УДК: 004.056:004.8

Чуркин Павел Юрьевич

Абитуриент магистратуры Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ), г. Москва

ИСПОЛЬЗОВАНИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ОБНАРУЖЕНИЯ И НЕЙТРАЛИЗАЦИИ ZERO-DAY УЯЗВИМОСТЕЙ

Аннотация

В условиях быстрого роста киберугроз zero-day уязвимости остаются одной из самых серьёзных проблем для критической информационной инфраструктуры и корпоративных систем. В этой статье разбирается, как искусственный интеллект может помочь быстро находить и нейтрализовать такие неизвестные уязвимости. Буду рассмотрены основные методы машинного обучения и поведенческого анализа, будут приведены реальные примеры их применения, а также обсудим комплексные меры защиты - технические, организационные и правовые. Особое внимание уделяется тому, что ИИ одновременно помогает и злоумышленникам, и тем, кто защищается. Статья предлагает опираться на международные стандарты, такие как NIST Cybersecurity Framework и OWASP, и показывает, почему важно развивать современные технологии кибербезопасности в эпоху цифровизации.

Abstract

In the context of the rapid growth of cyber threats, zero-day vulnerabilities remain one of the most serious challenges for critical information infrastructure and corporate systems. This article examines how artificial intelligence can help rapidly detect and neutralize such previously unknown vulnerabilities. The main machine learning and behavioral analysis techniques are discussed, real-world examples of their application are presented, and comprehensive protection measures—technical, organizational, and legal—are explored. Particular attention is paid to the fact that AI simultaneously empowers both attackers and defenders. The article advocates relying on international standards such as the NIST Cybersecurity Framework and OWASP and demonstrates why the development of modern cybersecurity technologies is essential in the era of digitalization.

Ключевые слова

искусственный интеллект, zero-day уязвимости, кибербезопасность, машинное обучение, поведенческий анализ, критическая информационная

инфраструктура, нейтрализация угроз, NIST Cybersecurity Framework, OWASP Top 10.

Keywords

artificial intelligence, zero-day vulnerabilities, cybersecurity, machine learning, behavioral analysis, critical information infrastructure, threat neutralization, NIST Cybersecurity Framework, OWASP Top 10.

Актуальность проблемы обнаружения и нейтрализации zero-day уязвимостей сегодня очень высока. По данным Google Threat Intelligence Group, в 2025 году в атаках использовали около 90 zero-day уязвимостей. Это на 15 % больше, чем годом ранее. Почти половина из них - примерно 48 % - была направлена на корпоративные технологии: сетевые устройства, системы безопасности и платформы виртуализации [1]. Среднее время от публикации уязвимости до её активной эксплуатации сильно сократилось. Теперь оно часто составляет всего несколько дней, а иногда и часы. Согласно отчёту Verizon Data Breach Investigations Report за 2026 год, эксплуатация уязвимостей стала самым частым способом проникновения в системы - 31 % всех инцидентов. Это впервые обогнало кражу паролей и учётных данных [2].

Поэтому классические антивирусы и базы известных уязвимостей (CVE) уже не справляются. Они хорошо работают против известных угроз, но почти бесполезны против новых, неизвестных. Особенно это опасно для критической инфраструктуры, где успешная атака может привести к серьёзным последствиям для бизнеса и общества.

Искусственный интеллект меняет ситуацию в обе стороны. С одной стороны, злоумышленники начали активно использовать ИИ. Они автоматизируют поиск уязвимостей, создают вредоносное ПО и даже генерируют эксплойты с помощью больших языковых моделей. Уже есть подтверждённые случаи, когда ИИ помогал создавать реальные zero-day эксплойты, в том числе для обхода двухфакторной аутентификации [1, 3].

С другой стороны, ИИ даёт защитникам мощный инструмент. Модели машинного обучения могут в реальном времени анализировать огромные объёмы данных - код, сетевой трафик и поведение пользователей - и находить аномалии, которые человек или обычные системы пропускают.

Один из самых эффективных инструментов защиты - автоэнкодеры. Они сжимают исходные данные и пытаются их восстановить. Если восстановление сильно отличается от оригинала, система сигнализирует об аномалии. На практике такие модели показывают точность выше 95 % при обнаружении эксплойтов типа buffer overflow или use-after-free [4]. Дополнительно используют генеративно-состязательные сети (GAN) и

графовые нейронные сети (GNN). GAN помогают создавать синтетические примеры атак для обучения. GNN хорошо работают с графами связей в инфраструктуре и могут предсказывать развитие атаки на ранних этапах.

Поведенческий анализ сегодня считается одним из самых перспективных направлений. Вместо того чтобы искать конкретные сигнатуры вредоносного кода, система изучает, как обычно ведут себя пользователи, приложения и устройства. Любое заметное отклонение от этой «нормы» считается подозрительным.

Современные платформы обнаружения и реагирования на конечных точках (EDR) и анализа сетевого трафика (NTA), построенные на базе искусственного интеллекта, формируют для каждой сущности в сети индивидуальную динамическую модель её нормального поведения. При резком отклонении от этой модели система автоматически изолирует устройство, собирает данные для расследования и предлагает рекомендации по реагированию.

Типичная архитектура такой системы состоит из нескольких слоёв. Сначала собирается телеметрия с компьютеров, серверов и сети. Далее работает ансамбль моделей: автоэнкодер для поиска аномалий, классификаторы для известных угроз и графовые сети для анализа связей. Отдельный модуль на основе reinforcement learning помогает принимать решения о реагировании. В конце всё интегрируется с SIEM-системой для формирования понятного отчёта [5, 6].

Злоумышленники тоже не стоят на месте. Они используют большие языковые модели для автоматического поиска уязвимостей и создания эксплойтов. По данным Mandiant и Google, время от появления уязвимости до её массовой эксплуатации в 2025 году заметно сократилось [3, 7]. История уже показала, насколько опасными могут быть zero-day уязвимости. Пример - Stuxnet в 2010 году и Industroyer в 2016-м. Они использовали неизвестные уязвимости для атак на промышленные системы. Сегодня похожие атаки происходят против браузеров, VPN и межсетевых экранов.

При этом ИИ уже помогает и в защите. Платформа Trend Micro с помощью ИИ-агентов за короткое время нашла 21 новую zero-day уязвимость в инфраструктуре искусственного интеллекта [8]. Системы вроде Darktrace успешно останавливают как новые, так и известные атаки, просто замечая необычное поведение в сети [9]. Для борьбы с zero-day уязвимостями в реальном времени используют несколько подходов. Это защита во время выполнения программ (runtime protection), автоматическая микросегментация сети, динамический виртуальный патчинг и быстрый откат изменений. ИИ-

агенты могут прямо во время атаки анализировать эксплойт и предпринимать меры для блокировки.

Однако у таких систем есть и свои сложности. Одна из главных - ложные срабатывания. Если модель слишком чувствительная, она будет постоянно тревожить специалистов. Если слишком «спокойная» - пропустит реальную угрозу. Кроме того, сами системы ИИ нужно защищать от атак: злоумышленники могут «отравить» данные для обучения или манипулировать выводами модели.

Ещё одна важная задача - сделать работу ИИ понятной для человека. Специалист по безопасности должен понимать, почему система приняла то или иное решение, особенно если оно касается изоляции важных сегментов сети. Поэтому всё чаще используют гибридные решения: сочетают мощь глубокого обучения с экспертными правилами и обязательной проверкой человеком самых важных решений [10].

Чтобы эффективно защищаться, нужны меры на разных уровнях. На техническом уровне стоит внедрять SIEM-системы с ИИ, которые хорошо оценивают события и используют современные модели для анализа. Также помогают технологии сегментации сети и защиты приложений во время работы. На организационном уровне важно создавать центры компетенций по безопасности ИИ, проводить регулярные учения red team / blue team с имитацией атак с помощью ИИ и внедрять безопасный цикл разработки ИИ-систем.

С правовой стороны полезно ориентироваться на международные документы - NIST Cybersecurity Framework Profile for Artificial Intelligence и OWASP Top 10 for LLM Applications. Они помогают выстроить управление рисками как при использовании ИИ для защиты, так и при защите самих ИИ-систем [11, 12].

Эффективная защита также предполагает разумный подход к выбору технологий. Лучше не зависеть от одного поставщика, а диверсифицировать решения и сохранять контроль над ключевыми компонентами. Как показывают исследования Gartner и Darktrace, инвестиции в современные ИИ-инструменты кибербезопасности окупаются за счёт снижения количества и тяжести инцидентов [13]. Конечно, при внедрении ИИ возникают и этические вопросы. Нужно контролировать уровень ложных срабатываний, обеспечивать прозрачность работы моделей и защищать сами защитные системы от атак. Гибридные подходы, где ИИ помогает человеку, а не заменяет его в ключевых решениях, пока остаются самым надёжным вариантом.

В итоге использование искусственного интеллекта для борьбы с zero-day уязвимостями в реальном времени - это уже не просто перспективная идея, а реальная необходимость. Комплексный подход, который сочетает сильные технологии, грамотную организацию и хорошую подготовку специалистов, позволяет значительно повысить защищённость цифровой инфраструктуры.

Искусственный интеллект заметно ускоряет реакцию на угрозы и помогает небольшому числу экспертов справляться с большим объёмом работы. Но при этом важно помнить о рисках и не забывать о человеческом контроле. Нужно быстрее разрабатывать и внедрять стандарты по применению ИИ в защите критической инфраструктуры. Следует обязательно добавлять поведенческий анализ и поиск аномалий в системы защиты важных объектов. Требуются серьёзные инвестиции в исследования, подготовку кадров и тестовые площадки для отработки сценариев атак с ИИ. Полезно проводить регулярные аудиты с помощью ИИ-агентов и создавать центры, которые занимаются противодействием угрозам, усиленным искусственным интеллектом. Только такой всесторонний и проактивный подход позволит сохранить устойчивость цифровой инфраструктуры в условиях растущих киберугроз.

Список литературы

1. Google Threat Intelligence Group. Look What You Made Us Patch: 2025 Zero-Days in Review [Электронный ресурс]. – 2026. – URL: https://cloud.google.com/blog/topics/threat-intelligence/2025-zero-day-review

2. Verizon. 2026 Data Breach Investigations Report (DBIR) [Электронный ресурс]. – 2026. – URL: https://www.verizon.com/business/resources/reports/dbir/

3. Mandiant / Google GTIG. M-Trends 2026 Report [Электронный ресурс]. – 2026. – URL: https://cloud.google.com/blog/topics/threat-intelligence/m-trends-2026

4. Guo Y., et al. A Survey of Machine Learning-Based Zero-Day Attack Detection // PMC. – 2023. – URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC9890381

5. Trend Micro. Fault Lines in the AI Ecosystem: TrendAI™ State of AI Security Report [Электронный ресурс]. – 2026. – URL: https://www.trendmicro.com/vinfo/us/security/news/threat-landscape/fault-lines-in-the-ai-ecosystem-trendai-state-of-ai-security-report

6. Darktrace. The State of AI Cybersecurity Report 2025–2026 [Электронный ресурс]. – 2025–2026. – URL: https://www.darktrace.com/the-state-of-ai-cybersecurity-2025

7. Bright Defense. 80+ Zero-Day Exploit Statistics (March 2026) [Электронный ресурс]. – 2026. – URL: https://www.brightdefense.com/resources/zero-day-exploit-statistics/

8. Trend Micro. Introducing ÆSIR: Finding Zero-Day Vulnerabilities at the Speed of AI [Электронный ресурс]. – 2026. – URL: https://www.trendmicro.com/en/research/26/a/aesir.html

9. Darktrace. Self-Learning AI for Zero-Day and N-Day Attack Defense (case studies) [Электронный ресурс]. – 2025–2026. – URL: https://www.darktrace.com/blog/using-self-learning-ai-to-defend-against-zero-day-and-n-day-attacks

10. Ali S., et al. Comparative Evaluation of AI-Based Techniques for Zero-Day Malware Detection // Electronics. – 2022. – Vol. 11, No. 23. – URL: https://www.mdpi.com/2079-9292/11/23/3934

11. National Institute of Standards and Technology (NIST). Cybersecurity Framework Profile for Artificial Intelligence (NIST IR 8596) [Электронный ресурс]. – 2025. – URL: https://nvlpubs.nist.gov/nistpubs/ir/2025/NIST.IR.8596.iprd.pdf

12. OWASP. Top 10 for LLM Applications 2025 [Электронный ресурс]. – 2025. – URL: https://genai.owasp.org/llm-top-10/

13. Gartner. Cybersecurity Trends and AI Impact Reports 2025–2026 [Электронный ресурс]. – 2025–2026. – URL: https://www.gartner.com/en/information-technology/research/cybersecurity-trends