УДК 004.056.5

Ю.Є. Яремчук

Л.Ю. Рудь

ПІДВИЩЕННЯ ЗАХИЩЕННОСТІ КОРПОРАТИВНИХ КОМП'ЮТЕРНИХ МЕРЕЖ НА ОСНОВІ AI-АГЕНТІВ ДЛЯ АНАЛІЗУ ЗАГРОЗ У СЕРЕДОВИЩІ N8N ТА ГІБРИДНОГО МЕТОДУ АДАПТИВНОГО РЕАГУВАННЯ

Вінницький національний технічний університет

Анотація

У роботі представлено комплексний аналіз підвищення захищеності корпоративних мереж шляхом застосування мультиагентних систем штучного інтелекту, гібридних методів виявлення загроз та автоматизації реагування на базі платформи n8n. Запропоновано архітектуру «Dual-LLM Airlock», що поєднує агента-вартового та агента-аналітика для детектування Prompt Injection, фішингових атак та інших сучасних векторів загроз. Проведено експериментальну верифікацію на тестовому наборі понад 1200 сценаріїв; отримано підвищення стійкості до ін'єкцій до ≈93% та зниження рівня хибних спрацьовувань при виявленні фішингу.

Ключові слова: мультиагентні системи, Prompt Injection, фішинг, n8n, SOAR.

Abstract

The paper provides an extended analysis of methods to enhance corporate network security using multi-agent AI systems, hybrid threat detection approaches and automated response workflows implemented via the n8n platform. A Dual-LLM Airlock architecture is proposed and experimentally validated, demonstrating improved resilience to prompt injections and reduced false positives in phishing detection.

Keywords: multi-agent systems, prompt injection, phishing, n8n, SOAR.

Вступ

Сучасний ландшафт кіберзагроз характеризується безпрецедентною складністю, швидкістю еволюції та зростаючою частотою атак. Корпоративні мережі, які є основою цифрової економіки, стикаються з постійним тиском, що вимагає переходу від застарілих реактивних моделей до проактивних, інтелектуальних стратегій захисту. Традиційні центри управління безпекою (SOC), які покладаються на ручний аналіз та сигнатурні методи, виявилися недостатньо ефективними перед обличчям цієї нової реальності. Необхідність захисту критично важливих цифрових активів та забезпечення безперервності бізнесу в умовах, де кіберзлочинці також використовують передові

технології, такі як штучний інтелект (ШІ), зробила трансформацію операцій SOC не просто бажаною, а критичною необхідністю для стійкості організацій. Робота присвячена аналізу головних викликів, з якими стикаються корпоративні мережі, та їх порівнянню з можливостями, які надає інтеграція ШІ, автоматизації та інноваційних технологій, для побудови ефективного кіберзахисту. Актуальність. У зв'язку з бурхливим розвитком та інтеграцією агентних систем ШІ, що поєднують великі мовні моделі з доступом до зовнішніх інструментів та API, виникає новий, критичний клас загроз безпеці. Традиційні підходи до захисту, такі як доналаштування моделей чи використання зовнішніх класифікаторів, працюють за принципом «найкращих зусиль» (best-effort) і демонструють свою неефективність проти нових векторів атак, зокрема ін’єкцій запитів (prompt injections). Такі атаки можуть призводити до несанкціонованого витоку даних або віддаленого виконання коду. Це зумовлює гостру необхідність у розробці принципово нових, надійних механізмів захисту, здатних надавати формальні гарантії безпеки агентів ШІ, а не лише намагатися виявити загрозу. Об'єктом дослідження є процес функціонування агентних систем ШІ та їхня взаємодія з програмними інструментами. Предметом дослідження є методи, моделі та засоби забезпечення контрольованої та безпечної поведінки агентів ШІ шляхом застосування формально верифікованих політик безпеки. Метою магістерської кваліфікаційної роботи є підвищення рівня безпеки та контрольованості агентних систем ШІ шляхом розробки архітектури, що інтегрує агента з формальним аналізатором безпеки, який блокує небезпечні дії на основі заздалегідь визначених правил.

Для досягнення поставленої мети необхідно вирішити такі задачі: провести аналіз вразливостей сучасних агентних систем ШІ, зокрема загроз ін’єкцій запитів, що призводять до витоку даних; дослідити обмеження та недоліки існуючих «best-effort» методів захисту, таких як детектори ін’єкцій, та продемонструвати їхню ненадійність (наявність хибнопозитивних та хибнонегативних спрацювань); розробити концепцію та вимоги до системи з формальними гарантіями, що базується на зовнішньому аналізаторі політик, який перевіряє дії агента перед їх виконанням; створити формальну модель та архітектуру аналізатора, що оперує слідами дій агента t та використовує правила r(V,C) на основі аналізу інформаційних потоків (IFA); реалізувати прототип аналізатора з набором правил, визначених, для блокування специфічних вразливостей (наприклад, включення в електронні листи промт ін’єкцій та фішингові атаки).

Наукова новизна роботи полягає у запропонованій архітектурі поєднання агента ШІ з зовнішнім формальним аналізатором безпеки, на відміну від існуючих підходів, які намагаються виявити саму ін’єкцію запиту, запропонована система накладає жорсткі, формально верифіковані обмеження на дії агента, що дозволяє запобігати небезпечним наслідкам атаки, незалежно від її природи, та надавати строгі, доказові гарантії безпечної поведінки системи.

Практична цінність роботи полягає у розробці бібліотеки предикатів (наприклад, детектори особистої інформації (PII), секретних ключів, небезпечного коду) для опису політик безпеки. Створений прототип аналізатора може бути інтегрований у реальні робочі процеси агентів ШІ, забезпечуючи надійний захист від широкого класу атак, зокрема витоку конфіденційних даних та несанкціонованого виконання коду, тим самим мінімізуючи потенційні збитки.

Дослідження

Корпоративні мережі, що захищаються за допомогою традиційних підходів до кібербезпеки, страждають від трьох ключових взаємопов'язаних проблем: перевантаження даними та неефективність аналізу, нездатність виявляти просунуті загрози та критично повільна реакція на інциденти. Корпоративна мережа є складною, інтегрованою системою, що об'єднує різноманітні технології зв'язку,

методи підключення ресурсів та комунікаційні протоколи. В узагальненому випадку її структура може поділятися на основний та віддалені фрагменти.

Ключова роль центрального сегмента полягає в реалізації централізованого керування всією інфраструктурою. Для спрощення архітектури та посилення безпеки, критичні функції, такі як адміністрування, розміщення інформаційних серверів та контрольоване підключення до глобальних мереж (наприклад, Інтернету), часто концентруються саме в основному фрагменті. Таку комп'ютерну інфраструктуру прийнято представляти у вигляді багатошарової моделі. Вона складається з робочих станцій, а також різноманітних типів комп’ютерів, що диктують загальні можливості. Важливою складовою є апаратне забезпечення для комунікації, що включає мережеві карти, кабелі, міжмережеві екрани, а також проміжне обладнання, як-от комутатори та маршрутизатори. Функціонування мережі забезпечують операційні системи, що працюють поверх транспортної системи, розподіляючи ресурси та підтримуючи мережеві додатки.

Особливе місце серед програмного забезпечення займають системи управління базами даних (СУБД), оскільки вони відповідають за зберігання та пошук всієї ключової інформації підприємства. Використовуючи СУБД, працюють різноманітні системні сервіси, наприклад, електронна пошта чи служба WWW, які надають дані кінцевим користувачам. Верхівкою цієї ієрархії є спеціалізовані програмні комплекси, орієнтовані на вирішення конкретних завдань підприємства, як-от системи автоматизації проектування чи банківські системи. Успішна робота прикладних програм верхнього рівня повністю залежить від коректної та стабільної роботи усіх нижчих підсистем. Для забезпечення взаємодії цього комплексу обладнання були прийняті стандартизовані правила – мережеві протоколи, що диктують алгоритми передачі даних. Оскільки один протокол не може описати всю взаємодію, була затверджена багаторівнева модель, наприклад, семирівнева OSI. Набір протоколів, достатній для організації зв'язку, називається стеком; найпоширенішим сьогодні є стек TCP/IP. Критична проблема обсягу даних та втоми аналітиків. Однією з найбільших оперативних проблем, з якою стикаються традиційні SOC, є критичне перевантаження обсягом сповіщень.

Мережеві пристрої, кінцеві точки, хмарні середовища та системи моніторингу генерують масивні потоки даних безпеки, які необхідно обробляти в режимі реального часу. Аналітики вручну переглядають і сортують ці величезні масиви інформації. Як наслідок, виникає феномен «втоми від сповіщень» (alert fatigue). Велика кількість хибних спрацювань (false positives) – тобто сповіщень, які ідентифікують нормальну активність як потенційну загрозу, – призводить до того, що аналітики витрачають значний час на розслідування доброякісних подій. Це не тільки знижує загальну ефективність роботи SOC, але й створює ризик пропуску справжніх, критично важливих загроз серед шуму. Замість того, щоб зосередитися на складних задачах, які вимагають експертного судження, аналітики перевантажені рутинною роботою з тріажування та класифікації, що збільшує ризик людської помилки та оперативного уповільнення. Традиційні системи виявлення загроз значною мірою покладаються на правила, засновані на сигнатурах відомих шкідливих програм та векторів атак. Цей підхід ефективний лише проти відомих загроз. Однак сучасні кіберзагрози постійно еволюціонують, стаючи більш складними та витонченими.

Головною проблемою традиційних SOC є їхня нездатність виявляти: Атаки нульового дня (Zero-day attacks): Це нові, раніше невідомі експлойти, для яких ще не створено захисних сигнатур. Оскільки традиційні системи не мають відповідних правил, ці атаки можуть залишатися непоміченими протягом тривалого часу. Просунуті стійкі загрози (APTs): Ці загрози характеризуються повільними, прихованими та багатоетапними кампаніями, часто спрямованими на крадіжку даних або тривалий несанкціонований доступ до мережі. APTs можуть маскуватися під нормальну діяльність користувачів або систем, що

робить їх майже невидимими для простих правил безпеки. Таким чином, традиційні методи фіксують лише симптоми, але не можуть передбачити чи ідентифікувати нові, еволюціонуючі тактики кіберзлочинців. Ризики людського фактору та оперативна неефективність. Неефективність традиційних SOC також корениться в залежності від ручного втручання на всіх етапах інциденту. Це призводить до критично повільних показників: середнього часу виявлення (MTTD) та середнього часу реагування (MTTR). У традиційних операціях SOC аналітики повинні вручну: розслідувати кожне спрацювання, оцінюючи його серйозність, корелювати дані з різних систем (SIEM, файрволи, журнали), визначати план реагування та виконувати дії (ізоляція, анулювання доступу, усунення шкідливого ПЗ). Ці ручні процеси створюють значні затримки в ліквідації кіберінцидентів.

Навіть невеликі затримки дозволяють зловмисникам закріпитися в мережі, збільшуючи потенційну шкоду та час простою бізнесу. Крім того, людська помилка (від неправильної конфігурації до неправильної оцінки загрози) є однією з головних причин інцидентів безпеки. Основна відмінність полягає у методології виявлення. Традиційні системи виявлення загроз покладаються переважно на правила та сигнатури відомих шкідливих програм та векторів атак. Цей підхід ефективний проти вже задокументованих загроз, але він є реактивним і не здатний адекватно протистояти швидко еволюціонуючим атакам. На противагу цьому, сучасне ШІ-орієнтоване програмне забезпечення використовує алгоритми машинного навчання (ML), обробку природної мови (NLP) та поведінкову аналітику (Behavioral Analytics). Замість пошуку відомих сигнатур, ці системи встановлюють базовий рівень нормальної поведінки у корпоративній мережі (включаючи трафік, активність користувачів та кінцевих точок). ШІ аналізує величезні масиви даних у реальному часі і виявляє будь-які аномалії – тобто відхилення від цієї норми. Це дозволяє ШІ-системам виявляти раніше невідомі загрози, включаючи атаки нульового дня (zero-day attacks) та просунуті стійкі загрози (APTs), які часто маскуються під звичайну діяльність. Крім того, ШІ забезпечує предиктивний аналіз загроз, оцінюючи історичні моделі атак для прогнозування потенційних майбутніх загроз, що дозволяє застосовувати проактивні стратегії захисту. Інший критичний порівняльний аспект стосується ефективності роботи аналітиків. Традиційні SOC страждають від перевантаження сповіщеннями та високої кількості хибних спрацювань (false positives), що призводить до «втоми аналітиків». Сучасне програмне забезпечення, особливо інтегроване з Security Information and Event Management (SIEM), вирішує цю проблему за допомогою ШІ.

Алгоритми корелюють дані з численних джерел і використовують контекстно-орієнтований аналіз, щоб фільтрувати низькопріоритетні оповіщення та пріоритезувати високоризикові інциденти. Це забезпечує, що аналітики зосереджуються виключно на справжніх загрозах. Щодо реагування, то тут домінує Security Orchestration, Automation, and Response (SOAR), керований ШІ. У той час як традиційні операції вимагають ручного втручання для розслідування та стримування, що призводить до значних затримок. SOAR дозволяє миттєво виконувати заздалегідь визначені сценарії реагування (automated playbooks). Це різко скорочує середній час виявлення (MTTD) та середній час реагування (MTTR), мінімізуючи шкоду. Автоматизоване реагування також зменшує залежність від людського фактору та мінімізує ризик помилки, забезпечуючи послідовність дій. Сучасне програмне забезпечення для виявлення кіберзагроз у корпоративних мережах переживає трансформаційний зсув від традиційних, статичних рішень до інтелектуальних систем, керованих штучним інтелектом (ШІ) та автоматизацією. Цей порівняльний аналіз зосереджується на ключових відмінностях між традиційним і ШІ-орієнтованим програмним забезпеченням, що використовується в Центрах управління безпекою (SOC). Таким чином, сучасне програмне забезпечення забезпечує перехід від повільного, схильного до помилок реагування до швидкого, послідовного та проактивного управління кіберінцидентами. Корпоративні комп'ютерні мережі мають багато критичних елементів, які часто стають жертвами кіберзагроз не через складні

підходи, а через прості, але невідворотні помилки, такі як хибні конфігурації, слабкі паролі та «непатчені» системи. Автоматизовані тести внутрішніх мереж виявили, що ці "базові прогалини" є постійною проблемою. Загалом, 50% критичних вразливостей походять від хибних конфігурацій, 30% – від відсутності патчів, і 20% – від слабких паролів. Критичні системи управління ідентифікацією та облікові дані. Елементи управління ідентифікацією та доступом є основними цілями, оскільки їх компрометація може поставити під загрозу всю інформаційну систему.

Незважаючи на обізнаність, слабкі та повторно використані паролі є однією з найбільш поширених вразливостей. Понад 82% зломів, проаналізованих у звіті Verizon DBIR, були пов’язані зі скомпрометованими або слабкими обліковими даними. Крім того, багато служб, як-от Firebird Servers або Redis Service (з оцінкою CVSS 9.9), можуть приймати стандартні (дефолтні) облікові дані або взагалі не вимагати аутентифікації, дозволяючи зловмисникам легко отримати доступ до чутливих даних або ескалювати привілеї. Active Directory є ядром інфраструктури, і його неправильне налаштування має катастрофічні наслідки. Вразливість Необмеженого Делегування (Unconstrained Delegation) є ризиковою конфігурацією, що дозволяє скомпрометованій машині узурпувати ідентичність користувачів, які до неї підключаються. Якщо в систему зайде доменний адміністратор, зловмисник може отримати його квиток (TGT) і повний контроль над середовищем.

Крім того, атаки Kerberoasting і AS-REP Roasting експлуатують слабкі паролі облікових записів служб або облікові записи, які не вимагають попередньої аутентифікації Kerberos, дозволяючи викрадати хеші паролів. Недостатня аутентифікація: використання однофакторної аутентифікації є суттєвою вразливістю, оскільки вона залишає двері відчиненими, навіть якщо пароль є відносно надійним. Впровадження багатофакторної аутентифікації (MFA) може блокувати понад 99% парольних атак. Проблеми з програмним забезпеченням та обслуговуванням складають значну частку критичних вразливостей. Застаріле програмне забезпечення залишається найбільш поширеною точкою входу для атак. Відсутність своєчасних оновлень дозволяє зловмисникам використовувати відомі, легко експлуатовані вразливості.

У таблиці 1 наведено аналіз вразливостей системних характеристки корпоративних мереж до сучасних кіберзагроз.

Таблиця 1 - Аналіз вразливості системних характеристик корпоративних мереж до сучасних кіберзагроз

Категорія/ Характеристика	**Ключова вразливість системи**	**Кіберзагрози**	**Категорія/ Характеристика**
Обсяг мережевих та системних даних	Втома аналітиків (Alert Fatigue): Критичне перевантаження, спричинене надмірною кількістю сповіщень, багато з яких є хибними спрацюваннями.	Маскування загроз: Справжні, високопріоритетні загрози можуть бути пропущені або проігноровані серед загального "шуму" низькопріоритетних сповіщень.	Обсяг мережевих та системних даних
Централізоване	Цілісність даних:	Маніпуляція	Централізоване

зберігання даних та журналів (Logs)	Уразливість до несанкціонованого доступу, маніпуляцій та підробки даних безпеки.	аудиторським слідом: Спроби зловмисників змінити або видалити записи безпеки, щоб приховати свою активність або уникнути виявлення. Інсайдерські загрози.	зберігання даних та журналів (Logs)
Ручні процеси реагування на інциденти	Оперативна неефективність: Критично повільні показники MTTD (середній час виявлення) та MTTR (середній час реагування). Ризик людської помилки.	Швидке поширення загроз: Затримка дозволяє зловмисникам закріпитися в мережі, збільшуючи потенційну шкоду від атак. Ransomware-атаки: Швидке стримування є критично важливим для мінімізації збитків.	Ручні процеси реагування на інциденти
Кінцеві точки та поведінка користувачів	Складність ідентифікації прихованої активності: Важко відрізнити легітимну дію від злочинної, якщо атака маскується під нормальну поведінку.	Інсайдерські загрози: Зловмисники або скомпрометовані облікові записи демонструють аномалії в поведінці (наприклад, доступ до даних у незвичний час чи з незвичної локації).	Кінцеві точки та поведінка користувачів
Хмарні середовища (cloud security)	Динамічний та розподілений характер: Складність контролю та моніторингу в реальному часі.	Misconfigurations (неправильні конфігурації): Виникають через складність розподілених інфраструктур. Складні зовнішні атаки (DDoS, APTs): Вимагають адаптивного реагування.	Хмарні середовища (cloud security)

Висновки

У статті було розглянуто основні моделі штучного інтелекту та автоматизації, що використовуються в системах захисту корпоративних мереж. Проведений аналіз продемонстрував, що різні типи моделей мають свої унікальні переваги та недоліки, які впливають на їх ефективність у конкретних задачах. Наприклад, мультиагентні системи забезпечують високоточний контекстуальний аналіз трафіку, проте

вимагають значних обчислювальних ресурсів. Моделі на основі ШІ є швидкими й простими у використанні, але менш ефективними при роботі зі складними даними. Значний потенціал мають агентні системи, що дозволяють виявляти аномалії у поведінкових даних без потреби у мітках, а також інтеграція з n8n для аналізу контенту. Окрему увагу варто приділити моделям на основі Reinforcement Learning і GAN, які забезпечують адаптацію політик безпеки та генерують нові дані для навчання систем, відповідно. Результати проведеного дослідження свідчать, що комплексне використання цих моделей у системах захисту мереж дозволяє підвищити їхню точність, адаптивність і здатність реагувати на нові загрози. Таким чином, інтеграція сучасних моделей штучного інтелекту у рішення для кібербезпеки є ключем до забезпечення ефективного захисту конфіденційної інформації в умовах динамічного інформаційного середовища.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Cisco Talos. 2023 year in review. URL: https://blog.talosintelligence.com/cisco-talos2023-year-in-review/ (Last accessed: 08.10.2025).
2. CrowdStrike. Global Threat Report 2024. URL: https://go.crowdstrike.com/rs/281-OBQ-266/images/GlobalThreatReport2024.pdf (Last accessed: 08.10.2025).
3. ENISA. Threat Landscape 2024. URL: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024 (Last accessed: 08.10.2025).
4. What Is SOAR? Palo Alto Networks Cyberpedia. (Last accessed: 05.10.2025).
5. What is incident response? IBM. (Last accessed: 05.10.2025).
6. Incident management for high-velocity teams. Atlassian. (Last accessed: 05.10.2025).
7. IBM QRadar SOAR. Product page. (Last accessed: 03.10.2025).
8. Splunk SOAR. Product brief. (Last accessed: 03.10.2025).
9. Cortex XSOAR. Palo Alto Networks. (Last accessed: 03.10.2025).

Рудь Любомир Юрійович — студент групи 1КІТС-24м, факультет менеджменту та інформаційної безпеки, Вінницький національний технічний університет, Вінниця, e-mail: lybomurrud@gmail.com

Науковий керівник: **Яремчук Юрій Євгенович** — доктор технічних наук, професор кафедри менеджменту та безпеки інформаційних систем, Вінницький національний технічний університет, Вінниця, e-mail: yurevyar@vntu.edu.ua

Rud Liubomyr Y.— student of group 1KITS-24m, Faculty of Management and Information Security, Vinnytsia National Technical University, Vinnytsia, e-mail: lybomurrud@gmail.com

Supervisor: **Yaremchuk Yuriy Y.** — Doctor of Technical Sciences, Professor of the Department of Management and Information Systems Security, Vinnytsia National Technical University, Vinnytsia, e-mail: yurevyar@vntu.edu.ua