**УДК 004.056.2**

**Ткачева Елизавета Григорьевна,** студент, Московский государственный технический университет им. Н.Э. Баумана, г. Москва

**Калашников Василий Сергеевич,** студент, Московский государственный технический университет им. Н.Э. Баумана, г. Москва

# АТАКА ОТРАВЛЕНИЯ ДАННЫХ

**Аннотация.** Статья посвящена анализу атак отравления данных (data poisoning) — метода манипуляции обучающими наборами для нарушения работы систем ИИ. Рассмотрены механизмы целевых и нецелевых атак, их последствия для безопасности критически важных систем (медицина, финансы) и стратегии защиты, включая аудит моделей, робастные алгоритмы и проверку целостности данных. Подчёркивается необходимость комбинирования классической кибербезопасности и инноваций в машинном обучении для минимизации рисков.

**Annotation.** The article examines data poisoning attacks — a method of manipulating training datasets to compromise AI systems. It explores targeted and non-targeted attack mechanisms, their impact on critical infrastructure (healthcare, finance), and defense strategies, including model auditing, robust algorithms, and data integrity verification. The study emphasizes the need to combine classical cybersecurity and machine learning innovations to mitigate risks.

**Ключевые слова:** биометрическая аутентификация, спуфинг-атаки, двухфакторная аутентификация, кибербезопасность, защита персональных данных.

**Keywords:** biometric authentication, spoofing attacks, two-factor authentication, cybersecurity, personal data protection.

В эпоху, когда данные становятся основой обучения искусственного интеллекта (ИИ), их целостность оказывается под угрозой из-за новых форм кибератак. Отравление данных, метод манипуляции обучающими наборами, способен незаметно искажать работу алгоритмов, ставя под сомнение их

надежность. Понимание механизмов таких атак и разработка эффективных контрмер становятся важнейшими задачами для обеспечения доверия к технологиям будущего и их безопасности.

**Что такое отравление данных?**

Data poisoning (в переводе с англ. – отравление данных) предполагает преднамеренное и злонамеренное «загрязнение» данных с целью ухудшения работы систем ИИ и машинного обучения (МО). В отличие от других методов атак, которые нацелены на модель во время выполнения (например, adversarial perturbations attacks), атаки отравления данных происходят на этапе обучения. Путем добавления, изменения или удаления определенных точек данных в обучающем наборе злоумышленники могут вызвать смещения, ошибки или уязвимости, которые проявляются, когда скомпрометированная модель принимает решения или делает прогнозы.

**Механизм отравления данных**

Механизм заключается в преднамеренном внесении искажений в обучающие данные с целью нарушения работы моделей МО. Атаки отравления данных можно разделить на две основные категории в зависимости от их цели: целевые атаки, при которых злоумышленник стремится повлиять на поведение модели для определенных входных данных, не ухудшая ее общую производительность, например, добавление отравленных данных может заставить систему распознавания лиц неправильно классифицировать лицо конкретного человека, и нецелевые атаки, направленные на снижение общей производительности модели через добавление шума или нерелевантных данных, что уменьшает точность, прецизионность или полноту модели для различных входных данных. Успех отравления данных зависит от трех ключевых компонентов: скрытности, при которой отравленные данные не должны быть легко обнаружимыми, чтобы избежать механизмов очистки или предварительной обработки данных; эффективности, требующей, чтобы атака приводила к желаемому ухудшению производительности модели или целевому неправильному поведению; и

последовательности, предполагающей, что эффекты атаки должны стабильно проявляться в различных контекстах или средах, где работает модель. Эти принципы делают отравление данных сложным и опасным инструментом, требующим комплексных мер защиты.

**Последствия для безопасности ИИ**

Отравление данных представляет собой серьезную угрозу для безопасности ИИ-систем, вызывая ряд критических проблем. Во-первых, компрометация целостности моделей: когда обучение происходит на отравленных данных, прогнозы и решения модели теряют свою надежность, даже если сама архитектура модели остается технически исправной и безопасной. Во-вторых, эволюция поверхности атаки: традиционные подходы кибербезопасности, направленные на защиту кода и инфраструктуры, оказываются недостаточными, поскольку отравление данных расширяет зону риска, включая в нее обучающие данные, что требует разработки новых методов защиты. В-третьих, использование отравленных моделей в критически важных системах, таких как здравоохранение, финансы или оборона, может привести к катастрофическим последствиям, так как ошибки в принятии решений способны повлечь за собой значительный ущерб. Эти аспекты подчеркивают необходимость комплексного подхода к защите данных и моделей, чтобы минимизировать риски, связанные с отравлением данных, и обеспечить безопасность ИИ-систем в различных сферах применения.

**Стратегии защиты**

Для эффективного противодействия отравлению данных необходим комплексный подход, сочетающий превентивные и реактивные меры. Первым шагом является внедрение строгих методов проверки данных, таких как статистический анализ, обнаружение аномалий и кластеризация, которые позволяют выявлять и устранять подозрительные точки данных до начала обучения модели. Регулярный аудит моделей машинного обучения помогает отслеживать их производительность и своевременно обнаруживать

отклонения, вызванные потенциальным отравлением. Использование разнообразных источников данных снижает зависимость от единого набора, минимизируя влияние отравленных выборок, а применение робастных алгоритмов обучения, включая усеченную среднеквадратичную ошибку или методы на основе медианных оценок, повышает устойчивость моделей к выбросам и искажениям. Кроме того, прозрачное отслеживание происхождения данных – фиксация источников, изменений и шаблонов доступа – обеспечивает основу для постфактумного анализа в случае подозрений на атаку. Сочетание этих стратегий формирует многоуровневую защиту, способную противостоять растущим угрозам в сфере безопасности искусственного интеллекта.

По мере того, как системы ИИ и машинного обучения проникают во все сферы жизни, способы реализации угроз становятся более изощренными, требуя сочетания классических методов кибербезопасности, глубокого понимания принципов МО и постоянного внедрения инноваций. Несмотря на то, что отравление данных представляет собой серьезную угрозу, оно также открывает новые возможности для развития технологий защиты и укрепления доверия к ИИ-системам.

**Литература**

1. Biggio B., Nelson B., Laskov P. Poisoning Attacks against Support Vector Machines // Proceedings of the 29th International Conference on Machine Learning (ICML). – 2012. – URL: https://arxiv.org/abs/1206.6389 (дата обращения: 09.02.2025).

2. Chen X., Liu C., Li B., Lu K., Song D. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning // IEEE Symposium on Security and Privacy. – 2020. – URL: https://ieeexplore.ieee.org/document/9152775 (дата обращения: 10.02.2025).

3. Barreno M., Nelson B., Joseph A. D., Tygar J. D. The Security of Machine Learning // Machine Learning. – 2010. – Vol. 81. – P. 121–148.

4. Организация по стандартизации ISO/IEC TR 24027:2021 Информационные технологии — Искусственный интеллект — Уязвимости и атаки в системах машинного обучения. – URL: https://www.iso.org/standard/77608.html (дата обращения: 11.02.2025).

5. Шнайер Б. Искусственный интеллект и безопасность: вызовы и решения. – М.: Издательский дом «Вильямс», 2023. – 320 с.

**Literature**

1. Biggio B., Nelson B., Laskov P. Poisoning Attacks against Support Vector Machines // Proceedings of the 29th International Conference on Machine Learning (ICML). – 2012. – URL: https://arxiv.org/abs/1206.6389 (access date: 02/09/2025).

2. Chen X., Liu C., Li B., Lu K., Song D. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning // IEEE Symposium on Security and Privacy. – 2020. – URL: https://ieeexplore.ieee.org/document/9152775 (access date: 02/10/2025).

3. Barreno M., Nelson B., Joseph A. D., Tygar J. D. The Security of Machine Learning // Machine Learning. – 2010. – Vol. 81. – P. 121–148.

4. ISO/IEC TR 24027:2021 Information technology — Artificial intelligence — Vulnerabilities and attacks in machine learning systems. – URL: https://www.iso.org/standard/77608.html (accessed: 11.02.2025).

5. Schneier B. Artificial intelligence and security: challenges and solutions. – M.: Williams Publishing House, 2023. – 320 p.