DATA SECURITY AND PRIVACY: SECURITY BEHAVIOR IN ARTIFICIAL INTELLIGENCE

Ablizova Gulzahiryam Alimovna

Senior Lecturer, Department of Modern Information Technologies, Uzbek State University of World Languages

Aripova Culchehra Ishankulovna,

Lecturer, Department of Modern Information Technologies, Uzbek State University of World Languages

ABSTRACT

This paper analyzes the theoretical foundations, technical mechanisms, and principles for managing security behavior to ensure data security and privacy in artificial intelligence systems. It examines in detail the types of risks that arise when AI models operate on data, including potential threats, adversarial attacks, data leakage, and model reconstruction. The study demonstrates the effectiveness of advanced security approaches such as differential privacy, federated learning, encrypted computing, permission management, and security by design. The paper proposes an integrated security architecture for the stable and reliable operation of AI systems, based on ethical, legal, and technical requirements. The results identify complex security challenges in AI systems and propose scientifically based solutions to address them.

Keywords: artificial intelligence, data security, privacy, security behavior, differential privacy, federated learning, encrypted computing, adversarial attacks, data protection, security design, algorithmic resilience, risk analysis.

БЕЗОПАСНОСТЬ ДАННЫХ И КОНФИДЕНЦИАЛЬНОСТЬ: ПОВЕДЕНИЕ БЕЗОПАСНОСТИ В ИСКУССТВЕННОМ ИНТЕЛЛЕКТЕ

Аблизова Гульзахирям Алимовна

Старший преподаватель кафедры современных информационных технологий Узбекского государственного университета мировых языков

Арипова Кулчехра Ишанкуловна,

Преподаватель кафедры современных информационных технологий Узбекского государственного университета мировых языков

АННОТАЦИЯ

В данной статье анализируются теоретические основы, технические механизмы и принципы управления поведением безопасности для обеспечения безопасности данных и конфиденциальности в системах искусственного интеллекта. Подробно изучаются виды рисков, возникающих при работе моделей ИИ с данными, такие как потенциальные угрозы, враждебные атаки, утечка данных и реконструкция моделей. Исследование демонстрирует эффективность передовых подходов к защите, таких как дифференциальная приватность, федеративное обучение, шифрованные вычисления, управление разрешениями и проектируемая безопасность. В статье предлагается интегрированная архитектура безопасности для

стабильной и надежной работы систем ИИ, основанная на этических, правовых и технических требованиях. Результаты выявляют сложные проблемы безопасности систем ИИ и предлагают научно обоснованные решения для их решения.

Ключевые слова: искусственный интеллект, безопасность данных, конфиденциальность, поведение безопасности, дифференциальная приватность, федеративное обучение, шифрованные вычисления, враждебные атаки, защита данных, проектирование систем безопасности, алгоритмическая устойчивость, анализ рисков.

Введение

Стремительное развитие цифровых технологий в глобальном масштабе приводит к качественным изменениям во всех аспектах человеческой деятельности. В частности, технологии искусственного интеллекта (ИИ) широко используются в экономике, здравоохранении, образовании, управлении, безопасности и повседневной жизни. Изначально рассматриваемый как средство упрощения автоматизированных процессов, ИИ теперь превратился в интеллектуальную платформу, способную принимать самостоятельные решения, обрабатывать данные в режиме реального времени и управлять сложными системами. Наряду с расширением этих возможностей возрастают новые риски, растёт число киберугроз и вероятность нарушения конфиденциальности данных. Поэтому обеспечение безопасности данных в системах ИИ признано не только технологической проблемой, но и одним из ключевых факторов национальной безопасности, экономической стабильности и общественного доверия. Основным ресурсом моделей ИИ являются данные, которые отличаются широким охватом, разнообразием качества и часто содержат конфиденциальную персональную информацию. Алгоритмы, построенные на этих данных, часто управляют

такими процессами, как идентификация пользователей, прогнозирование поведения и автоматизация принятия решений. Следовательно, незащищённая обработка данных или их незаконное распространение нарушает неприкосновенность частной жизни, порождает социальную несправедливость и резко снижает доверие к системам ИИ. Сегодня многие страны принимают строгие нормативно-правовые акты для защиты данных, в частности, такие мировые стандарты, как GDPR и CCPA. Это предъявляет особые требования не только к технической, но и к этической и юридической ответственности систем ИИ. Сложная архитектура систем ИИ еще больше увеличивает угрозы безопасности. Существуют такие факторы риска, как враждебные атаки на модель, внедрение данных, восстановление модели путем обратной разработки и косвенное извлечение данных. Эти угрозы могут иметь не только технические, но и социальные последствия. Например, алгоритмическое решение, принятое на основе неверной информации, может нанести прямой ущерб правам и свободам граждан. В результате обеспечение безопасности и конфиденциальности данных имеет первостепенное и стратегическое значение при внедрении систем ИИ.

Поведение безопасности ИИ относится к тому, как алгоритм реагирует на различные условия, как он адаптируется к угрозам и как он поддерживает согласованность обработки данных. Системы без должным образом разработанного поведения безопасности принимают неверные решения, подвержены манипуляциям и приводят к неожиданным результатам. Таким образом, разработка систем ИИ требует опоры на принципы «безопасности при проектировании» и приоритета безопасности на всех этапах, от обучения модели до этапов эксплуатации и мониторинга. Кроме того, ответственное использование систем ИИ, типы используемых в них данных, согласие пользователей, алгоритмическая прозрачность и интеграция понятий объяснимого ИИ (XAI) являются неотъемлемыми элементами обеспечения безопасности данных. С этой точки зрения, данная статья направлена на

разработку научно обоснованного подхода к архитектуре безопасности систем ИИ, механизмам обеспечения конфиденциальности данных, анализу рисков и формированию безопасного поведения.

Научно-методической основой данного исследования является изучение механизмов обеспечения безопасности данных и конфиденциальности в системах искусственного интеллекта с использованием концептуального, системного и аналитического подходов. Предметом исследования является изучение угроз безопасности, видов рисков, механизмов защиты и последовательного управления алгоритмическим поведением моделей ИИ при работе с данными. Методология основывалась на теоретическом моделировании, сравнительном анализе, научном обобщении, оценке рисков и методах концептуального проектирования. На первом этапе исследования были изучены научная литература, международные стандарты (ISO/IEC 27001, ISO/IEC 23894), правовые нормы (GDPR, CCPA) и передовые технические исследования с использованием метода контент-анализа. Данный подход позволил выявить мировые научные парадигмы в области безопасности ИИ и синтезировать повторяющиеся концепции, связанные с конфиденциальностью данных. Также было проведено системное сравнение преимуществ и ограничений существующих технологических решений.

Следующий этап методологии включал функциональный анализ архитектуры ИИ и потоков данных. В ходе этого процесса были выявлены факторы риска, возникающие на этапах жизненного цикла данных – сбор, хранение, передача, обработка и уничтожение. Для каждого этапа была разработана матрица потенциальных угроз, уровень которых был оценен с использованием качественных и аналитических показателей. В результате были выявлены технические угрозы, напрямую влияющие на данные (состязательные атаки, инъекции, обратная разработка, утечка данных), и социально-этические риски, косвенно влияющие на данные (алгоритмическая

дискриминация, нарушения конфиденциальности, неверные решения). Третьим методологическим подходом исследования стало концептуальное моделирование механизмов защиты. Теоретически исследовались такие механизмы, как дифференциальная приватность, федеративное обучение, шифрованные вычисления, системы управления разрешениями, автоматический аудит политик безопасности, и оценивались их потенциальные области применения. На данном этапе была разработана модель интеграции принципа «безопасность по проектированию» в системы ИИ. Модель определила согласованные механизмы контроля безопасности на входных, обрабатывающих и выходных звеньях системы ИИ. Важной частью методологии стал подход к анализу поведения безопасности. Теоретически оценивались такие показатели, как реакция алгоритма в различных условиях, его подверженность манипулятивной информации и уровень согласованности в принятии решений. Для моделирования поведения использовался системный подход, а в качестве основных критериев были выбраны коэффициент устойчивости модели, предсказуемость решений и скорость реагирования на риск. На заключительном этапе методологии были обобщены полученные научные выводы и разработана идеализированная архитектурная концепция, рекомендуемая для обеспечения безопасности данных систем ИИ. Эта архитектура представляет собой интегрированное состояние технических защит, политик безопасности, нормативных требований и этических стандартов. Такой многоэтапный и многоуровневый подход к методологии позволил всесторонне охватить вопросы безопасности ИИ.

Заключение

Широкое использование технологий искусственного интеллекта стало важным фактором экономического, социального и культурного развития

современного общества. Однако по мере роста возможностей систем ИИ возрастают и требования к их безопасности. В данном исследовании представлен углубленный анализ теоретических, технических и методологических аспектов обеспечения безопасности и конфиденциальности данных в системах ИИ. Полученные результаты показывают, что защита данных является неотъемлемой частью архитектуры ИИ, определяющей надежность, стабильность и социальное принятие системы. В ходе исследования было выявлено наличие факторов риска на всех этапах обработки данных – от сбора до хранения, передачи, обработки и уничтожения – и предложены эффективные механизмы защиты от них. В частности, было показано, что такие подходы, как дифференциальная приватность, федеративное обучение, шифрованные вычисления, строгое управление разрешениями и непрерывный аудит политик безопасности, способны обеспечить высокий уровень защиты в системах ИИ. Интеграция таких механизмов снижает вероятность нарушения конфиденциальности данных и обеспечивает прозрачность и согласованность работы алгоритмов.

Использованная литература:

1. Ahmed, S., & Shapiro, A. (2022). *Artificial intelligence governance and data protection in digital ecosystems*. Journal of Cybersecurity Studies, 14(2), 85–103.
2. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
3. Brundage, M., Avin, S., Clark, J., Toner, H., & Eckersley, P. (2020). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. AI Policy Foundation.

4. European Union Agency for Cybersecurity (ENISA). (2023). *AI cybersecurity challenges and safeguards*. ENISA Publications.
5. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1), 1–15.