УДК: —
МРНТИ: —

[1] Мубарак Максат Серикулы*, Түлкібай Дінмұхамед Әбділдаұлы
[1]Национальный университет обороны Республики Казахстан
**maks.pvl7182@mail.ru*

РАЗРАБОТКА МОДЕЛИ ОБЕСПЕЧЕНИЯ КИБЕРБЕЗОПАСНОСТИ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА НА ОСНОВЕ МНОГОУРОВНЕВОЙ АРХИТЕКТУРЫ ЗАЩИТЫ И ИНТЕЛЛЕКТУАЛЬНОЙ ОЦЕНКИ РИСКОВ

Аннотация. Аннотация. Стремительное развитие технологий искусственного интеллекта (ИИ) и их широкое внедрение в государственное управление, промышленность, здравоохранение, образование, финансовый сектор и оборонную сферу сопровождаются возникновением новых угроз информационной безопасности. Современные системы искусственного интеллекта, включая большие языковые модели, генеративные модели и интеллектуальные системы поддержки принятия решений, обладают специфическими уязвимостями, которые не учитываются традиционными средствами киберзащиты. Наиболее распространенными угрозами являются атаки на обучающие данные, извлечение параметров модели, внедрение состязательных примеров, обход механизмов безопасности посредством Prompt Injection и Jailbreak, а также утечка конфиденциальной информации. В статье проведен анализ современных подходов к обеспечению кибербезопасности систем искусственного интеллекта, рассмотрены международные стандарты и рекомендации в области AI Security, включая NIST AI Risk Management Framework, MITRE ATLAS и OWASP Top 10 for Large Language Model Applications. На основе анализа существующих исследований выявлены ограничения существующих методов защиты, заключающиеся в отсутствии комплексного подхода к обеспечению безопасности на всех этапах жизненного цикла интеллектуальной системы. Предложена авторская многоуровневая модель обеспечения кибербезопасности систем искусственного интеллекта, включающая уровни защиты данных, моделей машинного обучения, пользовательского взаимодействия, вычислительной инфраструктуры, мониторинга безопасности и управления рисками. В отличие от существующих решений, модель интегрирует классические механизмы информационной безопасности с интеллектуальными средствами анализа угроз и адаптивной оценкой рисков, что позволяет повысить устойчивость AI-систем к современным кибератакам. Практическая значимость исследования заключается в возможности применения предложенной модели при разработке защищенных интеллектуальных информационных систем, функционирующих в государственных, корпоративных и военных организациях.

Ключевые слова: Ключевые слова: искусственный интеллект, кибербезопасность, информационная безопасность, большие языковые модели, Security, машинное обучение, оценка рисков, Prompt Injection, Zero Trust, MITRE ATLAS.

[1] Мубарак Максат Серикулы, Түлкібай Дінмұхамед Әбділдаұлы

[1]Национальный университет обороны Республики Казахстан

КӨП ДЕҢГЕЙЛІ ҚОРҒАНЫС АРХИТЕКТУРАСЫ МЕН ТӘУЕКЕЛДЕРДІ ИНТЕЛЛЕКТУАЛДЫ БАҒАЛАУ НЕГІЗІНДЕ ЖАСАНДЫ ИНТЕЛЛЕКТ ЖҮЙЕЛЕРІНІҢ КИБЕРҚАУІПСІЗДІГІН ҚАМТАМАСЫЗ ЕТУ МОДЕЛІН ӘЗІРЛЕУ

Аңдатпа. Аннотация. Жасанды интеллект (АИ) технологияларының қарқынды дамуы және оларды мемлекеттік басқаруға, өнеркәсіпке, денсаулық сақтауға, білімге, қаржы секторына және қорғаныс саласына кеңінен енгізу ақпараттық қауіпсіздікке жаңа қатерлердің пайда болуымен қатар жүреді. Қазіргі заманғы жасанды интеллект жүйелері, соның ішінде үлкен тілдік модельдер, генеративті модельдер және шешімдерді қолдаудың интеллектуалды жүйелері дәстүрлі киберқауіпсіздік құралдарымен ескерілмейтін ерекше осалдықтарға ие. Ең жиі кездесетін қауіптер-оқу деректеріне шабуыл жасау, үлгі параметрлерін шығару, қарсылас мысалдарды енгізу, Prompt Injection және Jailbreak арқылы қауіпсіздік механизмдерін айналып өту және құпия ақпараттың ағып кетуі. Мақалада жасанды интеллект жүйелерінің киберқауіпсіздігін қамтамасыз етудің заманауи тәсілдеріне талдау жасалды, NIST ai Risk Management Framework, MITRE ATLAS және LASP Top 10 for Large Language Model Applications сияқты AI Security саласындағы халықаралық стандарттар мен ұсынымдар қарастырылды. Қолданыстағы зерттеулерді талдау негізінде зияткерлік жүйенің өмірлік циклінің барлық кезеңдерінде қауіпсіздікті қамтамасыз етудің кешенді тәсілінің болмауынан тұратын қолданыстағы қорғау әдістерінің шектеулері анықталды. Деректерді қорғау, Машиналық оқыту модельдері, пайдаланушылардың өзара әрекеттесуі, есептеу инфрақұрылымы, қауіпсіздік мониторингі және тәуекелдерді басқару деңгейлерін қамтитын жасанды интеллект жүйелерінің киберқауіпсіздігін қамтамасыз етудің авторлық көп деңгейлі моделі ұсынылған. Қолданыстағы шешімдерден айырмашылығы, модель ақпараттық қауіпсіздіктің классикалық механизмдерін зияткерлік қауіп-қатерді талдау құралдарымен және тәуекелдерді адаптивті бағалаумен біріктіреді, бұл AI жүйелерінің заманауи кибершабуылдарға төзімділігін арттыруға мүмкіндік береді. Зерттеудің практикалық маңыздылығы мемлекеттік, корпоративтік және әскери ұйымдарда жұмыс істейтін қорғалған зияткерлік ақпараттық жүйелерді әзірлеу кезінде ұсынылған модельді қолдану мүмкіндігі болып табылады.

Түйінді сөздер: Түйін сөздер: жасанды интеллект, киберқауіпсіздік, Ақпараттық қауіпсіздік, үлкен тілдік модельдер, Машиналық оқыту, тәуекелдерді бағалау, Prompt Injection, Zero Trust, MITRE ATLAS.

[1] Мубарак Максат Серикулы, Түлкібай Дінмұхамед Әбділдаұлы

[1]Национальный университет обороны Республики Казахстан

DEVELOPMENT OF A CYBERSECURITY MODEL FOR ARTIFICIAL INTELLIGENCE SYSTEMS BASED ON A MULTI-LEVEL PROTECTION ARCHITECTURE AND INTELLIGENT RISK ASSESSMENT

Abstract. Annotation. The rapid development of artificial intelligence (AI) technologies and their widespread adoption in public administration, industry, healthcare, education, the financial sector and the defense sector are accompanied by the emergence of new threats to information security. Modern artificial intelligence systems, including large language models, generative models, and intelligent decision support systems, have specific vulnerabilities that are not taken into account by traditional cyber defenses. The most common threats are attacks on training data, extraction of model parameters, implementation of adversarial examples, circumvention of security mechanisms through Prompt Injection and Jailbreak, as well as leakage of confidential information. The article analyzes modern approaches to ensuring cybersecurity of artificial intelligence systems, examines international standards and recommendations in the field of AI Security, including the NIST AI Risk Management Framework, MITRE ATLAS and OWASP Top 10 for Large Language Model Applications. Based on the analysis of existing research, the limitations of existing protection methods have been identified, which consist in the lack of an integrated approach to ensuring security at all stages of the life cycle of an intelligent system. The author's multilevel model of ensuring cybersecurity of artificial intelligence systems is proposed, including levels of data protection, machine learning models, user interaction, computing infrastructure, security monitoring and risk management. Unlike existing solutions, the model integrates classical information security mechanisms with intelligent threat analysis tools and adaptive risk assessment, which makes it possible to increase the resilience of AI systems to modern cyber attacks. The practical significance of the research lies in the possibility of applying the proposed model in the development of secure intelligent information systems operating in government, corporate and military organizations.

Keywords: Keywords: artificial intelligence, cybersecurity, information security, large language models, Security, machine learning, risk assessment, Prompt Injection, Zero Trust, MITRE ATLAS.

Введение. В последние годы технологии искусственного интеллекта стали одним из ключевых факторов цифровой трансформации общества. Алгоритмы машинного обучения и глубокие нейронные сети активно применяются при автоматизации производственных процессов, анализе больших объемов данных, медицинской диагностике, интеллектуальном управлении транспортными системами, финансовом прогнозировании, обработке естественного языка и поддержке принятия решений. Особое развитие получили генеративные модели искусственного интеллекта, способные создавать текст, изображения, программный код и аналитические материалы, что значительно расширило область применения интеллектуальных технологий.

Наряду с очевидными преимуществами широкое внедрение искусственного интеллекта сопровождается возникновением новых угроз информационной безопасности. В отличие от традиционных информационных систем, функционирование которых определяется заранее заданными алгоритмами, системы искусственного интеллекта принимают решения на основе анализа больших массивов данных и результатов обучения математических моделей. Это создает дополнительные поверхности атак, связанные с возможностью воздействия на обучающие выборки, параметры моделей, процессы принятия решений и взаимодействие пользователей с интеллектуальными сервисами.

В последние годы исследователями выделена отдельная область кибербезопасности — **AI Security**, охватывающая вопросы защиты систем искусственного интеллекта на всех этапах их жизненного цикла. Особое внимание уделяется угрозам, характерным исключительно для интеллектуальных систем: внедрению вредоносных данных в обучающие выборки (Data Poisoning), состязательным атакам (Adversarial Attacks), извлечению параметров моделей (Model Extraction), восстановлению обучающих данных (Membership Inference), а также атакам на большие языковые модели, включая Prompt Injection и Jailbreak. Реализация подобных атак способна привести к нарушению конфиденциальности, целостности и доступности информации, снижению точности принимаемых решений и компрометации интеллектуальных сервисов.

В ответ на появление новых угроз международными организациями разработан ряд нормативных и методических документов, определяющих принципы безопасного применения искусственного интеллекта. К ним относятся NIST AI Risk Management Framework, ISO/IEC 23894:2023, MITRE ATLAS, OWASP Top 10 for Large Language Model Applications и рекомендации ENISA. Несмотря на значительный прогресс в данной области, большинство существующих подходов ориентировано на решение отдельных задач защиты, таких как управление рисками, безопасность данных или обеспечение устойчивости моделей к отдельным видам атак. При этом отсутствует единая архитектура, объединяющая защиту всех компонентов интеллектуальной системы в рамках комплексной модели обеспечения кибербезопасности.

Для государственных информационных систем и объектов критической информационной инфраструктуры проблема обеспечения безопасности искусственного интеллекта приобретает особую актуальность. Использование интеллектуальных технологий в системах государственного управления, обороны, здравоохранения и энергетики требует разработки новых методов защиты, способных учитывать особенности функционирования моделей искусственного интеллекта, характер современных угроз и необходимость непрерывного мониторинга безопасности.

Целью исследования является разработка многоуровневой модели обеспечения кибербезопасности систем искусственного интеллекта, основанной на интеграции механизмов защиты данных, моделей машинного обучения, пользовательского взаимодействия, вычислительной инфраструктуры и интеллектуальной оценки рисков.

Для достижения поставленной цели определены следующие задачи исследования:

провести анализ современных угроз кибербезопасности систем искусственного интеллекта;

исследовать существующие международные подходы к обеспечению AI Security;

разработать многоуровневую модель защиты систем искусственного интеллекта;

предложить метод интеллектуальной оценки рисков;

определить практические рекомендации по внедрению разработанной модели в защищенные информационные системы.

Научная новизна исследования заключается в разработке интегрированной модели обеспечения кибербезопасности систем искусственного интеллекта, объединяющей механизмы защиты данных, моделей машинного обучения, пользовательских запросов, вычислительной инфраструктуры и процессов управления рисками. В отличие от существующих подходов предложенная модель обеспечивает комплексную защиту интеллектуальных систем на протяжении всего жизненного цикла и может быть адаптирована для использования в государственных, корпоративных и военных информационных системах.

Материалы и методы исследования. Развитие технологий искусственного интеллекта привело к появлению нового класса киберугроз, направленных непосредственно на алгоритмы машинного обучения, большие языковые модели и интеллектуальные информационные системы. В отличие от традиционных объектов информационной безопасности, модели искусственного интеллекта обладают рядом особенностей, связанных с использованием обучающих данных, механизмов самообучения и адаптации, что требует применения специализированных методов защиты.

В рамках настоящего исследования использован комплексный подход, объединяющий методы системного анализа, анализа рисков, сравнительного анализа международных стандартов и моделирования архитектуры безопасности интеллектуальных систем.

Методологической основой исследования послужили международные стандарты и рекомендации в области информационной безопасности и искусственного интеллекта, включая NIST AI Risk Management Framework (AI RMF 1.0), ISO/IEC 23894:2023 «Artificial Intelligence – Risk Management», MITRE ATLAS, OWASP Top 10 for Large Language Model Applications, а также положения серии стандартов ISO/IEC 27001 и ISO/IEC 27005, регламентирующих процессы управления информационной безопасностью и рисками.

Для определения современного состояния проблемы проведен анализ научных публикаций отечественных и зарубежных исследователей, посвященных вопросам защиты моделей машинного обучения, безопасности больших языковых моделей, состязательных атак (Adversarial Attacks), защиты обучающих данных и управления рисками искусственного интеллекта. Анализ литературных источников показал, что большинство существующих исследований сосредоточено на отдельных аспектах безопасности ИИ, таких как предотвращение атак на модели, защита данных или повышение устойчивости алгоритмов машинного обучения. При этом комплексные архитектурные решения, обеспечивающие защиту интеллектуальных систем на всех этапах жизненного цикла, представлены в ограниченном количестве.

В работе применен метод системной декомпозиции, позволивший представить систему искусственного интеллекта в виде совокупности взаимосвязанных компонентов, каждый из которых характеризуется собственными угрозами, уязвимостями и механизмами защиты. Такой подход обеспечивает возможность построения многоуровневой архитектуры безопасности, учитывающей специфику функционирования интеллектуальных систем.

Для оценки защищенности компонентов искусственного интеллекта использован риск-ориентированный подход, основанный на определении вероятности реализации угроз и потенциального ущерба от их воздействия. В рамках исследования риск рассматривается как функция вероятности возникновения угрозы и тяжести последствий для конфиденциальности, целостности и доступности данных, а также корректности функционирования интеллектуальной модели.

С целью структурирования угроз проведена их классификация по объектам воздействия. Предлагается выделить следующие основные группы угроз:

- угрозы, направленные на обучающие данные (Data Poisoning, Data Leakage);
- угрозы, связанные с компрометацией модели искусственного интеллекта (Model Extraction, Model Inversion, Membership Inference);
- угрозы пользовательского взаимодействия (Prompt Injection, Jailbreak, Prompt Leakage);
- угрозы вычислительной инфраструктуры (атаки на облачные сервисы, API, контейнерные платформы и цепочки поставок программного обеспечения);
- угрозы нарушения доступности интеллектуальных сервисов (Denial of Service, Resource Exhaustion);
- организационные угрозы, связанные с недостаточным уровнем управления безопасностью, отсутствием контроля жизненного цикла моделей и несоблюдением требований нормативных документов.

Проведенный анализ показал, что существующие методы защиты преимущественно ориентированы на предотвращение отдельных видов атак и не обеспечивают комплексного управления безопасностью систем искусственного интеллекта. Например, использование механизмов фильтрации пользовательских запросов позволяет снизить вероятность успешной реализации атак типа Prompt Injection, однако не предотвращает компрометацию обучающих данных или извлечение параметров модели. Аналогично применение технологий Differential Privacy обеспечивает защиту персональных данных, но не гарантирует устойчивость модели к состязательным воздействиям.

Таким образом, эффективное обеспечение кибербезопасности интеллектуальных систем возможно только при использовании интегрированного подхода, объединяющего защиту всех компонентов жизненного цикла искусственного интеллекта в рамках единой архитектуры.

На основании результатов проведенного анализа в настоящем исследовании разработана концептуальная многоуровневая модель обеспечения кибербезопасности систем искусственного интеллекта. В отличие от существующих решений предложенная модель объединяет механизмы защиты данных, моделей машинного обучения, пользовательского взаимодействия, вычислительной инфраструктуры, мониторинга безопасности и управления рисками, что позволяет обеспечить комплексную защиту интеллектуальной системы от современных киберугроз.

Результаты исследования

Разработка многоуровневой модели обеспечения кибербезопасности систем искусственного интеллекта

Проведенный анализ существующих методов обеспечения безопасности искусственного интеллекта показал, что большинство современных подходов ориентировано на защиту отдельных компонентов интеллектуальной системы. В частности, одни методы направлены на повышение устойчивости моделей машинного обучения к состязательным воздействиям, другие – на защиту обучающих данных или предотвращение несанкционированного доступа к вычислительной инфраструктуре. Такой подход не обеспечивает комплексного противодействия современным киберугрозам, поскольку успешная атака на один из компонентов может привести к компрометации всей интеллектуальной системы.

Для устранения данного недостатка разработана многоуровневая модель обеспечения кибербезопасности систем искусственного интеллекта (Multi-Layer Artificial Intelligence Security Model – MAISM). Модель основана на принципе эшелонированной защиты и предусматривает последовательную реализацию механизмов безопасности на всех этапах жизненного цикла интеллектуальной системы – от подготовки данных до эксплуатации и сопровождения модели.

В отличие от существующих решений предложенная модель рассматривает систему искусственного интеллекта как совокупность взаимосвязанных уровней безопасности, каждый из которых выполняет самостоятельные функции, но одновременно взаимодействует с другими уровнями посредством обмена информацией о состоянии защищенности и выявленных угрозах.

Архитектура модели включает шесть функциональных уровней (рисунок 1).

Первый уровень – уровень безопасности данных (Data Security Layer). На данном уровне обеспечивается контроль качества и достоверности данных, используемых при обучении и эксплуатации моделей искусственного интеллекта. Основными задачами являются проверка происхождения данных, обнаружение аномальных записей, предотвращение атак типа *Data Poisoning*, защита конфиденциальной информации и контроль целостности обучающих выборок. Для реализации указанных функций могут применяться методы статистического анализа, криптографической защиты, цифровых подписей, технологий *Differential Privacy* и автоматизированной очистки данных.

Второй уровень – уровень безопасности моделей искусственного интеллекта (Model Security Layer). Данный уровень предназначен для защиты параметров модели от несанкционированного доступа и модификации. Здесь реализуются механизмы контроля целостности модели, проверки цифровых подписей, защиты интеллектуальной собственности разработчиков, а также средства обнаружения атак *Model Extraction, Model Inversion, Membership Inference* и внедрения скрытых уязвимостей (*Backdoor Attack*). Кроме того, выполняется периодическое тестирование модели на устойчивость к состязательным воздействиям с использованием специальных тестовых наборов данных.

Третий уровень – уровень безопасности взаимодействия (Interaction Security Layer). Особенностью современных генеративных моделей является непосредственное взаимодействие с пользователями посредством текстовых запросов. Именно поэтому одним из наиболее уязвимых компонентов являются механизмы обработки пользовательских инструкций. Для минимизации риска эксплуатации языковых моделей на данном уровне реализуются фильтрация запросов, анализ семантического содержания команд, контроль системных инструкций и автоматическое выявление попыток обхода встроенных механизмов защиты. Особое внимание уделяется обнаружению атак *Prompt Injection, Jailbreak, Prompt Leakage* и других способов манипулирования поведением модели.

Четвертый уровень –уровень безопасности вычислительной инфраструктуры (Infrastructure Security Layer). Он включает традиционные механизмы информационной безопасности, обеспечивающие защиту серверов, облачной инфраструктуры, сетевых соединений, прикладных интерфейсов (API), контейнерных платформ и средств виртуализации. На данном уровне реализуются механизмы управления доступом, сегментации сети, многофакторной аутентификации, защиты каналов передачи данных, а также принципы архитектуры Zero Trust, предусматривающей постоянную проверку подлинности всех субъектов взаимодействия независимо от их местоположения.

Пятый уровень – уровень интеллектуального мониторинга безопасности (AI Security Monitoring Layer). В отличие от традиционных систем мониторинга информационной безопасности данный уровень ориентирован на непрерывный анализ поведения модели искусственного интеллекта. Он обеспечивает сбор журналов взаимодействия пользователей, анализ аномалий, корреляцию событий безопасности, выявление нетипичного поведения модели и прогнозирование возможных атак. Для повышения эффективности мониторинга могут использоваться методы машинного обучения, позволяющие обнаруживать ранее неизвестные угрозы на основе анализа отклонений от нормального режима функционирования.

Шестой уровень – уровень управления рисками (AI Risk Governance Layer). Верхний уровень модели обеспечивает стратегическое управление кибербезопасностью интеллектуальной системы. Он включает процессы оценки рисков, планирования

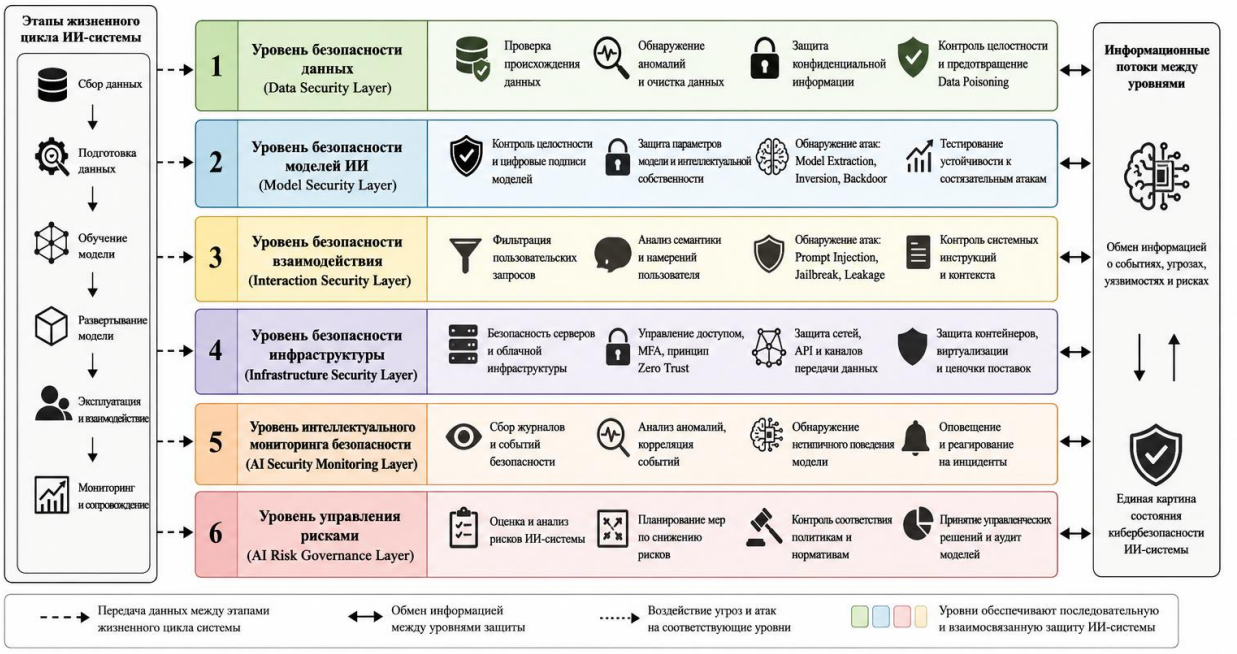
мероприятий по снижению угроз, аудита моделей, контроля соответствия требованиям нормативных документов, а также принятия управленческих решений по обеспечению безопасности. На данном уровне формируется единая картина состояния защищенности системы и осуществляется координация работы остальных компонентов архитектуры.

Предлагаемая модель обеспечивает комплексную защиту интеллектуальной системы благодаря последовательному взаимодействию всех уровней. Информация о выявленных инцидентах, аномалиях и уязвимостях передается между уровнями в режиме реального времени, что позволяет оперативно реагировать на изменение уровня угроз и адаптировать механизмы защиты к текущей обстановке.

По сравнению с существующими архитектурами предложенная модель обладает рядом преимуществ. Во-первых, она охватывает весь жизненный цикл системы искусственного интеллекта – от подготовки данных до эксплуатации модели. Во-вторых, архитектура интегрирует традиционные средства информационной безопасности с современными механизмами защиты искусственного интеллекта, включая анализ пользовательских запросов, мониторинг поведения моделей и интеллектуальную оценку рисков. В-третьих, модульный принцип построения обеспечивает возможность адаптации модели к различным типам интеллектуальных систем без существенной перестройки их программной и аппаратной инфраструктуры.

Полученные результаты свидетельствуют о том, что применение многоуровневой архитектуры позволяет значительно повысить устойчивость интеллектуальных систем к современным киберугрозам и создать основу для дальнейшей автоматизации процессов управления безопасностью искусственного интеллекта.

Рисунок 1 – Многоуровневая модель обеспечения кибербезопасности систем искусственного интеллекта (MAISM)



Предлагаемая модель обеспечивает комплексную защиту интеллектуальной системы благодаря последовательному взаимодействию всех уровней. Информация о выявленных инцидентах, аномалиях и уязвимостях передается между уровнями в режиме реального времени, что позволяет оперативно реагировать на изменение уровня угроз и адаптировать механизмы защиты к текущей обстановке.

По сравнению с существующими архитектурами предложенная модель охватывает весь жизненный цикл системы искусственного интеллекта, интегрирует традиционные и специализированные механизмы защиты и обеспечивает возможность адаптации к различным типам интеллектуальных систем.

Метод интеллектуальной оценки рисков в системе искусственного интеллекта

Эффективность обеспечения кибербезопасности систем искусственного интеллекта определяется способностью своевременно выявлять угрозы, оценивать их потенциальное влияние и принимать необходимые меры по снижению уровня риска. В связи с этим в рамках разработанной модели предлагается использовать риск-ориентированный подход, позволяющий проводить комплексную оценку защищенности системы искусственного интеллекта на всех этапах ее жизненного цикла.

В отличие от традиционных информационных систем, где объектами защиты являются преимущественно вычислительная инфраструктура, сети и базы данных, системы искусственного интеллекта требуют оценки безопасности значительно большего числа компонентов. К ним относятся обучающие данные, модели машинного обучения, пользовательские запросы, вычислительная инфраструктура, внешние сервисы, программные интерфейсы, а также процессы мониторинга и управления.

Предлагаемый метод интеллектуальной оценки рисков основан на последовательном анализе всех компонентов системы искусственного интеллекта и определении уровня их защищенности. Оценка осуществляется с учетом вероятности возникновения угроз, возможных последствий их реализации и эффективности применяемых механизмов защиты. Такой подход позволяет не только определить наиболее уязвимые элементы системы, но и установить приоритетность мероприятий по повышению уровня кибербезопасности.

Процесс оценки включает несколько последовательных этапов.

На первом этапе выполняется идентификация объектов защиты. Определяются компоненты интеллектуальной системы, требующие обеспечения безопасности: обучающие данные, модели искусственного интеллекта, программное обеспечение, вычислительная инфраструктура, пользовательские интерфейсы и средства мониторинга.

На втором этапе осуществляется идентификация потенциальных угроз. Для каждого объекта определяются наиболее вероятные сценарии атак с использованием рекомендаций MITRE ATLAS, OWASP Top 10 for Large Language Model Applications и других современных методик анализа угроз.

Следующим этапом является оценка уровня защищенности каждого компонента. Анализируются существующие средства защиты, оценивается их эффективность, а также выявляются возможные уязвимости, которые могут быть использованы нарушителем.

После определения уязвимостей производится комплексная оценка риска. При этом учитываются вероятность реализации угроз, возможные последствия для функционирования системы искусственного интеллекта и потенциальный ущерб для организации. Полученные результаты позволяют определить наиболее критичные направления совершенствования системы защиты.

Заключительным этапом является формирование рекомендаций по снижению уровня риска. В зависимости от результатов оценки могут быть предложены дополнительные меры защиты, включающие совершенствование политики информационной безопасности, внедрение средств фильтрации пользовательских запросов, усиление контроля обучающих данных, применение технологий Zero Trust, расширение возможностей мониторинга безопасности и проведение регулярного тестирования моделей искусственного интеллекта.

Использование риск-ориентированного подхода обеспечивает непрерывный контроль состояния защищенности интеллектуальной системы и позволяет своевременно адаптировать механизмы защиты к изменению внешней киберугрозовой обстановки.

Таблица 1 – Классификация основных угроз систем искусственного интеллекта

Вид угрозы	**Объект воздействия**	**Возможные последствия**	**Основные меры защиты**
Data Poisoning	Обучающие данные	Искажение результатов обучения модели	Проверка источников данных, очистка обучающих выборок
Model Extraction	Модель искусственного интеллекта	Кража модели и интеллектуальной собственности	Контроль доступа, защита программного кода, шифрование
Prompt Injection	Пользовательские запросы	Изменение поведения модели, обход ограничений	Фильтрация запросов, анализ контекста, контроль системных инструкций

Jailbreak	Большие языковые модели	Получение запрещённых ответов	Ограничение доступа, мониторинг запросов
Membership Inference	Обучающие данные	Раскрытие конфиденциальной информации	Анонимизация данных, ограничение доступа
Adversarial Attack	Алгоритм искусственного интеллекта	Ошибочные результаты классификации	Регулярное тестирование модели, использование защитных алгоритмов
Атаки на API	Программные интерфейсы	Несанкционированный доступ к системе	Аутентификация пользователей, защита API, журналирование
Data Leakage	Базы знаний	Утечка служебной информации	Шифрование данных, контроль прав доступа

Анализ представленных угроз показывает, что обеспечение кибербезопасности систем искусственного интеллекта требует применения комплексного подхода, объединяющего организационные, программные и технические меры защиты. Предложенная модель позволяет учитывать взаимосвязь между различными компонентами интеллектуальной системы и своевременно реагировать на изменение уровня киберугроз.

Практическое применение разработанной модели

Предлагаемая многоуровневая модель может применяться при проектировании и эксплуатации информационных систем, использующих технологии искусственного интеллекта. Наиболее эффективным является ее использование в государственных информационных системах, объектах критической информационной инфраструктуры, финансовом секторе, медицинских организациях, а также в образовательных и военных учреждениях.

При внедрении модели обеспечивается комплексная защита всех компонентов AI-системы: обучающих данных, моделей искусственного интеллекта, пользовательского взаимодействия, вычислительной инфраструктуры и процессов управления безопасностью. Использование риск-ориентированного подхода позволяет своевременно выявлять уязвимости, оценивать уровень угроз и принимать меры по предотвращению киберинцидентов.

Внедрение предложенной модели способствует повышению устойчивости интеллектуальных систем к современным кибератакам, снижению вероятности компрометации данных и повышению надежности принимаемых системой решений.

Обсуждение результатов. Проведенное исследование показало, что существующие подходы к обеспечению безопасности искусственного интеллекта преимущественно направлены на защиту отдельных компонентов системы. Вместе с тем современные угрозы требуют комплексного подхода, учитывающего особенности всего жизненного цикла интеллектуальной системы.

Предлагаемая модель отличается интеграцией механизмов защиты данных, моделей, инфраструктуры, пользовательского взаимодействия и процессов управления рисками. В отличие от существующих решений она обеспечивает непрерывный мониторинг состояния безопасности и позволяет адаптировать механизмы защиты в зависимости от изменения уровня киберугроз.

Практическая ценность модели заключается в возможности ее интеграции с существующими средствами информационной безопасности организации, включая системы мониторинга событий безопасности (SIEM), архитектуру Zero Trust и корпоративные политики управления информационной безопасностью.

Заключение. В статье рассмотрены современные угрозы кибербезопасности систем искусственного интеллекта и проведен анализ существующих подходов к их защите. Установлено, что применение традиционных средств информационной безопасности не обеспечивает достаточного уровня защиты интеллектуальных систем вследствие появления новых видов атак, направленных на модели искусственного интеллекта и обучающие данные.

Разработана многоуровневая модель обеспечения кибербезопасности систем искусственного интеллекта, включающая уровни защиты данных, моделей, пользовательского взаимодействия, вычислительной инфраструктуры, мониторинга безопасности и управления рисками. Предложенная модель обеспечивает комплексную защиту AI-систем на протяжении всего жизненного цикла и может быть использована при создании защищенных информационных систем различного назначения.

Практическая значимость исследования заключается в возможности применения разработанной модели при проектировании и модернизации интеллектуальных информационных систем в государственных, корпоративных и военных организациях. Перспективным направлением дальнейших исследований является разработка программного средства автоматизированной оценки уровня кибербезопасности систем искусственного интеллекта и экспериментальная проверка эффективности предложенной модели в реальных условиях эксплуатации.

СПИСОК ЛИТЕРАТУРЫ

1 National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework (AI RMF 1.0). – Gaithersburg, MD: NIST, 2023.

2 International Organization for Standardization. ISO/IEC 23894:2023 Information technology – Artificial intelligence – Risk management. – Geneva: ISO, 2023.

3 MITRE Corporation. MITRE ATLAS (Adversarial Threat Landscape for Artificial Intelligence Systems). – McLean, VA, 2024.

4 OWASP Foundation. OWASP Top 10 for Large Language Model Applications. – Version 1.1. – 2024.

5 Частикова В.А., Гуляй В.Г., Жерлицын С.А. Технологии искусственного интеллекта в кибербезопасности: учебник для вузов. – Москва: Горячая линия – Телеком, 2024. – 508 с.