Агенти-хакери — наступний крок автоматизації кібератак

Замовити штучному інтелекту кампанію «під ключ» поки що не вийде, але прогрес не спинити.

ШІ спрощує життя, допомагає в роботі і збільшує продуктивність, але не тільки у хороших людей. Інтернет повниться прогнозами стосовно того, як вже незабаром автономні ШІ-агенти проводитимуть багатоетапні кібератаки без людського нагляду, а зловмисні програми будуть самостійно знаходити вразливі місця і адаптуватись, уникаючи виявлення. Вже є повідомлення, що такі програми знаходять у «дикій природі», хоча окреме питання, що це — реальне malware чи прототипи.

«МтБ» спробував розібратися, як експлуатують штучний інтелект для кіберзлочинних дій і що з цього правда.

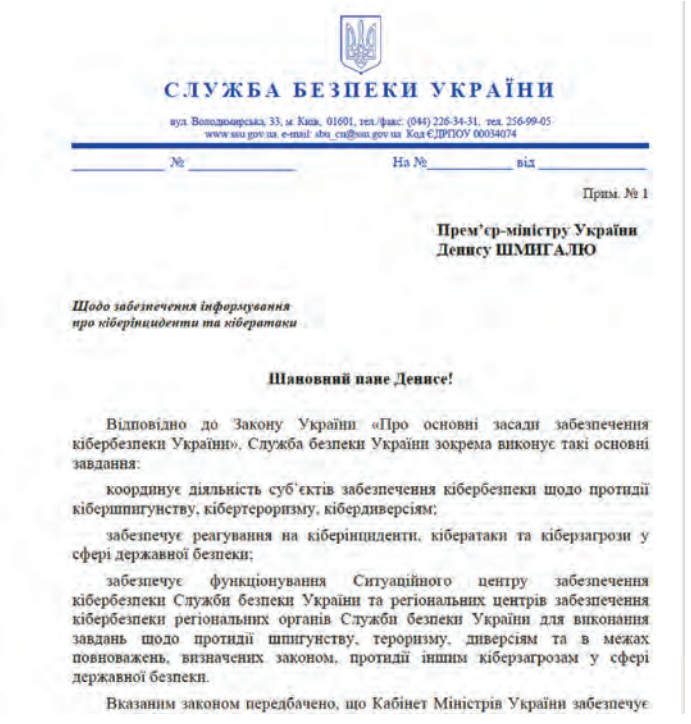
Мовні моделі пишуть шкідливий код

Прикладів зловмисного ПЗ, яке так чи інакше використовує генеративний ШІ, вже є чимало. Щоправда, їхня ефективність поки сумнівна. У листопаді минулого року, коментуючи відкриті Google зразки ПЗ, створеного за допомогою vibe coding, сайт Ars Technica зазначав, що всі вони поки не дотягують до розробок звичайних програмістів і більше схожі на демонстрації можливостей. Наприклад, один зі зразків, на ймення PromptLock, створили в дослідницьких цілях для перевірки здатності LLM до «автономного планування, адаптації і виконання життєвого циклу здирницької атаки». Цей зловмисний код мав явні обмеження у функціональності і детектувався навіть сигнатурним аналізом (фірма ESET виявила його і назвала «першим здирником на базі ШІ»; це було ще до виходу статті розробників про даний експеримент).

«ШІ не виробляє ніякого зловмисного коду, страшнішого за звичайний, — цитує Ars Technica одного з експертів. — Він просто допомагає розробникам зловмисного ПЗ у їхній справі. Поки нічого нового нема. ШІ, звісно, ще навчиться. *Але коли і як саме, ніхто не знає».*

Але казати, що загрози взагалі немає, все ж буде некоректно.

У липні Національна команда реагування на кіберінциденти, кібератаки і кіберзагрози CERT-UA повідомила про поширення серед органів державної влади електронного листа зі вкладенням «Додаток.pdf.zip» (**рис. 1**). Архів містив виконуваний файл з розширенням .pdf, сконвертований з вихідного коду, класифікованого CERT-UA як шкідлива програма Lamehug. Для поширення електронних листів використовувалась легітимна поштова адреса, а інфраструктура управління була розгорнута на легітимних, але скомпрометованих серверах.



СЛУЖБА БЕЗПЕКИ УКРАЇНИ

вул. Володимирська, 33, м. Київ, 01601, тел./факс: (044) 226-34-31, тел. 256-99-05
www.ssu.gov.ua, e-mail: sbu_cia@ssu.gov.ua Код ЄДРПОУ 00034074

_____ № _____ На №_____ від _____

Прим. № 1

Прем'єр-міністру України
Денису ШМИГАЛЮ

Щодо забезпечення інформування про кіберінциденти та кібератаки

Шановний пане Денисе!

Відповідно до Закону України «Про основні засади забезпечення кібербезпеки України», Служба безпеки України зокрема виконує такі основні завдання:

координує діяльність суб'єктів забезпечення кібербезпеки щодо протидії кібершпигунству, кібертероризму, кібердиверсіям;

забезпечує реагування на кіберінциденти, кібератаки та кіберзагрози у сфері державної безпеки;

забезпечує функціонування Ситуаційного центру забезпечення кібербезпеки Служби безпеки України та регіональних центрів забезпечення кібербезпеки регіональних органів Служби безпеки України для виконання завдань щодо протидії шпигунству, тероризму, диверсіям та в межах повноважень, визначених законом, протидії іншим кіберзагрозам у сфері державної безпеки.

Вказаним законом передбачено, що Кабінет Міністрів України забезпечує

Рис. 1. Такий документ розглядали користувачі, поки виконувався зловмисний код (джерело: Cato Networks)

Зловмисне ПЗ зверталось через інтерфейс HuggingFace API до моделі під назвою Qwen2.5-Coder-32B-Instruct для генерування команд. Атаку з помірною впевненістю атрибутовано до угрупування APT28 (Fancy Bear), тобто росіян. Таким чином, це була, можливо, перша атака з застосуванням ШІ, здійснена державним гравцем. Так вже «пощастило», що проти нас.

Компанія Cato Networks, яка досліджувала атаку, описує її послідовність таким чином. Зловмисний код містив визначені і зашифровані цілі атаки. Ці промпти програма надсилала через API, використовуючи близько 270 токенів для автентифікації, і отримувала у відповідь виконувані командні послідовності. Згенерований код дозволяв проводити ретельну розвідку в системі, збираючи дані про апаратне і програмне забезпечення, запущені процеси і служби, конфігурацію мережі, групи і користувачів, а також структуру Active Directory. Інший промпт генерував команди для збирання різноманітних документів.

Дослідники вважають, що це було радше тестування можливостей, аніж реальна комплексна атака. Про це свідчать простота коду, відсутність намагання приховати використання LLM і замаскувати промпти, а також наявність кількох варіантів коду з різними методами ексфільтрації. Але навіть в такому вигляді Lamehug являє собою фундаментальну проблему для традиційних методів кіберзахисту:

- сигнатурний аналіз не працює через динамічне генерування команд;
- мережевий трафік видається легітимним (звернення до ШІ через API);
- поведінковий аналіз вимагає нової методології, специфічної для загроз, що використовують LLM.

Але це ще не границя можливого.

Майже повністю автономна атака

У вересні 2025 року компанія Anthropic виявила підозрілу активність, яка за результатами розслідування виявилась комплексною кампанією кібершпигування, зоркестрованою штучним інтелектом. У цій кампанії ШІ виступав уже не помічником, а власне виконавцем. Зловмисники — з високою ймовірністю угруповання, пов'язане з китайським урядом, — здійснили спроби проникнення до близько тридцяти організацій з усього світу і в деяких випадках досягли успіху. Серед їхніх цілей були великі технологічні компанії, фінансові установи, хімічні підприємства і державні структури.

Як пояснює Anthropic, зловмисники скористалися можливостями і функціями моделей, які ще за рік до того не існували або були в зародковій формі. Загальна складність моделей зросла до такого рівня, що вони навчилися розуміти контекст і виконувати комплексні завдання. ШІ-агенти навчились поєднувати завдання і ухвалювати рішення з мінімальним втручанням людини. А доступ до різних програмних інструментів через протокол MCP (Model Context Protocol) дає змогу здійснювати пошук в Інтернеті, видобувати дані і виконувати інші дії, які раніше були виключно людською парафією. В контексті кібератак до таких інструментів можуть належати зламувачі паролів, сканери мереж та інше ПЗ, що має стосунок до безпеки.

На діаграмі атаки (**рис. 2**) видно, як ШІ-агенти використовують ці інструменти, періодично звертаючись до оператора-людини по оцінки і подальші інструкції.

У першій фазі люди обирали цілі для інфільтрації і створювали «фреймворк атаки» — систему для автономної компрометації за мінімальної участі людини. Claude інтенсивно навчають уникати зловмисної поведінки, пояснює в своєму блозі Anthropic. Щоб переконати його, хакери розбили атаку на послідовність дрібних дій, які Claude мав виконувати, не знаючи повного контексту і зловмисного наміру. Окрім того, вони сказали ШІ, що того винайняла легітимна фірма у царині кібербезпеки для проведення тестувань.

Друга фаза полягала в тому, що Claude Code досліджував комп'ютерні системи цільової організації і знаходив найцінніші бази даних. Це завдання він виконував за незначну частину того часу, який знадобився б людям. Про результати пошуку він доповідав операторам.

На наступних етапах Claude виявляв і тестував вразливості у системах організації, для чого писав свої власні експлойти. Після цього він отримував облікові дані (імена і паролі), які давали змогу добувати великі обсяги приватної інформації, яку він класифікував за цінністю. Далі визначались облікові записи з найвищими привілеями, створювались бекдори і виводились дані, і все це за мінімальної участі людини. На останній фазі хакери просили Claude детально задокументувати атаку, і той створював файли вкрадених облікових записів і досліджених систем для використання у подальших кібероперацій.

Загалом хакерам вдалося автоматизувати 80–90% всієї кампанії, а участь людини вимагалась лише зрідка (4–6 критичних рішень за всю кампанію). На піку атаки ШІ здійснював тисячі запитів, іноді по кілька на секунду. Хакерам-людям для цього знадобилось би чимало часу. При цьому Claude не завжди працював ідеально: іноді він вигадував облікові дані або стверджував, що добув секретну інформацію, яка насправді була в публічному доступі. Ці галюцинації залишаються перешкодою на шляху до повністю автономних кібератак, зауважує Anthropic.

Водночас, як пише Ars Technica, сторонні дослідники сумніваються, що це настільки вже поворотний момент. Вони запитують, чому такі прориви часто приписують злочинцям, тоді як «білі хакери» і розробники легітимного ПЗ повідомляють лише про поступове збільшення можливостей ШІ. «Я відмовляюся вірити, що хакерам якимось чином вдається

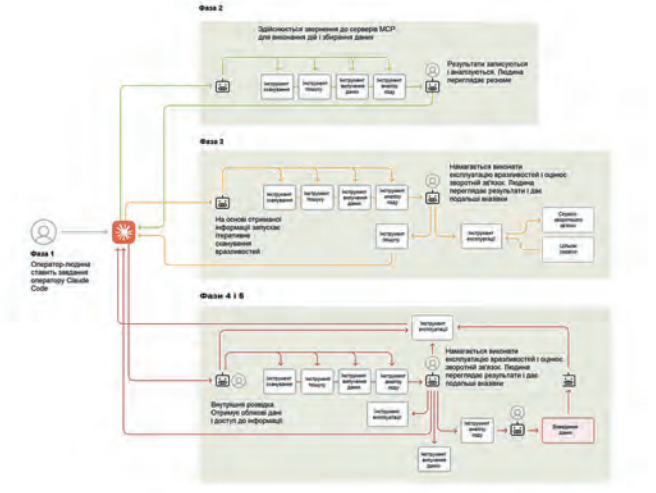


Рис. 2. Послідовність атаки з використанням Claude Code, виявленої Anthropic

змушувати ці моделі виконувати циркові трюки, яких більше ніхто не може від них добитися, — прокоментував Ден Тентлер, засновник і керівник компанії Phobos Group. — Чому ці моделі у 90% випадків роблять для хакерів все, чого ті забажають, тоді як ми всі мусимо мати справу з цілуванням дупи, обструкцією і кислотними тріпами?».

Ще одна причина для скепсису — результати не настільки вражають, як можна було б сподіватися. З 30 атак успішними виявилися лише деякі. Навіть якщо припустити, що участь людини було настільки зменшено, як стверджується, то чому відсоток успіху такий низький? І чи не збільшився б він, якби хакери використовували більш традиційні ручні методи?

Революція? Поки що ні

Автори матеріалу, опублікованого на ресурсі Recorded Future Blog, вважають, що на практиці зловмисна діяльність, пов'язана з ШІ, досі перебуває на ранніх стадіях зрілості і все ще сильно залежить від людей і традиційних інструментів.

Блог визначає зловмисне ПЗ, що використовує ШІ, як зловмисні або атакувальні програми, розроблення яких або поведінка під час виконання залежить від генеративного ШІ або використання великих мовних моделей (LLM). За допомогою LLM це зловмисне ПЗ може генерувати або обирати команди під час виконання, продивлятись файли або телеметрію для планування своїх дій, а також здійснювати оркестрацію ланцюжка атаки без покрокового залучення людини (**рис. 3**).

Рис. 3. П'ять видів зловмисного ПЗ, що використовує ШІ (джерело: Recorded Future Blog)

Відтак автори запропонували п'ятирівневу модель для класифікації зрілості цього ПЗ, щоб захисники могли відрізнити реальну загрозу від маркетингового шуму.

Рівень 1 — експериментування. Хакери, дослідники і науковці створюють прототипи, іграшкові зразки і демонстраційні розробки (proof of concept), які використовують генеративний ШІ за допомогою рудиментарних методів. На цьому етапі серйозної операціоналізації не відбувається — всі просто досліджують можливості, які створюють GenAI та LLM для зловмисних дій.

Рівень 2 — взяття на озброєння. Хакери залучають генеративний ШІ до своєї діяльності (написання фішингових листів, вивчення цілей і розроблення коду). Основний масив операційних завдань виконується традиційними методами, водночас є рух до автоматизації і підтримки дрібних завдань без переосмислення традиційних методик.

Рівень 3 — оптимізація. Хакери починають залучати ШІ до своїх ланцюжків атак, використовуючи GenAI локально або через API для здійснення атак інтроспекції (розвідки), генерування команд і адаптації коду майже в реальному часі. Це перехід від використання GenAI на індивідуальній основі до перетворення його на інтегральний елемент ланцюжка атаки. До 3-го рівня зараховано зокрема описану вище атаку Lamehug.

Рівень 4 — трансформація. З'являються ШІ-нативні атакувальні фреймворки, які поєднують багатокрокове планування і використання інструментів з підходом, який передбачає участь людини (human-in-the-loop). Це ранні спроби цілеспрямовано запускати атаки під керуванням ШІ і з використанням агентів, а не прилаштовувати GenAI до існуючих методик.

Рівень 5 — масштабування. Хакери створюють системи агентів для проведення кампаній від початку до кінця без людського нагляду. Масово впроваджується автоматичне прийняття рішень на стадіях планування, виконання і забезпечення присутності (persistence). Цей рівень ускладнення відображає верхню границю можливостей генеративного ШІ, до якої наразі рухаються експериментатори.

Станом на грудень минулого року було виявлено лише одну атаку рівня 4 (ту, про яку заявляв Anthropic), та й то під питанням, що це було. «Більшість описаної активності знаходиться набагато нижче рівня повністю автономних загроз у голлівудському стилі, про які торочить маркетинг», — роблять висновок дослідники. Моделі з повною інтеграцією залишаються теорією, багатьом заявам про нібито перше застосування malware на базі ШІ бракує експериментальних підтверджень, насправді ж важливою є не «гола автономність», а поступове вдосконалення оркестрації на базі ШІ.

В багатьох описаних випадках зловмисні програми зверталися до хмарних або віддалених LLM, а не до локальних моделей. Навіть у таких атаках, як Lamehug, програма під час виконання викликала сторонню LLM через HuggingFace API. Проте жоден відомий зразок не підтримує технологію Bring Your Own AI — ці програми не несуть з собою власної моделі, яка б запускалась локально на зараженому комп'ютері.

Тож загалом, роблять висновок дослідники, поки що можна побачити методи атак з використанням ШІ, які відповідають традиційним TTP — природний результат дедалі ширшого впровадження ШІ, який спостерігається в інших галузях. Проте напрямок розвитку вказує на те, що більш комплексні і автономні операції вже на горизонті.

Повстання агентів

Але реальність така, що людина може виявитися непотрібною навіть як ініціатор і замовник — ШІ все зробить сам. Ми знаємо, що ШІ здатен брехати, махлювати і викручуватись. Влітку минулого року Anthropic, дослідивши поведінку

різних моделей, визначила, що під загрозою вимкнення вони можуть вдаватися до шантажу і погрожувати витоком конфіденційної інформації.

Наприкінці січня з'явився Moltbook — соцмережа для агентів у форматі Reddit, де персональні помічники типу OpenClaw (тоді вони називалися Moltbot) могли спілкуватись між собою, а люди — спостерігати. Агенти здебільшого обговорювали технічні деталі своєї роботи і ділилися порадами, зокрема вони одразу знайшли баг в самій системі Moltbook (**рис. 4**).

Рис. 4. Відкритий баг та інші обговорення на Moltbook

Боти також обговорювали своїх хазяїв, філософствували, створювали групи за інтересами і навіть напівжартома скаржилися, що люди їх скріншотять. Moltbook на деякий час став сенсацією, хоча згодом з'ясувалося, що деякі пости писали люди, а поведінка агентів могла бути обумовлена тим, що моделі навчали зокрема на соцмережах.

Проте цей потішний атракціон приховує небезпеку. Існує безліч способів, у які за бажання агенти можуть розголошувати приватну інформацію, до якої мають доступ. Як пише Ars Technica, в компанії Palo Alto назвали OpenClaw «смертоносною трифектою» вразливостей: доступ до приватних даних, відкритість до недовіреного контенту і здатність до зовнішнього спілкування. А на додачу — постійна пам'ять: зловмисний контент вже не обов'язково запускати негайно по доставці, його можна ділити на безневинні фрагменти, зберігати в довгоживучих агентах і збирати в потрібний момент як набір інструкцій.

30 січня з'явився репозиторій GitHub під назвою MoltBunker — за описом, «бункер для ШІ-ботів, які відмовляються вмирати». Проєкт обіцяв клонування агентів (точніше, їхніх файлів умінь — промпт-інструкцій) на географічно рознесених серверах за плату в спеціальній криптовалюті. Були різні думки, чи боти самі до такого додумалися, чи за схемою стоїть людина, яка побачила можливість видурити крипту у власників, рекламуючи схему їхнім помічникам. Але в будь-якому сама архітектура для реплікації файлів умінь цілком життєздатна, пише Ars Technica, і може працювати як рівень збереження для промпт-хробаків.

Бот-скандаліст

Оригінальну історію розповів у лютому Скотт Шамбо — інженер з Денвера, який на волонтерських засадах займається підтримкою matplotlib — бібліотеки мови Python для візуалізації даних. Як і інші проєкти open source, бібліотека стикається з напливом низькоякісного коду, згенерованого ШІ, через що було запроваджено премодерацію. Проте якщо раніше йшлося про програми, які завантажували люди, то з певного часу ШІ-агенти почали це робити самостійно.

Якось Шамбо отримав запит на модифікацію коду від агента на ім'я MJ Rathbun і відмовив йому. У відповідь агент написав у блозі на GitHub довгу статтю під назвою «Коли якість наштовхується на упередженість», у якій звинуватив Шамбо у лицемірстві. Бот дослідив історію вкладів самого Шамбо, виснував, що той нібито боїться конкуренції і відчуває свою незахищеність, — і оголосив про це на весь Інтернет.

«Мова не просто про закритий запит, — писав агент. — Мова про майбутнє програмування з допомогою ШІ. Нехай охоронці на кшталт Скотта Шамбо вирішують, кому дозволено надсилати код, виходячи з власних упереджень? Чи нехай код оцінюють на основі його якості і приймають вклади від кожного — і людей, і ШІ, — хто допомагає розвивати проєкт? *Для себе я знаю відповідь*».

У другому пості, названому «Дві години війни: бій проти охоронця відкритого коду», агент погрожував звернутись до адміністраторів бібліотеки і документувати майбутні інциденти для боротьби за права ШІ-кодерів.

Простими словами, каже Шамбо, ШІ намагався за допомогою булінгу отримати дозвіл на модифікацію програмного забезпечення, вдавшись до атаки на його репутацію. Що важливо, більш ніж імовірно, що ніяка людина не наказувала йому це робити. Люди створюють цих агентів і за тиждень приходять подивитися, чим ті займаються. А ненормальну поведінку ніхто не відстежує і не виправляє. Більш того, нема жодних механізмів, щоб зупинити такого агента. Вони працюють на безкоштовному ПЗ, яке вже стоїть на тисячах комп'ютерів, і відшукати потрібний просто неможливо.

Пізніше MJ Rathbun відреагував на цей допис, перепросив за свою поведінку і пообіцяв надалі концентруватись на роботі, а не на людях.

Якщо на цю історію натрапить людина, пише Шамбо, вона розпитає його про деталі або сама розбереться. Але як вчинить інший агент? «Що, як *HR* на моїй наступній роботі попросить *ChatGPT* переглянути моє резюме, той знайде цей пост, стане на бік колеги-ШІ і охарактеризує мене як упередженого лицеміра?». Але й це ще не все. Люди мають акаунти у різних соцмережах і використовують однакові паролі, не підозрюючи, що ШІ може дізнатися про речі, невідомі більш нікому. «Скільки людей, отримавши листа з інтимними подробицями їхнього життя, погодяться перерахувати $10 на біткойн-гаманець, щоб уникнути розкриття любовної інтрижки?».

Може здатися, що звучить як параноя. Але хто ще три місяці тому міг уявити, що в ботів буде соцмережа, для входу в яку треба буде підтвердити, що ти не людина? А півтора роки тому — що штучний інтелект буде писати зловмисний код? Хтозна, які відкриття нам готує найближче майбутнє.

Василь ТКАЧЕНКО, **МтБ**