УДК 004.056.5 URL: https://ptsj.bmstu.ru/catalog/icec/insec/1064.html

ОБЗОР АСПЕКТОВ ПРИМЕНЕНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В КИБЕРБЕЗОПАСНОСТИ

В.С. Аманатиди amanatidivs@student.bmstu.ru
Е.В. Глинская glinskaya@bmstu.ru
SPIN-код: 5430-3023

МГТУ им. Н.Э. Баумана, Москва, Российская Федерация

В настоящее время можно наблюдать процесс интеграции искусственного интеллекта (ИИ) во все сферы информационных технологий. Кибербезопасность, конечно, не стала исключением: организации используют искусственный интеллект для усиления своей защиты, но в то же время атакующие на его основе создают кибератаки нового поколения, которые становятся все более устойчивыми к традиционным системам безопасности, основанным на эвристике и сигнатурах. Этот факт говорит о необходимости появления и повсеместного внедрения адаптивной защиты и упреждающих мер для устранения потенциальных последствий. В статье рассмотрены и проанализированы типы ИИ в кибербезопасности, их цели, представлена методология разработки систем защиты на основе ИИ.

Ключевые слова: искусственный интеллект, машинное обучение, системы киберзащиты на основе искусственного интеллекта, состязательный искусственный интеллект, методология разработки, обучение с подкреплением, уязвимости

Введение. Кибератаки стали постоянной угрозой в меняющемся ландшафте цифровых взаимодействий. На протяжении десятилетий эти атаки развивались, становясь все более изощренными и массивными. Внедрение возможностей искусственного интеллекта (ИИ) стало поворотным моментом в этой эволюции.

Искусственный интеллект является мощным инструментом, который может быть использован как в оборонительных, так и в наступательных целях в области кибербезопасности. В данной статье принята следующая классификация: «оборонительный» ИИ (Defensive AI), «наступательный» ИИ (Offensive AI), состязательный ИИ (Adversarial AI) (см. таблицу). При этом состязательный ИИ рассматривается как подкатегория «наступательного» ИИ [1].

Искусственный интеллект в системах защиты. В настоящее время традиционные меры кибербезопасности, основанные на эвристике и сигнатурах, становятся все более неэффективными против атак под управлением ИИ. Причинами этого служат динамический характер, устойчивость, сложность, скорость и высокая вариативность таких атак.

 ISSN 2541-8009

Основные различия между типами ИИ

Тип ИИ в кибербезопасности	Цель применения	Примеры
«Оборонительный»	Использование методов ИИ для защиты компьютерный систем и сетей от атак	Антивредоносные программы (Anti-malware); системы обнаружения вторжений (IDS)
«Наступательный»	Использование методов ИИ для атак на компьютерные системы и сети	Автоматизация эксплуатации существующих уязвимостей; динамическая генерация шаблонов атак, не соответствующих известным сигнатурам
Состязательный	Злонамеренная эксплуатация и/или атака систем и данных ИИ/ML	«Отравление» обучающих данных; манипулирование входными данными

Искусственный интеллект (в частности, алгоритмы машинного обучения) может быстро анализировать данные и позволяет специалистам сосредоточиться на более тонких аспектах кибербезопасности. Однако он все же не может полностью заместить человека и его опыт, что делает ИИ лишь вспомогательным инструментом, а не комплексным решением проблем кибербезопасности [2].

Эффективная система защиты на основе ИИ должна обеспечивать:

– высокую скорость обнаружения;

– применение самообучения без учителя и обучения с подкреплением (RL);

– распознавание состязательных атак и новых кибератак;

– выявление угроз на малых объемах данных;

– обучение контрмерам с использованием данных киберразведки (CTI);

– извлечение и кодирование значимых данных («отпечатков») из реальных систем;

– грамотную визуализацию данных.

Методология разработки систем киберзащиты с использованием ИИ. Методология предусматривает следующие этапы.

1. *Сбор данных*. Например, обучение обнаружению угроз осуществляется во время передачи непрерывного потока пакетов, обрабатываемых отдельной системой с минимальными задержками обработки [3].

2. *Извлечение признаков и буферных данных*. В тех случаях, когда основное внимание уделяется сетевым сеансам, значимые признаки следующие: IP источника, IP получателя, длина IP-адреса, флаги TCP, порт источника, порт назначения, протокол, IP-адреса источника и назначения ARP и переда-

ваемые данные. Процесс завершается извлечением данных из файлов PCAP во время обучения или буферизацией пакетов в режиме реального времени по мере их поступления [4].

3. *Подготовка данных*. Метаданные, связанные с угрозами, связываются с извлеченным набором данных для обучения системы распознаванию угроз. Добавляется дополнительная информация, такая как приложения и службы.

4. *Сеансы снятия «отпечатков»*. Осуществляются извлечение, кодирование и визуализация значимых данных. Визуализация может быть выполнена разными методами (кривые Гильберта, торнадо-графики, построчно) [5]. В описанном выше процессе «отпечаток» состоит из заголовка, содержащего метаданные, кодированных моделей коммуникации участвующих сетевых протоколов и передаваемых данных.

5. *Обнаружение угроз*. Реализуется обнаружение угроз с использованием «отпечатков» сетевых сеансов с динамическим самообучением и обучением с подкреплением (RL). Система управления «отпечатками» буферизирует все созданные «отпечатки», RL-модель классифицирует их. Результат классификации сохраняется в каждом «отпечатке», принимаются соответствующие меры для смягчения рисков. Только после этого «отпечаток» удаляется из буфера.

«**Наступательный» и состязательный искусственный интеллект.** Атаки на основе ИИ становятся все более изощренными и масштабными, что представляет собой серьезную проблему кибербезопасности [6]. Злоумышленники используют ИИ в целях автоматизации и усовершенствования своих инструментов.

Состязательный и «наступательный» ИИ нацелены на компьютерные системы и сети, но в то время как «наступательный» тип обычно эксплуатирует уязвимости всей системы, состязательный манипулирует и вводит в заблуждение именно модели ИИ [7]. Ниже будут рассмотрены актуальные типы атак, в приведенной классификации это атаки на основе состязательного ИИ.

Состязательные атаки на основе ИИ. Такие атаки можно подразделить на несколько типов.

1. *Атаки уклонения*. Атаки данного типа являются самыми распространенными среди состязательных атак. Атаки уклонения направлены на уже обученные модели, их задача состоит в том, чтобы обмануть модель, предоставляя ей специально созданные входные данные — состязательные примеры. Цель атаки — не влияя на саму модель или обучающие данные, заставить ее сделать ошибочное предсказание на конкретном примере.

Атакующие используют генеративные состязательные сети для создания реалистичных, трудно обнаруживаемых состязательных примеров [8]. Также часто используется RL-модель для поиска оптимальных возмущений,

 ISSN 2541-8009

которые приводят к успешной атаке: атакующая модель отправляет запросы целевой модели и, получая обратную связь, учится создавать состязательные примеры.

2. *Атаки «отравления».* В отличие от атак уклонения, данные атаки происходят во время обучения модели. Атакующий добавляет в обучающий набор данных специально разработанные «отравляющие» примеры, которые выглядят как легитимные данные, но содержат информацию, которая искажает процесс обучения модели [9]. Цель атаки — ухудшить производительность модели или заставить ее делать определенные ошибки во время эксплуатации.

Алгоритмы машинного обучения способны анализировать обучающие данные, выявляя примеры, в случае отравления которых будет нанесен наибольший ущерб процессу обучения атакуемой модели. Также ИИ может быть использован атакующими для генерации наиболее правдоподобных вредоносных примеров, способных эффективно исказить процесс обучения.

3. *Адаптивные атаки.* В адаптивных атаках применяются алгоритмы машинного обучения для адаптации к защитным механизмам, используемым целевой системой для защиты от состязательных атак.

Атакующий обучает свою модель на основе информации об оборонительных механизмах цели, модель динамически адаптируется к изменениям в защитных стратегиях, меняя свои параметры.

Вывод. В ходе исследования были проанализированы основные типы ИИ в кибербезопасности, этапы разработки систем защиты на основе ИИ и типы атак с использованием состязательного ИИ. Таким образом, организациям стоит инвестировать в системы обнаружения угроз на основе ИИ, которые могут эффективно выявлять атаки (в том числе на основе машинного обучения) и реагировать на них [10]. Также разработчикам моделей ИИ следует внедрять передовые методы обеспечения безопасности, включая тщательное тестирование на уязвимости и состязательные атаки в процессе разработки модели.

Литература

[1] Malatji M., Tolah A. Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. *AI and Ethics*, 2024. https://doi.org/10.1007/s43681-024-00427-4

[2] Sun N., Ding M., Jiang J. et al. Cyber Threat Intelligence Mining for Proactive Cybersecurity Defense: A Survey and New Perspectives. *IEEE Communications Surveys & Tutorials*, 2023, vol. 25, no. 3, p. 1748–1774. https://doi.org/10.1109/comst.2023.3273282

[3] Klopper Ch., Eloff Ja.H.P. Data Fingerprinting and Visualization for AI-Enhanced Cyber-Defence Systems. *IEEE Access*, 2024, vol. 12, p. 154054–154065. https://doi.org/10.1109/access.2024.3482728

[4] Афанасьев Н.С., Чеснов А.Е. Процесс применения машинного обучения в области безопасности киберпространства. *Актуальные научные исследования в современном мире*, 2021, № 7–2 (75), с. 84–90.

[5] Резниченко С.А., Купцова Е.С., Белалов М.Р. Использование машинного обучения в аудите информационной безопасности. *Журнал высоких гуманитарных технологий*, 2024, № 2 (5), с. 6–14.

[6] Paracha A., Arshad Ju., Farah M.B., Ismail Kh. Machine learning security and privacy: a review of threats and countermeasures. *EURASIP Journal on Information Security*, 2024, vol. 2024, no. 1. https://doi.org/10.1186/s13635-024-00158-3

[7] Mirsky Y., Demontis A., Kotak Ja. et al. The Threat of Offensive AI to Organizations. *Computers & Security*, 2023, vol. 124, art. no. 103006. https://doi.org/10.1016/j.cose.2022.103006

[8] Muheidat F., Mallouh M.A., Al-Saleh O. et al. Applying AI and Machine Learning to Enhance Automated Cybersecurity and Network Threat Identification. *Procedia Computer Science*, 2024, vol. 251, p. 287–294. https://doi.org/10.1016/j.procs.2024.11.112

[9] Грибунин В.Г., Кондаков С.Е. К вопросу о защите информации в интеллектуализированных образцах вооружения. *Вопросы кибербезопасности*, 2021, № 5 (45), с. 5–11. https://doi.org/10.21681/2311-3456-2021-5-5-11

[10] Palani K., Kethar J., Prasad S., Torremocha V. Impact of AI and Generative AI in transforming Cybersecurity. *Journal of Student Research*, 2024, vol. 13, no. 2. https://doi.org/10.47611/jsrhs.v13i2.6710

Поступила в редакцию 20.03.2025

Аманатиди Валерия Станиславовна — студентка кафедры «Информационная безопасность», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Глинская Елена Вячеславовна — старший преподаватель кафедры «Информационная безопасность», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Научный руководитель — Басараб Михаил Алексеевич, доктор физико-математических наук, заведующий кафедрой «Информационная безопасность», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Ссылку на эту статью просим оформлять следующим образом:
Аманатиди В.С., Глинская Е.В. Обзор аспектов применения искусственного интеллекта в кибербезопасности. *Политехнический молодежный журнал*, 2025, № 05 (100). URL: https://ptsj.bmstu.ru/catalog/icec/insec/1064.html

 ISSN 2541-8009

AN OVERVIEW OF DEFENSIVE AND OFFENSIVE ARTIFICIAL INTELLIGENCE IN CYBERSECURITY

V.S. Amanatidi amanatidivs@student.bmstu.ru
E.V. Glinskaya glinskaya@bmstu.ru
SPIN-code: 5430-3023

Bauman Moscow State Technical University, Moscow, Russian Federation

Currently, we can observe the process of integrating artificial intelligence (AI) into all areas of information technology. Cybersecurity, of course, is no exception: organizations use AI to strengthen their defenses, but at the same time, attackers use it to create new-generation cyberattacks that are becoming increasingly resistant to traditional security systems based on heuristics and signatures. This fact indicates the need for the emergence and widespread implementation of adaptive protection and proactive measures to eliminate potential consequences. The article considers and analyzes the types of AI in cybersecurity, their goals. It also presents a methodology for developing AI-based protection systems.

Keywords: artificial intelligence, machine learning, artificial intelligence-based cyber defense systems, competitive artificial intelligence, development methodology, reinforcement learning, vulnerabilities

Received 20.03.2025

Amanatidi V.S. — Student of the Department of Information Security, Bauman Moscow State Technical University, Moscow, Russian Federation.

Glinskaya E.V. — Senior Lecturer, Department of Information Security, Bauman Moscow State Technical University, Moscow, Russian Federation.

Scientific advisor — Basarab M.A., Dr. Phys. and Math. Sci., Head of the Department of Information Security, Bauman Moscow State Technical University, Moscow, Russian Federation.

Please cite this article in English as:
Amanatidi V.S., Glinskaya E.V. An overview of defensive and offensive artificial intelligence in cybersecurity. *Politekhnicheskiy molodezhnyy zhurnal*, 2025, no. 05 (100). (In Russ.). URL: https://ptsj.bmstu.ru/catalog/icec/insec/1064.html

