УДК 004.056

Д. В. Залевський
В. С. Катаєв

АВТОНОМНІ AI-АГЕНТИ В СОС: НОВЕ ПОЛЕ БОЮ МІЖ КІБЕРАТАКОЮ ТА ЗАХИСТОМ

Вінницький національний технічний університет

Анотація

Робота присвячена дослідженню впливу автономних агентів штучного інтелекту на безпеку центрів моніторингу кібербезпеки (SOC). Показано, що перехід від класичної автоматизації до автономних AI-SOC-агентів створює як нові можливості для виявлення атак у реальному часі, так і нові вектори зламу через компрометацію агентів, маніпуляцію їхнім контекстом та підміну ланцюгів постачання моделей. Проаналізовано типові сценарії атак на AI-агентів у SOC, AI-підсилені фішингові та ransomware-кампанії, а також підходи до red teaming та побудови «автономної SOC-безпеки».

Ключові слова: автономні AI-агенти, SOC, кібербезпека, AI-powered SOC, AI red teaming, LLM, штучний інтелект.

Abstract

The paper explores how autonomous AI agents are transforming Security Operations Centers (SOCs) from rule-based automation to AI-driven autonomy. We show that AI SOC agents can dramatically reduce detection and response times by automating triage, correlation, and incident handling, but they also introduce new attack surfaces: prompt and tool injection, poisoned model supply chains, and AI-enabled phishing and ransomware campaigns. The work summarizes recent industry cases of AI-powered SOC deployments, discusses AI red teaming approaches against LLM-based agents, and outlines key directions for building secure, human-overseen autonomous SOCs.

Keywords: autonomous AI agents, SOC, cybersecurity, AI-powered SOC, AI red teaming, LLM, artificial intelligence.

Вступ

У 2025–2026 роках центри моніторингу кібербезпеки (SOC) стають одними з перших майданчиків, де автономні AI-агенти виходять за межі «чат-ботів» і починають самостійно приймати рішення: пріоритизувати інциденти, корелювати журнали, запускати плейбуки реагування[1]. Аналітики говорять про «рік автономного SOC» та про те, що до 40 % великих підприємств впровадять AI-агентів у своїх SOC уже найближчим часом. Паралельно з цим зростає й offensive-вимір штучного інтелекту: LLM-подібні моделі допомагають автоматизувати фішинг, створювати варіанти шкідливого ПЗ, проводити AI-red teaming проти захисних систем. Це означає, що SOC одночасно отримує «свого» AI-захисника й стикається з хвилею AI-підсилених атак, які намагаються обманути або скомпрометувати цього захисника. Тому актуальним є аналіз ролі автономних AI-агентів у сучасних SOC, виявлення нових векторів атак, пов'язаних із впровадженням таких агентів, а також узагальнення підходів до тестування та зміцнення безпеки AI-SOC (AI red teaming, захист ланцюгів постачання моделей, контроль дій агентів). У роботі використано аналітичні звіти провідних вендорів SOC-рішень щодо впровадження автономних AI-агентів у 2025 році, огляди інструментів AI red teaming для тестування LLM-систем, дослідження щодо використання великих мовних моделей для генерування варіантів шкідливого ПЗ, а також матеріали з практики побудови AI-підсилених SOC. Застосовано методи порівняльного аналізу класичних SOC-процесів та AI-орієнтованих сценаріїв, аналіз атак на LLM-агентів, а також узагальнення підходів до побудови безпечних автономних систем.

Результати дослідження

Традиційно SOC спирався на кореляційні правила, сигнатурний аналіз та напівавтоматизовані плейбуки реагування. Сучасні AI-SOC-агенти переходять до моделі, де системи самі «мислять» як досвідчений аналітик: збирають контекст, об'єднують журнали з різних джерел, оцінюють критичність інциденту та пропонують чи навіть виконують дії реагування [2]. AI-платформи безпеки декларують

можливість автоматичного проведення розслідування: агент, отримавши тригер, збирає логи з endpoint, мережеві події, історію користувача, виконує кореляцію, формує гіпотезу щодо типу атаки та пропонує блокування облікових записів або сегментацію мережі. У найрадикальніших сценаріях автономні агенти можуть виконувати зміни в конфігурації без прямого підтвердження людини, що суттєво скорочує час реагування, але створює нову зону ризику.

Перенесення логіки ухвалення рішень у AI-агентів відкриває для зловмисників нові цілі. До класичних вразливостей SOC (невірні правила кореляції, помилки в SIEM/EDR, відсутність сегментації) додаються:

Prompt- та tool-ін'єкції проти SOC-агентів. Якщо агент має доступ до логів, тікет-системи, знань бази інцидентів, то будь-які дані, які він «читає», потенційно можуть містити інструкції, що змінюють його поведінку. Атака через спеціально сформований рядок у журналі або тікеті здатна змусити агента приховати певний клас подій чи, навпаки, створити «шум» із фальшивих інцидентів.

Компрометація ланцюга постачання моделей. Використання відкритих чи сторонніх моделей для SOC-платформ створює ризик отруєних датасетів, моделей із прихованими тригерами або підроблених «оновлень», які змінюють реакції агента в критичних ситуаціях.

Атаки на інструменти, якими керує агент. Навіть якщо сама модель відносно захищена, зловмисник може цілитися в API, скрипти та плейбуки, які викликає агент: змінюючи їх логіку, щоб перетворити «автономного захисника» на мимовільного співучасника атаки.

З боку наступу штучний інтелект так само змінює правила гри. Дослідження показують, що LLM можна використати для генерування варіантів шкідливого коду, які обминають частину сигнатурних та евристичних детекторів, при цьому зберігаючи функціональність ПЗ [3]. Паралельно платформи на кшталт WormGPT- чи FraudGPT-подібних рішень пропонують автоматизовані фішингові й BEC-кампанії, де тексти максимально імітують стиль конкретних керівників чи колег [4].

Для SOC це означає збільшення обсягу «правдоподібного шуму»: фішингові листи та комунікація зловмисників стають настільки схожими на легітимну, що класичні правила детекції втрачають ефективність. Водночас AI може використовуватись для маскування lateral movement, автоматичного підбору оптимальних моментів для атаки та адаптації технік під конкретне середовище. У відповідь на нові ризики формується практика AI red teaming, спрямована на тестування стійкості LLM-агентів і AI-процесів у SOC до ворожих впливів. Інструменти AI red teaming дозволяють моделювати prompt-ін'єкції, спроби обійти політики безпеки, а також зловживання автономними діями агентів. У контексті SOC це означає заплановані «атаки» на аналітичних агентів: перевірку, чи здатні вони розпізнавати маніпульований контекст, не виконувати небезпечні запити та коректно ескалювати інциденти людині.

Практика показує, що без систематичного AI red teaming навіть добре захищені SOC-платформи можуть мати «сліпі зони», де агент надто довіряє вхідним даним, неправильно інтерпретує політики або приймає рішення без належної валідації. Аналіз тенденцій дозволяє виділити кілька ключових напрямів, без яких впровадження AI-агентів у SOC стає надто ризикованим:

Принцип «human-in-the-loop» для критичних дій. Навіть за високого рівня автоматизації фінальні рішення щодо блокування облікових записів, зміни політик доступу чи зупинки бізнес-критичних сервісів мають підтверджуватися людиною.

Жорстка ізоляція та валідація дій агентів. Усі виклики до внутрішніх API та плейбуків повинні проходити через шар політик, що перевіряють контекст, параметри та можливі побічні ефекти.

Захист ланцюга постачання моделей. Перевірка походження моделей, контроль оновлень, сканування контейнерів із AI-компонентами, моніторинг для виявлення аномальної поведінки агентів.

Вбудований AI red teaming. Регулярне тестування SOC-агентів на стійкість до prompt-ін'єкцій, отруєних даних, спроб примусу порушити політики безпеки.

Висновки

Автономні AI-агенти в SOC стають одночасно найпотужнішим інструментом оборони й новою ціллю для атак. Перехід від ручних плейбуків до автономних рішень дозволяє скоротити час виявлення та реагування, але вимагає переосмислення безпеки: від захисту правил та скриптів — до захисту моделей, їхнього контексту та життєвого циклу. Подальші дослідження повинні зосередитися на формальних методах верифікації дій SOC-агентів, стандартах безпечного впровадження LLM у критичну інфраструктуру, а також на інтеграції AI red teaming у повсякденну практику кіберзахисту.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Palo Alto Networks. 2025: The Year of the Autonomous SOC. The Year of XSIAM. 2025. URL: https://www.paloaltonetworks.com/blog/security-operations/2025-the-year-of-the-autonomous-soc-the-year-of-xsiam/.
2. Vertu. Top AI Red Teaming Tools for 2025. 2025. URL: https://vertu.com/ai-tools/top-ai-red-teaming-tools-2025/.
3. Delamotte M., Bernade-Shapiro Y. LLM-Enabled Malware In the Wild. SentinelLABS, 2026. URL: https://www.sentinelone.com/labs/labscon25-replay-llm-enabled-malware-in-the-wild/.
4. Radiant Security. Real-World Use Cases of AI-Powered SOC. 2025. URL: https://radiantsecurity.ai/learn/soc-use-cases/.

Залевський Дмитро Володимирович – студент групи 3КІТС-24б, факультет менеджменту та інформаційної безпеки, Вінницький національний технічний університет, м. Вінниця, e-mail: fareinheits@gmail.com

Катаєв Віталій Сергійович – асистент кафедри менеджменту та безпеки інформаційних систем, Вінницький національний технічний університет, Вінниця, kataev@vntu.net

Dmytro Zalevskyi – student in Faculty of Management and Information Security, Vinnytsia National Technical University, Vinnytsia, e-mail: fareinheits@gmail.com

Vitaliy Kataiev – assistant of the Department of Management and Security of Information Systems; Vinnytsia National Technical University, Vinnytsia, kataev@vntu.net