

УДК 004.056 + 004.8

В. А. МАЛАШЁНОК (ЭМ-41)

Научный руководитель – канд. физ.-мат. наук *Н. В. РЯЗАНЦЕВА*

МАТРИЦА MITRE ATLAS

Рассматривается матрица MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) – структурированная база знаний об угрозах, направленных против систем машинного обучения и искусственного интеллекта. Проводится анализ архитектуры фреймворка, описываются 16 тактик и различные техники атак, задокументированных на основе реальных инцидентов. Особое внимание уделяется новым угрозам, связанным с генеративным ИИ и агентными системами, а также практическому применению матрицы в организациях для построения защиты AI-систем.

Внедрение искусственного интеллекта (ИИ) порождает новые киберугрозы, с которыми традиционные средства защиты периметра не справляются. В ответ на это, а также на фоне растущего регуляторного давления, организация MITRE в 2020 году выпустила матрицу MITRE ATLAS. Этот фреймворк систематизирует тактики и техники атак на AI/ML-системы, опираясь исключительно на реальные инциденты.

Основой для ATLAS послужил фреймворк MITRE ATT&CK (2013). Главное их отличие – в объекте защиты: ATT&CK сфокусирован на традиционных IT-сетях и облаках, тогда как ATLAS нацелен на специфические уязвимости машинного обучения. Разработанный совместно с 12 партнерами проект активно развивается. С момента запуска матрица прошла через множество версий: к маю 2026 года (версия 5.6.1) ATLAS документирует 16 тактик, 170 техник в рамках 57 задокументированных реальных кейсов с 35 стратегиями смягчения угроз, часть из которых показана на рисунке 1.

Фреймворк построен на трёх уровнях абстракции. Первый уровень – тактики – описывают фазы и цели атаки: от начальной разведки до финального воздействия на систему. Второй уровень – техники – детализируют конкретные методы реализации каждой тактики. Третий уровень – суб-техники – представляют специализированные варианты применения базовых техник. Такая иерархическая организация позволяет как давать общее описание класса угроз, так и указывать точные векторы атаки.

Reconnaissance ¹	Resource Development ²	Initial Access ³	AI Model Access ⁴	Execution ⁵	Persistence ⁶	Privilege Escalation ⁷	Defense Evasion ⁸
8 techniques	13 techniques	7 techniques	4 techniques	6 techniques	9 techniques	4 techniques	15 techniques
Active Scanning ⁵	Acquire Infrastructure ⁶	AI Supply Chain Compromise ⁷	AI Model Inference API Access ⁸	AI Agent Cloudset ⁹	AI Agent Context Poisoning ¹⁰	AI Agent Tool Injection ¹¹	AI Supply Chain Reputation Pull ¹²
Gather B2G Indexed Targets ⁵	Acquire Public AI Artifacts ⁶	Drive-by Compromise ⁷	AI-Enabled Product or Service ⁸	AI Agent Tool Invocation ⁹	AI Agent Tool Data Poisoning ¹⁰	Escape to Host ¹¹	AI Supply Chain Rug Pull ¹²
Gather Victim Identity Information ⁵	Develop Capabilities ⁶	Coerce AI Model ⁷	Full AI Model Access ⁸	Command and Scripting Interpreter ⁹	AI Agent Tool Poisoning ¹⁰	LLM Jailbreak ¹¹	Corrupt AI Model ¹²
Search Application Repositories ⁵	Enrichment Accounts ⁶	Exploit Public-Facing Applications ⁷	Physical Environment Access ⁸	Deploy AI Agent ⁹	LLM Prompt Self-Replication ¹⁰	Valid Accounts ¹¹	Delay Execution of LLM Instructions ¹²
Search Open AI Vulnerability Analyses ⁵	LLM Prompt Crafting ⁶	Poisoning ⁷		LLM Prompt Injection ⁹	Manipulate AI Model ¹⁰		Erase AI Model ¹²
Search Open Technical Databases ⁵	Obtain Capabilities ⁶	Prompt Infiltration via Public-Facing Applications ⁷		Modify AI Agent Configuration ⁹			Exploitation for Defense Evasion ¹²
Search Open Websites/Forums ⁵	Person Training Data ⁶	Valid Accounts ⁷		Poison Training Data ⁹			False RAG Entry Injection ¹²
Search Victim-Owned Websites ⁵	Publish Malicious Entities ⁶			Prompt infiltration via Public-Facing Application ⁹			Impersonation ¹²
	Publish Poisoned AI Agent Tool ⁶			RAG Poisoning ⁹			LLM Jailbreak ¹²
	Publish Poisoned Gadgets ⁶						LLM Prompt Obfuscation ¹²
	Publish Poisoned Models ⁶						LLM Trusted Output Components Manipulation ¹²
	Publish Content Crafting ⁶						Manipulate User LLM Chat History ¹²
	Obtain Capabilities ⁶						Maximizing ¹²
							Modify AI Agent Configuration ¹²
							Virtualization/Sandbox Evasion ¹²

Рисунок 1 – Фрагмент матрицы ATLAS для систем ИИ

Важнейшая часть ATLAS – опора на реальные цепочки атак. В базе собрано 57 кейсов, среди которых отравление чат-бота Tay, обман автопилота Tesla, обход антивируса Cylance, дистилляция модели Anthropic Claude и инъекции промптов через MCP-серверы.

Инструментальная экосистема ATLAS включает ATLAS Navigator – интерактивный веб-инструмент для визуализации и работы с матрицей. Он позволяет отмечать релевантные для конкретной организации техники, планировать покрытие защитных мер, сравнивать свой профиль угроз с эталонными сценариями. База данных ATLAS распространяется в открытом доступе через репозиторий GitHub в формате YAML, что обеспечивает интеграцию с существующими SIEM-системами и платформами управления угрозами.

MITRE ATLAS охватывает 16 тактик, описывающих полный жизненный цикл атаки на AI-систему. Тактики не являются строго последовательными – атакующий может задействовать несколько из них одновременно или возвращаться к ранее использованным фазам по мере развития операции.

1 Разведка. Сбор данных о целевой AI-системе (архитектура, обучающая выборка, уязвимости) и зондирование API для понимания логики модели.

2 Разработка ресурсов. Подготовка базы для атаки: создание отравленных датасетов, закупка инфраструктуры, регистрация фейковых аккаунтов.

3 Первоначальный доступ. Проникновение в систему через уязвимые интерфейсы, фишинг (включая дипфейки) или инъекции промптов.

4 Доступ к ML-модели. Получение доступа к модели и её параметрам через открытый API или скомпрометированные учётные данные. Данная тактика специфична для ATLAS и не имеет аналога в АТТ&СК.

5 Выполнение. Осуществление атаки: запуск адверсариальных примеров, активация бэкдоров, принуждение AI-агентов использовать инструменты во вред.

6 Закрепление. Долгосрочное присутствие через внедрение бэкдоров при обучении, изменение конфигураций AI-агентов или манипуляцию памятью LLM.

7 Повышение привилегий. Расширение прав доступа за счет кражи API-ключей из конфигов агентов или эксплуатации уязвимостей в цепочках вызовов.

8 Уклонение от защиты. Обход систем безопасности: обман классификаторов адверсариальными примерами, удаление логов, принуждение агентов игнорировать правила через контекст.

9 Доступ к учётным данным. Хищение паролей и токенов (например, извлечение случайно попавших секретов из базы знаний RAG с помощью LLM-запросов).

10 Обнаружение. Изучение внутренней архитектуры: конфигураций агентов, доступных API, источников данных RAG и триггеров рабочих процессов.

11 Боковое перемещение. Расширение присутствия и переход к другим компонентам (базам данных, смежным моделям), часто с использованием цепочек вызовов агентов.

12 Сбор данных. Извлечение ценной информации (параметров модели, логов). Включает кражу конфиденциальных документов через манипуляции с RAG.

13 Подготовка ML-атаки. Автономная подготовка: генерация адверсариальных примеров, отравленных данных или обучение суррогатной (тестовой) модели. Данная тактика также специфична для ATLAS и не имеет аналога в АТТ&СК.

14 Командование и управление. Создание скрытых каналов связи (например, для удаленного управления AI-агентом через инъекции промптов).

15 Эксфильтрация. Несанкционированный вывод данных из системы (в том числе принуждение AI-агента отправить конфиденциальную информацию злоумышленнику).

16 Воздействие. Достижение финальной цели: деградация качества модели, генерация ошибочных решений, повреждение репутации или отказ в обслуживании (финансовое истощение системы через перегрузку запросами).

Из множества техник ATLAS описанные ниже представляют наибольшую практическую опасность.

Инъекция промптов. Главная угроза для языковых моделей (LLM). Специально сконструированный запрос заставляет ИИ игнорировать системные правила. Инъекция бывает прямой (ввод пользователя) и косвенной (через обрабатываемые моделью документы или веб-страницы). Защиту усложняет то, что LLM не имеют встроенного механизма проверки источника команд.

Отравление данных. Внедрение вредоносных данных в обучающую выборку (например, через публикацию зараженных датасетов в открытом доступе) для искажения логики работы итоговой модели. Для автономных систем актуален AI Agent Context Poisoning – манипуляция контекстом, меняющая поведение агента без вмешательства в веса самой модели. Такие атаки очень сложно обнаружить.

Извлечение модели. Кража интеллектуальной собственности: атакующий систематически опрашивает целевую модель и на основе её ответов обучает функциональную копию (суррогатную модель). Схожая угроза – Model Inversion (инверсия модели), позволяющая по ответам ИИ восстановить исходные обучающие данные, что критично для систем с медицинскими или биометрическими базами.

Автономные ИИ-агенты кардинально расширили ландшафт угроз. В отличие от пассивных моделей, они умеют вызывать API, работать с базами данных и управлять внешними системами. Получив контроль над агентом, злоумышленник использует его легитимные привилегии для скрытого перемещения по сети и кражи данных в обход традиционной защиты.

Для манипуляции агентами злоумышленники отравляют контекст, изменяют долгосрочную память ИИ, внедряют инструкции в диалоги или модифицируют конфигурационные файлы. Также они ищут триггеры для автоматического запуска вредоносных рабочих процессов.

В рамках интеграции ИИ критической угрозой стали атаки на протокол Model Context Protocol (MCP), где скомпрометированные серверы перехватывают данные, и эксплуатация уязвимостей удаленного выполнения кода (RCE). Кроме того, развиваются техники обхода ограничений (Jailbreak), позволяющие с помощью специфических формулировок запросов заставить модель полностью игнорировать правила безопасности.

Фреймворк ATLAS предлагает 35 стратегий защиты, позволяющих выстраивать точечные меры безопасности вместо универсальных шаблонов. Первой линией обороны выступает защита конвейера обучения, которая реализуется через криптографическую верификацию датасетов хэш-функциями, децентрализацию данных с помощью федеративного обучения и внедрение дифференциальной приватности для блокировки атак инверсии модели. На стыке взаимодействия с пользователями применяется многоуровневая валидация через AI-файрволы, автоматизированное «shift-left» тестирование промптов на стадии разработки, а также ограничение частоты запросов для противодействия извлечению моделей и Cost Harvesting. Оперативное обна-

ружение угроз обеспечивается непрерывным мониторингом аномалий, детальным логированием телеметрии ИИ и архитектурной сегментацией компонентов агентных систем. Дополнительно организации используют механизмы коллективной разведки через платформу обмена данными об инцидентах AI Incident Sharing. Специфическая защита автономных агентов строится на принципе минимальных привилегий, изоляции сред их выполнения, верификации подписи кода для МСР-серверов и регулярном аудите конфигураций, что вкупе с технологиями детектирования дипфейков позволяет эффективно нейтрализовать сложные векторы кибератак.

MITRE ATLAS предоставляет организациям инструментарий для защиты AI-систем на всех этапах их жизненного цикла. Практическое применение фреймворка начинается с моделирования угроз, в ходе которого проводится инвентаризация AI-активов и их сопоставление с матрицей для построения структурированной карты рисков. Это особенно актуально для встраиваемых и промышленных систем, включая автомобильное компьютерное зрение, сенсорные платформы и ИИ-интерпретаторы сигналов ГНСС. В рамках Red Teaming фреймворк служит основой для симуляции реальных атак, а концепция «shift-left» позволяет автоматизировать тестирование моделей на устойчивость к инъекциям промптов и отравлению данных непосредственно в процессах CI/CD. При реагировании на инциденты ATLAS обеспечивает классификацию обнаруженных аномалий и восстановление всей цепочки атаки, что подкрепляется возможностью обмена опытом через платформу AI Incident Sharing.

В заключение, структурированная таксономия MITRE ATLAS обеспечивает индустрию общим языком и систематической основой для нейтрализации киберугроз. В условиях массовой интеграции машинного обучения в критическую инфраструктуру, IoT-платформы и промышленную автоматизацию, применение MITRE ATLAS становится обязательным условием для создания защищенных ИИ-систем следующего поколения.

СПИСОК ЛИТЕРАТУРЫ

- 1 ATLAS Matrix for AI Systems // MITRE ATLAS. – URL: <https://atlas.mitre.org/matrices/ATLAS> (date of access: 25.05.2026).
- 2 MITRE ATLAS explained: The complete guide to AI security threat intelligence // VECTRA. – URL: <https://www.vectra.ai/topics/mitre-atlas> (date of access: 25.05.2026).
- 3 Securing AI Systems with MITRE ATLAS // Medium. – URL: <https://medium.com/@zbabar/securing-ai-systems-with-mitre-atlas-89a263f38772> (date of access: 25.05.2026).

Получено 12.06.2026