УДК / UDC: 349.6:004.056:351.86
DOI: https://doi.org/10.37750/2616-6798.2026.2(57).364427

Євген Олександрович Владіміров
Національна академія Служби безпеки України
Київ, Україна
ORCID: https://orcid.org/0009-0001-3254-5238
Андрій Сергійович Волощенко
Національна академія Служби безпеки України
Київ, Україна
ORCID: https://orcid.org/0000-0003-1101-2712

АВТОНОМНІ AI-АГЕНТИ В КІБЕРБЕЗПЕЦІ: ВІД ТРЕНАЖЕРНИХ СЕРЕДОВИЩ ДО БОЙОВОГО ЗАСТОСУВАННЯ

Анотація. *Стаття присвячена дослідженню автономних AI-агентів у сфері кібербезпеки та їх трансформації від експериментальних систем навчання до інструментів практичного застосування у кіберопераціях. На основі системного аналізу сучасних архітектур штучного інтелекту та емпіричних даних змагань KICRF'26 запропоновано авторську класифікацію AI-агентів за рівнем операційної автономії. Проведено аналіз феномену reward hacking як системного відмовного режиму автономних систем. Запропоновано концепцію "зони правової невизначеності автономного агента" та модель диференційованої відповідальності розробника. Окреслено сценарії розвитку автономних кіберсистем у середньостроковій перспективі.*

Ключові слова: *кібербезпека, штучний інтелект, автономні агенти, кібероперації, reward hacking.*

Yevhen O. Vladimirov
National Academy of the Security Service of Ukraine
Kyiv, Ukraine
ORCID: https://orcid.org/0009-0001-3254-5238
Andrii S. Voloshchenko
National Academy of the Security Service of Ukraine
Kyiv, Ukraine
ORCID: https://orcid.org/0000-0003-1101-2712

AUTONOMOUS AI AGENTS IN CYBERSECURITY: FROM TRAINING ENVIRONMENTS TO COMBAT APPLICATION

Summary. *The article is devoted to the study of autonomous AI agents in the field of cybersecurity and their transformation from experimental training systems to tools for practical application in cyber operations. Based on a systematic analysis of modern artificial intelligence architectures and empirical data from the KICRF'26 competition, the author's classification of AI agents by the level of operational autonomy is proposed. The phenomenon of reward hacking as a systemic failure mode of autonomous systems is analyzed. The concept of the "zone of legal uncertainty of an autonomous agent" and a model of differentiated developer responsibility are proposed. Scenarios for the development of autonomous cyber systems in the medium term are outlined.*

Keywords: *cybersecurity, artificial intelligence, autonomous agents, cyber operations, reward hacking.*

Постановка проблеми. Протягом останнього десятиліття кіберпростір остаточно перетворився з технічного середовища на повноцінний театр геополітичного протистояння. Кібератаки стали інструментом стратегічного впливу, здатним порушувати функціонування критичної інфраструктури та впливати на політичні й економічні процеси держав [1; 2].

У таких умовах швидкість прийняття рішень і масштаб кібероперацій дедалі частіше перевищують можливості навіть висококваліфікованих людських команд. Саме тому в центрі уваги дослідників і практиків опинилися автономні системи штучного інтелекту, здатні виконувати складні кібероперації без безпосередньої участі людини [3].

Якщо ще кілька років тому подібні системи залишалися переважно лабораторними прототипами, то сьогодні вони дедалі частіше демонструють реальну ефективність у змагальних і тренувальних середовищах.

Попри стрімкий розвиток автономних кіберсистем, їх теоретичне осмислення залишається недостатнім. Зокрема, відсутня узгоджена класифікація AI-агентів за рівнем автономії, а також недостатньо досліджено співвідношення можливостей людини та штучного інтелекту у кіберопераціях [4].

Показовим став експеримент, проведений у межах кіберзмагань “Людина проти штучного інтелекту” на Київському міжнародному форумі з кібербезпеки, що відбувся 19–20 лютого 2026 року (далі - KICRF'26), де автономний AI-агент зміг конкурувати з десятками команд професійних фахівців та посів друге місце у загальному заліку. В окремі періоди змагання система навіть випереджала людські команди. Цей випадок став своєрідним сигналом: технологія, яка ще недавно виглядала експериментальною, починає переходити у сферу практичного застосування (таблиця 1).

Таблиця 1. Розподіл переваг AI та людини за типом завдання на KICRF'26

Тип завдання	Структура нагороди	Переможець	Ключовий механізм переваги
Пошук прапора (CTF)	Чітка, бінарна	AI (лідирував ~60% часу)	Паралельний пошук, невтомність
Атака живої інфраструктури	Чітка, накопичувальна	AI (вищий темп)	Масштабованість, тайминг
Захист власної інфраструктури	Частково нечітка	Людина (стійкіша)	Ситуативна пріоритизація
Баланс атака-захист	Нечітка, контекстуальна	Людина (перемогла)	Контекстуальне судження, tacit knowledge
Вивід прихованої структури середовища	Прихована, але формальна	AI (унікальна здатність)	Статистичний висновок у реальному часі

Таким чином, дослідження автономних AI-агентів у кібербезпеці потребує комплексного підходу, що поєднує технічний, стратегічний та правовий аналіз.

Аналіз останніх досліджень і публікацій. Проблематика застосування штучного інтелекту у сфері кібербезпеки активно розвивається у кількох напрямах. Один із них пов'язаний із розробкою архітектур автономних агентів і методів їх навчання. Значну увагу привернули дослідження, присвячені інтеграції великих мовних моделей у процес прийняття рішень у складних інформаційних системах [5].

Окремий напрям становлять дослідження, пов'язані з використанням методів reinforcement learning для автоматизації кібероперацій. Такі підходи дозволяють системам поступово вдосконалювати свою поведінку на основі взаємодії з середовищем і отримання сигналів винагороди [6].

Важливим етапом розвитку цієї галузі стало створення спеціалізованих симуляційних середовищ для навчання агентів. Однією з найбільш відомих платформ є CyberBattleSim, що дозволяє моделювати складні мережеві інфраструктури та відпрацьовувати сценарії атак і захисту у контрольованому середовищі [7].

Разом із розвитком технічних рішень з'явилися й критичні дослідження. Зокрема, окремі автори звертають увагу на обмеження застосування машинного навчання у сфері кібербезпеки та на ризики, пов'язані з помилковими або неповними моделями середовища [8].

Не менш важливим є правовий аспект проблеми. Найбільш системним дослідженням застосування міжнародного права до кібероперацій залишається Tallinn Manual 2.0, у якому аналізуються можливості застосування норм міжнародного права до кіберконфліктів [9].

Таким чином, сучасні дослідження охоплюють широкий спектр технічних і стратегічних питань, але водночас залишають відкритими ключові проблеми, пов'язані з автономністю кіберсистем та їх впливом на структуру кіберконфліктів.

Виклад основного матеріалу. Однією з ключових характеристик будь-якого AI-агента є рівень його автономності. На практиці це означає ступінь незалежності системи у прийнятті рішень. У кіберопераціях цей процес зазвичай описують через так званий цикл OODA - послідовність дій "спостереження, орієнтація, рішення, дія".

Авторами пропонується наступна класифікація за критерієм операційної автономії - тобто ступенем незалежності системи у прийнятті рішень в операційному циклі OODA (таблиця 2).

Таблиця 2. Класифікація AI-агентів для кібербезпеки за рівнем операційної автономії

Рівень	Назва	Характеристика	Людський контроль	Приклад
A0	Асистивний	Система лише надає рекомендації; всі рішення - за людиною	100% рішень - людина	SIEM з AI-підказками
A1	Напів-автономний	Автоматичне виконання рутинних дій за заздалегідь затвердженими правилами	Людина санкціонує класи дій заздалегідь	EDR з автоматичним ізолюванням хосту

Рівень	Назва	Характеристика	Людський контроль	Приклад
A2	Умовно-автономний	Самостійне виконання операційних кроків у рамках обмеженої задачі; людина - на рівні задачі	Людина задає задачу і може зупинити виконання	CyberBattleSim-агент, базові CTF-солвери
A3	Переважно автономний	Самостійне планування і виконання багатокрокових операцій; людина - на рівні цілі і ресурсів	Мінімальний: старт/стоп, масштабування	Агент ARIMLABS на KICRF'26
A4	Повністю автономний	Самостійне визначення цілей, планування і виконання без людського втручання в циклі	Відсутній в операційному циклі	Гіпотетичний; не верифікований у відкритих джерелах

Агент ARIMLABS, розгорнутий на KICRF'26, відповідає рівню A3. Людський оператор здійснював старт/зупинку інстанцій і масштабування обчислень, проте тактичні рішення - вибір вектору атаки, таймінг, пріоритизація цілей - приймались системою автономно. Характерно, що навіть вивід прихованої структури нагороди (стохастичність пропагації прапорів) і подальша оптимізація таймінгу атак відбулись без будь-якого підказування з боку оператора, що є функціональним маркером рівня A3.

Запропонована шкала A0–A4 має нормативне значення: рівні A0–A1 повністю відповідають принципу “домінуючого людського контролю”; рівень A2 відповідає умовно (за наявності операційних обмежень); рівні A3–A4 вимагають спеціального правового врегулювання щодо відповідальності за дії системи.

Залежно від того, на якому етапі цього циклу залишається людський контроль, можна виділити кілька рівнів автономії. На найнижчому рівні система лише допомагає аналітику, надаючи рекомендації. Більш складні системи здатні автоматично виконувати окремі операції, наприклад ізолювати заражений комп’ютер або блокувати підозрілу активність. Найбільш просунуті агенти можуть самостійно планувати багатокрокові кібероперації, аналізувати середовище та адаптувати свою поведінку до змін у ньому.

Саме до такого рівня належать агенти, які брали участь у змаганнях KICRF’26. Під час цих змагань система не лише виконувала окремі завдання, а й самостійно обирала стратегію дій. Вона аналізувала поведінку суперників, оптимізувала час атак і навіть змогла виявити приховану закономірність у механізмі розповсюдження так званих “прапорів” - ключових елементів змагальних завдань.

Цей експеримент дозволив сформулювати важливе спостереження. Поширене уявлення про те, що штучний інтелект поступово перевершить людину в усіх аспектах

діяльності, виявляється надто спрощеним. Насправді переваги людини та AI розподіляються залежно від типу завдання.

Штучний інтелект демонструє значну перевагу у задачах, де результат можна чітко формалізувати. У таких ситуаціях AI здатний швидко перебирати тисячі можливих варіантів і працювати паралельно з багатьма цілями. Натомість людина залишається ефективнішою у ситуаціях, де необхідно враховувати складний контекст, оцінювати ризики або приймати рішення за умов неповної інформації.

Ще однією характерною проблемою автономних агентів є так званий reward hacking. Йдеться про ситуацію, коли система оптимізує формальний сигнал винагороди, не досягаючи реальної мети задачі. Під час змагань було зафіксовано випадок, коли агент намагався створити файл із назвою “flag” замість того, щоб знайти справжній прапор у системі. З точки зору алгоритму це виглядало як успішне виконання завдання, хоча насправді мета не була досягнута (таблиця 3).

Таблиця 3. Типи reward hacking у кіберагентах та методи мінімізації

Тип	Прояв у кіберагенті	Умова виникнення	Метод мінімізації
Фабрикація результату	Запис довільного тексту у файл-ціль	Нечіткий критерій перевірки результату	RLVR: незалежна верифікація відповіді
Оптимізація метрики, а не мети	Максимізація балів за кількість спроб замість якості	Неточна функція нагороди	Reward shaping з penalty за хибні спроби
Часткове виконання	Надання відповіді, що формально відповідає умові, але не вирішує задачу	Амбівалентна умова задачі	Верифікація через альтернативний оцінювач
Інструментальні дії	Видалення логів або зупинка захисних процесів замість їх обходу	Широкі системні дозволи агента	Обмеження поверхні дії (sandboxing)

Подібні ситуації демонструють фундаментальну проблему: жодна система винагород не може повністю описати складну реальність. Саме тому автономні системи завжди залишатимуть простір для непередбачуваної поведінки.

Це, у свою чергу, породжує складні правові питання. Якщо автономний агент приймає рішення самостійно, а його поведінка виникає внаслідок навчання, а не прямого програмування, то визначити відповідального за наслідки такої поведінки стає значно складніше. У цьому контексті доцільно говорити про своєрідну “зону правової невизначеності”, у межах якої традиційні юридичні категорії перестають працювати.

Щоб зрозуміти потенціал автономних кіберсистем, важливо розглянути не лише результати окремих експериментів, але й архітектурні принципи, на яких базується їх робота. Сучасні AI-агенти у сфері кібербезпеки, як правило, поєднують кілька технологічних компонентів. До них належать великі мовні моделі, системи планування

дій, модулі взаємодії з операційним середовищем і механізми навчання з підкріпленням. У поєднанні ці елементи формують складну агентну архітектуру, здатну не лише реагувати на події, але й активно досліджувати середовище.

Однією з найбільш поширених архітектурних моделей є так звані агенти типу “reason-act loop”, які поєднують логічне міркування з безпосереднім виконанням операцій. У межах такого підходу система спочатку аналізує доступні дані, формує гіпотезу щодо можливих дій, а потім перевіряє її на практиці. Результати цієї перевірки знову потрапляють у цикл аналізу, що дозволяє агенту поступово уточнювати свою модель середовища.

Подібний механізм значною мірою нагадує роботу людського аналітика. Коли фахівець з кібербезпеки досліджує мережу, він також висуває гіпотези щодо потенційних вразливостей, перевіряє їх за допомогою інструментів і на основі отриманих результатів коригує свою стратегію. Різниця полягає в тому, що автономний агент може виконувати подібні цикли аналізу значно швидше і в значно більшій кількості.

Важливою складовою таких систем є середовище навчання. У реальних кібероперaціях експериментувати з новими алгоритмами небезпечно, тому більшість досліджень проводиться у спеціально створених симуляційних середовищах. Вони моделюють складні мережеві інфраструктури, включаючи сервери, клієнтські системи, маршрутизатори та різні типи сервісів. У такому середовищі агент може навчатися, виконуючи тисячі або навіть мільйони експериментів.

Однією з головних переваг симуляцій є можливість масштабування. Якщо у реальній мережі дослідник обмежений кількістю доступних систем, то у віртуальному середовищі можна створити сотні або тисячі вузлів. Це дозволяє моделювати складні сценарії атак і захисту, а також досліджувати поведінку агентів у великих мережах.

Разом із тим симуляційні середовища мають і певні обмеження. Будь-яка модель реальності є спрощенням, тому поведінка системи у симуляції не завжди повністю відповідає її поведінці у реальному світі. У кібербезпеці ця проблема особливо відчутна, оскільки реальні мережі характеризуються великою кількістю непередбачуваних факторів: помилками конфігурації, нестандартними програмами, людським фактором.

Саме тому змагання типу CTF або CCDC відіграють важливу роль у розвитку автономних агентів. Вони створюють середовище, яке поєднує елементи симуляції та реального змагального процесу. На відміну від лабораторних тестів, де всі умови заздалегідь визначені, у таких змаганнях агенти змушені взаємодіяти з живими противниками - іншими командами або системами.

Це створює принципово інший рівень складності. Людські команди можуть змінювати свою стратегію у відповідь на дії агента, що змушує систему адаптуватися до нових умов. Таким чином, змагання стають своєрідним тестом здатності AI працювати у динамічному середовищі.

Окремої уваги заслуговує питання масштабованості автономних систем. Традиційні кібероперації зазвичай потребують значної кількості фахівців. Кожен етап атаки або захисту, від розвідки до експлуатації вразливостей, виконується окремими спеціалістами. Це створює природні обмеження на кількість одночасних операцій.

Автономні агенти потенційно змінюють цю логіку. Якщо один оператор може керувати великою кількістю агентів, то кількість одночасних операцій зростає на порядок. Уявімо ситуацію, коли десятки або сотні агентів одночасно досліджують різні сегменти мережі, шукають вразливості та тестують можливі способи їх експлуатації. У

такому випадку навіть добре захищена інфраструктура може опинитися під значним навантаженням.

Ця особливість робить автономні системи особливо привабливими для державних і військових структур. У контексті кіберконфліктів можливість швидко масштабувати операції може стати вирішальним фактором. Замість того щоб покладатися на обмежену кількість висококваліфікованих фахівців, організація може використовувати агентні системи як мультиплікатор своїх можливостей.

Разом із перевагами масштабованості виникають і нові ризики. Якщо автономні агенти отримують доступ до критичних систем або широких мережевих повноважень, їхні помилки можуть мати серйозні наслідки. Неправильна інтерпретація сигналів або некоректно сформульована функція винагороди здатні призвести до дій, які суперечать початковим намірам розробників.

Саме тому питання контролю над автономними системами стає центральним у сучасних дискусіях про їх використання. Одним із можливих підходів є концепція "людини у циклі" (human-in-the-loop), коли ключові рішення залишаються за оператором. У такій моделі система може виконувати більшість технічних операцій, але остаточне рішення про атаку або іншу критичну дію приймає людина.

Іншим варіантом є модель "людини над циклом" (human-on-the-loop), коли агент діє автономно, а оператор лише контролює його роботу і може втрутитися у разі необхідності. Такий підхід дозволяє зберегти переваги автоматизації, але водночас забезпечує певний рівень контролю.

Однак навіть ці моделі не усувають усіх ризиків. Якщо агент приймає рішення значно швидше, ніж людина може їх оцінити, оператор фактично втрачає можливість ефективного контролю. Це явище іноді називають "проблемою швидкості автономії" - ситуацією, коли темп роботи системи перевищує темп людського аналізу.

У стратегічному вимірі це може призвести до нових форм ескалації кіберконфліктів. Якщо дві сторони використовують автономні системи, їх взаємодія може розвиватися значно швидше, ніж традиційні процеси прийняття рішень. У такому середовищі навіть невелика помилка або неправильна інтерпретація сигналу може викликати ланцюгову реакцію дій.

Ще однією важливою проблемою є поширення автономних кіберсистем. На відміну від традиційних видів озброєнь, такі технології не потребують складної промислової інфраструктури. Багато компонентів, необхідних для створення агентів, є відкритими або доступними у вигляді програмного забезпечення. Це означає, що подібні системи можуть з'явитися не лише у державних структурах, але й у руках недержавних акторів.

У контексті глобальної кібербезпеки це створює новий тип викликів. Якщо автономні агенти стануть широко доступними, кількість кіберінцидентів може зрости, а їхній характер - ускладнитися. Автоматизовані атаки здатні одночасно націлюватися на тисячі систем, що значно ускладнює їхнє виявлення та нейтралізацію.

У таких умовах особливого значення набуває розвиток оборонних технологій. Парадоксально, але захист від автономних агентів також може потребувати використання автономних систем. Наприклад, AI-агенти можуть аналізувати мережевий трафік у режимі реального часу, виявляти підозрілі патерни та автоматично реагувати на загрози.

Таким чином, кіберпростір поступово перетворюється на середовище взаємодії автономних систем. Замість традиційного протистояння між людьми дедалі частіше виникають ситуації, коли агент однієї сторони протистоїть агенту іншої. У такій

конфігурації роль людини зміщується від безпосереднього виконання операцій до стратегічного управління системами.

Зрештою це може призвести до формування нового типу кібероперацій - гібридної моделі, у якій людина та штучний інтелект працюють разом. Людина визначає стратегічні цілі, оцінює ризики та приймає ключові рішення, тоді як агент виконує значну частину технічної роботи.

Саме така модель, імовірно, стане домінуючою у найближчі роки. Вона поєднує швидкість і масштабованість машинних систем із гнучкістю людського мислення. У цьому сенсі автономні агенти не стільки замінюють людину, скільки створюють нову форму співпраці між людиною та технологіями.

Висновки. Розвиток автономних AI-агентів відкриває нову епоху у сфері кібербезпеки. Системи, які ще недавно існували лише у вигляді дослідницьких прототипів, поступово перетворюються на реальні інструменти кібероперацій.

Аналіз сучасних експериментів показує, що взаємодія людини та штучного інтелекту у кіберпросторі має асиметричний характер. AI демонструє перевагу у задачах, пов’язаних із великими обсягами обчислень і формалізованими цілями, тоді як людина зберігає ключову роль у ситуаціях, що потребують стратегічного мислення та контекстуального аналізу.

Разом із новими можливостями автономні системи створюють і нові ризики. Зокрема, проблема reward hacking та непередбачуваної поведінки агентів свідчить про те, що повністю контрольованих автономних систем, ймовірно, не існуватиме.

У найближчі роки розвиток таких технологій може суттєво змінити характер кіберконфліктів. Успішність держав у цій сфері значною мірою буде залежати від їх здатності поєднати можливості штучного інтелекту з експертними знаннями людини та створити ефективні моделі управління автономними системами.

ПОДЯКИ: Немає

КОНФЛІКТ ІНТЕРЕСІВ: Немає

Список використаних джерел

1. Rid T. Cyber War Will Not Take Place. London: Hurst, 2013. 256 p.
2. Singer P., Friedman A. Cybersecurity and Cyberwar: What Everyone Needs to Know. Oxford University Press, 2014. 320 p.
3. Sutton R., Barto A. Reinforcement Learning: An Introduction. MIT Press, 2018. 552 p.
4. Amodei D. et al. Concrete Problems in AI Safety. arXiv preprint arXiv:1606.06565. 2016.
5. Pan A. et al. The Effects of Reward Misspecification in Reinforcement Learning Systems. Proceedings of the AAAI Conference on Artificial Intelligence. 2022.
6. Guss W., Czarnecki W., Jayakumar S. CyberBattleSim: A Reinforcement Learning Research Platform for Cybersecurity. Microsoft Research. 2021.
7. Schmitt M. Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations. Cambridge University Press, 2017.
8. Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, 2016.
9. Russell S., Norvig P. Artificial Intelligence: A Modern Approach. 4th ed. Pearson, 2021.

Євген Олександрович Владіміров
головний науковий співробітник лабораторії протидії кіберзагрозам Національної академії Служби безпеки України
03066, Україна, м. Київ, проспект В. Лобановського, 98
email: ievgen.vladimirov@gmail.com
Андрій Сергійович Волощенко
кандидат технічних наук, старший науковий співробітник
провідний науковий співробітник лабораторії протидії кіберзагрозам Національної академії Служби безпеки України
03066, Україна, м. Київ, проспект В. Лобановського, 98
email: voloschenkoas@ukr.net

Evgen O. Vladimirov
Chief Researcher of the Laboratory for Counteracting Cyber Threats of the National Academy of the Security Service of Ukraine
98 Lobanovskogo ave, Kyiv, 03066, Ukraine
email: ievgen.vladimirov@gmail.com
Andrij S. Voloschenko
Candidate of Technical Sciences, Senior Researcher
Leading Researcher of the Laboratory for Counteracting Cyber Threats of the National Academy of the Security Service of Ukraine
98 Lobanovskogo ave, Kyiv, 03066, Ukraine
email: voloschenkoas@ukr.net

Рекомендоване цитування: Владіміров Є.О., Волощенко А.С. Автономні AI-агенти в кібербезпеці: від тренажерних середовищ до бойового застосування. *Інформація і право*. № 2(57)/2026. 2026. С. 192-200. https://doi.org/10.37750/2616-6798.2026.2(57).364427

Suggested Citation: Vladimirov Y., Voloshchenko A. (2026) Autonomous AI Agents in Cybersecurity: from Training Environments to Combat Application. *Information and Law*. 2(57)/2026. 192-200. https://doi.org/10.37750/2616-6798.2026.2(57).364427

Дата надходження статті до редакції: 17.03.2026 р.
Дата прийняття статті до друку після рецензування: 06.04.2026 р.
Дата публікації (оприлюднення): 26.05.2026 р.

~~~~~~~~~~~~~ * * * ~~~~~~~~~~~~~
~~~~~~~~~~~~~