DOI 10.37882/2223-2966.2026.02-2.33

АНАЛИЗ ВОЗМОЖНОСТЕЙ ПРИМЕНЕНИЯ ИИ-АГЕНТОВ ДЛЯ ПОВЫШЕНИЯ УРОВНЯ КИБЕРУСТОЙЧИВОСТИ ИНФОРМАЦИОННОЙ ИНФРАСТРУКТУРЫ ОРГАНИЗАЦИИ

ANALYSIS OF THE POSSIBILITIES OF USING AI AGENTS TO INCREASE THE LEVEL OF CYBER RESILIENCE OF AN ORGANIZATION'S INFORMATION INFRASTRUCTURE

G. Shipulin
I. Kalutsky
A. Shornikov

Summary. The article discusses the possibilities of using AI agents to solve problems of improving the cyber resilience of modern information infrastructure, especially in monitoring and respond to information security incidents. The analysis showed a significant reduction in MTTD and MTTR time metrics when AI agents were introduced into an organisation's information security management system, but also revealed an increase in type I and type II errors in the detection and identification of information security incidents. The authors highlighted the advantages and limitations of AI agents, identified the risks associated with their use, and made a well-founded conclusion about the advisability of considering AI agents as a promising addition to a mature information security management system and processes. The analytical work led to a reasonable conclusion about the advisability of considering AI agents as a promising addition to a mature information security management system and processes for monitoring and responding to information security incidents, especially in the context of a dynamically changing cyber threat landscape.

Keywords: artificial intelligence, AI agents, information security, information security incidents, monitoring of information security incidents, information security tools.

Шипулин Георгий Фаризович
кандидат юридических наук, доцент, РТУ МИРЭА;
доцент, Московский Политехнический Университет
podumai_nad@mail.ru
Калуцкий Игорь Владимирович
кандидат технических наук, доцент,
Московский Политехнический Университет;
Доцент, РТУ МИРЭА
kalutsky_igor@mail.ru
Шорников Андрей Валерьевич
Ассистент, Московский Политехнический Университет
a.vshornikov@yandex.ru

Аннотация. В статье рассмотрены возможности применения ИИ-агентов для решения задач повышения уровня киберустойчивости современной информационной инфраструктуры, а именно при мониторинге и реагировании на инциденты информационной безопасности. Проведенный анализ показал существенное снижение значений временных метрик MTTD и MTTR при внедрении ИИ-агентов в систему управления информационной безопасности организации, однако также был выявлен рост ошибок первого и второго рода при обнаружении и идентификации инцидентов информационной безопасности. Авторами были выделены преимущества и ограничения применимости ИИ-агентов, а также определены риски, связанные с их эксплуатацией. Проведенная аналитическая работа позволила сделать обоснованный вывод о целесообразности рассмотрения ИИ-агентов в качестве перспективного дополнения зрелой системы управления информационной безопасностью и процессов мониторинга и реагирования на инциденты информационной безопасности особенно в условиях динамично изменяющегося ландшафта киберугроз.

Ключевые слова: искусственный интеллект, ИИ-агенты, информационная безопасность, инциденты информационной безопасности, мониторинг инцидентов информационной безопасности, средства защиты информации.

Введение

Актуальность темы обусловлена тем, что наиболее распространенные подходы к обеспечению информационной безопасности (ИБ), включающие в себя построение многоуровневой защиты, применение сигнатурного анализа, ручное расследование инцидентов ИБ и др., становятся менее результативным в условиях постоянно усложняющейся информационной инфраструктуры организации (включающей в себя, как правило, облачные сервисы, клиентские и серверные рабочие станции, средства защиты информации (СЗИ) и др.), а также качественной трансформации современных компьютерных атак. В контексте чего, предлагается использование подходов обеспечения ИБ с акцентом на киберустойчивость, под которой понимается способность информационной инфраструктуры (системы) предвидеть, противостоять и продолжать выполнение критически важных функций в условиях реализации угроз безопасности информации, компьютерных инцидентов или компьютерных атак, а также оперативно восстанавливаться после них и адаптироваться к изменяющейся среде угроз за счёт комплекса технических и организационных мер обнаружения, реагирования, восстановления и непрерывного совершенствования процесса обеспечения ИБ. В контексте данного исследо-

вания под ИИ-агентом понимается интеллектуальная система, использующая технологии искусственного интеллекта (ИИ) для обучения на массивах данных, принятия решений и выполнения задач без постоянного участия человека с целью достижения определенных целей [1, 2].

Целью данного исследования является анализ возможностей и оценка сопутствующих ограничений и рисков, связанных с использованием автономных ИИ-агентов в решении задач повышения уровня киберустойчивости при мониторинге и реагировании на инциденты ИБ информационной инфраструктуры организации.

Анализ применения ИИ-агентов для мониторинга и реагирования на инциденты ИБ

Применение ИИ-агентов в системах управления ИБ для повышения уровня киберустойчивости покрывает полный спектр задач, от предвидения угроз безопасности информации (а равно и компьютерных инцидентов или компьютерных атак, далее — инциденты ИБ) до восстановления и адаптации информационной инфраструктуры к инцидентам ИБ.

Однако, в рамках данной работы, область решаемых задач сокращена до мониторинга и реагирования на инциденты ИБ в связи с наличием как массивов «сырых» данных (такие как, AWS CloudTrail, UCF-Crime video frames и др.) [3], так и специализированных наборов данных для обучения моделей ИИ (такие как, NSL-KDD, CICIDS2017, UNSW-NB15, CSE-CIC-IDS2018 и др.) [4–8] и практически подтвержденных результатов для проведения сравнительного анализа показателей эффективности применения как классических СЗИ, так и с применением ИИ-агентов. В то же время, для решения других задач в рамках повышения киберустойчивости информационной инфраструктуры как подобные наборы данных, так и практически значимые результаты, отсутствуют. С точки зрения поставленных задач, ИИ-агенты демонстрируют положительные результаты, а именно значительное сокращение среднего времени обнаружения (MTTD, Mean Time to Detect) и реагирования (MTTR, Mean Time to Respond) инцидента ИБ [1]. Результаты [9] демонстрируют снижение обозначенных выше метрик в среднем на, примерно, 95 % с 48 часов и 24 часов до 3 часов и 1 часа соответственно. Менее оптимистично выглядят результаты оценки, приведенной в [2], отмечающие снижение MTTR после внедрения ИИ-агентов на 30 %. Помимо изменения временной метрики, MTTR, немаловажным является процент ошибок первого и второго типа, создаваемых ИИ-агентами в ходе мониторинга и реагирования на инциденты ИБ.

Современные системы обнаружения вторжений (IDS, Intrusion detection system) активно применяют технологии ИИ для повышения эффективности обнаружения признаков инцидентов ИБ. Особое значение приобретают мультиагентные системы, способные обрабатывать разнообразные потоки данных в реальном времени. К примеру, решение AgenticCyber, разработанное на основе генеративного ИИ, демонстрирует значительные результаты в обнаружении признаков инцидентов ИБ, достигая 96,2 % F1-метрики (представляющей собой гармоническое среднее точности (*Precision*) и полноты (*Recall*), отображенное в формуле 1) при снижении MTTR на 65 % по сравнению с классическими IDS [3].

$$F_1 = 2 \times \frac{Percision \times Recall}{Percision + Recall} \quad (1)$$

Также, исследования показывают, что интеграция механизмов внимания, более подробно описанных в [10] с мультиагентными системами, основанными на глубоком обучении с подкреплением (DQN, Deep Q-networks) обеспечивает существенное повышение качества обработки массивов данных и вычислительной эффективности моделей ИИ, в частности улучшение результатов более чем на 2 BLEU (Bilingual Evaluation Understudy) и сокращение времени обучения до 12 часов, повышает точность обнаружения признаков инцидента ИБ и снижает количество ошибок первого и второго типа примерно на 42 % [11]. Такие системы обладают повышенной устойчивостью к атакам в отношении как моделей ИИ, так и самих ИИ-агентов, а также могут быстрее масштабироваться при увеличении размера сети. Ряд экспериментов, проведенных в [12] показывает, что мультиагентные системы обеспечивают более высокую точность обнаружения аномалий и могут адаптироваться к новым типам атак без необходимости полного переобучения системы, что особенно важно в условиях постоянно эволюционирующего ландшафта киберугроз, где скорость адаптации системы защиты информации напрямую влияет на уровень киберустойчивости информационной инфраструктуры организации.

Применяемые для реагирования на инциденты ИБ средства защиты информации выходят на качественно новый уровень благодаря применению мультиагентных систем на основе больших языковых моделей. Исследования в данной сфере показывают, что различные конфигурации мультиагентных систем, как правило, выделяют следующие классы: централизованные, децентрализованные и гибридные, демонстрирующие различную эффективность в зависимости от типа инцидента ИБ.

Централизованные структуры с «единоличным» лидерством показывают наивысший процент успешного реагирования (около 70 %) в сценариях, требующих быстрого принятия решений. В то же время гибридные структуры обеспечивают более высокую адаптивность к иным типам инцидентов ИБ [13]. Анализ работ, посвященных мультиагентному обучению с подкреплением в сфере ИБ подтверждает, что гибридные архитектуры обеспечивают оптимальный баланс между скоростью

реакции и устойчивостью к частичным отказам компонентов системы [12–14]. В свою очередь, эксперименты в условиях частично автономной системы управления ИБ показывают, что децентрализованные структуры демонстрируют на 25 % [15] более высокую устойчивость к целенаправленным атакам на отдельные узлы по сравнению с централизованными аналогами, что особенно важно при защите, к примеру, критической информационной инфраструктуры.

Обобщение результатов проведенного анализа различий классических СЗИ, а также СЗИ с применением технологий ИИ и отдельно на базе ИИ-агентов, в контексте представленных в работе исследований, приводится в табл. 1. Более наглядная оценка результатов приведена на диаграмме 1.

Риски и ограничения применимости ИИ-агентов в рамках решения задач мониторинга и реагирования на инциденты ИБ

Несмотря на обозначенные выше преимущества и возможности от применения ИИ-агентов и, построенных на их базе, мультиагентных систем необходимо также отметить риски и ограничения подобных решений в рамках задач мониторинга и реагирования на инциденты ИБ.

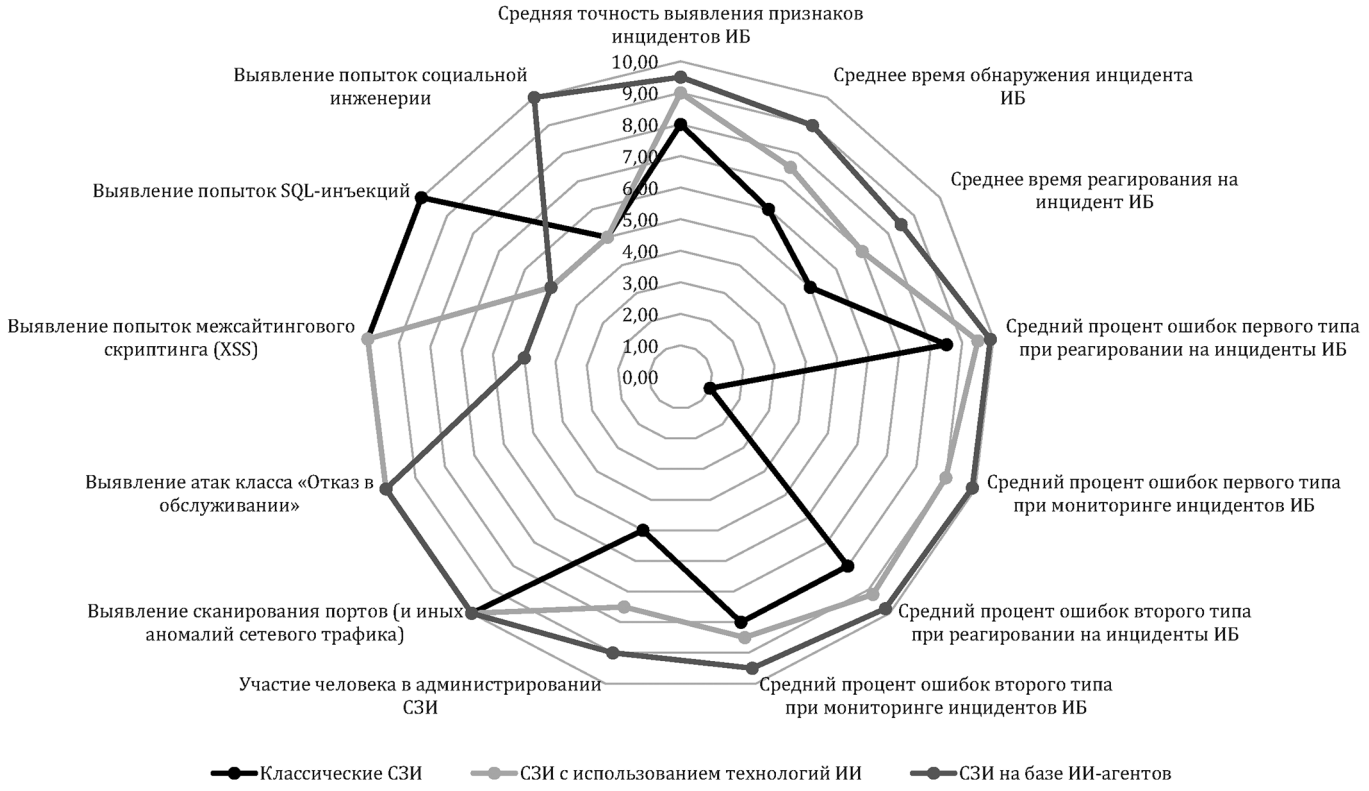
В первую очередь, важно отметить, что такая система (как на базе одного ИИ-агента, так и мультиагентная) не лишена участия человека, специалиста по защите информации или специалиста по работе с ИИ, и требует контроля за результатом. Выявленные в исследованиях

Таблица 1.

Сравнительный анализ классических СЗИ и решений на базе технологий ИИ (в т.ч. на базе ИИ-агентов)

Критерий оценки	Классические СЗИ	СЗИ с применением технологий ИИ (не считая ИИ-агентов)	СЗИ на базе ИИ-агентов
Средняя точность выявления признаков инцидентов ИБ	~80 %	~90 %	~92, в редких случаях ~95 %
Среднее время обнаружения инцидента ИБ	~15–20 минут	~5-7 минут	~3–5 минут
Среднее время реагирования на инцидент ИБ	~30 минут	~10 минут	~5 минут
Средний процент ошибок первого типа при реагировании на инциденты ИБ	~10–15 %	~5 %	<~1 %
Средний процент ошибок первого типа при мониторинге инцидентов ИБ	может достигать 90 %	5–10 %	<~1 %
Средний процент ошибок второго типа при реагировании на инциденты ИБ	~15–20 %	~5–8 %	<~2 %
Средний процент ошибок второго типа при мониторинге инцидентов ИБ	~15–20 %	~10–15 %	<~5 %
Участие человека в администрировании СЗИ	Непосредственное, выполняет полный цикл взаимодействия с СЗИ	Непосредственное, администрирует и контролирует функционирование СЗИ	Опосредованное, контролирует результаты работы СЗИ и, при необходимости, осуществляет дополнительную настройку.
Выявление сканирования портов (и иных аномалий сетевого трафика)	+	+	+
Выявление атак класса «Отказ в обслуживании»	+	+	+
Выявление попыток межсайтингового скриптинга (XSS)	+	+	±
Выявление попыток SQL-инъекций	+	±	±
Выявление попыток социальной инженерии	±	±	+

Диаграмма 1. Результаты сравнительного анализа классических СЗИ и решений с использованием технологий ИИ



[4, 9, 22] показатели ошибок первого типа могут достигать значений в диапазоне 1–5 %.

Также важно выделить возможность адаптации тактик, техник и процедур злоумышленника под особенности данных систем. Что, в частности, выражается в построении вектора атаки через негативное воздействие на модель ИИ, обучающие наборы данных, на оператора ИИ-агентов (мультиагентных систем) и др. способы обхода механизмов защиты информации, построенных на базе ИИ-агентов.

Помимо этого, часть исследователей сходится во мнении, что данная технология и построенные на её базе СЗИ на данный момент времени могут применяться для мониторинга и реагирования на инциденты ИБ в ограниченных случаях, а именно при идентификации, оценке и противодействии наиболее типовым сценариям компьютерных атак [4].

Помимо этого, ряд исследований [2, 4, 16, 17] выделяет следующие риски применения ИИ-агентов (мультиагентных систем) для решения задач мониторинга и реагирования на инциденты ИБ, наиболее острые из которых:

1. Уязвимость к атакам типа «prompt injection» (использование определенным образом подготовленных запросов к модели ИИ) и эксплойтам протоколов, позволяющим злоумышленникам манипулировать поведением агентов и заставлять их выполнять деструктивные действия, а также действия противоречащие изначальной цели работы [16].
2. Увеличение количества ошибок первого и второго типа при мониторинге и реагировании на нетипичные инциденты ИБ, что может привести к их пропуску, несмотря на общее снижение числа ошибок первого и второго типа [17, 18].
3. Проблемы объяснимости, прозрачности и риски, связанные с автономным принятием решений ИИ-агентами, затрудняющим анализ причин принятых решений, проверку корректности их работы, а также сложности в определении ответственности за неправильные действия при отсутствии достаточного человеческого контроля [2, 13, 18].
4. Уязвимость к атакам на обучающие наборы данных, когда злоумышленники специально создают некорректные или особым образом размеченные массивы данные для «обмана» моделей ИИ, снижая их эффективность в мониторинге и реагировании на инциденты ИБ [19].
5. Зависимость от качества обучающих наборов данных, где некорректные или предвзятые данные могут привести к систематическим ошибкам в работе ИИ-агентов [17, 20].
6. Рост сложности инфраструктуры, ведущий к увеличению поверхности атаки и появлению новых точек отказа в системах защиты [2].

7. Недостаточная адаптивность к новым типам угроз, особенно когда агенты обучаются на устаревших данных и не могут эффективно реагировать на изменяющиеся тактики и техники атакующих [12, 19].

Таким образом, проведенная аналитическая работа демонстрирует сильные и слабые стороны предлагаемой технологии повышения уровня киберустойчивости информационной инфраструктуры за счет внедрения СЗИ на базе ИИ-агентов для решения задач мониторинга и реагирования на инциденты ИБ. Полученные результаты позволяют сделать вывод о значительных преимуществах данного подхода по сравнению с применением классических или использующих технологии ИИ (машинное, глубокое обучение) СЗИ, а также отражают его ключевые риски и ограничения.

ЛИТЕРАТУРА

1. Намиот, Д.Е. О кибербезопасности ИИ-агентов / Д.Е. Намиот, Е.А. Ильюшин // International Journal of Open Information Technologies. — 2025. — Т. 13, № 9. — С. 13–24. — EDN YMDPLP (дата обращения: 04.01.2026).
2. Kshetri N. Transforming cybersecurity with agentic AI to combat emerging cyber threats // Telecommunications Policy. — 2025. — Vol. 49, iss. 6. — P. 102976. — ISSN 0308-5961. — DOI: 10.1016/j.telpol.2025.102976. — URL: https://www.sciencedirect.com/science/article/pii/S0308596125000734 (дата обращения: 20.01.2026).
3. Roy S. AgenticCyber: A GenAI-Powered Multi-Agent System for Multimodal Threat Detection and Response. — 2025. — URL: https://arxiv.org/html/2512.06396v1 (дата обращения: 20.01.2026).
4. Kamalakanta Sethi, Y. Venu Madhav, Rahul Kumar, Padmalochan Bera: Attention based multi-agent intrusion detection systems using reinforcement learning — 2021. — URL: https://www.sciencedirect.com/science/article/abs/pii/S2214212621001411 (дата обращения: 20.01.2026).
5. Amani Bacha, Farah Barika Ktata, Faten Louati: Improving Intrusion Detection Systems with Multi-Agent Deep Reinforcement Learning: Enhanced Centralized and Decentralized Approaches — 2023. — URL: https://www.scitepress.org/Papers/2023/121246/121246.pdf (дата обращения: 20.01.2026).
6. CSE-CIC-IDS2018 on AWS. A collaborative project between the Communications Security Establishment (CSE) & the Canadian Institute for Cybersecurity (CIC). — 2018. — URL: https://www.unb.ca/cic/datasets/ids-2018.html (дата обращения: 20.01.2026).
7. CIC UNSW-NB15 Augmented Dataset. — 2024. — URL: https://www.unb.ca/cic/datasets/cic-unsw-nb15.html (дата обращения: 20.01.2026).
8. Samed AL, Seref Sagiroglu. Explainable artificial intelligence models in intrusion detection systems. — 2025. — URL: https://www.sciencedirect.com/science/article/abs/pii/S0952197625001459 (дата обращения: 20.01.2026).
9. Maryam Roshanaei, Mahir R. Khan, Natalie N. Sylvester. Enhancing Cybersecurity through AI and ML: Strategies, Challenges, and Future // Challenges, and Future Directions. Journal of Information Security, 15, 320–339. doi: 10.4236/jis.2024.153019 — 2024. — URL: https://www.scirp.org/journal/paperinformation?paperid=134347 (дата обращения: 20.01.2026).
10. Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. — 2017. — Т. 30. — URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (дата обращения: 20.01.2026).
11. Bacha A., Ktata F.B.., Louati F. Improving Intrusion Detection Systems with Multi-Agent Deep Reinforcement Learning. SCITEPRESS. — 2023. — URL: https://www.scitepress.org/Papers/2023/121246/121246.pdf (дата обращения: 20.01.2026).
12. Liu Z., Wang L., Chen X., Zhang Y. Multi-Agent Collaboration in Incident Response with Large Language Models. — 2024. — URL: https://arxiv.org/html/2412.00652v1 (дата обращения: 20.01.2026).
13. Wang M., Dechene R. Multi-Agent Actor-Critics in Autonomous Cyber Defense. — 2024. — URL: https://arxiv.org/html/2410.09134v1 (дата обращения: 20.01.2026).
14. Kaur J. AI Agents Re-Define Security Operations Testing and Verification Tasks. Akira AI Blog. — 2025. — URL: https://www.akira.ai/blog/ai-agents-for-verification-tasks (дата обращения: 20.01.2026).
15. Ferrag M.A., Tihanyi N., Hamouda D., Maglaras L., Lakas A., Debbah M. From prompt injections to protocol exploits: Threats in LLM-powered AI agents workflows. Computers & Security, 2025, Vol. 145, P. 103025. — 2025. — URL: https://www.sciencedirect.com/science/article/pii/S2405959525001997 (дата обращения: 20.01.2026).
16. Намиот Д.Е. Искусственный Интеллект в Кибербезопасности. Хроника. Выпуск 4. — 2026. — URL: https://cyberleninka.ru/article/n/iskusstvennyy-intellekt-v-kiberbezopasnosti-hronika-vypusk-4 (дата обращения: 20.01.2026).
17. Verizon. 2025 Data Breach Investigations Report. — 2025. — URL: https://www.verizon.com/business/resources/T5ef/reports/2025-dbir-data-breach-investigations-report.pdf (дата обращения: 20.01.2026).
18. Schroeder de Witt C. Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents. — 2025. — URL: https://arxiv.org/html/2505.02077v1 (дата обращения: 20.01.2026).
19. AL S., Sagiroglu S. Explainable artificial intelligence models in intrusion detection systems. Engineering Applications of Artificial Intelligence, 2025, Vol. 139, P. 109548. — 2025. — URL: https://www.sciencedirect.com/science/article/abs/pii/S0952197625001459 (дата обращения: 20.01.2026).
20. Kotte G. Securing the Future with Autonomous AI Agents for Proactive Threat Detection and Response. International Research Journal of Engineering and Management Sciences, 2025, Vol. 4, Issue 5, P. 123–135. — 2025. — URL: https://irjems.org/Volume-4-Issue-5/IRJEMS-V4I5P123.pdf (дата обращения: 20.01.2026).
21. Ляпунцова Е.В., Арм А.А.С. Использование искусственного интеллекта для повышения сетевой безопасности: стратегии обнаружения аномалий и перспективы. — 2024. — URL: https://cyberleninka.ru/article/n/ispolzovanie-iskusstvennogo-intellekta-dlya-povysheniya-setevoy-bezopasnosti-strategii-obnaruzheniya-anomaliy-i-perspektivy (дата обращения: 20.01.2026).
22. R. da Silveira Lopes, J.C. Duarte and R.R. Goldschmidt, «False Positive Identification in Intrusion Detection Using XAI», in IEEE Latin America Transactions, vol. 21, no. 6, pp. 745–751, June 2023, doi: 10.1109/TLA.2023.10172140-2023 — URL: https://ieeexplore.ieee.org/abstract/document/10172140 (дата обращения: 21.01.2026).