К ВОПРОСУ ПРИМЕНЕНИЯ В СРЕДСТВАХ ОБЕСПЕЧЕНИЯ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ ЭЛЕМЕНТОВ ДОВЕРЕННОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

***М.М. Добрышин**, к.т.н., Академия ФСО России, dobrithin@ya.ru.*

УДК 004.942

Аннотация. Совершенствование средств и способов реализации компьютерных атак требуют от разработчиков средств защиты применения самых современных технологий. Одним из направлений повышения эффективности современных средств обеспечения информационной безопасности является внедрение систем, использующих различные технологии искусственного интеллекта. В статье затронуты вопросы регламентирования разработки систем с искусственным интеллектом, в том числе доверенного искусственного интеллекта. Проведен анализ широкого спектра известных технологий, их применимости для решения различных задач в области информационной безопасности, который включает краткое описание типовых задач обеспечения безопасности с указанием применимости конкретного метода для решения указанных задач. Определены потенциальные угрозы безопасности непосредственно для систем с искусственным интеллектом, учитывающие угрозы, реализуемые за счет уязвимостей сред разработки программного обеспечения, процессов обучения и выполнения функциональных задач. Предложены мероприятия, позволяющие минимизировать риски сформулированных угроз.

Ключевые слова: доверенный искусственный интеллект; компьютерные атаки; угрозы безопасности; нарушитель информационной безопасности; средства обеспечения информационной безопасности.

ON THE ISSUE OF USING ELEMENTS OF TRUSTED ARTIFICIAL INTELLIGENCE IN INFORMATION SECURITY TOOLS

***M.M. Dobryshin**, Candidate of Technical Science, Academy of the Federal Tax Service of Russia.*

Annotation. Improving the means and methods of implementing computer attacks require the developers of security tools to use the most modern technologies. One of the ways to increase the effectiveness of modern information security tools is the introduction of systems using various artificial intelligence technologies. The article touches upon the issues of regulating the development of artificial intelligence systems, including trusted artificial intelligence. An analysis of a wide range of well-known technologies and their applicability for solving various tasks in the field of information security is carried out, which includes a brief description of typical security tasks indicating the applicability of a specific method for solving these tasks. Potential security threats directly for artificial intelligence systems have been identified, taking into account the threats posed by vulnerabilities in software development environments, the learning process, and the process of performing functional tasks. A group of measures is proposed to minimize the risks of formulated threats.

Keywords: trusted artificial intelligence; computer attacks; security threats; information security violator; information security tools.

Введение

Необходимость применения элементов «искусственного интеллекта» (ИИ) в современных технических системах обуславливается не трендовыми веяниями в обществе, а обоснованной необходимостью. Необходимость определяется текущим этапом развития систем, а именно тем, что решаемые задачи требуют применения адаптивных систем, которые обрабатывают значительные объемы данных для принятия адекватных решений. Ключевой особенностью является необходимость принятия решения в течение малого времени на основании существенного объема данных. В качестве примера скорости принятия решения можно выбрать реакцию роботизированного транспортного средства на изменение дорожной обстановки, обслуживание роботизированных касс самообслуживания (распознающих покупки и покупателей) или системы контроля доступа и видеонаблюдения. Инструменты искусственного интеллекта активно и, самое главное, крайне эффективно применяются в финансовой сфере и позволяют на основе динамики изменения множества параметров прогнозировать изменения курсов ценных бумаг [1].

Вместе с этим ряд отраслей в силу консерватизма, непроработанности нормативной базы, а самое главное – недоверия к самим решениям, только начинают внедрять ИИ для решения актуальных задач. Одной из таких отраслей является информационная безопасность (ИБ). Отрицать наличие и применение решений, использующих ИИ, конечно, нельзя, так программные продукты «Лаборатории Касперского», *Positive Technologies* и ряд других компаний используют элементы ИИ в средствах обеспечения информационной безопасности (СрОИБ). В отдельных решениях ИИ является вспомогательным инструментом, в отдельных ИИ применяется как самостоятельно действующая система [2, 3].

Консерватизм применения ИИ в ИБ существует из-за того, что в настоящее время применение ИИ с правовой точки зрения определено недостаточно конкретно, а понятие «доверенный искусственный интеллект» в настоящий момент не регламентировано вовсе. Задачей регламентирования всех стадий жизненного цикла ИИ в международном пространстве занимается объединенный технический комитет *SC* 42 *Artificial Intelligence* (*ISO/IEC JTC* 1 *Information Technologies*), который объединяет постоянно действующие группы – *WG*1-*WG*5 и консультационные группы *AG*1, *AG*3. Особое место в этих группах, исходя из рассматриваемой предметной области, следует уделить подкомитету *ISO/IEC JTC* 1/*SC* 27, рассматривающему вопросы кибербезопасности и защиты частной жизни, группе *WG*-5.

В России в качестве базового документа используется Концепция технологического развития на период до 2030 г., утвержденная распоряжением Правительства Российской Федерации № 1315-р от 20 мая 2023 г. Разработкой регламентирующих документов в области применения ИИ занимается Технический комитет по стандартизации № 164 «Искусственный интеллект» (ТК 164).

На основе совместной работы отечественных и международных профильных комитетов сформирован начальный набор стандартов. Эти документы не охватывают всю предметную область, однако работа по дополнению и расширению функционала документов сейчас проводится[1][2][3][4].

[1] ISO/IEC TR 24027 «Информационные технологии. Искусственный интеллект (AI)» Предвзятость в системах искусственного интеллекта и принятии решений с помощью искусственного интеллекта.

[2] ISO/IEC PDTR 24028 «Информационные технологии – искусственный интеллект. Обзор надежности в области искусственного интеллекта».

Одним из направлений является формирование требований и условий применения доверенного ИИ.

Постановка задачи

Однако консерватизм применения ИИ в ИБ вызван не только, а, скорее, не столько непроработанностью нормативной базы, а наличием технических и научных задач, в качестве некоторых из них можно выделить следующие [2-6]:

- модели ИИ выступают в качестве «черных ящиков», т. е. разработчик потенциально знает, что он может получить на выходе при определенном наборе данных на входе, но как происходит принятие решения, он не в полной мере понимает (с детализацией каждого этапа расчетов), помимо этого, разработчик не в полной мере способен повлиять на определение (задание) критериев принятия решения;
- значительное влияние на правильность функционирования СрОИБ с ИИ влияет обучающая выборка, т.е. от объема, полноты и обоснованности сформированной выборки исходных данных в значительной мере зависит правильность обучения и функционирования СрОИБ;
- сложности с определением уязвимостей и потенциально возможных техник реализации компьютерных атак в отношении самой СрОИБ с ИИ;
- СрОИБ с ИИ, реализующие функцию непрерывного обучения в ходе эксплуатации, требуют дополнительного инструмента контроля анализируемых данных, а для систем с доверенным ИИ возникает угроза злонамеренного переобучения в интересах нарушителя ИБ;
- рассматривая СрОИБ как сложное программное средство и, следовательно, потенциальный объект воздействия нарушителя ИБ, возникает необходимость контроля правильности принятия решений и правильности поведения данного средства.

Обзор методов машинного обучения в контексте решения задач обеспечения информационной безопасности

Указанные задачи с развитием ИИ (в том числе доверенного ИИ) в перспективе будут устранены (с накоплением опыта применения), а снижение потенциальной опасности на текущем этапе, возможно, обеспечить за счет обоснованного выбора моделей ИИ и порядка их применения. На основе анализа известных работ [7-10] представлен вариант применимости методов, алгоритмов и технологий ИИ:

1. Машинное обучение *(Machine Learning – ML)*

1.1. Контролируемое обучение (*Supervised Learning*) – обучение на размеченных данных, где для каждого примера известен правильный ответ (метка класса).

Возможная область применения. Обнаружение вредоносного ПО – классификация файлов как вредоносных или безопасных на основе признаков (например, сигнатур, структуры файла, поведения). Выявление фишинга – определение фишинговых писем на основе анализа содержания, отправителя, ссылок и других параметров. Анализ журналов событий – выявление аномалий в журналах на основе исторических данных. Выявление фактора реализации атак на основе изменения значений параметров защищаемого ресурса.

[3] ISO/IEC TR 24029-1 «Искусственный интеллект (ИИ). Оценка надежности нейронных сетей». Часть 1.
[4] ISO/IEC 23894 – Информационные технологии. Искусственный интеллект. Управление рисками.

Применяемые методы, алгоритмы и технологии: линейная регрессия (*Linear Regression*); логистическая регрессия (*Logistic Regression*); деревья решений (*Decision Trees*); случайный лес (*Random Forest*); метод опорных векторов (*Support Vector Machines – SVM*); наивный байесовский классификатор (*Naive Bayes Classifier*).

1.2. Обучение без учителя (*Unsupervised Learning*) – обучение на неразмеченных данных, цель которого – найти скрытые закономерности, структуру или аномалии.

Возможная область применения. Обнаружение аномалий: выявление необычного поведения пользователей, сетевого трафика или системных процессов. Кластеризация: группировка схожих событий или объектов для упрощения анализа. Уменьшение размерности данных: сокращение количества признаков для ускорения обработки и визуализации [11].

Применяемые методы, алгоритмы и технологии: метод *k*-средних (*k-Means Clustering*); иерархическая кластеризация (*Hierarchical Clustering*); алгоритм *DBSCAN* (пространственная кластеризация приложений на основе плотности с шумом); метод главных компонент (*Principal Component Analysis – PCA*); алгоритм обнаружения выбросов (*Isolation Forest*).

1.3. Обучение с подкреплением (*Reinforcement Learning – RL*) – обучение агента принимать решения в среде для максимизации вознаграждения.

Возможная область применения. Автоматическое реагирование на инциденты: оптимизация стратегий реагирования на атаки в режиме реального времени. Анализ уязвимостей: автоматическое выявление и использование уязвимостей для тестирования систем. Управление сетевой безопасностью: оптимизация настроек сетевых экранов и систем обнаружения вторжений [12].

Применяемые методы, алгоритмы и технологии: *Q*-обучение (*Q-Learning*); алгоритм *SARSA* (*State-Action-Reward-State-Action*); метод глубокого *Q*-обучения (*Deep Q-Network – DQN*); алгоритмы на основе политики *(Policy Gradient Methods*), генетические алгоритмы.

2. Глубокое обучение *(Deep Learning – DL)* – использование многослойных нейронных сетей для извлечения сложных признаков из данных [13].

Возможная область применения. Обнаружение вредоносного ПО – анализ двоичного кода, структуры файлов и поведения программ с использованием сверточных нейронных сетей (*Convolutional Neural Networks – CNN*) и рекуррентных нейронных сетей (*Recurrent Neural Networks – RNN*). Анализ сетевого трафика: обнаружение аномалий и атак с использованием *RNN* и долгой краткосрочной памяти (*Long Short-Term Memory – LSTM*) для анализа временных последовательностей. Обработка естественного языка (*Natural Language Processing – NLP*): анализ текста фишинговых писем и документов с использованием трансформеров (*Transformers*) и *BERT* для выявления угроз. Распознавание образов: идентификация вредоносных изображений и видео с помощью *CNN*.

Типы нейронных сетей: сверточные нейронные сети (*CNN*); рекуррентные нейронные сети (*RNN*); долгая краткосрочная память (*LSTM*); трансформеры (*Transformers*); генеративно-состязательные сети (*Generative Adversarial Networks – GAN*).

3. Анализ поведения *(Behavioral Analysis)* –мониторинг и анализ действий пользователей, процессов и сетевого трафика для выявления аномалий и подозрительного поведения.

Применение. Обнаружение инсайдерских угроз – выявление необычного поведения сотрудников, указывающего на злонамеренные действия. Выявление компрометации учетных записей – обнаружение аномальной активности учетных записей (например, вход с необычного *IP*-адреса, доступ к нетипичным ресурсам). Анализ сетевого трафика – выявление подозрительных сетевых подключений и передачи данных.

Технологии: статистический анализ (*Statistical Analysis*); машинное обучение (*ML*); анализ временных рядов (*Time Series Analysis*); сетевой анализ поведения (*Network Behavior Analysis - NBA*); анализ поведения пользователей и сущностей (*User and Entity Behavior Analytics – UEBA*).

4. Обработка естественного языка *(Natural Language Processing – NLP)* – использование методов ИИ для анализа и понимания текста.

Применение. Анализ фишинговых писем – выявление фишинговых писем на основе анализа текста, заголовков и вложений. Анализ отчетов об инцидентах – автоматическое извлечение информации из отчетов об инцидентах [14]. Анализ угроз в социальных сетях – выявление упоминаний о компании и ее продуктах, которые могут указывать на угрозы.

Техники: токенизация (*Tokenization*); частеречная разметка (*Part-of-Speech Tagging*); определение именованных сущностей (*Named Entity Recognition – NER*); анализ тональности (*Sentiment Analysis*); классификация текста (*Text Classification*); использование языковых моделей (*Language Models*).

5. Автоматизация и управление безопасности *(Security Automation and Orchestration – SOAR)* – автоматизация рутинных задач по обеспечению безопасности и управление различных инструментов безопасности для ускорения реагирования на инциденты.

Применение. Автоматическое реагирование на инциденты – автоматическое блокирование вредоносных *IP*-адресов, изоляция скомпрометированных хостов и выполнение других действий по реагированию. Автоматизация анализа журналов событий – автоматический сбор, анализ и корреляция журналов событий из различных источников [15]. Автоматизация управления уязвимостями – автоматическое сканирование на уязвимости, приоритизация и исправление.

Технологии: скрипты и автоматизация (*Scripts and Automation*); интеграция с другими инструментами безопасности (*Integration with Security Tools*); управление рабочими процессами (*Workflow Management*); индикаторы компрометации (*Indicators of Compromise – IOC*).

6. Другие технологии:

Анализ графов (*Graph Analysis*) – использование графов для визуализации и анализа связей между объектами (например, пользователями, устройствами, *IP*-адресами) для выявления атак.

Имитационное моделирование (*Simulation Modeling*) – использование моделей для имитации атак и тестирования эффективности мер безопасности.

Квантовые вычисления (*Quantum Computing*) – исследование применения квантовых алгоритмов для решения задач криптографии и анализа безопасности.

Некоторые угрозы ИБ для средств обеспечения ИБ, использующих искусственный интеллект

Средства обеспечения информационной безопасности с искусственным интеллектом, несмотря на свои преимущества, подвержены ряду угроз:

1. Атакующие примеры (*Adversarial Examples*) – специально разработанные входные данные, которые вводят в заблуждение модели машинного обучения, заставляя их выдавать неправильные результаты. Нарушитель ИБ вносит небольшие, практически незаметные изменения во входные данные (например, в изображение, звук или код), которые приводят к тому, что модель ИИ классифицирует их неправильно. Данные действия способствуют обходу систем обнаружения ВПО, систем распознавания лиц, систем обнаружения фишинговых атак.

2. Отравление данных (*Data Poisoning*) – злонамеренное добавление вредоносных данных в обучающий набор данных, используемый для обучения модели ИИ. Злоумышленник внедряет данные, которые искажают процесс обучения, в результате чего модель начинает выдавать неверные результаты или становится более уязвимой к определенным атакам. Такое воздействие снижает точность обнаружения угроз, создает возможности для обхода защиты, ухудшение производительности модели.

3. Уклонение от модели (*Model Evasion*) – манипулирование входными данными таким образом, чтобы обойти обученную модель ИИ, не изменяя ее саму. Нарушитель ИБ изучает особенности модели ИИ и разрабатывает входные данные, которые не вызывают срабатывания системы обнаружения и потенциально повышает риски ИБ.

4. Эксфильтрация модели (*Model Extraction*) – извлечение информации о структуре и параметрах обученной модели ИИ без прямого доступа к ней. Злоумышленник отправляет запросы к модели и анализирует ответы, чтобы реконструировать ее внутреннюю структуру и алгоритмы. Данный метод анализа систем позволяет собрать набор исходных данных для разработки более эффективных атак, обхода средств защиты, копирования и использования модели в злонамеренных целях.

5. Атаки на конфиденциальность *(Privacy Attacks*) – извлечение конфиденциальной информации из обучающих данных или из самой модели ИИ. Используя атаки на основе членства (*Membership Inference Attacks*) или дифференциальная конфиденциальность (*Differential Privacy Attacks*), нарушитель выявляет информацию о конкретных объектах в обучающем наборе данных, что приводит к раскрытию персональных данных.

6. Атаки на доступность (*Availability Attacks*) – нарушение доступности систем ИИ, используемых в целях безопасности (*DoS/DDoS*-атаки, направленные на перегрузку системы запросами, приводящие к отказу в обслуживании). Атаки на ресурсы – истощение вычислительных ресурсов, необходимых для работы модели ИИ, приводящие к отключению систем защиты, создание «слепых зон» в системе безопасности.

7. Обратные атаки (*Replay Attacks*) – повторное использование ранее записанных взаимодействий с моделью ИИ. Злоумышленник записывает запросы и ответы, а затем повторяет их, чтобы обойти защиту или получить несанкционированный доступ к модели ИИ.

8. Специфические уязвимости алгоритмов ИИ – уязвимости, связанные с особенностями конкретных алгоритмов машинного обучения и нейронных сетей. В качестве примера возможно рассмотреть уязвимости в алгоритмах кластеризации, приводящие к неправильной классификации аномалий;

уязвимости в алгоритмах обучения с подкреплением, позволяющие манипулировать процессом обучения.

9. Перенос атак между моделями (*Transferability of Attacks*) – возможность использования атак, разработанных для одной модели ИИ, против других моделей. Атакующие примеры, созданные для одной модели, могут успешно обходить защиту других моделей, даже если они обучены на разных данных или используют разные архитектуры.

Для минимизации указанных угроз возможно применять следующие меры защиты систем с искусственным интеллектом:

- использование методов защиты от атакующих примеров – состязательное обучение, защитная дистилляция;
- проверка и очистка обучающих данных – обеспечение качества и надежности обучающих данных;
- мониторинг поведения модели – выявление аномального поведения и попыток обойти защиту;
- регулярное переобучение модели – обновление модели с использованием новых данных для адаптации к меняющимся угрозам;
- разработка объяснимых моделей ИИ (*Explainable AI – XAI*) – создание моделей, решения которых можно понять и проверить;
- ограничение доступа к моделям ИИ – защита моделей от несанкционированного доступа и эксфильтрации;
- использование техник дифференциальной конфиденциальности – защита конфиденциальности данных при обучении и использовании моделей ИИ.

Заключение

Технологии ИИ играют всё более важную роль в обеспечении информационной безопасности, позволяя автоматизировать задачи обнаружения, анализа и реагирования на угрозы. Постоянное развитие и совершенствование этих технологий позволяют организациям более эффективно защищать свои информационные активы в условиях постоянно меняющегося ландшафта угроз. Текущий этап развития методов позволяет применять ИИ для разрешения различных задач. Применимость ИИ подтверждается эффективностью известных решений, разработанных ведущими компаниями в данной области. В качестве активно применяемых и достаточно эффективных решений сейчас на рынке ИБ присутствуют:

- *Kaspersky Anti APT* – решение для обнаружения и предотвращения целевых атак (*Advanced Persistent Threat, APT*). За счет применения ИИ и продвинутых методов анализа трафика и поведения системы сервис выявляет угрозы, характерные для сложных и продолжительных кибератак.
- *PT Sandbox* использует комплексный анализ системных процессов для высокоточного обнаружения неизвестных угроз. Данное решение реализует функции защиты от масштабных кибератак с применением современных вредоносных программ. Включено в Единый реестр российских программ как продукт, использующий ИИ.
- *Darktrace* осуществляет непрерывный мониторинг сетевой активности и поведения пользователей для выявления аномалий и потенциальных атак. Предоставляет автоматизированные инструменты для реагирования на угрозы и предотвращения утечек данных.

Проведенный анализ возможностей применения ИИ при решении задач обеспечения информационной безопасности показывает, что, помимо традиционных видов деятельности ИИ – выполнение рутинных операций, ИИ в средствах защиты информации, способен выявлять и противодействовать новым, ранее неизвестным атакам. Эффективность известных решений по противодействию новым атакам безусловно превышает традиционные сигнатурные и поведенческие методы, что позволяет утверждать о целесообразности применения ИИ в средствах и системах обеспечения ИБ.

Вместе с тем, применение ИИ порождает новые научные задачи, а именно возникает необходимость определения степени доверия к системам с ИИ, а также решения задач по защите систем с ИИ от новых атак на процесс обучения и функционирования этих систем.

Текущий этап развития теории и практики применения инструментов ИИ и систем с ИИ, безусловно, требует изучения. Получаемые системами с ИИ решения сложных задач, безусловно, требуют если не постоянного, то периодического контроля. Однако, применение систем защиты с ИИ, даже с учетом выявленных недостатков, позволяет повысить защищенность активов компаний и корпораций.

Таким образом, применение средств и систем защиты информации и систем обеспечения ИБ со встроенными модулями ИИ позволяют повысить защищенность активов, но требуют наложения на эти системы определенных ограничений по функциональным возможностям и продолжения исследований в предметной области.

Литература

1. Coinspot: электрон. ресурс. Mirocana – нейронная сеть для торговли на крипторынке. 2017 URL: https://coinspot.io/technology/mirocana-nejronnaya-set-dlya-torgovli-na-kriptorynke/ (дата обращения 24.07.2025).
2. Positive Technologies. О машинном обучении с точки зрения ИБ: реальная обстановка / https://habr.com/ru/companies/pt/profile/ (дата обращения 24.03.2025).
3. Солдатов С. ИИ против рутины в ИБ / www.kaspersky.ru/blog/ai-role-in-cybersecurity-automation/38412/ (дата обращения 24.07.2025).
4. Намиот Д.Е., Ильюшин Е.А., Чижов И.В. Искусственный интеллект и кибербезопасность // International Journal of Open Information Technologies ISSN: 2307-8162, 2022. – V. 10. – № 9. – С. 135-144.
5. Гайфулина Д.А., Котенко И.В. Применение методов глубокого обучения в задачах кибербезопасности // Вопросы кибербезопасности. 2020. – № 3 (38). – С. 76-86.
6. Гайфулина Д.А., Котенко И.В. Применение методов глубокого обучения в задачах кибербезопасности. Часть 2 // Вопросы кибербезопасности, 2020. – № 4 (38). – С. 11-18.
7. Колодка А.А., Метелкин П.В., Леоненко В. В. Искусственный интеллект: защитить и защититься // Интеллектуальные транспортные системы России. Информационно-аналитический портал. https://www.itsjournal.ru/ (дата обращения 24.03.2025).
8. Исаков А.А. Искусственный интеллект и расследование киберпреступлений // Вестник науки, – № 5 (62). – Т. 3 – С. 597-602.
9. Рассел, Стюарт, Норвиг, Питер. Искусственный интеллект: современный подход, 4-е изд., том 1. Решение проблем: знания и рассуждения: Пер. с англ. – СПб.: ООО «Диалектика», 2021. – 704 с.

10. Рассел, Стюарт, Норвиг, Питер. Искусственный интеллект: современный подход, 4-е издание, том 3. Обучение, восприятие и действие: Пер. с англ. – СПб.: ООО «Диалектика», 2022. – 640 с. - Парал. тит. англ.
11. Добрышин М.М., Шугуров Д.Е., Слесарева К.М., Юркин Д.А. Методика сокращения признакового пространства систем обнаружения атак на основе применения генетического алгоритма // Известия Тульского государственного университета. Технические науки, 2023. – № 11. – С. 110-113.
12. Добрышин М.М., Левичева Ю.А., Реформат А.Н. Предложения по повышению эффективности выбора стратегий защиты корпоративной сети связи // В сборнике: Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО 2022). Сборник научных статей XI Международной научно-технической и научно-методической конференции. Санкт-Петербург, – 2022. – С. 199-202.
13. Белов А.С., Добрышин М.М., Борзова Н.Ю. Формирование модели угроз информационной безопасности на среднесрочный период // Приборы и системы. Управление, контроль, диагностика, 2021. – № 7. – С. 41-48.
14. Добрышин М.М., Шугуров Д.Е., Кирикова Ю.А., Погодин Н.В. Программа автоматического обновления и формирования техник реализации компьютерных атак для системы обеспечения информационной безопасности / Свидетельство о государственной регистрации программы для ЭВМ № 2023688299 от 14.12.2023 Бюл. № 1.
15. Добрышин М.М., Свечников Д.А., Маркин Д.О., Шугуров Д.Е., Горохов Д.Е., Уваров Д.С., Попыванов А.И. Программный модуль обработки событий мониторинга информационной безопасности / Свидетельство о регистрации программы для ЭВМ RU 2022610076, 10.01.2022.