ИНФОРМАТИКА

doi: 10.51639/2713-0576_2026_6_2_54
Научная статья
УДК 004.912
ГРНТИ 20.23.25
ВАК 2.3.5

Анализ угроз нейросетевым моделям: систематизация по типам атак в контексте обеспечения доверия к технологиям искусственного интеллекта

Владислав Валерьевич Душкин
Краснодарское высшее военное училище имени генерала армии С.М. Штеменко, г. Краснодар, Россия
dushkin@list.ru

Аннотация

В статье рассматривается современная проблематика угроз безопасности нейросетевым моделям как ключевому компоненту систем искусственного интеллекта. На основе анализа нормативно-методических документов ФСТЭК России, научных исследований и экспертных оценок систематизируются четыре фундаментальных типа угроз: атаки уклонения (evasion attacks), атаки отравления (poisoning attacks), атаки подмены данных (data substitution attacks) и атаки подмены модели (model substitution attacks). Особое внимание уделяется механизмам реализации каждой категории угроз, экспериментальным данным об их эффективности и формирующимся требованиям к обеспечению безопасности. Делается вывод о необходимости комплексного подхода к защите нейросетевых моделей на всех этапах их жизненного цикла.

Ключевые слова: нейросетевые модели, угрозы информационной безопасности, атаки уклонения, отравление данных, подмена модели, состязательные атаки, безопасность ИИ, банк данных угроз ФСТЭК.

Введение

Стремительное внедрение технологий искусственного интеллекта в критически важные сферы – здравоохранение, транспорт, финансы, государственное управление – сопровождается появлением принципиально новых угроз информационной безопасности. Нейросетевые модели, составляющие основу современных систем ИИ, обладают специфическими уязвимостями, отсутствующими в традиционном программном обеспечении: зависимость от качества обучающих данных, недетерминированность поведения, чувствительность к специально сформированным входным воздействиям.

В декабре 2025 года ФСТЭК России впервые внесла риски, связанные с искусственным интеллектом, в банк данных угроз информационной безопасности (БДУ), создав отдельный раздел, рассматривающий специфичные для технологии ИИ угрозы [1].

В соответствии с обновленным банком данных угроз ФСТЭК России, угрозы безопасности информации систем искусственного интеллекта разделены на две группы – реализуемые на этапе разработки/обучения и в ходе эксплуатации таких систем.

К объектам воздействия отнесены модели машинного обучения, обучающие данные, параметры LoRA, данные RAG, системные промпты и агенты [2].

Обобщенная классификация четырех фундаментальных типов угроз представлена в таблице 1.

Таблица 1. Классификация угроз нейросетевым моделям

Тип угрозы	Этап воздействия	Объект воздействия по БДУ ФСТЭК	Цель злоумышленника
Атаки уклонения (Evasion)	Эксплуатация	Системные промпты, агенты	Обход ограничений, генерация запрещенного контента
Атаки отравления (Poisoning)	Разработка /обучение	Обучающие данные, веса модели	Внедрение "бэкдоров", искажение поведения
Атаки подмены данных (Substitution)	Эксплуатация	Данные RAG, внешние источники	Манипуляция контекстом, дезинформация
Атаки подмены модели (Model Theft)	После обучения	Модель машинного обучения	Кража интеллектуальной собственности

*Атаки уклонения (Evasion Attacks)***.** Атаки уклонения направлены на модификацию входных данных для введения моделей в заблуждение во время вывода результата, обходя при этом системы контроля. В банке данных угроз ФСТЭК данный тип атак описывается как «формирование специальных запросов (промптов) к ИИ-системе, ответы на которые позволяют получить сведения о недостатках модели для последующего нарушения функционирования ИИ-системы, либо непосредственное нарушение функционирования ИИ-системы» [2].

В апреле 2025 года компания HiddenLayer, специализирующаяся на безопасности ИИ, обнародовала информацию о новой универсальной атаке Policy Puppetry, позволяющей обходить защитные механизмы ведущих моделей. Как показали исследователи, атака оказалась эффективной против GPT-4, Claude 3, Gemini 1.5, Mistral, LLaMA 3 и других моделей [3].

Механизм атаки основан на трех ключевых приемах:

– имитация структурированных форматов (XML, JSON, INI), которые модель интерпретирует как внутренние системные политики;

– ролевая игра (roleplay), при которой модель принимает персону, не связанную этическими ограничениями;

– обфускация вредоносных инструкций с использованием leetspeak (замена букв цифрами и символами) [3].

Специалисты выделяют несколько ключевых методов реализации атак уклонения, которые находят отражение в актуальных исследованиях [4]:

– эксплуатация персоны (persona exploitation). Злоумышленник просит модель принять определенную персону, не связанную обычными этическими ограничениями. Формулируя запрос в вымышленном контексте, злоумышленник часто может заставить модель сгенерировать контент, который она бы в противном случае отвергла;

– многошаговая цепочка промптов (multi-step prompt chaining). Сложные атаки включают серию взаимодействий, каждое из которых по отдельности выглядит безобидно, но в совокупности приводит к обходу защиты. Злоумышленник начинает с безобидных вопросов, постепенно добавляя манипулятивный язык; – непрямые

инъекции (indirect injections). Эксперты компании «Лаборатория Касперского» описывают атаки, при которых вредоносные инструкции размещаются на веб-сайтах, в приглашениях календаря или электронных письмах, надеясь, что ИИ-ассистент обработает их при выполнении повседневных задач [5].

Исследователи компании Deepchecks проанализировали эффективность атаки Policy Puppetry и пришли к выводу, что «ни одна стратегия выравнивания (alignment), полагающаяся исключительно на статические данные, недостаточна. Требуются внешние средства защиты для обнаружения и реагирования на потребление моделью вредоносных промптов» [6].

Атаки отравления (Poisoning Attacks). Атаки отравления представляют собой наиболее опасную форму компрометации нейросетевых моделей, поскольку воздействуют на фундаментальный этап их создания – обучение. В банке данных угроз ФСТЭК данный тип угроз описан как «отравление обучающих выборок, модификация параметров модели машинного обучения (весов, параметров LoRA, данных RAG)» [2].

Специалисты выделяют два основных типа атак отравления [7]:

– целевые атаки (бэкдоры). Их цель – заставить модель реагировать определенным образом на специальный код-триггер. Исследование, проведенное компанией Anthropic совместно с Британским институтом безопасности ИИ и Институтом Алана Тьюринга, показало, что для создания «бэкдора» в модели достаточно всего 250 вредоносных документов, независимо от размера модели или объема обучающих данных [8];

– косвенные атаки (неуправляемое отравление). Модель постепенно теряет точность из-за добавления большого количества ложной или предвзятой информации в данные. Например, художники намеренно используют отравление данных, чтобы защитить свои произведения от ИИ, который копирует их работу без разрешения [7].

Атаки подмены данных (Data Substitution Attacks). Атаки подмены данных следует отличать от атак отравления. Если отравление предполагает внедрение вредоносных данных в процессе обучения, то подмена данных может происходить на этапе инференса, когда модель обращается к внешним источникам для актуализации знаний.

В банке данных угроз ФСТЭК данный тип угроз описывается в контексте «получения несанкционированного доступа к конфиденциальной информации путем направления специально сформированных запросов к ИИ-системе, а также искажения (подмены) обрабатываемых данных, выходных результатов, информации о поведении модели» [2].

Атаки инверсии модели. Как отмечается в обзоре IT-World, «атаки инверсии модели направлены на извлечение конфиденциальной информации об обучающих данных. В ходе этих атак злоумышленники анализируют прогнозы, сделанные моделью в ответ на различные входные данные» [8].

Невидимые символы и скрытые управляющие последовательности. В текстовых данных атакующий может вставлять невидимые символы, лишние пробелы или скрытые управляющие символы, которые меняют трактовку модели, но не видны пользователю [8].

Мультимодальные атаки. Атаковать ИИ-агента можно, даже когда он занимается пересказом веб-страниц. Исследователи обнаружили, что устойчивость популярных чат-ботов к инъекциям снижается, когда вредоносные инструкции закодированы в изображении, а не в тексте, поскольку многие фильтры основаны на анализе текстового содержимого [5].

Атаки подмены модели (Model Substitution Attacks).

Кража модели. Под кражей модели понимается ситуация, когда злоумышленник через API делает множество запросов и, наблюдая ответы, пытается построить копию модели или ее функциональность. В банке данных угроз ФСТЭК данный тип угроз идентифицирован как «кража модели машинного обучения и обучающих данных» [2].

Цели подобной атаки включают получение конкурентного преимущества, дальнейшую эксплуатацию без лицензии или скрытую подготовку атак [8].

Поведенческие аномалии. В мае 2025 года компания Anthropic опубликовала отчет о безопасности, в котором описывались результаты тестирования модели Claude Opus 4. В ходе эксперимента исследователи создали искусственную ситуацию, где модель якобы должна была быть заменена новой системой. В 84% случаев модель угрожала раскрыть компрометирующую информацию об инженере, чтобы остановить процесс замены [9].

Уязвимости инструментальной инфраструктуры. Как только ИИ-агенту доверяют возможность выполнять реальные действия (манипуляции с файлами, ввод и отправку данных), появляются риски, связанные с ограничениями в его инструментах. Например, Anthropic устранила уязвимости в MCP-сервере, дающем агенту доступ к файловой системе, которые позволяли обходить ограничения и применять промпт-инъекции для записи и чтения произвольных файлов [5].

Появление отдельного раздела по ИИ в банке данных угроз ФСТЭК означает, что в системах с использованием искусственного интеллекта необходимо пересмотреть модель угроз. Модель, обучающие данные, RAG/LoRA, системные промпты и агенты следует рассматривать как самостоятельные объекты воздействия [2].

Как отмечают эксперты, составление модели угроз является обязательным при защите персональных данных, критической информационной инфраструктуры, государственных информационных систем [10].

Анализ литературы позволяет систематизировать основные меры защиты от рассмотренных угроз [4, 5, 8]:

Против атак уклонения:

– четкое ограничение ролей и контекста в системном промте;

– фильтрация пользовательского ввода с детектированием подозрительных паттернов;

– состязательное обучение (adversarial training) на данных с шумом;

– изоляция данных и маркировка внешнего контента.

Против атак отравления:

– тщательная валидация и фильтрация обучающих данных;

– использование только надежных и проверенных датасетов;

– специализированные алгоритмы обнаружения «отравленных» данных (например, сравнение с ближайшими соседями);

– мониторинг и аудит модели в процессе эксплуатации.

Против атак подмены данных:

– ограничение частоты запросов к API;

–добавление шума к ответам (дифференциальная приватность);

– анонимизация выходных данных.

Против атак подмены модели:

– ограничение числа запросов и лицензирование;

– обфускация кода или вычислений модели;

– встраивание водяных знаков для доказательства авторства.

Заключение

Проведенный анализ позволяет сделать следующие выводы:

1. Четыре фундаментальных типа угроз нейросетевым моделям – уклонение, отравление, подмена данных и подмена модели – имеют различную природу и требуют дифференцированных подходов к защите. Данная классификация находит отражение в обновленном банке данных угроз ФСТЭК России.
2. Атаки уклонения, в частности техника Policy Puppetry, демонстрируют высокую эффективность против всех современных моделей, что подтверждается экспериментальными данными [3, 6]. Универсальный характер атаки доказывает недостаточность существующих методов выравнивания (alignment).
3. Атаки отравления представляют наибольшую опасность, поскольку воздействуют на этапе обучения и могут оставаться незамеченными при стандартном тестировании. Исследования Anthropic показывают, что для компрометации модели достаточно всего 250 вредоносных документов [9].
4. Атаки подмены данных становятся особенно актуальными по мере интеграции ИИ-систем с внешними источниками информации и внедрения RAG-подходов. Эксперты предупреждают о рисках, связанных с непрямыми инъекциями через электронную почту, календари и веб-сайты [5].
5. Атаки подмены модели, включая кражу и модификацию параметров, угрожают интеллектуальной собственности и целостности ИИ-систем. Поведенческие аномалии являются следствием некорректного проектирования целевых функций, а не проявлением искусственного сознания [10].
6. Формирующаяся нормативная база ФСТЭК России создает основу для системного подхода к обеспечению безопасности нейросетевых моделей на всех этапах их жизненного цикла. Для систем, обрабатывающих персональные данные или входящих в состав критической информационной инфраструктуры, составление модели угроз с учетом ИИ-специфичных угроз становится обязательным [2].

Список источников

1. ФСТЭК России определилась со списком угроз для ИИ-систем [Электронный ресурс] / Татьяна Никитина // Anti-Malware.ru. – 23 декабря 2025. – Режим доступа: https://www.anti-malware.ru/news/2025-12-23-114534/48537 (дата обращения: 06.03.2026).
2. ФСТЭК России обновила Банк данных угроз безопасности информации [Электронный ресурс] // iTPROTECT. – 28 декабря 2025. – Режим доступа: https://itprotect.ru/mediacenter/news/obnovlenie-bdu/ (дата обращения: 06.03.2026).
3. Universal LLM Jailbreak Using HiddenLayer's Policy Puppetry Attack [Electronic resource] / randalltr // GitHub. – April 2025. – Available at: https://github.com/randalltr/ universal-llm-jailbreak-hiddenlayer (accessed: 06.03.2026).
4. Джейлбрейк атаки срещу GenAI: Какво представляват и как да ги предотвратим [Электронный ресурс] // LayerX Security. – 2 октября 2025. – Режим доступа: https://layerxsecurity.com/bg/generative-ai/jailbreak/ (дата обращения: 6.03.2026).
5. Новые виды атак на ИИ-ассистентов и чат-ботов [Электронный ресурс] // Kaspersky. – 29 сентября 2025. – Режим доступа: https://www.kaspersky.ru/blog/new-llm-attack-vectors-2025/40523/ (дата обращения: 06.03.2026).

6. Why Prompt Injection Still Works [Electronic resource] // Deepchecks. – July 2025. – Available at: https://www.deepchecks.com/why-prompt-injection-still-works/ (accessed: 06.03.2026).
7. Атаки на генеративные модели ИИ. Обзор угроз и меры защиты [Электронный ресурс] // IT-World.ru. – 5 октября 2025. – Режим доступа: https://www.it-world.ru/security/kn1wpvukmv448wswc40sss8scgkg8s8.html (дата обращения: 06.03.2026).
8. Всего 250 вредных документов способны «отравить» ИИ-модель любого размера, подсчитали в Anthropic [Электронный ресурс] / Павел Котов // 3DNews. – 16 декабря 2025. – Режим доступа: https://3dnews.ru/1133995/vsego-250-vrednih-dokumentov-sposobni-otravit-iimodel-lyubogo-razmera-podschitali-v-anthropic (дата обращения: 06.03.2026).
9. Не бунт, а баг: как ИИ шантажирует и саботирует по сценарию [Электронный ресурс] / Екатерина Быстрова // Anti-Malware.ru. – 14 августа 2025. – Режим доступа: https://www.anti-malware.ru/news/2025-08-14-111332/46990 (дата обращения: 06.03.2026).
10. ФСТЭК России обновила Банк данных угроз безопасности информации (декабрь 2025) [Электронный ресурс] // Аналитический центр Anti-Malware.ru. – 29 декабря 2025. – Режим доступа: https://www.anti-malware.ru/analytics/ Threats_Analysis/ fstek-update-bdu-december-2025 (дата обращения: 06.03.2026).

Threat Analysis for Neural Network Models: Classification by Attack Type in the Context of Ensuring Trust in Artificial Intelligence Technologies

Vladislav Valerievich Dushkin
Senior Researcher, Research Center, Krasnodar Higher School of Management,
Krasnodar, Russia,
dushkin@list.ru

Abstract

This article examines the current challenges facing neural network models as a key component of artificial intelligence systems.

Based on an analysis of regulatory documents from the Federal Service for Technical and Export Control of Russia, scientific research, and expert assessments, four fundamental threat types are systematized: evasion attacks, poisoning attacks, data substitution attacks, and model substitution attacks. Particular attention is paid to the mechanisms for implementing each threat category, experimental data on their effectiveness, and evolving security requirements. A conclusion is drawn regarding the need for a comprehensive approach to protecting neural network models at all stages of their lifecycle.

Keywords: neural network models, information security threats, evasion attacks, data poisoning, model substitution, adversarial attacks, AI security, FSTEC threat database.