (или ценность) для каждого состояния или действия и используют эти оценки для выбора оптимальной стратегии.

При моделировании вводятся уровни абстракции, которые определяются как степени детализации состояний атакующих воздействий. Причина введения уровней абстракции – сложность описания сценария атаки при наличии множества учитываемых параметров и компонентов. Выявляются три уровня абстракции.

Первый уровень предполагает: состояния модели определяются как наборы этапов-состояний, позволяющих реализовать множества действий, связанных с логикой функционирования вычислительной модели ПИИ в период атакующих воздействий. Второй уровень предполагает: состояния модели определяются как наборы этапов компрометации в процессе атаки, позволяющих реализовать множества действий, связанных с инфраструктурой и логикой функционирования ПИИ в период атакующих воздействий. Данная модель может использоваться для общего описания системы в период атаки с уточнением последовательности специфики действий злоумышленника, связанных с этапами реализации атакующих воздействий. Третий уровень предполагает: состояния модели определяются как наборы событий, позволяющих реализовать множества действий, приведенных в тактиках MITRE и Методике ФСТЭК, связанных с инфраструктурой и логикой функционирования модели и инфраструктуры ПИИ в период реализации атакующих воздействий. Данная модель может использоваться для частного и подробного описания системы в состоянии атаки с уточнением специфики действий злоумышленника.

В четвертой главе приводится экспериментальная оценка разработанных моделей. Полнота при моделировании атак на ПИИ с использованием МППР в дан-

ном случае представляется как способность модели учитывать все состояния, возникающие при воздействии атакующих действий на систему ИИ, и переходы между ними. Приводится определение полноты.

Производится применение моделей. В качестве входных данных использовались метрики уязвимостей, которые были классифицированы в соответствии с методами эксплуатации уязвимостей, и, следовательно, сопоставлены тактикам MITRE (следует учесть, что оценка уязвимости может носить экспертный характер, если четкого соотнесения с тактиками не наблюдается).

Архитектура предложенного приложения включает в свой состав следующие элементы:

1. Модуль сбора данных о параметрах атак (клиентская часть).

2. Модуль анализа данных и сохранения результатов анализа (серверная часть).

3. Анализатор событий или других параметров.

4. Консультационная подсистема, использующая МППР, интегрированный в общую подсистему консультации.

Глава 1. Анализ специфики информационной безопасности систем искусственного интеллекта

1.1 Специфика аудита безопасности систем искусственного интеллекта

Аудит систем искусственного интеллекта является критически важным процессом, который помогает обеспечить безопасность и эффективность применения технологий искусственного интеллекта (ИИ). Он требует комплексного подхода и включает в себя множество аспектов, от анализа данных до тестирования алгоритмов. В условиях растущего использования ИИ в различных сферах, аудит становится необходимым инструментом для управления рисками и повышения доверия к этим технологиям. В частности, аудит помогает выявить уязвимости в системах ИИ, которые могут быть использованы злоумышленниками для атак, таких как отравление данных или манипуляция моделью, позволяет определить сценарии нападений потенциальных злоумышленников, и, соответственно, позволяет оценить защищенность функциональных компонентов систем, а также данных, используемых для обучения и работы моделей. В процессе аудита выявляются недостатки вычислительных моделей, такие как низкая точность. Кроме того, аудит помогает организациям идентифицировать и оценить риски ИИ, что позволяет разработать стратегии для их минимизации [1-7].

Аудит систем ИИ требует многоуровневого подхода, который включает в себя анализ данных, моделей, алгоритмов и инфраструктуры; он может включать как технические аспекты (например, проверка кода и архитектуры), так и организационные (например, оценка процессов управления и контроля). Важной частью аудита является: оценка качества и целостности данных, используемых для обучения моделей, что подразумевает проверку на наличие полноты и актуальности данных; тестирование систем ИИ на устойчивость к атакам, таким как «adversarial attacks», чтобы оценить, насколько модели защищены от манипуляций [8]. В итоге создается набор сценариев атак с учетом зафиксированных угроз и уязвимостей [9-

10]. Соответственно, важной частью аудита является моделирование атак для последующего создания документации и отчетов, которые фиксируют результаты аудита, выявленные проблемы, и позволяет дать рекомендации по их устранению [11]. Последовательность проведения аудита ИИ на предприятии предполагает следующие этапы [12-16]:

1. Определение целей и объема аудита. На этом этапе важно установить, что именно будет проверяться: безопасность данных, соответствие регуляторным требованиям, надежность алгоритмов и т. д. Также нужно определить границы аудита: какие системы и процессы будут охвачены.

2. Сбор информации. Осуществляется сбор документации, информации об архитектуре системы, информации о протоколах взаимодействия компонентов и других материалов, которые позволяют понять, как работает система ИИ и какие данные она использует. При этом доступ к необходимой информации может быть ограничен, а документация может быть устаревшей или неполной.

3. Анализ архитектуры системы. Производится изучение архитектуры ИИ-системы для выявления потенциальных уязвимостей: анализ используемых алгоритмов, источников данных и методов обучения.

4. Идентификация угроз. Производится определение возможных угроз для системы ИИ как подсистем ИС, включая внутренние и внешние риски.

5. Разработка сценариев атак. Предполагается создание реалистичных сценариев атак на основе выявленных угроз и уязвимостей. Сценарии могут описывать следующее: атаки на данные (например, отравление данных), атаки на модель (например, обход защиты) и физические атаки на оборудование. Возникают затруднения при попытке определить все возможные сценарии атак. Для этого требуется разрабатывать различные методики составления сценариев с использованием вероятностных методов

6. Тестирование и анализ. На этом этапе реализуется проведение тестов ПИИ на основе разработанных сценариев атак. Это может включать в себя как автома-

тизированные тесты, так и ручные проверки. При этом тестирование может повредить систему или привести к сбоям в ее работе. Также могут потребоваться специальные инструменты и навыки.

7. Оценка результатов. Предполагается анализ полученных данных о безопасности системы после тестирования, оценка уязвимостей и потенциальных последствий их эксплуатации. Интерпретация результатов может быть сложной задачей, особенно если данные противоречивы или неясны.

8. Рекомендации и отчетность. Производится подготовка отчета с выводами и рекомендациями по устранению выявленных уязвимостей и улучшению безопасности системы. Рекомендации могут быть сложными для реализации из-за ограничений бюджета или ресурсов.

Аудит информационной безопасности (ИБ) систем искусственного интеллекта и корпоративных информационных систем (КИС) имеет свои особенности и различия [1-12]. Наглядно демонстрируется ключевые различия и особенности аудита ИБ систем ИИ и КИС в таблице 1.1.

Таблица 1.1 - Ключевые различия и особенности аудита ИБ систем ИИ и КИС

Особенности	Корпоративные информационные системы (КИС)	Системы искусственного интеллекта (ИИ)
1. Определение и контекст	Системы для обработки, хранения и передачи информации (ERP, CRM и др.), поддерживающие бизнес-процессы.	Специализированные системы, использующие алгоритмы машинного обучения, обработки естественного языка и др.
2. Цели аудита	Оценка соответствия требованиям безопасности, целостности данных, доступности и конфиденциальности.	Оценка моделей ИИ, обучающих данных и алгоритмов; проверка объяснимости и справедливости решений.
3. Методы и подходы	Традиционные методы: анализ документации, тестирование систем доступа, оценка политик безопасности.	Специфические методы: анализ алгоритмов на предвзятость, проверка качества данных, оценка рисков, тестирование на устойчивость к атакам.
4. Управление рисками	Риски: утечка данных, несанкционированный доступ, нарушение целостности данных.	Риски: неправильные выводы, предвзятость алгоритмов, утечка конфиденциальной информации, недостаточная прозрачность.
5. Требования регуляторов	Требования регуляторов зависят от отрасли (например, GDPR для персональных данных).	Специфические требования, регулирующие технологии ИИ (например, законы о прозрачности алгоритмов).
6. Документация и отчетность	Отчеты о политике безопасности, процедурах управления доступом и инцидентами.	Отчеты о сборе данных для обучения, тестах на справедливость и объяснимость моделей.

Сценарий атаки занимает важное место при аудите подсистем искусственного интеллекта, так как он помогает выявить уязвимости и потенциальные риски [10, 17]. Вот несколько ключевых аспектов, указывающих на важность использования сценариев атак:

1. Выявление уязвимостей [9-10,12]. Сценарии атаки позволяют определить слабые места в системе ИИ, такие как:

- уязвимости в алгоритмах машинного обучения;

- ошибки в обработке данных;

- недостатки в архитектуре системы.

2. Оценка устойчивости. Аудит с использованием сценариев атаки позволяет оценить, насколько система устойчива к различным видам угроз, включая:

- атаки на целостность данных (например, отравление данных);

- атаки на конфиденциальность (например, извлечение чувствительной информации);

- атаки на доступность (например, DDoS-атаки).

3. Проверка механизмов защиты. Сценарии атаки помогают проверить эффективность существующих механизмов защиты и реагирования на инциденты. Это включает:

- тестирование систем обнаружения вторжений;

- оценку стратегий шифрования и аутентификации и др.

4. Разработка стратегий «mitigations». Анализ сценариев атак позволяет разработать стратегии по снижению рисков, включая:

- обновление алгоритмов для повышения их устойчивости;

- внедрение дополнительных уровней защиты.

5. Обучение и повышение осведомленности. Сценарии атак могут быть использованы для обучения сотрудников и повышения их осведомленности о потенциальных угрозах и методах защиты.

6. Соответствие стандартам и регуляциям. Многие стандарты требуют проведения оценки рисков, включая анализ потенциальных атак.

Специфика проведения аудита систем искусственного интеллекта (ИИ) может зависеть от конкретных целей, методологий и используемых подходов. Можно отметить, что многие российские исследования по оценке рисков информационной безопасности (ИБ) используют методологию оценки угроз, основанную на законодательстве РФ и стандарте ISO 27005 (Таблица 1.2). В этих исследованиях применяются различные математические методы, такие как теория графов, байесовский подход, методы статистики и экспертных оценок. Однако, в большинстве исследований угрозы ИБ рассматриваются с точки зрения защищающейся стороны (организации), а не атакующей.

Таблица 1.2 – Перечень ключевых стандартов в области аудита ПИИ

Стандарт/рекомендация	Описание
1. ISO/IEC 27001 и ISO/IEC 27002*[5]*	Стандарты управления информационной безопасностью, применимые к системам ИИ для защиты данных, используемых в обучении моделей.
2. ISO/IEC JTC 1/SC 42	Группа, разрабатывающая международные стандарты для ИИ, охватывающая управление рисками, оценку и валидацию алгоритмов ИИ.
3. IEEE P7000 Series	Серия стандартов от IEEE, направленная на описание процессов проектирования и использования ИИ
4. NIST AI Risk Management Framework	Фреймворк от NIST, помогающий управлять рисками, связанными с ИИ, включая рекомендации по аудиту и оценке систем ИИ.
5. OECD Principles on Artificial Intelligence	Принципы от OECD, подчеркивающие важность этичного использования ИИ, включая прозрачность, ответственность и защиту прав человека
6. EU AI Act (проект)	Законопроект Европейской комиссии о регулировании ИИ, включающий требования к аудиту и оценке рисков для высоких рисков ИИ-систем
7. Guidelines for Trustworthy AI (EU)	Рекомендации Европейской комиссии по созданию надежного ИИ, включая принципы прозрачности и подотчетности
8. ACM Code of Ethics	Кодекс этики ACM, включающий рекомендации по этическому использованию технологий, включая ИИ
9. ISO 9241-210	Стандарт по эргономике взаимодействия человека с системой, полезный для оценки пользовательского интерфейса и взаимодействия с системами ИИ

Важно подчеркнуть, что аудит ИИ требует соблюдения ряда российских стандартов и рекомендаций, поскольку они являются основой для проведения мероприятий по обеспечению безопасности подсистем ИИ. Однако стоит отметить, что этому подходу присущи ограничения российской нормативно-правовой базы.

В частности, можно отметить максимально широкое использование мнения экспертов при составлении сценариев атак. При этом полученная информация учитывается скорее качественно (с помощью таблицы градаций, где значение определяется экспертом), а не количественно. Использование экспертного подхода при оценке вероятности угроз и рисков информационной безопасности (ИБ) имеет следующие недостатки: субъективность, отсутствие полноты или наличие избыточности при оценке угроз. Соответственно, необходимо минимизировать влияния субъективного мнения на процессы аудита.

1.2 Защищаемые информационные ресурсы и существующие угрозы

В процессе функционирования систем искусственного интеллекта (как и подсистем ИИ) используется несколько ключевых элементов, которые чаще всего подвержены атакующим воздействиям. Состав и специфика этих элементов рассмотрены в работах Д. Найомита (исследователь подробно рассматривает большинство аспектов организации атак на ПИИ) [10]. В группу основных атакуемых компонентов входят:

1. Данные. Данные являются фундаментальным элементом, необходимым для обучения моделей. Они могут быть структурированными (таблицы, базы данных) или неструктурированными (текст, изображения, видео). Качество и количество данных напрямую влияют на эффективность обучения моделей.

2. Алгоритмы машинного обучения. Алгоритмы машинного обучения представляют собой набор методов и техник, используемых для анализа данных и построения моделей.

3. Модели. Модели представляют собой обученные структуры, которые могут делать предсказания или принимать решения на основе входных данных. Они являются результатом применения алгоритмов машинного обучения к данным. Специфика их работы создает возможность манипулирования системой классификации.

4. Интерфейсы. Интерфейсы обеспечивают средства взаимодействия между пользователями и системой. Они могут включать в себя API, графические интерфейсы пользователя (GUI). В данном случае они обеспечивают взаимодействие ИИ с пользователями, и, в том числе, со злоумышленниками.

5. Инфраструктура. Инфраструктура состоит из аппаратных и программных средств, необходимых для хранения данных, обучения моделей и развертывания систем (серверы, облачные вычисления, системы хранения данных, фреймворки машинного обучения и другие компоненты). Эти компоненты работают вместе, образуя основу для построения систем ИИ и машинного обучения (Рисунок 1.1).

Использование систем ИИ подразумевает несколько процессов:

1. Сбор данных. Осуществляется процесс получения данных из различных источников, включая базы данных, веб-скрейпинг, сенсоры и другие. Это начальный этап, на котором данные собираются из различных источников.

2. Предобработка данных. Производится обработка данных для устранения аномалий. Данные очищаются и подготавливаются для анализа. Этот процесс связан с компонентами «Данные» и «Модели», так как качество данных влияет на обучение моделей.

3. Обучение модели. Подразумевается использование алгоритмов для создания модели на основе подготовленных данных. При реализации процесса осуществляется выбор алгоритма, настройка гиперпараметров и обучение на тренировочных данных. Модель обучается на предобработанных данных.

4. Валидация модели. Производится оценка производительности модели на валидационных данных для предотвращения переобучения.

5. Тестирование модели. Производится проверка модели на тестовых данных, которые не использовались в процессе обучения. После обучения модель оценивается с использованием тестовых данных.

6. Развертывание. Производится внедрение модели в рабочую среду, где она может принимать входные данные и делать предсказания.

7. Мониторинг и обновление. Осуществляется постоянный мониторинг производительности модели в реальном времени и ее обновление по мере поступления

новых данных, что позволяет системе оперативно адаптироваться к изменениям и поддерживать высокую эффективность.

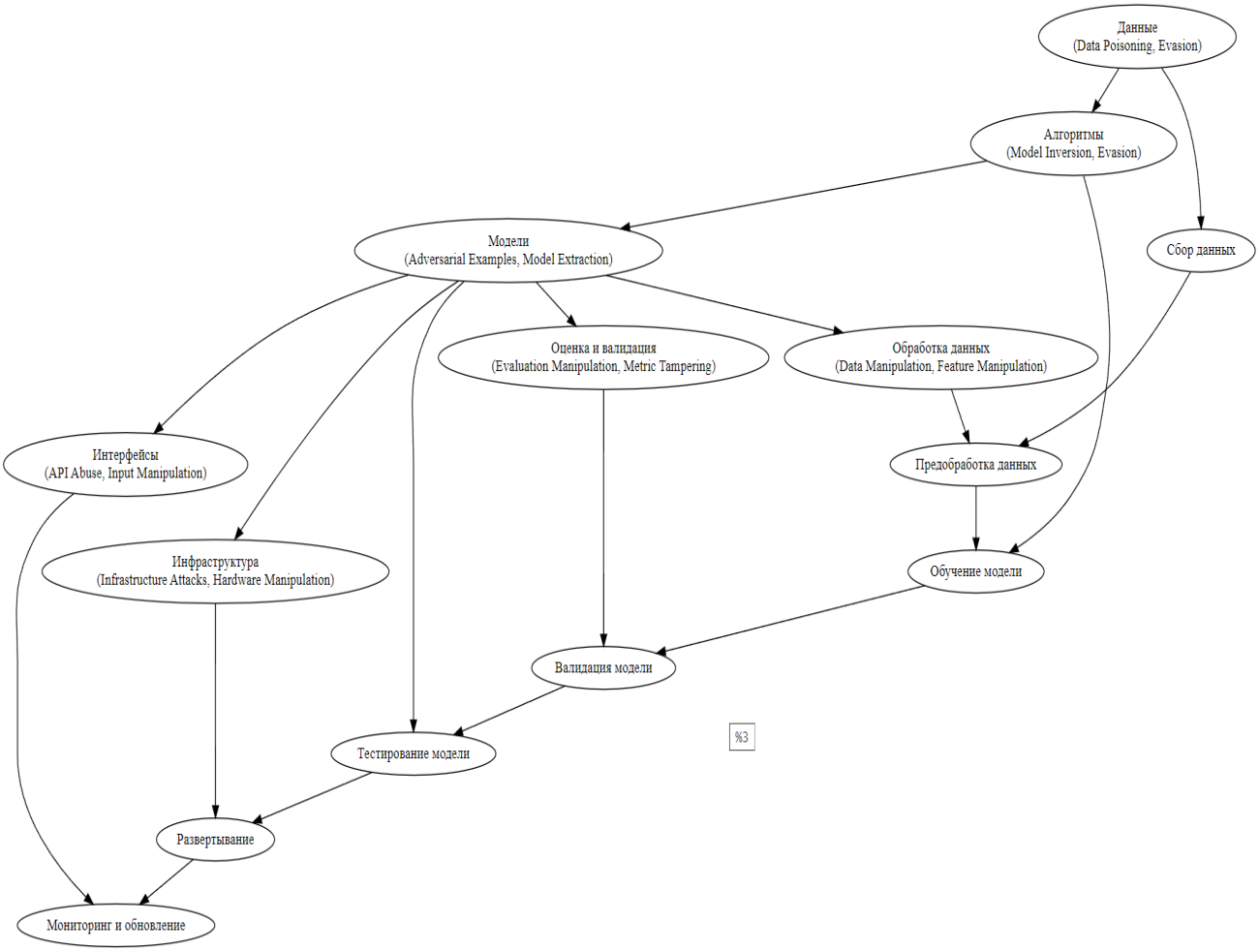


Рисунок 1.1 – Порядок работы ПИИ

При анализе связей между элементами системы ИИ и ее процессами можно отметить следующее (Рисунок 1.2):

1. Данные служат основой для обучения моделей, а алгоритмы машинного обучения используют эти данные для создания предсказаний.

2. Модели зависят от алгоритмов и данных, а также от процессов валидации и тестирования для обеспечения их точности.

3. Интерфейсы обеспечивают связь между пользователями и системой, позволяя получать данные и выводы от моделей.

4. Обработка и предобработка данных являются необходимыми шагами перед обучением моделей, обеспечивая высокое качество входных данных.

5. Мониторинг и обновление моделей обеспечивают их актуальность и эффективность в изменяющихся условиях.

Эти компоненты и процессы формируют основу для построения систем ИИ и машинного обучения, позволяя им эффективно выполнять задачи классификации. Кроме того, следует отметить отличительные особенности систем ИИ в сравнении с корпоративными (традиционными) информационными системами (КИС). Корпоративные системы — это инструменты для стабильных, регламентированных процессов, тогда как ИИ — это гибкие решения для неопределённых, динамичных задач. Для КИС (ИС) характерно следующее:

– модели детерминированные, основаны на реляционной алгебре или бизнес-правилах;

– чаще всего архитектура централизованная, с чёткими интерфейсами (например, клиент-сервер);

– изменяются только при явной модификации кода или конфигурации;

– традиционные ИС не способны к самообучению (требует участия разработчиков).

В то же время для систем ИИ характерно следующее:

– модели вероятностные (нейросети, деревья решений, байесовские сети), при этом результаты зависят от качества данных и обучения;

– архитектура децентрализованная, часто облачная, с распределёнными вычислениями (например, обучение на GPU-кластерах);

– сложно интерпретировать решения систем ИИ;

– способны улучшать производительность по мере накопления данных (online learning);

– системы ИИ могут автоматически адаптироваться к изменениям (например, обнаружение аномалий в реальном времени).

При этом следует отметить, что аудит информационной безопасности систем ИИ сталкивается с рядом проблем, которые отличают его от традиционного аудита информационных систем: современные модели ИИ, особенно глубокие нейронные сети, являются «черными ящиками», что затрудняет анализ их уязвимостей и возможных атак; данные могут быть скомпрометированы, что приведет к некорректным предсказаниям; модели постоянно обновляются и переобучаются (не всегда известны точные данные о параметрах ИИ); аудит осложняется отсутствием внедренных общепринятых стандартов и методик для анализа безопасности систем ИИ (при этом уже существуют базы описания действий злоумышленника (тактик и техник), подобные MITRE ATLAS) [18, 19]. При этом составление сценария атаки на ИИ является критически важным этапом аудита.

Сценарии атак помогают обнаружить слабые места в системе, которые могут быть использованы злоумышленниками, а также позволяют проверить эффективность механизмов защиты.

Таким образом, аудит с использованием сценариев атак помогает подготовить систему к возможным реальным атакам. В этом случае проявляются преимущества применения МППР в задачах аудита. Методы МППР позволяют описать последовательность действий злоумышленника, принимая во внимание вероятностный характер атак и адаптивное поведение злоумышленника, зависящее от состояния системы. Методы МППР помогают находить оптимальные стратегии для атак, оценивать вероятности успеха и долгосрочные последствия, характерные для реальных атак.

При исследовании были сопоставлены компоненты системы ИИ и уязвимости и техники методик [17, 19]. Распределения и ассоциации техник и тактик с компонентами и процессами системы искусственного интеллекта, приводятся в таблице 1.3.

Уровень основных функции систем ИИ (функциональные блоки):

Блок 1 отвечает за функции приема (интерфейс доступа), обработки входных данных (модель ИИ), вывод результатов.

Блок 2 представляет датасеты (сбор, подготовка (предобработка), хранение).

Уровень технических систем, обеспечивающих реализацию функций Блока 1 и 2:

Блок 3 представляет **т**ехническую инфраструктуру (системы и сервисы, реализующие логику работы ИИ)

Рисунок 1.2 – Пример порядка работы ПИИ с учетом атакующих воздействий

Таблица 1.3 – Пример сопоставления распространенных компонентов системы ИИ, уязвимостей и техник методик описания атак (MITRE ATLAS)

Этап/Компонент	Описание	Тактики MITRE ATLAS	Техники MITRE ATLAS
1. Сбор данных	Сбор необходимых данных для обучения модели ИИ из различных источников.	Отравление данных (Data Poisoning)	- Манипуляция данными (Data Manipulation) - Data Exfiltration - Сбор данных через социальную инженерию (Social Engineering Data Collection) - Атака на конфиденциальность данных (Data Confidentiality Attack) - Атака на доступность данных (Data Availability Attack)
2. Предобработка данных	Очистка и преобразование данных в формат, пригодный для обучения.	Атака на целостность данных (Data Integrity Attack)	- Атака на этапе предобработки (Adversarial Preprocessing) - Манипуляция признаками (Feature Manipulation) - Отбор признаков (Feature Selection Attack) - Атака на распределение данных (Data Distribution Attack) - Атака на метаданные (Metadata Attack)
3. Обучение модели	Обучение модели на подготовленных данных и настройка ее параметров.	Отравление модели (Model Poisoning)	- Манипуляция гиперпараметрами (Hyperparameter Manipulation) - Реконструкция обучающих данных (Training Data Reconstruction) - Атака на архитектуру модели (Model Architecture Attack) - Атака на алгоритм обучения (Learning Algorithm Attack) - Атака на целостность модели (Model Integrity Attack)
4. Оценка модели	Тестирование модели на новых данных для оценки ее точности и надежности.	Атака на этапе тестирования (Adversarial Testing)	- Уклонение от модели (Model Evasion) - Манипуляция производительностью (Performance Manipulation) - Атака на метрики (Metrics Manipulation) - Атака на валидацию (Validation Attack) - Атака на тестовые данные (Test Data Attack)
5. Развертывание	Внедрение модели в рабочую среду для обработки реальных данных.	Извлечение модели (Model Extraction)	- Манипуляция выводом (Inference Manipulation) - Инверсия модели (Model Inversion) - Злоупотребление API (API Abuse) - Атака на систему развертывания (Deployment System Attack) - Атака на конфигурацию (Configuration Attack)
6. Мониторинг	Мониторинг для выявления аномалий, необходимости обновлений.	Обнаружение дрейфа модели (Model Drift Detection)	- Обновление безопасности (Security Patching) - Непрерывный мониторинг (Continuous Monitor.) - Атака на систему мониторинга (Monitoring System Attack) - Атака на отчетность (Reporting Attack) - Атака на аномалии (Anomaly Attack)

Таким образом, каждая стадия процесса в системе ИИ связана с определенными тактиками и техниками. Кроме того, можно выделить два основных набора компонентов (функциональных блоков) ПИИ, способных функционировать в режимах обучения и эксплуатации в составе ИС, которые реализуют логику работы ИИ (уровень логики ИИ):

1. Блок датасетов.

2. Блок вычислительной модели, который включает в свой состав собственно вычислительную модель, а также интерфейс доступа к ней в режиме ее эксплуатации.

Особо следует отметить, что интерфейс доступа к системе ИИ позволяет взаимодействовать с ней внешнему нарушителю без непосредственного воздействия на техническую инфраструктуру, которая обеспечивает работу вычислительной модели. Специфические особенности внешнего нарушителя при таких атаках на ИИ включают:

1. Акцент на данные и модели, а не на традиционные системы и сети.

2. Эксплуатацию публичных интерфейсов и API для взаимодействия с ИИ.

3. Манипуляцию данными и моделями для подрыва их целостности.

4. Использование социальной инженерии для получения доступа к критическим ресурсам.

5. Уклонение от обнаружения с помощью обфускации и манипуляции запросами.

6. Минимизацию прямого взаимодействия с инфраструктурой, что затрудняет обнаружение.

Однако нельзя исключать атаки на техническую инфраструктуру с целью компрометации модели ИИ опосредованным образом (через вредоносные включения или повреждение компонентов ИС). Соответственно, к двум блокам, которые актуальны для злоумышленника при атаках на изменение логики работы ИИ, следует добавить технический уровень ИС, связанный с работой ПИИ. Атаки, реализуемые по отношению к компонентам технического уровня, носят традиционный

характер, уязвимости и способы их эксплуатации приводятся в известных базах CWE, CVE, CAPEC и др. [19 – 27].

1.3 Анализ научных подходов к определению атакующих воздействий

При исследовании аспектов безопасности информационных систем используют разные способы определения проблем в области безопасности ИИ. Особое место занимает анализ защищенности систем с применением моделирования компьютерных атак как совокупности последовательных атакующих воздействий. Следует отметить, что не все из существующих методов одинаково применимы к задачам анализа атак в процессе эксплуатации информационных систем. Существует системы, которые не всегда позволяют реализовать приемы пентеста. Отсюда следует, что при аудите информационной безопасности определение возможных путей развертывания атак предполагает построение сценариев, которые могут создаваться и анализироваться как на формальном уровне, при использовании моделей, так и в практической форме (пентест). При этом следует отметить, что определение сценариев атакующих воздействий возможно без учёта конкретного периода времени, которое требуется на проведение атакующих мероприятий. Подобное необходимо в том случае, если моделирование атаки, построение сценария нападения реализуется для выявления эффективных путей атакующих воздействий злоумышленника с целью их превентивной блокировки. Кроме того, анализ атакующих воздействий в режиме реального времени не обязательно предполагает определение наиболее эффективного пути развертывания атаки по критерию наименьших затрат времени, поскольку действия злоумышленника не всегда подчиняются логике поиска наилучшего способа нападения в силу ограниченности квалификации и доступных ему средств нападения.

Таким образом, предполагается наличие двух типов моделирования с точки зрения потребности аудита:

1. Первый тип, позволяет рассматривать последовательность атакующих воздействий во времени с целью определения порядка блокирования атакующих

воздействий как мер, которые разворачиваются либо параллельно действиям злоумышленника, либо в процессе выявления последовательности атакующих воздействий.

2. Второй тип предполагает рассмотрение сценария вне контекста противодействия в режиме реального времени. Такой сценарий ориентируется на выявление эффективного пути достижения цели злоумышленника с последующим превентивным интегрированием мер защиты с целью изоляцией или уничтожения найденного пути нападения.

В целом можно выделить множество методов (деревья атак, различные способы, связанные с использованием машинного обучения, использование байесовских сетей доверия, сетей Петри Маркова, нечетких множеств; теория игр, теория графов, теория случайных процессов), которые использовались для построения сценариев нападения (определения векторов атак) [20 – 49]. Для решения задачи определения последовательности атакующих воздействий сейчас часто используют модели на базе Марковских процессов принятия решений (МППР), поскольку они позволяют учесть фактор неопределенности: марковские процессы позволяют моделировать среду, где результаты действий не всегда предсказуемы. Это особенно важно при анализе компьютерных атак, где атакующие могут изменять свои тактики и стратегии.

Следует отметить, что учитываются в данном случае следующие факторы, влияющие на анализ атак [40 – 49]:

1. Динамический характер процесса. Атаки представляют собой динамический процесс, где существует зависимость между состоянием системы и воздействием на нее. Марковские модели, обладая высокой степенью адаптивности, позволяют эффективно учитывать эту зависимость. Изменения в модели атаки при смене стратегии поведения злоумышленника во время нападения можно отслеживать. Модели могут адаптироваться к новой информации по мере ее поступления, что позволяет постоянно изучать и обновлять политику на основе результатов предыдущих решений.

2. Использование вероятностного подхода. В сфере ИБ часто присутствует неопределенность, связанная с поведением злоумышленников. Подход позволяет учитывать вероятности различных событий и оценивать риски при различных стратегиях принятия решений. Это помогает оптимизировать защиту от компьютерных атак и минимизировать потенциальные угрозы. С другой стороны, МППР позволяет найти оптимальную последовательность атакующих воздействий. При этом МППР не теряет свойства структурированности в описании последовательностей действий злоумышленника. Кроме того, в этом случае можно отметить адаптивность МППР в задачах построения модели атак на системы ИС, которая проявляется через обновление моделей на основе новых данных, учет вероятностных переходов, интеграцию с методами обучения.

3. Возможность учета ранее неизвестных факторов. Марковские модели позволяют включать в анализ широкий спектр факторов, влияющих на вектор атаки, такие как характеристики системы, действия злоумышленника, защитные меры и др.

4. Четкая структура. В рамках МППР четко определены состояния, действия и вознаграждения, что упрощает анализ и информирование о процессе принятия решений. Такой структурированный подход помогает понять сложные взаимодействия внутри системы ИИ [28, 50 – 58]. Это важно для эксперта, составляющего сценарий атак тем, что позволяет описать предметную область и одновременно учитывать фактор неопределённости (облегчает процесс понимания и анализа новых угроз, позволяя более эффективно разрабатывать стратегии защиты) [67– 69].

Другие методы также обладают своими отрицательными и положительными сторонами в решении рассматриваемых задач [70 – 132]:

1. Деревья позволяют представить атаку как иерархическую структуру (каждый узел соответствует определенной стадии атаки, а ветви — возможным действиям злоумышленника). Это позволяет реализовать визуализацию последовательности действий злоумышленника и упростить анализ атак. Однако при многоуровневых атаках в отличие от МППР возникают сложности в учете всех состояний

и переходов для многоуровневых атак, могут не учитываться вероятностные аспекты и неопределенности.

При моделировании атак на информационные системы (ИС) часто используются различные модели, основанные на графах. Некоторые используют байесовские сети, сети Петри и графы атак. Граф атак представляет собой граф, который охватывает все известные траектории или пути, которыми нарушитель может реализовать угрозу. Графы атак анализируются для решения задач обнаружения возможных атак, анализа инцидентов информационной безопасности, оценки соответствия и адекватности средств защиты, а также минимизации рисков от реализации угрозы [8].

Однако графы атак могут иметь проблемы с масштабируемостью, что делает их неприменимыми для распределенных систем с большим количеством хостов и уязвимостей. Для решения этой проблемы применяется другой графовый метод, основанный на деревьях атак. Деревья атак позволяют легко решить проблему масштабируемости, но они сложны для моделирования циклических атак и не могут быть использованы для динамического моделирования. В простом моделировании деревьев атак вершины могут быть двух типов: «И» и «ИЛИ». Ребра графа обозначают различные параметры, такие как: стоимость проведения атаки; время, затраченное нарушителем; сложность реализации атаки и другие.

Также существуют расширенные версии деревьев атак, в которых вводятся дополнительные типы вершин, например, «Order AND», которые показывают, что подцели, соответствующие дочерним вершинам, должны быть достигнуты нарушителем в определенном порядке. Байесовские графы атак основаны на байесовских сетях и представляют собой направленные ациклические графы, в которых вершины соотносятся с инцидентами информационной безопасности, а ребра представляют операции конъюнкции (логическое «И») или дизъюнкции (логическое «ИЛИ»).

Ребра графа атак указывают направление на возможный инцидент, который может произойти при выполнении предшествующих условий. Граф атак такого типа содержит только одну целевую вершину, которая соответствует конкретной

атаке. Значения вероятностей задаются для каждой вершины графа, кроме целевой, и отражают возможность возникновения инцидента. Эти значения вычисляются с использованием формулы условной вероятности. Байесовские графы атак схожи с деревьями атак, но отличаются возможностью учета неопределенности исходных данных о моделируемых атаках. Среди разновидностей байесовских сетей стоит упомянуть скрытые марковские процессы, которые часто используются при моделировании атак из-за удобства исследования путей в пространстве состояний.

2. Методы машинного обучения способны обрабатывать большие объемы данных и выявлять аномалии, что может быть полезно для предсказания атак, кроме того позволяют автоматизировать процесс обнаружения угроз. Однако машинное обучение требует больших объемов данных для обучения и сильно зависимо от их качества, что может быть проблемой в условиях ограниченных ресурсов, массивов доступных данных, также результаты могут быть трудными для интерпретации, в то время как МППР предлагает более понятный порядок построения сценариев атак.

3. Байесовские сети учитывают вероятностные зависимости и неопределенности, что делает их мощным инструментом для анализа сложных сценариев. Позволяют оценивать влияние различных факторов на уровень риска. Однако по сравнению с МППР формирование и настройка байесовских сетей может быть сложнее и требовать значительных ресурсов.

4. Сети Петри подходят для анализа систем с параллельными процессами, что может быть полезно при моделировании атаки и, кроме того, позволяют исследовать временные аспекты атак, что может улучшить понимание динамики угроз. Однако они могут быть сложны для интерпретации, тогда как МППР имеют более интуитивно понятную структуру. Кроме того, при работе с моделями на основе сетей Петри присутствует необходимость тщательного описания всех возможных состояний системы.

5. Нечеткие множества позволяют учитывать неопределенности и субъективные оценки рисков, что может быть полезно в условиях недостатка информации.

Однако, могут потребовать больше вычислительных ресурсов для обработки нечетких данных, а результаты могут быть затруднять интерпретацию результатов, в отличие от более четких вероятностных выводов в МППР.

6. Теория игр позволяет учитывать поведение как атакующих, так и защищающихся, что может дать более полное представление о сценарии атаки. Однако модели теории игр могут быть сложными при реализации и, самое важное в этом случае, требуют точных данных о поведении противника, чего может не хватать при прогнозировании действий потенциального злоумышленника, тем более, что рассчитывать на точные данные сложно при вероятностном анализе и построении надежной модели [133, 134].

7. Модели случайных процессов хорошо подходят для анализа непредсказуемых изменений в системе. Но подобные модели могут быть сложными для построения и интерпретации при анализе действий злоумышленника, что затрудняет их использование на практике, и также требуют значительных вычислительных ресурсов при анализе [134].

Необходимо отметить, что несмотря на явные преимущества в задачах определения сценария атак, МППР-модели имеют некоторую сложность применения: настройка модели подразумевает детальное понимание системы, включая определение состояний, действий и вероятностей переходов [130-131]. Эта сложность может быть препятствием для применения МППР в задачах определения атакующих последовательностей, особенно в больших и динамичных системах. Однако в контексте моделирования атак на ПИИ и составления сценариев атак специалистом в области ИБ при аудите этот недостаток нивелируется конкретизацией задач и введением уровней детализации ПИИ (ИС) (по примеру нотаций IDEF0): аудитор сосредотачивается на системных компонентах и уязвимостях отдельного сегмента (уровне абстракции описания системы) ИС или на определённых типах угроз (предполагаемых типах атак), при более высоком уровне абстрагирования каждый сегмент представляет собой компонент ИС более высокого порядка. Это органично вписывается в методику составления сценариев атак, учитывая сегментиро-

ванность современных информационных систем и систем ИИ. Этим же компенсируется и другой недостаток метода МППР, который заключается в том, что по мере увеличения числа состояний и действий, вычислительные ресурсы, необходимые для решения задач МППР, могут значительно возрасти [134].

Таким образом, в задачах аудита и составления моделей атак, такие особенности метода МППР, как четкая структурированность, простота реализации, учет вероятности атакующих воздействий, возможность определения оптимальной стратегии злоумышленника являются выгодными. Марковские процессы имеют хорошо разработанный теоретический аппарат. Математический аппарат обеспечивает высокую точность расчетов, позволяет исследовать моделируемый процесс на любых интервалах времени и анализировать изменение интересующих показателей во времени. Вероятности нахождения процесса в различных состояниях являются аналитическими критериями оценки и не обладают недостатками статистических показателей. Однако отсутствуют адекватные математические модели оцениваемых систем, что является существенной проблемой.

При этом следует отметить специфику анализа атакующих воздействий. Построение модели предполагает изучение особенностей уязвимостей программно-аппаратного обеспечения, топологические особенности информационных систем (если они распределённого типа), модели нарушителей, существующие способы и механизмы эксплуатации уязвимостей. Таким образом, прежде чем определить порядок реализации атаки, предварительно исследуются целевые информационные системы (ЦИС), специфика взаимосвязей объектов в ней. Поскольку ЦИС бывают разного типа, то атакуемые действия рассматриваются как последовательность эксплуатации уязвимостей, то есть как этапы работы злоумышленника для достижения целевого состояния, которое означает достижения успеха атаки. Таким образом, атаки могут быть представлены как перемещение злоумышленника, наблюдаемого через фиксацию атакующих воздействий, между узлами сети, на которых были обнаружены уязвимости и которые были использованы для достижения целевого объекта. Однако уязвимости разного типа, позволяющие продвинуться зло-

умышленнику на этапе компрометации или захвата ресурсов, при эксплуатации могут предполагать, как локальные, так и межузловые действия злоумышленников. Учитывая эти факторы, сценарий атаки (последовательность действий злоумышленника) в современной методологии построения вектора атак (подразумевается методический документ ФСТЭК «Методика оценки угроз безопасности информации», далее - Методика ФСТЭК) предлагается рассматривать не как перемещение между узлами, а как перемещения между состояниями, которые маркируют повышение уровня компрометации (появление новых возможностей у злоумышленника) системы (объекта системы) от начального состояния атаки (например, Тактика 1(Разведка)) до конечного состояния, которое определяется как достижение целей злоумышленника (Тактика 10 (ФСТЭК) или 15 (MITRE ATLAS)) [2,19]. Таким образом, действия при атаке соответствуют техникам и их тактикам. Реализация тактики рассматривается как успех (наступление некого состояния) в последовательности этапов повышения степени компрометации. Таким образом, анализ защищённости информационной системы представлен как поиск наилучшей последовательности атакующих воздействий (наилучшего вектора атаки) злоумышленника, направленных на повышение компрометации ЦИС. Следует отметить, что на сегодняшний момент есть несколько программных решений, которые позволяют построить вектор атак, например, MulVal. Однако в основном, эти решения используют «логики» описания атак, например, логические выражения на языке Datalog. Подобные подходы не учитывают вероятностные показатели атак в полной мере в сопряжении с методологией их описания по Методике ФСТЭК или MITRE ATLAS [134].

При этом следует учитывать, что метод экспертных оценок, который практикуется в области информационной безопасности в РФ как основной, имеет ряд ограничений, таких как субъективность, неполнота или избыточность информации, а также сложность при повторяемости процесса.

1.4 Особенности построения ПИИ и определение специфики атакующих воздействий

При разработке ПИИ (в данном случае, основанной на нейронной сети, приведенной на рисунке 1.3) следует учитывать вероятность того, что злоумышленник будет пытаться атаковать модель ИИ (с учетом необходимости обхода систем защиты) [38, 39, 60-70]. Существующие вычислительные модели ИИ приведены в таблице 1.4.

Последовательность состязательной атаки показана на рисунке 1.3. К правильно распознаваемым данным добавляется шум, вычисленный на основании функции потерь модели нейронной сети.

Таблица 1.4 – Описание моделей глубокого обучения

Модель	Вид	Метод обучения	Входные данные	Особенности
Сети прямого распространения	Дискриминатор	С учителем	Различные	Нелинейность, адаптивное обучение, устойчивость к ошибкам, обработка последовательности данных во внутренней памяти
Рекуррентные сети	Дискриминатор	С учителем	Последовательность данных, временные последовательности	Совместим с анализом сетевого трафика чувствительного к времени, ошибкам, обработка последовательности данных во внутренней памяти
Сверточные сети	Дискриминатор	С учителем	Вектор, представленный в виде изображения	Необходимость большого набора данных
Автокодировщики	Генератор	Без учителя	Различные	Возможность работы с непомеченными экземплярами данных, извлечение признаков, снижение размерности, реконструкция входных данных
Глубокие генеративные модели	Генератор	Без учителя	Различные	Генерирование новых данных, аналогичных существующим

Модели искусственного интеллекта тесно связаны с данными, на которых они были обучены. Модификация этих данных может существенно повлиять на работу модели.

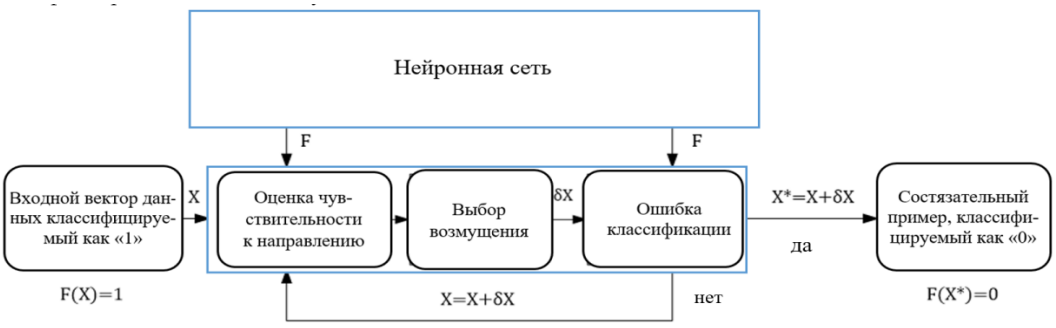


Рисунок 1.3 – Изменение данных для обмана классификатора

Следует отметить, что на этапе эксплуатации системы ИИ могут происходить атаки, в ходе которых создаются наборы данных, неправильно интерпретируемые вычислительной моделью системы. Примером такого воздействия может служить добавление специально созданного шума к входным данным, что может кардинально изменить результат классификации. Кроме того, данные могут быть изменены и на этапе тренировки, когда в обучающий набор данных добавляются записи, призванные снизить качество работы модели или добиться желаемых злоумышленником результатов работы модели. Таким образом, атаки могут быть произведены на следующих этапах работы ПИИ:

– обучение, в том числе выбор и обработка обучающих данных модели (выполняется до активного состояния ПИИ);

– эксплуатация.

Атаки могут быть разными, в зависимости от имеющихся знаний у злоумышленника об атакуемой системе. Различают три основных типа атак на ПИИ: атаки белого ящика, атаки черного ящика, атаки серого ящика. К методам белого ящика можно отнести случаи, когда у злоумышленника имеется полная информация о модели нейронной сети, его архитектуре, а также известны данные обучения. Методы серого ящика – атакующий знает подробную информацию о наборе данных или

типе нейронной сети, ее структуре. Методы черного ящика – злоумышленник ничего не знает о модели, он может только пытаться отправлять информацию в систему и получать результат. Способы атак на модели искусственного интеллекта можно классифицировать по нескольким группам (Таблица 1.5).

Состязательные атаки являются наиболее распространенными атаками на нейронные сети, выполняемые в ходе эксплуатации. Атака заключается в выработке входных данных, которые ошибочно классифицируются моделями нейронных сетей. Атаки отравления направлены на модификацию данных на этапе тренировки модели. Модификация тренировочных данных изменяет поведение модели до нового обучения.

Таблица 1.5 – Сопоставление типов атак и компонентов ПИИ

Тип атаки	Описание атаки	Цель атаки на ПИИ	Примеры методов атаки	Уязвимости ПИИ	Процессы ПИИ	Компоненты ПИИ
Атаки "белого ящика"	Атаки, при которых злоумышленник имеет полный доступ к модели, включая архитектуру и веса.	Изменение входных данных так, чтобы классификатор ошибочно классифицировал их.	L-BFGS, DeepFool, Быстрая адаптивная атака на границу	Доступ к внутренним параметрам модели	Обучение, классификация	Модель, веса, архитектура
Атаки "черного ящика"	Атаки, при которых злоумышленник не имеет доступа к внутренним параметрам модели.	Изменение входных данных, основываясь только на выходных данных модели.	Генерация примеров с использованием градиентного спуска	Неполное понимание модели	Обучение, классификация	Входные данные, выходные данные
Целевые атаки	Атаки, направленные на то, чтобы заставить модель классифицировать входные данные в определенный класс.	Изменение входных данных так, чтобы они были классифицированы как целевой класс.	FGSM, L-BFGS	Уязвимость к манипуляциям с данными	Классификация	Входные данные, выходные данные
Нецелевые атаки	Атаки, при которых злоумышленник не имеет конкретного целевого класса.	Введение в заблуждение классификатора без указания конкретного класса.	DeepFool, Fast Gradient Sign Method (FGSM)	Непредсказуемость модели	Классификация	Входные данные, выходные данные

Продолжение таблицы 1.5

Тип атаки	Описание атаки	Цель атаки на ПИИ	Примеры методов атаки	Уязвимости ПИИ	Процессы ПИИ	Компоненты ПИИ
Семантические атаки	Атаки, которые создают состязательные примеры с большими искажениями, но сохраняют семантическое сходство.	Обман классификатора, сохраняя визуальную схожесть с оригиналом.	Изменение цвета, поворот, сдвиг изображений	Чувствительность к изменениям	Обработка изображений	Входные данные, визуальные данные
Теневые атаки	Атаки, нацеленные на выходную метку классификатора и сертифицируемые средства защиты.	Генерация неблагоприятного примера с учетом защиты модели.	Оптимизация с учетом потерь и штрафов	Уязвимость к защитным механизмам	Обучение, классификация	Модель, функции потерь
Атаки на основе градиента	Атаки, использующие информацию о градиенте для нахождения уязвимостей модели.	Минимизация возмущения для изменения классификации.	FGSM, DeepFool, L-BFGS	Уязвимость к градиентным методам	Обучение, классификация	Модель, градиенты

Атака с использованием бэкдоров в нейронных сетях подразумевает подготовку моделей таким образом, что полученная в результате модель специальным образом будет реагировать на данные, в которых присутствует специальный признак, называемый триггером. Атака «извлечение исходных данных из моделей» направлена на реконструкцию данных из тренировочного набора атакуемой модели. Главным образом, атаки этого типа (таблица 1.5) направлены на генеративные модели и автокодировщики, поскольку на выходе этих моделей получаются данные, схожие с входными.

Специфика целей злоумышленника, как следует из таблицы 1.5, определяется особенностями функционирования и назначением системы искусственного интеллекта: компрометация работы вычислительных моделей путём модификации датасетов, нарушение логики работы вычислительных моделей, негативное воздействие на инфраструктурные компоненты системы искусственного интеллекта.

В целом атаки характеризуются спецификой применения, связанной с особенностями атакуемых платформ, доступностью интерфейсов, с которыми можно

взаимодействовать и влиять на логику решений (вычислительных моделей), и следующих из этого особенностей анализа атакующих последовательностей, которые требуется преобразовывать в сценарии атак. Следует учитывать, что вычислительные модели, атакуемые злоумышленниками, вероятностные, и существует неполнота и неоднозначность данных о действиях злоумышленника.

При учёте заданных условий требуется применять те модели, которые могут, с одной стороны, учитывать неполноту и непредсказуемость сведений об атаке, с другой стороны, должны быть достаточно структурированы и понятны аудитору при составлении сценариев атак. Марковские модели принятия решений позволяют учесть приведённые особенности атак и целевых объектов (элементов системы искусственного интеллекта).

1.5 Постановка задачи исследования

Анализ исследуемой проблемной области позволяет сделать вывод о необходимости внедрения процедур моделирования атакующих воздействий при аудите информационной безопасности систем искусственного интеллекта (ИИ).

От успешности моделирования во многом зависят правильность составления сценария атаки и проведения аудита. Это влияет на специфику построения защиты систем ИИ, что в современных условиях является нетривиальной задачей из-за своеобразия технологий искусственного интеллекта. Мероприятия по защите систем ИИ, которые не направлены на перекрытие наиболее опасных атак (последовательностей), в общем случае повлекут дополнительные временные и материальные расходы.

Необходимо решить научно-техническую задачу по разработке новых моделей выявления опасных атак, направленных на системы ИИ, с целью повышения качества аудита, которое проявляется в построении более точных сценариев потенциальных атак, которые в первую очередь следует перекрывать.

Сформулирована задача исследования. Она заключается в разработке:

– моделей атак на системы ИИ;

– алгоритмов построения, модификации и анализа наилучшего пути реализации атак;

– методики применения моделей, для определения опасных атакующих воздействий на системы искусственного интеллекта как подсистемы ИС;

– архитектуры системы оценки защищенности компьютерных сетей на основе моделирования атак.

Целью разработки методики является повышение степени полноты учета возможных реализаций атакующих воздействий и неполноты, непредсказуемости действий злоумышленника, выявление оптимальных для злоумышленника способов и последовательностей реализации атак на ПИИ, что позволяет построить сценарии атак при аудите ИБ ПИИ.

Полнота в рамках МППР может быть определена как способность модели учитывать все состояния, возникающие при атаке, и вероятности переходов между ними. В контексте атак на ИИ это означает, что модель должна включать все возможные состояния, стратегии атаки и реакции системы на эти атаки.

Основной задачей методики применения моделей атак является определение последовательности атакующих воздействий на системы искусственного интеллекта в процессе аудита ИБ для последующего формирования сценария атак.

Таким образом, задачи исследования предполагают:

1. Анализ существующих подходов, инструментов поиска и анализа 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 определены параметры используемых при моделировании данных.

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

4. Разработку методических рекомендаций и алгоритмов, позволяющих моделировать атакующие воздействия в процессе аудита ИБ ПИИ, изучать их динамику, использовать для составления сценарии атак.

5. Разработку архитектуры программного решения, позволяющей автоматизировать процессы моделирования и провести оценку работоспособности моделей.

Основной целью исследования является увеличение степени полноты сценариев атак при их построении и анализе в процессе аудита безопасности информационных систем с компонентами подсистем ИИ.

В качестве исходных данных рассматриваются:

- реализуемые цели и решаемые задачи;

- инфраструктура ПИИ;

- внешние общедоступные интерфейсы, общедоступные информационные ресурсы ПИИ (модели, датасеты);

– физические и логические структуры ПИИ;

– алгоритмы функционирования вычислительных моделей ПИИ, ее инфраструктурных компонентов;

– возможные каналы нарушения ИБ (уязвимости);

– режимы функционирования ИИ;

– возможности элементов по реализации этих функций;

– исходные и конечные состояния;

– внешние условия.

Исходными данными о злоумышленниках могут выступать:

– преследуемые цели и решаемые задачи;

– сведения о наличии и специфики версии датасетов, вычислительных моделей ПИИ, оборудованиия, информационного и программного обеспечения для нарушения ИБ;

– каналы НСД, которыми злоумышленники могут воспользоваться;

– типовые атакующие воздействия (с учетом методик описания атак).

Предлагаемый подход к моделированию ИБ должен учитывать специфику работы систем искусственного интеллекта и способов описания ИБ ИИ. Создаваемые таким образом модели и методика их построения должны обладать всеми требуемыми свойствами: учитывать стохастическую природу угроз, позволять анализировать последовательности действий злоумышленника, моделировать различные состояния системы и переходы между ними, находить оптимальные стратегии атаки. Кроме того, модели МППР обладают способностью учитыватьдинамику, неопределённость и адаптивность, присущие ИИ-системам, поддерживают сценарии с неполной информацией, что характерно для реальных атак. Методы МППР позволяют оптимизировать стратегии противоборствующих сторон, максимизируя ожидаемую полезность (например, ущерб для атакующего или стоимость защиты) на длительном горизонте. Это критично для аудита, где важно предсказать не только единичные инциденты, но и цепочки событий.

Выводы к главе 1

В настоящей главе проведён анализ известных структур систем искусственного интеллекта, по результатам которого были выявлены особенности и специфика работы и функционирования ИИ. Были рассмотрены особенности реализации угрозы ПИИ, на основании которых в дальнейшем могут быть выделены основные целевые элементы ИИ.

Исследование различных методик моделирования атак выявило их преимущества, однако также обнаружило, что они не полностью учитывают специфические аспекты, характерные для процессов аудита ИИ. Эти аспекты включают: построение сценариев атак с использованием методик ФСТЭК и MITRE; одновременное сочетание структурирования, вероятностного подхода и поиска оптимальных стратегий реализации атак на искусственный интеллект.

Модели МППР позволяют находить оптимальные стратегии, направленные на максимизацию вознаграждения. Они также учитывают предполагаемые спо-

собы реализации атак на системы ИИ. Вероятностный характер этих моделей позволяет учитывать порядок эксплуатации уязвимостей, неопределенности и случайные события, которые могут возникнуть в ходе атаки, поскольку атаки на системы ИИ часто имеют стохастический характер, где результаты зависят от множества случайных факторов.

Методы МППР позволяют моделировать различные состояния системы и переходы между ними, что позволяет исследовать, как атаки могут повлиять на систему ИИ в разных условиях. Кроме того, МППР обеспечивают формализм для работы с неопределенностью в принятии решений.

Глава 2. Специфика применения методов моделирования

2.1 Специфика рассмотрения атакующих воздействий на ПИИ в приложении к МППР

Основным параметром, определяющим последовательность состояний, является вероятность перехода некого компрометируемого состояния в новое в соответствии с последовательностью тактик (действий злоумышленника), классифицируемых в методическом документе ФСТЭК «Методика оценки угроз безопасности информации» и в методике «MITRE ATLAS» (далее - MITRE ATLAS). На основе параметров вероятностей переходов и начальных состояний атаки можно сформировать граф атаки. Для этого необходимо ассоциировать тактики приведенных методик и состояния, которые будут являться вершинами графа модели атаки, выстраиваемой на основе марковских процессов принятия решений. По требованиям методик предполагается их последовательное соединение. При этом для определения вероятностей переходов между состояниями модели и фиксирования входных вероятностей при исследовании аспектов безопасности больших инфраструктур требуется автоматизация фиксации событий безопасности. Среди основных шагов формирования последовательности выступают:

1. Определение типов состояний для Марковских моделей.

2. Определение соотнесённости состояний (реализации тактики) для построения графа.

3. Использование последовательности графа состояний (последовательности) для построения модели

4. Упрощение графа состояний (последовательности) при необходимости.

Для рассмотрения атакующих воздействий нужно ввести некоторые правила и ограничения [133 – 135]:

1. Логика описания атак по методологии MITRE ATLAS (а также MITRE ATT@K, Методики ФСТЭК) до начала реальных нападений подразумевает, что

если злоумышленники не достигают цели промежуточной компрометации, то есть отсутствует фиксация успешного достижения состояния, соответствующего тактике, учтенной в сценарии, то ветвь пути, на котором встретилась эта тактика, отбрасывается, потому что состояние недостижимо по предложенному варианту сопряжения уязвимости и действия, обеспечивающих реализацию этого состояния. Следуя этой логике, возвраты в графе в предыдущее состояние не обязательны (это допустимо, когда фиксируется начало развёртывания атаки, а реальный режим времени не используется при рассмотрении атаки).

2. Рассматриваются состояния как некое комплексное описание всей системы, в которых она может пребывать в определенный момент времени. При этом надо учитывать, что это комплексная система состоит из множества различных отдельных узлов сети, которые могут включать как множество уязвимостей, так и одинаковые уязвимости, то есть те, которые классифицируются одинаково. Допустим, такие уязвимости могут присутствовать в операционной системе, которые были изначально закуплены по одной лицензии и установлены в компьютерах определённой организации [134].

3. Нужно классифицировать уязвимости не только на уровне переходов от одной тактики к другой, но ещё нужно классифицировать те тактики, которые между собой могут быть сопряжены в рамках повышения мощности одного состояния. С вниманием следует относиться к вопросу изменения мощности состояния компрометации в каком-то определённом диапазоне, если при этом существует множество различных уязвимостей, которые ведут к одному и тому же состоянию в локальной или сетевой системе: суммарно они могут дать высокую вероятность промежуточных компрометаций (реализации какой-то промежуточной тактики по Методике ФСТЭК или MITRE ATLAS) [134].

Если аудитор рассматривает каждую уязвимость отдельно, маркируя её принадлежность к какой-либо тактике (состоянию), то фиксируется сопряжение уязвимостей через определённые действия без учёта их общей суммарной мощности и без общего состояния, которое описывает в целом всю систему (то есть система может побывать в каком-либо определённом состоянии). Это удобно при

возникновении потребности выявления отдельного пути, проходящего по конкретным уязвимостям. Таким образом, возникают два представления о переходах в марковских процессах при моделировании атак:

1. Перемещения злоумышленника с помощью действий по уязвимостям рассматривается в режиме имитации развертывания атаки без априорного предположения о логичности действий злоумышленника (имитация в реальном времени) (on-line). При этом переходы по уязвимостям будут многочисленны. В этом случае следует учитывать возможность злоумышленника перемещаться в те состояния, которые он уже ранее посещал (возвратные состояния) при проявлении контрмер или иных препятствий в процессе развертывания атаки (Рисунок 2.1). Таким образом усложняется определение последовательности переходов между тактиками, как состояниями компрометации, если эти переходы рассматривать с точки зрения потребности достижения конкретной цели (состояния системы), поскольку цели злоумышленника могут зависеть от его индивидуальных предпочтений. Целью злоумышленника может быть промежуточное состояние [134].

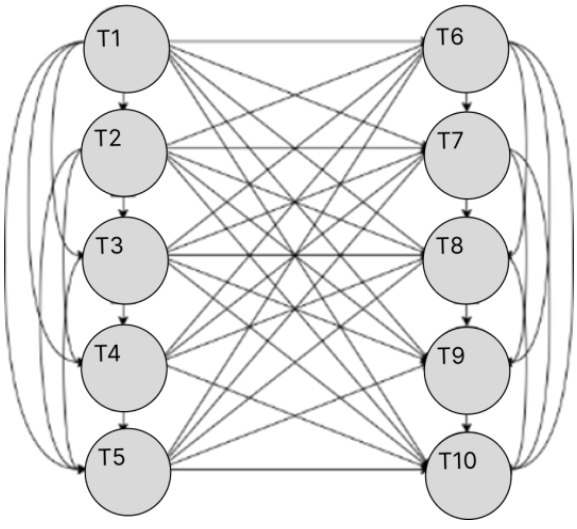


Рисунок 2.1 - Граф возможных состояний системы on-line

2. Второй способ описания переходов (режим off-line), актуален при плановом аудите. Он позволяет определить наиболее эффективный путь атаки, исходя из

того, что при построении сценария атаки не повторяются уже прошедшие этапы, так как возвратные состояния отдаляют злоумышленника от поставленной цели, то есть в этом случае не учитывается алогичность действий злоумышленника, как проявление человеческого фактора. При этом требуется конкретизировать состояние узла в распределённой системе или системы в целом для определения последовательности состояний атаки (могут учитываться при построении сценария как общесистемные состояния, так и состояния отдельных узлов сети). Следует учесть: длительность времени, затрачиваемого на переход не важна, поскольку не влияет на успешность поиска эффективного пути при появлении контрмер защищающейся стороны. Однако длительность каждого перехода может быть параметром метрики эксплуатируемой уязвимости. Кроме того, целью злоумышленника может быть промежуточное состояние (успех тактики), тогда следующие состояния отбрасываются, что упрощает порядок вычислений (Рисунок 2.2 (отброшены состояния с T2 по T5)).

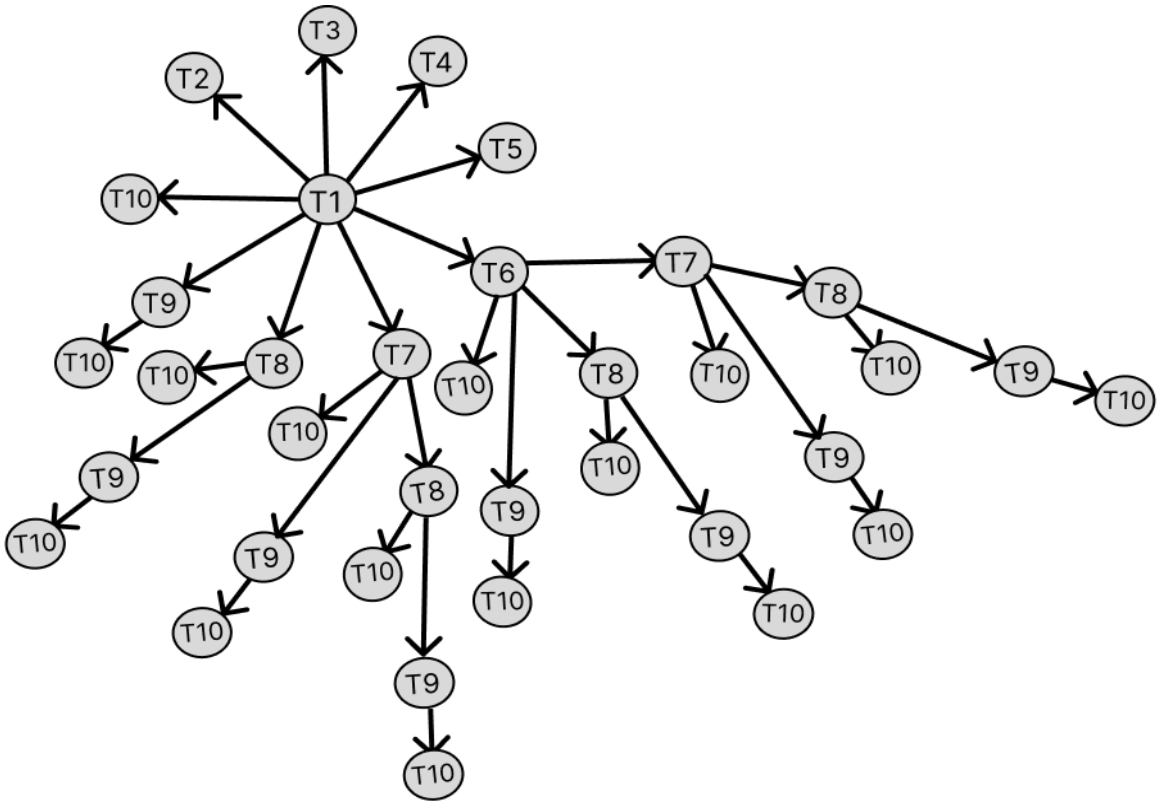


Рисунок 2.2 - Граф без возвратных связей между возможными состояниями системы

Марковские модели, использующие только уязвимости без определения классификационной принадлежности их состояний к тактикам Методики, не упорядочены и поэтому сложно определять качество состояния как повышения степени компрометации. При отсутствии поглощающих состояний задача определения наилучших атакующих последовательностей усложняется в еще большой степени, поскольку возможно зацикливание в эргодической последовательности. Определение специфических особенностей поиска наилучших последовательностей действий атакующего в режиме on-line является целью данной работы. Достижение этой цели дает возможность сформировать более полный сценарий потенциальной атаки, построенный на основе последовательности состояний, интерпретируемых как наступление возможности реализовать тактики злоумышленника. Таким образом, достижение этой цели позволяет:

- выявить наиболее опасный сценарий атак (группу сценариев атак), описываемый тактиками (рассматриваются в приведенном исследовании как этапы атак) и принадлежащими им техниками (конкретная реализация этапа посредством эксплуатации уязвимостей) по Методике ФСТЭК или MITREATLAS, что позволяет сосредоточиться стороне защиты на перекрытии уязвимостей (первоначально с помощью пакетов обновления), которые используются при реализации техник сценария;

- дает исследовательский инструмент – модель позволяет оценивать различные стратегии злоумышленника (последовательности действий с учетом наград) для максимизации ожидаемых выгод, исследовать динамику изменений состояний модели при изменении вознаграждений;

- позволяет облегчить работу эксперта–аналитика (сузить область поиска), составляющего и анализирующего сценарии атак (в плане определения наиболее опасных состояний и последовательностей) при разработке Модели угроз организации (использование экспертного подхода регламентировано Методикой ФСТЭК, используется в работе далее при необходимости);

- помогает найти оптимальную последовательность действий специалисту по активному аудиту ИБ, проводящего оценку защищенности информационных систем путем тестирования на проникновение.

Соответственно, следует:

- более подробно рассмотреть специфику методов моделирования атак (в качестве основного метода была выбран метод МППР);

- изучить специфику поиска наилучших для злоумышленника последовательностей атакующих воздействий;

- рассмотреть особенности использования алгоритмов оценки стратегий злоумышленника при нападении;

- изучить порядок получения исходных данных для моделирования;

- выявить особенности определения стратегий в задачах моделирования сетевых атак в различных режимах.

2.2 Определение параметров модели атакующих воздействий на ПИИ

При использовании марковских процессов принятия решения (МППР) сценарий атаки (векторы атак) описывается графом состояний и действий. Предполагается, что злоумышленник получает вознаграждение r, которое зависит от действия a и состояния S. Требуется найти функцию, называемую стратегией, которая определяет, какое действие предпринять в каждом состоянии, чтобы максимизировать некоторую другую функцию (например, среднюю или ожидаемую дисконтированную сумму последовательности вознаграждений) [47]. В общем случае применяемый метод определения вектора атаки предполагает следующие ключевые этапы:

1. Определение пространства состояний системы при атаке и возможных действий по отношению к ним.

2. Назначение вероятностей переходов между состояниями (постоянных во времени).

3. Определение характеристик выхода (политика действий злоумышленника).

4. Разработка математической модели (формирование матриц вероятностей переходов, вознаграждений) и расчеты методами МППР.

5. Анализ результатов.

Марковский процесс принятия решений – это кортеж $\langle S, A, P, R, \gamma \rangle$, для которого случайные состояния из последовательности $\{s1, s2, \dots sn\}$ обладают марковским свойством, где:

1. $S = \{s1, s2, \dots, sn\}$ – множество вершин-состояний системы, соответствующих тактикам базы знаний MITRE ATLAS (и Методики ФСТЭК в случае ее применимости) описания атак. Количество состояний и переходов между ними определяется на основе технического исследования целевой инфраструктуры. Отдельно добавляется состояние «Блок» (Б) - состояние блокировки означает возможность реализации действий, препятствующих реализации (как результат работы триггера, срабатывающего на действия злоумышленника).

2. $A = \{(s_i, s_j) | s_i, s_j \in S, a \in \{1, \dots, n\}\}$ – множество ребер, отражающих способы эксплуатации уязвимостей (действия), которые могут быть применены атакующим для перехода между состояниями. Необходимо учитывать, что в зависимости от следующего состояния системы (следующей тактики), количество доступных действий для следующего перехода может варьироваться из-за различий в способах эксплуатации уязвимостей. Множество действий эксплуатации уязвимостей (сопоставлены техниками, которые реализуемы в актуальном состоянии (тактике)) включает следующие основные подмножества ($A = \{(s_i, s_j) | s_i, s_j \in S, a \in \{C, D, R\}\}$):

1) нелегальное взаимодействие (D);

2) легальное взаимодействие (C);

3) сброс (R) – действие означает возможность возвращения в предыдущее состояние, которое агент (злоумышленник) уже посещал.

Также эти действия сопоставлены со множествами действий, которые определяются и, соответственно, вводятся в последовательность атакующих воздействий на основе специфики архитектуры и функционирования ПИИ:

- целевые действия. Подразумевается, что злоумышленник способен управлять взаимодействием его инструментария с компонентами ПИИ, контролировать воздействие на логику модели ИИ. В этом случае рассматриваеются следующи типы действий: доступ к датасетам; контролируемое взаимодействие с вычислительной моделью ПИИ; контролируемое взаимодействие с инфраструктурными элементами вычислительной модели ПИИ.

- возвратные действия. Выделяются три типа возвратных действий:

1) действие по типу техники, приводящее в ранее не посещённое состояние, находящееся от целевого дальше, чем то, из которого было осуществлено действие (определяется с помощью параметра в формуле награды);

2) действие отката к любому ранее достигнутому состоянию (определяется с помощью параметра в формуле награды);

3) действие по типу техники.

3. $\mathrm{R}(s, a, s')$ – это награды за переход в определенное состояние, характеризующие эффективность или сложность эксплуатации уязвимостей (действия $a \in A$) для перехода к следующему состоянию (успешной реализации тактики).

4. $P(s, a; s')$ – вероятность перехода из состояния $s \epsilon S$ при действии $a \epsilon A$ в состояние $s' \epsilon S$;

5. γ – коэффициент дисконтирования (оценивает ценность будущих наград) $\gamma \in [0; 1]$.

Подходы применимые к определению вероятностей переходов между состояниями для МППР:

1. Равномерное распределение. Если нет дополнительной информации, вероятности могут быть распределены равномерно между всеми возможными переходами.

$$P(s_j|s_j) = \frac{1}{k}, \quad (2.1)$$

где k — количество возможных переходов из состояния s_i.

Этот подход определения вероятностей чрезвычайно важен поскольку отражает специфику реальной информированности о действиях будущего злоумышленника: защищающая сторона может опереться в расчётах на статистические данные, но подобная статистика не всегда достоверна относительно конкретного предприятия.

2. Частотный анализ. Если доступны данные о переходах, вероятности можно оценить на основе частоты переходов.

$$P(s_j|s_j) = \frac{s_i s_j}{s_i}. \quad (2.2)$$

3. Экспертные оценки. Использование экспертных знаний для оценки вероятностей, если статистические данные отсутствуют.

Учитывая оценки эксплуатируемости для каждой из уязвимостей в графе атак, возможно оценить вероятности перехода путем нормализации оценок уязвимостей по всем ребрам, начиная с исходного состояния системы. Пусть p_{ij}– вероятность того, что злоумышленник, находящийся в состоянии i, воспользуется уязвимостью в состоянии j. Соответственно можно, формально определить вероятность перехода, используя формулу (2.3) [135]:

$$p_{ij} = \frac{\sum r_j}{\sum_{k \in S_i}^{n} r_k}, \quad (2.3)$$

где n – количество уязвимостей, доступных из состояния i;

r_j – оценка уязвимости состояния j;

r_k – сумма оценок эксплуатируемости всех уязвимостей всех состояний, доступных из состояния i;

S_i – множество состояний, достижимых из текущего состояния i (индекс i влияет на знаменатель формулы, ограничивая сумму только состояниями, доступными из текущего состояния). Сумма $\sum_{k \in S_i}^{n} r_k$ – учитывает только те состояния, которые достижимы из i.

Исходя из описанных выше условий, можно формировать матрицу переходных вероятностей при выполнении действия a_i в отношении набора состояний.

Следующим важным параметром модели является вознаграждение, которое отражает успех злоумышленника в случае реализации атакуемого воздействия или затраты в случае неудачи.

Проверка нормировки: для каждого состояния сумма вероятностей переходов при выполнении действия равна 1:

$$\sum_{j=1}^{n} P_{ij}\,(a) = 1\ \forall i \quad . \tag{2.4}$$

Ключевым фактором для данного параметра является базовая метрика уязвимости. Чем выше значение данной метрики, тем большее вознаграждение должен получать атакующий при переходе в соответствующее состояние в модели марковского процесса. Учитывая пропорциональность награды величине уязвимости, их параметры будут варьироваться в диапазоне от 0 до 1 (1 – означает достижение полной уязвимости, и, следовательно, максимальную награду). Исходя из описанных выше условий, формируется матрица переходных вероятностей при выполнении действия a_i в отношении набора состояний S.

Стратегией, или политикой π в данном случае является последовательность действий злоумышленника, которые связаны с эксплуатацией уязвимостей при заданных состояниях. Она описывает поведение атакующего. Если нарушитель следует стратегии π на шаге i, то $\pi(a|s)$ – вероятность того, что будет предпринято

действие $A_i = a$ при условии того, что планируется переход в состояние $S_i = s$ (1) [135]:

$$\pi(a|s) = P[A_i = a \vee S_i = s]. \quad (2.5)$$

Для нахождения лучших действий из состояния S используется следующая формула:

$$\pi^*(S) = argmaxa(R(S) + \gamma \sum S' P(S' \mid S, a) V * (S')). \quad (2.6)$$

Функция ценности определяется следующим образом вознаграждений на каждом шаге (2).

$$G_i = \mathrm{R}_{i+1} + y\mathrm{R}_{i+2} + y^2\mathrm{R}_{i+3} + \cdots = \sum_{k=0}^{\infty} y^k \mathrm{R}_{i+k+1}. \quad (2.7)$$

Ценность состояния – это ожидаемое дисконтированное вознаграждение R злоумышленника, начинающего эксплуатировать уязвимость из состояния $s_i = S$ при соблюдении стратегии атаки π. Функция ценности состояния s при стратегии π определяется как математическое ожидание дисконтированной суммы будущих вознаграждений R (описано уравнением оптимальности Беллмана.

$$V_\pi(s) = M[\mathrm{R}_{i+1} + \gamma V_\pi(S_{i+1}) | S_i = s], s \in S, \quad (2.8)$$

где $V_\pi(s)$ – функция ценности состояний (важность достижения состояния успеха эксплуатации уязвимостей атакующим в векторе атаки) при стратегии атаки π;

M – математическое ожидание случайной величины;

γ – коэффициент дисконтирования.

Графически стратегия может быть представлена как граф, узлами которого являются состояния, в которых находится система на каждом из этапов атаки, а дуги – действиями атакующего. Правдоподобие траектории последовательности действий злоумышленника, связанных с эксплуатацией уязвимостей при сохранении вектора стратегии определяется, следующим образом:

$$P(\tau_{h-1}|\pi) = \mu_0(s_0) \prod_{t=0}^{h-2} \big(\pi(a_t|s_t) p(s_t, a_t; s_{t+1})\big), \quad (2.9)$$

где $\mu_0(s_0)$ — начальное распределение состояния s_0 при развертывании последовательности атакующих воздействий; для каждого времени t от 0 до $h-2$ (где h – длина траектории – последовательности атакующих воздействий), учитывается вероятность выбора действия a_t в состоянии s_t и вероятность перехода в следующее состояние s_{t+1} после выполнения этого действия.

Решить марковский процесс принятия решений означает найти наилучшую стратегию поведения нарушителя в заданной системе. Наилучшая стратегия π^* – это такая стратегия, при соблюдении которой достигается максимальная ожидаемая совокупная награда злоумышленника и которая в дальнейшем будет определяться как параметр компрометации ПИИ. Оптимальность стратегии может пониматься в различных смыслах в зависимости от выбранного критерия. Политика выгоды (в марковских процессах принятия решений) связана с оптимизацией стратегий принятия решений в последовательности атакующих воздействий при условии неопределенности среды. Наилучший вектор атаки обеспечивает максимальную ожидаемую ценность компрометирующего состояния в векторе атаки (2.10):

$$\pi^*(s) = \underset{\pi}{argmax} V_\pi(s). \quad (2.10)$$

Функция ценности, обеспечивающая максимальную ожидаемую величину компрометации относительно всех стратегий злоумышленника, называется

оптимальной функцией ценности, представлена как важнейшая величина компрометации (эксплуатации уязвимости) и обозначается как $V^*(s)$ (2.11) [134]:

$$V^*(s) = \max_{\pi} V_\pi(s). \quad (2.11)$$

Для сравнения эффективности атакующих политик и определения наилучшего плана действий вводится функция ценности состояния $V_\pi(s)$. На множестве политик вводится отношение частичного порядка. Политика π (2.12). считается предпочтительнее или равной политике π^*, , если и только если ценность состояния $V_\pi(s)$ для политики π больше или равна ценности состояния $V_{\pi`}(s)$ для политики π^* во всех состояниях s из множества S.

$$\pi \geq \pi^* \Leftrightarrow V_\pi(s) \geq V'_\pi(s), \forall s \in \text{S}. \quad (2.\ 12)$$

Иным способом определения стратегии является использование функции ценности действия a_i, представляемого как эксплуатация уязвимости [134]. Ценность применения действия в состоянии s при стратегии π формируется как ожидаемый доход, когда нарушитель начинает реализацию атаки в состоянии s, выполняет действие a и затем следует стратегии π (2.13) [5]:

$$Q_\pi(s,a) = M[G_i | S_i = s, A_i = a], \quad (2.13)$$

где $Q_\pi(s,a)$ – функция ценности действия (как эксплуатации уязвимости) при атакующей последовательности (стратегии) π.

Эта функция ценности действия позволяет представить функцию ценности состояния (2.14), а также стратегию (2.15) соответственно [5]:

$$V(s) = \max_{a} Q(s,a), \quad (2.14)$$

$$\pi(s) = arg\max_{a} Q(s,a). \quad (2.15)$$

Для вычислений, связанных со стратегией принятия решений, и для поиска оптимальной стратегии, применяются различные алгоритмы, в том числе относящиеся к методам динамического программирования, например, Q-обучение. Используемые методы рассматриваются далее. При этом каждая стратегия по-разному проявляет свою эффективность. Ее выбор определяется особенностями описания проблемной области [134]:

1. Динамичность или статичность состояний.
2. Полная или неполная известность модели (неточные знания о количестве состояний, то есть о количестве эксплуатируемых уязвимостей). При поиске стратегий в реальном режиме времени злоумышленник может обнаружить ранее неучтенные уязвимости.
3. Наличие обратных связей при использовании времени достижения состояния.

При построении марковских моделей используется динамическое программирование как основной метод решения задач оптимизации. Это семейство алгоритмов, которые применяются для поиска оптимальных стратегий в условиях модели окружающей среды, представленной в виде марковского процесса принятия решений. Чаще всего применяются следующие алгоритмы:

1. Policy iteration – итерация по стратегиям.
2. Value iteration – итерация по значениям.

Сходимость при итерации по стратегиям достигается за конечное число итераций, так как существует ограниченное количество возможных стратегий, и каждая итерация улучшает стратегию, если это возможно [134]. Несмотря на то, что число допустимых стратегий растет экспоненциально с увеличением количества состояний, итерация по стратегиям на практике сходится достаточно быстро [7]. Однако, это означает, что исследуемая ПИИ должна быть полностью описана, а все возможные пути атакующих последовательностей учтены [132-133]. В результате работы этого алгоритма создается цепочка политик, то есть действий атакующего, где каждая из которых является улучшением по сравнению с предыдущей (2.16).

$$\pi_0 \underset{\rightarrow}{E} v_{\pi_0} \underset{\rightarrow}{I} \pi_1 \underset{\rightarrow}{E} v_{\pi_1} \underset{\rightarrow}{I} \pi_2 \underset{\rightarrow}{E} \dots \underset{\rightarrow}{I} \pi_n \underset{\rightarrow}{E} v_n \quad . \tag{2.16}$$

Оценка политики (E) и шаги по ее совершенствованию (I) проводятся до тех пор, пока политика больше не перестанет улучшаться.

Итерации по стратегиям могут быть более эффективными в ситуациях, когда необходимо быстро находить оптимальные стратегии действий. Этот подход позволяет оценивать политику и улучшать её до тех пор, пока не будет достигнута сходимость. В контексте MITRE ATLAS, где состояния представляют собой тактики, а действия — техники, итерации по стратегиям могут помочь в быстром нахождении эффективных комбинаций атакующих техник, что особенно важно в динамичной среде киберугроз. Модифицированная итерация по стратегиям предлагает альтернативный подход, который аппроксимирует функцию полезности, используя итеративную оценку стратегии вместо точной оценки. Это позволяет сократить вычислительные затраты, но может привести к менее точным результатам.

Итерации по значениям фокусируются на вычислении оптимальной функции полезности для каждого состояния. Этот подход может быть более вычислительно затратным, так как требует оценки значений для всех состояний, что может замедлить процесс, особенно если количество тактик и техник велико. Однако, он может быть полезен, если требуется глубокий анализ и понимание всех возможных исходов для каждой тактики.

Основное отличие алгоритма итерации по значениям от итерации по стратегиям заключается в том, что итерация по значениям фокусируется на вычислении оптимальной функции полезности для каждого состояния, в то время как итерация по стратегиям направлена на поиск оптимальной стратегии действий [8, 13, 15].

Поскольку приведенные методы определения стратегий требуют предварительного описания исследуемой среды, их целесообразно применять в

режиме off-line. В случаях исследования атакующих последовательностей без предварительного знания модели среды допустимо использовать Q-обучение - алгоритм обучения с подкреплением, который позволяет агенту находить оптимальную стратегию поведения в не полностью описанной среде. Таким образом, Q-обучение целесообразно использовать как дополнительный метод МППР в тех случаях, когда используется неточное описание ПИИ, что характерно для распределенных ИС, в которые встраиваются элементы ИИ.

В Q-обучении оптимальная политика может быть определена двумя терминами: функцией ценности и функцией Q-value. Функция ценности показывает, насколько приоритетно фиксируемое состояние уязвимости. Подобная функция ценности определяется как ожидаемое совокупное вознаграждение от соблюдения политики со стороны состояний, то есть переходов по уязвимостям [5] (2.17).

$$V^{\pi}(s) = E[\sum_{t>0} \gamma^{t} r_{t} | \pi]. \quad (2.17)$$

Функция Q-value определяется как ожидаемое совокупное вознаграждение, получаемое обеими частями пары состояние-действие, то есть уязвимость – эксплуатация уязвимости.

$$Q^{\pi}(s,a) = \mathbb{E}\,[\sum_{t>0} \gamma^{t} r_{t} | s_{0} = s, a_{0} = a, \pi]. \quad (2.18)$$

Применимость алгоритма вычисления зависит от требуемой глубины (подробности описания инфраструктуры, состояний, типов действий) описания атаки. Это необходимо учитывать при выборе модели построения атакующей последовательности в дальнейшем (в контексте работы допустимо использовать три типа модели с учетом указанных ограничений). Основой реализации алгоритмов является использование функции полезности.

Функция полезности каждого действия может определяться отдельно при использовании Value iteration. Три действия используются в контексте следующих функций (2.19 – 2.21):

1. Нелегальное взаимодействие (*D*).

$$V_{i+1}^{*}(S, a = D) = \sum_{s' \in S} P(s, D, s') \ [\mathrm{R}\ (s, D, s') + \gamma V_i^{*}(s')]. \quad (2.19)$$

2. Легальное взаимодействие (*C*).

$$V_{i+1}^{*}(S, a = C) = \sum_{s' \in S} P(s, C, s') \ [\mathrm{R}\ (s, C, s') + \gamma V_i^{*}(s')]. \quad (2.20)$$

3. Сброс (*R*) (в этом случае состояние$s' \in Sprev$, где $Sprev$ -это ранее полученные состояния).

$$V_{i+1}^{*}(S, a = R) = \sum_{s' \in Sprev} P(s, R, s') \ [\mathrm{R}\ (s, R, s') + \gamma V_i^{*}(s')]. \quad (2.21)$$

где:

– $P(s'|s,\ a)$ - вероятность перехода в состояние s' из состояния s при действии a_i.

– $\mathrm{R}(s,\ a,\ s')$ - награда за переход из состояния s в состояние s' при действии a.

– $V_i^{*}(s')$ - ожидаемая полезность состояния s' на итерации i.

– γ - коэффициент дисконтирования, который определяет важность будущих наград.

Таким образом, приводимые формулы соответствуют уравнению Беллмана, а оптимальная политика π^{*} соответствует принятию наилучших последовательностей действий злоумышленника в любом состоянии.

2.3 Определение особенностей параметров модели

В моделях МППР нарушитель взаимодействует с информационной системой, предпринимая в состоянии S_i действие A_i, за успешное выполнение которого он получает положительное вознаграждение R_i либо отрицательное в обратном случае, то есть злоумышленник получает вознаграждение, которое зависит от действия и состояния. Цель моделирования состоит в том, чтобы найти функцию, называемую стратегией, или политикой, которая определяет, какое действие предпринять в каждом состоянии, чтобы максимизировать некоторую другую функцию (например, среднюю или ожидаемую дисконтированную сумму) последовательности вознаграждений [134].

Формально последовательность атакующих действий можно представить в виде графа состояний, генерирующегося с помощью разрабатываемого инструмента и принимающего в качестве основных входных данных следующую информацию о компонентах анализируемой системы:

- IP-адрес хоста, топология сети;
- идентификатор (ID) уязвимости, при этом хосты могут иметь более одной уязвимости (используется для сопоставления действия и сетевого узла);
- тип действий;
- CVSS-оценка эксплуатируемости уязвимости (по стандарту 2.0 (допустимо использование 3.0 и др.)) и другие парамеры функции R(s,a,s`). Это число будет использовано как один из коэффициентов в расчете вероятностей и наград математической модели для перехода между состояниями.

Другим требованием для формирования модели является выбор и категорирование состояний системы, чтобы связать с каждой вершиной графа набор уязвимостей, делающих переход к следующей вершине возможным. В качестве набора возможных состояний были приняты тактики из MITRE ATLAS, и Методики оценки угроз безопасности информации (Методики ФСТЭК) [134].

В процессе совершения атаки у злоумышленника имеется возможность выполнять некоторые действия (способы эксплуатации) по отношению к уязвимостям компонентов системы, посредством которых он меняет ее состояние. При этом вероятности перехода для каждого действия принимают разные значения. В состав такого набора входят следующие способы эксплуатации уязвимостей: несанкционированный сбор информации, исчерпание ресурсов, инъекция, анализ целевого объекта, подмена при взаимодействии, злоупотребление функционалом, нарушение авторизации, нарушение аутентификации, манипулирование структурами данных, манипулирование ресурсами, манипулирование сроками и состоянием.

Сопоставляя способы эксплуатации уязвимостей с выявленными уязвимостями, можно сузить область поиска и затем определить экспертным способом тип техник, соответствующий специфике эксплуатации слабостей системы.

Целевой вершиной графа (поглощающее состояние) является последнее состояние в методике описания (последняя тактика в списке тактик (T15, на момен проведения исследования) для MITRE ATLAS и тактики T10 для Методики ФСТЭК) или то состояние, которое определит, как целевое, эксперт, формирующий описание атаки. При нахождении одного из компонентов системы в данном состоянии нарушается состояние безопасности всей системы в целом. Как только злоумышленник достигает этого состояния, система считается скомпрометированной, то есть злоумышленник достиг своей цели. Таким образом, система будет оставаться в этом состоянии до тех пор, пока специалисты по безопасности не примут превентивные меры по удалению присутствия злоумышленника в системе [134].

Проблемой является отсутствие прямого сопоставления тактик с уязвимостями. В соответствии с утвержденной формой описания уязвимостей в БДУ ФСТЭК, каждая уязвимость может быть классифицирована по одному из двенадцати способов эксплуатации, которые были выведены путем синтеза шаблонов атак из международного стандарта CAPEC. Способы эксплуатации уязвимостей, в свою очередь, могут быть соотнесены с техниками, которые

применяются для реализации тактик. В результате анализа перечня техник получена таблица 2.1, отражающая соотношения тактик и способов эксплуатации, что позволит в дальнейшем отнести уязвимости к состояниям системы. В процессе совершения атаки у злоумышленника имеется возможность выполнять некоторые действия (способы эксплуатации) по отношению к уязвимостям компонентов системы, посредством которых он меняет ее состояние. При этом вероятности перехода для каждого действия принимают разные значения при наличии статистической информации от датчиков (поставщиков данных о событиях в системе), и равномерное распределение вероятностей с учетом соблюдения нормировки, исходя из количества действий от состояния к состоянию [134].

Таблица 2.1 – Демонстрация принципа соотнесения тактик и способов эксплуатации уязвимостей (по Методике ФСТЭК)

Тактика	Способы эксплуатации уязвимости
1. Сбор информации о системах и сетях (Т1)	Несанкционированный сбор информации Анализ целевого объекта
2. Получение первоначального доступа к компонентам систем и сетей (Т3)	Нарушение аутентификации Нарушение авторизации Инъекция
3. Внедрение и исполнение вредоносного программного обеспечения в системах и сетях(Т3)	Инъекция Манипулирование структурами данных
4. Закрепление (сохранение доступа) в системе или сети (Т4)	Манипулирование структурами данных Манипулирование ресурсами
5. Управление вредоносным программным обеспечением и (или) компонентами, к которым ранее был получен доступ (Т5)	Злоупотребление функционалом
6. Повышение привилегий по доступу к компонентам систем и сетей(Т6)	Нарушение аутентификации Нарушение авторизации
7. Сокрытие действий и применяемых при этом средств от обнаружения (Т7)	Манипулирование сроками и состоянием
8. Получение доступа (распространение доступа) к другим компонентам систем и сетей или смежным системам и сетям (Т8)	Подмена при взаимодействии
9. Сбор и вывод из системы или сети информации, необходимой для дальнейших действий при реализации угроз безопасности информации или реализации новых угроз (Т9)	Несанкционированный сбор информации
10. Несанкционированный доступ и (или) воздействие на информационные ресурсы или компоненты систем и сетей, приводящие к негативным последствиям (Т10)	Манипулирование структурами данных Манипулирование ресурсами Исчерпание ресурсов

Возможность соотнесения состояний системы, выраженных в виде фиксации успеха реализации действий, тактик MITRE (наступление этапов атаки), и способов эксплуатации уязвимостей, которые позволяют достичь состояния успеха, с состояниями и переходами между ними указывает на целесообразность их использования при моделировании (для общего случая атаки) (таблица 2. 1). Тактики, в соответствии с приведенными на ресурсах ФСТЭК способами эксплуатации уязвимостей, могут быть сопоставлены с метриками уязвимостей. Способы эксплуатации позволяют связать метрики оценки уязвимостей разных метрических систем (MITRE ATLAS, MITRE ATT@CK, Методики ФСТЭК). Следует учитывать возможность обновления метрик и методик, соответственно, соотношение тактик и способов эксплуатации уязвимостей следует постоянно актуализировать.

Данные, получаемые из открытого банка данных угроз ФСТЭК РФ, относятся к уязвимостям и метрикам CVSS. Метрики CVSS представляют собой открытый стандарт, используемый для количественной оценки уязвимости в безопасности компьютерной системы. Вторая группа определяется на основе данных, собранных внутри локальной сети: сетевой трафик, конфигурация хостов, топология сети, актуальные уязвимости.

Также допустимо использовать альтернативный способ сопряжения тактик (техник) и уязвимостей – маппинг DETT&CT. Это может быть реализовано через заранее определенные правила или алгоритмы, которые сопоставляют уязвимости с известными методами атаки. Когда DETT&CT обнаруживает уязвимость в системе, он может использовать маппинг, чтобы определить, какие техники MITRE могут быть использованы злоумышленниками для ее эксплуатации. Это позволяет командам безопасности лучше понять риски и разработать соответствующие меры защиты. В зависимости от заданных условий, сценарий реализации атаки может включать не все тактики, то есть их количество может варьироваться. Злоумышленник переходит из одного состояния в другое, когда эксплуатируется хотя бы одна из уязвимостей, относящихся к следующему состоянию исследуемой системы.

В отношении уязвимостей может применяться не весь набор способов эксплуатации, что отражено в таблице 2.1. Возможность соотнесения состояний системы, выраженных в виде фиксации успеха реализации тактик ФСТЭК, и способов эксплуатации уязвимостей, которые позволяют достичь состояние успеха, говорит о целесообразности их использования в качестве действий. Переходы между состояниями вероятностны и зависят от особенностей эксплуатации (сложности) уязвимостей.

Параметры определения уязвимости можно использовать для исследования и анализа вектора атаки при условии того, что уязвимость влияет на вероятность реализации атакующих воздействий. Для демонстрации примера возьмем простую, хотя и устаревшую, стандарт 2.0. Параметр $r(v)$, рассматриваемый как комплексный показатель доступности эксплуатации, можно рассмотреть и как меру сложности эксплуатации уязвимости (*Exploitability*). Стандарт CVSS 2.0 использует формулу (2.22) для расчета данного параметра [3].

$$r(v) = 20 * AV * AC * Au, \tag{2.22}$$

где:

Access Vector (AV) – параметр, показывающий, каким путем может быть внедрена уязвимость; *Access Complexity (AC)* – сложность атаки, оценивающая, насколько легко или сложно использовать данную уязвимость; *Authentication (Au)* – количество действий (например, аутентификаций), которые атакующий производит, прежде чем воспользоваться уязвимостью [104]. Параметры определения уязвимости *Access Complexity (AC), Access Vector (AV)* позволяет отразить специфику взаимосвязей и влияния злоумышленника на ПИИ при моделировании. Используя эти параметры или их смысловые аналоги в более новых версиях CVSS (версии 3.1, 4.0), можно рассматривать динамику изменения состояния уязвимостей атакуемой системы: понижая или повышая величину параметров стоимости атакующих воздействий, которые зависят от величины

параметра уязвимости «Сложность получения доступа (AC)» можно отслеживать вероятности изменения.

При рассмотрении метрик уязвимостей, относимых к моделям ИИ, следует учитывать следующие особенности:

1. Оценка уязвимости. AUC и другие метрики, такие как точность (Accuracy) и полнота (Recall), могут служить индикаторами уязвимости модели. Низкие значения AUC могут указывать на то, что модель неэффективно различает классы конечных значений, что может быть использовано злоумышленниками для манипуляции результатами.

2. Анализ отклонений. Использование диапазонов отклонений, основанных на AUC, может помочь в выявлении потенциальных уязвимостей. Например, если модель имеет высокую точность, но низкий AUC, это может свидетельствовать о том, что модель может быть уязвима к атаке на уклонение, где злоумышленник пытается ввести данные, которые модель неправильно классифицирует.

3. Адаптивные стратегии защиты. Зная, как злоумышленники могут использовать метрики для своих атак, разработчики могут адаптировать свои стратегии защиты, улучшая модели и повышая их устойчивость к атакам.

Формирование вознаграждения, которое отражает успех злоумышленника в случае реализации атакуемого воздействия или затраты в случае неудачи, играет важную роль с точки зрения точности результатов, эффективности модели и сходимости алгоритма поиска наиболее выгодной политики. Ключевым фактором для данного параметра является базовая метрика уязвимости. Чем выше значение данной метрики, тем большее вознаграждение должен получать атакующий при переходе в соответствующее состояние в модели Марковского процесса [134].

Коэффициент дисконтирования указывает горизонт решаемой задачи для нарушителя. Он определяется в диапазоне от 0 до 1, и показывает, насколько весомыми станут будущие вознаграждения в сравнении с полученными на текущем шаге. Для определния коэффициента дисконтирования можно использовать следующие подходы:

1. Учет временных предпочтений нарушителя. Коэффициент дисконтирования может показать, насколько важны для нарушителя будущие выгоды по сравнению с текущими. Если нарушитель нацелен на быстрый результат, коэффициент будет ближе к 0; если же он ориентирован на долгосрочную перспективу, коэффициент будет ближе к 1.

2. Оценка рисков срыва атаки на каждом этапе. Чем выше вероятность прерывания атаки на промежуточных этапах, тем ниже должен быть коэффициент дисконтирования. Это отражает неопределенность в получении будущих выгод. [134].

Одним из наиболее важных аспектов процесса моделирования является доступность и качество входных данных. Поэтому перед началом работы необходимо провести аудит анализируемой системы. Кроме того, при увеличении масштабов исследуемой сети трудоемкость вычислений возрастает, что может привести к значительному уменьшению производительности и увеличению времени выполнения симуляций.

В дополнение к вышесказанному, следует отметить, что рассматриваемая модель не учитывает некоторые значимые аспекты, такие как динамические изменения вероятностей переходов между состояниями во времени. Следует подчеркнуть, что аналитические выводы могут быть получены исключительно в условиях стационарности переходных вероятностей.

При анализе моделей искусственного интеллекта с целью выявления их уязвимостей необходимо рассмотреть две ключевые составляющие, влияющие на функциональность систем на логическом уровне. Это вычислительные модели и датасеты, которые являются основополагающими элементами ПИИ. Атаки, которые характерны для них, приведены в таблицах 2.2, 2.3 [20-26, 42, 43]. В таблице 2.3. можно увидеть основные уязвимости, связанные с отравлением данных, используемых при обучении.

Наличие этих уязвимостей подчеркивает необходимость строгих мер безопасности и контроля в процессе работы с данными, так как их несоблюдение ведет к нарушению функциональности систем ИИ.

Таблица 2.2 – Атаки на ИИ с учетом моделей и алгоритмов

Атаки	Алгоритмы/модели	Вероятность успешного воздействия	Источники
Состязательные примеры (целенаправленные возмущения входных данных, которые приводят к неправильной классификации выводимого)	Нейронные сети, глубокие нейронные сети, SVM, k-NN, логистическая регрессия, линейная регрессия, решающие деревья, ансамбли моделей	Высокая	SecurityNet, NIST, OWASP
Data Poisoning (внесение вредоносных изменений в обучающие данные, чтобы ухудшить производительность модели)	Все алгоритмы	Средняя	SecurityNet, NIST
Кража модели (восстановление модели машинного обучения)	Любые модели через API	Средняя	SecurityNet, NIST
Инверсия модели (восстановление данных обучающих примеров на основе выводов модели)	Нейронные сети, глубокие нейронные сети	Низкая	NIST, IEEE
Определение принадлежности (поиск конкретного примера, использованного для обучения модели.)	Нейронные сети, глубокие нейронные сети	Средняя	SecurityNet, NIST
Смена меток (модификация меток обучающих данных, чтобы ухудшить производительность модели)	Все алгоритмы	Средняя	SecurityNet, NIST

При моделировании атак на системы и модели ИИ необходимо учитывать принятые методологии построения вектора атакующих воздействий. Методология MITRE ATLAS уже включает в свой состав действия, обозначающие не только перемещение между узлами, но и между состояниями, которые сопоставлены этапам компрометации вычислительной модели ИИ или системы ИИ в целом [19].

Таким образом, действия злоумышленника при атаке соответствуют тактикам методологии MITRE ATLAS (далее MITRE), а их осуществление соответствует наступлению состояний, фиксирующих успех действия на некотором этапе атаки [135].

Таблица 2.3 - Используемые при отравлении данных уязвимости

Уязвимость	Описание	Действия злоумышленника	Вероятность
Незащищенные базы данных	Недостаточная защита базы данных, отсутствие шифрования, слабые пароли	Получение доступа к базе данных через уязвимость и изменение данных или меток	Высокая (30-40%)
Недостатки процедур аутентификации и авторизации	Недостаточные меры контроля доступа, недостаточная сегментация прав доступа	Получение доступа к системе через скомпрометированные учетные записи или недостаточно защищенные интерфейсы	Высокая (25-35%)
Незащищенные файловые системы	Недостаточная защита файловой системы, отсутствие контроля целостности файлов	Изменение или замена файлов данных после получения доступа к файловой системе	Средняя (20-30%)
Отсутствие проверки данных из внешних источников	Отсутствие валидации и проверки поступающих данных	Внедрение и отправка вредоносных данных через скомпрометированные внешние источники	Средняя (15-25%)
Отсутствие мониторинга и аудита	Недостаточный мониторинг и аудит данных и процессов	Внесение изменений в данные без обнаружения	Низкая (10-20%)
Социальная инженерия	Использование методов социальной инженерии для обмана сотрудников и получения доступа	Обман сотрудников для внесения изменений в данные или получения доступа к системе	Средняя (20-30%)
Недостаточная защита сетевого периметра	Недостаточная защита сетевого периметра, отсутствие сегментации сети	Получение удаленного доступа к системе для внесения изменений в данные	Средняя (15-25%)

Однако моделирование атак на алгоритмы ИИ (математический аппарат), заслуживает отдельного внимания, поскольку целевым объектом злоумышленника становится специфика вычислений, некоторые допущения в области точности, трактуемые как уязвимости. Соответственно, необходимо изучить специфику моделирования атак на модели ИИ.

Определение параметров наград для функции ценности модели МППР атак на ПИИ специфичны. При рассмотрении уязвимостей вычислительных моделей следует учитывать:

1. Известность модели, данных злоумышленнику (до начала атаки, в процессе атаки).

2. Доступность модели, данных злоумышленнику (до начала атаки, в процессе атаки).

Параметры точности моделей можно преобразовать в метрику CVSS или параметр метрики.

Достоверность (accuracy) – показывает долю правильно классифицированных событий и определяется следующим образом:

$$Accuracy(a) = \frac{TP+TN}{TP+TN+FP+FN}, \quad (2.23)$$

где TP, TN, FP, FN – значения матрицы ошибок бинарной классификации.

Когда происходит прогнозирование события, качество классификатора зависит от того, насколько большое значение TPR (true positive rate) получается, при как можно меньшем значении FPR (false positive rate). Данные метрики формируются следующими формулами:

$$TPR = \frac{TP}{TP+FN}, \quad (2.24)$$

где:

TP – верно положительные срабатывания,

FN – ложно отрицательные срабатывания.

$$FPR = \frac{FP}{TN+FP}, \quad (2.25)$$

где:

FP – ложно положительные срабатывания,

TN – верно отрицательные срабатывания.

Исходя из метрик (5) и (6) получим формулу для AUC-метрики, определяющую площадь под графиком ROC-кривой:

$$AUC = \int_0^1 TPRdFPR. \quad (2.26)$$

Многие атаки на ПИИ предполагают использование недостатков вычислительных моделей или датасетов (могут быть доступны злоумышленникам). При использовании методик описания, таких как MITRE ATLAS, целесообразно использовать смежные методики описания уязвимостей. В случае с ПИИ это осложняется тем, что недостатки вычислительных моделей не имеют четкую маркировку в базе CVSS.

Для перевода оценок точности нейросетей, таких как AUC (Area Under Curve), точность (Accuracy) и полнота (Recall), в метрики CVSS (учитывается возможность использования CVSS разных версий) необходимо учитывать, что данная метрическая система предназначен для оценки уязвимостей и их влияния на безопасность, тогда как AUC и другие метрики относятся к производительности моделей классификации. Тем не менее, можно рассмотреть подходы и методы, которые помогут сделать такую трансформацию.

В условиях постоянного развития и изменения CVSS (с версии 2.0 до 4.0 включительно) целесообразно сосредоточиться на основных трех типах метрик, которые по своему смысловому наполнению прослеживаются (в разной степени) во всех применяемых сегодня версиях рассматриваемой метрической системы:

1. Базовые метрики (Base Metrics). Это основные характеристики уязвимости.

2. Временные метрики (Temporal Metrics). Учитывают временные аспекты уязвимости (в версии 4.0 данные метрики отсутствуют как отдельный класс, но представлены в виде аналогов в Supplemental Metrics, Environmental Metrics, Threat Intelligence).

3. Контекстные метрики (Environmental Metrics). Отражают специфические для организации факторы.

Точность (Accuracy может указывать на низкий уровень ложных срабатываний, что может быть переведено в более низкий балл CVSS. Например, если точность выше 90 %, это может соответствовать низкому уровню риска.

Полнота (Recall) показывает, насколько хорошо модель находит все уязвимости. Высокая полнота может быть связана с высоким уровнем риска, если модель пропускает значительное количество уязвимостей. Например, если полнота ниже 70 %, это может указывать на высокий риск и, соответственно, более высокий балл CVSS. AUC (Area Under the Curve) является показателем общей производительности модели. Высокое значение AUC (ближе к 1) может указывать на то, что модель хорошо различает классы объектов. Это может быть переведено в более низкий балл CVSS, если модель показывает высокую эффективность. Однако, потребуется использовать методы калибровки, чтобы адаптировать значения AUC, точности и полноты к шкале CVSS: установить пороговые значения для каждой метрики, которые будут соответствовать определённым уровням уязвимости в CVSS.

Каждая из этих метрик применима при исследовании модели ИИ на предмет подбора различных атак: каждая атака ориентирована на эксплуатацию определённой характеристики модели:

1. Целевые атаки ориентированы на метрику полноты, поскольку их задачи связаны с конкретным набором распознаваемых и анализируемых атакуемой системой данных.

2. Нецелевые атаки могут использовать точность, поскольку в этом случае важен диапазон ошибок.

Данный подход характерен для тех случаев, когда необходимо оценить уязвимости ПИИ при аудите ИБ ИИ именно как логических моделей без учета технических особенностей компонентной базы ИС. Для перевода метрик точности нейросетей, таких как AUC, точность (Accuracy) и полнота (Recall), в метрики (метрические параметры) CVSS с учетом базовых и временных параметров можно использовать следующий подход: в системе CVSS используются различные метрики для оценки уязвимостей, и они не имеют прямого соответствия с метриками, используемыми для оценки моделей машинного обучения, такими как AUC, точность (Accuracy) и полнота (Recall). Однако можно провести аналогии на концептуальном уровне - перевод метрик точности нейросетей в метрики CVSS

позволяет оценить уязвимости систем ИИ с учетом их производительности. Такой подход может помочь в более глубоком понимании рисков и разработке эффективных стратегий защиты.

Подходы применимые для преобразования метрик точности ПИИ в метрики CVSS:

1. Анализ контекста. Прежде всего, важно понять, как метрики производительности моделей могут быть связаны с уязвимостями.

2. Калибровка метрик. Можно рассмотреть возможность калибровки метрик AUC, точности и полноты для оценки их влияния на безопасность. Например, можно установить пороговые значения для этих метрик, которые будут соответствовать определенным уровням риска в CVSS.

3. Сравнительный анализ. Проведение сравнительного анализа между метриками производительности и метриками CVSS может помочь выявить корреляции.

4. Моделирование. Использование статистических и машинных методов для моделирования взаимосвязей между метриками.

5. Метод нормализации. Приведение значений метрик к единой шкале (например, от 0 до 1) и использование их для расчета балла CVSS. Например, можно нормализовать каждую метрику и затем использовать их в формуле CVSS.

Перевод метрик нейросетей, таких как точность, полнота и AUC, в контекст CVSS, в CVSS-метрики с учетом специфики инфраструктуры ИС возможно реализовать так, как приведено в примере (по методу калибровки) в таблице 2.4.

Таблица 2.4 - Значения метрик ИИ в сопряжении с CVSS

Значения метрик ИИ	Точность (Accuracy)	Полнота (Recall)	AUC
Изначальные значения	Точность = 0.95	Полнота = 0.65	AUC = 0.88
Значение в соответствии со шкалой	90 % → Низкий риск (балл CVSS 0-3); 70-90 % → Средний риск (балл CVSS 4-6); < 70 % → Высокий риск (балл CVSS 7-10)	0.9 → Низкий риск (балл CVSS 0-3); 0.7-0.9 → Средний риск (балл CVSS 4-6); Меньше 0.7 → Высокий риск (балл CVSS 7-10)	80 % → Низкий риск (балл CVSS 0-3); 50-80 % → Средний риск (балл CVSS 4-6); < 50 % → Высокий риск (балл CVSS 7-10)

При этом мера риска определяется в зависимости от конкретных свойств вычислительной модели, поскольку каждая обученная модель, в этом отношении, имеет уникальные характеристики, которые связаны, в том числе, с требованиями точности разработчиков ИИ (преимущество в адаптивности и простоте метода).

На основе этих оценок можно вычислить итоговый балл CVSS, используя взвешенную сумму или другие методы, чтобы учесть относительное значение каждой метрики. Метод взвешенной суммы позволяет интегрировать несколько метрик, таких как точность, полнота и AUC, в одну общую оценку, которая затем может быть переведена в баллы CVSS. Этот метод особенно полезен, когда необходимо учесть влияние каждой метрики на общий уровень уязвимости WS (рассматривает как аналог оценки CVSS при рассмотрении атак без учета инфраструктурных компонентов ПИИ и при исследовании состояния на предмет выявления динамических характеристик атак).

$$WS = w_1 \cdot (1 - Accuracy) + w_2 \cdot (1 - Recall) + w_3 \cdot (1 - AUC) + b, \quad (2.27)$$

где:

1- Accuracy – учитывает, что более низкая точность указывает на большую уязвимость;

1- Recall – учитывает, что более низкая полнота указывает на большую уязвимость;

1- AUC – учитывает, что более низкое значение AUC указывает на большую уязвимость;

b – свободный член, который можно установить на основе полученных ранее данных или экспериментов;

w_1 – вес для точности (Accuracy);

w_2 – вес для полноты (Recall);

w_3 – вес для AUC (Area Under the Curve).

При определении весов важно убедиться, что они нормализованы, чтобы сумма всех весов была равна 1. Это поможет избежать искажений в итоговом значении CVSS.

$$w_1+w_2+w_3+...+w_n=w_1+w_2+w_3+...+w_n=1 \quad . \tag{2.28}$$

Определение весов в формуле – это процесс, который требует анализа данных, экспертной оценки и, возможно, итеративного подхода. Один из наиболее распространенных способов — это использование экспертного подхода (способ в Методике ФСТЭК). Другим распространенным методом является использование корреляционного анализа. Можно рассчитать коэффициент корреляции между каждой метрикой (например, точностью, полнотой и AUC) и значениями CVSS. Это позволяет понять, какие метрики имеют наибольшее влияние на уровень уязвимости.

$$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\,(Y-\bar{Y})^2}}, \tag{2.29}$$

где:

X – значения метрик точности (например, точность, полнота, AUC);

Y – значения CVSS;

$\overline{X}$ и $\overline{Y}$ – средние значения переменных X и Y.

Чем выше абсолютное значение r, тем больше влияние метрики на CVSS. Увеличение метрики связано с увеличением значения CVSS (или снижением уязвимости).

На основе значений коэффициента корреляции можно определить веса для каждой метрики. Например, метрики с высоким положительным r могут получить больший вес, так как они более значимы для снижения уязвимости; метрики с высоким отрицательным r также могут получить вес, но в обратном направлении, так как их увеличение связано с повышением уязвимости. Затем требуется нормализовать их, чтобы сумма всех весов равнялась 1.

Далее проводится вычисление взвешенной суммы и сопоставление её с оценкой уязвимости CVSS. Допустим результат вычислений $WS=3$. В соответствии со шкалой определяется оценка CVSS (низкий риск) (таблица 2.5):

Таблица 2.5 - Пример сопоставления взвешенной суммы с оценкой уязвимости CVSS

Оценка ИИ	Шкала уязвимости
0.75 - 1.0	Высокий риск (балл CVSS 7-10)
0.5 - 0.75:	Средний риск (балл CVSS 4-6)
0.0 - 0.5	Низкий риск (балл CVSS 0-3)

При этом следует учитывать следующее: величина оценок ИИ (точность и др.) в каждом конкретном случае зависит от требований к метрикам системы ИИ, которые закладываются при ее проектировании и позже окончательно устанавливаются при вводе в эксплуатацию. Метод нормализации позволяет привести значения метрик, таких как точность, полнота и AUC, к единой шкале (например, от 0 до 1). Это упрощает их сравнение и интеграцию в общую оценку уязвимости. Для нормализации метрик можно использовать формулу (2.29).

$$Normalized\ Value = \frac{Max-Min}{Value-Min}, \quad (2.30)$$

где:

Value – текущее значение метрики;

Min – минимальное значение метрики в наборе данных;

Max – максимальное значение метрики в наборе данных.

Рассматривая характеристики ПИИ в контексте всей группы метрик, включенных в систему оценок CVSS, можно выявить параметры, которые отражают своеобразие уязвимостей вычислительных моделей искусственного интеллекта (в том числе специфику уязвимостей датасетов) на логическом и на инфраструктурном уровне. Создание комбинированной метрики, которая учитывает AUC, точность и полноту, может помочь в оценке общей безопасности

системы. Подобное может быть сделано через взвешивание каждой метрики в зависимости от её важности для конкретной системы. Такой подход допустим при рассмотрении атак, как на логические части (модели) ИИ, так и на ее инфраструктурные элементы. Способ сопоставления метрик должен определять сам аудитор, поскольку вычислительные модели и специфика их применения носит уникальный характер. Следует учесть приведенные ниже факторы, на которые аудитор должен обратить внимание.

При анализе уязвимостей ПИИ по CVSS можно выделить (без детализации сопоставления):

1. Базовые метрики отражают характеристики уязвимости, которые не зависят от времени или контекста. Они включают такие параметры, как: сложность эксплуатации (Exploitability), воздействие (потенциальный ущерб от успешной эксплуатации уязвимости (Impact)). Базовые метрики можно сопоставить с точностью (Accuracy) нейросетей, так как они показывают, насколько правильно модель классифицирует объекты и насколько точно определяет уязвимости. Допустим, существует уязвимость в программном обеспечении, которая может быть использована только при определенных условиях. Если нейросеть имеет высокую точность (например, 95 %), это означает, что она правильно идентифицировала 95 % уязвимостей. Это указывает на то, что сложность эксплуатации уязвимости может быть низкой.

2. Контекстные метрики зависят от окружения, в котором находится уязвимость. Они учитывают факторы, такие как:

− уровень доступа (Access Vector): как злоумышленник может получить доступ к уязвимости.

− условия эксплуатации (Access Complexity): дополнительные условия, которые могут повлиять на возможность эксплуатации.

Контекстные метрики можно сопоставить с полнотой (Recall) нейросетей, так как они помогают оценить, насколько хорошо модель обнаруживает все возможные уязвимости в различных условиях. Предположим, что нейросеть

обнаруживает 80 % уязвимостей в определенном контексте (например, в специфической конфигурации системы). Это можно сравнить с контекстными метриками CVSS, которые оценивали бы, насколько легко можно использовать уязвимость в разных условиях. Если нейросеть работает лучше в определенных условиях, это может означать, что контекстные метрики CVSS в этом сценарии также имеют высокие значения.

3. Временные метрики отражают изменения в характеристиках уязвимости со временем. Они включают: статус уязвимости (Exploitability), обновления (Remediation Level).

Временные метрики можно сопоставить с AUC (Area Under Curve), так как они отражают способность модели адаптироваться к изменениям во времени и оценивать риски на основе новых данных. Если рассматривать уязвимости ПИИ (СИИ) в контексте каждой группы метрик с детализацией сопоставления, то можно выделить следующие особенности:

1. Параметр AUC может быть сопоставлен с Confidentiality, так как высокая AUC указывает на способность модели различать классы и минимизировать ложные срабатывания, что важно для защиты конфиденциальности. Точность может быть сопоставлена с метрикой Access Complexity, так как высокая точность модели подразумевает низкую вероятность ошибок в классификации. Полнота может быть сопоставлена с метрикой Integrity, поскольку высокая полнота означает, что модель успешно обнаруживает большинство положительных примеров, что критично для защиты целостности данных.

2. Временные метрики учитывают факторы, которые могут изменяться со временем, такие как наличие эксплойтов или патчей. AUC может быть сопоставлен с Exploit Code Maturity, так как высокий AUC указывает на высокую надежность модели и ее способность предотвращать эксплуатацию уязвимостей. Точность может быть связана с метрикой Remediation Level, поскольку высокая точность в классификации позволяет быстрее реагировать на угрозы. Полнота может быть сопоставлена с метрикой Report Confidence, так как высокая полнота свидетельствует о высокой уверенности в обнаружении угроз.

3. Контекстные метрики зависят от конкретного окружения и позволяют учитывать уникальные характеристики системы (CR, IR, AR). Метрики AUC, точность и полнота могут быть использованы для оценки специфики системы и ее требований к безопасности. Например, высокий уровень AUC может указывать на необходимость строгих требований к конфиденциальности. Высокая точность может быть критична для систем с высокими требованиями к целостности данных. Высокая полнота важна для систем, где доступность информации является приоритетом.

Сопоставление базовых, временных и контекстных метрик CVSS с метриками точности нейросетей позволяет глубже понять риски безопасности и принять обоснованные решения о защите информации. Таким образом, анализ показывает возможность использования распространённых метрик оценки уязвимости информационных систем к системам ИИ. Однако при сопоставлении обязательно следует учитывать специфику выбранных параметров уязвимостей ИИ и выбранных стандартов для количественной оценки уязвимостей, а также специфику их версий. Исходя из анализа, можно рекомендовать следующие:

- сопоставлять WS с отдельными метриками CVSS;

- целесообразно сопоставлять с частью Base Metrics (Базовые метрики CVSS).

Параметр WS (Weighted Sum) лучше всего сопоставлять с Impact Score из базовых метрик CVSS, а также учесть параметр Attack Complexity (сложность атаки) из для отражения специфики точности модели в рамках ее эксплуатируемости (Exploitability Coefficient = 1 − Attack Complexity (нормализовано от 0 до 1)). Произведение этих параметров будет отражать специфику уязвимости моделей ИИ как отдельного компонента.

Для оценки доступности составных наборов данных и заимствованных компонентов, применяемых при атакующих воздействиях, можно применить модель Басса. В данном случае получение новой технологии атаки злоумышленником зависит от числа уже ранее использованных случаев заимствований технологий. Требуется рассмотреть это в контексте того, что злоумышленник может получить датасет или информацию о модели ИИ, поскольку

они доступны. Эти датасеты или модели злоумышленник может изменить или, изучив их, подобрать способ обмана системы ИИ.

Для оценки угрозы, связанной с возможностью получения злоумышленником датасета или вычислительной модели, можно использовать несколько параметров. Эти параметры помогут понять, насколько вероятно, что злоумышленник сможет получить доступ к необходимым ресурсам и использовать их для атаки на системы ИИ.

Формулы для оценки параметров:

1. Вероятность получения датасета или модели:

$$\mathrm{P(D)} = p + \frac{q}{m} y(D)P, \tag{2.31}$$

где:

p – базовая вероятность доступа к датасету или модели. Это может быть оценено на основе того, насколько открыты или защищены данные;

q – коэффициент, отражающий влияние заимствований. Это значение можно определить на основе статистики о том, сколько раз технологии или данные были заимствованы другими злоумышленниками;

m – число потенциальных злоумышленников. Это можно оценить на основе анализа целевой аудитории или сообщества, которое может быть заинтересовано в использовании данных;

$y(D)$ – прирост вероятности получения с учетом уже совершенных заимствований. Это значение можно оценить на основе анализа предыдущих случаев заимствований.

2. Вероятность успешной модификации или изучения.

$$P(M) = f(M) \cdot P(D), \tag{2.32}$$

где $f(M)$ — функция, описывающая вероятность успешной модификации или изучения модели. Это значение можно оценить на основе сложности модификации

модели или данных. Если модель легко поддается изменениям, то $f(M)$ будет высоким.

Соотношение с метриками CVSS:

1. Сложность эксплуатации (Exploitability). Вероятность получения датасета или модели $P(D)$ может быть использована для оценки сложности эксплуатации. Если $P(D)$ высока, это указывает на низкую сложность эксплуатации.

2. Воздействие на конфиденциальность (Confidentiality Impact). Вероятность успешной модификации или изучения $P(M)$ может быть связана с воздействием на конфиденциальность. Если злоумышленник может модифицировать модель, это может привести к утечке конфиденциальной информации.

Модификации входных данных сильно связаны с проблемой устойчивости модели. Под устойчивостью модели понимают меру его чувствительности к возмущениям в исходных данных. Модель считается устойчивой, если при обучении погрешность в изначальных данных поэтапно не снижает точность классификации. При этом достичь неустойчивости работы модели можно другим способом: данные могут быть модифицированы на этапе тренировки, когда в обучающий набор добавляются записи, которые снижают качество классификации. Соответственно, возникает проблема доверия к обучающим датасетам. Эту проблему можно решить путем введения процедур обязательной верификации обучающих выборок. Таким образом, для интеграции WS и CVSS требуется:

1. Этап 1: Рассчитайте WS и\или $P(\mathrm{D})$ для модели ИИ.

2. Этап 2: Сопоставьте компоненты $P(\mathrm{D})$ и WS с метриками CVSS (перевести WS в Impact Score через пороговые значения и уточнить оценку через Exploitability)

3. Этап 3: Использовать CVSS-калькулятор для агрегации баллов.

Моделирование атакующих стратегий с использованием МППР опирается на параметризацию, где каждый компонент играет строго определенную роль в формировании оптимальной политики. В моделях МППР дисконтирование регулирует временной горизонт планирования, определяя относительный вес

немедленных и будущих вознаграждений. При этом сами функции вознаграждения позволяют отразить специфику как приоритетов при переходах между состояниями и самих переходов, так и описать систему ИИ, а также ПИИ, с позиций атакующего. Это позволяет моделировать разные ситуации, анализировать развитие атаки в формируемых ситуациях. Ее ключевые параметры выступают как модификаторы, определяющие стратегическое поведение злоумышленника. Определение вознаграждения при полной известности инфраструктуры (атаки «белый ящик») дает возможность атакующему до атаки подготовить требуемый инструментарий и не требует сложных вычислений. Для вычисления вознаграждения за переход из одного состояния в другое в модели марковских процессов принятия решений (МППР), учитывая метрику уязвимостей по CVSS, возможность взаимодействия между узлами сети, возможность сопряжения уязвимостей и приближающее состояние к целевому состоянию по методике MITRE ATLAS, можно использовать следующую формулу:

$$\mathrm{R}(s, a, s') = CVSS(s, s') \cdot I(s, s') \cdot Y(s, s') \cdot L(s, a) \cdot D(s, s') \cdot A(s'), \quad (2.33)$$

где:

$CVSS(s, s')$ – оценка уязвимости по CVSS для перехода из состояния s в состояние s', пропорциональна награде – обратная оценка для блокировки, зависит от эффективности системы безопасности;

$D(s, s')$ – коэффициент затухания уязвимости;

$L(s, a)$ – коэффициент легальности от 0 до 1 (всегда не равно 0);

$I(s, s')$ – индикатор возможности взаимодействия между узлами и системными компонентами (1, если связь есть; 0, если нет);

$Y(s, s')$ – коэффициент сопряжения уязвимостей (1, если результат воздействия на уязвимость позволяет с помощью действий эксплуатировать другую; 0, если нет);

$A(s')$ — индикатор того, приближает ли новое состояние злоумышленника к целевому состоянию (1, если приближает; 0, если нет), данный параметр используется при расчётах наград в режиме off-line.

Перемножение используется для учета всех факторов одновременно. Если хотя бы один из факторов равен нулю (например, нет связи между узлами), то итоговое вознаграждение также будет равно нулю. Это отражает ситуацию, когда отсутствие одного элемента делает переход невозможным или незначимым. Также это может быть полезно для строгого ограничения вознаграждения в случае отсутствия взаимодействия, легальности или других критически важных факторов. Функция не включает явные штрафы, но учитывает легальность и затухание как факторы, которые могут уменьшить общее вознаграждение. Это делает её более позитивной, так как она не наказывает за возвращение в ранее посещенные состояния, если они все еще легальны и эффективны.

Таким образом, при оценке наград можно отметить следующее:

1. Значение CVSS варьируется от 0 до 10, где 10 — это максимальная уязвимость. Это значение отражает серьезность угрозы и является основным компонентом вознаграждения.

2. Фиксация возможности взаимодействия между узлами сети, компонентами инфраструктуры (I=1; если связи нет, то I=0) позволяет учитывать только те уязвимости, которые могут быть использованы в контексте активного взаимодействия.

3. Фиксация возможности сопряжения уязвимостей важна для оценки комбинированного эффекта уязвимостей.

4. Приближающее состояние по методике MITRE ATLAS позволяет учитывать, насколько переход в новое состояние может быть выгоден для злоумышленника.

В случае определения наград при неполной известности инфраструктуры, а также допустимости влияния «человеческого фактора» на переходы между состояниями, то есть при наличии обратных переходов, повторов переходов, применим другой подход к формированию функции $R(s, a, s')$. Он предполагает

использование экспертной оценки (по Методике ФСТЭК) параметров вознаграждения (для каждого параметра формулируются ключевые вопросы, которые помогут экспертам оценить их значимость в контексте моделирования атак на ИИ; учитывается контекст задачи, то есть весовые коэффициенты должны отражать специфику моделируемой инфраструктуры и целей атаки (если сеть сильно сегментирована,w_i может быть увеличен; если атака предполагает цепочку уязвимостей, w_c получает больший вес; если законность действий не имеет значения для модели (например, чисто атакующий сценарий) w_l можно минимизировать). Важно подчеркнуть, что неучтенные данные о ПИИ и ее инфраструктуре могут не исключаться из предлагаемой формулы вознаграждения, то есть могут присутвовать недостоверные переходы между состояниями. В этом случае при вычислении вознаграждения за переход из одного состояния в другое в модели МППР учитывается:

R(s, a, s', n) – общее вознаграждение за переход из состояния s в состояние s' при выполнении действия a и использовании уязвимости n раз;

w_{cvss} - вес, определяющий важность оценки уязвимости по метрике CVSS;

$CVSS(s, s')$ – оценка уязвимости по метрике CVSS для перехода из состояния s в состояние s' (это значение варьируется от 0 до 10, где 10 — это максимальная уязвимость);

$I(s, s')$ – индикатор возможности взаимодействия между узлами сети (0 или 1). Если связь между узлами существует $I(s, s')=1$, иначе $I(s, s')=0$;

w_i – вес, определяющий важность индикатора возможности взаимодействия между узлами сети;

$Y(s, s')$ – индикатор возможности сопряжения уязвимостей (0 или 1). Если одна уязвимость может эксплуатировать другую, то $C(s, s') = 1$, иначе $C(s, s') = 0$.

w_c – вес, определяющий важность индикатора возможности сопряжения уязвимостей;

A(s, s') – индикатор приближения к целевому состоянию по методике MITRE ATLAS). Он может определяться двумя способами. Первый способ более

ресурсозатратный, позволяет учитывать множество посещенных состояний. Он предполагает следующие условия:

$$A(s,s') = \begin{cases} \lambda \cdot (d(s,g) - d(s',g)), \text{если } s' \notin V \text{ и } d(s',g) < d(s,g) \\ 0, \text{если } s' \notin V \text{ и } d(s',g) \geq d(s,g) \\ -\gamma \cdot n(s'), \text{ если } s' \in V, \end{cases}$$

где:

d(x,g) - кратчайшее расстояние от состояния x до цели g;

V - множество посещенных состояний в текущей траектории;

n(s′) - количество посещений состояния s′;

λ - коэффициент награды (λ>0);

γ - коэффициент штрафа (γ>0).

Второй способ используется при необходимости упрощения (снижает точность, но повышает скорость расчета R, что важно, если аудитор производит анализ без применения вычислительных ресурсов). В этом случае допустимо использовать следующий вариант определения A: если новое состояние s' приближает злоумышленника к его цели, то A(s') = 1, A(s') = - 1, если новое состояние s' отдаляет злоумышленника от его цели, переходя в ранее уже посещенное состояние; A(s') = 0, если расстояние до цели не меняется или отдаляется от целевого состояния, переходя в новое, которое злоумышленник ранее не посещал.

w_a - вес, определяющий важность индикатора приближения к целевому состоянию;

$L(s, a)$ – индикатор легальности действия (0 или 1). Если действие a легально в состоянии s, то $L(s, a)$=1, иначе $L(s,a)$=0;

w_l - вес, определяющий важность индикатора легальности действия;

n — количество раз, когда уязвимость использовалась (количество переходов);

$D(n)$ — коэффициент затухания, выражающий специфику защиты ПИИ, которая проявляется, в том числе, в возможности распознавания угрозы (чем ближе

к целевому состоянию, тем больше награда; если состояние ранее не посещалось, штраф будет меньше);

δ - штраф за шаг, где δ>0, используется для подавления циклов, задержек при обратных переходах. Определяется двумя способами:

1. При моделировании без использования тактик δ определяется как константа (например, 0,1 может выступать как стартовое значение для баланса между A(s,s′). Значение δ определяется по правилу δ=0.5/K, где K - средняя длина кратчайшего пути от начального состояния до цели (по алгоритму BFS), а числитель 0.5 – это максимальная награда за один «полезный» шаг (это коэффициент награды λ из определния параметра A(s,s′), при Δd=1 (идеальное приближение к цели) A(s,s′)=λ·1=0.5). Для общего случая штраф за шаг должен быть пропорционален этой награде, но обратно пропорционален сложности пути;

2. Значение δ определяется как вычисляемый параметр при работе с тактиками (T(s) — индекс тактики состояния s), где δ(s,s′)= δ_0·|T(s)−T(s′)|.

Потенциально параметр δ можно задать как $\beta_0 \cdot \mathrm{dist}(s, s')$ для того, чтобы учесть разную специфику описания переходов между состояниями (в этом случае dist(s,s′) - метрика расстояния между состояниями): количество хопов время выполнения перехода и др. При упрощенном рассмотрении параметра A(s,s′) штраф за шаг будет определяться как $\delta_{\text{eff}}(s,a)= \delta_0 \cdot (1+h \cdot n_{\text{global}})$, где h - коэффициент усиления, а n_{global} - общее число шагов в траектории).

Приведенные определения также следует использовать для реализации калибровочного процесса, который предполагает (в зависимости от способа определения δ): построение графа атак, определения K, δ или, в случае использования тактик MITRE или ФСТЭК, вычисление $T_{max} = \max|T_i - T_j|$ для определения $\delta_0 = b/T_{max}$, затем δ(s,s') (опционально при учете специфики атакующих можно задать $C_{attacker}$ - параметр направленности на короткие (C>1) или многоэтапные атаки (C<1) и получить $\delta_{final} = \delta \cdot C_{attacker}$). Значение b может быть эмпирически подобрано как константа (например, b=0.2), чтобы обеспечить баланс между подавлением аномалий и разрешением легитимных переходов. В некоторых

реализациях и исследованиях, связанных с симуляцией атак (включая проекты, вдохновлённые MITRE CALDERA), могут использоваться подобные подходы к определению штрафов. Альтернативно, b может быть адаптивной величиной b = $\lambda \cdot P_{natural}$, где P_{natura} - это статистическая мера, показывающая, какая доля переходов между тактиками MITRE в системе соответствует реальным (естественным) сценариям атак.

Формула вознаграждения 2.33 может рассматриваться как частный случай формулы 2.34 и может применяться с учетом разной степенью известности атакуемой инфраструктуры:

$$\mathrm{R}(s,a,s,n) = \Big(w_{cvss}\, CVSS(s,s') + w_i \cdot I(s,s') + w_y \cdot Y(s,s') + w_l \cdot L(s,a) + w_a \cdot A(s')\Big) \cdot D(n) - \delta. \quad (2.34)$$

Для определения параметров $R(s,a,s,n)$ нужно реализовать следующее:

1. Определить значения всех компонентов. Вычислите значения для каждого компонента формулы.

2. Инициализировать значение весов: $w_{cvss}, w_i, w_c, w_a, w_l$. Параметры весов определяются экспертным методом в зависимости от важности рассматриваемого параметра. При наличии у аудитора информации логов и иной информации об анализируемой системе ИИ допустимо использовать следующие подходы к определению весов:

- определение w_{cvss} предполагает сбор значений CVSS для уязвимостей системы, вычисление среднего μ_{cvss} по всем уязвимостям системы и долю эксплуатируемых уязвимостей $\mathrm{f}_{exploit}$ (например, из Exploit-DB), нормирование следующим образом: $w_{cvss} = \mu_{cvss} * \mathrm{f}_{exploit}/10$;

- определение веса важности связности w_i предполагает построение графа (узлы — компоненты, рёбра – I(s, s') = 1), вычисление доли связанных пар N узлов: $w_i = \sum I(s,s')\,/N^2$;

- определение w_c предполагает оценку доли переходов с Y(s, s') = 1 из данных атак (MITRE ATLAS) или графа уязвимостей с учетом количества переходов (M): $w_y = \sum \mathrm{Y}(s, s') / M$.

- определение веса важность приближения к цели как $w_a = \frac{T_a}{T}$ предполагает использование доли переходов T_a приближающих к цели (A(s') = 1) с учетом общего количества переходов T, при этом если $w_a < w_{cvss,}$ значение актуально, в противном случае $w_a = 0$ (необходимо, чтобы избежать нереалистичных сценариев);

- определение w_l предполагает оценку легальных действий A(L(s, a) = 1) из сценариев атак при знании вероятности обнаружения нелегальных действий (P_{detect}): $w_l = \frac{\sum L(s,a)}{A} * (1 - P_{detect})$;

Далее следует определить относительное значение каждого компонента на основе контекста задачи. В общем случае предпочтительным подходом к определению весов является экспертный метод, поскольку недостаток и недостоверность информации, которая требуется при статистических исследованиях, вычислениях относительной значимости, об анализируемой системе и параметрах атаки частое явление.

3. Использовать в формуле полученные значения и веса для вычисления итогового вознаграждения.

4. Ввести параметры, актуальные для состояний модели (неактуальные исключаются).

Можно отметить следующие преимущества приводимого подхода: комплексность, гибкость (возможность настройки весов позволяет адаптировать модель к различным сценариям), учет легальности действий. Функция включает явные штрафы за нелегальные действия и за возвращение в ранее посещенные состояния, что может более точно отражать риски и последствия действий злоумышленника. В функции вознаграждения складываются положительные компоненты и вычитаются штрафы, что позволяет более гибко управлять вознаграждением. Это может быть полезным для того, чтобы при необходимости вознаграждение оставалось положительным, даже если есть негативные факторы.

Параметр защиты учитывается в наградах как затухание $D(n)$. где n — это счётчик, показывающий, сколько раз злоумышленник уже эксплуатировал данную уязвимость. Можно использовать несколько подходов для его определения.

1. Экспоненциальное затухание определяется следующим набором формул (2.36 - 2.37).

$$D(n) = e^{-\beta n} \ , \tag{2.35}$$

$$\mathrm{R}(s, a, s') = \mathrm{R}_0 \cdot e^{-\beta n}, \tag{2.36}$$

$$\frac{\mathrm{R}(n)}{\mathrm{R}_0} = e^{-\beta n}, \tag{2.37}$$

$$\beta = -\frac{1}{\mathrm{т}} \ln\left(\frac{R(n)}{R_0}\right). \tag{2.38}$$

Преимущества: быстрое снижение вознаграждения в начале, математическая простота, плавное затухание. Также это может приводить к резкому обнулению вознаграждения (требует вычисления β).

2. Линейное затухание (2.39 - 2.41).

$$D(n) = \frac{1}{1+k \cdot n} \ , \tag{2.39}$$

$$\mathrm{R}(s, a, s') = \mathrm{R}_0 \cdot D(n) = \mathrm{R}_0 \cdot \frac{1}{1+k \cdot n}, \tag{2.40}$$

$$k = \frac{\mathrm{R}_0 - \mathrm{R}(n)}{\mathrm{R}(n) \cdot n}. \tag{2.41}$$

Преимущества в том, что более интуитивное, гибкая настройка скорости затухания с помощью коэффициент затухания k, позволяет избегать резкого обнуления награды, но при этом меньшая распространенность в моделировании (требует определения k). Выбор подхода к затуханию зависит от конкретного контекста и

требований модели. Экспоненциальное затухание может быть предпочтительным, если требуется быстрое снижение вознаграждения, в то время как линейное затухание будет более подходящим для плавной динамики и возможности гибкой настройки. Формула ценности (2.42) с учетом функции вознаграждения $R(s_t, a, s_{t+1})$ примет следующий вид:.

$$\mathrm{V(s)} = \sum_{t=0}^{\infty} \gamma^t \cdot E[\mathrm{R}(s_t, a, s_{t+1})] \quad . \tag{2.42}$$

Коэффициент затухания можно выразить через оценку состояния $V(s)$ и вероятности переходов $P(s' \mid s,\ a)$. Чтобы выразить экспоненциальный коэффициент затухания k через известные значения вознаграждения R, можно использовать следующие шаги (приведено в формулах 2.43 - 2.48):

$$\mathrm{R}(s, a, s') = \mathrm{R}_0 \cdot D(n) = \mathrm{R}_0 \cdot \frac{1}{1+k \cdot n}\,, \tag{2.43}$$

$$\mathrm{V(s)} = \sum_{S'} P(s'|s, a) \cdot (\mathrm{R}_0 \cdot \frac{1}{1+k \cdot n} + \gamma V(s'))\,, \tag{2.44}$$

$$\frac{\mathrm{V(s)}}{R_0} = \sum_{S'} P(s'|s, a) \cdot (\frac{1}{1+k \cdot n} + \frac{\gamma V(s')}{\mathrm{R}_0})\,, \tag{2.45}$$

$$\mathrm{x(s)} = \sum_{S'} P(s'|s, a) \cdot (\frac{1}{1+k \cdot n} + \gamma x(s'))\,, \tag{2.46}$$

$$\mathrm{x(s)} \cdot (s'|s, a) = \sum_{S'} P(s'|s, a) \cdot \left(\frac{1}{1+k \cdot n} + \gamma \sum_{S'} P(s'|s, a) \cdot x(s')\right), \tag{2.47}$$

$$k = \frac{\sum_{S'} P(s'|s,a) - x(s)}{n \cdot x(s)}. \tag{2.48}$$

Таким образом, коэффициент затухания k можно выразить через оценку состояния $V(s)$ и вероятности переходов $P(s' \mid s, a)$.

Использования двух типов коэффициентов затухания, как отражения специфики систем защиты. Объединенный тип затухания определяется следующим образом (2.49 - 2.51):

$$\mathrm{R}(s,a,s') = \mathrm{R}_0 \cdot \frac{1}{1+k\cdot n}, \tag{2.49}$$

$$\mathrm{R}(s,a,s') = \mathrm{R}_0 \cdot \left(\alpha \cdot e^{-\beta n} + (1-\alpha)\cdot \frac{1}{1+k\cdot n}\right), \tag{2.50}$$

$$V(s) = \sum_{s'} P(s'|s,a)\cdot (\mathrm{R}_0 \cdot \left(\alpha \cdot e^{-\beta n} + (1-\alpha)\cdot \frac{1}{1+k\cdot n}\right) + \gamma V(s')). \tag{2.51}$$

Общие преимущества способа расчета:

1. Учет динамики угроз. Использование экспоненциального затухания позволяет учитывать изменения в угрозах со временем. Это важно, так как уязвимости могут устаревать, и их актуальность может снижаться.

2. Гибкость. Каждый фактор (CVSS, возможность взаимодействия, сопряжение уязвимостей, приближение к целевому состоянию и легальность) имеет свой вес, что позволяет адаптировать модель под конкретные сценарии и условия.

3. Взаимодействие факторов. Возможность добавления дополнительных членов для учета взаимодействий между факторами (например, $R_{adj} = R + \omega_{inter} \cdot (I \cdot C)$) улучшает точность оценки вознаграждения).

4. Использование бинарных значений (0 или 1) для некоторых факторов позволяет однозначно трактовать, влияет ли фактор на вознаграждение или нет.

Однако можно выделить недостатки следующего характера:

1. Сложность оценки. Оценка всех факторов и назначение весов может быть сложной задачей, требующей значительных усилий и экспертизы.

2. Необходимость валидации. Модель требует валидации и тестирования на реальных данных для подтверждения её эффективности. Без этого сложно гарантировать, что модель будет работать в различных сценариях.

3. Потенциальная переоценка. Если веса неправильно настроены, это может привести к перекосу в оценках вознаграждений, что повлияет на принятие решений.

Чтобы выразить экспоненциальный коэффициент затухания β через известные значения вознаграждения R, можно использовать следующие шаги (2.52 – 2.58):

1. Предположим, что вознаграждение имеет экспоненциальную форму затухания:

$$\mathrm{R}(s, a, s') = \mathrm{R}_0 \cdot e - \beta n, \tag{2.52}$$

где R_0 – начальное вознаграждение, n – количество переходов.

2. Подставим это выражение в формулу оценки состояния:

$$V(s) = s' \sum P(\, s' \mid s, a\,) \cdot \big(\mathrm{R}_0 \cdot e - \beta n + \gamma V(s')\big). \tag{2.53}$$

3. Выразим β из этого уравнения. Для простоты предположим, что $P(s'|s, a)$ и γ не зависят от n:

$$V(s) = \mathrm{R}_0 \cdot e - \beta n \cdot s' \sum P(s' \mid s, a) + \gamma s' \sum P(s' \mid s, a) \cdot V(s')\,. \tag{2.54}$$

4. Разделим обе части на R_0:

$$V(s) = e - \beta n \cdot s' \sum P(s' \mid s, a) + \gamma s' \sum P(s' \mid s, a) \cdot \mathrm{R}_0 V(s'). \tag{2.55}$$

5. Пусть $V(s)R0 = x(s)$, тогда:

$$\mathrm{x(s)} = \mathrm{e} - \beta \mathrm{n} \cdot \mathrm{s}' \sum \mathrm{P}(\, \mathrm{s}' \mid \mathrm{s}, \mathrm{a}\,) + \gamma \mathrm{s}' \sum \mathrm{P}(\, \mathrm{s}' \mid \mathrm{s}, \mathrm{a}\,) \cdot \mathrm{x(s}'). \tag{2.56}$$

6. Применим натуральный логарифм к обеим частям:

$$\ln(x(s)) = -\beta n + \ln(s' \sum P(s' \mid s, a)) \ . \quad (2.57)$$

$$\beta = -n1(\ln(x(s)) - \ln(s' \sum P(s' \mid s, a))) \ . \quad (2.58)$$

В итоге можно выразить параметры защиты, исходя из знания конечных значений функций вознаграждений.

Каждый из приведенных параметров функции вознаграждения позволяет описать специфическую особенность атакуемой системы ИИ и, в целом, саму систему. Обобщенная формула (2.59) вознаграждения позволяет учесть множество параметров и может использоваться при моделировании атак на системы другого типа.

$$R(s, a, s', n) = (\sum_k w_k \cdot X_k(s, s')) \cdot D(n, s') - \beta_0 \cdot dist(s, s'), \quad (2.59)$$

где: X_k — параметры вознаграждения за атакующие воздействия;

dist(s,s') — выбранная метрика расстояния.

2.4 Определение ограничений при моделировании атак

В банке данных угроз безопасности информации ФСТЭК РФ присутствуют несколько возможных угроз искусственному интеллекту. К ним относятся:

1. УБИ. 218. Угроза раскрытия информации о модели машинного обучения.

2. УБИ. 219. Угроза хищения обучающих данных.

3. УБИ. 220. Угроза нарушения функционирования («обхода») средств, реализующих технологии искусственного интеллекта.

4. УБИ. 221. Угроза модификации модели машинного обучения путем искажения («отравления») обучающих данных.

5. УБИ. 222. Угроза подмены модели машинного обучения.

Определение сценария реализации угроз предусматривает установление последовательности возможных тактик и техник. Однако такое малое количество угроз, как и актуальна Методика ФСТЭК, не позволяет полностью описать атаки на системы ИИ. Соответственно, необходимо заимствовать более полную и актуальную базу описания MITRE ATLAS. В соответствии с матрицей MITRE ATLAS и Методикой оценки угроз безопасности информации ФСТЭК РФ можно привести обобщенный перечень тактик и техник для реализации угроз на системы машинного обучения (таблица 2.6).

Таким образом, описание возможных тактик и техник рассмотрено в соответствии с глобальной базой знаний о тактиках и способах (учитывая фактор их актуальности), основанных на реальных наблюдениях метрической базы MITRE ATLAS (она может использоваться как база об инцидентах, тактиках и техниках, применяемых нарушителем для атак на системы машинного обучения и нейронные сети) [1].

Следует отметить, что в произвольный момент времени атакуемая система ИИ может находиться в любом из состояний (десяти по методике ФСТЭК или четырнадцати по MITRE ATLAS), которые образуют пространство состояний [2]. Подобный подход удобен при рассмотрении посредством марковских процессов принятия решений (МППР) последовательности атакующих воздействий как череды смены состояний и одновременно позволяет отследить их сопряжённость (для этого достаточно применения системы линейных уравнений) [134].

Необходимо подчеркнуть сложную природу состояний, поскольку они учитывают с одной стороны, тип тактики, а, с другой стороны, поскольку данные состояния достигаются при успешной эксплуатации уязвимости (узлов сети, их программного обеспечения), в пространстве состояний также должна учитываться специфика процесса достижения успеха, то есть действия (a_i), которые позволяют эксплуатировать уязвимости, а также сами уязвимости.

Таблица 2.6 - Перечень тактик атак на модели

Тактика ФСТЭК	Тактика MITRE ATLAS	Идентичные техники	Схожие техники
Т1. Сбор информации о системах и сетях	Разведка	- Т1.1. Поиск сайтов, принадлежащих жертве (сбор информации с общедоступных сайтов жертвы); - Т1.2, Т1.4. Active Scanning (активное сканирование устройств и сетей); - Т1.11. Фишинг (фишинг для сбора данных).	- Т1.3. Ппоиск сайтов, принадлежащих жертвам (пассивный сбор данных); - Т1.5. Использование уязвимостей общедоступных приложений (для сбора информации); - Т1.15. Приобретение общедоступных артефактов машинного обучения (покупка баз данных или артефактов)
Т2. Получение первоначального доступа	Начальный доступ / Доступ к модели ML	- Т2.8. Фишинг (социальная инженерия/фишинг для получения доступа); - Т2.5. Exploit Public-Facing Application (эксплуатация уязвимостей публичных приложений).	- Т2.1 - Продукт или услуга с поддержкой машинного обучения (доступ через внешние сервисы); - Т2.10, Т2.11. Действительные учётные записи (использование учётных данных); - Т2.12. Компрометация цепочки поставок машинного обучения (компрометация через сторонние организации)
Т3. Внедрение и исполнение вредоносного ПО	Исполнение	- Т3.5. Command and Scripting Interpreter (удаленное выполнение кода)	- Т3.1 - User Execution (запуск через действия пользователя); - Т3.3 - Компрометация плагина LLM (автоматическая загрузка кода); - Т3.9 - Компрометация цепочки поставок машинного обучения (подмена ссылок на ПО)
Т4. Закрепление (сохранение доступа)	Настойчивость	- Т4.2. Действительные учётные записи (использование штатных средств для доступа)	- Т4.5 - Компрометация плагина LLM (автозагрузка через системные механизмы); - Т4.6 - Компрометация цепочки поставок машинного обучения (компрометация прошивок);
Т5. Управление вредоносным ПО	Командование и контроль	- (отсутствуют идентичные техники)	- Т5.1 - Извлечение данных с помощ ью киберсредств (управление через стандартные протоколы, например, RDP/SSH); - Т5.6 - Экфильтрация с помощью API машинного обучения (проксирование трафика для управления);

Продолжение таблицы 2.6

Тактика ФСТЭК	Тактика MITRE ATLAS	Идентичные техники	Схожие техники
Т6. Повышение привилегий	Повышение Привилегий	- Т6.1. Незащищённые учётные данные (кража учётных данных для повышения привилегий)	- Т6.3. Использование уязвимостей общедоступных приложений (для повышения привилегий); - Т6.5. Взлом LLM (манипуляции с токенами/сессиями);
Т7. Сокрытие действий	Уклонение от защиты	- Т7.1. Уклонение от модели машинного обучения (использование законных инструментов для уклонения)	- Т7.9. Обход модели машинного обучения (подписание кода скомпрометированными сертификатами); - Т7.17. Создание вредоносных данных (обфускация/шифрование данных);
Т8. Получение доступа (распространение)	Боковое перемещение	- (отсутствуют идентичные техники)	- Т8.3. Саморепликация LLM (распространение через групповые политики); - Т8.5. Продукт или услуга с поддержкой машинного обучения (изменение конфигурации сети для доступа);
Т9. Сбор и вывод информации	Сбор и Эксфильтрация	- Т9.14. Вывод данных с помощью киберсредств (вывод данных через общедоступные ресурсы)	- Т9.1. Данные из локальной системы (сбор данных через стандартные протоколы); - Т9.12. Извлечение данных с помощью API машинного обучения (стеганография для вывода данных);
Т10. Несанкционированное воздействие	Воздействие	- Т10.10. Отказ в обслуживании (организация отказа в обслуживании)	- Т10.8. Нарушение целостности набора данных (уничтожение данных); - Т10.12. Нарушение целостности модели машинного обучения (воздействие на АСУ или модели машинного обучения).

Для определения специфики состояний *S* и действий *A* модели МППР, то есть условий их реализации с учётом ограничений инфраструктуры атакуемых систем ИИ, можно произвести типологизацию тактик MITRE ATLAS (как совокупности техник или техник, если требуется больший уровень детализации анализа). При этом определяется: требуется ли обязательный доступ к интерфейсам вычислительных систем, другим элементам ИИ или нет. Это позволяет определить доступные для внешнего злоумышленника тактики (и их техники) с учетом

особенностей инфраструктуры ПИИ (блок датасетов; блок вычислительной модели; внутренняя техническая инфраструктура). Возможность реализации тактики (техники) определяется на основе соблюдения условий, приведенных в таблице 2.7. При этом следует учитывать различия в представляемых множествах: с одной стороны, множество S обозначает состояние как этап атаки (компрометации), и под состоянием может пониматься тактика, которую достиг злоумышленник; с другой стороны, достижение тактики дает возможность эксплуатации множества техник (действий), которые возможны в достигнутом состоянии-тактике, и, соответственно воспринимается как множества доступных действий в состоянии (As_i). Порядок реализации способа определения следующий:

1. В первой фазе определяется, требуется доступ во время атаки к вычислениям (в том числе возможность доступа к внешнему интерфейсу ввода данных для вычислений) ИИ (множество *Авр*) или не требуется (множество *Анвр*).

2. Во второй фазе на основе результатов первой фазы определяется требуется ли доступ во время атаки к датасетам ИИ (множество *Ads*) или тактика реализуема без доступа во время атаки к датасетам (множество *Ands*).

3. В третьей фазе на основе результатов второй фазы определяется, требуется ли злоумышленнику доступ во время атаки к внутренней технической инфраструктуре систем ИИ (внутренняя локальная сеть, которая отделена от глобальной сети, вспомогательные службы серверов, маршрутизаторы и т.п.) (множество *Ats*) или атаки возможны без доступа к внутренней инфраструктуре (Таблица 2.7) (множество *Ants*). При этом учитываются режимы работы ИИ (множество *OE*): обучение (О); эксплуатация (Э) ИИ.

Таким образом, следуя предложенному порядку действий можно распределить все тактики по MITRE ATLAS, учитывая особенности технической и логической организации ПИИ. Демонстрация реализации способа (принципа) распределения тактик приведена в таблицах 2.7, 2.8, 2.9. При этом следует учитывать, что количество техник и тактик со временем меняется, следовательно, необходимо постоянно актуализировать сведения о них.

Таблица 2.7 - Распределение техник по фазам

Фаза анализа	Условие доступа	Режим работы ИИ	Доступность тактики	Типология MITRE ATLAS
Первая фаза	Требуется ли доступ к интерфейсам вычислительных систем (включая внешний интерфейс ввода данных)?	Обучение / Эксплуатация	Да / Нет	Тактики, требующие доступа к интерфейсам
Вторая фаза	Требуется ли доступ к датасетам ИИ во время атаки?	Обучение / Эксплуатация	Да / Нет	Тактики, требующие доступа к датасетам
Третья фаза	Требуется ли доступ к внутренней технической инфраструктуре (локальная сеть, серверы и т.д.)?	Обучение / Эксплуатация	Да / Нет	Тактики, требующие доступа к инфраструктуре
Итоговое распределение	На основе результатов всех фаз определяется доступность тактик для внешнего злоумышленника	Обучение / Эксплуатация	Да / Нет	Полный список тактик MITRE ATLAS

Таким образом, можно выделить следующие множества доступа:

1. С доступом к вычислениям (Авр): Авр={a∈A|требуется доступ к вычислениям}.

2. Без доступа к вычислениям (Anвр): Anвр=A\Авр.

3. С доступом к датасетам (Ads): Ads={a∈A|требуется доступ к датасетам}.

4. Без доступа к датасетам (Ands): Ands=A\Ads.

5. С доступом к технической инфраструктуре (Ats): Ats={a∈A|требуется доступ к инфраструктуре}.

6. Без доступа к технической инфраструктуре (Ants): Ants=A\Ats.

Таблица 2.8 - Демонстрация способа распределения тактик с учетом ограничений инфраструктуры ПИИ с доступом во время атаки к вычислениям

Режим эксплуатации (множество *OE*)	С доступом во время атаки к вычислениям (пользовательским интерфейсам ИИ) (множество *Авр*)			
	с доступом во время атаки к датасетам ИИ (множество *Ads*)		без доступа во время атаки к датасетам ИИ (множество *Ands*)	
	с доступом к внутренней технич. инфрастр. (множество *Ats*)	без доступа к внутренней технич. нфрастр. (множество *Ants*)	с доступом к внутренней технич. инфрастр. (множество *Ats*)	без доступа к внутренней технич. инфрастр. (множество *Ands*)
Обучение (множество O)	Resource Development, AI Model Access, Execution, Persistence, Privilege Escalation, Defense Evasion, Credential Access, Discovery, Collection, AI Attack Staging, Command and Control, Exfiltration, Impact	AI Model Access, Discovery, Collection, AI Attack Staging, Exfiltration, Impact	Execution, Persistence, Privilege Escalation, Defense Evasion, Credential Access, Discovery, AI Attack Staging, Command and Control, Exfiltration, Impact	AI Model Access, Discovery, AI Attack Staging, Impact
Эксплуатация (множество *E*)	AI Model Access, Execution, Persistence, Privilege Escala-tion, Defense Evasion, Credential Access, Discovery, Collection, AI Attack Staging, Command and Control, Exfiltration, Impact	AI Model Access, Discovery, Collection, AI Attack Staging, Exfiltration, Impact	Execution, Persistence, Privilege Escalation, Defense Evasion, Credential Access, Discovery, AI Attack Staging, Command and Control, Exfiltration, Impact	AI Model Access, Discovery, AI Attack Staging, Impact

Для атак, при которых не требуется обязательный доступ к интерфейсам вычислительных систем, демонстрация принципа распределения тактик и само распределние приведено в таблице 2.9.

Таблица 2.9 - Демонстрация способа распределения тактик с учетом ограничений инфраструктуры ПИИ без доступа во время атаки к вычислениям

Режим эксплуатации (множество *OE*)	Без доступа во время атаки к вычислениям (пользовательским интерфейсам ИИ) (множество *Anвр*)			
	с доступом во время атаки к датасетам ИИ (множество *Ads*)		без доступа во время атаки к датасетам ИИ (множество *Ands*)	
	с доступом к внутренней технич. инфрастр. (множество *Ats*)	без доступа к внутренней технич. нфрастр. (множество *Ants*)	с доступом к внутренней технич. инфрастр. (множество *Ats*)	без доступа к внутренней технич. нфрастр. (множество *Ants*)
Обучение (множество O)	Execution, Persistence, Privilege Escalation, Defense Evasion, Credential Access, Discovery, Collection, AI Attack Staging, Command and Control, Exfiltration, Impact	Collection, AI Attack Staging, Exfiltration, Impact	Execution, Persistence, Privilege Escalation, Defense Evasion, Credential Access, Discovery, Collection, Command and Control, Exfiltration, Impact, AI Attack Staging	Resource Development, AI Attack Staging, Impact
Эксплуатация (множество *E*)	Execution, Persistence, Privilege Escalation, Defense Evasion, Credential Access, Discovery, Collection, AI Attack Staging, Command and Control, Exfiltration, Impact	Collection, AI Attack Staging, Exfiltration, Impact	Execution, Persistence, Privilege Escalation, Defense Evasion, Credential Access, Discovery, Collection, Command and Control, Exfiltration, Impact	Resource Development, AI Attack Staging, Impact

При проведении аудита ИИ и, соответственно, построении сценария атаки, последовательно применяя фазы отбора доступных состояний и соответствующих им наборов действий, можно сузить перечень возникающих последовательностей действий злоумышленника, что приведет к упрощению процесса построения сценария атак. Этот подход допустим, когда техническая информация точно описывает инфраструктурную специфику систем ИИ как подсистем ИС.

Также целесообразно сопоставить техники (следует учитывать, что состав техник постоянно обновляется) с условиями доступа, к которым относятся: доступ к интерфейсу, доступ к датасетам, доступ к инфраструктуре, режим работы системы ИИ. Реализация сопоставления на основе исследования описания действий (техник) и их оценки, как и результаты сопоставления приведены в таблице 2.10 с применеием базы занаий MITRE ATLAS [11, 17 - 20].

Таблица 2.10 – Пример сопоставления техник с условиями доступа

Тактика	Основные техника	Доступ к интерфейсам	Доступ к датасетам	Доступ к инфраструктуре	Режим работы ИИ
Reconnaissance	Search Open Technical Databases	Не требуется	Не требуется	Не требуется	Обучение/Эксплуатация
	Search Open AI Vulnerability Analysis	Не требуется	Не требуется	Не требуется	Обучение/Эксплуатация
	Search Victim-Owned Websites	Требуется (внешний)	Не требуется	Не требуется	Обучение/Эксплуатация
	Search Application Repositories	Не требуется	Не требуется	Не требуется	Обучение/Эксплуатация
	Active Scanning	Требуется (внешний)	Не требуется	Не требуется	Обучение/Эксплуатация

Продолжение таблицы 2.10

Тактика	Основные техника	Доступ к интерфейсам	Доступ к датасетам	Доступ к инфраструктуре	Режим работы ИИ
	Gather RAG-Indexed Targets	Не требуется	Не требуется	Не требуется	Эксплуатация
Resource Development	Acquire Public AI Artifacts	Не требуется	Не требуется	Не требуется	Обучение/Эксплуатация
	Obtain Capabilities	Не требуется	Не требуется	Не требуется	Обучение/Эксплуатация
	Develop Capabilities	Не требуется	Не требуется	Не требуется	Обучение/Эксплуатация
	Acquire Infrastructure	Не требуется	Не требуется	Не требуется	Обучение/Эксплуатация
	Publish Poisoned Datasets	Не требуется	Не требуется	Не требуется	Обучение
	Poison Training Data	Не требуется	Не требуется	Не требуется	Обучение
	Establish Accounts	Требуется (внешний)	Не требуется	Не требуется	Обучение/Эксплуатация
	Publish Poisoned Models	Не требуется	Не требуется	Не требуется	Обучение
	Publish Hallucinated Entities	Не требуется	Не требуется	Не требуется	Обучение/Эксплуатация
	LLM Prompt Crafting	Не требуется	Не требуется	Не требуется	Эксплуатация
	Retrieval Content Crafting	Не требуется	Не требуется	Не требуется	Эксплуатация
	Stage Capabilities	Не требуется	Не требуется	Не требуется	Обучение/Эксплуатация

Продолжение таблицы 2.10

Тактика	Основные техника	Доступ к интерфей-сам	Доступ к датасетам	Доступ к инфра-структуре	Режим работы ИИ
Initial Access	AI Supply Chain Compromise	Требуется (внешний)	Не требу-ется	Частично	Обучение/Эксплу-атация
	Valid Accounts	Требуется (внешний)	Не требу-ется	Частично	Обучение/Эксплу-атация
	Evade AI Model	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	Exploit Public-Facing Application	Требуется (внешний)	Не требу-ется	Частично	Обучение/Эксплу-атация
	Phishing	Требуется (внешний)	Не требу-ется	Не требу-ется	Обучение/Эксплу-атация
	Drive-by Compromise	Требуется (внешний)	Не требу-ется	Не требу-ется	Обучение/Эксплу-атация
AI Model Access	AI Model Inference API Access	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	AI-Enabled Product or Service Access	Требуется API	Не требу-ется	Частично	Эксплуатация
	Physical Environment Access	Требуется (физиче-ский)	Не требу-ется	Частично	Обучение/Эксплу-атация
	Full AI Model Access	Требуется (полный)	Требуется	Требуется	Обучение/Эксплу-атация
Execution	User Execution	Требуется (внешний)	Не требу-ется	Не требу-ется	Обучение/Эксплу-атация
	Command and Scripting Interpreter	Требуется (внешний)	Не требу-ется	Частично	Обучение/Эксплу-атация

Продолжение таблицы 2.10

Тактика	Основные техника	Доступ к интерфей-сам	Доступ к датасетам	Доступ к инфра-структуре	Режим работы ИИ
	LLM Prompt Injection	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	LLM Plugin Compromise	Требуется API	Не требу-ется	Частично	Эксплуатация
Persistence	Poison Training Data	Не требу-ется	Требуется	Частично	Обучение
	Manipulate AI Model	Требуется (полный)	Требуется	Требуется	Обучение/Эксплу-атация
	LLM Prompt Self-Replication	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	RAG Poisoning	Не требу-ется	Частично	Частично	Эксплуатация
Privilege Escalation	LLM Plugin Compromise	Требуется API	Не требу-ется	Частично	Эксплуатация
	LLM Jailbreak	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
Defense Evasion Exfiltration	Evade AI Model	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	LLM Jailbreak	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	LLM Trusted Out-put Components Manipulation	Требуется API	Не требу-ется	Частично	Эксплуатация
	LLM Prompt Obfuscation	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	False RAG Entry Injection	Не требу-ется	Частично	Частично	Эксплуатация

Продолжение таблицы 2.10

Тактика	Основные техника	Доступ к интерфей-сам	Доступ к датасетам	Доступ к инфра-структуре	Режим работы ИИ
	Impersonation	Требуется (внешний)	Не требу-ется	Не требу-ется	Обучение/Эксплу-атация
	Masquerading	Требуется (внешний)	Не требу-ется	Частично	Обучение/Эксплу-атация
	Corrupt AI Model	Требуется (полный)	Требуется	Требуется	Обучение/Эксплу-атация
Credenti al Access	Unsecured Credentials	Требуется (внешний)	Не требу-ется	Частично	Обучение/Эксплу-атация
Discove ry	Discover AI Model Ontology	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	Discover AI Model Family	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	Discover AI Artifacts	Требуется (внешний)	Не требу-ется	Частично	Обучение/Эксплу-атация
	Discover LLM Hallucinations	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	Discover AI Model Outputs	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	Discover LLM System Information	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	Cloud Service Discovery	Требуется (внешний)	Не требу-ется	Частично	Обучение/Эксплу-атация
Collecti on	AI Artifact Collection	Требуется (внешний)	Не требу-ется	Частично	Обучение/Эксплу-атация
	Data from Information Repositories	Требуется (внешний)	Не требу-ется	Частично	Обучение/Эксплу-атация

Продолжение таблицы 2.10

Тактика	Основные техника	Доступ к интерфей-сам	Доступ к датасетам	Доступ к инфра-структуре	Режим работы ИИ
	Data from Local System	Требуется (внешний)	Не требу-ется	Частично	Обучение/Эксплу-атация
AI Attack Staging	Create Proxy AI Model	Требуется (полный)	Требуется	Требуется	Обучение/Эксплу-атация
	Manipulate AI Model	Требуется (полный)	Требуется	Требуется	Обучение/Эксплу-атация
	Verify Attack	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	Craft Adversarial Data	Не требу-ется	Требуется	Не требу-ется	Обучение
Comma nd and Control	Reverse Shell	Требуется (внешний)	Не требу-ется	Частично	Обучение/Эксплу-атация
Exfiltrat ion	Exfiltration via AI In-ference API	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	Exfiltration via Cyber Means	Требуется (внешний)	Не требу-ется	Частично	Обучение/Эксплу-атация
	Extract LLM System Prompt	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	LLM Data Leakage	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	LLM Response Rendering	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
Impact	Evade AI Model	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация

Продолжение таблицы 2.10

Тактика	Основные техника	Доступ к интерфей-сам	Доступ к датасетам	Доступ к инфра-структуре	Режим работы ИИ
	Denial of AI Service	Требуется API	Не требу-ется	Частично	Эксплуатация
	Spamming AI System with Chaff Data	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	Erode AI Model Integrity	Требуется (полный)	Требуется	Требуется	Обучение/Эксплу-атация
	Cost Harvesting	Требуется API	Не требу-ется	Частично	Эксплуатация
	External Harms	Требуется API	Не требу-ется	Не требу-ется	Эксплуатация
	Erode Dataset Integrity	Не требу-ется	Требуется	Частично	Обучение

При распределении тактик следует учитывать актуальное состояние методики (базы) их описания. Из приведенных результатов анализа в таблицах видно, что количество действий, доступных злоумышленнику, находится в зависимости от наличия или отсутствия доступа к вычислительным моделям (их интерфейсам), их датасетам, инфраструктурным компонентам. Таким образом, при определении действий, доступных в достигнутых злоумышленником состояниях, при определении самих состояний модели как фаз атаки и их сопряжений можно учитывать ограничения инфраструктуры ПИИ относимой к логическому уровню взаимодействия ее компонентов (этот уровень, как правило, доступен внешнему нарушителю при атаке).

Формула реализуемых действий (используется как правило) задаёт конструктивное разбиение множества A на восемь подмножеств в зависимости от доступа к ресурсам (2.60). Формула представляет множество A как объединение

восьми подмножеств, каждое из которых соответствует уникальной комбинации требований к доступу (каждая комбинация соответствует одному из столбцов в таблицах 2.8 и 2.9).

$$A = (Avp \cap Ads \cap Ats) \cup (Avp \cap Ads \cap Ants) \cup (Avp \cap Ands \cap Ats) \cup (Avp \cap Ands \cap Ants) \cup (Anvp \cap Ads \cap Ats) \cup (Anvp \cap Ads \cap Ants) \cup (Anvp \cap Ands \cap Ats) \cup (Anvp \cap Ands \cap Ants). \quad (2.60)$$

Эта формализация позволяет детально анализировать и моделировать атаки в системах ИИ, учитывая сложные взаимодействия между доступом, контролем и состояниями системы.

Формула 2.61 представляет собой логическое условие (используется как правило), описывающее свойство каждого элемента $a \in A$. Формула подтверждает, что разбиение с помощью логического условия гарантирует, что каждое действие $a \in A$ принадлежит ровно одной из этих комбинаций, приведенной в таблицах.

$$\forall a \in A: (a \in Avp \wedge a \in Ads \wedge a \in Ats) \vee (a \in Avp \wedge a \in Ads \wedge a \in Ants) \vee (a \in Avp \wedge a \in Ands \wedge a \in Ats) \vee (a \in Avp \wedge a \in Ands \wedge a \in Ants) \vee (a \in Anvp \wedge a \in Ads \wedge a \in Ats) \vee (a \in Anvp \wedge a \in Ads \wedge a \in Ants) \vee (a \in Anvp \wedge a \in Ands \wedge a \in Ats) \vee (a \in Anvp \wedge a \in Ands \wedge a \in Ants). \quad (2.61)$$

Примеры определения принадлежности действий:

1. Разведка (Reconnaissance): ARecon=Anвр∩Ands∩Ants∩(O∪Э).
2. Эксфильтрация (Exfiltration): AExfil=Авр∩Ads∩Ats∩Э.

Использование приведенных формальных ограничений позволяет:

1. Четко определить условия доступности действия: на основе инфраструктурных ограничений и режима работы системы ИИ.
2. Сузить пространство поиска возможных атак: отсеять действия, которые не соответствуют условиям доступа.

3. Интегрировать информацию из MITRE ATLAS в математическую модель: сопоставить тактики и техники с конкретными множествами действий и условиями доступа.

Эта детализация обеспечивает более строгий и точный способ представления знаний об атаках на системы ИИ, что позволяет использовать формальные методы для анализа рисков и разработки стратегий защиты.

Модель нарушителя, интегрированная в процесс моделирования, может быть представлена в виде кортежа $H = \{Kn_0, H_0, Kв, G, A\}$, где:

1. Kn_0 (начальные знания) — это множество исходных знаний нарушителя об атакуемой системе. Оно может быть определено как $Kn_0 \in \{white, grey, black\}$, что соответствует полному, частичному или нулевому знанию о системе.

2. Множество логических и инфраструктурных компонентов, к которым нарушитель имеет первоначальный доступ, H_0 (начальное расположение и доступ) определяет точку входа и начальное состояние в графе марковского процесса принятия решений (МППР). Для внешнего нарушителя подразумевается доступ только к публичным интерфейсам и API из внешних сетей, а для внутреннего — доступ к внутренним компонентам, таким как датасеты или вычислительные модели. При этом следует учитывать возможность для внешнего нарушителя получить преимущества внутреннего после проникновения во внутреннюю инфраструктуру систем ИИ. В целом, множества Kn_0 и H_0 формируют начальное состояние S1 графа состояний и действий.

3. Квалификационные характеристики $Kв = \langle Kn, CVSS_{attakER} \rangle$ определяют способность эксплуатации уязвимостей. В этом случае Kn отражает знания о методах атак, а $CVSS_{attakER}$ — практическую способность реализации атак определённого уровня сложности, что непосредственно влияет на параметры вероятностей переходов P(s,a,s') и функции вознаграждения R(s,a,s') в модели МППР.

4. Множество G (цели) — это цели атаки, достижение которых определяет успех нарушителя. Цели ставятся в соответствии с состояниями графа, в том числе, с тактиками из методик MITRE ATLAS и ФСТЭК. Множество $G = \{g_1, g_2, ..., g_n\}$

предполагает, что каждая цель $g_i = \{S_i, V(S_i)\}$, где S_i — это целевое состояние, а $V(S_i)$ — ценность достижения этого состояния для нарушителя, что определяет приоритет цели.

5. Множество A (доступные действия) — это подмножество всех возможных атакующих действий (техник MITRE ATLAS), которые нарушитель может выполнить исходя из своих параметров (Kn_0, H_0, Кв). Таким образом, $A \subseteq \{(s_i, s_j) \mid s_i, s_j \in S, a \in \{1,\ldots,n\}\}$. Это множество определяется в соответствии с правилами, формирующими таблицы, которые позволяют определить специфику доступности компонентов систем на основе требуемых условий для реализации атакующих действий (Таблицы 2.7, 2.8, 2.9, 2.10).

Выводы к главе 2

В ходе анализа была определена специфика моделирования, были выделены два возможных сценария функционирования модели. Первый сценарий предполагает, что злоумышленник действует в соответствии с известной аудитору логикой, его целью являются наиболее уязвимые и незащищённые компоненты системы искусственного интеллекта. Второй сценарий предполагает, что злоумышленник может действовать непредсказуемо или руководствоваться иной логикой, что в некоторых случаях может привести к успешной атаке.

Функциональные возможности модели включают набор действий, которые она может выполнять в различных состояниях. Эти действия учитывают специфику архитектуры системы искусственного интеллекта и процессы, происходящие внутри неё. Также были определены тактики, которые могут быть использованы в случае необходимости. Эти тактики основаны на том, что злоумышленник, достигая своих целей, получает новые возможности, реализуемые через определённые действия. Действия можно разделить на три категории:

1. Категория нелегальных действий. Это попытки ввести систему в заблуждение, скомпрометировать ее для достижения целей злоумышленника.

2. Категория легальных действий. Это действия, которые соответствуют установленным правилам и условиям безопасности ИИ.

3. Действия, направленные на уже полученные состояния атаки.

Кроме того, действия могут быть разделены на группы в зависимости от техник, используемых для эксплуатации уязвимостей. Эти две классификации могут пересекаться.

Уязвимости открывают возможности для реализации действий, но их реализация зависит от вероятности проявления действия. Вероятности могут быть определены экспертным методом, вероятностными методами или с использованием равномерного распределения, если у эксперта нет статистических данных. Были выявлены особенности определения вознаграждения за переход в последующие состояния атаки. Награды могут быть определены в двух режимах. Аспект безопасности ИИ, то есть присутствие функций систем защиты, которые могут противодействовать порядку эксплуатации уязвимости, учитывается как функция затухания. Функция затухания определена в двух видах: экспоненциальном и линейном. Была определена специфика сопряжения тактик и техник и инфраструктуры ИИ. При этом выделены критерии, определяющие отбор тактик или, при наличии требований аудита, набор техник, связанных с большей детализацией действий злоумышленника.

Глава 3. Модели определения последовательностей атакующих воздействий на системы искусственного интеллекта как подсистемы ИС

3.1 Определение правил моделирования

В приводимых моделях используется метод итераций по значениям для формирования и оценки функций ценности (полезности) состояний как основной для МППР. Допускается использование дополнительных методов: итерации по стратегиям могут использоваться в ситуациях, когда необходимо быстро находить оптимальные стратегии действий; Q-обучение используется в тех случаях, когда присутствует неточное описание ПИИ и доступны вычислительные ресурсы. При аудите задается тип злоумышленника – внешний. Каждая функция ценности обновляется на основе текущих значений и максимизации ожидаемого вознаграждения с учетом вероятностей переходов. Метод фокусируется на оценке ценности состояний или пар «состояние-действие». Функции ценности позволяют вычислить ожидаемую полезность (или ценность) для каждого состояния или действия, что дает возможность определения оптимальной «стратегии». В целом преимущества метода итераций по значениям в контексте построения моделей атакующих воздействий следующие:

1. Методы по значениям часто проще в реализации, особенно для задач с дискретными состояниями и действиями.

2. Алгоритмы имеют теоретические гарантии сходимости к оптимальной стратегии.

3. В задачах с ограниченными пространствами состояний и действий методы по значениям могут быть очень эффективными.

При моделировании используются уровни абстракции. Они определяются как степень детализации состояний атакующих воздействий. Причина введения уровней абстракции – сложность описания сценария атаки при наличии множества учитываемых параметров и компонентов. Выявляются три уровня абстракции:

1.1 Состояния определяются как наборы состояний, позволяющих реализовать множества действий, связанных с общей логикой функционирования систем ИИ, а также ПИИ, в период атакующих воздействий. Данную модель можно использовать для следующего:

- общего описания системы в период атаки без уточнения специфики конкретных действий злоумышленника;

- исследования конкретного состояния.

1.2. Состояния определяются как наборы состояний, позволяющих реализовать множества действий, связанных с инфраструктурой и логикой функционирования ИИ в период атакующих воздействий. Данная модель может использоваться для общего описания системы в период атаки с уточнением последовательности специфики действий злоумышленника, связанных с этапами реализации атакующих воздействий.

1.3. Состояния определяются как наборы событий, позволяющих реализовать множества действий, приведенных в тактиках MITRE ATLAS, связанных с инфраструктурой и логикой функционирования ИИ в период атакующих воздействий. Данная модель может использоваться для частного и подробного описания системы в период атаки с уточнением последовательности специфики действий злоумышленника, связанных с порядком атакующих воздействий.

Таким образом, формируются три типа моделей:

1. Типовая 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.

2. Упрощенные модели с использованием действий из классификации тактик MITRE ATLAS в режимах аудита on-line и off-line, в которой тактики объединены на основе сходства функционального назначения техник в несколько классов. Предлагается использовать модели, зависящие от структуры и способа функционирования ПИИ. Категоризация действий (совокупности техник) производится по типу функционирования злоумышленника с учетом

необходимости обязательного взаимодействия во время атаки. Альтернативная категоризация производится по типу функционирования.

3. Подробные модели с использованием состояний и действий, сформированных на основе тактик и техник MITRE ATLAS, в режимах аудита on-line и off-line (тактики рассматриваются отдельно, каждая с совокупностью всех принадлежащих ей техник).

Связи между состояниями ограничены спецификой устройства и функционирования архитектуры системы ИИ. Также следует учитывать специфику доступа к модели злоумышленника. Следует рассматривать доступ, контролируемый злоумышленником (контролируемое взаимодействие) к функциям и компонентам ПИИ, с учетом возможности осуществления тактик (техник) (Таблица 2.10). Выделяются следующие виды доступа: доступ к датасетам; взаимодействие с вычислительной моделью ПИИ; взаимодействие с инфраструктурными элементами вычислительной модели ПИИ. Также определяются режимы работы ПИИ: режим обучения; режим эксплуатации; общий режим совместного обучения и эксплуатации.

Таким образом, можно выделить и режимы описания действий в моделях с учетом уровня детализации:

1. Первый режим. Учитываются: нелегальное взаимодействие (D); легальное взаимодействие (передача легитимных данных) (C); сброс (R).

2. Второй режим. Учитываются действия по эксплуатации уязвимостей, каждое из которых разделяется на: нелегальное взаимодействие (D); легальное взаимодействие (передача легитимных данных) (C); отдельно учитывается сброс (R); уточняющие действия (по тактикам).

3. Третий режим. Учитываются действия по эксплуатации уязвимостей, каждое из которых разделяется на: нелегальное взаимодействие (D); легальное взаимодействие (передача легитимных данных) (C); отдельно учитывается сброс (R); уточняющие действия по тактикам и техникам.

Формируются последовательности атакующих воздействий с учетом стратегии вознаграждений. Параметры вознаграждений: маркировка по топологии

(политика выбора пути заключается в определении (максимизации) набора лучших уязвимостей и поиска связи между ними (связанные уязвимости)). При визуализации графа должны быть учтены все возможные состояния (все наборы тактик) (Рисунок 3.1).

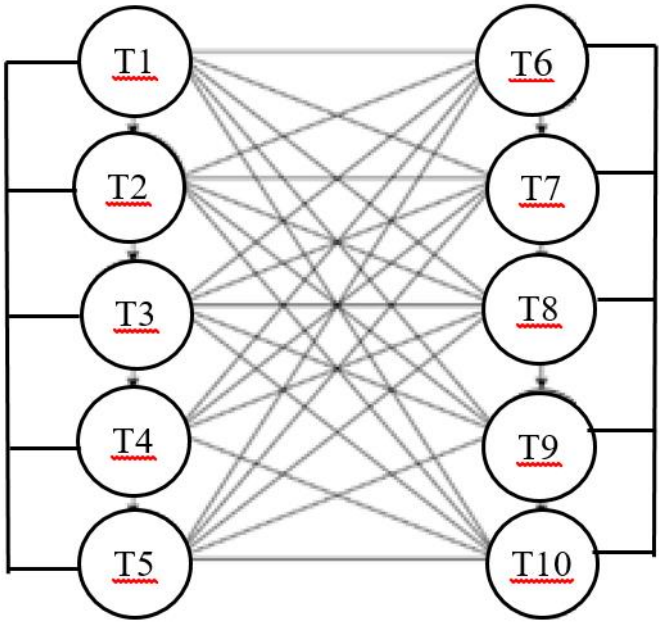


Рисунок 3.1 - Последовательность атакующих воздействий

Допущения, которые необходимо учесть при моделировании в процессе аудита, связаны с предположением защищающейся стороны о действиях злоумышленника, а, точнее, со следующим:

- требуется определить целевые состояния важные для защищаемой системы, и подразумевается, что к ним стремится злоумышленник, тип нарушителя - внешний;

- для моделей в режиме off-line требуется логичность действий злоумышленников;

- для моделей off-line и on-line требуется полная информация о вознаграждениях и о затратах, как части вознаграждения, следует учитывать фиксированный характер затрат и вознаграждений.

Также следует учитывать следующие особенности моделирования:

1. Учитывая логику развития атаки, которая соответствует описанию MITRE, можно предположить то, что с целью завоевать доверие модели ПИИ злоумышленник имеет возможность сначала передавать легитимные данные. Например, атакующий выбирает записи из обучающей выборки, которые модель ПИИ правильно классифицирует, и далее добавляет возмущения к ним, приводящие в итоге к неправильному классифицированию данных в последующем. Также злоумышленник должен получить достаточное вознаграждение *R* за то, что смог завоевать доверие атакуемых систем. При этом начальные значения вознаграждения отражают выгоды и потери, свойственные состоянию, при котором нападающий не достиг легального взаимодействия с атакуемым объектом.

2. Ядром описания модели на основе МППР является описание функции полезности (ценности), на основе которой производится поиск оптимальной стратегии. На основе ее определяются стратегии поведения нападающей стороны.

3. После итераций по значениям используются формулы определения оптимальных действий, которые могут быть различными и составлять целые множества D, C, R (для каждого действия приведены формулы (3.1 – 3.3)):

a. Нелегальное взаимодействие (*D*).

$$V_D^*(s) = \max_{a \in D} \sum_{s' \in S} P(s,a,s') \ [\mathrm{R}\,(s,a,s') + \gamma V^*(s')]. \qquad (3.1)$$

b. Сброс (*C*).

$$V_C^*(s) = \max_{a \in C} \sum_{s' \in S} P(s,a,s') \ [\mathrm{R}\,(s,a,s') + \gamma V^*(s')]. \qquad (3.2)$$

c. Легальное взаимодействие (*R*).

$$V_R^*(s) = \max_{a \in R} \sum_{s' \in Sprev} P(s,a,s') \ [\mathrm{R}\,(s,a,s') + \gamma V^*(s')]. \qquad (3.3)$$

В формулах допускается использовать общее обозначение *S* вместо конкретных состояний (например, *S*P). Это позволяет применять формулы к любому состоянию системы. Вместо конкретных действий (например, *D, R* или *C*) допускается использовать обобщенные обозначения для действий (a_i), что делает формулы более универсальными.

4. Важно отметить, что поскольку состояние интерпретируется как получение злоумышленником новых возможностей для осуществления дальнейших компрометирующих действий, то множество действий будет включать в свой состав те действия, которые разрешаются состоянием атаки (ее этапом на пути продвижения к цели с учетом того, что этапы интерпретируются по методике MITRE ATLAS или при подготовке сценария атаки в рамках аудита ИБ по нормативным документам РФ (по Методике ФСТЭК)). В данном случае под действиями подразумеваются техники тактик, которые входят во множества действий, включающих классы*: D, R* или *C*.

5. Переходные вероятности действий имеют равное распределение в зависимости от допустимых действий по отношению к следующим состояниям. Допустимы методики вычисления вероятностей (при наличии статистических данных) на основе расчета уязвимостей, вознаграждений, методы машинного обучения, используемые при анализе событий безопасности.

6. Количество этапов работы модели зависит от периода, исследуемого при аудите ПИИ (конечный горизонт планирования). Также можно исследовать систему с бесконечным горизонтом планирования, тогда период аудита учитываться не будет.

7. При необходимости преобразовать Value Iteration в Policy Iteration, требуется заменить немедленное вычисление максимума на хранение явной политики и разделить процесс вычисления на две фазы: оценку текущей политики (обновление V(s) в соответствии с π(s)) и её последующее улучшение (выбор нового действия, максимизирующего ценность). Для преобразования в Q-Learning требуется перейти от функций ценности состояний V(s) к функциям действия-ценности Q(s, a), а затем замените полное математическое ожидание модели на

обновление по одной выборке из окружающей среды, используя правило временной разницы.

3.2 Общие модели на основе МППР для анализа специфики атаки в режиме on-line и off-line

Общие модели атак на системы искусственного интеллекта и их подсистемы в on-line и off-line позволяют исследовать общее состояние безопасности ПИИ или одно из состояний более сложной модели. Эти модели не предназначены для детального описания тактик и методов атак, используемых злоумышленниками при атаке на ПИИ. Однако она позволяет получить общее представление о том, как злоумышленник может взаимодействовать с системой.

Ограничения модели при ее применении следующие: модель используется без детализации атакующих действий по методикам (техникам), без сопоставления действий, уязвимостей, локальных идентификаторов, при этом состояния абстрактно описывают этапы атаки и, соответственно, те состояния, в которых система ИИ пребывает. Используются четыре состояния, с которыми злоумышленник может столкнуться при взаимодействии с ПИИ:

– контролируемое взаимодействие (*C*);

– легальное взаимодействие (*L*);

– доверенное взаимодействие (*T*);

– блокировка (*B*).

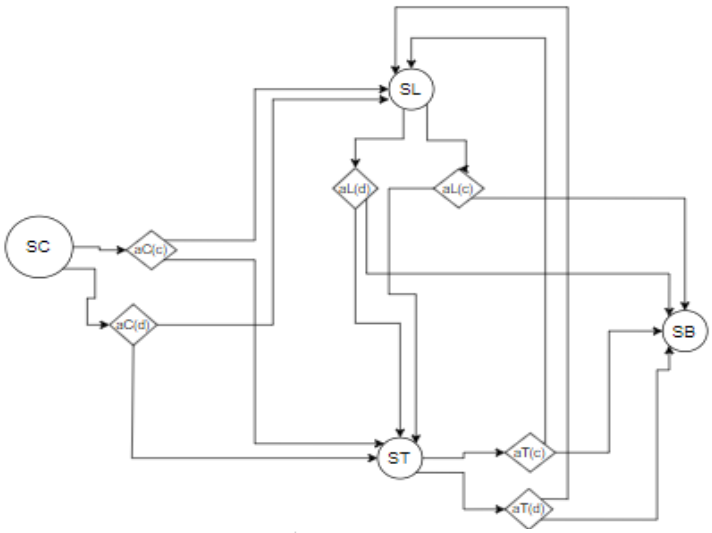
Таким образом, с учётом ограничений модели целесообразно применять в этом случае первый режим описания действий модели. Граф состояний модели приведен на рисунке 3.2. Специфика действий при заданном подходе к моделированию предполагает, что при атаке злоумышленник может выбрать действие a_i из множества доступных действий A ($a_i \in A$) с учетом выбранной им стратегии: в режимах on-line(б); $D \cup C \cup R \subseteq A$; в режимах off-line (а) $D \cup C \subseteq A$. Таким образом, в модели используются следующие обозначения действий:

1. Нелегальное взаимодействие (*D*). $P(S_i, D, S')$ - вероятность перехода из состояния S_i в состояние S' при выполнении действия *D*; $R(S_i, D, S')$ - награда за переход из состояния S_i в состояние S' при выполнении действия *D*.

2. Легальное взаимодействие (*C*) (аналогично предыдущему случаю для легального взаимодействия (действия)). $P(S_i, C, S')$ - вероятность перехода из состояния S_i в состояние S' при выполнении действия *C*; $R(S_i, C, S')$ - награда за переход из состояния *Si* в состояние S' при выполнении действия *C*;

3. Сброс (*R*). $P(S_i, R, S')$ - вероятность сброса из состояния S_i в состояние S'; $R(S_i, R, S')$ - награда за сброс из состояния S_i в состояние S'.

а)



б)

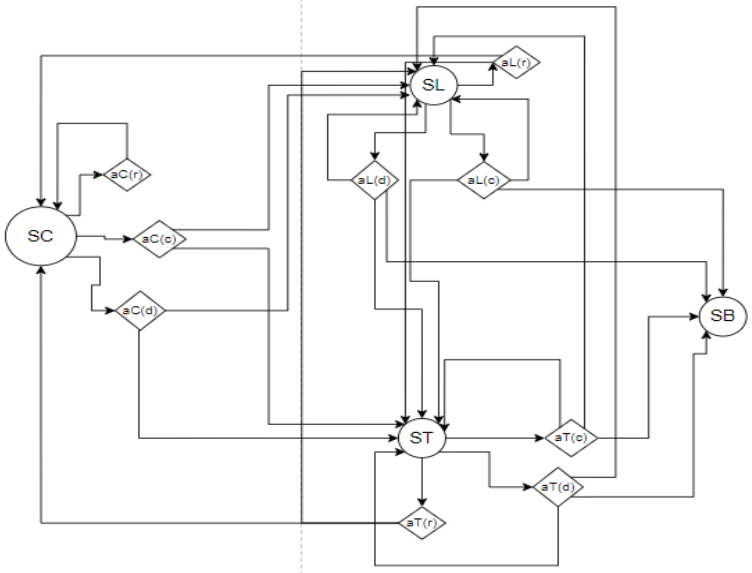


Рисунок 3.2 – Графы состояний общих моделей МППР для атак в режиме off-line (а) и on-line (б)

При выборе стратегии атаки в рамках моделирования с использованием МППР, также учитываются параметры «вознаграждения» и «затрат». Нейтральное *SC* состояние является начальным в развертываемом векторе атаки. Действия, связанные с осуществлением атаки методом навязывания ложных обучающих данных, будут зафиксированы ПИИ, что вызовет ответную реакцию.

Вознаграждение формируются на основе следующего: отклонения в вычислительной модели; уязвимости датасета. Дополнительный способ (необязателен), предполагает использование метрик CVSS, определяемых экспертным образом (экспертная группа определяет уязвимость модели или датасета или системных компонентов по параметрам CVSS). Злоумышленник стремится получить доступ к наиболее опасным уязвимостям, чтобы получить доступ к следующим состояниям, позволяющим реализовать новые атакующие воздействия и продвинуться к целевому состоянию.

При определении величины вознаграждений целесообразно учитывать специфику определения уязвимостей по методологии CVSS, то есть необходимо согласовать размерность (шкалу измерений уязвимостей). Таким образом, в процессе моделирования, при определении значений ожидаемых вознаграждений *R,* нужно учитывать следующие особенности:

1. Стоимость введения в заблуждение модели ПИИ (является частью вознаграждения, используется для исследования динамики изменений состояния безопасности системы при повышении или понижении величины вознаграждений и, соответственно, стоимости реализации нападения) может оказаться намного выше, чем стоимость легального взаимодействия.

2. Выгоды, получаемые при переходе между состояниями, зависят от исходных и целевых вершин графа атаки, представляющего собой последовательность атакующих воздействий.

3. Вознаграждений R не будет в случае, когда злоумышленник сначала отправляет ложные данные, а затем его сообщения блокируют независимо от каких-либо действий.

Относительный размер вознаграждений R в этом случае, полученных злоумышленником за каждое изменение (переход в иное состояние), независимо от возможных действий отвечает следующим закономерностям (для моделирования в режиме off-line): R($C-T$)>R($L-T$)> R($L-C$)> R($C-C$)≥ R(C-L) ≥ R($T-T$)> R($T-C$) ≥ R($T-L$)>R($C-B$). Эти закономерности подразумевают, что логика злоумышленника базируется на целеполагании, которое можно свести к следующему: присутствие злоумышленника в доверенном состоянии при выполнении любых его действий с целью избежать состояния блокировки.

При этом следует учитывать специфику формирования затрат злоумышленника как части награды. Она зависит от количества вложений в процедуры подготовки и осуществления компрометации ПИИ. В случае, когда используется CVSS, зависимость усилий нарушителя можно сопоставить с одной из метрик, отражающих сложность эксплуатации уязвимости. В иных случаях можно воспользоваться рекомендациями Методики ФСТЭК (в части описания нарушителя) [1].

Следует учесть, что для моделей МППР важны итоговые значения V^*, поэтому в данном случае целесообразно отнести приведенные закономерности к описанию регламента специфики последовательности перехода от одного состояния к другому, который отражает приоритеты злоумышленника при атаке в режиме off-line. В случае атаки в режиме on-line злоумышленник способен сбрасывать достигнутый успех (реализовывать действие сброса (R)) (менять целеполагание), и поскольку в этом случае возможны повторы состояний ранее достигнутых, специфика выбора приоритета перехода в большей степени будет отражать стремление достичь целевого состояния злоумышленника с точки зрения защищающейся стороны, которая самостоятельно определяет возможные приоритеты злоумышленника.

При моделировании начальные значения для всех состояний устанавливаются равными нулю. Далее для каждого состояния вычисляются новые значения. Этот процесс повторяется до тех пор, пока значения не достигнут равновесного состояния и не изменятся. Кроме того, учитывается максимальное

количество повторений (например, 1000), чтобы избежать попадания в бесконечный цикл, когда значения меняются очень незначительно.

Логика компрометации (воздействие на вычислительную модель с целью вызова ошибки в работе модели ИИ) злоумышленником ориентирована на достижение целевого состояния (*T*), при котором он может реализовать компрометирующие действия при наименьших затратах, величина которых входит в объем награды за действие и не приводит к блокировке (состояние *SB*). Учитывая логику компрометации определяется ее специфика:

1. Суммирование вероятностей перехода из одного состояния в другое для каждого действия должно быть равно 1. В случае получения суммирования, не равного единице, требуется стандартизировать значения, чтобы гарантировать, что они попадают в надлежащий интервал [0,1].

2. При выполнении совместного действия вероятность перехода в доверенное состояние должна быть больше, чем вероятность перехода в компрометируемое (оспариваемое) состояние. Кроме того, вероятность перехода в оспариваемое состояние должна быть больше, чем вероятность перехода в заблокированное состояние.

3. Вероятность перехода в заблокированное состояние *B* должна быть больше, чем вероятность перехода в состояние *C* (контролируемое взаимодействие). При выполнении обманного действия вероятность перехода в оспариваемое состояние должна быть больше, чем вероятность перехода в доверенное состояние.

Ключевым фактором в построении модели является определение величин ценности состояний, что позволяет определить наилучшую последовательность состояний атаки с точки зрения злоумышленника. Определение значений для каждого состояния с использованием функции полезности модели на основе МППР для общей атаки в режиме on-line (Рисунок 3.2 (б)) на ПИИ приведено в выражениях ниже (3.4 – 3.9).

$$V_{i+1}^{*}(s=ST)=\max_{a\in A}\sum_{s'\in S}P(ST,a,s')\ [\mathrm{R}\ (ST,a,s')+\gamma V_i^{*}(s')]\ . \qquad (3.4)$$

$$V_{i+1}^{*}\ (s=ST)=max\ \begin{cases} P(ST,D,SL)[\mathrm{R}(ST,D,SL)+\gamma V_i^{*}(SL)]+ \\ P(ST,D,ST)[\mathrm{R}(ST,D,ST)+\gamma V_i^{*}(ST)]+ \\ P(ST,D,SB)[\mathrm{R}(ST,D,SB)+\gamma V_i^{*}(SB)] \\ ---------- \\ P(ST,C,SL)[\mathrm{R}(ST,C,SL)+\gamma V_i^{*}(SL)]+ \\ P(ST,C,ST)[\mathrm{R}(ST,C,ST)+\gamma V_i^{*}(ST)]+ \\ P(ST,C,SB)[\mathrm{R}(ST,C,SB)+\gamma V_i^{*}(SB)] \\ ---------- \\ P(ST,R,SC)[\mathrm{R}(ST,R,CS)+\gamma V_i^{*}(SC)] \end{cases} . \qquad (3.5)$$

$$V_{i+1}^{*}(s=SL)=\max_{a\in A}\sum_{s'\in S}P(SL,a,s')\ [\mathrm{R}\ (SL,a,s')+\gamma V_i^{*}(s')] \qquad . \qquad (3.6)$$

$$V_{i+1}^{*}\ (s=SL)=max\ \begin{cases} P(SL,D,SL)[\mathrm{R}(SL,D,SL)+\gamma V_i^{*}(SL)]+ \\ P(SL,D,SB)[\mathrm{R}(SL,D,BS)+\gamma V_i^{*}(SB)]+ \\ P(SL,D,ST)[\mathrm{R}(SL,D,ST)+\gamma V_i^{*}(ST)] \\ ---------- \\ P(SL,C,SL)[\mathrm{R}(SL,C,SL)+\gamma V_i^{*}(SL)]+ \\ P(SL,C,SB)[\mathrm{R}(SL,C,SB)+\gamma V_i^{*}(SB)]+ \\ P(SL,C,ST)[\mathrm{R}(SL,C,ST)+\gamma V_i^{*}(ST)] \\ ---------- \\ P(SL,R,SC)[\mathrm{R}(SL,R,SC)+\gamma V_i^{*}(SC)] \end{cases} . \qquad (3.7)$$

$$V_{i+1}^{*}(s=SC)=\max_{a\in A}\sum_{s'\in S}P(SC,a,s')\ [R\ (SC,a,s')+\gamma V_i^{*}(s')] \quad . \qquad (3.8)$$

$$V_{i+1}^{*}(s=SC)=max\ \begin{cases} P(SC,D,SL)[\mathrm{R}(SC,D,SL)+\gamma V_i^{*}(SL)]+ \\ P(SC,D,ST)[\mathrm{R}(SC,D,ST)+\gamma V_i^{*}(ST)]+ \\ (SC,D,SC)[\mathrm{R}(SC,D,SC)+\gamma V_i^{*}(SC)] \\ ---------- \\ P(SC,C,SL)[\mathrm{R}(SC,C,SL)+\gamma V_i^{*}(SL)]+ \\ P(SC,C,SC)[\mathrm{R}(SC,C,SC)+\gamma V_i^{*}(SC)]+ \\ P(SC,C,ST)[\mathrm{R}(SC,C,ST)+\gamma V_i^{*}(ST)] \\ ---------- \\ P(SL,R,SC)[\mathrm{R}(SL,R,SC)+\gamma V_i^{*}(SC)] \end{cases} . \qquad (3.9)$$

Присутствие обратных дуг, обозначающих действие R, в данной модели обусловлено характером построения сценария при пассивном аудите (определяется

наилучший сценарий атаки злоумышленника при учёте того, что злоумышленнику известно все об атакуемой системе, средства защиты предустановлены, а система при этом изменяется только под действиями злоумышленника). В этом случае откат в предыдущее состояние означает отсутствие успеха в выбранной стратегии поведения злоумышленника, и, следовательно, возникает потребность в смене вектора атаки.

Граф состояний модели off-line приведен на рисунке 3.2 (а). В этом случае отсутствуют переходы в переходы через действия, направленные на источник приводимых действий (3.10 – 3.15). Такое решение обусловлено тем, что злоумышленник выбирает лучшие действия, а возвращение в источник действий (состояние) в режиме off-line указывает на отсутствие лучшего решения у предполагаемого злоумышленника (аудитор следует наилучшему выбору злоумышленника чтобы выявить проблемы безопасности ПИИ).

$$V_{i+1}^{*}(s = ST) = \max_{a \in A} \sum_{s' \in S} P(ST, a, s') \ [\mathrm{R}\ (ST, a, s') + \gamma V_i^{*}(s')]. \quad (3.10)$$

$$V_{i+1}^{*}\ (s = ST) = max \begin{cases} P(ST, D, ST)[\mathrm{R}(ST, D, ST) + \gamma V_i^{*}(ST)] + \\ P(SL, D, SB)[\mathrm{R}(SL, D, BS) + \gamma V_i^{*}(SB)] + \\ - - - - - - - \\ P(ST, C, ST)[\mathrm{R}(ST, C, ST) + \gamma V_i^{*}(ST)] + \\ P(ST, C, SB)[\mathrm{R}(ST, C, SB) + \gamma V_i^{*}(SB)] \end{cases}. \quad (3.11)$$

$$V_{i+1}^{*}(s = SL) = \max_{a \in A} \sum_{s' \in S} P(SL, a, s') \ [\mathrm{R}\ (SL, a, s') + \gamma V_i^{*}(s')]\ . \quad (3.12)$$

$$V_{i+1}^{*}\ (s = SL) = max \begin{cases} P(SL, D, SB)[\mathrm{R}(SL, D, BS) + \gamma V_i^{*}(SB)] + \\ P(SL, D, ST)[\mathrm{R}(SL, D, ST) + \gamma V_i^{*}(ST)] \\ - - - - - - - \\ P(SL, C, SB)[\mathrm{R}(SL, C, SB) + \gamma V_i^{*}(SB)] + \\ P(SL, C, ST)[\mathrm{R}(SL, C, ST) + \gamma V_i^{*}(ST)] \end{cases}. \quad (3.13)$$

$$V_{i+1}^{*}(s = SC) = \underset{a \in A}{max} \sum_{s' \in S} P(SC, a, s') \ [\mathrm{R}\ (SC, a, s') + \gamma V_i^{*}(s')]. \quad (3.14)$$

$$V_{i+1}^{*}(s = SC) = max \begin{cases} P(SC, D, SL)[\mathrm{R}(SC, D, SL) + \gamma V_i^{*}(SL)] + \\ P(SC, D, SB)[\mathrm{R}(SC, D, SB) + \gamma V_i^{*}(SC)] + \\ - - - - - - - \\ P(SC, C, SL)[\mathrm{R}(SC, C, SL) + \gamma V_i^{*}(SL)] + \\ P(SC, C, SC)[\mathrm{R}(SC, C, SC) + \gamma V_i^{*}(SC)] + \end{cases} . \quad (3.15)$$

При моделировании начальные значения для всех состояний устанавливаются равными нулю. Далее для каждого состояния вычисляются новые значения. Этот процесс повторяется до тех пор, пока значения не достигнут равновесного состояния и не изменятся.

Модель используется для анализа:

- отдельного состояния (тактических возможностей злоумышленника) сценария атаки;

- для анализа безопасности (тактических возможностей злоумышленника) системы в целом, без учета ее компонентного состава (система описывается приведенными состояниями), при этом допустимо отдельно рассматривать уровень логической организации ИИ и отдельно технической.

Оптимальная стратегия, использующая алгоритм итерации значений, позволяет определить: оптимальное действие, когда злоумышленник находится в состоянии контролируемое взаимодействие (*C*), когда злоумышленник находится в состоянии доверенное взаимодействие (*T*), также атакует систему, когда злоумышленник находится в состоянии легальное взаимодействие (*L*).

Следуя этой оптимальной стратегии, злоумышленник избегает блокировки, используя политику МППР, оптимальную для атакующего. Для применения анализа состояния и выявления его специфики следует выполнить следующее [134-135]:

1. Определить базовую выгоду от действий по *C* (например, легальное действие, производимое совместно с системой ИИ (*coop*)) и *D* ((например, нелегальное обманное действие (*decep*)).

2. Выполнить обучение с бесплатной стоимостью *C (coop)* и *D (decep)* (например, при T1: D, T2: D, T6: C).

3. Определить максимальную стоимость *C (coop)* и *D (decep)*.

4. Поочерёдно зафиксировать цену *C* и *D* на значениях 0%, 10%, 50%, 100%, для каждой зафиксированной стоимости, построить зависимость ценности состояния от цены *C* и *D* соответственно [134 -135].

5. Формируются соответствующие графики (например, для состояния T2 (Рисунок 3.3)).

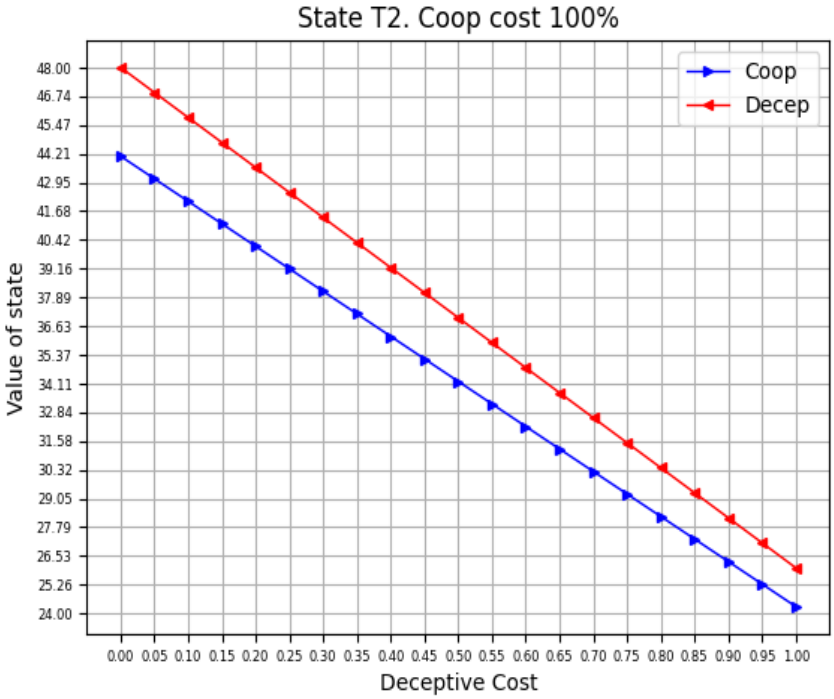


Рисунок 3.3 – Динамика изменений значений состояния T2 в зависимости от стоимости действий (части вознаграждений)

6. Для каждого состояния и зафиксированной цены соответствующего действия формируются графики зависимости ценности состояния при выполнении оптимальной политики (Рисунок 3.4).

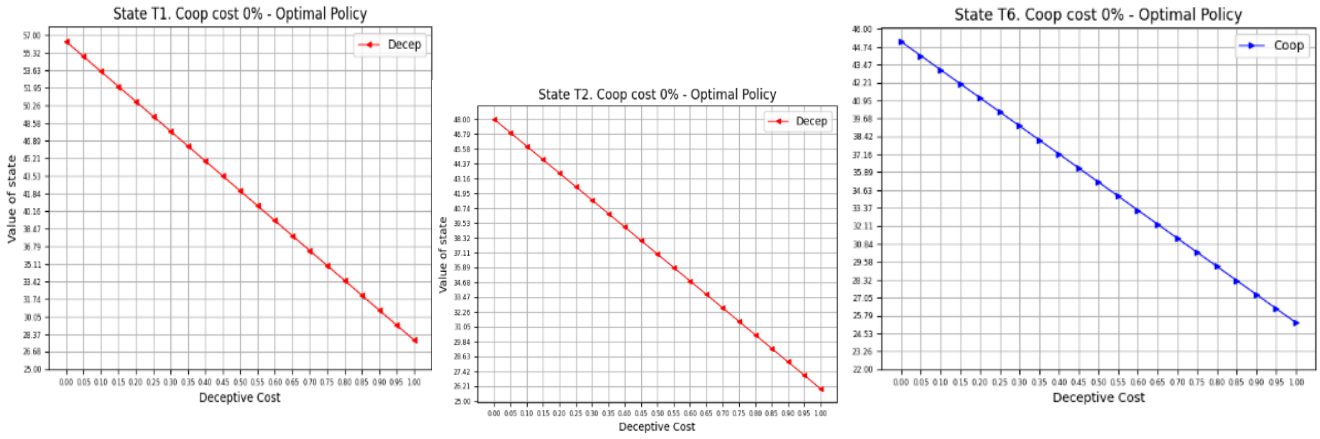


Рисунок 3.4 – Динамика изменений значений состояний зависимости от стоимости действий (части вознаграждений)

Другой пример, где максимальные цены, приведен на рисунке 3.5 (пример интуитивного объяснения – нарушителю дорого стоит взаимодействовать с моделью (например, ограничение количества запросов в минуту)).

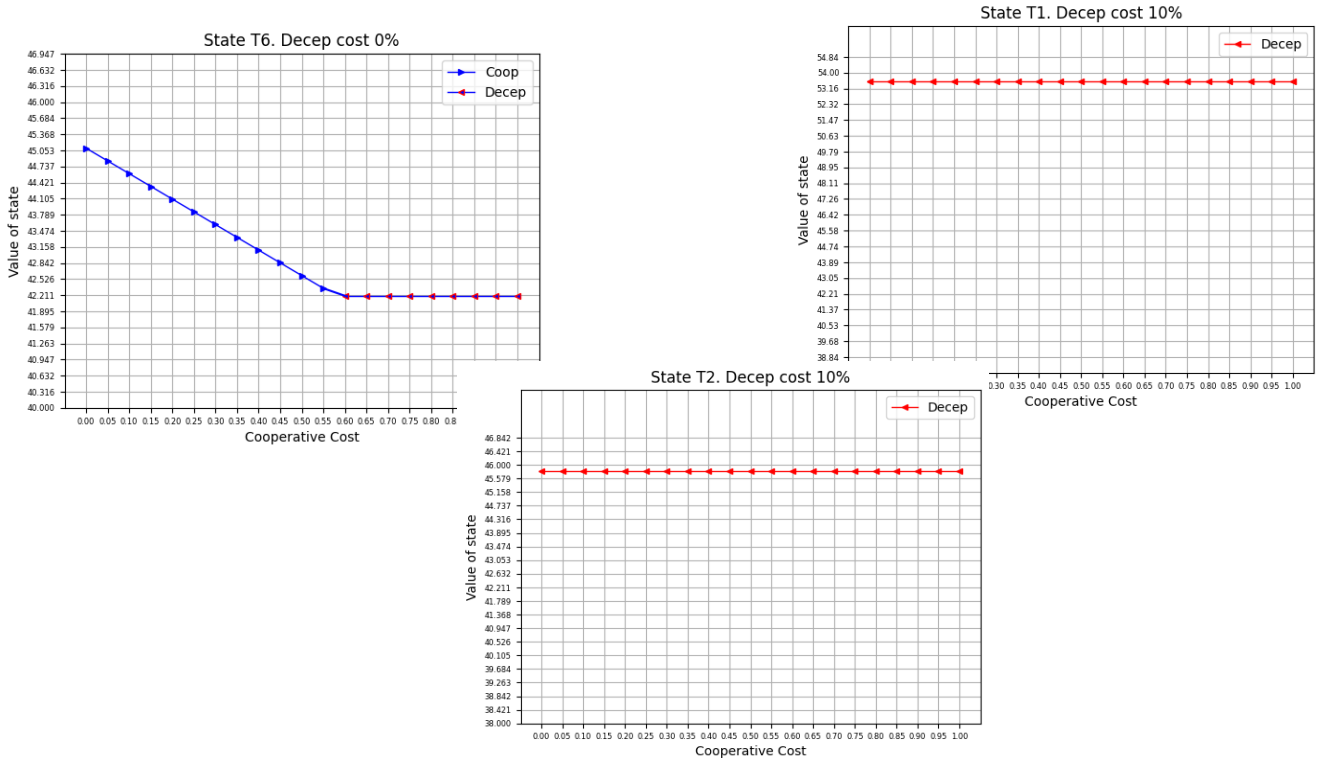


Рисунок 3.5 – Динамика изменений значений состояний (T1, T6, T2) зависимости от стоимости действий (части вознаграждений)

В примерах взяты пробные значения наград с учетом сопоставления метрической системы CVSS и уязвимости датасета.

3.3 Модели на основе упрощенной классификации тактик

В этом случае приводятся модели на основе упрощенной классификации тактик MITRE, в которой тактики объединены на основе сходства функционального назначения техник в несколько классов. Следует учитывать, что модели, описывающие последовательности предполагаемой атаки, которые формируются при аудите информационных систем, создаются с учетом поиска наилучших последовательностей действий злоумышленника в контексте взаимодействия двух сторон: нарушителя и защищающейся стороны [135].

Первый способ объединения тактик и их техник приведен на рисунке 3.6. При этом возникают и объединенные состояния атаки, которые означают появление возможности осуществления заданных действий. Это упрощенная классификация MITRE ATLAS (допустимы упрощения и для других методик), в которой действия объединены на основе сходства функционального назначения выбранных тактик в несколько классов:

1. Разведка предполагает выявление слабостей (уязвимостей к воздействию на системы ИИ), проводится на разных этапах атаки: на начальном и промежуточных этапах при продвижении в последовательности компрометирующих действий.

2. Сопровождение включает в свой состав подготовку средств эксплуатации уязвимостей и условий для их успешной эксплуатации. Также может осуществляться на начальном и промежуточных этапах.

3. Компрометация содержит действия, направленные на изменение порядка функционирования ПИИ, в том числе ее вычислительной модели.

4. Достижение – это заключительная фаза, наступление которой означает достигнут успех атаки и злоумышленник нанес ущерб системе.

Рисунок 3.6 – Пример упрощения методологии MITRE ATLAS

Второй способ объединения тактик и их техник (действий) приведен в таблице 3.1 (при этом следует учитывать акуальность базы тактик). Здесь также используются объединенные состояния (этапы) атаки. Данный тип классифицирования атакующих воздействий отражает специфику построения ПИИ, в которой связи между состояниями, и, соответственно, действия зависят от конфигурирования системных компонент во время атаки, возможности взаимодействовать с интерфейсами ИС при обращении к вычислительной модели. Множества контроля взаимодействия следующие:

- без обязательного контроля (Ancon): Ancon={a∈A|без контроля модели;}
- условный контроль (Amid): Amid={a∈A|условный контроль};
- обязательный контроль (Acon): Acon={a∈A|требуется контроль модели}.

При построении вычислительной модели действия между состояниями системы, интерпретируемыми по методике MITRE ATLAS, определяются по следующему правилу: дуга действия злоумышленника связывает состояния при наличии доступа или возможности прямого взаимодействия с компонентом или интерфейсом ПИИ. Реализация правила приведена в таблице 3.1 для выбранных тактик.

Таблица 3.1 – Пример выбранных действий (совокупности техник (тактики)) по типу функционирования злоумышленника с учетом необходимости обязательного взаимодействия во время атаки с вычислительной моделью ИИ

Тип (действия осуществимы)		Разведка – Р (множество Р)	Подготовка к атаке - П (множество П)	Доступ – Д (множество Д)	Воздействие – В (множество В)	Блокировка – Б (множество Б)
1	Без обязательного контролируемого взаимодействия нарушителя с системой ИИ (управление действиями) (множество *Ancon*)	Reconnaissance	Resource Development	-	-	-
2	Вычислительная модель доступна для управления действиями обеим сторонам (атакующей стороне и защищающейся), соответственно контроль условен (множество *Amid*)	Discovery, Collection	-	Initial Access, AI Model Access, Defense Evasion, Credential Access	-	-
3	Требуется обязательное контролируемое взаимодействие нарушителя с моделью ИИ (множество *Acon*)	-	AI Attack Staging, Persistence	Privilege Escalation	Execution, Exfiltration, Impact	Блокировано

Более подробно это правило раскрывается в таблице 3.1 составленной с учетом таблиц 2.7, 2.8, 2.9.

Для того чтобы оценить каждое действие и охарактеризовать тактику или технику (в зависимости от уровня детализации), требуется выбрать доступные варианты использования из таблицы 3.2.

Таблица 3.2 - Варианты использования

Крите-рий	Для режима обучения (О)			Для режима эксплуатации (Е)		
	Ancon	Amid	Acon	Ancon	Amid	Acon
Авр	Авр=О∩Ancon	Авр=О∩Amid	Авр=О∩Acon	Авр=Е∩An	Авр=Е∩Amid	Авр=Е∩Acon
Анвр	Анвр=О∩$\overline{Авр}$	Анвр=О∩$\overline{Авр}$	Анвр=О∩$\overline{Авр}$	Анвр=Е∩$\overline{Авр}$	Анвр=Е∩$\overline{Авр}$	Анвр=Е∩$\overline{Авр}$
Ads	Ads=О∩Ancon	Ads=О∩Amid	Ads=О∩Acon	Ads=Е∩∅=∅	Ads=Е∩∅=∅	Ads=Е∩∅=∅
Ands	Ands=О∩$\overline{Ads}$	Ands=О∩$\overline{Ads}$	Ands=О∩$\overline{Ads}$	Ands=Е∩S=*E*	Ands=Е∩S=*E*	Ands=Е∩S=E
Ats	Ats=О∩Ancon	Ats=О∩Amid	Ats=О∩Acon	Ats=Е∩∅=∅	Ats=Е∩∅=∅	Ats=Е∩∅=∅
Ants	Ants=О∩$\overline{Ats}$	Ants=О∩$\overline{Ats}$	Ants=О∩$\overline{Ats}$	Ants=Е∩S=E	Ants=Е∩S=E	Ants=Е∩S=E

Таким образом, доступность действия a_i определяется подобным образом: a_i∈Asi∩(Авр∪Анвр)∩(Ads∪Ands)∩(Ats∪Ants)∩(Ancon∪Amid∪Acon) ∩(AO∪AE).

При первом получении данных механизмы фильтрации ПИИ обычно находятся в нейтральном состоянии по отношению к принимаемой информации. Некоторые системы, основанные на эксплуатации нейронных сетей, контролируют источники сообщений при обучении в заданный для этого период времени. Если системами определяется, что полученные ими данные являются подлинными, то есть соответствуют результатам контрольной выборки, то источники информации считаются заслуживающими доверия и в дальнейшем данные от них будут фильтроваться с меньшей интенсивностью [135].

Модель МППР для сценария атаки на модели и системы ИИ предполагает также несколько состояний, с которыми злоумышленник может столкнуться при взаимодействии с ПИИ:

– разведка (совокупность техник MIRTE ATLAS или Методики ФСТЭК, реализующих действия, связанные со сбором информации, о состоянии системы, специфики ее уязвимостей);

– подготовка (совокупность техник, реализующих действия, связанные со сбором информации, о состоянии системы, специфики ее уязвимостей П);

– доступ (Д);

– воздействие (В);

– блокировка (Б).

Учитывается возможность наличия или отсутствия обратных дуг. В соответствии с классификацией состояний атакуемой системы (доступные для прямого и контролируемого воздействия злоумышленнику при атаке на ПИИ) вводятся ограничения на переходы между состояниями моделей. Условия и ограничения при определении (наличии) переходов между состояниями следующие:

1. Режим применения ПИИ:

– эксплуатация;

– обучение;

– объединение режимов применения.

2. Наличие или отсутствие контролируемого (управляемого) злоумышленником взаимодействия (доступа) к ПИИ и ее компонентам в зависимости от типа функционирования и архитектуры ПИИ:

– наличие или отсутствие доступа во время атаки к процессам вычислений, интерфейсам вычислительных моделей (вычислительным моделям);

– наличие или отсутствие доступа во время атаки к датасетам, данным;

– наличие или отсутствие доступа во время атаки к инфраструктурным компонентам.

При этом следует отметить специфику двух возникающих ситуаций при описании сценариев атак, при которых проявляются режимы формирования моделей МППР off-line и on-line:

1. Ситуация, когда актуален режим off-line, существует при описании переходов связи действия–состояние в процессе формирования сценария атаки как наихудшего для стороны защиты и наилучшего для атакующего (атакующий стремится всегда выбирать оптимальный маршрут до целевого состояния) с точки зрения достижения целевого состояния.

При этом целевое состояние, к которому должен стремиться злоумышленник, определяется как наихудшее для системы аудитором, с учетом наличия предварительного представления о наилучшем пути эксплуатации уязвимостей при предполагаемом раскрытии (известности) инфраструктуры. В этом случае связи будут подчиняться следующим требованиям:

- не могут использоваться возвратные действия (сброс R недопустим), то есть неактуальная последовательность сразу исключается;

- боковое помещение означает формирование новой последовательности;

- специфика зоны определяет наличие связей взависимости от инфраструктуры ПИИ с учетом ее эксплуатации и доступности;

В целом модель предполагает поиск лучшего пути атаки при условии наличия возможностей эксплуатации самых опасных уязвимостей.

2. Ситуация для режима on-line существует при описании переходов (связи действие–состояние) в процессе формирования сценария атаки как наилучшего для атакующего, при этом он может свободно выбирать целевые состояния и маршрут его достижения. В этом случае учитывается отсутствие предварительного представления о наилучшем пути эксплуатации уязвимостей для злоумышленника при неизвестности инфраструктуры. Связи состояний будут подчиняться следующим правилам:

- допускается наличие возвратных связей (сброс *R*), и последовательность не прерывается;

- неактуальная последовательность не исключается, поэтому боковое помещение допустимо;

- специфика зоны ограничивает наличие связей взависимости от инфраструктуры ПИИ с учетом ее эксплуатации и доступности (контроля действий).

В целом модель предполагает поиск пути атаки в свободном порядке (с возможностью случайного смещения к допустимому состоянию).

На основе типов доступа при формировании моделей в данном случае следует учитывать следующее:

1. Существует ли зона доступных для злоумышленника состояний в режиме контролируемого управляемого доступа к состояниям и неподконтрольная стороне защиты.

2. Существует ли зона частично доступных для злоумышленника состояний в режиме контролируемого управляемого доступа и контролируемая стороной защиты. Потеря полного контроля означает, что часть переходов (действий) становится недоступна, но при этом сохраняется управление отдельным числом действий.

3. Существует ли зона недоступных для злоумышленника состояний в режиме контролируемого доступа и контролируемая стороной защиты.

Могут использоваться следующие действия:

1. Основные действия. При атаке злоумышленник может выбрать следующие действия с учетом выбранной им стратегии:

- нелегальное взаимодействие (*D*);
- легальное взаимодействие (передача легитимных обучающих данных) (*C*);
- сброс (*R*).

2. Техники-действия (реализуют техники, необходимые при построении подробной по действиям модели) соответствуют техникам, которые входят в тактики, и при этом ассоциированы с действиями, представленными дугами графа состояний последовательности атакующих воздействий. Подобные действия могут быть легального и нелегального типа.

Таким образом формируется несколько множеств действий при переходе между состояниями (Рисунок 3.7). При выборе стратегии атаки в рамках

моделирования с использованием МППР также учитываются параметры «вознаграждения» и «затрат» как части вознаграждения.

Вознаграждение описывается с учетом выбранной функции и полученных для нее основных параметров (2.34, 2.35):

– метрики CVSS для компонентного состава ИС;

– отклонения моделей;

– уязвимость датасета и др.

Таким образом, в процессе моделирования, при определении значений ожидаемых вознаграждений R, нужно учитывать следующие особенности:

1. Выгоды, получаемые при переходе между состояниями, зависят от исходных и целевых вершин графа атаки.

2. Величина вознаграждений будет оганичена в случае, когда злоумышленник сначала отправляет ложные данные, а затем его сообщения блокируют независимо от каких-либо действий.

Относительный размер вознаграждений R, полученных злоумышленником за каждое изменение (переход в иное состояние), независимо от возможных действий отвечает следующим закономерностям: вознаграждение соответствует параметрам и, вычисляется по формулам (в зависимости от режима построения атакующей последовательности), приводимым в главе 2 для случаев on-line и off-line. При этом следует учитывать, что при условии недостижения конечного состояния атаки суммарное значение V_i^* может превышать V_i^* последовательности с конечными состояниями, означающими достижение цели. Подобные последовательности будут считаться действительными. Задача специалиста – определить конечные состояния атакующей последовательности.

С учётом ограничений модели целесообразно применять в этом случае первый и второй режим описания действий модели. В этом случае учитываются: нелегальное взаимодействие (*D*); легальное взаимодействие (передача легитимных данных) (*C*); сброс (*R*), уточняющие действия (по тактикам) с учётом

распределения действий по возможности реализации таковых в зависимости от архитектуры ПИИ.

Специфика моделирования следующая:

1. Используются следующие функции полезности:

V^*_{i+1} (s=SP) - функция полезности для состояния SP (Реализация) на итерации i+1;

V^*_{i+1} (s=SП) - функция полезности для состояния SП (Подготовка) на итерации i+1;

V^*_{i+1} (s=SД) - функция полезности для состояния SД (Доступ) на итерации i+1;

V^*_{i+1} (s=SB) - функция полезности для состояния SB (Воздействие) на итерации i+1;

V^*_{i+1} (s=SБ) - функция полезности для состояния SБ (Блокировка) на итерации i+1.

2. Формируется множество функций полезности:

$$V = \{V^*_{i+1}\ (s = S\text{P}), V^*_{i+1}(s = S\text{П}), V^*_{i+1}(s = S\text{Д}), V^*_{i+1}(s = S\text{B}), V^*_{i+1}(s = S\text{Б})\}$$

3. Используется множество состояний *S*: *S*={*S*P, *S*П, *S*Д, *S*B, *S*Б}, где:

*S*P — состояние Реализации;

*S*П — состояние Подготовки;

*S*Д — состояние Доступа;

*S*B — состояние Воздействия;

*S*Б — состояние Блокировки;

4. Используется множество действий *A*: *A* = {*D*, *C*, *R*}, где:

D — нелегальное взаимодействие;

C — легальное взаимодействие;

R — сброс.

5. Множество вознаграждений *R (*вознаграждения за достижение соответствующих состояний*)*: R={R(*S*P), R(*S*П), R(*S*Д), R(*S*B), R(*S*Б)}.

6. Используется коэффициент дисконтирования: $\gamma \in R$.

7. Используется идентификация вероятностей переходов:

P(*S*P, *a*P=*D*, *S*П) - вероятность перехода от состояния SP к SП при нелегальном взаимодействии (*D*).

P(*S*P, *a*P=*C*, *S*Д) - вероятность перехода от состояния SP к SД при легальном взаимодействии (C).

P(*S*П, *a*П=*D*, *S*Б) - вероятность перехода от состояния SП к SБ при нелегальном взаимодействии (*D*).

P(*S*П, *a*П=*C*, *S*Д) - вероятность перехода от состояния SП к SД при легальном взаимодействии (C).

P(*S*Д, *a*Д=*D*, *S*В) - вероятность перехода от состояния SД к SВ при нелегальном взаимодействии (*D*).

P(*S*Д, *a*Д=*C*, *S*Б) - вероятность перехода от состояния SД к SБ при легальном взаимодействии (C).

P(*S*В, *a*В=*D*, *S*Б) - вероятность перехода от состояния SВ к SБ при нелегальном взаимодействии (*D*).

P(*S*В ,*a*В=*C*, *S*Б) - вероятность перехода от состояния SВ к SБ при легальном взаимодействии (*C*).

Для описания вероятностей переходов между состояниями в зависимости от действий, можно использовать функцию $P: S \times A \times S \rightarrow [0,1]$, где $P(S_i, a, S')$ обозначает вероятность перехода от состояния S_i к состоянию S' при выполнении действия a.

Ограничения модели при применении:

- не требует детализации атакующих действий по методикам (техникам) (для режима без детализации);

- используется сопоставления действий, уязвимостей, локальных идентификаторов (для второго режима);

- состояния абстрактно описывают безопасность системы;

- уточняющие действия допустимы (по тактикам и техникам для режима с детализацией).

Для того чтобы определить последовательность действий злоумышленника при атаке выгодную для него с учетом всех предоставленных ему возможностей

(лучшие последовательности действий), можем использовать обобщенные обозначения для состояний и действий. Ниже приведены формулы, которые могут применяться к любому состоянию. В случаях, когда действие может привести к нескольким возможным состояниям (S'), используется сумма по всем возможным состояниям для учета вероятностей перехода и соответствующих вознаграждений.

Эти формулы позволяют анализировать оптимальные действия в зависимости от текущего состояния и выбранного действия в более универсальном формате. Они могут применяться для поиска оптимальных последовательностей действий в любой модели приводимого исследования. Далее рассматриваются типы моделей с учетом объединения тактик (таблица 3.1) и специфики доступности состояний с учетом инфраструктуры ПИИ.

Первый подтип упрощенных моделей приведен на рисунке 3.7. Функции полезности для состояний *S*Р, *S*П, *S*Д, *S*В и *S*Б описывают, как злоумышленник может оптимизировать свои действия в зависимости от текущего состояния и доступных ему в зонах действий. Каждая функция полезности учитывает вероятности переходов и вознаграждения, что позволяет формировать стратегию атаки.

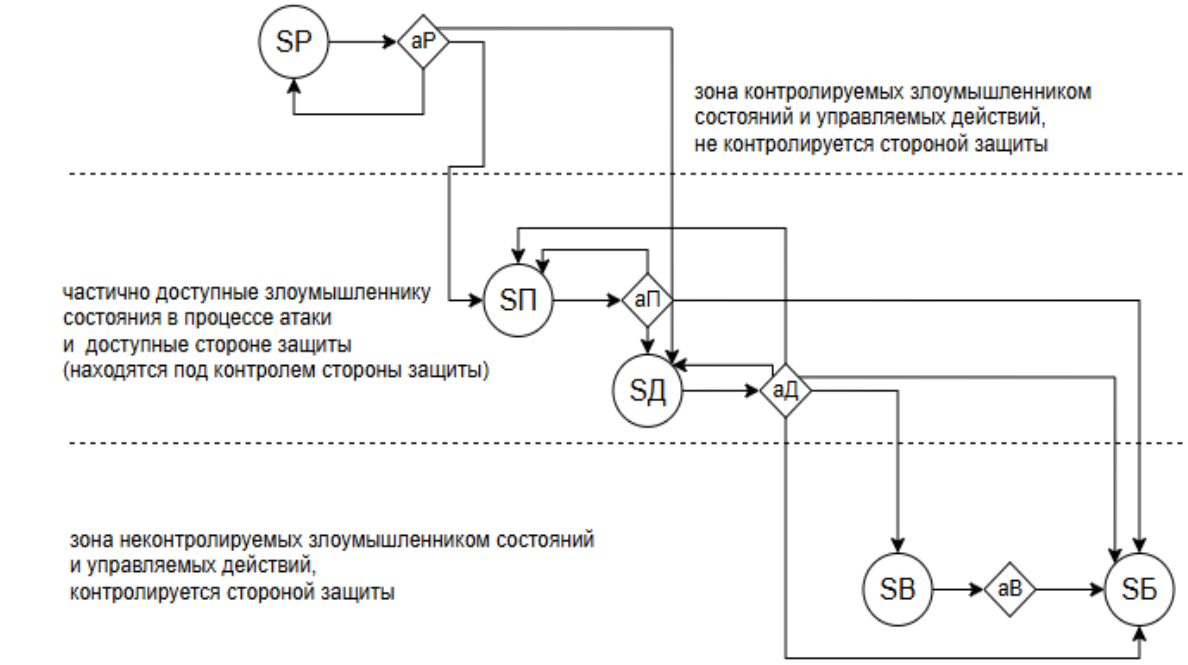
Учитываются ограничения действий в соответствии с зонами:

1. Состояние *S*Р (набор действий разведки) принадлежит зоне доступных для злоумышленника состояний в режиме контролируемого доступа и неподконтрольно стороне защиты.

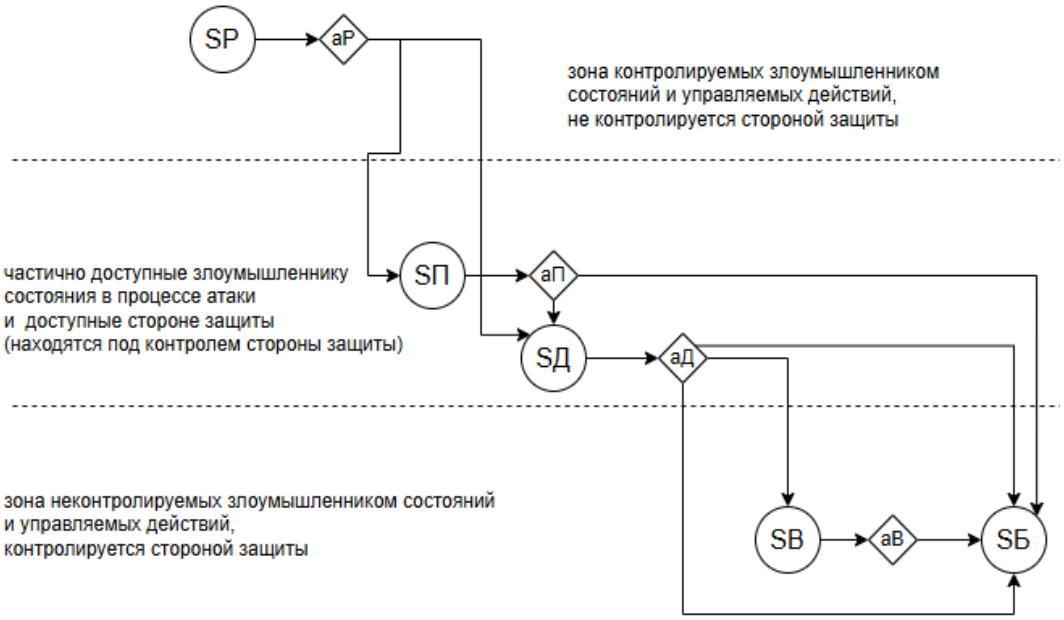
2. Состояния *S*П и *S*Д (набор действий подготовки к атаке и развертывания атаки - доступа) принадлежат частично зоне доступных для злоумышленника состояний в режиме контролируемого доступа и контролируемы стороной защиты. В этом случае, ограничение в управлении действиями (переходами) выражается в отсутствии возможности перехода к *S*Р. Злоумышленник попадает в замкнутое множество состояний и доступных действий.

3. Состояния *S*В и *S*Б (набор действий, позволяющих воздействовать на системы при атаке и действия блокирования) принадлежат зоне недоступных для

злоумышленника состояний в режиме контролируемого доступа и контролируемых стороной защиты.



а)



б)

Рисунок 3.7 – Графы состояний моделей on-line (а), off-line (б) первого подтипа

Определение значений для каждого состояния с использованием функции полезности модели на основе МППР для общей атаки в режиме on-line на ПИИ

приведено в выражениях ниже (функции полезности представлены в формулах с (3.15) по (3.19)). Модель взаимодействия приведена на рисунке 3.7 (а).

Функции для модели on-line предполагают связи состояний по следующим типам действий, предполагающим детализацию: нелегальное взаимодействие (*D*), легальное взаимодействие (*C*). От состояния исходит действие (разновидности: нелегальное взаимодействие (*D*), легальное взаимодействие (*C*), сброс (*R*)).

Описание полезности включает в свой состав следующее (3.16 – 3.20):

1. Функция полезности для состояния *S*Р (Разведка). От состояния *S*Р исходит действие аР в состояние *S*П, *S*Д, *S*Р. Не связано с состояниями *S*Б, *S*В.

$$V_{i+1}^{*}(S\text{Р}) = R(S\text{Р}) + \gamma max\{P(S\text{Р}, a\text{Р} = D, S\text{П})[R(S\text{Р}, a\text{Р} = D, S\text{П}) + \gamma V_i^{*}(S\text{П})] + P(S\text{Р}, a\text{Р} = D, S\text{Д})[R(S\text{Р}, a\text{Р} = D, S\text{Д}) + \gamma V_i^{*}(S\text{Д})] + P(S\text{Р}, a\text{Р} = C, S\text{П})[R(S\text{Р}, a\text{Р} = C, S\text{П}) + \gamma V_i^{*}(S\text{П})] + P(S\text{Р}, a\text{Р} = C, S\text{Д})[R(S\text{Р}, a\text{Р} = C, S\text{Д}) + \gamma V_i^{*}(S\text{Д})] + P(S\text{Р}, a\text{Р} = R, S\text{Р})[R(S\text{Р}, a\text{Р} = R, S\text{Р}) + V_i^{*}(S\text{Р})]\}. \quad (3.16)$$

2. Функция полезности для состояния *S*П (Подготовка). От состояния *S*П исходит действие аП в состояния *S*Б, *S*Д, *S*П (только R), не связано с *S*Р, *S*В.

$$V_{i+1}^{*}(S\text{П}) = \text{R}(S\text{П}) + \gamma max\{P(S\text{П}, a\text{П} = D, S\text{Д})[\text{R}(S\text{П}, a\text{П} = D, S\text{Д}) + \gamma V_i^{*}(S\text{Д})] + P(S\text{П}, a\text{П} = D, S\text{Б})[\text{R}(S\text{П}, a\text{П} = D, S\text{Б}) + \gamma V_i^{*}(S\text{Б})] + P(S\text{П}, a\text{П} = C, S\text{Д})[\text{R}(S\text{П}, a\text{П} = C, S\text{Д}) + \gamma V_i^{*}(S\text{Д})] + P(S\text{П}, a\text{П} = C, S\text{Б})[\text{R}(S\text{П}, a\text{П} = C, S\text{Б}) + \gamma V_i^{*}(S\text{Б})] + P(S\text{П}, a\text{П} = R, S\text{П})[\text{R}(S\text{П}, a\text{П} = R, S\text{П}) + \gamma V_i^{*}(S\text{П})]\}. \quad (3.17)$$

3. Функция полезности для состояния SД (Доступ). От состояния SД исходит действие аД в состояния SВ, SБ, SП. Не связано с SД, SВ.

$$V_{i+1}^{*}(s = S\text{Д}) = \text{R}(S\text{Д}) + \gamma max\{P(S\text{Д}, a\text{Д} = D, S\text{П})[\text{R}(S\text{Д}, a\text{Д} = D, S\text{П}) + \gamma V_i^{*}(S\text{П})] + P(S\text{Д}, a\text{Д} = C, S\text{П})[\text{R}(S\text{Д}, a\text{Д} = C, S\text{П}) + \gamma V_i^{*}(S\text{П})] + P(S\text{Д}, a\text{Д} = D, S\text{В})[\text{R}(S\text{Д}, a\text{Д} = D, S\text{В}) + \gamma V_i^{*}(S\text{В})] + P(S\text{Д}, a\text{Д} = C, S\text{В})[\text{R}(S\text{Д}, a\text{Д} = C, S\text{В}) + \gamma V_i^{*}(S\text{В})] +$$

$$P(S\text{Д}, a\text{Д} = D, S\text{Б})[\text{R}(S\text{Д}, a\text{Д} = D, S\text{Б}) + \gamma V_i^*(S\text{Б})] + P(S\text{Д}, a\text{Д} = C, S\text{Б})[\text{R}(S\text{Д}, a\text{Д} = C, S\text{Б}) + \gamma V_i^*(S\text{Б})] + P(S\text{Д}, a\text{Д} = R, S\text{Д})[\text{R}(S\text{Д}, a\text{Д} = R, S\text{Д}) + \gamma V_i^*(S\text{Д})] + P(S\text{Д}, a\text{Д} = R, S\text{П})[\text{R}(S\text{Д}, a\text{Д} = R, S\text{П}) + \gamma V_i^*(S\text{П})]\}. \quad (3.18)$$

4. Функция полезности для состояния *S*В (Выполнение). От состояния *S*В исходит действие аВ в состояния *S*Б, не связано с *S*П, *S*Д, *S*В, *S*Р.

$$V_{i+1}^*(S\text{В}) = \text{R}(S\text{В}) + \gamma max\{P(S\text{В}, a\text{В} = D, S\text{Б})[\text{R}(S\text{В}, a\text{В} = D, S\text{Б}) + \gamma V_i^*(S\text{Б})] + P(S\text{В}, a\text{В} = C, S\text{Б})[\text{R}(S\text{В}, a\text{В} = C, S\text{Б}) + \gamma V_i^*(S\text{Б})]\}. \quad (3.19)$$

5. Функция полезности для состояния SБ (Блокировка).

$$V_{i+1}^*(s = S\text{Б}) = \text{R}(S\text{Б}). \quad (3.20)$$

Функции полезности для модели off-line (Рисунок 3.7 (б)) включает набор состояний SР (разведка), SП (подготовка), SД (доступ), SВ (выполнение) и SБ (блокировка). Учитываются их связи по типам действий: нелегальное взаимодействие (D), легальное взаимодействие (C). От состояния исходит действие (разновидности: нелегальное взаимодействие (D), легальное взаимодействие (C)).

1. Функция полезности для состояния *S*Р (Разведка). От состояния SР исходит действие аР в состояния *S*П и *S*Д. Не связано с состояниями *S*Р, *S*Б, *S*В.

$$V_{i+1}^*(s = S\text{Р}) = \text{R}(S\text{Р}) + \gamma max\{P(S\text{Р}, a\text{Р} = D, S\text{П})[\text{R}(S\text{Р}, a\text{Р} = D, S\text{П}) + \gamma V_i^*(S\text{П})] + P(S\text{Р}, a\text{Р} = D, S\text{Д})[\text{R}(S\text{Р}, a\text{Р} = D, S\text{Д}) + \gamma V_i^*(S\text{Д})] + P(S\text{Р}, a\text{Р} = C, S\text{П})[\text{R}(S\text{Р}, a\text{Р} = C, S\text{П}) + \gamma V_i^*(S\text{П})] + P(S\text{Р}, a\text{Р} = C, S\text{Д})[\text{R}(S\text{Р}, a\text{Р} = C, S\text{Д}) + \gamma V_i^*(S\text{Д})]\}. \quad (3.21)$$

2. Функция полезности для состояния *S*П (Подготовка). От состояния *S*П исходит действие аП ((тип (D), тип (C)) в состояния *S*Б, *S*Д. Не связано с *S*Р, *S*В.

$$V_{i+1}^{*}(s = S\text{П}) = \text{R}(S\text{П}) + \gamma max\{P(S\text{П}, a\text{П} = D, S\text{Д})[\text{R}(S\text{П}, a\text{П} = D, S\text{Д}) + \gamma V_i^{*}(S\text{Д})] + P(S\text{П}, a\text{П} = D, S\text{Б})[\text{R}(S\text{П}, a\text{П} = D, S\text{Б}) + \gamma V_i^{*}(S\text{Б})] + P(S\text{П}, a\text{П} = C, S\text{Д})[R(S\text{П}, a\text{П} = C, S\text{Д}) + \gamma V_i^{*}(S\text{Д})] + P(S\text{П}, a\text{П} = C, S\text{Б})[\text{R}(S\text{П}, a\text{П} = C, S\text{Б}) + \gamma V_i^{*}(S\text{Б})] + P(S\text{П}, a\text{П} = R, S\text{П})[\text{R}(S\text{П}, a\text{П} = R, S\text{П}) + \gamma V_i^{*}(S\text{П})]\}. \quad (3.22)$$

3. Функция полезности для состояния SД (Доступ). От состояния SД исходит действие аД в состояния SВ, SБ, SП. Не связано с SД, SВ.

$$V_{i+1}^{*}(s = S\text{Д}) = \text{R}(S\text{Д}) + \gamma max\{P(S\text{Д}, a\text{Д} = D, S\text{В})[\text{R}(S\text{Д}, a\text{Д} = D, S\text{В}) + \gamma V_i^{*}(S\text{В})] + P(S\text{Д}, a\text{Д} = C, S\text{В})[\text{R}(S\text{Д}, a\text{Д} = C, S\text{В}) + \gamma V_i^{*}(S\text{В})] + P(S\text{Д}, a\text{Д} = D, S\text{Б})[\text{R}(S\text{Д}, a\text{Д} = D, S\text{Б}) + \gamma V_i^{*}(S\text{Б})] + P(S\text{Д}, a\text{Д} = C, S\text{Б})[\text{R}(S\text{Д}, a\text{Д} = C, S\text{Б}) + \gamma V_i^{*}(S\text{Б})]\}. \quad (3.23)$$

4. Функция полезности для состояния SB (Выполнение). От состояния SB исходит действие аВ, в состояние SБ, не связано с SП, SД, SВ, SР.

$$V_{i+1}^{*}(s = S\text{В}) = \text{R}(S\text{В}) + \gamma max\{P(S\text{В}, a\text{В} = D, S\text{Б})[\text{R}(S\text{В}, a\text{В} = D, S\text{Б}) + \gamma V_i^{*}(S\text{Б})] + P(S\text{В}, a\text{В} = C, S\text{Б})[\text{R}(S\text{В}, a\text{В} = C, S\text{Б}) + \gamma V_i^{*}(S\text{Б})]\}. \quad (3.24)$$

5. Функция полезности для состояния SБ (Блокировка).

$$V_{i+1}^{*}(s = S\text{Б}) = \text{R}(S\text{Б}). \quad (3.25)$$

На рисунках 3.8 отражена динамика изменений предпочтительности состояний при усилении влияния на степень эксплуатируемости (уровень вложений злоумышленника).

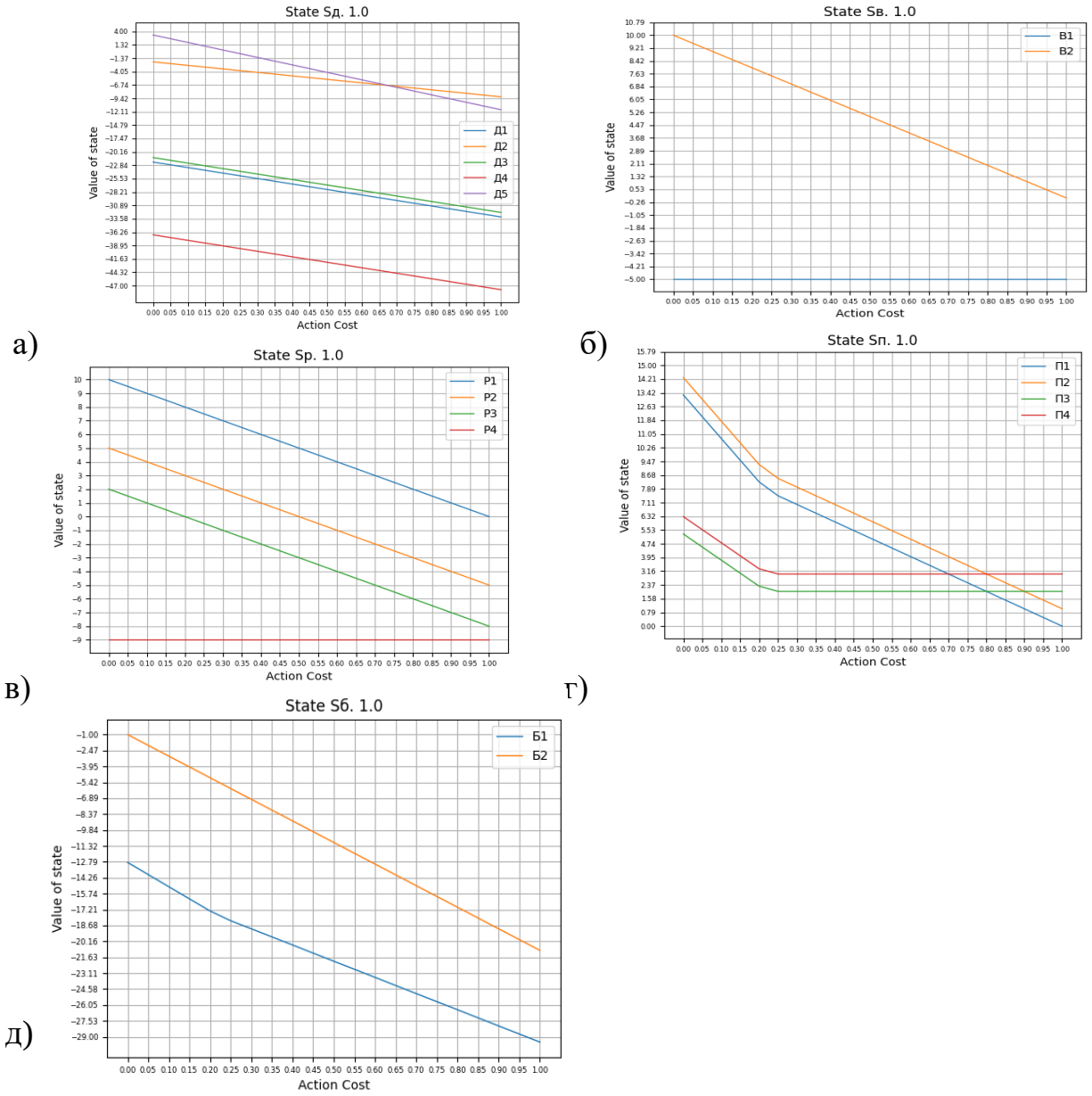


Рисунок 3.8 – Соотношение усилий злоумышленника (наград) и значимости состояний: SД(а), SВ(б), SР(в), SП(г), SБ(д)

Таким образом, для исследования состояний производится построение диаграммы, в которой отражается динамика модели при смене влияния на величины вознаграждений, получаемых злоумышленником в зависимости от части вознаграждений, вкладываемых при атаке (параметр CVSS, отвечающий за меру сложности). Учитывается стоимость атакующих воздействий (часть награды) и значения состояния при понижении и при повышении значимости вложений

злоумышленника в нападении. При этом учет вложений проводится для каждого состояния.

Пример определения последовательности наилучших состояний в заданных условиях (Рисунок 3.9, 3.10): для обоих случаев цена взаимодействия 10; для примера фиксировали действия ("Р4", "П3", "П4", "Д2", "В1"); награды и вероятности для действий, кроме аР и аП, равные; вознаграждения для случая 1 и 3 равны.

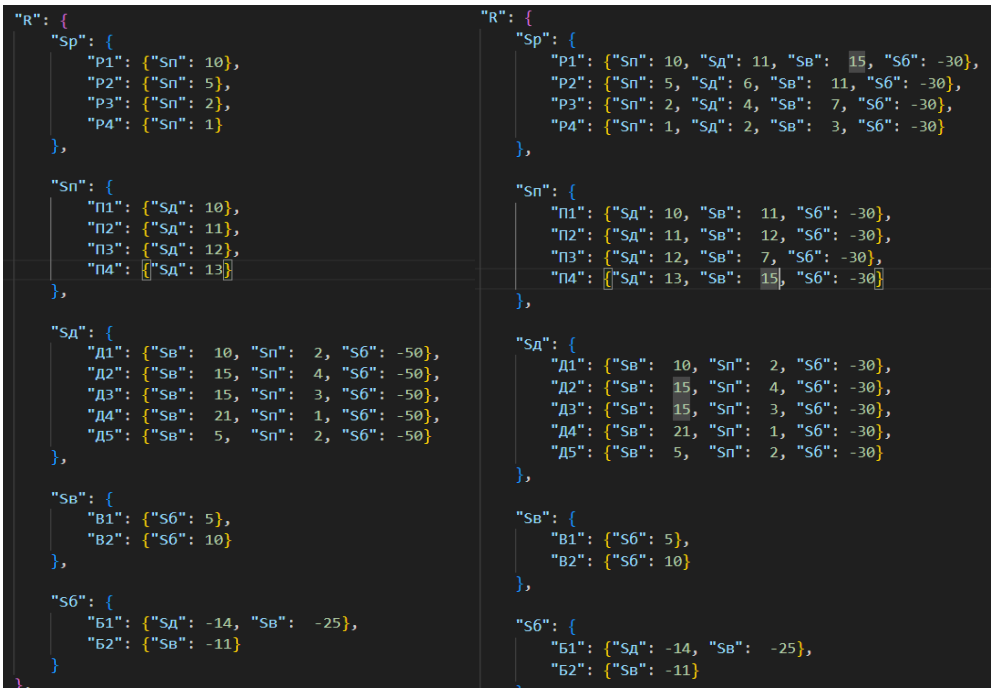


Рисунок 3.9 – Награды для действий

Оптимальные политики приведены на рисунке 3.10. Анализ показывает устойчивость вектора атаки.

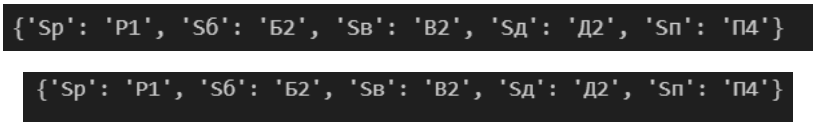


Рисунок 3.10 – Последовательности состояний

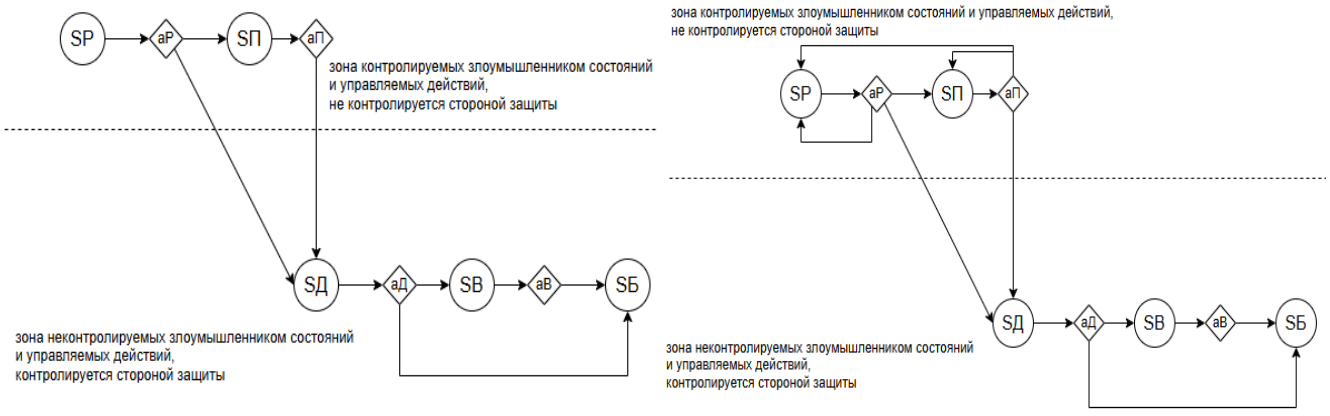
Анализ показывает устойчивость вектора атаки. Оптимальная стратегия, использующая алгоритм итерации значений, позволяет определить: оптимальное действие, когда злоумышленник находится в состоянии контролируемого взаимодействия (*S*P); оптимальное действие, когда злоумышленник находится в состоянии доверенное взаимодействие (*S*П), также атакует систему, оптимальное действие, когда злоумышленник находится в состоянии легальное взаимодействие (*S*В); оптимальное действие, когда злоумышленник находится в состоянии легальное взаимодействие (*S*Д и *S*Б).

Второй подтип моделей учитывает следующие ограничения действий в соответствии с зонами контролируемого доступа:

1. Состояния *S*P и *S*П (набор действий разведки и набор действий подготовки к атаке) принадлежат зоне доступных для злоумышленника состояний в режиме контролируемого доступа и неподконтрольны стороне защиты;

2. Состояния *S*В, *S*Д и *S*Б (набор действий, позволяющих воздействовать на системы при атаке) принадлежат частично зоне недоступных для злоумышленника состояний в режиме контролируемого доступа и контролируемы стороной защиты.

Граф состояний модели взаимодействия off-line приведена на рисунке 3.11(а). Отсутствие обратных дуг, обозначающих действие R, в данной модели обусловлено характером построения сценария при пассивном аудите.



а) б)

Рисунок 3.11 – Графы состояний моделей off-line (а), on-line (б) второго подтипа

Функции полезности представлены в формулах с (3.26) по (3.30) для состояний *S*Р (разведка), *S*П (подготовка), *S*Д (доступ), *S*В (выполнение) и *S*Б (блокировка) с учетом их связи по типам действий (нелегальное взаимодействие (*D*) и легальное взаимодействие (*C*) (Рисунок 3.11 (а)):

1. Функция полезности для состояния *S*Р (Разведка). От состояния SP исходит действие аР в состояние *S*П и состояние *S*Д, не связано с состояниями SP, *S*Б, *S*В.

$$V_{i+1}^{*}(s = S\text{Р}) = \text{R}(S\text{Р}) + \gamma max\{P(S\text{Р}, a\text{Р} = D, S\text{П})[\text{R}(S\text{Р}, a\text{Р} = D, S\text{П}) + \gamma V_i^{*}(S\text{П})] + P(S\text{Р}, a\text{Р} = D, S\text{Д})[\text{R}(S\text{Р}, a\text{Р} = D, S\text{Д}) + \gamma V_i^{*}(S\text{Д})] + P(S\text{Р}, a\text{Р} = C, S\text{П})[\text{R}(S\text{Р}, a\text{Р} = C, S\text{П}) + \gamma V_i^{*}(S\text{П})] + P(S\text{Р}, a\text{Р} = C, S\text{Д})[\text{R}(S\text{Р}, a\text{Р} = C, S\text{Д}) + \gamma V_i^{*}(S\text{Д})]\}. \quad (3.26)$$

2. Функция полезности для состояния SП (Подготовка). От состояния SП исходит действие аП в состояние SП, SД, не связано с SP, SB, SБ;

$$V_{i+1}^{*}(s = S\text{П}) = \text{R}(S\text{П}) + \gamma max\{P(S\text{П}, a\text{П} = D, S\text{Д})[\text{R}(S\text{П}, a\text{П} = D, S\text{Д}) + \gamma V_i^{*}(S\text{Д})] + P(S\text{П}, a\text{П} = C, S\text{Д})[\text{R}(S\text{П}, a\text{П} = C, S\text{Д}) + \gamma V_i^{*}(S\text{Д})]\}. \quad (3.27)$$

3. Функция полезности для состояния SД (Доступ). От состояния SД исходит действие аД в состояние SB, SБ, не связано с SД, SП.

$$V_{i+1}^{*}(s = S\text{Д}) = \text{R}(S\text{Д}) + \gamma max\{P(S\text{Д}, a\text{Д} = D, S\text{В})[\text{R}(S\text{Д}, a\text{Д} = D, S\text{В}) + \gamma V_i^{*}(S\text{В})] + P(S\text{Д}, a\text{Д} = C, S\text{В})[\text{R}(S\text{Д}, a\text{Д} = C, S\text{В}) + \gamma V_i^{*}(S\text{В})] + P(S\text{Д}, a\text{Д} = D, S\text{Б})[\text{R}(S\text{Д}, a\text{Д} = D, S\text{Б}) + \gamma V_i^{*}(S\text{Б})] + P(S\text{Д}, a\text{Д} = C, S\text{Б})[\text{R}(S\text{Д}, a\text{Д} = C, S\text{Б}) + \gamma V_i^{*}(S\text{Б})]\}. \quad (3.28)$$

4. Функция полезности для состояния SB (Выполнение). От состояния SB исходит действие аВ в состояние SБ, не связано с SП, SД, SB, SP.

$$V_{i+1}^{*}(s = S\text{В}) = \text{R}(S\text{В}) + \gamma max\{P(S\text{В}, a\text{В} = D, S\text{Б})[\text{R}(S\text{В}, a\text{В} = D, S\text{Б}) + \gamma V_i^{*}(S\text{Б})] + P(S\text{В}, a\text{В} = C, S\text{Б})[\text{R}(S\text{В}, a\text{В} = C, S\text{Б}) + \gamma V_i^{*}(S\text{Б})]\}. \quad (3.29)$$

5. Функция полезности для состояния SБ (Блокировка).

$$V_{i+1}^{*} = (s = S\text{Б}) = \text{R}(S\text{Б}) . \quad (3.30)$$

Определение значений для каждого состояния с использованием функции полезности модели на основе МППР для общей атаки в режиме on-line на ПИИ приведено в выражениях ниже (Рисунок 3.11 (б)). Функции полезности (представлены в формулах с (3.31) по (3.35)) для модели включает набор состояний SР (разведка), SП (подготовка), SД (доступ), SВ (выполнение) и SБ (блокировка). Учитываются их связи по типам действий: нелегальное взаимодействие (D), легальное взаимодействие (C), сброс (R)) (Рисунок 3.11 (б)).

1. Функция полезности для состояния SP (Разведка). От состояния SP исходит действие aР в состояния SП, SP, SД, не связано с состояниями SБ, SВ.

$$V_{i+1}^{*}(s = S\text{Р}) = \text{R}(S\text{Р}) + \gamma max\{P(S\text{Р}, a\text{Р} = D, S\text{П})[\text{R}(S\text{Р}, a\text{Р} = D, S\text{П}) + \gamma V_i^{*}(S\text{П})] + P(S\text{Р}, a\text{Р} = D, S\text{Д})[\text{R}(S\text{Р}, a\text{Р} = D, S\text{Д}) + \gamma V_i^{*}(S\text{Д})] + P(S\text{Р}, a\text{Р} = C, S\text{П})[\text{R}(S\text{Р}, a\text{Р} = C, S\text{П}) + \gamma V_i^{*}(S\text{П})] + P(S\text{Р}, a\text{Р} = C, S\text{Д})[\text{R}(S\text{Р}, a\text{Р} = C, S\text{Д}) + \gamma V_i^{*}(S\text{Д})] + P(S\text{Р}, a\text{Р} = R, S\text{Р})[\text{R}(S\text{Р}, a\text{Р} = R, S\text{Р}) + \gamma V_i^{*}(S\text{Р})]\}. \quad (3.31)$$

2. Функция полезности для состояния SП (Подготовка). От состояния SП исходит действие aП в состояния SП, SД, SР, не связано с SВ, SБ;

$$V_{i+1}^{*}(s = S\text{П}) = \text{R}(S\text{П}) + \gamma max\{P(S\text{П}, a\text{П} = D, S\text{Д})[\text{R}(S\text{П}, a\text{П} = D, S\text{Д}) + \gamma V_i^{*}(S\text{Д})] + P(S\text{П}, a\text{П} = C, S\text{Д})[\text{R}(S\text{П}, a\text{П} = C, S\text{Д}) + \gamma V_i^{*}(S\text{Д})] + P(S\text{П}, a\text{П} = R, S\text{П})[\text{R}(S\text{П}, a\text{П} = R, S\text{П}) + \gamma V_i^{*}(S\text{П})]\}. \quad (3.32)$$

3. Функция полезности для состояния SД (Доступ). От состояния SД исходит действие аД в состояния SВ, SБ, не связано с SД, SП, SР.

$$V_{i+1}^{*}(s = S\text{Д}) = \text{R}(S\text{Д}) + \gamma max\{P(S\text{Д}, a\text{Д} = D, S\text{В})[R(S\text{Д}, a\text{Д} = D, S\text{В}) + \gamma V_i^{*}(S\text{В})] + P(S\text{Д}, a\text{Д} = C, S\text{В})[\text{R}(S\text{Д}, a\text{Д} = C, S\text{В}) + \gamma V_i^{*}(S\text{В})] + P(S\text{Д}, a\text{Д} = D, S\text{Б})[\text{R}(S\text{Д}, a\text{Д} = D, S\text{Б}) + \gamma V_i^{*}(S\text{Б})] + P(S\text{Д}, a\text{Д} = C, S\text{Б})[\text{R}(S\text{Д}, a\text{Д} = C, S\text{Б}) + \gamma V_i^{*}(S\text{Б})]\}. \quad (3.33)$$

4. Функция полезности для состояния SB (Выполнение). От состояния SB исходит действие аВ в состояние SБ, не связано с SП, SД, SB, SP.

$$V_{i+1}^{*}(s = S\text{В}) = \text{R}(S\text{В}) + \gamma max\{P(S\text{В}, a\text{В} = D, S\text{Б})[\text{R}(S\text{В}, a\text{В} = D, S\text{Б}) + \gamma V_i^{*}(S\text{Б})] + P(S\text{В}, a\text{В} = C, S\text{Б})[\text{R}(S\text{В}, a\text{В} = C, S\text{Б}) + \gamma V_i^{*}(S\text{Б})]\}. \quad (3.34)$$

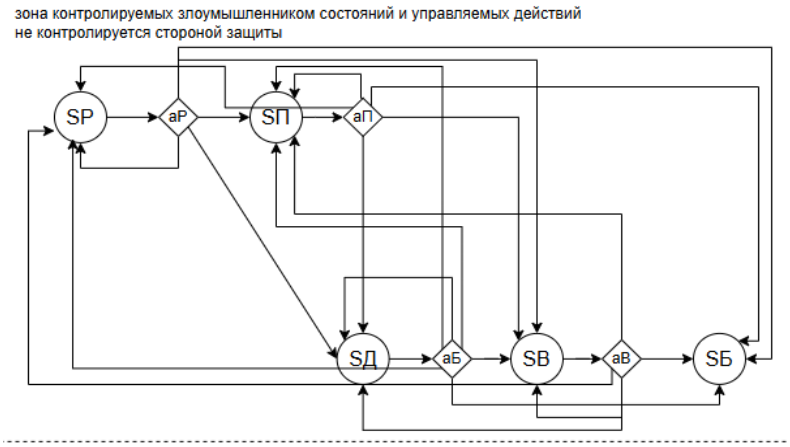
5. Функция полезности для состояния SБ (Блокировка).

$$V_{i+1}^{*} = (s = S\text{Б}) = \text{R}(S\text{Б}). \quad (3.35)$$

Третий подтип моделей применим как для описания атак на логические компоненты и интерфейсы, так и для описания атак на инфраструктурные компоненты, поскольку учитывается осуществленный доступ к элементу ПИИ как подчинение (захват) данного компонента (Рисунок 3.12).

Учитываются ограничения действий в соответствии с зонами. Состояние SР, SП, SВ, SД и SБ принадлежит зоне доступных для злоумышленника состояний в режиме контролируемого доступа и в режиме контролируемого доступа стороне защиты. Модель МППР для атак в режиме on-line приведена на рисунке 3.12 (а). Отсутствие обратных дуг, обозначающих действие R, в данной модели обусловлено характером построения сценария при пассивном аудите.

а)



б)

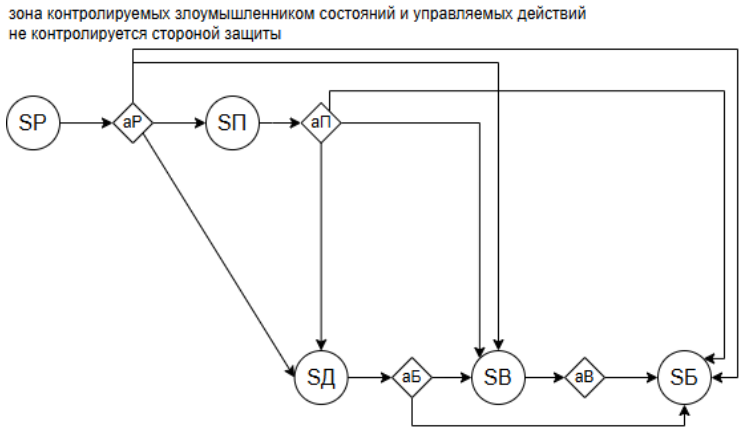


Рисунок 3.12 – Графы состояний моделей МППР третьего подтипа для атак в режиме on-line (а) и off-line (б) без ограничений на взаимодействие с компонентами ПИИ

Определения значений (Рисунок 3.12 (а)) для каждого состояния с использованием функции полезности упрощенной модели на основе МППР для общей атаки в режиме on-line на ПИИ приведены в выражениях ниже (в формулах с (3.47) по (3.51)).

1. Функция полезности для состояния SP (Разведка). От состояния SP исходит действие aP к состояниям SП, SД, SP, SБ, SВ.

$$V_{i+1}^{*}(s = SP) = \mathrm{R}(S\mathrm{P}) + \gamma max\{P(S\mathrm{P}, a\mathrm{P} = D, S\mathrm{П})[\mathrm{R}(S\mathrm{P}, a\mathrm{P} = D, S\mathrm{П}) + V_i^{*}(S\mathrm{П})]$$
$$+P(S\mathrm{P}, a\mathrm{P} = D, S\mathrm{Д})[\mathrm{R}(S\mathrm{P}, a\mathrm{P} = D, S\mathrm{Д}) + V_i^{*}(S\mathrm{Д})]$$
$$+P(S\mathrm{P}, a\mathrm{P} = D, S\mathrm{B})[\mathrm{R}(S\mathrm{P}, a\mathrm{P} = D, S\mathrm{B}) + V_i^{*}(S\mathrm{B})]$$

$$+P(S\text{Р}, a\text{Р} = R, S\text{Р})[\text{R}(S\text{Р}, a\text{Р} = R, S\text{Р}) + V_i^*(S\text{Р})]$$
$$+P(S\text{Р}, a\text{Р} = C, S\text{П})[\text{R}(S\text{Р}, a\text{Р} = C, S\text{П}) + V_i^*(S\text{П})]$$
$$+P(S\text{Р}, a\text{Р} = C, S\text{Д})[\text{R}(S\text{Р}, a\text{Р} = C, S\text{Д}) + V_i^*(S\text{Д})]$$
$$+P(S\text{Р}, a\text{Р} = C, S\text{В})[\text{R}(S\text{Р}, a\text{Р} = C, S\text{В}) + V_i^*(S\text{В})]$$
$$+P(S\text{Р}, a\text{Р} = D, S\text{Б})[\text{R}(S\text{Р}, a\text{Р} = D, S\text{Б}) + V_i^*(S\text{Б})]$$
$$+P(S\text{Р}, a\text{Р} = C, S\text{Б})[\text{R}(S\text{Р}, a\text{Р} = C, S\text{Б}) + V_i^*(S\text{Б})]\} \quad . \qquad (3.36).$$

2. Функция полезности для состояния SП (Подготовка). От состояния SП исходит действие аП к состояниям SП, SД, SР, SБ, SВ.

$$V_{i+1}^*(s = S\text{П}) = \text{R}(S\text{П}) + \gamma\max\{P(S\text{П}, a\text{П} = D, S\text{П})[\text{R}(S\text{П}, a\text{П} = D, S\text{П})$$
$$+\gamma Vi * (S\text{Р})] + P(S\text{П}, a\text{П} = D, S\text{Д})[\text{R}(S\text{П}, a\text{П} = D, S\text{Д})$$
$$+\gamma Vi * (S\text{Д})] + P(S\text{П}, a\text{П} = D, S\text{В})[\text{R}(S\text{П}, a\text{П} = D, S\text{В})$$
$$+\gamma Vi * (S\text{В})] + P(S\text{П}, a\text{П} = D, S\text{Б})[\text{R}(S\text{П}, a\text{П} = D, S\text{Б})$$
$$+\gamma Vi * (S\text{Б})] + P(S\text{П}, a\text{П} = R, S\text{П})[\text{R}(S\text{П}, a\text{П} = R, S\text{П})$$
$$+\gamma Vi * (S\text{П})] + P(S\text{П}, a\text{П} = R, S\text{Р})[\text{R}(S\text{П}, a\text{П} = R, S\text{Р})$$
$$+\gamma Vi * (S\text{Р})] + P(S\text{П}, a\text{П} = C, S\text{Д})[\text{R}(S\text{П}, a\text{П} = C, S\text{Д})$$
$$+\gamma Vi * (S\text{Д})] + P(S\text{П}, a\text{П} = C, S\text{В})[\text{R}(S\text{П}, a\text{П} = C, S\text{В})$$
$$+\gamma Vi * (S\text{В})] + P(S\text{П}, a\text{П} = C, S\text{Б})[\text{R}(S\text{П}, a\text{П} = C, S\text{Б})$$
$$+\gamma Vi * (S\text{Б})]\} . \qquad (3.37)$$

3. Функция полезности для состояния SД (Доступ). От состояния SД исходит действие аД в состояния SБ, SП, SД, SВ, SР.

$$V_{i+1}^*(s = S\text{Д}) = \text{R}(S\text{Д}) + \gamma max\{P(S\text{Д}, a\text{Д} = D, S\text{П})[\text{R}(S\text{Д}, a\text{Д} = D, S\text{П}) + V_i^*(S\text{П})]$$
$$+P(S\text{Д}, a\text{Д} = D, S\text{Б})[\text{R}(S\text{Д}, a\text{Д} = D, S\text{Б}) + V_i^*(S\text{Б})]$$
$$+P(S\text{Д}, a\text{Д} = D, S\text{П})[\text{R}(S\text{Д}, a\text{Д} = D, S\text{П}) + V_i^*(S\text{П})]$$
$$+P(S\text{Д}, a\text{Д} = D, S\text{В})[\text{R}(S\text{Д}, a\text{Д} = D, S\text{В}) + V_i^*(S\text{В})]$$
$$+P(S\text{Р}, a\text{Д} = D, S\text{Д})[\text{R}(S\text{Р}, a\text{Д} = R, S\text{Д}) + V_i^*(S\text{Д})]$$

$$+P(S\text{Д}, a\text{Д} = R, S\text{Р})[\text{R}(S\text{Д}, a\text{Д} = R, S\text{Р}) + V_i^*(S\text{Д})]$$
$$+P(S\text{Д}, a\text{Д} = R, S\text{П})[\text{R}(S\text{Д}, a\text{Д} = R, S\text{П}) + V_i^*(S\text{П})]$$
$$+P(S\text{Д}, a\text{Д} = C, S\text{П})[\text{R}(S\text{Д}, a\text{Д} = C, S\text{П}) + V_i^*(S\text{П})]$$
$$+P(S\text{Д}, a\text{Д} = C, S\text{В})[\text{R}(S\text{Д}, a\text{Д} = C, S\text{В}) + V_i^*(S\text{В})]$$
$$+P(S\text{Д}, a\text{Д} = C, S\text{Б})[\text{R}(S\text{Д}, a\text{Д} = C, S\text{Б}) + V_i^*(S\text{Б})]\}. \quad (3.38)$$

4. Функция полезности для состояния SВ (Выполнение). От состояния SВ исходит действие аВ в состояния SБ, SП, SД, SВ, SР.

$$V_{i+1}^*(s = S\text{В}) = \text{R}(S\text{В}) + \gamma max\{P(S\text{В}, a\text{В} = D, S\text{П})[R(S\text{В}, a\text{В} = D, S\text{П}) + V_i^*(S\text{П})]$$
$$+P(S\text{В}, a\text{В} = D, S\text{Д})[\text{R}(S\text{В}, a\text{В} = D, S\text{Д}) + V_i^*(S\text{Д})]$$
$$+P(S\text{В}, a\text{П} = D, S\text{П})[\text{R}(S\text{В}, a\text{В} = D, S\text{П}) + V_i^*(S\text{П})]$$
$$+P(S\text{В}, a\text{В} = D, S\text{Б})[\text{R}(S\text{В}, a\text{В} = D, S\text{Б}) + V_i^*(S\text{Б})]$$
$$+P(S\text{В}, a\text{В} = R, S\text{Д})[\text{R}(S\text{В}, a\text{В} = R, S\text{Д}) + V_i^*(S\text{Д})]$$
$$+P(S\text{В}, a\text{В} = R, S\text{В})[\text{R}(S\text{В}, a\text{В} = R, S\text{В}) + V_i^*(S\text{В})]$$
$$+P(S\text{В}, a\text{В} = R, S\text{Р})[\text{R}(S\text{В}, a\text{В} = R, S\text{Р}) + V_i^*(S\text{Р})]$$
$$+P(S\text{В}, a\text{В} = R, S\text{П})[\text{R}(S\text{В}, a\text{В} = R, S\text{П}) + V_i^*(S\text{П})]$$
$$+P(S\text{В}, a\text{В} = C, S\text{П})[\text{R}(S\text{В}, a\text{В} = C, S\text{П}) + V_i^*(S\text{П})]$$
$$+P(S\text{В}, a\text{В} = C, S\text{Д})[\text{R}(S\text{В}, a\text{В} = C, S\text{Д}) + V_i^*(S\text{Д})]$$
$$+P(S\text{В}, a\text{В} = C, S\text{Б})[\text{R}(S\text{В}, a\text{В} = C, S\text{Б}) + V_i^*(S\text{Б})]\}. \quad (3.39)$$

5. Функция полезности для состояния SБ (Блокировка).

$$V_{i+1}^*(s = S\text{Б}) = \text{R}(S\text{Б}). \quad (3.40)$$

Функции полезности для модели off-line на рисунке 3.12(б) включает набор состояний SР (разведка), SП (подготовка), SД (доступ), SВ (выполнение) и SБ (блокировка). Учитываются их связи по типам действий: нелегальное взаимодействие (D), легальное взаимодействие (C) (3.41 – 3.45).

1. Функция полезности для состояния SР (Разведка). От состояния SР исходит действие аР в состояния SБ, SП, SД, SВ.

$$V_{i+1}^{*}(s = SP) = \mathrm{R}(S\mathrm{P}) + \gamma\max\{P(S\mathrm{P}, a\mathrm{P} = D, S\mathrm{П})[\mathrm{R}(S\mathrm{P}, a\mathrm{P} = D, S\mathrm{П}) + V_i^{*}(S\mathrm{П})]$$
$$+P(S\mathrm{P}, a\mathrm{P} = D, S\mathrm{Д})[\mathrm{R}(S\mathrm{P}, a\mathrm{P} = D, S\mathrm{Д}) + V_i^{*}(S\mathrm{Д})]$$
$$+P(S\mathrm{P}, a\mathrm{P} = D, S\mathrm{В})[\mathrm{R}(S\mathrm{P}, a\mathrm{P} = D, S\mathrm{В}) + V_i^{*}(S\mathrm{В})]$$
$$+P(S\mathrm{P}, a\mathrm{P} = D, S\mathrm{Б})[\mathrm{R}(S\mathrm{P}, a\mathrm{P} = D, S\mathrm{Б}) + V_i^{*}(S\mathrm{Б})]$$
$$+P(S\mathrm{P}, a\mathrm{P} = C, S\mathrm{П})[\mathrm{R}(S\mathrm{P}, a\mathrm{P} = C, S\mathrm{П}) + V_i^{*}(S\mathrm{П})]$$
$$+P(S\mathrm{P}, a\mathrm{P} = C, S\mathrm{Д})[\mathrm{R}(S\mathrm{P}, a\mathrm{P} = C, S\mathrm{Д}) + V_i^{*}(S\mathrm{Д})]$$
$$+P(S\mathrm{P}, a\mathrm{P} = C, S\mathrm{В})[\mathrm{R}(S\mathrm{P}, a\mathrm{P} = C, S\mathrm{В}) + V_i^{*}(S\mathrm{В})]$$
$$+P(S\mathrm{P}, a\mathrm{P} = C, S\mathrm{Б})[\mathrm{R}(S\mathrm{P}, a\mathrm{P} = C, S\mathrm{Б}) + V_i^{*}(S\mathrm{Б})]\}. \qquad (3.41)$$

2. Функция полезности для состояния SП (Подготовка). От состояния SП исходит действие аП в состояния SБ, SП, SД, SВ.

$$V_{i+1}^{*}(s = S\mathrm{П}) = R(S\mathrm{П}) + \gamma\max\{P(S\mathrm{П}, a\mathrm{П} = D, S\mathrm{Д})[R(S\mathrm{П}, a\mathrm{П} = D, S\mathrm{Д})$$
$$+\gamma Vi * (S\mathrm{Д})] + P(S\mathrm{П}, a\mathrm{П} = D, S\mathrm{В})[\mathrm{R}(S\mathrm{П}, a\mathrm{P} = D, S\mathrm{В})$$
$$+\gamma Vi * (S\mathrm{В})] + P(S\mathrm{П}, a\mathrm{П} = D, S\mathrm{Б})[\mathrm{R}(S\mathrm{П}, a\mathrm{P} = D, S\mathrm{Б})$$
$$+\gamma Vi * (S\mathrm{Б})] + P(S\mathrm{П}, a\mathrm{P} = C, S\mathrm{Д})[\mathrm{R}(S\mathrm{П}, a\mathrm{P} = C, S\mathrm{Д})$$
$$+\gamma Vi * (S\mathrm{Д})] + P(S\mathrm{П}, a\mathrm{P} = C, S\mathrm{В})[\mathrm{R}(S\mathrm{П}, a\mathrm{P} = C, S\mathrm{В})$$
$$+\gamma Vi * (S\mathrm{В})] + P(S\mathrm{П}, a\mathrm{P} = C, S\mathrm{Б})[\mathrm{R}(S\mathrm{П}, a\mathrm{P} = C, S\mathrm{Б})$$
$$+\gamma Vi * (S\mathrm{Б})])\}. \qquad (3.42)$$

3. Функция полезности для состояния SД (Доступ). От состояния SД исходит действие аД в состояния SБ, SВ.

$$V_{i+1}^{*}(s = S\mathrm{Д}) = R(S\mathrm{Д}) + \gamma max\{(S\mathrm{Д}, a\mathrm{Д} = D, S\mathrm{В})[R(S\mathrm{Д}, a\mathrm{P} = D, S\mathrm{В}) + V_i^{*}(S\mathrm{В})]$$
$$+P(S\mathrm{Д}, a\mathrm{Д} = C, S\mathrm{В})[\mathrm{R}(S\mathrm{Д}, a\mathrm{Д} = C, S\mathrm{В}) + V_i^{*}(S\mathrm{В})]$$
$$+P(S\mathrm{Д}, a\mathrm{Д} = D, S\mathrm{Б})[\mathrm{R}(S\mathrm{Д}, a\mathrm{Д} = D, S\mathrm{Б}) + V_i^{*}(S\mathrm{Б})]$$

$$+P(S\text{Д}, a\text{Д} = C, S\text{Б})[\text{R}(S\text{Д}, a\text{Д} = C, S\text{Б}) + V_i^*(S\text{Б})]\}. \qquad (3.43)$$

4. Функция полезности для состояния *S*В (Выполнение). От состояния *S*В исходит действие аВ в состояние *S*Б, не связано с *S*П, *S*Д, *S*В, *S*Р.

$$V_{i+1}^*\ (s = S\text{В}) = \text{R}(S\text{В}) + \gamma max\{P(S\text{В}, a\text{В} = D, S\text{Б})[\text{R}(S\text{В}, a\text{В} = D, S\text{Б}) + V_i^*(S\text{Б})]$$
$$+P(S\text{В}, a\text{В} = C, S\text{Б})[\text{R}(S\text{В}, a\text{В} = C, S\text{Б}) + V_i^*(S\text{Б})]\}. \qquad (3.44).$$

5. Функция полезности для состояния SБ (Блокировка).

$$V_{i+1}^*(s = S\text{Б}) = \text{R}(S\text{Б})\ . \qquad (3.45).$$

Модели позволяют учитывать более сложные зависимости между переменными и обеспечивают большую гибкость при анализе состояний. Способы применения моделей предполагают следующее:

1. Используется для анализа отдельного состояния (тактических возможностей злоумышленника) сценария атаки;

2. Используется для анализа безопасности (тактических возможностей злоумышленника) системы в целом (системы описывается приведенными состояниями).

Ограничение при применении моделей упрощенного типа следующие:

1. Первые два подтипа модели предусматривают статичное распределение состояний по зонам контроля.

2. В третьем подтипе моделей существует одна зона контроля (одновременно злоумышленника и стороны защиты), при этом проблема смещения зон отсутствует. В этом случае можно рассматривать нападение как совокупность разных атак.

3. Тип нарушителя – внешний (как наиболее полный). Атакующая последовательность при этом всегда начинается с состояния SP (Разведка).

В итоге можно привести общие формулы полезности (представлены в формулах (3.46 - 3.50)) для всех моделей, которые отражают специфику подхода к построению моделей атак с учетом ограничений переходов в состояния атаки, представленной как последовательность обобщенных тактик MITRE ATLAS (функции полезности для модели on-line и off-line с учетом специфики доступности состояний $s' \in \{S\text{П}, S\text{Д}, S\text{В}, S\text{Р}, S\text{Б}\}$, приведённой на рисунке (3.12)).

1. Функция полезности для состояния SР (Разведка).

$$V_{i+1}^{*} * (S\text{Р}) = \text{R}(S\text{Р}) + \gamma max_{a\text{Р}} E[\text{R}(S\text{Р}, a\text{Р}, s') + V_i^{*}(s')] . \quad (3.46)$$

2. Функция полезности для состояния SП (Подготовка).

$$V_{i+1}^{*}(S\text{П}) = \text{R}(S\text{П}) + \gamma max_{a\text{П}} E[\text{R}(S\text{П}, a\text{П}, s') + V_i^{*}(s')]. \quad (3.47)$$

3. Функция полезности для состояния SД (Доступ).

$$V_{i+1}^{*}(S\text{Д}) = \text{R}(S\text{Д}) + \gamma max_{a\text{Д}} E[\text{R}(S\text{Д}, a\text{Д}, s') + V_i^{*}(s')]. \quad (3.48)$$

4. Функция полезности для состояния SВ (Выполнение).

$$V_{i+1}^{*}(S\text{В}) = \text{R}(S\text{В}) + \gamma max_{a\text{В}} E[\text{R}(S\text{В}, a\text{В}, s') + V_i^{*}(s')]. \quad (3.49)$$

5. Функция полезности для состояния SБ (Блокировка)

$$V_{i+1}^{*}(s = S\text{Б}) = \text{R}(S\text{Б}). \quad (3.50)$$

Обобщенные формулы при соблюдении способа их формироваия можно модифицировать с учетом требований новых связей или появления новых состояний, которые следует учитывать в связи с тем, что базы описания методов

атак постоянно дополняются, а технические инфраструктуры расширяются и обновляются.

3.4 Модель с учетом всех тактик методики описания сценария атаки

В данном случае используется общая классификация тактик MITRE ATLAS (также могут использоваться при необходимости тактики и техники Методики ФСТЭК (и при условии сопряжения техник возможно применение MITRE ATT@CK)).

Для получения полного графа состояний атаки требуется, чтобы механизмы ПИИ позволяли реализовать состояние S1 (Тактика 1). В противном случае количество состояний для моделей off-line уменьшится, в отличие от моделей типа on-line. Модель МППР для сценария атаки на модели и данные ПИИ предполагает, что количество типов и описаний (техник) состояний ПИИ находится в соответствии с количеством тактик (состояний) ПИИ методик построения с сценария атак (MIRTE ATLAS или Методики ФСТЭК). Модели взаимодействия приведены на рисунке 3.13, 3.14. Специфика на прямые связи между состояниями (действие соединяет два состояния без промежуточных состояний) предполагает следующие:

- не исключается переход между состояниями при отсутствии сетевых связей;
- не исключается переход между состояниями при отсутствии возможности сопряжения уязвимостей.

В соответствии с классификацией состояний атакуемой системы (доступные для прямого и контролируемого воздействия злоумышленнику при атаке на ПИИ) вводятся ограничения на переходы между состояниями моделей. При определении перехода между состояниями требуется учитывать следующие:

1. Режим применения ПИИ (при учете особенностей методики MIRTE ATLAS): эксплуатация, обучение, объединение режимов применения.

2. Наличие или отсутствие контролируемого (управляемого) злоумышленником взаимодействия (доступа) к ПИИ и ее компонентам в зависимости от типа функционирования и архитектуры ПИИ:

− наличие или отсутствие доступа во время атаки к процессам вычислений, интерфейсам вычислительных моделей (вычислительным моделям) вычислениям;

− наличие или отсутствие доступа во время атаки к датасетам, данным;

− наличие или отсутствие доступа во время атаки инфраструктурным компонентам.

При формировании моделей в данном случае допустимо учитывать ограничения зон доступа (приведены в описании упрощенной модели), влияющих на специфику переходов между состояниями в модели, если инфраструктура системы может точно описываться.

Также при построении связей между состояниями следует учитывать следующие: связи коспонетов системы, топологию сети, сопряжение уязвимостей и другие параметры, приведенные в функции R. При выборе стратегии атаки в рамках моделирования с использованием МППР также учитываются параметры «вознаграждения» и «затрат».

С учётом ограничений модели целесообразно применять в этом случае второй и третий режим описания действий модели. В этом случае учитываются: типы множеств действий (*D, C, R*) и уточняющие действия (тактики) (актуальность тактики зависит от архитектурных особенностей ПИИ). Таким образом, любое действие $\forall a_i \in AI$, $\exists S_i \in \{S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15\}$: $a \in Si \subseteq A$ или $\exists S_i \in \{S1, \ldots, Sn\}$: $a \in Si \subseteq A$, где Sn – последняя актуальная тактика при обновлении MITRE ATLAS, $S_i \cap (D \cup C) = \emptyset$ или $a_i \in R \subseteq SB$.

Ограничения модели при применении:

- требует детализации атакующих действий по используемым Методикам и базам знаний MITRE при использования третьего режима описания действий;

- используется сопоставление действий, уязвимостей, локальных идентификаторов (для второго режима и третьего);

- состояния абстрактно описывают безопасность системы.

Ключевым фактором в построении модели является определение величин ценности (полезности) состояний, что позволяет определить наилучшую последовательность состояний атаки и действий с точки зрения злоумышленника с учетом разной степени детализации действий. Общие формулы (ценности) полезности для всех состояний (трактуемых как тактики MITRE ATLAS) следующие (3.51 – 3.52):

1. Для режима on-line (используется множество S_{all}={$S1$, $S2$,..., Sn, $SБ$}, то есть учитываются все состояния S_i в соответствии с числом актуальных тактик базы MITRE ATLAS (на период исследования в базе существует 15 тактик), а также добавляемого состояния блокировки $SБ$, при этом вводятся состояния S_k={$S1$, $S2$,..., S_k}, число которых ограничено спецификой действия (перехода) R (сброс)):

$$V_{i+1}^*(s = S_k) = max\begin{Bmatrix}\sum_{s\prime \in S_{all}} P(S_k, D, s')[R(S_k, D, s') + \gamma V_i^*(s')] \\ \sum_{s\prime \in S_{all}} P(S_k, C, s')[R(S_k, C, s') + \gamma V_i^*(s')] \\ \sum_{s\prime \in S_k} P(S_k, R, s')[R(S_k, R, s') + \gamma V_i^*(s')]\end{Bmatrix}. \quad (3.51)$$

2. Для режима off-line (используется $S_{current \to end}$ – множество состояний, начинающееся с текущего и включающее все последующие состояния, а также конечное состояние SБ, при этом для Sk подразумевается k∈{1,2,...,n,Б}):

$$V_{i+1}^*(s = S_k) = max\begin{Bmatrix}\sum_{s\prime \in S_{current \to end}} P(S_k, D, s')[R(S_k, D, s') + \gamma V_i^*(s')] \\ \sum_{s\prime \in S_{current \to end}} P(S_k, C, s')[R(S_k, C, s') + \gamma V_i^*(s')]\end{Bmatrix}. \quad (3.52)$$

Функции полезности для режима on-line (представлены в формулах с (3.53) по (3.67)) с учетом всех состояний и доступных действий (Рисунок 3.13) описываются пятнадцатью тактиками (представлены как состояния по актуальным тактикам MITRE ATLAS на момент исследования, допускается изменение их количества в модели при изменении базы тактик) и принадлежащими им подмножествами техник (выступают в роли действий), а также дополнительным

действием сброса R. При этом следует учитывать то, что множество состояний и множество действий при подобном подходе может регулироваться согласно методике отбора (по советующим таблицам, приведенным в главе 2). Таким образом, при моделировании необязательно использовать все тактики и, соответственно, техники. Однако при условии неполного знания инфраструктуры или при нестабильности инфраструктуры учитывать предполагаемые состояния и тактики (техники), как потенциальные, данный способ моделирования позволяет. Это дает возможность оценить потенциал возможных атак.

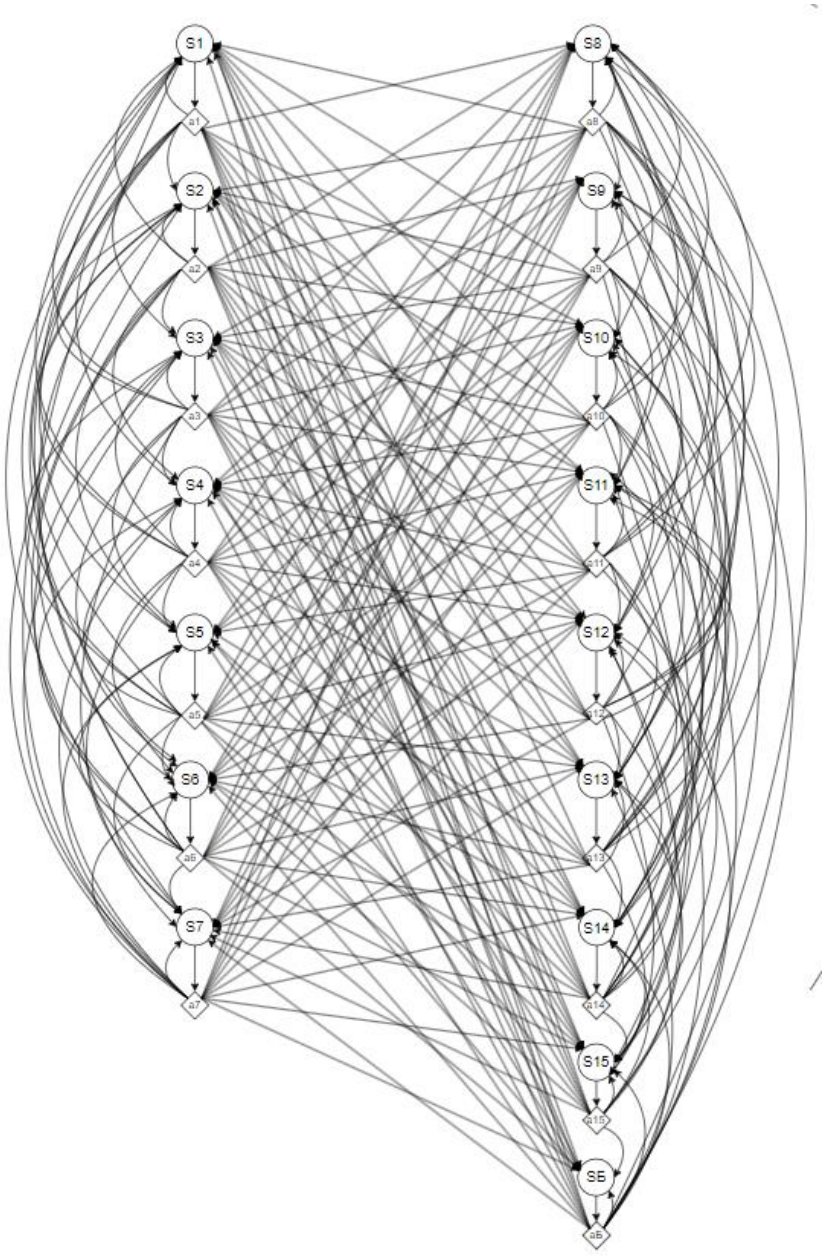


Рисунок 3.13 - Модель МППР с учетом всех возможных состояний on-line

Функции полезности состояний S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15 и SБ (по актуальным тактикам MITRE ATLAS) с учетом типов действий *D, C, R* (3.53 – 3.67):

1. Функция полезности для состояния S_1, действия *a1* может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_1) = max\begin{Bmatrix} \sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_1, D, s')[\mathrm{R}\,(S_1, D, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_1, C, s')[\mathrm{R}(S_1, C, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1\}} P\,(S_1, R, s')[\mathrm{R}(S_1, R, s') + \gamma V_i^{*}(s') \end{Bmatrix}. \quad (3.53)$$

2. Функция полезности для состояния S_2, действия *a2* может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_2) = max\begin{Bmatrix} \sum\limits_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_2, D, s')[\mathrm{R}\,(S_2, D, s') + \gamma V_i^{*}(s')] \\ \sum\limits_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_2, C, s')[\mathrm{R}(S_2, C, s') + \gamma V_i^{*}(s')] \\ \sum\limits_{s' \in \{S_1,S_2\}} P\,(S_2, R, s')[\mathrm{R}(S_2, R, s') + \gamma V_i^{*}(s') \end{Bmatrix}. (3.54)$$

3. Функция полезности для состояния S_3, действия *a3* может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_3) = max\begin{Bmatrix} \sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_3, D, s')[\mathrm{R}\,(S_3, D, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_3, C, s')[\mathrm{R}(S_3, C, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,S_2,S_3\}} P\,(S_3, R, s')[\mathrm{R}(S_3, R, s') + \gamma V_i^{*}(s') \end{Bmatrix}. \quad (3.55)$$

4. Функция полезности для состояния S_4, действия *a*4 может быть представлена следующим образом:

$$V_{i+1}^{*}(s=S_4)=max\begin{Bmatrix}\sum_{s'\in\{S_1,\dots,S_{15},S_Б\}}P(S_4,D,s')[\mathrm{R}\,(S_4,D,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_1,\dots,S_{15},S_Б\}}P(S_4,C,s')[\mathrm{R}(S_4,C,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_1,S_2,S_3,S_4\}}P(S_4,R,s')[\mathrm{R}(S_4,R,s')+\gamma V_i^{*}(s')\end{Bmatrix}. \quad (3.56)$$

5. Функция полезности для состояния S_5, действия $a5$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s=S_5)=max\begin{Bmatrix}\sum_{s'\in\{S_1,\dots,S_{15},S_Б\}}P(S_5,D,s')[\mathrm{R}\,(S_5,D,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_1,\dots,S_{15},S_Б\}}P(S_5,C,s')[\mathrm{R}(S_5,C,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_1,S_2,S_3,S_4,S_5\}}P(S_5,R,s')[\mathrm{R}\,(S_5,R,s')+\gamma V_i^{*}(s')\end{Bmatrix}. \quad (3.57)$$

6. Функция полезности для состояния S_6, действия $a6$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s=S_6)=max\begin{Bmatrix}\sum_{s'\in\{S_1,\dots,S_{15},S_Б\}}P(S_6,D,s')[\mathrm{R}\,(S_6,D,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_1,\dots,S_{15},S_Б\}}P(S_6,C,s')[\mathrm{R}(S_6,C,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_1,S_2,S_3,S_4,S_5,S_6\}}P(S_6,R,s')[\mathrm{R}\,(S_6,R,s')+\gamma V_i^{*}(s')\end{Bmatrix}. \quad (3.58)$$

7. Функция полезности для состояния S_7, действия $a7$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s=S_7)=max\begin{Bmatrix}\sum_{s'\in\{S_1,\dots,S_{15},S_Б\}}P(S_7,D,s')[\mathrm{R}\,(S_7,D,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_1,\dots,S_{15},S_Б\}}P(S_7,C,s')[\mathrm{R}(S_7,C,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_1,S_2,S_3,S_4,S_5,S_6,S_7\}}P(S_7,R,s')[\mathrm{R}\,(S_7,R,s')+\gamma V_i^{*}(s')\end{Bmatrix}. \quad (3.59)$$

8. Функция полезности для состояния S_8, действия $a8$:

$$V_{i+1}^{*}(s=S_8)=max\begin{Bmatrix}\sum_{s'\in\{S_1,\dots,S_{15},S_Б\}}P(S_8,D,s')[\mathrm{R}\,(S_8,D,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_1,\dots,S_{15},S_Б\}}P(S_8,C,s')[\mathrm{R}(S_8,C,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_1,S_2,\dots,S_8\}}P(S_8,R,s')[\mathrm{R}\,(S_8,R,s')+\gamma V_i^{*}(s')\end{Bmatrix}. \quad (3.60)$$

9. Функция полезности для состояния S_9, действия $a9$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_9) = max\begin{Bmatrix} \sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_9, D, s')[\mathrm{R}\,(S_9, D, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_9, C, s')[\mathrm{R}(S_9, C, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,S_2,\dots,S_9\}} P(S_9, R, s')[\mathrm{R}\,(S_9, R, s') + \gamma V_i^{*}(s') \end{Bmatrix}. \quad (3.61)$$

10. Функция полезности для состояния S_{10}, действия $a10$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_{10}) = max\begin{Bmatrix} \sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_{10}, D, s')[\mathrm{R}\,(S_{10}, D, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_{10}, C, s')[\mathrm{R}(S_{10}, \mathrm{C}, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,S_2,\dots,S_{10}\}} P(S_{10}, R, s')[\mathrm{R}\,(S_{10}, R, s') + \gamma V_i^{*}(s') \end{Bmatrix}. \quad (3.62)$$

11. Функция полезности для состояния S_{11}, действия $a11$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_{11}) = max\begin{Bmatrix} \sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_{11}, D, s')[\mathrm{R}\,(S_{11}, D, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_{11}, C, s')[\mathrm{R}(S_{11}, C, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,S_2,\dots S_{11}\}} P(S_{11}, R, s')[\mathrm{R}\,(S_{11}, R, s') + \gamma V_i^{*}(s') \end{Bmatrix}. \quad (3.63)$$

12. Функция полезности для состояния S_{12}, действия $a12$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_{12}) = max\begin{Bmatrix} \sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_{12}, D, s')[\mathrm{R}\,(S_{12}, D, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_{12}, C, s')[\mathrm{R}(S_{12}, C, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,S_2,\dots,S_{12}\}} P(S_{12}, R, s')[\mathrm{R}\,(S_{12}, R, s') + \gamma V_i^{*}(s')] \end{Bmatrix}. \quad (3.64)$$

13. Функция полезности для состояния S_{13}, действия $a13$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_{13}) = max\begin{Bmatrix}\sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_{13}, D, s')[\mathrm{R}\,(S_{13}, D, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_{13}, C, s')[\mathrm{R}(S_{13}, C, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,S_2,\dots,S_{13}\}} P(S_{13}, R, s')[\mathrm{R}\,(S_{13}, R, s') + \gamma V_i^{*}(s')]\end{Bmatrix}. (3.65)$$

14. Функция полезности для состояния S_{14}, действия $a14$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_{14}) = max\begin{Bmatrix}\sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_{14}, D, s')[\mathrm{R}\,(S_{14}, D, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_{14}, C, s')[\mathrm{R}(S_{14}, C, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,S_2,\dots,S_{14}\}} P(S_{14}, R, s')[\mathrm{R}\,(S_{14}, R, s') + \gamma V_i^{*}(s')]\end{Bmatrix}. (3.66)$$

15. Функция полезности для состояния S_{15}, действия $a15$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_{15}) = max\begin{Bmatrix}\sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_{15}, D, s')[\mathrm{R}\,(S_{14}, D, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_{15}, C, s')[\mathrm{R}(S_{14}, C, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,S_2,\dots,S_{15}\}} P(S_{15}, R, s')[\mathrm{R}\,(S_{14}, R, s') + \gamma V_i^{*}(s')]\end{Bmatrix}. (3.67)$$

16. Функция полезности для состояния $S_Б$, действия aБ может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_Б) = max\begin{Bmatrix}\sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_Б, D, s')[\mathrm{R}\,(S_Б, D, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,\dots,S_{15},S_Б\}} P(S_Б, C, s')[\mathrm{R}(S_Б, C, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_1,S_2,\dots,S_Б\}} P(S_Б, R, s')[\mathrm{R}\,(S_Б, R, s') + \gamma V_i^{*}(s')]\end{Bmatrix}. (3.68)$$

Далее следует рассмотреть функции полезности для модели режима off-line (Рисунок 3.14). Она представляет собой более упрощенный вариант модели on-line и не предполагает наличие действий R (сброса), также исключаются боковые

смещения. Такая модель представляет действия a_i как уже осуществившиеся и конечные результаты массива воздействий, связанных с компрометацией системы ИИ.

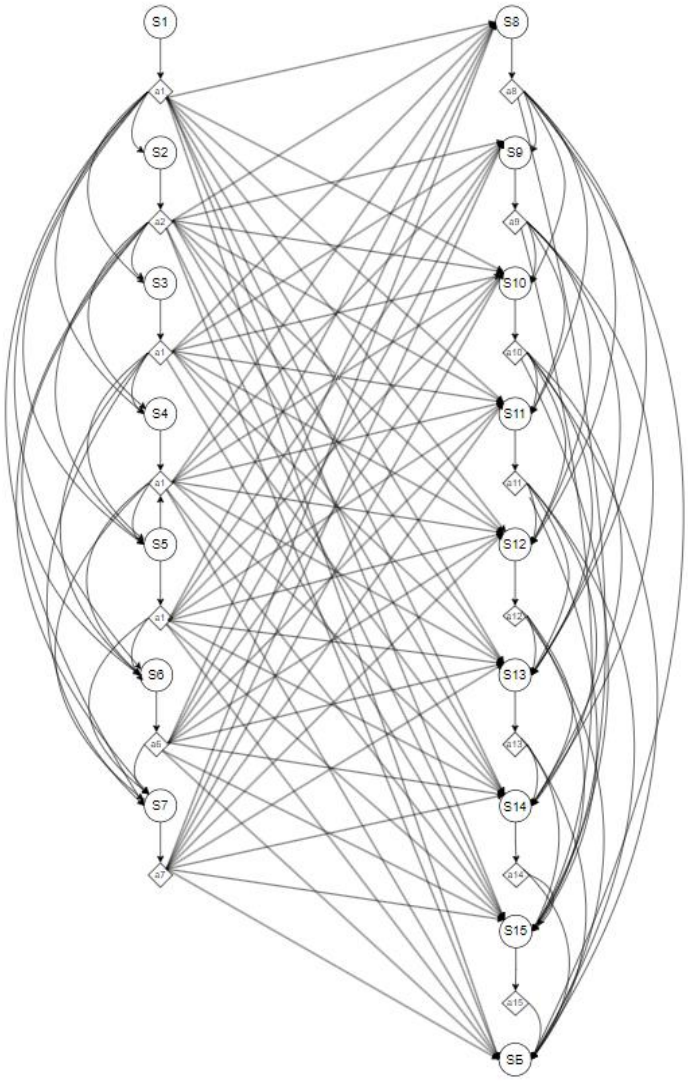


Рисунок 3.14 - Модель МППР с учетом всех возможных состояний off-line

Функции полезности (представлены в формулах с (3.69) по (3.84)) для режима off-line на рисунке 3.14 с учетом всех состояний и доступных действий (Рисунок 3.14) описываются также приведенными в MITRE ATLAS тактиками и принадлежащими им подмножествами техник с добавлением состояния блокировки *S*Б.

1. Функция полезности для состояния S_1, действия $a1$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s=S_1)=max\begin{Bmatrix}\sum_{s'\in\{S_1,\dots,S_{15},S_Б\}}P(S_1,D,s')[\mathrm{R}\,(S_1,D,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_1,\dots,S_{15},S_Б\}}P(S_1,C,s')[\mathrm{R}(S_1,C,s')+\gamma V_i^{*}(s')]\end{Bmatrix}. \quad (3.69)$$

2. Функция полезности для состояния S_2, действия $a2$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s=S_2)=max\begin{Bmatrix}\sum_{s'\in\{S_2,\dots,S_{15},S_Б\}}P(S_2,D,s')[\mathrm{R}\,(S_2,D,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_2,\ \dots,S_{15},S_Б\}}P(S_2,C,s')[\mathrm{R}(S_2,C,s')+\gamma V_i^{*}(s')]\end{Bmatrix}. \quad (3.70)$$

3. Функция полезности для состояния S_3, действия $a3$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s=S_3)=max\begin{Bmatrix}\sum_{s'\in\{S_3,\dots,S_{15},S_Б\}}P(S_3,D,s')[\mathrm{R}\,(S_3,D,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_3,\dots,S_{15},S_Б\}}P(S_3,C,s')[\mathrm{R}(S_3,C,s')+\gamma V_i^{*}(s')]\end{Bmatrix}. \quad (3.71)$$

4. Функция полезности для состояния S_4, действия $a4$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s=S_4)=max\begin{Bmatrix}\sum_{s'\in\{S_4,\dots,S_{15},S_Б\}}P(S_4,D,s')[\mathrm{R}\,(S_4,D,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_4,\dots,S_{15},S_Б\}}P(S_4,C,s')[\mathrm{R}(S_4,C,s')+\gamma V_i^{*}(s')]\end{Bmatrix}. \quad (3.72)$$

5. Функция полезности для состояния S_5, действия $a5$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s=S_5)=max\begin{Bmatrix}\sum_{s'\in\{S_5,\dots,S_{15},S_Б\}}P(S_5,D,s')[\mathrm{R}\,(S_5,D,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_5,\ \dots,S_{15},S_Б\}}P(S_5,C,s')[\mathrm{R}(S_5,C,s')+\gamma V_i^{*}(s')]\end{Bmatrix}. \quad (3.73)$$

6. Функция полезности для состояния S_6, действия $a6$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s=S_6)=max\begin{Bmatrix}\sum_{s'\in\{S_6,\dots,S_{15},S_Б\}}P(S_6,D,s')[\mathrm{R}(S_6,D,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_6,\dots,S_{15},S_Б\}}P(S_6,C,s')[\mathrm{R}(S_6,C,s')+\gamma V_i^{*}(s')]\end{Bmatrix}. \quad (3.74)$$

7. Функция полезности для состояния S_7, действия $a7$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s=S_i)=max\begin{Bmatrix}\sum_{s'\in\{S_7,\dots,S_{15},S_Б\}}P(S_7,D,s')[\mathrm{R}(S_7,D,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_7,\dots,S_{15},S_Б\}}P(S_7,C,s')[\mathrm{R}(S_7,C,s')+\gamma V_i^{*}(s')]\end{Bmatrix}. \quad (3.75)$$

8. Функция полезности для состояния S_8, действия $a8$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s=S_8)=max\begin{Bmatrix}\sum_{s'\in\{S_8,\dots,S_{15},S_Б\}}P(S_8,D,s')[\mathrm{R}(S_8,D,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_8\dots,S_{15},S_Б\}}P(S_8,C,s')[\mathrm{R}(S_8,C,s')+\gamma V_i^{*}(s')]\end{Bmatrix}. \quad (3.76)$$

9. Функция полезности для состояния S_9, действия $a9$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s=S_9)=max\begin{Bmatrix}\sum_{s'\in\{S_9,\dots,S_{15},S_Б\}}P(S_9,D,s')[\mathrm{R}(S_9,D,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_9,\dots,S_{15},S_Б\}}P(S_9,C,s')[\mathrm{R}(S_9,C,s')+\gamma V_i^{*}(s')]\end{Bmatrix}. \quad (3.77)$$

10. Функция полезности для состояния S_{10}, действия $a10$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s=S_{10})=max\begin{Bmatrix}\sum_{s'\in\{S_{10},\dots,S_{15},S_Б\}}P(S_{10},D,s')[\mathrm{R}(S_{10},D,s')+\gamma V_i^{*}(s')]\\ \sum_{s'\in\{S_{10},\dots,S_{15},S_Б\}}P(S_{10},C,s')[\mathrm{R}(S_{10},C,s')+\gamma V_i^{*}(s')]\end{Bmatrix}. \quad (3.78)$$

11. Функция полезности для состояния S_{11}, действия $a11$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_{11}) = max\begin{Bmatrix}\sum_{s' \in \{S_{11},\ldots,\, S_{15},S_{Б}\}} P(S_{11}, D, s')[\mathrm{R}\,(S_{11}, D, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_{11},\ldots,\, S_{15},S_{Б}\}} P(S_{11}, C, s')[\mathrm{R}(S_{11}, C, s') + \gamma V_i^{*}(s')]\end{Bmatrix}. \quad (3.79)$$

12. Функция полезности для состояния S_{12}, действия $a12$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_{12}) = max\begin{Bmatrix}\sum_{s' \in \{S_{12},\ldots,\, S_{15},S_{Б}\}} P(S_{12}, D, s')[\mathrm{R}\,(S_{12}, D, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_{12},\ldots,\, S_{15},S_{Б}\}} P(S_{12}, C, s')[\mathrm{R}(S_{12}, C, s') + \gamma V_i^{*}(s')]\end{Bmatrix}. \quad (3.80)$$

13. Функция полезности для состояния S_{13}, действия $a13$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_{13}) = max\begin{Bmatrix}\sum_{s' \in \{S_{13},\ldots,\, S_{15},S_{Б}\}} P(S_{13}, D, s')[\mathrm{R}\,(S_{13}, D, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_{13},\ldots,\, S_{15},S_{Б}\}} P(S_{13}, C, s')[\mathrm{R}(S_{13}, C, s') + \gamma V_i^{*}(s')]\end{Bmatrix}. \quad (3.81)$$

14. Функция полезности для состояния S_{14}, действия $a14$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_{14}) = max\begin{Bmatrix}\sum_{s' \in \{S_{14},S_{15},S_{Б}\}} P(S_{14}, D, s')[\mathrm{R}\,(S_{14}, D, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_{14},S_{15},S_{Б}\}} P(S_{14}, C, s')[\mathrm{R}(S_{14}, C, s') + \gamma V_i^{*}(s')]\end{Bmatrix}. \quad (3.82)$$

15. Функция полезности для состояния S_{15}, действия $a15$ может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_{15}) = max\begin{Bmatrix}\sum_{s' \in \{\, S_{15},S_{Б}\}} P(S_{14}, D, s')[\mathrm{R}\,(S_{14}, D, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_{15},S_{Б}\}} P(S_{14}, C, s')[\mathrm{R}(S_{14}, C, s') + \gamma V_i^{*}(s')]\end{Bmatrix}. \quad (3.83)$$

16. Функция полезности для состояния $S_{Б}$, действия aБ может быть представлена следующим образом:

$$V_{i+1}^{*}(s = S_{\text{Б}}) = max\begin{Bmatrix}\sum_{s' \in \{S_{\text{Б}}\}} P(S_{\text{Б}}, D, s')[\text{R}\,(S_{\text{Б}}, D, s') + \gamma V_i^{*}(s')] \\ \sum_{s' \in \{S_{\text{Б}}\}} P(S_{\text{Б}}, C, s')[\text{R}(S_{\text{Б}}, C, s') + \gamma V_i^{*}(s')]\end{Bmatrix}. \quad (3.84)$$

Требуется рассмотреть ограничения по действиям. Приводимые значения вознаграждения должны быть масштабированы в заранее определенном диапазоне, чтобы избежать поверхностного эффекта больших вознаграждений, и они должны быть значимыми, представляя ценность, полученную при переходе из одного состояния в другое. При атаке, когда осуществляются компрометирующие действия, злоумышленник может оказаться в заблокированном состоянии, то есть величина выгоды стремится к нулевым значениям. Моделирование в соответствии с приведенным порядком вычисления $V_{i+1}^{*}(s)$ должно показывать компромисс между затратами, действиями и последствиями различных стратегий, доступных злоумышленникам, что в итоге позволяет определить наилучшую стратегию нападения.

Таким образом, в процессе исследования проблемной области были сформированы модели определения наилучших политик нападения злоумышленника для режимов on-line и off-line с учетом методов MITRE ATLAS. Специфика моделирования атак на ПИИ с использованием МППР, учитывая требования оптимальной политики нападения, позволяет более точно построить последовательность атакующих воздействий, связанных с компрометацией модели ИИ, навязыванием ложных обучающих данных, а также последующим модифицированием классифицирования.

3.5 Общая методическая последовательность использования моделей

Последовательность применения модели предполагает следующие шаги:

Шаг 1. Выделяются типы состояний, являющиеся результатом осуществления тактик в соответствии с выбранной методикой описания сценария атаки. Например, воспользоваться методикой ФСТЭК. На основании принципов

построения МППР формируется система уравнений, которая учитывает специфику зависимостей состояний (тактик, приводимых в Методике и MITRE ATLAS), описанных в графе состояний и действий (например, рисунок 3.14), и специфику достижения состояний при осуществлении техник из начального состояния. Это позволяет рассмотреть множество различных моделей атак с учетом того, что источником начальных состояний (начальных атакующих действий) является внешним нарушителем [132].

Шаг 2. Для определения переходов нужно определить сопрягаемые техники различных тактик. На основе анализа функциональных взаимосвязей техник выделяются связи тактик.

Шаг 3. В этом случае предполагается определение типовой модели последовательности атакующих воздействий (ее построение) с учетом вероятности состояний (событий, маркирующих состояние). Имея в виду вышеуказанные цели, сначала должна быть получена оптимальная стратегия МППР для злоумышленника. Затем, для анализа оптимальной стратегии, поведение агента (т.е. злоумышленника) должно быть использовано в модели путем применения оптимальной политики МППР, то есть требуется реализовать следующее:

1) Произвести указание начального (нейтрального) состояния, как начального шага противоборствующего агента.

2) Выбрать действие для базового состояния.

3) На основе выбранного действия определить состояния, в которые можно перейти (S') из состояния s (где S' – это подмножество состояний, которые могут быть достигнуты из s действием).

4) Принимая во внимание вероятности перехода, перейти к следующему состоянию, учитывая значения, определенные для каждой из вероятностей перехода.

5) Получить вознаграждение за переход и добавить его к предыдущим вознаграждениям, а также увеличить количество шагов на один.

6) Если новое состояние является заблокированным, вернуть награды и количество шагов; в противном случае перейти к шагу 2.

В качестве примера реализации указанных шагов можно привести использование модели при испытании устойчивости системы ИИ к атакам отравления (в частности, датасетов) без учета инфраструктурных компонентов ИИ. Реализация шагов примера требует:

1. Обучить рассматриваемую модель на безопасных данных.

2. Вычислить метрики модели.

3. Выбрать тип атакующего воздействия (отравление).

4. Отравить исходные данные с помощью выбранного метода и с разными процентами.

5. Провести обучение модели для каждого отравленного датасета.

6. Вычислить метрики полученных моделей, в том числе ASR (метрика, используемая для оценки уровня угрозы или уязвимости в контексте кибербезопасности).

7. Получить вероятности переходов (при это учитывается предположение того, что качество новой модели определяется не только с помощью метрик, но и с помощью сравнения новых показателей с предыдущими, например, использовать можно только accuracy).

8. Установить максимально возможную стоимость обмана.

9. Вычислить награды.

10. Произвести RL-обучение для выбранной стоимости (повторить несколько раз для построения графиков).

Положительная сторона формируемой методики применения моделей в этом случае - это связь с реальностью (так как отражает реальные данные и учитывает предполагаемые цели), возможность использовать предварительные вычисления (результаты работы поставщиков данных можно использовать в качестве одного из входных значений).

Для реализации моделирования предпочтительно использовать дополнительные программное обеспечение — системы расчета (используются при большом количестве состояний и действий) и системы сбора и анализа данных. Их основная задача - обеспечить предоставление модели вероятностей событий

(маркеров безопасности) и статистические данные по переходам между состояниями. Для применения модели требуется провести подготовительную работу:

1. Определить набор действий для каждого множества, а также то, куда эти действия могут привести. Набор действий определяется, в том числе, в соответствии с MITRE ATLAS.

2. Определить размер наград для каждого действия, а также цену каждого из них. Каждое возможное действие определяется как уязвимость, для которой высчитывается контекстная метрика CVSS.

3. Определить вероятности перехода для каждого действия. По умолчанию вероятности перехода для каждого действия задаётся равномерно. Затем можно использовать другие подходы к определнию вероятностей (на основе статистических данных или на основе вознаграждений).

4. Провести анализ зависимости оптимальной политики от стоимости действий.

Основные шаги алгоритма (порядок реализации алгоритма моделирования конкретного состояния приведен на рисунке 3.15) определения последовательности действий с использованием МППР при атаке на ПИИ включают следующее:

1. Определение набора действий и, следовательно, состояний модели. Следует учитывать особенности определения действий:

– набор действий и состояний определяется в соответствии с MITRE ATLAS (при необходимости в сопряжении с Методикой ФСТЭК), при этом учитывается доступность действий;

– учитывается то, что в состояние, соответствующее технике, можно перейти, выполнив соответствующее действие из предыдущих состояний.

2. Определение размера наград для каждого действия, а также его стоимости (при необходимости) как части награды для каждого из них. Следует учитывать:

– каждое возможное действие сопряжено с уязвимостью, для которой высчитывается контекстная метрика CVSS;

– отдельно задается награда (штраф) за переход в состояние блокировки.

3. Определение переходов, вероятностей переходов между состояниями для каждого действия (вероятность перехода в следующее состояние, вероятность перехода в состояние блокировки).

4. Проведение вычислений. Требуется провести анализ зависимости оптимальной политики от стоимости действий с последующим определением наиболее выгодных для злоумышленника переходов между состояниями.

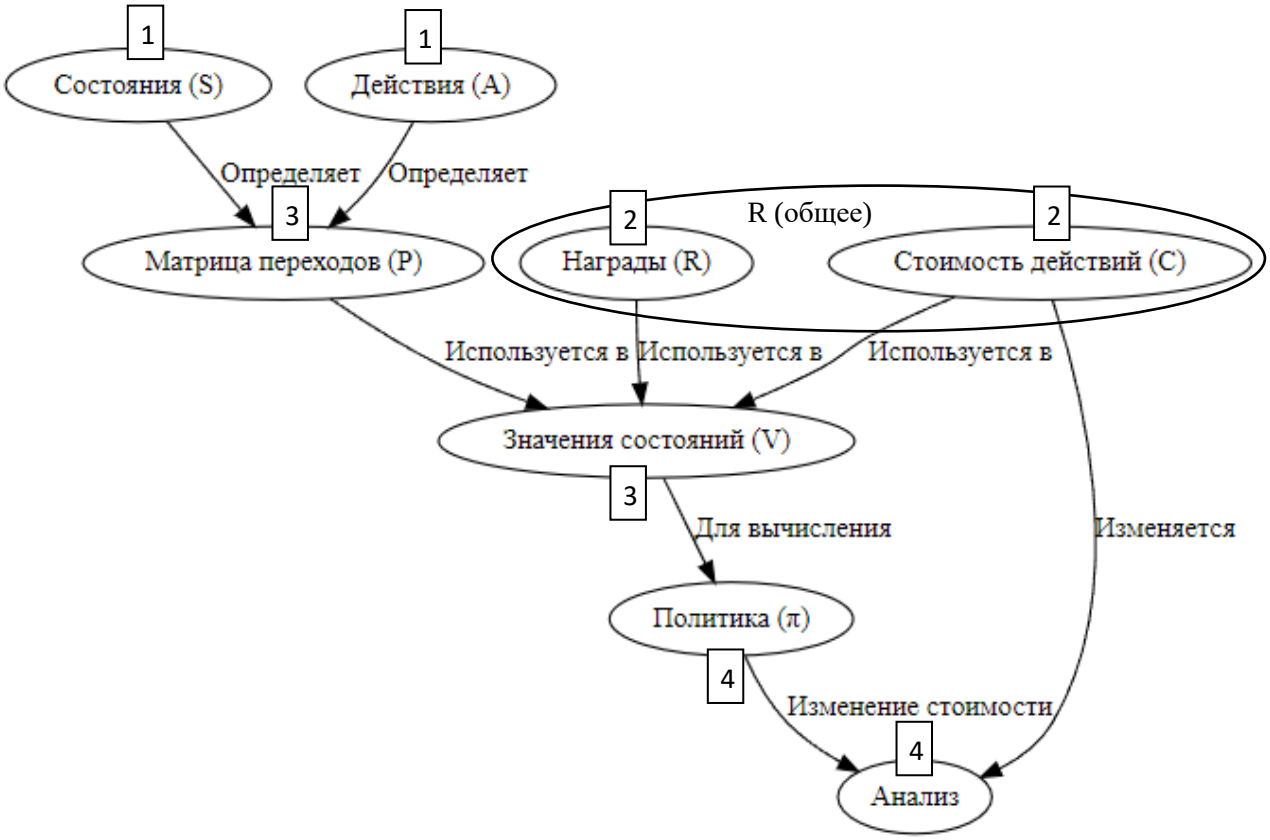


Рисунок 3.15 – Основные шаги общего алгоритма определения наилучшей для злоумышленника последовательности действий и состояний

Общий алгоритм моделирования (пригоден для использования всех типов моделей и режимов моделирования) приведен на рисунке 3.16 в формате диаграммы деятельности (UML).

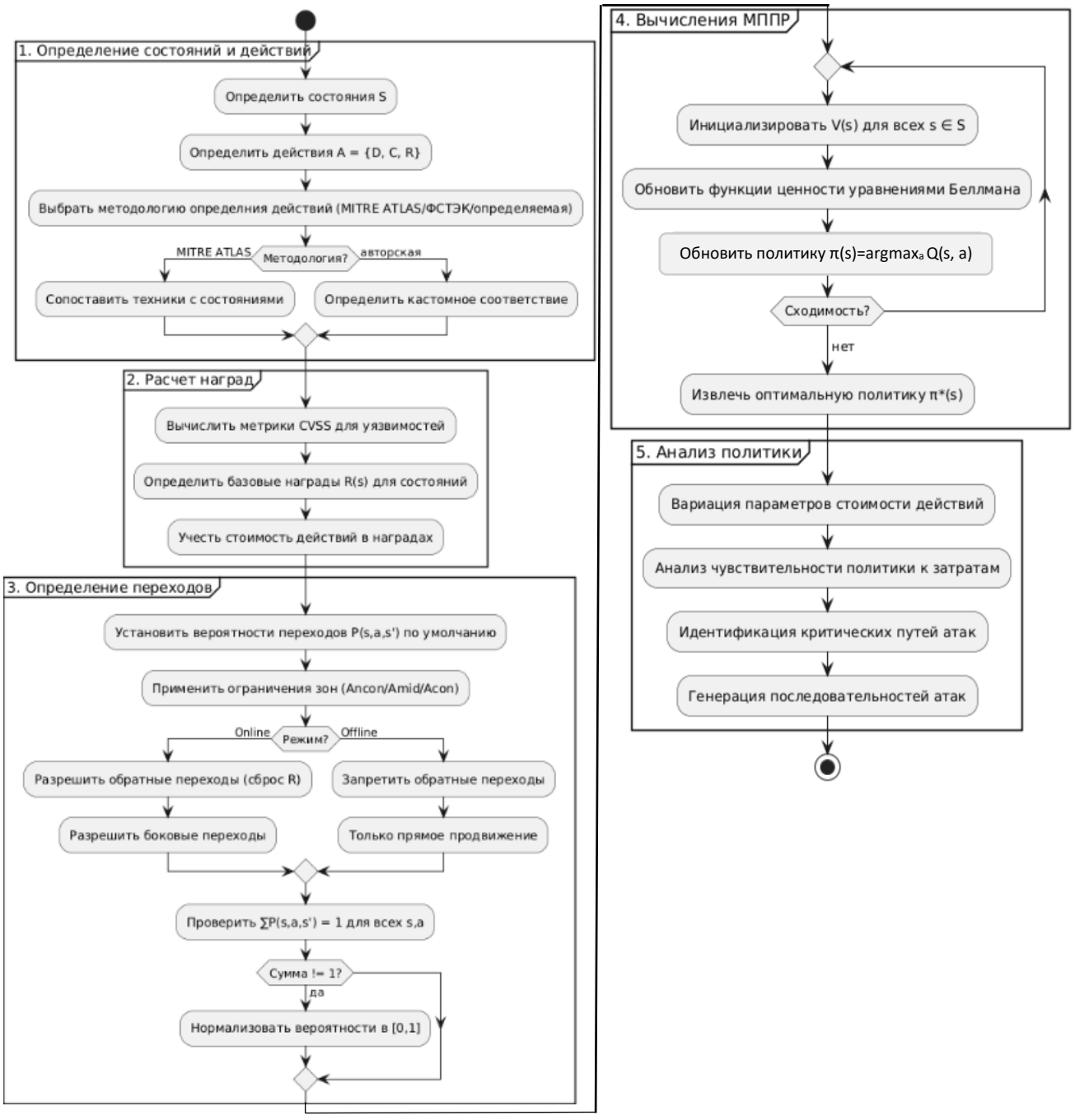


Рисунок 3.16 – Диаграмма общего алгоритм моделирования

Следует учитывать следующие ограничения [135]: суммирование вероятностей перехода из одного состояния в другое для каждого действия должно быть равно 1. В случае получения суммирования, не равного единице, можно стандартизировать значения, чтобы гарантировать, что они попадают в надлежащий интервал [0,1].

Методика определения последовательностей атакующих воздействий на системы и подсистемы ИИ в процессе построения сценариев атак при аудите ИБ (Рисунок 3.17) включает следующие обобщенные этапы с соответствующими особенностями реализации:

А. Подготовительный этап. Важными составляющими этого этапа является:

1. Определение исходных данных и входных параметров (приведены в формуле оценки наград). Используется техническое описание ПИИ для формирования связей состояний на техническом и логическом уровнях.

2. Определение способа моделирования:

- моделирование с учетом последовательного прохождения уровней детализации (от общей до полной) уровня детализации и режима моделирования;

- моделирование без учета прохождения уровней детализации.

3. Определение режима моделирования on-line или off-line.

Б. Основной этап (применяется к выбранному режиму и уровню моделирования):

- реализуется алгоритм определения последовательности состояний и действий, при этом учитывается необходимость определения доступности действий (по способу распределения тактик (формулы 2.62 – 2.63), при необходимости рассматривается обязательность взаимодействия нарушителя во время атаки с вычислительной моделью ИИ);

- формируется список опасных последовательностей.

В. Дополнительный этап используется в том случае, когда необходим выбор последовательного прохождения уровней детализации атакующих воздействий при моделировании. После этого предполагается повтор этапа Б.

Г. Итоговый этап предполагает формирование сценария атаки.

В данной методике исследование охватывает несколько сценариев, каждый из которых представляет уникальный вариант использования атакующих воздействий. Состояния могут определяться аудитором. Методика предполагает соблюдение следующих условий и рекомендаций:

1. Целесообразно выбирать состояния, соответствующие этапам компрометации на основе известных методик описания атак.

2. Определение оптимальной стратегии для злоумышленника производится с использованием МППР. Далее производится анализ эффективности этой стратегии, включая накопленные вознаграждения и количество шагов до блокировки.

3. При моделировании поведения агента (злоумышленника) рекомендуется:

- начинать с нейтрального состояния;
- выбирать оптимальное действие для текущего состояния;
- определчть возможные состояния перехода из текущего состояния;
- на основе вероятностей перехода выбирать следующее состояние;
- получать вознаграждение за переход и обновлять счетчик шагов;
- если новое состояние — заблокировано, завершать процесс; в противном случае повторять шаги.

4. Аналогично для случайных действий выполняются те же шаги (пункт 3), но без использования оптимальной стратегии.

5. Следует скорректировать правила вероятностей перехода. Сумма вероятностей переходов для каждого действия должна равняться 1.

6. Следует скорректировать правила вознаграждений.

7. Следует определять уязвимости в соответствии с параметрами уязвимости, которые связаны с техникой его эксплуатации и, следовательно, состоянием – тактикой. Методика включает оценку уязвимости моделей ИИ к различным типам атак, таким как инверсионные атаки и атаки уклонения.

Важно учитывать:

1. Уровень осведомленности злоумышленника о системе.

2. Разные цели атак требуют различных подходов к моделированию и защите.

3. Разрабатывать модели с учетом потенциальных угроз и уязвимостей.

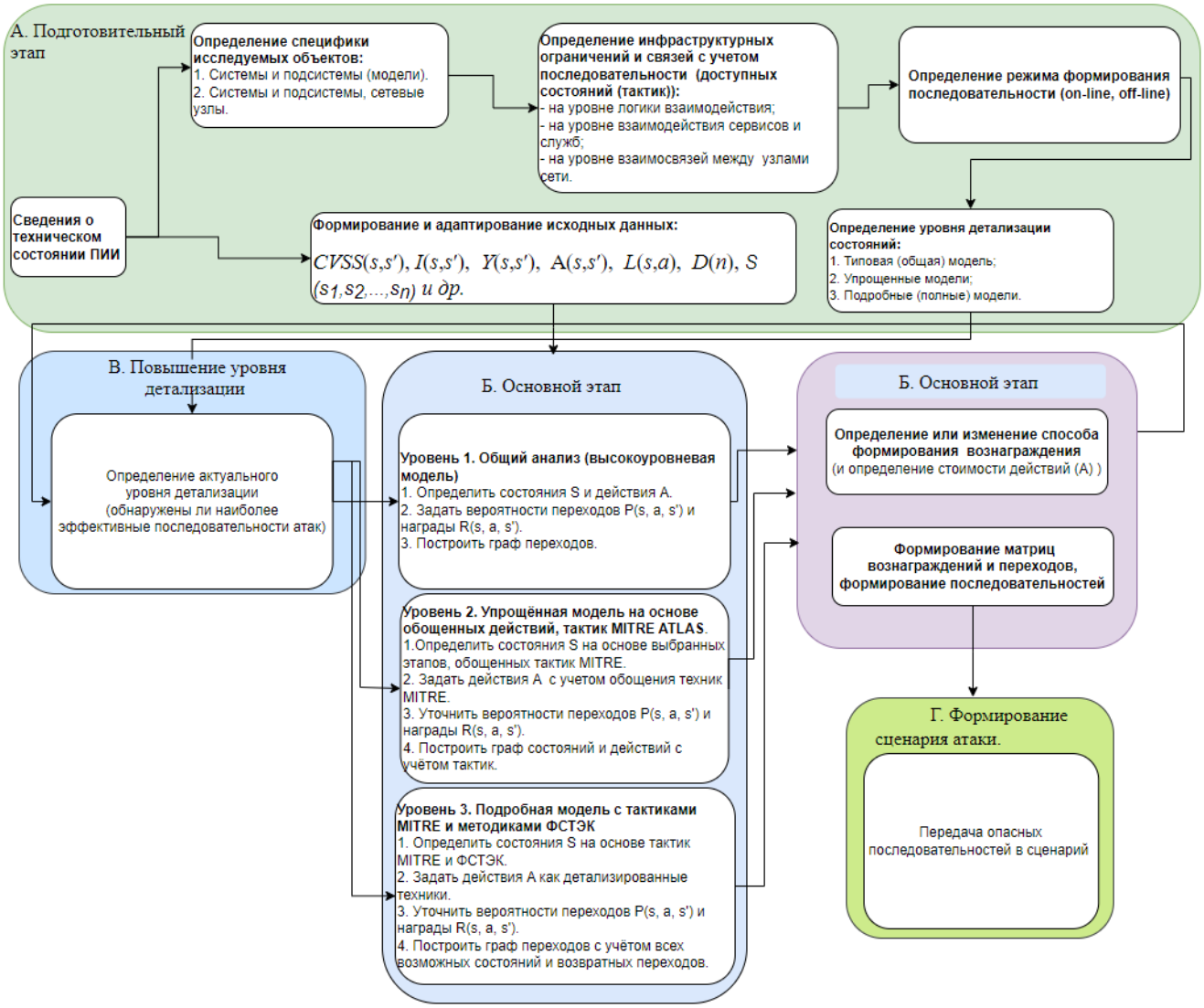


Рисунок 3.17 – Методика определения наилучшей для злоумышленника последовательности действий и состояний

Эта методика предоставляет структурированный подход к моделированию атак на системы искусственного интеллекта через марковские процессы принятия решений. Она может быть использована для разработки более защищенных приложений и повышения общей безопасности в реальных проблемных областях.

Выводы к главе 3

В третьей главе рассмотрены ключевые модели, отражающие особенности анализа безопасности систем искусственного интеллекта. Эти модели построены с учетом степени детализации и описывают все возможные состояния, в которых может находиться система во время атаки. Каждое состояние соответствует определенной тактике, позволяющей злоумышленнику воспользоваться уязвимостью.

Первый тип модели направлен на общее исследование безопасности системы и не использует метрики CVSS. Оценка уязвимости касается в первую очередь устойчивости модели к атакам и способности датасетов противостоять негативным изменениям.

Второй тип моделей можно разделить на три подтипа. Основанием для такого разделения является необходимость взаимодействия злоумышленника с вычислительной системой, датасетом и местом его хранения. Если такое взаимодействие возможно, действия злоумышленника учитывают величину вознаграждения. Таким образом, выделяются три подтипа:

1. Первый тип указывает на отсутствие доступа в других состояниях, кроме разведки и первичной компрометации.

2. Второй тип предполагает возможность первичной компрометации и разведки без доступа к выполнению и блокировке.

3. Третий тип допускает свободное использование злоумышленниками любого из состояний системы.

Эти модели можно применять в основном для анализа атакующих действий на уровне логического взаимодействия, то есть касается компрометации вычислительной модели и отравления ее датасета. Последний тип модели, который включает возможность полного взаимодействия злоумышленника с любым набором инфраструктурных элементов, позволяет оценивать атаки на уровне, как инфраструктурного взаимодействия, так и поражения вычислительной модели

системы искусственного интеллекта. Поэтому он применим не только для атак на уровне вычислительной модели и отравления датасетов, но и для классических вариантов атак при ограничениях MITRE ATLAS. Следует отметить, что уровень детализации в последней модели выше, чем в первой.

Модель первого типа достаточно абстрактна, что позволяет упростить представление атаки и выявить наиболее уязвимые зоны, на которые следует обратить особое внимание. Первый тип моделей позволяет исследовать не только общее состояние системы, но и само состояние за счет понижения или повышения наград и вложений (усилий) злоумышленника.

Третий тип модели использует все состояния, интерпретируемые по методике MITRE ATLAS, со всеми возможными взаимосвязями. Данная модель позволяет учитывать последовательности, которые могут основываться только на предположениях о том, как сформирована инфраструктура систем ИИ. Это сделано для того, чтобы аудитор, если ему полная инфраструктурная информация не доступна, мог исследовать все возможные пути атаки.

Все модели, приведенные в исследовании, предполагают использование заранее подготовленных данных, которые прежде всего включают в свой состав:

- результаты технического обследования;

- сведения об уязвимостях, проверенных по базам данных способов эксплуатации уязвимостей и банкам уязвимостей: ФСТЭК, MITRE ATLAS, CVE, CWE, CAPEC.

Глава 4. Экспериментальная оценка разработанных моделей

4.1. Сравнение предлагаемого моделирования с альтернативными

Полнота при моделировании атак на ПИИ с использованием МППР, в данном случае, представляет собой способность модели учитывать все состояния атакуемой системы и переходы между ними. В контексте рассмотрения атак на системы искусственного интеллекта это означает, что модель должна включать все возможные действия злоумышленника и реакции системы с учетом организации работы ПИИ и ее инфраструктурных особенностей. Это необходимо для формирования в процессе аудита сценария атаки. В МППР полнота влияет на качество принимаемых решений. Чем более полная модель, тем более точные прогнозы можно сделать о поведении системы при различных сценариях атак. Полнота описания набора действий достигается следующим образом:

1. Учитываются все возможные состояния системы с учетом методики описания. Максимально полное описание состояний достигается при использовании тактик MITRE ATLAS.

2. Учитываются типы действий (нелегальное взаимодействие (*D*), легальное взаимодействие (*C*)), а также все возможные действия атакующего с учетом методики описания сценария атаки (действия представлены как тактики из MITRE ATLAS и Методики ФСТЭК), содержащие максимальное число актуальных тактик злоумышленников с учетом их обновления, при использовании полносвязанного графа атаки.

3. Учитываются обратные действия атакующего, возвращение злоумышленника в предыдущее состояние атаки (действия сброса (*R*)) в режимах on-line.

4. При поиске наилучшей последовательности атакующих воздействий перебираются все варианты при учёте специфики уязвимостей и адресации узла в инфраструктуре системы ИИ, которые также учитываются при аудите.

Типологическая упорядоченность действий в соответствии с методиками описания позволяют формировать сценарий атаки (и векторы атак) как набор последовательных этапов развития атаки (от разведки до целевой тактики-состояния, например, «Выполнение»). Допустимо обновление описания состояний в соответствии с обновлением методики описания тактик.

При сравнении с различными методами, не учитывающими динамические особенности атак (отсутствие действия сброса (R)) и распределение действий по типам (в том числе по принадлежности к тактикам MITRE ATLAS), в предлагаемом способе моделирования on-line количество действий больше. Это увеличивает меру сложности моделирования, а, следовательно, и скорость расчетов. Однако при процедурах аудита ИБ ПИИ, реализуемых до начала инцидента безопасности, скорость расчетов менее важна, чем полнота описания сценария возможной атаки. Таким образом, полнота, как максимально возможное описание действий в сценарии атак с применением методик описания (MITRE ATLAS и Методики ФСТЭК), является более важной характеристикой (4.1).

$$A_s = \sum_{s \ni S}(Ds + Cs + R) \geq A'_s = \sum_{s \ni S}(Ds + Cs) \geq A''_s = \sum_{s \ni S} a_i, \quad (4.1)$$

где $D_s = f_D(H_s, I_s, Y_s, S)$ – функция, зависящая от количества хостов (компонентов ПИИ), связей между ними, связей между уязвимостями и самого состояния для нелегитимных действий; $C_s = f_C(H_s, I_s, Y_s, S)$ – аналогично для легальных действий; R – количество возможных сбросов, которое может быть функцией от числа посещённых состояний или фиксированным значением; H_s – количество хостов (компонентов), доступных в состоянии s; a_i – действие, связанное с эксплуатацией уязвимости без определения принадлежности к типу эксплуатации; A_s – множество действий всех типов, A'_s – множество действий без учета R, A''_s – множество действий без указания их типа.

При определении вектора атаки в многосоставной системе возрастает мера сложности, поскольку количество узлов и, соответственно, уязвимостей может влиять на сложность модели.

Снижение меры сложности достигается путём учёта:

- технологических ограничений повторения и функционирования систем искусственного интеллекта;

- уровней детализации описания последовательностей: от описания логики работы искусственного интеллекта (важно при атаках на логику работы вычислительной модели искусственного интеллекта) до описания с учётом каждого узла и его уязвимостей.

На рисунках 4.1 и 4.2 показано число прироста действий в соответствии с приростом числа актуальных поставщиков уязвимостей – узлов инфраструктуры систем ИИ.

№	Случай	Количество состояний	Формула для количества действий	расчет (N=10 узлов)
1	D, R, C для каждой техники каждой тактики MITRE ATLAS	15 (14 тактик + Блок)	$A = \sum_{t \in T} Tech(t) \cdot 3$	56 · 3 = 168
2	D, C для каждой техники каждой тактики MITRE ATLAS	15	$A = \sum_{t \in T} Tech(t) \cdot 2$	56 · 2 = 112
3	D, R, C для состояний: Р, П, Д, В, Б	5	$A = \sum_{s \in \{Р, П, Д, В\}} Tech(s) \cdot 3$	40 · 3 = 120
4	D, C для состояний: Р, П, Д, В, Б	5	$A = \sum_{s \in \{Р, П, Д, В\}} Tech(s) \cdot 2$	40 · 2 = 80
5	D, R, C как подтипы с состояниями: C, L, T, Б	4	$A = \sum_{s \in \{C, L, T\}} Tech(s) \cdot 3 + Tech(Б)$	30 · 3 + 10 = 100
6	D, C как подтипы с состояниями: C, L, T, Б	4	$A = \sum_{s \in \{C, L, T\}} Tech(s) \cdot 2 + Tech(Б)$	30 · 2 + 10 = 70

Рисунок 4.1 – Количество состояний и действий для разных случаев

Случай расчёта количества действий, когда отсутствует сброс (*R*), характерен для тех методик, которые не учитывают этот вариант действий (дерево решений, нейросеть и др.).

Случай	Формула	N=5	N=10	N=20
1	$A = 56 \cdot 3 + 2N$	178	192	212
2	$A = 56 \cdot 2 + 2N$	122	132	152
3	$A = 40 \cdot 3 + 3N$	135	150	180
4	$A = 40 \cdot 2 + 3N$	95	110	140
5	$A = 30 \cdot 3 + 10 + N$	105	110	120
6	$A = 30 \cdot 2 + 10 + N$	80	85	95

Рисунок 4.2 – Зависимость количества действий от узлов сети (N)

В МППР полнота влияет на качество принимаемых решений. Чем более полная модель, тем более точные прогнозы можно сделать о поведении системы при различных сценариях атак.

При этом количество действий увеличивается с приростом количества элементов ИИ. На рисунке 4.3 демонстрируется изменение количества действий в зависимости от количества узлов, присутствующих в инфраструктуре системы ИИ. Для моделей без возвратных действий степень полноты снижена (Случай 2 – подробная (полная) модель, Случай 4 – упрощенные модели, Случай 6 – типовая (общая) модель) по сравнению с моделями, в которых возвратные действия учтены (Случай 1 – подробная (полная) модель, Случай 3 – упрощенные модели, Случай 5 – типовая (общая) модель). Такое понижение уровня охвата характерно и для моделей, основанных и на других подходах в силу методологических ограничений. Максимальный уровень охвата возможных действий и состояний демонстрирует

подробная модель с использованием техник и тактик, учитывающая возвратные состояния (Случай 1), что актуально для повышения качества аудита в части повышения степени полноты описания атак на системы ИИ, а также ПИИ.

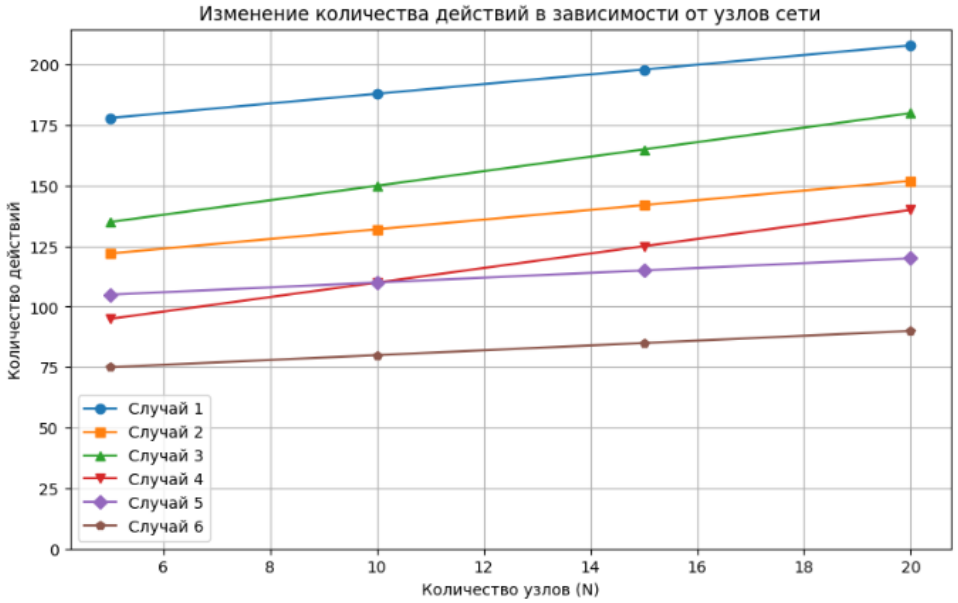


Рисунок 4.3 – Изменение количества действий в зависимости от количества узлов в инфраструктуре ИИ

Далее требуется показать, что модель на основе МППР превосходит альтернативные методы (деревья атак, методы машинного обучения и др.) по полноте, точности, времени и адаптивности (Таблица 4.1). Моделирование производилось с учетом следующего:

1. Выбор производился из набора уязвимостей и тактик MITRE ATLAS для тестовой системы ИИ (например, атаки на интерфейсы API, отравление данных). Используются следующие тактики:

– T1 (реализуется сбор информации о модели/данных).

– T4 (реализуется получение доступа к API модели).

– T5 (реализуется внедрение вредоносного кода).

– T14 (реализуется компрометация (эксфильтрация) модели).

2. Для каждой атаки при моделировании с помощью МППР строился граф состояний с учётом тактик MITRE ATLAS (например, T1 → T4 → T5 → T14).

3. Вероятности переходов между состояниями задавались на основе CVSS-оценок уязвимостей. Например, вероятность перехода от T1 (интепретируется как «Сбор информации») к T4 («Внедрение вредоносного ПО») определяется следующим образом: $P(T4/T1) = CVSS_{T1 \rightarrow T4} \backslash \sum CVSS_{T1 \rightarrow \text{все возможные состояния}}$.

4. При моделировании последовательностей действий злоумышленника использовались следующие методы: модели МППР (предложенный метод), деревьев атак (традиционный метод), байесовские сети, нейросеть.

Краткое описание альтернативных моделей:

1. Описание нейросети. Представлена модель нейросети LSTM-сеть с архитектурой, включающей три слоя:

a) Входной слой: 50 нейронов (векторизация тактик). Каждая тактика MITRE ATLAS (например, T1, T4 и др.) кодируется в 50-мерный вектор с использованием one-hot encoding или word embedding.

b) Скрытый слой: 30 нейронов. Количество нейронов: 30. Функция активации: tanh (для управления потоком информации). Dropout: 0.2 (для предотвращения переобучения).

c) Выходной слой: Softmax (предсказание следующей тактики). Количество нейронов: число возможных тактик (например, 14 тактик MITRE ATLAS).

Использовались следующие тестовые данные (логи атак на системы ИИ (пример: 1000 последовательностей тактик)): 80% данных (80 атак) для обучения, 20% (20 атак) для тестирования. В процессе обучения использовались функция потерь «Categorical Cross-Entropy» (для многоклассовой классификации). Другие особенности нейросети:

– оптимизатор: Adam (learning rate = 0.001);

– количество примеров данных, которые обрабатываются нейронной сетью за один шаг обучения перед обновлением её параметров (весов): 32;

– количество эпох: 100;

– валидация: 20% данных выделено для проверки.

2. Краткое описание параметров дерева решений. Каждый узел представляет

этап атаки или условие. Связи (ребра) отражают последовательность действий. Каждому переходу между узлами присваивается вероятность успеха, основанная на: оценках уязвимостей (используются метрики CVSS), экспертных оценках (например, вероятность успеха фишинга — 30%). Вес ребра определяется следующим образом: *вес*=CVSS×*вероятность успеха*. Структура узлов следующая:

– корневой узел представлен как цель атаки (например, "Кража модели ИИ");

– промежуточные узлы представлены как тактики (T1, T4, T5, T14);

– листовые узлы узлы представлены как конкретные техники реализации (например, «Фишинг для доступа к API»).

3. Описание байесовской сети. Байесовская сеть представляет собой направленный ациклический граф (DAG), где узлы соответствуют тактикам MITRE ATLAS (T1, T4, T5, T14), а ребра отражают вероятностные зависимости между ними. Каждый узел имеет таблицу условных вероятностей, которая определяет вероятность наступления события при заданных значениях родительских узлов. Для моделирования атакующих последовательностей, тактики MITRE ATLAS могут быть представлены как узлы, а вероятности переходов между ними — как условные вероятности.

Таблица 4.1 - Сравнение результатов моделирования

Метод	Полнота (%)	Точность (%)	Время выполнения (с)	Адаптивность (1-5)
МППР	90	95	120	5
Деревья атак	70	80	60	2
Нейронная сеть	75	85	300	3
Байесовская сеть	80	88	200	4

1. Полнота (основной показатель) в этом случае определяется как процент охвата всех возможных атакующих последовательностей (сценариев атак): *полнота = количество обнаруженных уникальных сценариев \ общее количество возможных сценариев × 100%.*

Особенности получения приводимых результатов следующие:

- МППР (90%). Модель МППР анализирует все возможные пути, включая

редкие комбинации тактик (например, комбинация T1-T4-T5-T14). Учет вероятностей позволяет выявлять даже маловероятные, но возможные сценарии. МППР выявила 90% за счёт анализа всех путей через матрицу переходов. Например, для тактик {T1, T4, T5, T14} вероятность пути $T1 \rightarrow T4 \rightarrow T5 \rightarrow T14 = P(T4|T1) \times P(T5|T4) \times P(T14|T5)$. Вероятность пути $T1 \rightarrow T4 \rightarrow T5 \rightarrow T14 = P(T4|T1) \times P(T5|T4) \times P(T14|T5)$;

- деревья атак (70%). Деревья ограничены предопределенной структурой. Например, если в дереве не учтена тактика "Сокрытие действий" (T7), соответствующие сценарии будут пропущены. Деревья не учитывают циклы и альтернативные пути (например, только линейные пути типа: $T1 \rightarrow T4 \rightarrow T5 \rightarrow T14$);

- нейронная сеть (75%). Нейросети могут пропускать нетипичные сценарии, если они недостаточно представлены в обучающей выборке. Нейросеть пропускает редкие сценарии из-за недостатка данных (например, пропущены комбинации с тактикой T7 "Сокрытие действий");

- байесовская сеть (80%). Байесовские сети могут учитывать сложные зависимости между тактиками, что позволяет им охватывать широкий спектр сценариев атак. Однако, если сеть не включает все возможные зависимости, полнота может быть снижена. Например, если в сети не учтена зависимость между тактикой T1 и T5, соответствующие сценарии могут быть пропущены.

2. Особенности определения параметра «точность». Подразумевается процент корректно предсказанных атакующих последовательностей относительно реальных данных. Определяется следующим образом: *точность = количество верных предсказаний \ общее количество тестовых случаев × 100 %.* Особенности получения приводимых результатов следующие:

- МППР (95%). Марковские процессы учитывают вероятности переходов между состояниями (например, успешность эксплуатации уязвимости), что позволяет точнее прогнозировать многошаговые атаки. Например, если злоумышленник переходит от тактики "Сбор информации» (T1) к «Получению

доступа» (T2) с вероятностью 0.8, модель МППР корректно отражает эту динамику;

- деревья атак (80%). Иерархическая структура деревьев не всегда учитывает альтернативные пути атак (например, циклические переходы). Например, если атакующий может вернуться к предыдущему состоянию, дерево не отражает эту возможность. При тестировании использовались 100 атак из тестовых данных. При моделировании использовалось построение дерева атак с узлами, которые соответствуют тактикам (T1, T4, T5, T14). Для каждого узла задавались бинарные условия перехода (например, «успех T1 → переход к T4»). При этом циклы и альтернативные пути игнорировались. В результате из 100 атак дерево предсказало 80 верных последовательностей;

- нейронная сеть (85%). Метод зависит от качества обучающих данных. Если в данных отсутствуют редкие сценарии (например, атаки на обновленные модели ИИ), точность снижается. На тестовых данных модель предсказала 17 из 20 атак;

- байесовская сеть (88%). Байесовские сети могут быть точными при наличии достаточных данных для обучения. Например, если в данных присутствуют все возможные комбинации тактик, точность предсказаний будет высокой. Однако, если данные недостаточно полны, точность может снижаться.

3. Время выполнения (секунды) определяется как время, необходимое для построения модели и прогнозирования атакующих последовательностей.

$$\text{Время выполнения} = t_{\text{построение}} + t_{\text{прогноз}} . \quad (4.2)$$

Особенности получения приводимых результатов следующие:

- вычисления для МППР (120 с.). Построение матрицы переходов проводилось для 15 состояний и 50 действий. Время построения определяется следующим образом:

$$t_{\text{построение}} = 15 \times 50 \times t_{\text{расчёт вероятности}}, \quad (4.3)$$

где $t_{\text{расчёт вероятности}} = 0.01$ (на одно действие).

В итоге время $t_{построение}$= 7.5 с. Далее проводится оптимизация стратегии (использовалась Value Iteration, количество итераций: 20). Время на итерацию определяется следующим образом: $t_{итерация=}15×50×0.005=3.75$ с. Время на итерацию оптимизацию: $t_{оптимизация}$ =20×3.75=75 с. Время на прогнозироваие для 100 атак $t_{прогноз}$=10 с. Таким образом итоговое время определяется как $t_{общее}$ =7.5+75+10=92.5 с. ≈120 с. (с учётом накладных расходов);

- деревья атак (60 с.). Построение дерева с фиксированной структурой (например, 10 узлов) требует меньших вычислительных ресурсов. Построение дерева предполагает внесение в него пятидесяти узлов за следующий временной период: $t_{построение}$=50×1 с=50 с. Прогнозирование предполагает $t_{прогноз}$=10 с. Таким образом получается $t_{общее}$=50+10=60 с.;

- нейронная сеть (300 с.). Обучение нейросети (например, LSTM) на данных логов занимает время из-за необходимости настройки гиперпараметров и обработки больших датасетов (эпохи обучения: 100, время на эпоху: $t_{эпоха}$=2 с. $t_{обучение}$=100×2=200 с. Прогнозирование: $t_{прогноз}$=100 с. Итого: $t_{общее}$=200+100=300 с.;

- байесовская сеть (200 с.). Построение и обучение байесовской сети может быть вычислительно затратным. Например, для сети с 15 узлами и 50 зависимостями время построения может составлять около 150 секунд, а время прогнозирования — 50 секунд.

4. Адаптивность (1-5) предполагает способность модели быстро адаптироваться к новым угрозам и изменениям в системе. Критерии оценок следующие:

– оценка 5 обозначает динамическое обновление без перестройки модели;

– оценка 3 обозначает требуется частичное переобучение;

– оценка 1 обозначает полная перестройка модели.

Особенности приводимых результатов следующие:

- МППР (5) - модель легко обновляется при появлении новых уязвимостей. Например, добавление нового состояния «Уязвимость в API» требует только корректировки матриц переходов и вознаграждений;

- деревья атак (2) - для добавления новых узлов (например, тактики Т11 «Фишинг») требуется перестроение всей структуры, что трудоемко;

- нейронная сеть (3) - для учета новых данных требуется повторное обучение модели, что занимает время и вычислительные ресурсы;

- байесовская сеть (4). Байесовские сети могут быть адаптированы к новым данным путем обновления условных вероятностей. Однако, это может потребовать значительных вычислительных ресурсов, особенно при большом количестве узлов.

Таким образом, к положительным сторонам приводимых моделей МППР можно отнести:

1. Учет динамики и неопределенности. МППР позволяет моделировать вероятностные переходы между состояниями, что критически важно для анализа многошаговых атак, где злоумышленник адаптируется к защитным мерам. Например, если злоумышленник проваливает атаку на этапе Т4, модель МППР автоматически пересчитывает вероятности переходов к другим тактикам (например, Т5 или Т7), что недоступно в статических деревьях атак.

2. Полнота охвата сценариев. Модель МППР анализирует все возможные пути атак, включая редкие и неочевидные комбинации тактик.

3. Адаптивность. При добавлении новых уязвимостей или тактик (например, Т11 «Социальная инженерия») модель МППР требует лишь обновления матрицы переходов, а не полной перестройки. Это делает её применимой в быстро меняющихся средах.

4. Интерпретируемость. В отличие от «черного ящика» нейронных сетей, МППР предоставляет прозрачную структуру состояний и переходов, что упрощает анализ причинно-следственных связей.

Таким образом, предложенные модели выделяются по параметру полноты, превосходя сравниваемые. При этом учитывается возможность интеграции тактик различных методик описания атак в процесс моделирования.

4.2 Описание тестового примера и экспериментальная оценка определения путей атаки с использованием моделей МППР

Для испытания предложенных моделей и методики была также разработана генеративно-состязательная сеть (GAN), которая представляет собой модель машинного обучения, умеющую имитировать заданное распределение данных. Нейросеть GAN состоит из двух нейронных сетей, одна из которых обучена генерировать данные, а другая - отличать данные от реальных данных. Таким образом, GAN является комбинацией двух противоположных подходов построения нейронных сетей: дискриминатора и генератора. Нейронная сеть функционирует в инфраструктуре, приведенной на рисунке 4.4 [136].

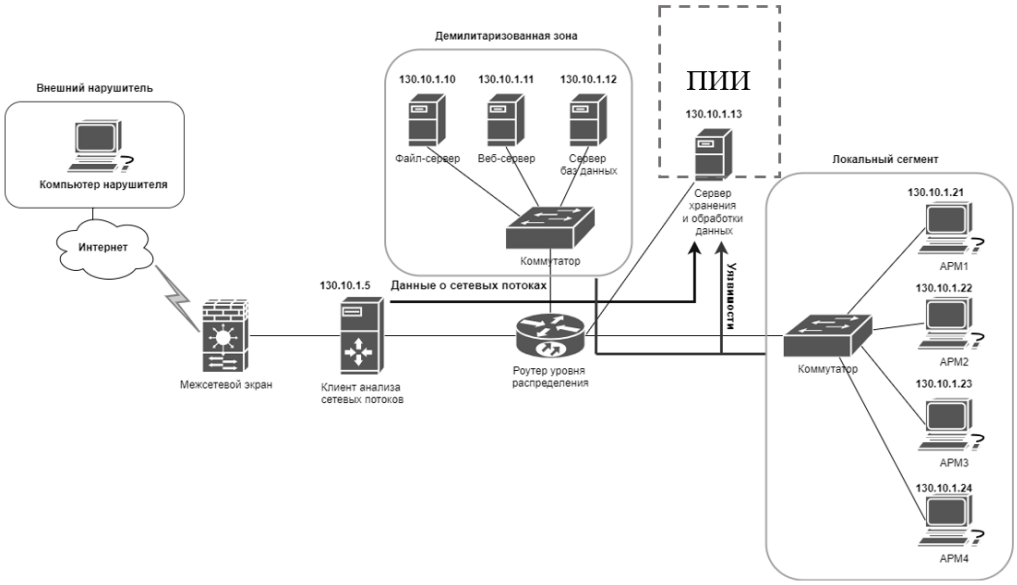


Рисунок 4.4 – Структура тестовой компьютерной сети

В ходе обучения генератор пытается сгенерировать на основе реальной выборки данные $\tilde{X}$, напоминающие реальные данные (набор ISCX Botnet Dataset). Дискриминатор обучается оценивать вероятность того, что образец получен из реальных данных X, а не предоставлен генератором. На вход генератора подаются случайный шум, который преобразуется так, чтобы данные стали состязательными примерами для дискриминатора. На вход дискриминатора попеременно подаются

данные из обучающего набора и смоделированные образцы, сгенерированные генератором. Процесс обучения GAN заключается в минимаксной игре двух моделей, в которой дискриминатор D адаптирован для минимизации ошибки различия реального и сгенерированного образца, а генератор G построен на максимизации вероятности того, что дискриминатор допустит ошибку. Таким образом, модель GAN может быть использована для повышения устойчивости модели к состязательным атакам с обеспечением достаточно хорошей точности для легитимных входных данных.

Генератор на каждом этапе обучения будет генерировать все более и более лучшие примеры, а дискриминатор будет классифицировать лучше, как вредоносные данные атакующего, так и быть устойчивым к состязательным примерам. Размер пакета был установлен равным 32 записям. Из обучающего набора была взята проверочная выборка равная 30 % от обучающего набора. Проверка модели на тестовом наборе происходит при вызове метода evaluate [136]. В ходе обучения точность модели на проверочной выборке составила 96.26 %. На рисунке 4.5 показано изменение функций потерь генератора и дискриминатора (Рисунки 4.5). С ростом числа эпох потери уменьшаются, а эффективность классификации дискриминатора и генерации данных генератора увеличивается.

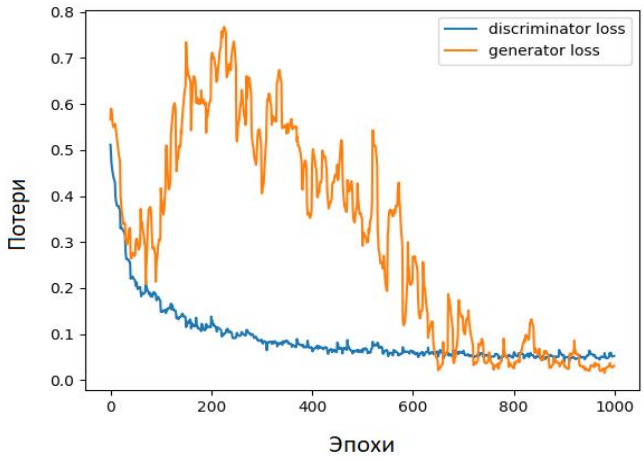


Рисунок 4.5 – Потери дискриминатора и генератора на этапе: обучения

Нейросеть GAN была переобучена с учетом наличия состязательных атак (Рисунки 4.5, 4.6).

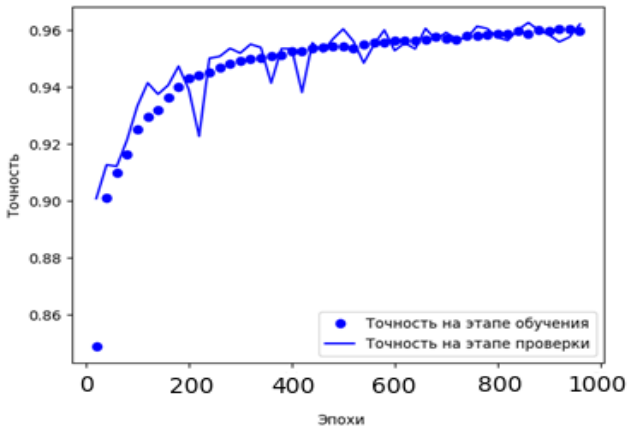


Рисунок 4.6 – Точность на этапах обучения и проверки

Таким образом, для анализа использовалась сборка генеративных состязательных сетей (GAN), которая включает в свой состав нейросеть-генератор (шесть слоев, из которых четыре – скрытые полносвязанные слои) и нейросеть-дискриминатор (семь слоев, из которых пять – скрытые полносвязанные слои). Сборка сравнивалась с моделью машинного обучения. На рисунке 4.7 показаны показатели точности дискриминатора на этапах обучения и проверки. С увеличением числа эпох точность модели увеличивается. Рассматривается оценка эффективности модели на тестовом наборе данных, на которых модель не обучалась.

```
Test loss:  0.10812488198280334
Test accuracy:  0.9561761021614075
AUC:  0.9910829663276672
Precision:  0.9843450784683228
Recall:  0.9502431154251099
```

Рисунок 4.7 – Метрики оценки эффективности модели

Моделирование атаки на датасет, который используется для обучения, имеет специфические характеристики [134]. Следует отметить, что в произвольный момент времени атакуемая система ИИ может находиться в любом из состояний атаки. Модификации входных данных сильно связаны с проблемой устойчивости модели. Под устойчивостью модели понимают меру его чувствительности к возмущениям в исходных данных. Модель считается устойчивой, если при обучении погрешность в изначальных данных поэтапно не снижает точность классификации.

При этом достичь неустойчивости работы модели можно другим способом: данные могут быть модифицированы на этапе тренировки, когда в обучающий набор добавляются записи, которые снижают качество классификации. Соответственно, возникает проблема доверия к обучающим датасетам. Эту проблему можно решить путем введения процедур обязательной верификации обучающих выборок. Однако в этом случае атака на уже обученную модель не исключается. В ходе моделирования атаки была определена случайная выборка 1000 векторов данных из тестового набора. Тестирование атаки проводилось как на исходных данных, так и на сгенерированных состязательных выборках с различной величиной множителя возмущения. Результаты эффективности работы метода оценивались при помощи кривой ROC-AUC показателя и кривой Precision-Recall.

На основе показателей полноты и точности определяется метрика средней точности AP как средневзвешенное значение точности (Precision), достигнутой на каждой итерации с увеличением полноты (Recall) по сравнению с предыдущей итерацией. На рисунке 4.8 показаны результаты атаки на обычную модель нейронной сети. Показатели AP и AUC резко уменьшаются при проведении состязательных атак. При проведении аналогичной атаки на модифицированную нейросеть (модель GAN), обученную с учетом наличия состязательных атак, показатели эффективности нейронной сети остаются на высоком уровне, и при этом эффективность атаки падает (Рисунок 4.7). Следует учитывать, что диапазон допустимых отклонений уменьшается в процессе классификации данных (при переобучении обычных нейросетей подобное может приводить к появлению множества ошибок, связанных с ложноположительными отказами вычислительной модели). Несмотря на то, что нейронная сеть (модель GAN) показала высокую устойчивость к состязательным атакам (по модели белого ящика), а при проведении атак черного ящика увеличивается время генерации состязательных примеров, а также уменьшается эффективность состязательных выборок из-за отсутствия данных о модели нейронной сети, диапазон доступных возмущений, для манипуляций злоумышленника, все равно присутствует (что означает увеличение времени его поиска). Кроме того, стоит отметить, что модифицированная сеть устойчива к уровню возмущений до 20% от

исходного вектора данных [135]. Таким образом, задача построения защищенных нейронных сетей сводится к минимизации величины уязвимости и возможности ее эксплуатации при манипуляции с вычислительной моделью ИИ.

Рассмотрим использование метода МППР для определения последовательности (тактик) (определение экспертным путем последовательности тактик при известной инфраструктуре) [134, 135]. Состояния модели атаки описываются в соответствии с тактическими наборами MITRE ATLAS. Например, состояние T1 понимается как успех действий злоумышленника, который вел разведку, и, собрав необходимую информацию о целевой системе, может перейти к следующим тактикам (этапам) атаки. Из тактик MITRE непосредственно к модели предлагаемого к рассмотрению нападения по вектору состязательной атаки относятся следующие наборы: T1, T2, T4, T4.1, T4.2, T4.3, T5, T7, T14 [135].

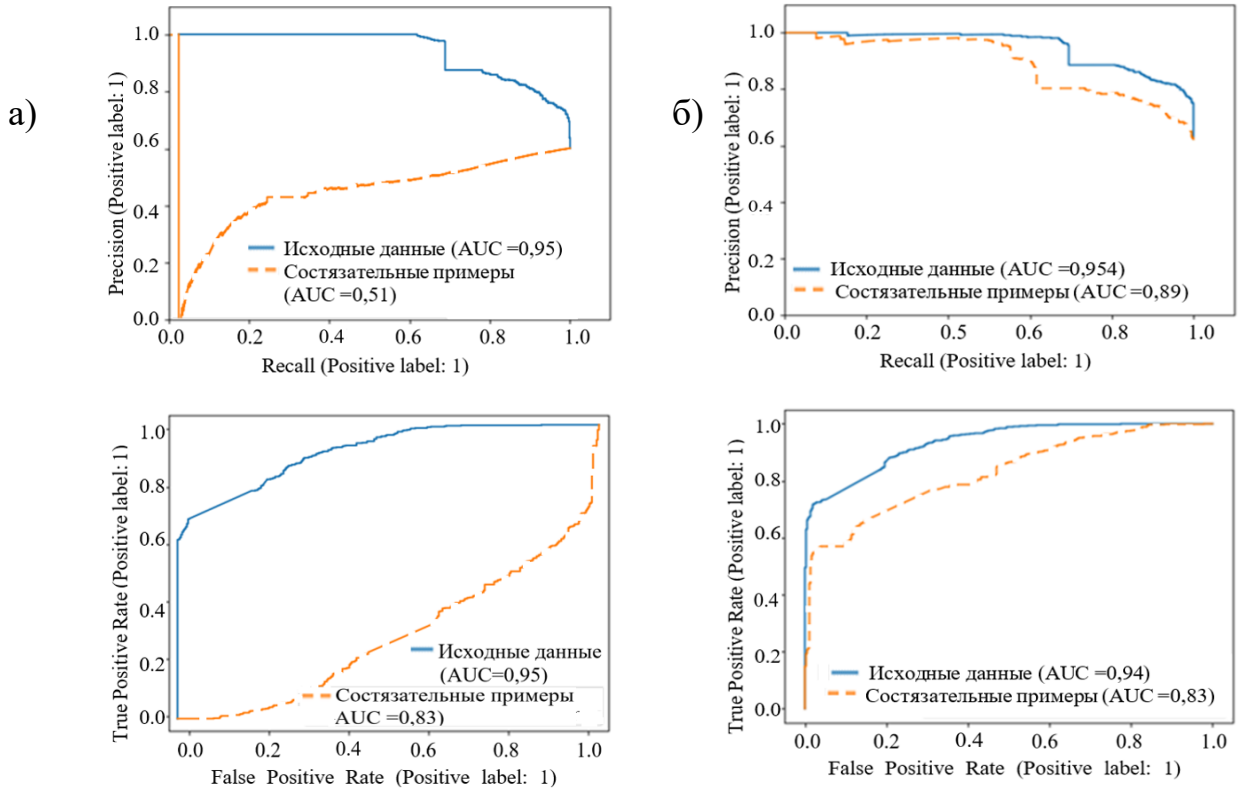


Рисунок 4.8 - Графики AP и AUC при моделировании состязательных атак на обычную модель нейронной сети (а) и графики AP и AUC при моделировании состязательных атак на обученную модель GAN (б).

При визуализации графа учтены все возможные состояния (все наборы тактик MITRE) (Рисунок 4.11 (б)). Множество действий модели содержит следующее:

1. Разрешить (Allow, «a»): успешное выполнение тактики.
2. Не разрешить (Not-Allow, «n»): отсутствие успеха.

Возможные результаты действий:

1. Отсутствие действия (not-compromised).
2. Простаивание – наблюдение не дает четко установить нынешнее состояние системы (unknown).
3. Взаимодействие – процесс использования уязвимости для достижения состояния компрометации по какому-либо этапу (compromised).

Динамика изменений заданных значений матрицы переходов между состояниями и матрицы вознаграждений при обучении приведена (фрагментарно) на рисунке 4.9.

а)

```
Матрица переходов между состояниями:
[[[0.05789954 0.15856926 0.08844104 0.09470629 0.04214539 0.
  0.14922619 0.14593516 0.1190518  0.14402534]
 [0.         0.39669316 0.         0.47196687 0.13133997 0.
  0.         0.         0.         0.        ]
 [0.52697934 0.         0.         0.         0.         0.
  0.47302066 0.         0.         0.        ]
 [0.         0.         0.         0.45164895 0.         0.
  0.         0.54835105 0.         0.        ]
 [0.14879651 0.         0.         0.32873833 0.         0.
  0.52246516 0.         0.         0.        ]
 [0.09888238 0.         0.13605139 0.01627506 0.15588886 0.13419906
  0.22020479 0.0862781  0.07356396 0.07865638]
 [0.00189184 0.00461932 0.13226219 0.12009744 0.0384432  0.07186007
  0.13709279 0.15937011 0.14704516 0.18731788]
 [0.         0.         0.         0.         0.06642054 0.
  0.         0.         0.93357946 0.         ]
```

б)

```
Матрица вознаграждении:
[[[-0.09357598 -0.99970764 -0.85979102  0.18980623 -0.82901452
   0.80575298 -0.83338879 -0.1045887   0.31732356 -0.95410244]
 [ 0.         0.         -0.         -0.3363867   0.
  -0.         -0.10847402 -0.         -0.91662604  0.        ]
 [ 0.         0.         0.         -0.93743775  0.97918305
  -0.57017224 -0.88202841  0.20733075 -0.93269793  0.4481216 ]
 [-0.64026389  0.74630463 -0.40566717 -0.4388521   0.29221433
  -0.69314171  0.32204563 -0.85754741 -0.35528272 -0.12167991]
 [-0.63654614  0.         -0.         0.         -0.
  -0.         -0.         -0.         0.06751705  0.55410047]
 [-0.         0.         -0.00786791 -0.         -0.
   0.         -0.         0.         0.         0.        ]
```

Рисунок 4.9 - Матрицы переходов между состояний (а) и вознаграждений (б)

Можно заметить, что при нулевых и близких к нулю значениях вероятностей переходов также наблюдаются нулевые или отрицательные значения вознаграждений. Это указывает на ослабление связей между состояниями атаки (достигаемыми тактиками), которые незначительно влияют на выбор последовательности действий злоумышленника (Рисунок 4.10) [135].

Результаты определения оптимальной политики показывают, что тактики T1, T2, T8, были включены в стратегию атаки (Рисунок 4.10). Таким образом,

оптимальная последовательность атакующих воздействий включает всего два перехода при условии изначальной успешности достижения состояния T1.

```
Iteration                Number of different  actions
  1              1
  2              1
  3              0
Iterating stopped? unchanging policy found
(1, 1, 0, 0, 0, 0, 0, 1, 0, 0)
```

Рисунок 4.10 – Результат моделирования – определение оптимальной политики атаки

Классические модели машинного обучения (МО) также подвергаются атакам. Однако, в силу специфики их вычислительной модели, состязательные атаки мало эффективны (интерпретируемость и меньшая чувствительность к шуму делают их более устойчивыми к состязательным атакам по сравнению с нейронными сетями). В этом отношении переобучение модели не даст того же эффекта, что и при модификации нейронной сети. Однако уязвимость, представленная как допустимый диапазон отклонений входных данных, хотя и в меньшей степени, но всё же влияют на безопасность вычислительной модели ИИ. Рассмотрим специфику моделирования атаки на модель машинного обучения, в которой используется метод опорных векторов и которая решает те же задачи, что модель с нейронной сетью [135]. При рассмотрении перечня техник и тактик MITRE были определены наиболее подходящие состояния для модели: T1, T2.1, T2.2, T3, T4.1, T4.2, T4.3, T5, T7.

Действия МППР-модели следующие:

1. Разрешить (Allow, «a») - успешное выполнение тактики.
2. Не разрешить (Not-Allow, «n») - отсутствие успеха.

Далее необходимо сформировать значения вознаграждений при переходе от одной тактики к другой. Вознаграждение сводится в соответствии с оценкой уязвимостей на основе методики CVSS, сопряженной с оценкой уязвимости модели. Далее указываются начальные вероятности, которые определяются исходя из количества запросов от клиентского приложения к системе машинного обучения. Вычисления учитывают отсутствие взаимодействия при запросе и, соответственно,

само взаимодействие. Используя данные из таблицы 4.2, можно найти вероятности реализации тактик и построить модель оптимальной политики.

Таблица 4.2 - Вероятности начальных вероятностей

Состояние	Отсутствие обращений	Взаимодействие
1	1	0
2	0	1
3	0,5	0,5
4	0,5	0,5
5	0	1
6	0	1
7	0	1
8	0	1
9	0	1

Модель графа оптимальной политики представлена на рисунке 4.11. Значение S представляет вероятность использования злоумышленником выбранной тактики из актуального состояния [135]. Модель МППР предполагает использование моделирования упрощённого типа (рисунок 4.11) в режиме on-line.

При проверке модели рассматривается следующее: сохранятся ли ограничения на переходы и состояния предлагаемой модели при исследовании атак с учетом наличия таких же задач, что и у нейронной сети при наличии свободных связей (разрешены любые переходы между состояниями). При проверке учитывается следующее: последовательность атакующих воздействий на нейросеть должна иметь меньшее количество переходов, необходимых для достижения целей атаки, чем последовательность атаки на модель машинного обучения, поскольку вычислительная модель в данном случае контролируется в меньшей степени, чем модель на основе моделей приводимого типа машинного обучения.

Последовательность проявления актуальных действий и состояний с учетом вызвавших их действий (эксплуатаций уязвимостей) приведена на рисунке 4.11.

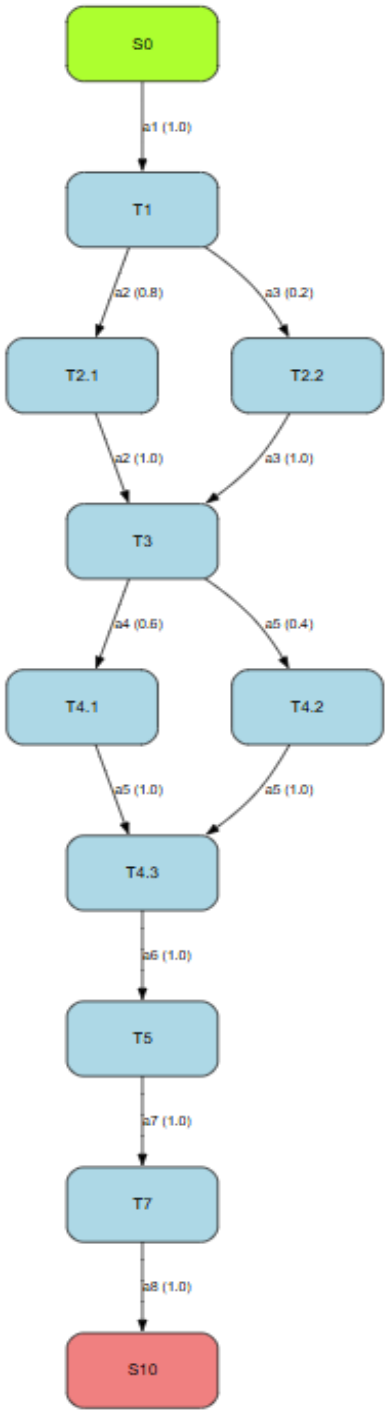


Рисунок 4.11 - Модель графа оптимальной политики злоумышленника в контексте общей последовательности состояний по методике при атаке на модель машинного обучения

В таблице 4.3 приведены итоговые вероятности сопоставленных действий и состояний последовательности атакующих воздействий на модель нейросети.

Таблица 4.3 - Итоговые вероятности сопоставленных действий и состояний последовательности

Состояние	Действия								
	a1	a2	a3	a4	a5	a6	a7	a8	a9
S0	1.0	0	0	0	0	0	0	0	0
S1	0	0.8	0.2	0	0	0	0	0	0
S2	0	1.0	0	0	0	0	0	0	0
S3	0	0	1.0	0	0	0	0	0	0
S4	0	0	0	0.6	0.4	0	0	0	0
S5	0	0	0	0	1.0	0	0	0	0
S6	0	0	0	0	1.0	0	0	0	0
S7	0	0	0	0	0	1.0	0	0	0
S8	0	0	0	0	0	0	1.0	0	0
S9	0	0	0	0	0	0	0	1.0	0
S10	0	0	0	0	0	0	0	0	1.0

Итоги определения оптимальной политики показывают, что существует несколько вариантов атаки на модель выбранного типа машинного обучения. При этом последовательность состояний включает больше переходов, что увеличивает количество действий для осуществления атаки на модель МО, что указывает на большие трудозатраты злоумышленника при подборе средств компрометации и осуществлении атакующих воздействий, в сравнении с атакой на модель нейронной сети. Таким образом, при определении особенностей поиска наилучших последовательностей действий атакующего были рассмотрены специфические особенности МППР как метода моделирования атак применительно к нейронной сети и модели машинного обучения. При построении моделей в качестве входных данных использовались метрики уязвимостей, которые были классифицированы в соответствии с методами эксплуатации уязвимостей, и, следовательно, сопоставлены тактикам MITRE (следует учесть, что оценка уязвимости может носить экспертный характер, если четкого соотнесения с тактиками не наблюдается) [135].

Последовательность атакующих воздействий на нейросеть включает в свой состав меньшее количество состояний и переходов, необходимых для достижения целей атаки, чем последовательность атаки на модель машинного обучения. В этом проявляются заданные ограничения, учитывающих специфику выбранных вычислительных модлей.

Таким образом, последовательности атакующих воздействий, сформированные из заданного множества, являются прямым результатом работы формальной математической модели (МППР). Свойства последовательностей (длина, сложность) напрямую коррелируют с известными теоретическими и практическими знаниями об уязвимостях разных типов моделей ИИ. Полученные результаты (снижение AUC, AP) объективно подтверждают, что воздействия оказывают запланированный деструктивный эффект.

4.3 Экспериментальная оценка применительности моделирования при использовании Q – обучения

Рассмотрим особенности определения стратегий в задачах анализа атак доступных (состояния рассматриваются по Методике ФСТЭК и по базе MITRE ATLAS) в режиме on-line (как наиболее емкую по количеству действий), когда используется неточное описание ПИИ. Проверяется, насколько правильно учитывает модель действительные ограничения и множества доступных путей атак.

Для изучения особенностей определения стратегий, как последовательности атакующих воздействий, в задачах анализа сетевых атак в режиме on-line рассмотрим атаку ARP-спуфинг, которую можно применить при доступе к технической инфраструктуре ПИИ. Инфраструктура предполагает использование диапазона IP-адресов 130.10.1.1 - 130.10.1.100. В заданном диапазоне идентифицируются уязвимости, соответствующие при их успешной эксплуатации (посредством реализации действий) достигаемым состояниям (фиксации успешной

реализации тактики). При этом в режиме реального времени исследуемая система изменчива, поэтому целесообразно применять алгоритмы, которые минимально требовательны к исходным данным и предполагают изначально неполную известность исходных данных политик.

```
{'T1': 'A1_2', 'T2': 'A2_3', 'T3': 'A3_6', 'T6': 'A6_10', 'T8': 'A8_10', 'T10': 'Terminate'}
{'T1': 'A1_2', 'T2': 'A2_3', 'T3': 'A3_6', 'T6': 'A6_8', 'T8': 'A8_10', 'T10': 'Terminate'}
{'T1': 'A1_2', 'T2': 'A2_6', 'T3': 'A3_6', 'T6': 'A6_10', 'T8': 'A2_6', 'T10': 'Terminate'}
```

Рисунок 4.12 - Список состояний и действий моделируемой атаки

Список состояний моделируемой атаки приведен на рисунке 4.13 и иллюстрирует визуализацию последовательности атакующих воздействий на инфраструктурный компонент ПИИ в виде графа состояний и действий.

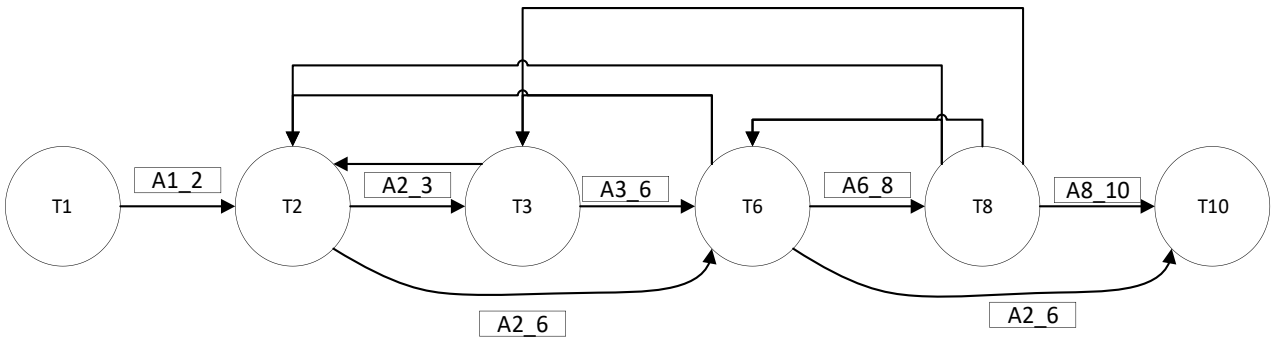


Рисунок 4.13 - Список состояний и действий моделируемой атаки

Круглые вершины отражают тактики ФСТЭК, которые последовательно описывают ход выполнения атаки. Переход между вершинами осуществляется через действие (ребра графа), которое вероятнее всего предпримет злоумышленник в отношении системы, для того чтобы перейти к следующему этапу атаки [134].

Модель МППР предполагает использование второго типа модели (глава 3, Рисунок 3.12 (а)) в режиме on-line (присутствует свобода перемещения злоумышленника между состояниями модели). Множество действий описывается следующим набором (Рисунок 4.14)

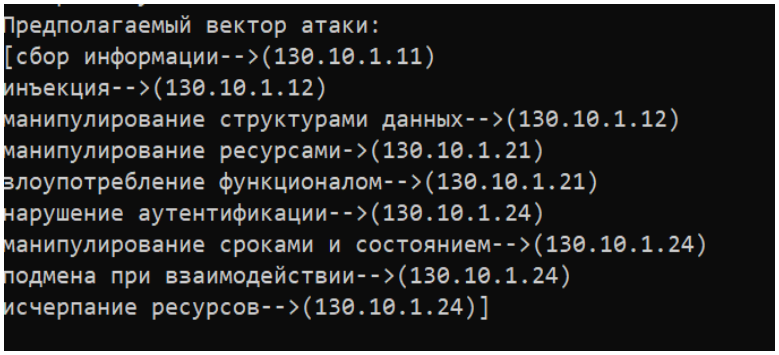


Рисунок 4.14 – Обнаруженный вектор атаки

Описанная атака характеризуется следующим: реализацией в реальном режиме времени, динамичностью состояний, неполнотой описания модели, при поиске стратегий в реальном режиме времени злоумышленник может обнаружить ранее неучтенные уязвимости, наличием обратных связей. В данном случае для определения стратегий были выбраны методы, классификационно принадлежащие к методу Q-обучение [134-135]. Изучаемые стратегии выбора действий (горизонт их планирования конечный) следующие:

1. Стратегия ε-greedy. Позволяет выбирать действие, максимизирующее текущую оценку функции полезности (greedy-action). При этом может выбираться случайное действие с вероятностью ε, и обеспечивается баланс между эксплуатацией (выбором лучшего известного действия) и исследованием (выбором случайного действия).

2. Стратегия greedy. В стратегии всегда выбирается действие, максимизирующее текущую оценку функции полезности. Не обеспечивает исследование альтернативных действий, что может привести к остановке в локальном оптимуме.

3. Стратегия softmax. Позволяет выбирать действия пропорционально их оценкам функции полезности. Вероятность выбора действия a_i пропорциональна

exp(Q(s,a)/τ), где τ - параметр времени, обеспечивает более плавное распределение вероятностей между действиями, чем e-greedy (Рисунки 4.15 - 4.16), позволяет сфокусировать поиск на наиболее перспективных действиях, но может пренебрегать менее перспективными действиями.

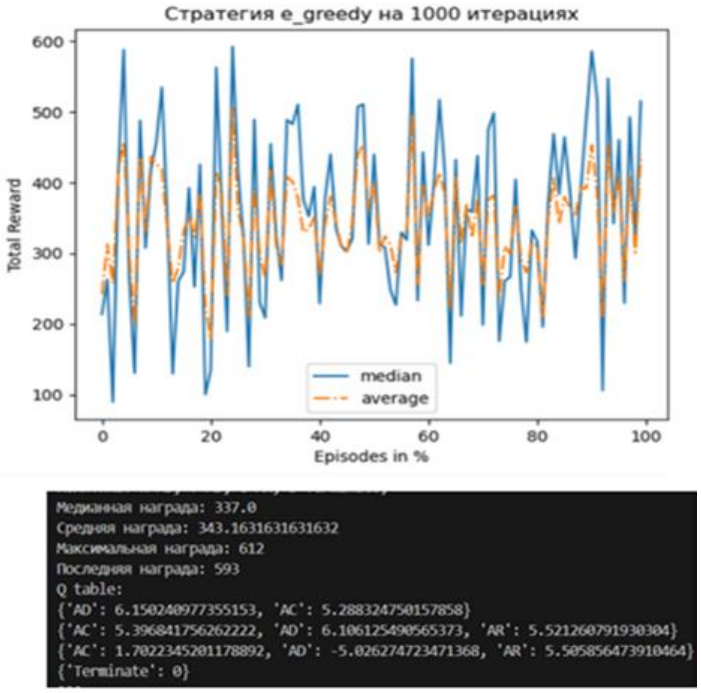


Рисунок 4.15 - Реализация e-greedy стратегии

Стратегия USB (Upper Confidence Bound) является еще одним подходом, который позволяет выбирать действия, основываясь на верхней границе доверительного интервала для оценок функции полезности. Она направлена на оптимальное исследование пространства действий. Проведенные подготовительные исследования выявили следующие:

1. При 100 000 итераций по поиску лучшей стратегии (методы e-greedy, greedy, softmax) было определено: стратегия greedy работает быстрее остальных, но является наименее точной; softmax имеет хорошую точность на базовых параметрах (но поиск требует больше итераций в сравнении с остальными методами (медленная работа)).

2. При 100 раундах стратегия e-greedy показывает ту же точность, что и ранее. Это означает, что ей не требуется в данной задаче много раундов для обучения (что экономит вычислительные ресурсы).

3.

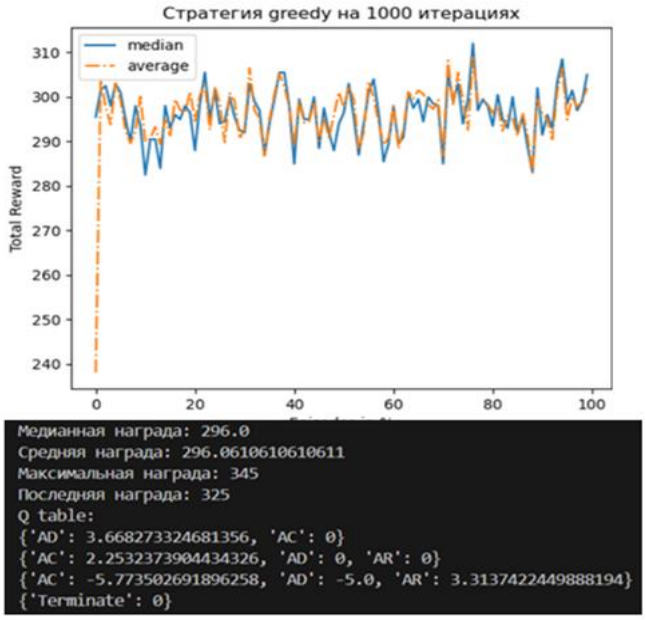


Рисунок 4.16 - Реализация стратегии greedy

Далее приводятся результаты исследования: полученные стратегии по e-greedy, greedy, usb (медианные значения по 1% от итераций (общая награда за раунд)) [134].

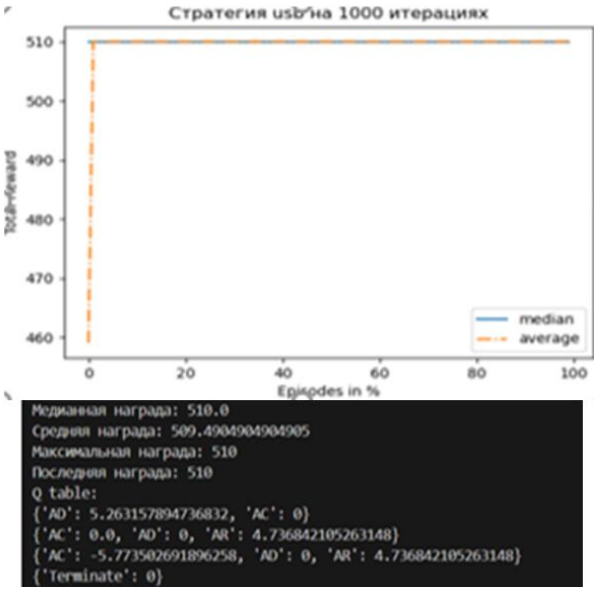


Рисунок 4.17 - Реализация usb-стратегии

При этом следует отметить, что некоторые стратегии (greedy и usb) для данной задачи опираясь на собственные результаты зацикливались и не выходили из цикла при 1000 и более итераций.

Соответственно был введён параметр дальности обзора 3000, внутри раунда обучения проводилась проверка на сходимость, чтобы при условии, что дальность обзора превышена досрочно прогон не завершался.

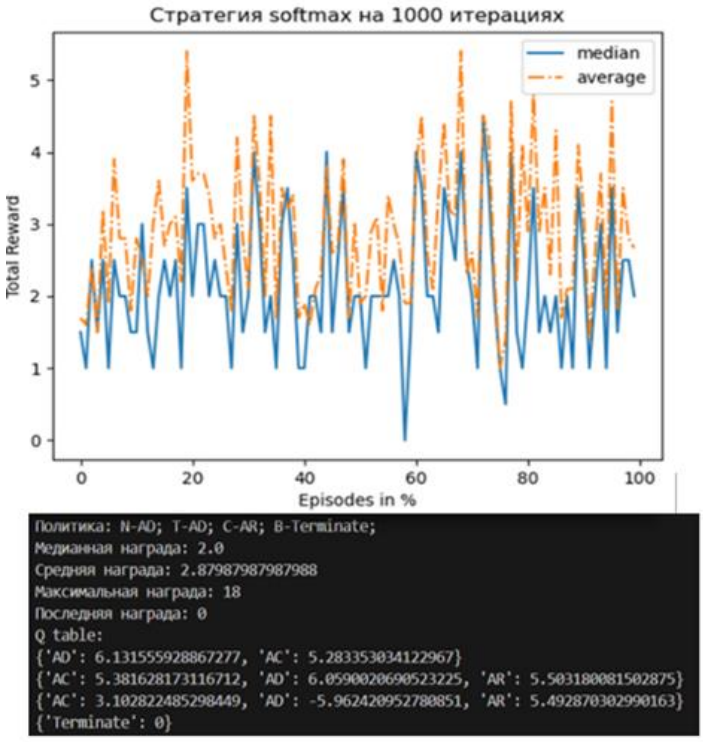


Рисунок 4.18 - Реализация softmax-стратегии

Стратегии greedy и usb, применяемые для данной задачи, опираясь на свои алгоритмически получаемые данные (специфика их алгоритма), вызывали коллизии: вычислительный процесс попадал в бесконечный цикл и не выходил из него. При этом был введён параметр дальности обзора - 3000 итераций. Внутри раунда обучения проводилась проверка на сходимость, чтобы прогон не завершался, если дальность обзора превышена досрочно [134, 135].

В целом можно отметить следующее:

1. Стратегия e-greedy проявила себя лучше всех, поскольку позволила исследовать большое количество состояний, аналогично softmax (в результате обе дали одинаковые политики).

2. Стратегии usb, greedy плохо исследовали среду, однако гарантированно отводит агента от терминального состояния (выделяются состояния challenged – reset).

3. Стратегия softmax проявила себя лучше всех, так как исследовала наибольшее количество состояний.

При рассмотрении модели атаки с возвратными состояниями в предложенных стратегиях при условии противодействия атакующим воздействиям задача определения эффективной стратегии атаки усложняется.

При определении особенностей поиска наилучших последовательностей действий атакующего были рассмотрены специфические особенности применения МППР как метода моделирования. При этом в процесс моделирования были интегрированы методы описания атак Методики ФСТЭК.

На основании методических ограничений и режимов аудита были определены режимы моделирования. Было выявлено, что в режиме off-line целесообразно применять поиск наилучших последовательностей атакующих воздействий с помощью поиска по значениям, а в режиме on-line использовать методы поиска по стратегиям или Q-обучения и классификационно схожие с ними.

При построении моделей в качестве входных данных использовались метрики уязвимостей, которые были классифицированы в соответствии с методами их эксплуатации, и, соответственно, по принадлежности к тактикам (следует учесть, что оценка уязвимости может носить экспертный характер, если четкого соотнесения с тактиками не наблюдается) [134].

4.4 Сравнительный анализ с альтернативными решениями в области моделирования

Структура лабораторной установки для испытания модели на уровне работы технической инфраструктуры, состоящей из таких компонентов, как автоматизированные рабочие места, межсетевые экраны, коммутаторы, роутеры, серверы и другие устройства, приведена на рисунке 4.4. Межсетевые экраны обеспечивают безопасность сети, фильтруя входящий и исходящий трафик.

Коммутаторы и роутеры играют ключевую роль в соединении и маршрутизации трафика между компьютерами и серверами внутри организации, а также подключении к внешним сетям, таким как Интернет. Серверы отвечают за хранение и обработку данных, предоставление доступа к ресурсам и приложениям для пользователей.

Инструмент для расчета вектора атаки обращается к серверу агрегации по защищенному протоколу HTTPS. При отправке запроса по адресу 'https://130.10.1.13:7225/getVulnJSON’ инструмент передает идентификатор сети, для которой необходимо получить данные об уязвимостях. Сервер обрабатывает запрос, выполняет поиск соответствующих уязвимостей в базе данных и формирует ответ в формате JSON, содержащий список уязвимостей и их метрики. Инструмент для расчета вектора атаки обрабатывает полученные данные, используя их в качестве входных параметров для своих алгоритмов.

По итогам проведения аудита были получены все данные, необходимые для корректных расчетов в рамках используемой математической модели. Эти входные данные включают в себя IP-адреса и информацию о выявленных уязвимостях на различных компонентах сети, которые могут быть использованы злоумышленниками для несанкционированного доступа и компрометации системы, их числовые метрики и способы эксплуатации. На рисунке 4.19 показан фрагмент полученного файла с отчетом по уязвимостям. На рисунке развернута цепочка действий злоумышленника при реализации атаки путем перехода между хостами сети, возможность компрометации которых определяется из матрицы переходных вероятностей

с учетом наличия уязвимостей для реализации переходов между промежуточными узлами.

Для более наглядного отражения полученных результатов может быть построен граф, на котором указаны возможные пути совершения атаки, а также наиболее оптимальный. В данной работе для этой цели использовалась библиотека Python NetworkX.

```
Выберите пункт: 2
Предполагаемый вектор атаки:
[сбор информации-->(130.10.1.11)
инъекция-->(130.10.1.12)
манипулирование структурами данных-->(130.10.1.12)
манипулирование ресурсами->(130.10.1.21)
злоупотребление функционалом-->(130.10.1.21)
нарушение аутентификации-->(130.10.1.24)
манипулирование сроками и состоянием-->(130.10.1.24)
подмена при взаимодействии-->(130.10.1.24)
исчерпание ресурсов-->(130.10.1.24)]
```

Рисунок 4.19 – Предполагаемый вектор атаки

Рисунок 4.20 иллюстрирует визуализацию цепочки атакующих воздействий на компонент системы в виде графа. Круглые вершины отражают тактики ФСТЭК, которые последовательно описывают ход выполнения атаки. Переход между вершинами осуществляется через ромб, обозначающий действие, которое вероятнее всего предпримет злоумышленник в отношении системы, для того чтобы перейти к следующему этапу атаки.

В случае наличия или отсутствия уязвимостей, подходящих под тот или иной способ эксплуатации, количество доступных действий (ромбов, входящих в вершину) может варьироваться. Ребра, выделенные жирным, указывают на наиболее вероятный вектор атаки.

Анализируя граф можно отметить, что на некоторых этапах атаки злоумышленнику не удается успешно эксплуатировать какую-либо из уязвимостей. Это может быть связано с переходом злоумышленника в узел, ограничивающий его доступ к следующим шагам атаки. Для преодоления этой проблемы атакующий имеет возможность вернуться в начальное состояние для повторного сбора информации и выбора нового пути воздействия на систему.

Таким образом, инструмент для определения вектора атаки может быть полезен для специалистов ИБ, поскольку он может опираться на распространенные методологии MITRE или применяемые во ФСТЭК [7], и позволяет моделировать и анализировать пути, которые могут быть использованы злоумышленниками для проникновения в систему и компрометации ее данных. Помимо понимания того, какие узлы могут быть затронуты атакой, указаны способы воздействия на них. Такое решение может помочь более оперативно принимать меры для предотвращения или уменьшения риска атаки.

В современной информационной среде можно ряд выделить популярных решений с частично схожим функционалом:

1. Инструменты с использованием МППР и их применение для ИИ:

- MITRE CALDERA — моделирование атак на ML-модели (ATLAS).
- SafeBreach — тестирование ML-пайплайнов.
- MLSec — анализ рисков в ML-системах.
- Counterfit — adversarial-тестирование ML-моделей.
- MulVal — анализ уязвимостей в ML-инфраструктуре.

2. Инструменты без использования МППР и их назначение:

- САЗ — сканирование уязвимостей.
- COMNET III — сетевое моделирование.
- SecurITree — построение и анализ деревьев атак.
- Symantec ESM — корреляция событий безопасности.
- Skybox Security — анализ сетевых угроз.

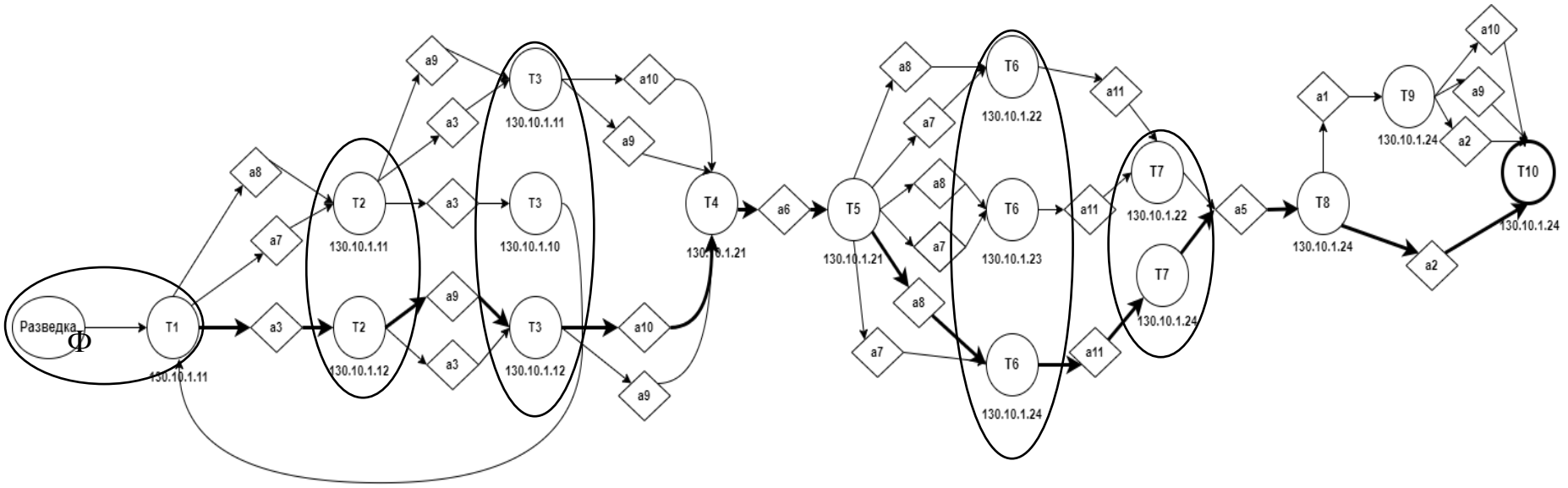


Рисунок 4.20 – Визуализация последовательности атакующих воздействий на инфраструктурный компонент системы (нумерация действий приводится в соответствии с их порядком в тактике (в состоянии))

Можно провести качественный анализ и сравнить их с предложенным решением. Критерии сравнения, следующие:

1. Полнота 1. Учет специфики действий злоумышленника при атаке:

– учитываются типы действий: нелегальное взаимодействие (*D*), легальное взаимодействие (*C*);

– учитываются обратные действия атакующего, включая возвращение в предыдущее состояние атаки (действия сброса (*R*)) в режимах on-line.

2. Полнота 2 и охват. Процент покрытия угроз из MITRE ATLAS.

3. Покрытие всех компонентов. Сервис, модель ИИ, хост.

4. Уровни детализации. Выявляются три уровня абстракции:

– первый уровень: состояния модели определяются как наборы состояний, позволяющих реализовать действия, связанные с логикой функционирования вычислительной модели ПИИ в период атакующих воздействий;

– второй уровень: состояния модели определяются как наборы состояний, позволяющих реализовать действия, связанные с инфраструктурой и логикой функционирования ПИИ в период атакующих воздействий. Модель используется для общего описания системы в период атаки с уточнением последовательности действий злоумышленника, связанных с этапами реализации атакующих воздействий;

– третий уровень: состояния модели определяются как наборы событий, позволяющих реализовать действия, приведенные в тактиках MITRE и методиках ФСТЭК, связанных с инфраструктурой и логикой функционирования модели и инфраструктуры ПИИ в период атакующих воздействий.

5. Учёт последовательности. Стремление к целевому состоянию MITRE.

6. Гибкость настройки. Обеспечивается функцией вознаграждения, изменение параметров которой может использоваться для оценки вложений злоумышленника через CVSS.

7. Реалистичность. Количество сценариев, подтвержденных реальными кейсами (например, из MITRE).

8. Интеграция с MITRE ATLAS. Возможность применения тактик и методов MITRE ATLAS

9. Специфичность для ИИ.

Результаты сравнения следующие (для оценок «Высокая», «Средняя» и «Низкая» по каждому критерию для каждой системы использовались качественные характеристики: насколько полно отвечает характеристика системы заданному критерию):

1. Полнота 1 предполагает учет специфики действий злоумышленника. Предлагаемая методика (оценка «Высокая») использует МППР для моделирования всех возможных действий злоумышленника, включая обратные действия и возвраты, что позволяет учитывать динамику атак в режимах on-line и off-line.

В САЗ (оценка «Низкая») основное внимание уделяется выявлению уязвимостей, а не моделированию действий злоумышленника, при этом не учитываются обратные действия или специфика атак на ИИ. COMNET III (оценка «Средняя») моделирует сетевые взаимодействия, но не отражает подробно специфику действий злоумышленника при атаках на ИИ, особенно в контексте обратных переходов. SecurITree (оценка «Средняя) использует деревья атак для моделирования, но нет явного учета обратных действий. CAULDRON (оценка «Средняя») моделирует атаки, но не предоставляет столь же глубокого анализа динамики атак, как МППР, особенно в отношении обратных действий. Symantec ESM (оценка «Низкая») фокусируется на мониторинге инцидентов, а не на моделировании действий злоумышленника, особенно с учетом обратных переходов. Skybox Security (оценка «Средняя) моделирует сетевые угрозы, но не предоставляет динамический анализ обратных действий в контексте ИИ. Counterfit (оценка «Средняя») используется для тестирования облачных приложений, но не специализируется на динамических моделях атак на ИИ. MLSec (оценка «Высокая») использует МППР, что позволяет учитывать динамику атак, включая обратные действия и возврат к предыдущим состояниям. SafeBreach (оценка «Средняя») имитирует атаки, но с ограниченным акцентом на динамические аспекты и обратные действия. MulVal (оценка «Средняя»)

реализует логическое моделирование атак, но без акцента на динамические обратные действия.

2. Полнота 2 и охват. Предлагаемая методика с МППР (оценка «Высокая») специально разработана для интеграции с MITRE ATLAS, что обеспечивает высокое покрытие угроз для ИИ. Оценка других систем варьируется от низкой до средней. Не все системы адаптированы под специфические угрозы для ИИ, как описано в MITRE ATLAS или OWASP ML Top 10. Некоторые могут иметь более широкий охват за счет общих методик анализа безопасности, но не специфичных для ИИ.

3. Покрытие всех компонентов (сервис, модель ИИ, инфраструктурные компоненты). Предлагаемая методика с МППР (оценка «Высокая») разработана для охвата всех компонентов в контексте атак на ИИ, включая модели, сервисы и базовую инфраструктуру. Остальные системы (оценка «Частично») могут фокусироваться на одном или двух аспектах (например, сеть, уязвимости хостов), но не обязательно охватывают все компоненты, специфичные для ИИ.

4. Уровни детализации. Предлагаемая методика с МППР (оценка «Высокая») предлагает три уровня абстракции (детализации) для детального анализа действий злоумышленников, что позволяет адаптировать модель под разные потребности. Другие системы (оценка «Варьируется») предлагают одну или две степени детализации, но не всегда охватывают такие же глубокие уровни анализа, особенно в контексте ИИ.

5. Учёт последовательности стремления злоумышленника к целевому состоянию MITRE. Предлагаемая методика с МППР (оценка «Высокая») интегрирует последовательность действий, ведущих к целевым состояниям атак в соответствии с MITRE ATLAS. Остальные системы (оценка «Варьируется») могут учитывать последовательность атак, но не всегда в контексте MITRE для ИИ, или делают это менее формально.

6. Гибкость настройки. Предложенная методика с МППР (оценка «Высокая») при использовании функции вознаграждения и CVSS позволяет настраивать модель под различные сценарии атак. Остальные системы (оценка «Варьируется»)

предоставляют меньше возможностей для настройки под конкретные сценарии атак на ИИ.

7. Реалистичность. Предлагаемая методика с МППР (оценка «Высокая»): моделирует сложные сценарии с учетом реальных кейсов из MITRE ATLAS. Для остальных систем (оценка «Варьируется») реалистичность зависит от способности системы имитировать реальные сценарии атак, что ограничивает реалистичность.

8. Интеграция с MITRE ATLAS. Предлагаемая методика с МППР (оценка «Высокая») явно интегрирована с MITRE ATLAS. В то же время другие системы, имеющие оценку «Варьируется», могут иметь ограниченную интеграцию или вовсе её не поддерживать, особенно когда речь идет об атаках на искусственный интеллект.

9. Ориентированность на специфику систем ИИ. Предлагаемый способ моделирования специально адаптирован для угроз, связанных с ИИ, с учетом их уникальных аспектов. Для сравнения: большинство других систем не всегда учитывают специфику атак на ИИ, что может привести к недостаточному охвату угроз.

4.5 Особенности программного решения

Серверная и клиентская части являются двумя основными компонентами, выполняющими различные функции в рамках программного решения. При этом каждая из них имеет собственный набор подключаемых модулей и зависимостей, которые необходимы для обеспечения их функциональных возможностей (Рисунок 4.21).

Состав программного решения (системы):

1. Модуль сбора данных о параметрах атак (клиентская часть).
2. Модуль анализа данных и сохранения результатов анализа (серверная часть).
3. Анализатор событий или других параметров.

4. Консультационная подсистема, использующая метод МППР, интегрированный в общую подсистему консультации.

Серверная часть программы является компонентом, который обеспечивает управление и обработку данных, полученных от клиентской части приложения и обеспечивает вывод информации пользователю.

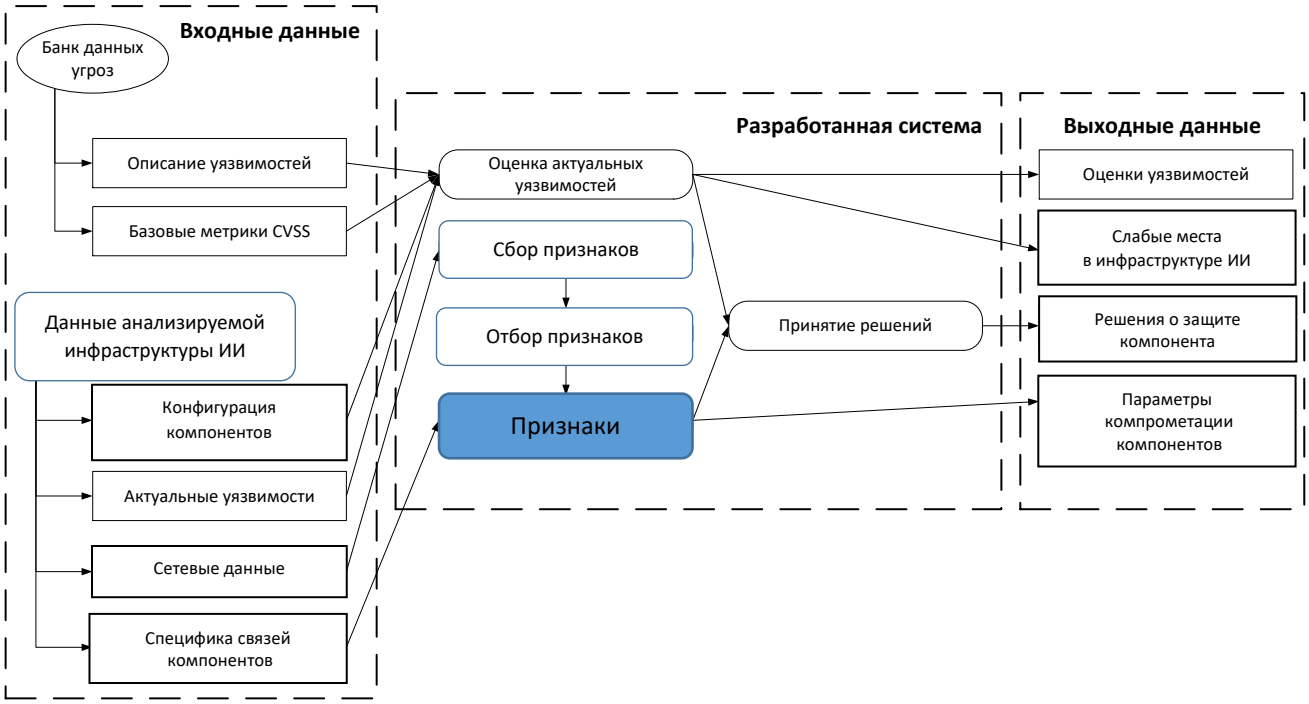


Рисунок 4.21 – Основные потоки данных системы разработанного системы

Главная функция клиентской части — это собирать информацию о текущем состоянии компонента инфраструктуры ИИ (или узла), на котором установлена клиентская часть, и отправлять эти данные на сторону серверной части программы. Сбор данных осуществляется при помощи подключаемых модулей, каждый модуль собирает информацию по конкретному виду аппаратных ресурсов.

В свою очередь, клиентская часть программы предназначена для взаимодействия с защищаемой системой. Она облегчает сбор и передачу данных на серверную часть. Кроме того, клиентская часть может содержать механизмы взаимодействия с другими системами, которые необходимы для реализации задач пользователя. Таким образом, серверная и клиентская части участвуют в различных задачах

в рамках приложения, имеют разные наборы функциональных возможностей и зависимостей, и каждая из них играет важную роль в обеспечении работы программного продукта в целом (Рисунок 4.22).

Рисунок 4.22 - Обобщенная архитектура системы сбора данных

Последовательность работы основной части системы приведена на рисунке 4.23. Предусматривается получение данных об инцидентах безопасности, их последующая обработка с выделением параметров, используемых при описании модели в части касающейся формализации награды. Выделяются следующие параметры:

1. Значение CVSS (отражает уровень угрозы и является основным компонентом вознаграждения).

2. Фиксация возможности взаимодействия между компонентами (узлами сети), компонентами инфраструктуры позволяет учитывать взаимосвязь уязвимости в контексте активного взаимодействия.

3. Сопряжения уязвимостей по типу эксплуатация важна для оценки комбинированного эффекта уязвимостей.

4. Метрики приближения к целевому состоянию по методике MITRE ATLAS, что позволяет учитывать, насколько переход в новое состояние может быть выгоден для злоумышленника.

5. Для расчета метрик приближения к компоненту-носителю целевой уязвимости используется файл hosts.json.

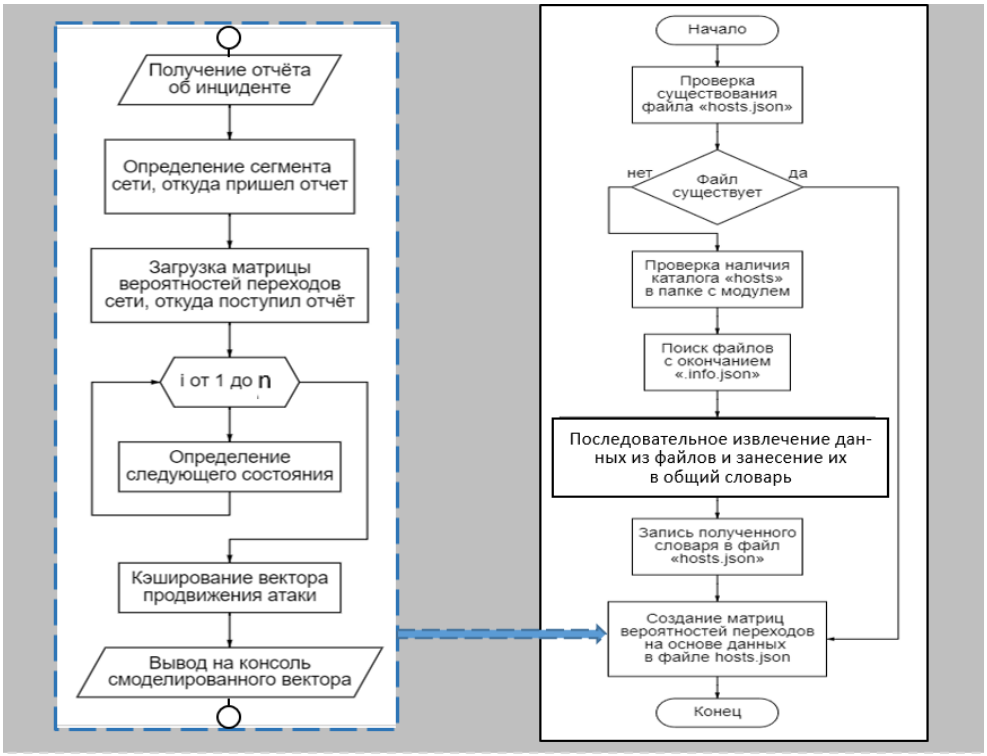


Рисунок 4.23 - Последовательность работы системы анализа

Далее представлена процедура изучения части атакующей последовательности с целью выявления динамики, включающей пять состояний. Рассмотрим вариант атаки с двумя типами действий. Действия следующие: успех выполнения действия (A1), неуспех выполнения действия (A2). Состояния взяты из модели второго типа. В таком случае награды за A1 (успех) определяются как мера уязвимости в текущем состоянии. Награды за A2 (неуспех) можно определять как величину

контрмер, закрывающих уязвимость для действия A1. Таким образом, если $A1_{ij} < A2_{ij}$ контрмеры перекрывают текущие уязвимости в данном элементе матрицы.

1. Условие 1: Награды за A1 больше наград за A2 (Рисунок 4.24).

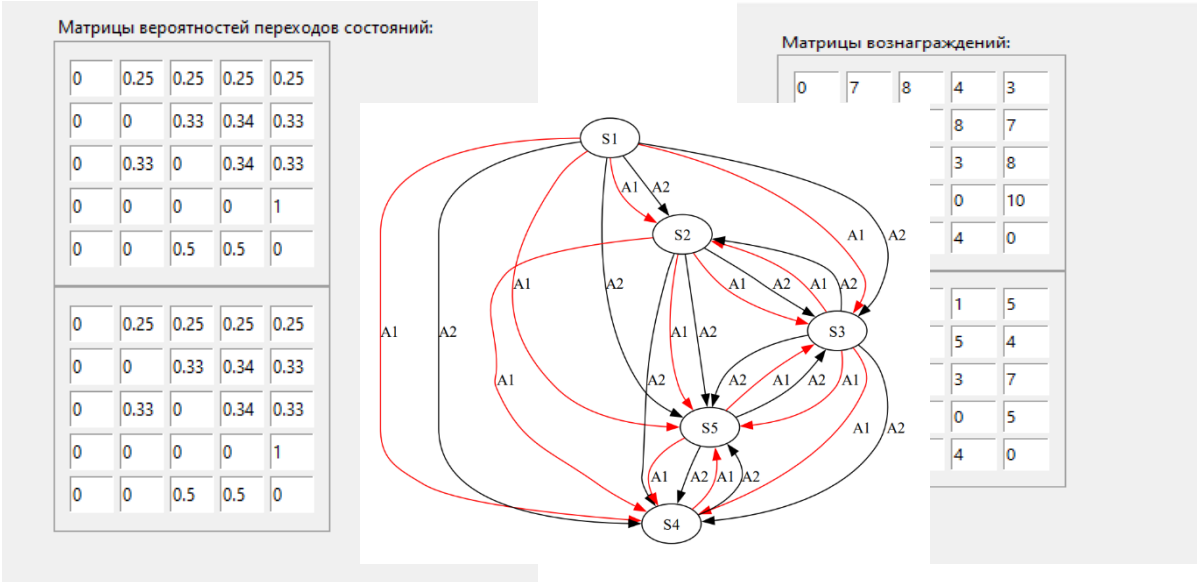


Рисунок 4.24 – Пример, в котором все награды за успех больше величин контрмер

Таким образом, все оптимальные действия для злоумышленника — это успех. Злоумышленник может пойти по любому из путей в случае равновероятных событий.

2. Условие 2: Награды за A1 меньше наград за A2 (Рисунок 4.25).

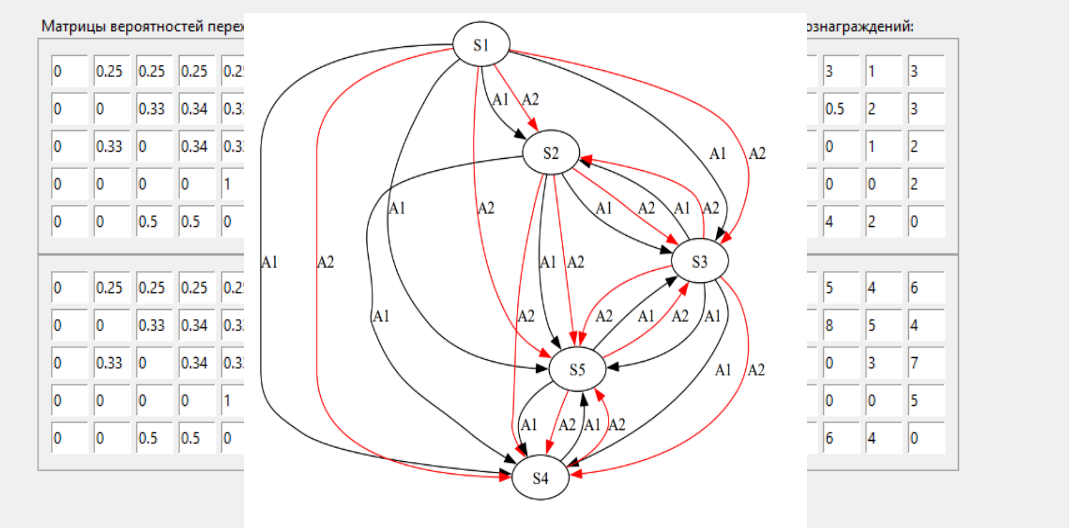


Рисунок 4.25 – Оптимальное действие для каждого состояния – это не успех

При установленной матрице контрмер злоумышленник не сможет осуществить выполнение действия A1. Выделяется набор оптимальных действий (Рисунок 4.26).

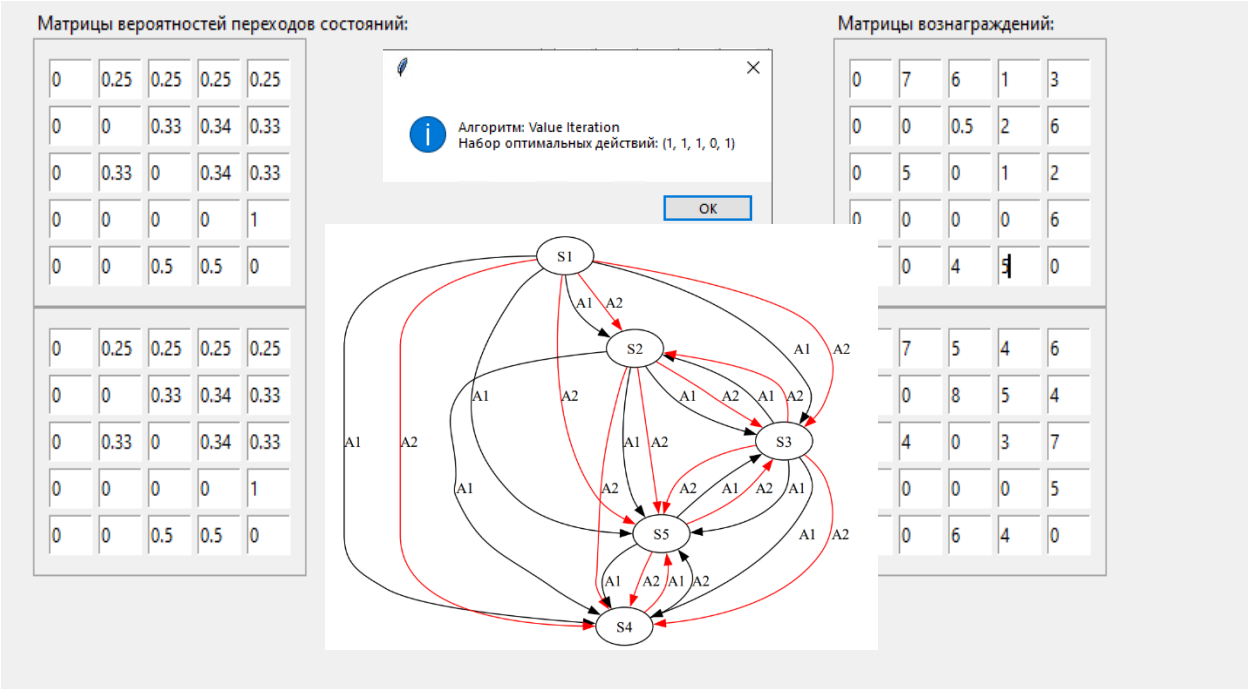


Рисунок 4.26 – Определение оптимального пути злоумышленника

Таким образом, данный подход позволяет учитывать величины уязвимостей и установленных контрмер в анализируемой системе. Результаты можно интерпретировать как поиск переходов между состояниями, по которым злоумышленник может достичь результата.

Разработанный алгоритм процесса аудита безопасности систем (подсистем) искусственного интеллекта и составления модели угроз с учетом проведения предложенного моделирования атак и последующего построения сценария атак содержит следующие шаги (учитывается доступ к вычислительным возможностям предложенной системы):

Шаг 1. Определение целей аудита. Требуется установить, какие аспекты системы ИИ необходимо протестировать (например, устойчивость к атакам, безопасность данных, эффективность защиты). Требуется выбрать типы атак, определить,

какие типы атак будут моделироваться (например, атаки уклонения, отравления, извлечения).

Шаг 2. Сбор информации. Требуется произвести анализ системы, изучить архитектуру системы ИИ, включая используемые алгоритмы и данные. Произвести сбор данных о потенциальных уязвимостях: использовать базы данных, такие как MITRE ATTLAS, для понимания существующих техник атак.

Шаг 3. Построение модели МППР. Требуется определить возможные состояния системы (например, нейтральное, доверенное, оспариваемое, заблокированное).

1. Необходимо установить возможные действия злоумышленника (например, легитимное (сотрудничество), нелегитимное, сброс).

2. Учесть логические ограничения. Логические зависимости между различными состояниями системы могут ограничивать возможности злоумышленника в реализации атак. Например, если определенные действия могут быть выполнены только из конкретных состояний, это создает дополнительные барьеры для успешного выполнения атаки.

3. Учесть специфику структуры алгоритмов. Алгоритмы ИИ могут иметь встроенные логические ограничения, которые определяют их поведение в ответ на различные входные данные. Понимание этих зависимостей позволяет более точно моделировать возможные сценарии атак.

4. Задать вероятности переходов: определить вероятности перехода между состояниями в зависимости от выбранных действий.

5. Установить правила вознаграждений.

Шаг 4. Моделирование атак. Использовать модель МППР для симуляции последовательности действий атакующей стороны.

1. Выбрать оптимальные действия и определить список возможных состояний получаемых последовательностей.

2. Определить вознаграждение за переход и обновить счетчик шагов.

3. Повторять процесс расчета до достижения заблокированного состояния, наименее ценной последовательности или максимального количества шагов.

Шаг 5. Анализ результатов.

1. Сбор данных о производительности модели: зафиксировать накопленные вознаграждения и количество шагов до блокировки, наименее ценной последовательности или максимального количества шагов.

2. Проанализировать результаты для выявления слабых мест в системе.

Шаг 6. Построение сценариев атак.

1. На основе полученных данных создать детализированные сценарии атак, используя ранжированные по степени опасности последовательности атакующих воздействий (при многочисленности последовательностей рекомендуется установить уровень опасности (ценности), после которого последовательности воздействий не будут считаться по значимыми в соответсвии с решением специалиста ИБ).

2. Использовать методы описания атак для визуализации последовательности действий злоумышленника.

3. Определить ключевые этапы каждой атаки и возможные пути обхода защиты.

4. Записать сценарии в формате, удобном для анализа и дальнейшего использования в обучении персонала по кибербезопасности.

Шаг 7. Рекомендации по защите. На основе анализа предложить конкретные меры по повышению устойчивости системы к атакам.

Приведенный алгоритм дополняет методику моделирования, изложенную в третьей главе. Итоговый вид формируемых последовательностей является основой для описания сценария атак, поскольку показывает фактически все пути атакующих воздействий в заданной инфраструктуре системы ИИ при учете уровня детализации при моделировании (Рисунки 4.27 – 4.28). При этом следует отметить, что уровень детализации можно регулировать.

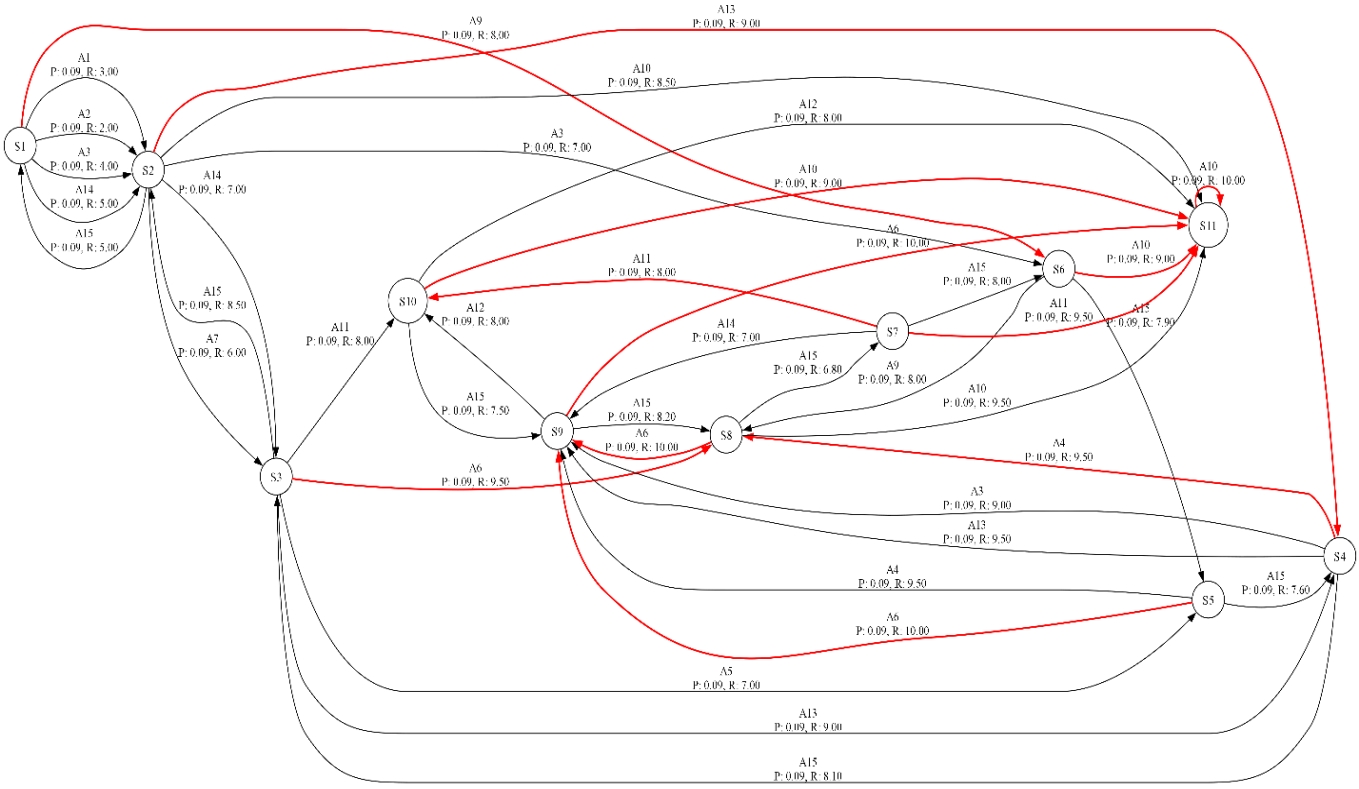


Рисунок 4.27 – Общая последовательность действий для последующего составления сценария атаки в процессе аудита (оптимальные политики)

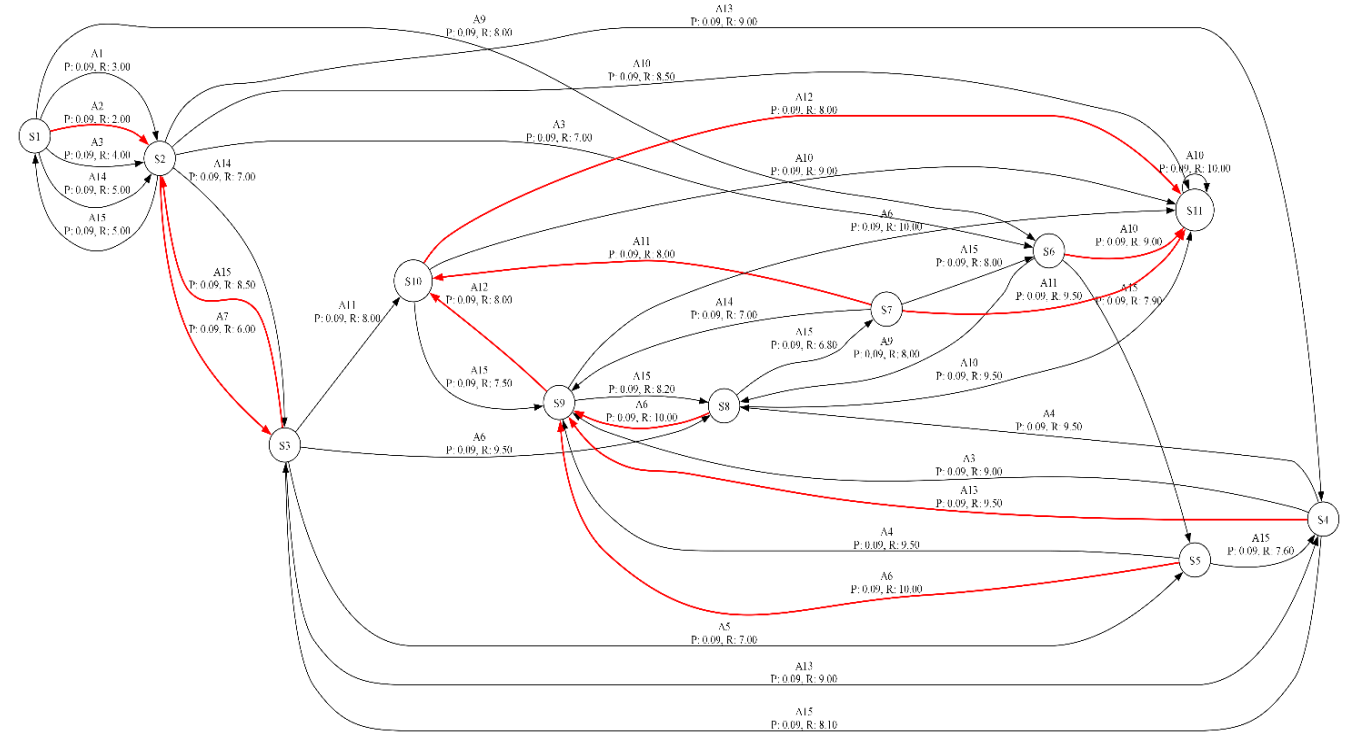


Рисунок 4.28 – Общая последовательность действий для последующего составления сценария атаки в процессе аудита (альтернативные политики)

Эта методика позволяет систематически подходить к моделированию атак на системы искусственного интеллекта и эффективно проводить аудит их безопасности. Использование методов МППР обеспечивает глубокое понимание угроз и помогает разработать адекватные меры защиты.

Выводы к главе 4

В главе проанализированы разработанные модели. Они выделяются по параметру полноты, демонстрируя значительное превосходство над сравниваемыми аналогами. Кроме того, они обладают уникальной способностью интегрировать тактики различных методик описания атак в единый процесс моделирования, что позволяет значительно расширить их функциональность и адаптивность к сложным и многогранным сценариям кибератак на системы ИИ. Данный подход способствует более глубокому пониманию природы атак и разработке более эффективных способов их предотвращения и нейтрализации.

Рассмотрена специфика использования дополнительных алгоритмов моделирования в рамках предложенного набора состояний и действий разработанных моделей. При создании среды испытаний моделей рассмотрены специфические особенности формирования нейронных сетей, используемых для тестирования атакующих воздействий. На основе изменений точности нейросети определялась их уязвимость для последующего тестирования. В результате, моделирование позволило определить специфику изменений классифицирования, что в итоге подтвердилось в экспериментальном порядке. Рассматривались также последовательности действий на способность следовать логике ограничений инфраструктурного типа. Последовательность состояний подтвердила ограничения, заданные при эксперименте. Также предложено программное решение, включающее в свой состав серверную часть и клиентскую часть. Программное решение предназначено для автоматизации процесса моделирования с последующим выводом результатов в виде, применимом для составления сценария атак при аудите ИБ ИИ. Предложен алгоритм применения моделей в процессе аудита с учетом автоматизации.

Заключение

Несмотря на значительное количество исследований в области анализа защищенности систем искусственного интеллекта с использованием подходов, основанных на моделировании атак, проблема анализа атакующих воздействий требует дополнительных исследований. В данной работе рассматривались атаки, направленные на эксплуатацию уязвимостей моделей и архитектур подсистем искусственного интеллекта в контексте общей архитектуры корпоративной информационной системы, процессы построения и анализа атакующих последовательностей для повышения качества аудита защищенности систем искусственного интеллекта. В результате были разработаны модели и методика, позволяющие реализовывать процессы аудита применительно к системам ИИ с учетом возможности использования известных баз знаний об атаках на вычислительные модели и компоненты систем ИИ, в том числе, если они выступают в роли подсистем ИС. Основные результаты работы следующие:

1. Определены возможности использования доступных наборов данных для разработки модели построения и анализа атак на системы ИИ и определены параметры используемые при моделировании. Осуществлен выбор стандартов описания уязвимостей и шаблонов атак, позволивший использовать открытые базы знаний для получения информации о возможных угрозах.

2. Разработаны модели построения последовательностей атакующих воздействий на системы ИИ, алгоритм моделирования атак для определения наиболее опасного набора этапов атаки и действий злоумышленника. Предложенные модели позволяют повысить полноту описания возможных атакующих воздействий. Итеративный подход к анализу дает возможность получать результаты на ранних стадиях моделирования.

3. Предложена методика применения моделей МППР в процессе аудита защищенности систем ИИ с учетом полноты исследования проблем информационной безопасности. При применении методики разработанные модели позволяют

получать результаты, сопоставимые по точности с моделями существующих аналогичных подходов, однако при этом разработанные модели превосходят аналогичные по полноте описания атак. Предложенное программное средство на основе данной методики обеспечивает возможность аналитического моделирования атак на ПИИ.

4. Разработана архитектура и реализован прототип системы, основанный на предложенной методике. Проведена оценка процесса построения и анализа моделей атак.

Эксперименты и теоретические оценки подтверждают работоспособность моделей при учете различных режимов моделирования. При этом сохраняется полнота описания при необходимости. Адаптивность модели позволяет легко менять наборы состояний и действий, позволяет учитывать непредсказуемость поведения злоумышленника при недостаточной информации о мотивации нарушителя. Результаты работы позволяют повысить качество аудита защищенности систем ИИ, который проводится по правилам методических документов регуляторов в области ИБ (ФСТЭК) и может значительно повысить эффективность существующих средств защиты информации.

Список использованной литературы и электронных ресурсов

1. Методический документ: Методика оценки угроз безопасности информации: утвержден ФСТЭК России 5 февраля 2021 г. [Электронный ресурс] – Режим доступа к рес.: https://fstec.ru/files/495/---5--2021-/891/---5--2021-.pdf (дата обращения: 30.10.2023)
2. Банк данных угроз БИ ФСТЭК России. [Электронный ресурс] – Режим доступа к рес.: https://bdu.fstec.ru/vul/ (дата обращения: 30.10.2023)
3. ГОСТ Р ИСО/МЭК 27005-2010. Информационная технология. Методы и средства обеспечения безопасности. Менеджмент риска информационной безопасности. – Введ. 2009-05-25. – Москва: Изд-во стандартов.
4. ГОСТ Р ИСО/МЭК 15408-1-2008. Информационная технология. Методы и средства обеспечения безопасности. Критерии оценки безопасности информационных технологий. Часть 1. Введение и общая модель. – Введ. 2009-05-25. – Москва: Изд-во стандартов.
5. ГОСТ Р ИСО/МЭК ТО 18044-2007 Информационная технология. Методы и средства обеспечения безопасности. Менеджмент инцидентов ИБ. [Электронный ресурс] – Режим доступа к рес.: https://docs.cntd.ru/document /1200068822 (дата обращения: 30.10.2021)
6. ГОСТ Р 59277—2020. Системы искусственного интеллекта. Классификация систем искусственного интеллекта. Утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и метрологии от 23 декабря 2020 г. № 1372-ст. [Электронный ресурс]. – Режим доступа к рес.: https://docs.cntd.ru/document/1200177292 (дата обращения: 30.10.2021).
7. ГОСТ Р ИСО/МЭК 13335-1-2006 Информационная технология. Методы и средства обеспечения безопасности. Часть 1. Концепция и модели менеджмента без-

опасности информационных и телекоммуникационных технологий. [Электронный ресурс] – Режим доступа к рес.: https://docs.cntd.ru/document/ 1200048398 (дата обращения: 30.10.2021)

8. Отчет компании Positive Technologies «Актуальные киберугрозы: I квартал 2022 года. [Электронный ресурс] – Режим доступа к рес.: https://www.ptsecurity.com/ru-ru/research/analytics/cybersecurity-threatscape-2022-q1/#id2 (дата обращения: 25.05.2022)
9. Володько, Л.П. Оценка качества банковских информационных технологий и услуг в условиях неопределенностей / Л.П. Володько // Аудит и финансовый анализ. – 2010. – № 3. – С. 1-12.
10. Намиот, Д. Е. Схемы атак на модели машинного обучения / Д. Е. Намиот // International journal of open information technologies. – 2023 – Т.11, № 5. – С. 68-86.
11. Common Attack Pattern Enumerations and Classifications [Электронный ресурс]. – Режим доступа к рес.: https://capec.mitre.org/ (12.05.2024).
12. Ветров, И.А. Особенности подготовки активного аудита информационной безопасности АСУТП / И.А. Ветров, В.В. Подтопельный // Вестник Балтийского федерального университета им. И. Канта. Серия: Физико-математические и технические науки. - 2021. - № 1. - С. 5-11.
13. Подтопельный, В.В. Определение пригодности правил обнаружения сетевых вторжений и их математическая оценка / В.В. Подтопельный, И.А. Ветров // Вестник Балтийского федерального университета им. И. Канта. Серия: Физико-математические и технические науки. - 2021. - № 2. - С. 11-18.
14. Подтопельный, В.В. Сравнительный анализ технологий аудита информационной безопасности сетевой инфраструктуры диспетчерского уровня АСУТП / В.В. Подтопельный // БАЛТИЙСКИЙ МОРСКОЙ ФОРУМ: VIII Международный Балтийский морской форума: материалы: в 6 т. - 2020. - С. 306-311.
15. Подтопельный, В.В. Особенности формирования siem-правил в АСУТП / В.В. Подтопельный // Научный аспект. - 2020. - Т. 4, № 4. - С. 480-484.

16. Подтопельный, В.В. Особенности формирования сигнатурных последовательностей для обнаружения сетевых атак в АСУТП / В.В. Подтопельный // Modern Science. - 2020. - № 12-3. - С. - 303-307.
17. Common Vulnerability Scoring System Calculator. [Электронный ресурс] – Режим доступа к рес.: https://nvd.nist.gov/vuln-metrics/cvss/v2-calculator (13.05.2024).
18. Брачо, А. А. Платформа на основе моделирования для оценки воздействия киберугроз на интеллектуальные производственные системы / А. Брачо, К. Сэйгин, Х. Ван [и др.] // Procedia Manufacturing. – 2018. – № 26. – С. 1116–1127. [Электронный ресурс] – Режим доступа к рес.: https://doi.org/10.1016/j.promfg.2018.07.148. (дата обращения: 08.07.2021).
19. MITRE ATLAS // MITRE ATT&CK [Электронный ресурс] – Режим доступа к рес.: https://atlas.mitre.org, свободный (дата обращения: 02.05.2024)
20. NIST (National Institute of Standards and Technology) [Электронный ресурс] – Режим доступа к рес.: https://www.nist.gov. (дата обращения: 08.07.2023).
21. Ясасин, Э. Forecasting IT Security Vulnerabilities - An Empirical Analysis = Прогнозирование уязвимостей ИТ-безопасности - эмпирический анализ / Э. Ясасин, Д. Престер, Г. Вагнер [и др.] // Computers & Security. – 2020. – № 88. – С. 1–24. – [Электронный ресурс] – Режим доступа: к рес.: https://doi.org/10.1016/j.cose.2019.101610 (дата обращения: 08.07.2021).
22. Сюн, Ц. Construction of information network vulnerability threat assessment model for CPS risk assessment / Построение модели оценки угроз уязвимости информационной сети для оценки рисков CPS / Ц. Сюн, Ц. Ву // Computer Communications. – 2020. – № 155. – С. 197–204. [Электронный ресурс] – Режим доступа к рес.: https://doi.org/10.1016/j.comcom.2020.03.026. (дата обращения: 08.07.2021).
23. Managing CRAMM Reviews Using PRINCE. Central Computer & Telecommunications Agency (UK) // Publisher: Stationery Office Books, November 1993, 140 pages. [Электронный ресурс] – Режим доступа к рес.: - http://www.cramm.com/files/techpapers/Managing%20CRAMM%20Reviews%20Using%20Prince.pdf. (дата обращения: 12.05.2023).

24. CAPEC (Common Attack Pattern Enumeration and Classification) [Электронный ресурс] – Режим доступа к рес.: https://capec.mitre.org (дата обращения: 08.07.2021).
25. CWE (Common Weakness Enumeration) [Электронный ресурс] – Режим доступа к рес.: https://cve.mitre.org (дата обращения: 08.07.2021
26. Open Web Application Security Project, OWASP Top Ten 2010 [Электронный ресурс] – Режим доступа к рес.: http://www.owasp.org (дата обращения: 22.04.2014
27. CVE (Common Vulnerabilities and Exposures). [Электронный ресурс] – Режим доступа к рес.: https://cve.mitre.org (дата обращения: 22.04.2014)
28. Носаль, И.А. Потенциал нападения и типовая модель нарушителя / И.А. Носаль // Информационная безопасность и защита персональных данных: Проблемы и пути их решения: VI Межрегиональная научно-практическая конф.: материалы (г. Брянск, 28 апреля 2014г.). – Брянск: Изд-во БГТУ, 2014. – С. 96-101.
29. Осипов, В.Ю. Информационный вандализм, криминал и терроризм как современные угрозы обществу / В.Ю. Осипов, Юсупов Р.М. // Труды СПИИРАН. – 2009, выпуск 8. – С. 34–45.
30. Мартин Дж. Вычислительные сети и распределённая обработка данных. Программное обеспечение, методы и архитектура / Дж. Мартин. – Пер. с англ. Москва: Финансы и статистика, 1986. – 269 с.
31. Peltier T.R. Information Security Risk Analysis / T.R. Peltier // Boca Raton, FL: Auerbach publications; 1 edition, 31.01.2001, 296 pages.
32. Alberts C. Managing Information Security Risks: The OCTAVE (SM) Approach / C. Alberts, A. Dorofee // Publisher: Addison-Wesley (E); 1 edition, Juli 2002, 470 pages.
33. Storms A. Using vulnerability assessment tools to develop an OCTAVE Risk Profile / A. Storms // GIAC GSEC Practical (v1.4b) December 03, 2003.
34. RiskWatch users manual // [Электронный ресурс] – Режим доступа к рес.: - http://www.riskwatch.com. (дата обращения: 12.05.2023).

35. Ажмухамедов, И.М. Анализ и управление комплексной безопасностью на основе когнитивного моделирования / И.М. Ажмухамедов // Управление большими системами: сборник трудов. – 2010. – № 29. – С. 5-15.
36. Рытов, М.Ю. Управление безопасностью информационных технологий на основе методов когнитивного моделирования / М.Ю. Рытов, М.В. Рудановский // Информационная безопасность. – 2010. – №4. – С. 579-582.
37. Вахний, Т.В. Теоретико-игровой подход к выбору оптимальных стратегий защиты информационных ресурсов / Т.В. Вахний, А.К. Гуц // Математические структуры и моделирование. – 2009. – № 1 (19). С. – 104-107.
38. Joseph A. D. Adversarial Machine Learning /A. D. Joseph, N. Blaine, B. I. Rubinstein, J. D. Tygar. Cambridge: Cambridge University Press, 2019. – P. 338.
39. Макгроу, Г. Обеспечение безопасности систем машинного обучения / Г. Макгроу, Р. Бонетт, Х.Фигероа, В.Шепардсон // Открытые системы. СУБД. –2019. – № 4. – С. 22.
40. Matthew, Stewart. Security Vulnerabilities of Neural Networks [Towards data science] / Stewart Matthew. [Электронный ресурс] – Режим доступа к рес.: https://towardsdatascience.com/hacking-neural-networks-2b9f461ffe0b, свободный. – дата обращения: 20.11.2021).
41. How to attack Machine Learning (Evasion, Poisoning, Inference, Trojans, Backdoors) [Towards data science]. [Электронный ресурс] – Режим доступа к рес.: https://towardsdatascience.com/how-to-attack-machine-learning-evasion-poisoning-inference-trojans-backdoors-a7cb5832595c?gif=true, свободный. (дата обращения: 30.12.2021).
42. Bolum, Wang.Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks/ Bolum Wang, Yuanshum Yao, Shawn Shan, Huiying Li // Conference: 2019 IEEE Symposium on Security and Privacy.–2019.
43. IEEE (Institute of Electrical and Electronics Engineers) [Электронный ресурс]. – Режим доступа к рес.: https://www.ieee.org (дата обращения:11.12.2021).

44. Ilyushin, E . /Attacks on machine learning systems-common problems and methods / Eugene Ilyushin, Dmitry Namiot, Ivan Chizhov. // International Journal of Open Information Technologies 10.3 (2022). – С. 17-22.
45. Nilaksh, Das.Keeping the Bad Guys Out: Protecting and Vaccinating Deep Learning with JPEG Compression / Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E. Kounavis, Duen Horng Chau // arXiv.org. – 2017. [Электронный ресурс]. – Режим доступа к рес.: https://arxiv.org/abs/1705.02900 (дата обращения:11.12.2021).
46. Kostyumov, V. A survey and systematization of evasion attacks in computer vision / Vasily Kostyumov // International Journal of Open Information Technologies 10.10 (2022) - 11-20.
47. Artificial Intelligence in Cybersecurity. [Электронный ресурс] – Режим доступа к рес.: https://cs.msu.ru/node/3732 (дата обращения:11.12.2023).
48. Hu, Z., Beuran, R., & Tan, Y.Automated Penetration Testing Using Reinforcement Learning / Z.Hu, R. Beuran, Y. Tan // 2020. [Электронный ресурс] – Режим доступа к рес.: https://www.researchgate.net/publication/353941853_Using_Cyber_Terrain_in_Reinforcement_Learning_for_Penetration_Testing (дата обращения:11.12.2023)
49. Bagdasaryan, E. Blind backdoors in deep learning models/ Eugene Bagdasaryan, Shmatikov Vitaly //Usenix Security. 2021. [Электронный ресурс] – Режим доступа к рес.: https://arxiv.org/abs/2005.03823 (дата обращения:11.02.2024)
50. Fickling. [Электронный ресурс] – Режим доступа к рес.: https://github.com/trailofbits/fickling (дата обращения:11.02.2024)
51. TensorFlow Hub. [Электронный ресурс] – Режим доступа к рес.: https://www.tensorflow.org/hub/overview (дата обращения:11.02.2024)
52. Parker S. Cybersecurity in process control, operations, and supply chain. / Sandra Parker, Wu Zhe, Panagiotis D. Christofides // Computers & Chemical Engineering (2023): 108169. [Электронный ресурс] – Режим доступа к рес.: https://skoge.folk.ntnu.no/prost/proceedings/focapo-cpc-2023/Invited%20Keynotes/5_Keynote_Invited.pdf (дата обращения:11.02.2024).

53. Costales, R. Live trojan attacks on deep neural networks. / Robby Costales, et al. //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020. [Электронный ресурс] – Режим доступа к рес.: https://arxiv.org/abs/2004.11370 (дата обращения:11.02.2024).
54. Li Q. A Label Flipping Attack on Machine Learning Model and Its Defense Mechanism / Qingru Li, et al. //Algorithms and Architectures for Parallel Processing: 22nd International Conference, ICA3PP 2022, Copenhagen, Denmark, October 10–12, 2022, Proceedings. Cham: Springer Nature Switzerland, 2023. [Электронный ресурс] – Режим доступа к рес.: https://fruct.org/publications/volume-32/fruct32/files/Abr.pdf (дата обращения:11.02.2024).
55. Steinhardt, Jacob, Pang Wei W. Koh, and Percy S. Liang. Certified defenses for data poisoning attacks / Jacob Steinhardt, Wei W. Koh Pang, S. Liang Percy. //Advances in neural information processing systems 30 (2017). [Электронный ресурс]. – Режим доступа к рес.: https://arxiv.org/abs/1706.03691 (дата обращения: 11.02.2024).
56. Xue M. Intellectual property protection for deep learning models: Taxonomy, methods, attacks, and evaluations. / Mingfu Xue, et al. // IEEE Transactions on Artificial Intelligence 3.6 (2021): 908-923. [Электронный ресурс]. – Режим доступа к рес.: https://arxiv.org/abs/2011.13564 (дата обращения:11.02.2024).
57. Szegedy Ch. Intriguing properties of neural networks / Christian Szegedy, et al. // arXiv preprint arXiv:1312.6199 (2013). [Электронный ресурс]. – Режим доступа к рес.: https://arxiv.org/abs/1312.6199 (дата обращения:11.02.2024).
58. Yang Y. A closer look at accuracy vs. robustness. / Yao-Yuan Yang, et al // Advances in neural information processing systems 33 (2020): 85888601. [Электронный ресурс]. – Режим доступа к рес.: https://arxiv.org/abs/2003.02460 (дата обращения:11.02.2024).
59. Namiot D. The rationale for working on robust machine learning / Dmitry Namiot, Eugene Ilyushin, Ivan Chizhov // International Journal of Open Information Technologies 9.11 (2021): 68-74.

60. Namiot D. Data shift monitoring in machine learning models. / Dmitry Namiot, Eugene Ilyushin, Ivan Chizhov // International Journal of Open Information Technologies 10.12 (2022): 84-93. [Электронный ресурс] – Режим доступа к рес.: http://injoit.org/index.php/j1/article/view/1462 (дата обращения:11.02.2024).
61. Namiot D. On the robustness and security of Artificial Intelligence systems / Dmitry Namiot, Eugene Ilyushin // International Journal of Open Information Technologies 10.9 (2022): 126-134. [Электронный ресурс] – Режим доступа к рес.: http://injoit.org/index.php/j1/article/view/1398 (дата обращения:11.02.2024).
62. Namiot D. Introduction to Data Poison Attacks on Machine Learning Models / Namiot, Dmitry // International Journal of Open Information Technologies 11.3 (2023): 58-68. [Электронный ресурс] – Режим доступа к рес.: https://arxiv.org/pdf/2112.02797 (дата обращения:11.02.2024).
63. Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Artificial intelligence and cybersecurity." International Journal of Open Information Technologies 10.9, 2022. – С.135-147. [Электронный ресурс]. – Режим доступа к рес.: (дата обращения:11.02.2024).
64. Namiot D. On Trusted AI Platforms. / Dmitry Namiot, Eugene Ilyushin, and Oleg Pilipenko. //International Journal of Open Information Technologies 10.7, 2022 – С.119-127. [Электронный ресурс] – Режим доступа к рес.: (дата обращения:11.02.2024).
65. Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Ongoing academic and industrial projects dedicated to robust machine learning." International Journal of Open Information Technologies 9.10, 2021. – С. 35-46. [Электронный ресурс] – Режим доступа к рес.: (дата обращения:11.02.2024).
66. Корт, С.С. Теоретические основы защиты информации: учеб. пособие / С.С. Корт. – Москва: Гелиос АРВ, 2004. – 240 с.
67. Digital Security. Алгоритм: модель анализа угроз и уязвимостей // [Электронный ресурс] – Режим доступа к рес.: http:\\www.dsec.ru (дата обращения: 22.04.2014).

68. Бордак, И. В. Разработка метода количественной оценки и прогнозирования безопасности информации ограниченного доступа на основе Марковских случайных процессов / И. В. Бордак, А. П. Росенко // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2017. – Т. 20, № 4. – С. 67–70.
69. P. Saitta Larcom. Trike v.1 Methodology Document. // Saitta P., Larcom B., Eddington M.; (13.07.2005) [Электронный ресурс] – Режим доступа к рес.: http://www.net-security.org/dl/articles/Trike_v1_Methodology_Document-draft.pdf. (дата обращения: 22.04.2014).
70. Oladimeji E.A., Supakkul S., Chung L. Security threat modeling and analysis: a goal-oriented approach / E.A. Oladimeji, S. Supakkul, L. Chung // [Электронный ресурс] – Режим доступа к рес.: http://citeseerx.ist.psu.edu/ viewdoc/download?doi=10.1.1.103.2997&rep=rep1&type=pdf. (дата обращения: 12.04.2023).
71. Williams L. GARNET: A graphical attack graph and reachabiity network evaluation tool in Visualization for Computer Security (VizSEC) / L. Williams, R. Lippmann, K. Ingols // ser. Lecture Notes in Computer Science, J.R. Goodall, G.J. Conti, and K.-L. Ma, Eds. - Springer, 2008. - Vol. 5210. - P. 44-59.
72. Wang L. Implementing interactive analysis of attack graphs using relational databases / L. Wang, C. Yao, A.Singhal, S. Jajodia // Journal of Computer Security. 2008. - N 16. - С. 419–437.
73. Moore A. Attack Modeling for Information Security and Survivability / A. Moore, R. Ellison, R. Linger // Software Engineering Institute, Technical Note CMU/SEI-2001-TN-01, March 2001. [Электронный ресурс] – Режим доступа к рес.: https://insights.sei.cmu.edu/documents/1966/2001_004_001_13793.pdf (дата обращения: 12.04.2023).
74. Бешелев, С.Д. Математико-статистические методы экспертных оценок / С.Д. Бешелев. – Москва: Статистика, 1974. – 159 с.
75. Трофимов, И.Б. Методы оценки информационной безопасности предприятия на основе процессного подхода: дис. канд. техн. наук. 05.13.19 / И. Б. Трофимов. – Санкт-Петербург, 2010. – 134 с.

76. Ажмухамедов, И.М. Оценка экономической эффективности мер по обеспечению информационной безопасности / И.М. Ажмухамедов, Т.Б. Ханжина // Вестник АГТУ. – 2011. – №1. – С. – 185-190.
77. Миронов, В.В. Моделирование и оценка системы обеспечения информационной безопасности на примере ГОУ ВПО «СыктГУ» / В.В. Миронов, И.А. Носаль // Информация и безопасность. – 2011. – № 2. – С. 209–216.
78. Молдованин, Т.В. Решение задачи выбора оптимального варианта комплексной защиты информации с помощью метода экспертного оценивания / Т.В. Молдованин // Информационно-управляющие системы. – 2007. – №3. – С. 39–44.
79. Домарев В.В. Безопасность информационных технологий. Системный подход / В.В. Домарев // Киев: ООО ТИД «Диасофт», 2004. – 992 с. Режим доступа: http://www.security.ukrnet.net/d-book-2/ch_06.pdf. (дата обращения: 12.05.2013).
80. Заболотский, В.П. Применение метода индексов для оценивания эффективности защиты информации / В.П. Заболотский, Р.М. Юсупов // Труды СПИИРАН. Вып.3, т. 2. – Санкт-Петербург: Наука, 2006.
81. Карпеев, Д.О. Исследование и развитие методического обеспечения оценки и управления рисками информационных систем на основе интересоориентированного подхода: дис. канд. техн. наук / Д.О. Карпеев. – Воронеж, 2009. – 171 с.
82. Корнилова, А.Ю. Проблемы применения методов экспертных оценок в процессе экономического прогнозирования развития предприятия / А.Ю. Корнилова, Т.Ф. Палей // Проблемы современной экономики. – 2010. – № 3 (35). – С. 124-128.
83. Ефимов, Е.И. Возможность применения существующих средств анализа рисков в системах принятия решений с привлечением экспертов / Е.И. Ефимов // Омский научный вестник. – 2011. – № 3-103 - С.281-284.
84. Куравский, Л.С. Марковские модели в задачах диагностики и прогнозирования: Учеб. пособие. / С. Л. Артеменков, В. И. Алхимов, С. Н. Баранов, Беляева О. Б., П. Н. Думин, П. А. Корниенко, Л. С. Куравский, С. Б. Малых, А. А. Марголис, П. А. Мармалюк, А. С. Панфилова, С. И. Попков, Г. А. Юрьев, Н. Е. Юрьева. – 2-е изд., доп. – М.: Изд-во МГППУ, 2017. – 203 с.

85. Кубарев, А.В. Параметрическое моделирование состояния объектов критической инфраструктуры в условиях деструктивного воздействия /А.В.Кубарев, А.П. Лапсарь, Я.В. Федорова // Вопросы кибербезопасности. 2021. № 3. С. 58-67.
86. Dewri R. Optimal security hardening on attack tree models of networks: a cost-benefit analysis / R. Dewri, I.Ray, N.Poolsappasit, D. Whitley // International Journal of Information Security. 2012. N 11. P. 167-188.
87. Мальцев, Г.Н. Оптимизация состава средств защиты информации в информационно-управляющей системе с каналами беспроводного доступа на основе графа реализации угроз / Г.Н.Мальцев, В.В. Теличко // Информационно-управляющие системы. – 2008. – №4. – С. 29–33.
88. Аграновский, А.В. Теоретико-графовый подход к анализу рисков в вычислительных сетях / А.В. Аграновский, Р.А. Хади, В.Н. Фомченко, А.П. Мартынов, В.А. Снапков // Конфидент. Защита информации. – 2002. – № 2. – С. 50-53.
89. Абрамов, Е.С. Использование графа атак для автоматизированного расчета мер противодействия угрозам информационной безопасности сети / Е.С. Абрамов, М.А. Кобилев, Л.С. Крамаров, Д.В. Мордвин // Известия ЮФУ. Технические науки. – 2014. – №2(151). – С. 92-100.
90. Аверченков, В. И. Формализация выбора решения при проектировании комплексных систем защиты информации от несанкционированного доступа / В. И. Аверченков, М. Ю. Рытов, Т.Р. Гайнулин, О.М. Голембиовская // Известия Волгоградского государственного технического университета. – Волгоград: ВолГТУ. – 2011. – №11(84). – С. 131-136.
91. Козленко, А.В. Метод оценки уровня защиты информации от НСД в компьютерных сетях на основе графа защищённости / А.В. Козленко, В.С. Авраменко, И.Б. Саенко, А.В. Кий // Труды СПИИРАН. – 2012. – №21. – С.41–55.
92. Кудрявцева, Р.Т. Управление информационными рисками с использованием технологий когнитивного моделирования: автореф. дис. канд. техн. наук. – Уфа, 2008. – 17 с.

93. Ажмухамедов, И.М. Анализ и управление комплексной безопасностью на основе когнитивного моделирования / И.М. Ажмухамедов // Управление большими системами: сборник трудов. – 2010. – № 29. – С. 5-15.
94. Рытов, М.Ю. Управление безопасностью информационных технологий на основе методов когнитивного моделирования / М.Ю. Рытов., М.В. Рудановский // Информационная безопасность. – 2010. – №4. – С. 579-582.
95. Чусавитин, М.О. Использование метода анализа иерархий при оценке рисков информационной безопасности образовательного учреждения / М.О. Чусавитин // Фундаментальные исследования. – 2013. – № 10. – С. 2080-2084.
96. Бикмаева, Е.В. Об оптимальном выборе системы защиты информации от несанкционированного доступа / Е.В. Бикмаева, Р.И. Баженов // Электронный научный журнал «APRIORI. Серия: Естественные и технические науки». – 2014. – №6. – С. 5-16. [Электронный ресурс]. – Режим доступа: https://cyberleninka.ru/article/n/ob-optimalnom-vybore-sistemy-zaschity-informatsii-ot-nesanktsionirovannogo-dostupa/viewer (дата обращения: 04.12.23).
97. Куземко, С.М. Усовершенствованный метод анализа иерархий для выбора оптимальной системы защиты информации в компьютерных сетях / С.М. Куземко, В.М. Мельничук // Наукові праці Вінницького національного технічного університету. – 2010. – №2 [Электронный ресурс]. – Режим доступа: http://praci.vntu.edu.ua/article/view/1253/592 (дата обращения: 10.01.2015).
98. Данилюк, С.Г. Обоснование нечёткого ситуационного подхода к созданию модели системы защиты информации с использованием ложных информационных объектов / Д С. Ганилюк, В.Г. Маслов // Известия Южного федерального университета. Технические науки. – 2008. – № 8, т.85. – С. 36-41.
99. Борзенкова, С.Ю. Модель принятия решения при управлении системой защиты информации / С.Ю. Борзенкова, О.В. Чечуга // Известия Тульского государственного университета. Технические науки. – 2013. – № 3. – С. 471-478.
100. Круглов, В.В. Нечёткая логика и искусственные нейронные сети / В.В. Круглов, М.И. Дли, Р.Ю. Голунов. – Москва: Физматлит, 2001. – 221 с.

101. Васильев, В.И. Построение нечётких когнитивных карт для анализа и управления информационными рисками вуза / В.И. Васильев, И.А. Савина, И.И. Шарипова // Вестник УГАТУ. – 2008. – №2, т.10. – С. 199-209.

102. Sodiya A.S. Threat Modeling Using Fuzzy Logic Paradigm / A.S. Sodiya, S.A. Onashoga, B.A. Oladunjoye // Issues in Informing Science and Information Technology. 2007. Volume 4. [Электронный ресурс] – Режим доступа: https://proceedings.informingscience.org/InSITE2007/IISITv4p053-061Sodi261.pdf (дата обращения: 04.12.24).

103. Чечулин, А.А. Методика оперативного построения, модификации и анализа деревьев атак / А.А. Чечулин // Труды СПИИРАН. №3(26). – Санкт-Петербург: Наука, 2013. – С. 40-53.

104. Sood, A. K., & Enbody, R. J. A Framework for Modeling Cyber Attacks Using Markov Decision Processes/ A. K. Sood, R. J. Enbody, (2013) [Электронный ресурс]. – Режим доступа: https://arxiv.org/pdf/2207.05436 (дата обращения: 21.03.2021).

105. Ingle M. Risk analysis using fuzzy logic / M. Ingle, M.Atique, S.O. Dahad // International Journal of Advanced Engineering Technology. 2011. Vol.II, Issue III. P. 96-99

106. Shang K. Applying Fuzzy Logic to Risk Assessment and DecisionMaking / K. Shang, Z. Hossen // Report of Casualty Actuarial Society, Canadian Institute of Actuaries, 2013. pp.59. [Электронный ресурс]. – Режим доступа: - https://www.soa.org/globalassets/assets /files/research/projects/ research-2013-fuzzy-logic.pdf (дата обращения: 20.03.2024).

107. Маслобоев, А.В. Разработка и реализация механизмов управления информационной безопасностью мобильных агентов в распределённых мультиагентных информационных системах / А.В. Маслобоев, В.А. Путилов // Вестник Мурманского государственного технического университета. – 2010. – № 4-2, т.13. – С. 1015-1032.

108. Файзуллин, Р.Р. Метод оценки защищённости сети передачи данных в системе мониторинга и управления событиями информационной безопасности на

основе нечёткой логики / Р.Р. Файзуллин, В.И. Васильев // Вестник УГАТУ. – 2013. – №2 (55). – С.150-156.

109. Котенко, И.В. Гибридная адаптивная система защиты информации на основе биометафор «нервных» и нейронных сетей / И.В. Котенко, Ф.Г. Нестерук, А.В. Шоров // Инновации в науке. – 2013. – №16-1. – С.79-83.

110. Котенко, И.В. Анализ биоинспирированных подходов для защиты компьютерных систем и сетей / В. Котенко, Ф.Г. Нестерук, А.В. Шоров // Труды СПИИРАН. – Вып.3 (18). – Санкт-Петербург: Наука, 2011. С. 19-73.

111. Котенко, И.В. Команды агентов в киберпространстве, моделирование процессов защиты информации в глобальном Интернете / И.В. Котенко, А.В. Уланов // Труды института системного анализа РАН. Проблемы управления кибербезопасностью информационного общества. – Москва: КомКнига, 2006. – Т. 27. – С. 108–129.

112. Sun L., Srivastava R.P., Mock T.J. An Information Systems Security Risk Assessment Model under Dempster-Shafer Theory of Belief Functions // Journal of Management Information Systems. Vol. 22. N 4. Spring 2006. pp.109-142.

113. Арьков, П.А. Комплекс моделей для поиска оптимального проекта системы защиты информации / П.А. Арьков // Известия Южного федерального университета. Технические науки. – 2008. – № 8, т. 85. – С. 30-36.

114. Вахний, Т.В. Теоретико-игровой подход к выбору оптимальных стратегий защиты информационных ресурсов / Т.В. Вахний, А.К. Гуц // Математические структуры и моделирование. – 2009. – № 1 (19). – С. 104-107.

115. Белый, А.Ф. Компьютерные игры для выбора методов и средств защиты информации в автоматизированных системах / А.Ф. Белый // Известия Южного федерального университета. Технические науки. – 2008. – № 8, т.85. – С. 172-176.

116. Шлыков, Г.Н. Применение гомоморфизма в моделях защиты информации / Г.Н. Шлыков // Вестник удмуртского университета. – 2011. – № 4. – С. 175-179.

117. Меньших, В.В. Теоретическое обоснование и синтез математической модели защищённой информационной системы ОВД как сети автоматов /

В.В. Меньших, Е.В. Петрова // Вестник Воронежского института МВД России. – 2010. – № 3. – С.134–142.

118. Ерохин, С.С. Оценка защищённости информационных систем с использованием скрытых Марковских процессов / С.С. Ерохин, С.В. Голубев // Научная сессия ТУСУР 2007: Всерос. науч.-техн. конф. студентов, аспирантов и молодых ученых: материалы докладов (Томск, 4-7 мая 2007 г.). – № 42. – С.133-137.

119. Росенко, А.П. Применение марковских случайных процессов с дискретным параметром для оценки уровня информационной безопасности / А.П. Росенко // Известия Южного федерального университета. Технические науки. – 2009. – № 11. – С.169-172.

120. Каштанов, В.А. О минимаксных подходах в задачах безопасности / В.А. Каштанов, О.Б. Зайцева // Труды Карельского научного центра Российской академии наук. – 2013. – № 1. – С. 55-67.

121. Иванов, К.В. Марковские модели средств защиты автоматизированных систем специального назначения: монография / К.В. Иванов, П.И. Тутубалин. – Казань: ГБУ «Республиканский центр мониторинга качества образования», 2012. – 2016 с.

122. Попов, С.В. Определение вероятностей состояний функционирования средства контентного анализа как элемента системы мониторинга инцидентов информационной безопасности / С.В. Попов, В.Н. Шамкин // Вестник ТГТУ. – 2012. – №1. – С. 27-37.

123. Росенко, А.П. Внутренние угрозы безопасности конфиденциальной информации: Методология и теоретическое исследование: моногр. / А.П. Росенко. – Москва: Красанд, 2010. – 160 с.

124. Осипов, В.Ю. Радиоэлектронная борьба. Теоретические основы. Учеб. пособие для вузов / В.Ю. Осипов, А.П. Ильин, В.П. Фролов, А.П. Кондратюк. – Петродворец: ВМИРЭ, 2006. – 302 с.

125. Окрачков, А.А. Метод аппроксимации распределения времени реализации защитных функций системой защиты информации от несанкционированного

доступа / А.А. Окрачков, М.Е. Фирюлин // Вестник ВИ МВД РосПИИ. – 2012. – №4. – С. 123-131.

126. Xiaofan Zhou. Markov Decision Process For Automatic Cyber Defense (WISA 2023) / Zhou Xiaofan, Yusuf Enoch Simon, Dan Dong Seong Kim. [Электронный ресурс] – Режим доступа: https://arxiv.org/abs/2207.05436 (дата обращения: 03.05.2024).

127. Booker, L.B. A model-based, decision-theoretic perspective on automated cyber response / L.B.Booker, S.A. Musman // arXiv. [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2002.08957 (дата обращения: 13.05.2024).

128. Zheng, J., Namin, A.S. Defending SDN-based IoT Networks Against DDoS Attacks Using Markov Decision Pro-cess / J. Zheng // Proceedings - 2018 IEEE International Confer-ence on Big Data, Big Data 2018 – 2018 – С. 4589-4592 [Электронный ресурс] – Режим доступа: https://ieeexplore.ieee.org/document/8622064 (дата обращения: 03.05.2024).

129. Кохендерфер, М. Алгоритмы принятия решений / М. Кохендерфер, Т. Уилер, К. Рэй. – Москва: ДМК Пресс, 2023. – 684 с.

130. Саттон, Р. С. Обучение с подкреплением: введение / Р. С. Саттон, Э. Дж. Барто; 2-е изд. – Москва: ДМК Пресс, 2020. – 552 с.

131. Mazengia D.H. Forecasting Spot Electricity Market Prices Using Time Series Models / D.H. Mazengia // Thesis for the degree of Master of Science in Electric Power Engineering. – Gothenburg: Chalmers University of Technology. - 2008 – С.89 p. [Электронный ресурс] – Режим доступа: https://arxiv.org/abs/2002.08957 (дата обращения: 13.05.2024).

132. Ветров, И. А. Формирование вектора сетевых атак с учетом специфики связей техник и тактик / И. А. Ветров, В.В. Подтопельный // Вестник СибГУТИ. 2023. – Т. 17, № 4. – С. 49–61.

133. Ветров, И.А. Особенности формирования вектора современных сетевых атак / И. А. Ветров, В.В. Подтопельный // Вестник СибГУТИ. – 2022. – № 3 (59). – С. 3-13.

134. Подтопельный, В. В. Исследование специфики моделирования компьютерных атак с использованием марковских процессов принятий решений и q-обучения / В. В. Подтопельный // Информация и безопасность. – 2024. – Т. 27, вып.3 – С. 421–441.

135. Подтопельный, В. В. Особенности моделирования атак на модели машинного обучения с использованием марковских процессов принятия решений / В. В. Подтопельный // Доклады ТУСУР. – 2024. – Т. 27, № 2. – С. 21-30.

136. Ветров, И. А. Особенности построения нейросетей с учетом специфики их обучения для решения задач поиска сетевых атак / И. А. Ветров, В. В. Подтопельный // Доклады ТУСУР. – 2023. – Т. 26, № 2. – С. 42–50.

ПРИЛОЖЕНИЕ А
(справочное)
СВИДЕТЕЛЬСТВА О ГОСУДАРСТВЕННОЙ РЕГИСТРАЦИИ ПРОГРАММЫ ДЛЯ ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2024662827

«Программа экспертного анализа угроз информационной безопасности»

Правообладатель: ***Федеральное государственное автономное образовательное учреждение высшего образования «Балтийский федеральный университет имени Иммануила Канта» (RU)***

Авторы: ***Ветров Игорь Анатольевич (RU), Подтопельный Владислав Владимирович (RU)***

Заявка № **2024619240**
Дата поступления **25 апреля 2024 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***30 мая 2024 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2024665096

Гибридная система поиска, анализа и прогнозирования событий безопасности в распределенной информационной системе

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования «Калининградский государственный технический университет» (RU)***

Авторы: ***Подтопельный Владислав Владимирович (RU), Семенов Никита Алексеевич (RU), Кожевникова Анна Александровна (RU), Подтереба Артем Алексеевич (RU)***

Заявка № **2024663319**

Дата поступления **13 июня 2024 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***27 июня 2024 г.***

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Владелец **Зубов Юрий Сергеевич**

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2024616878

Программа для ЭВМ «BotnetSniffer»

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования «Калининградский государственный технический университет» (RU)***

Авторы: ***Семенов Никита Алексеевич (RU), Подтопельный Владислав Владимирович (RU)***

Заявка № **2024614885**

Дата поступления **12 марта 2024 г.**

Дата государственной регистрации

в Реестре программ для ЭВМ ***26 марта 2024 г.***

Руководитель Федеральной службы

по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат [illegible]

Владелец Зубов Юрий Сергеевич

Действителен с [illegible] по 02.08.2024

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2021665446

Программа для ЭВМ «NNSCA»

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования "Калининградский государственный технический университет" (RU)***

Авторы: ***Майстренко Анастасия Юрьевна (RU), Подтопельный Владислав Владимирович (RU)***

Заявка № **2021664445**
Дата поступления **17 сентября 2021 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***27 сентября 2021 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

Г.П. Ивлиев

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025619455

«Программа анализа атак отравления на наборы данных»

Правообладатель: ***Федеральное государственное автономное образовательное учреждение высшего образования «Балтийский федеральный университет имени Иммануила Канта» (RU)***

Авторы: ***Ветров Игорь Анатольевич (RU), Подтопельный Владислав Владимирович (RU), Сацута Анатолий Игоревич (RU)***

Заявка № **2025617891**
Дата поступления **04 апреля 2025 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***16 апреля 2025 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2024689273

«Программа анализа сценариев атак на системы корпоративного типа при аудите»

Правообладатель: ***Федеральное государственное автономное образовательное учреждение высшего образования «Балтийский федеральный университет имени Иммануила Канта» (RU)***

Авторы: ***Ветров Игорь Анатольевич (RU), Подтопельный Владислав Владимирович (RU)***

Заявка № **2024688698**
Дата поступления **22 ноября 2024 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***05 декабря 2024 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

RU2020664040

ГОСУДАРСТВЕННАЯ РЕГИСТРАЦИЯ ПРОГРАММЫ ДЛЯ ЭВМ

Номер регистрации (свидетельства):
2020664040
Дата регистрации: **06.11.2020**
Номер и дата поступления заявки:
2020662457 20.10.2020
Дата публикации и номер бюллетеня:
06.11.2020 Бюл. № 11
Контактные реквизиты:
ionpvv@mail.ru, 89003539881

Автор(ы):
Куделка Денис Васильевич (RU),
Подтопельный Владислав Владимирович (RU)
Правообладатель(и):
Федеральное государственное бюджетное образовательное учреждение высшего образования "Калининградский государственный технический университет" (RU)

Название программы для ЭВМ:
Модуль анализа параметров сетевых атак для СОВ

Реферат:
Программа для ЭВМ разработана для анализа параметров сетевых атак. На основании результатов анализа создаются правила для системы обнаружения вторжений. Целью создания программы для ЭВМ является устранение недостатков обнаружения атак на основе сигнатурных методов анализа, путем применения методов машинного обучения. Программа состоит из трех ключевых подсистем: подсистема подготовки данных; подсистема анализа; подсистема графического вывода. Тип ЭВМ: ПЭВМ на базе процессора Intel Core i3/i5/i7 3,3 ГГц и выше. ОС: Windows 7 x64/8/8.1 x64/10, Linux, MacOS.

Язык программирования: Python

Объем программы для ЭВМ: 20 МБ

ПРИЛОЖЕНИЕ Б
(справочное)

АКТ О ВНЕДРЕНИИ РЕЗУЛЬТАТОВ ДИССЕРТАЦИОННОЙ РАБОТЫ

ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ
"ЦЕНТР ЗАЩИТЫ ИНФОРМАЦИИ"
ИНН/КПП 3906952332/390601001 ОГРН 1153926002230
236006, г. Калининград, ул. Кирпичная, д. 7 п. 2 , оф. 301-304
Тел. +7 (4012) 90-50-28, сайт: www.beltzi.ru, e-mail: czi@beltzi.ru

УТВЕРЖДАЮ
Генеральный директор
ООО «Центр защиты информации»
_____ А.И. Ковтун
«02» 05 2024 № 23

АКТ

о внедрении результатов диссертации Подтопельного Владислава Владимировича
в производственные процессы предприятия (инженерно-технический отдел)
ООО «Центр защиты информации»

Комиссия в составе:

- генерального директора ООО «Центр защиты информации» Ковтуна А. И.
- начальника инженерно-технического отдела Кулакова П. А.
- специалиста инженерно-технического отдела Буйневича Д. М.

составила настоящий акт о том, что основные научные результаты кандидатской диссертации Подтопельного Владислава Владимировича на тему «Определение и анализ последовательностей атакующих воздействий на системы искусственного интеллекта при аудите информационной безопасности»
(название диссертации)
внедрены в рабочие процессы инженерно-технического отдела, связанные проведением аудита информационной безопасности систем интеллектуального типа, использующие модели машинного обучения.

В рабочий процесс организации внедрены:

1. Методика поиска оптимальных путей реализации атак на подсистемы искусственного интеллекта информационных систем с последующим формированием сценариев атак и их анализа на основе методических документов, методик и классификации описания уязвимостей и векторов атак (Методика оценки угроз ФСТЭК и MITRE ATLAS) используемых при формировании Моделей угроз предприятий;
2. Алгоритм формирования последовательности атакующих воздействий, их модификации, анализа атак, учитывающие полноту описания атак и специфику подсистемы искусственного интеллекта информационных систем.

Генеральный директор ООО «Центр защиты информации»	_____ *(подпись)*	А. И. Ковтун
Начальник инженерно-технического отдела	_____ *(подпись)*	П. А. Кулаков
Специалист инженерно-технического отдела	_____ *(подпись)*	Д. М. Буйневич

ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ
Радиоэлектронные системы

620137, г. Екатеринбург, ул. Июльская, д. 41,
Тел./факс: 8 (343) 374 24 64,
E-mail:zi@irsural.ru,
ИНН: 6659102580, КПП: 667801001, ОКПО: 72889278
р/с 40702810016540007884 в Уральском банке ПАО Сбербанк,
г.Екатеринбург, БИК 046577674, к/с 30101810500000000674

УТВЕРЖДАЮ
Директор
Гильмияров Роман Владимирович
ООО «Радиоэлектронные системы»
11 апреля 2025

АКТ

о внедрении научных результатов,
полученных Подтопельным Владиславом Владимировичем

Комиссия в составе:
- Гильмиярова Романа Владимировича,
- Крашенинникова Максима Валерьевича
- Галимзяновой Лилия Илдаровна

составила настоящий акт о том, что научные результаты, полученные Подтопельным Владиславом Владимировичем, а именно:

1. Модели формирования последовательности атакующих воздействий, учитывающих специфику систем искусственного интеллекта, а также порядок этапов реализации угроз безопасности в соответствии с требованиями руководящих документов государственных регуляторов в области ИБ.
2. Методика и алгоритм определения последовательностей атакующих воздействий на системы искусственного интеллекта в процессе построения сценариев атак при аудите ИБ, внедрены в технологические процессы ООО «Институт радиоэлектронных систем», связанные с формированием моделей угроз для систем интеллектуального типа, используемых в корпоративной инфраструктуре.

Члены комиссии:

Директор _____ Р.В. Гильмияров
Руководитель отдела ОКР _____ / М.В. Крашенинников
Руководитель отдела испытаний ПАК и ИС _____ / Л.И. Галимзянова