РИЗИКИ ГАЛЮЦИНАЦІЙ ШТУЧНОГО ІНТЕЛЕКТУ В СИСТЕМАХ МОНІТОРИНГУ КІБЕРБЕЗПЕКИ ТА МЕТОДИ ЇХ МІНІМІЗАЦІЇ

Іванський І.О., Нехороших Д.М.
Харківський національний університет радіоелектроніки, Харків, Україна

В сучасних умовах безперервного зростання кількості та складності кіберзагроз ефективне управління інформаційною безпекою (ISMS) є важливим завданням для організацій. Процес оперативного виявлення та менеджменту вразливостей є цінною складовою загальної системи управління інформаційною безпекою, що визначає здатність інфраструктури адаптуватися та реагувати на ризики [1]. Одну з основних функцій у цій структурі відіграють системи класу SIEM, які забезпечують централізований збір і аналіз подій безпеки. Оскільки традиційні рішення мають обмеження у виявленні складних та невідомих атак, інтеграція технологій штучного інтелекту (ШІ) дозволяє підвищити їхню ефективність за рахунок автоматизації аналізу великих масивів даних, виявлення аномалій та прискорення реагування.

Водночас інтеграція інструментів штучного інтелекту супроводжується специфічним викликом – ризиком ШІ-галюцинацій. Генеративні моделі можуть створювати технічно правдоподібну, але насправді хибну інформацію. У системах моніторингу кібербезпеки такі помилки особливо небезпечні, оскільки результати автоматизованого аналізу можуть використовуватися для ідентифікації інцидентів, оцінювання вразливостей і формування рекомендацій щодо реагування.

Метою доповіді є дослідження ризиків, пов’язаних із галюцинаціями штучного інтелекту в системах моніторингу кібербезпеки, визначення їхнього впливу на достовірність результатів аналізу та ефективність виявлення кіберзагроз, а також обґрунтування методів мінімізації таких ризиків для підвищення надійності й точності функціонування систем моніторингу.

Загалом впровадження генеративного ШІ в процеси кібербезпеки супроводжується проблемою перевірки достовірності згенерованої інформації. Окрім генерації коректних результатів, ще однією проблемою є явище «галюцинацій», коли мовні моделі штучного інтелекту генерують текст, який здається правдоподібним, але є повністю вигаданим або не має фактичного підґрунтя [2]. Ці помилки не є тривіальними, оскільки це не дрібні помилки; це явні неправдиві твердження про API, яких не існує, небезпечні налаштування та бібліотеки, яких не існує. Масштаб цієї загрози підтверджується дослідженнями надійності генерації коду, в рамках яких дослідники згенерували 2,23 мільйона зразків коду за допомогою 16 популярних моделей генерації коду для Python та JavaScript. З цих зразків 440 445 (19,7 %) містили принаймні одне вигадане ім’я пакета [3].

У контексті SIEM/SOC галюцинації становлять ризик через необхідність аналізувати великі обсяги телеметрії, яка може бути неповною, суперечливою або позбавленою необхідного контексту. Специфіка кібербезпеки полягає в

тому, що системи безперервно працюють із такими даними, тоді як, якщо не вказано інше, штучний інтелект буде припускати, що працює з ідеальними даними, стабільною телеметрією та чистими вхідними даними. Проблема полягає в тому, що більшість збоїв у роботі SOC не відбуваються в ідеальних умовах [4]. Ситуацію також значно ускладнюють непослідовні підказки, які підвищують ризик виникнення галюцинацій [4]. У результаті в сфері кібербезпеки такі помилки створюють серйозні ризики: від введення аналітиків в оману до можливого зловживання з боку зловмисників [5]. Описуючи вплив на системи рівня SIEM, можна зазначити, що, наприклад, коли системи штучного інтелекту використовуються для узагальнення журналів та формування звітів про загрози, «галюцинації» можуть призвести до неправильної ідентифікації загроз безпеці. Це, у свою чергу, може спричинити застосування стратегій усунення неіснуючих вразливостей та загроз, що призведе до неконтрольованих помилкових спрацьовувань, марнування ресурсів і, можливо, невиправданої паніки [6].

Вплив галюцинацій штучного інтелекту на моніторинг та реагування проявляється в кількох основних категоріях ризиків. По-перше, хибнопозитивні висновки можуть спричиняти ідентифікацію неіснуючих загроз, зайві перевірки та невиправдане використання ресурсів. По-друге, хибнонегативні висновки можуть призводити до пропуску реальних атак або недооцінювання їхньої критичності. По-третє, модель може сформувати вигадані індикатори компрометації (IoC), неправдиві відомості про вразливості або некоректні зв'язки між подіями. Окремий ризик становлять рекомендації щодо реагування, які за відсутності належної перевірки можуть спричинити блокування легітимних ресурсів, порушення доступності сервісів або виконання необґрунтованих змін у конфігурації систем.

Вибір заходів мінімізації має визначатися характером ризику та можливими наслідками помилкового висновку. Для запобігання використанню вигаданих індикаторів компрометації доцільно перевіряти їх за авторитетними зовнішніми джерелами та даними внутрішньої телеметрії. Перевірка відомостей про вразливості має передбачати підтвердження відповідних ідентифікаторів CVE та умов експлуатації. Для зниження ризику пропуску дійсних загроз необхідно поєднувати результати генеративної моделі з незалежними механізмами виявлення й перевірки інцидентів. Небезпечні рекомендації щодо реагування потребують окремого контролю, зокрема обмеження автономності системи та підтвердження критичних дій відповідальним фахівцем.

Для мінімізації зазначених ризиків необхідний комплексний інженерний підхід до побудови та використання ШІ-систем. Наразі дослідження пропонують кілька напрямів для інтеграції в SIEM: генерацію з розширеним пошуком із використанням індексів у сфері безпеки, гібридні символьно-нейронні моделі, що включають перевірку CVE на основі правил та логічні обмеження послідовності використання вразливостей, а також генерацію з урахуванням невизначеності за допомогою конформного прогнозування,

калібрувальних моделей, механізмів утримання від відповіді та оцінки галюцинацій за допомогою метрик ентропії та самоузгодженості [5].

Одним із важливих напрямів є використання технології RAG, яка поєднує генеративні можливості мовних моделей із пошуком інформації у зовнішніх джерелах. Завдяки тому, що відповіді ґрунтуються на перевірених зовнішніх джерелах знань, ці системи можуть значно зменшити частоту виникнення галюцинацій [6]. Попри це, застосування RAG не усуває потреби перевіряти актуальність і достовірність знайдених матеріалів, оскільки якість результату залежить також безпосередньо від джерел, на яких ґрунтується відповідь.

Також існує обмеження результатів генерації за допомогою визначених правил і формалізованих вимог до змісту. Однією з перспективних методик є декодування на основі обмежень під час процесу генерації тексту. Завдяки застосуванню заздалегідь визначених правил та семантичних обмежень можна скерувати штучний інтелект на створення результатів, які краще відповідають перевіреним фактам [6]. Рішення допомагає обмежувати недопустимі результати, однак його необхідно поєднувати з перевіркою фактів і правил, що застосовуються у конкретній системі моніторингу.

Для управління поведінкою моделі в нестандартних ситуаціях, промтбуки мають містити чіткі структуровані інструкції, що визначають порядок дій за умов недостатності даних, суперечливості вхідних сигналів, а також критерії утримання від виконання дій і передачі завдання на вищий рівень опрацювання.

Технічні заходи необхідно доповнювати організаційним контролем у рамках ISMS. Інструменти ШІ мають допомагати фахівцям у прийнятті рішень, а не замінювати їх, тому код і конфігурації, згенеровані ШІ, повинні проходити перевірку людиною. Також необхідно перевіряти запропоновані моделями програмні залежності перед встановленням: підтверджувати існування пакета в цільовому реєстрі, дату його реєстрації та репутацію видавця. Організації мають визначати правила відповідального використання ШІ, допустимі сценарії його застосування та процедури перевірки результатів.

Результати застосованих методів мінімізації ризиків необхідно підтверджувати шляхом відповідного оцінювання. Для цього використовуються сукупність показників, зокрема частоту підтверджених галюцинацій, точність виявлення загроз, кількість хибнопозитивних і хибнонегативних висновків, обґрунтованість рекомендацій, частоту коректного утримання моделі від відповіді за недостатності доказів і кількість небезпечних дій, зупинених до виконання.

Це дає змогу визначити, чи знижують впроваджені механізми ризик хибних висновків і чи сприяють вони безпечнішому використанню автоматизації в процесах SOC.

Для забезпечення відтворюваності оцінювання необхідно визначити спосіб розрахунку кожного показника та порядок формування контрольної вибірки. Частоту галюцинацій доцільно визначати як відношення кількості відповідей із підтвердженими фактичними помилками до загальної кількості

перевірених відповідей. Частку хибнопозитивних висновків розраховують як відношення кількості хибнопозитивних результатів до кількості всіх фактично негативних випадків, а частку хибнонегативних висновків – як відношення кількості пропущених загроз до кількості всіх фактично позитивних випадків. Частоту небезпечних дій можна визначати як відношення кількості запропонованих або виконаних дій, що порушують установлені правила безпеки, до загальної кількості відповідних дій. Контрольна вибірка має містити приклади як із наявними загрозами, так і без них, а оцінювання до та після впровадження захисних механізмів необхідно проводити за однаковими критеріями.

Інтеграція штучного інтелекту в системи моніторингу кібербезпеки підвищує можливості аналізу загроз, однак галюцинації генеративних моделей можуть призводити до хибних висновків і небезпечних рішень. Мінімізація зазначених ризиків передбачає застосування різних методів контролю достовірності результатів роботи штучного інтелекту, обмеження генерації, регулювання рівня його автономності та залучення фахівців до прийняття критичних рішень.

Ефективність запропонованих заходів доцільно оцінювати за показниками достовірності відповідей, якості виявлення загроз і безпечності реагування в системах SIEM/SOC.

Список літератури

1. Поддубний В. О., Сєвєрінов О. В., Пустомельник О. С. Менеджмент вразливостей як складова частина політики безпеки ІТС. Системи управління, навігації та зв'язку. 2020. Т. 4, № 62. С. 55–58. DOI: https://doi.org/10.26906/SUNZ.2020.4.055.

2. CERT-EU. Generative AI in Cybersecurity: Balancing Innovation and Risk. 2025. URL: https://cert.europa.eu/publications/security-guidance/generative-ai-in-cybersecurity-balancing-innovation-and-risk/.

3. Cloud Security Alliance (CSA). Slopsquatting: AI Code Hallucinations Fuel Supply Chain Attacks. 2026. URL: https://labs.cloudsecurityalliance.org/wp-content/uploads/2026/04/CSA_research_note_slopsquatting-ai-supply-chain_20260419-csa-styled-1.pdf.

4. Capgemini. From AI-powered tools to an AI-powered SOC. 2026. URL: https://www.capgemini.com/wp-content/uploads/2026/08/EBook_AI-Powered-SOC-ipdf7_120826.pdf.

5. Ainsley W. J. H. Hallucinations at the Firewall. The Fortieth AAAI Conference on Artificial Intelligence. 2026. URL: https://www.researchgate.net/publication/402660549_Hallucinations_at_the_Firewall.

6. Njenga K., Madzinga R. M. AI Hallucinations in Information Security: A Bibliometric and Grounded Study Perspective. Journal of Information Security and Cybercrimes Research. 2025. URL:https://www.researchgate.net/publication/399572555_AI_Hallucinations_in_Information_Security_A_Bibliometric_and_Grounded_Study_Perspective.