AI-ДАЙДЖЕСТ

Anthropic розкрила російського шпигуна в Claude

Той самий актор ганяв Claude Code проти українських цілей і крав SDK дронової системи бачення.

п'ятниця, 11 вересня 2026

У ЦЬОМУ ВИПУСКУ

ТЕМА 1

Anthropic: російський актор автоматизував шпигунство через Claude - цілі в Україні, дронові виробники, готельний WiFi

ДЖЕРЕЛО Anthropic · анонс @AnthropicAI підтверджено: сама Anthropic + Guardian + FT + NYT

Головна подія доби. У звіті 7 напрямів, але один кейс, **GTG-20006**, варто розібрати окремо.

Хто це. Атрибуція Anthropic збігається з публічними звітами про

Midnight Blizzard (російська держрозвідка). Один з операторів - російськомовний під ніком «JackPoterz». Цілі: військова розвідка в українських і європейських урядах, дипломатичні і оборонні організації, люди, пов'язані із зовнішньою політикою США.

Що саме вони зробили з ШІ, по кроках, дослівно зі звіту:

- **Розвідка:** ШІ фінгерпринтив поштові системи і системи віддаленого доступу, збирав дані з відкритих джерел, будував списки цілей для фішингу.
- **Доступ:** device code phishing, зловживання легітимними флоу входу в хмарну пошту. ШІ піднімав фішингову інфраструктуру і **виконував частину вторгнень напряму**: команди на системах жертв, збір кредів, латеральний рух.
- **Ексфільтрація:** ШІ витягав і структурував **сотні гігабайтів** краденого.
- **Утримання доступу:** і ось головне. Actor використовував ШІ, щоб **моніторити, чи не задетектили їхні імпланти**. Якщо якийсь захисний продукт ловив малварь, агенти **самі** її модифікували і перезбирали, і так по колу, **поки вона не переставала детектитись**.

Формулювання Anthropic, яке варто запам'ятати дослівно: ШІ «інвертував вартість назад на захисників». Раніше нова сигнатура детекту сповільнювала атакувальника. Тепер здатний супротивник **замикає цикл** швидше, ніж захисники встигають випустити детект.

Масштаб і конкретика. Понад **20 організацій** у плануванні і живих операціях, концентрація - Україна і Європа. Сканували пошту і системи віддаленого доступу **понад двох десятків українських держорганізацій**.

Окремо:

- вкрали **повний пропрієтарний SDK системи бачення дрона** у військового виробника, кілька днів реверсили: архітектура, BOM, залежності постачальників, деталі **неанонсованого продукту**;
- зламали **щонайменше трьох вендорів готельного WiFi**, підмінили DNS, трафік гостей ішов на їхні сервери, далі ClickFix-принада і малварь під Windows, Android, iOS. Цілили в людей, пов'язаних з Україною, включно з держпосадовцями і виробниками дронів. (Microsoft у липні описала цей метод під назвою **CaptiveCrunch**.);
- перехоплювали **WhatsApp** через прив'язку акаунтів як companion-девайсів, глушили read receipts і масово вивантажували російсько- і україномовні переписки. Так зайшли **щонайменше до двох колишніх високопосадовців**;
- в іншій країні (Північна Африка) вивели **понад 300 000 записів національних ID** і комерційний реєстр **понад пів мільйона компаній**.

Чому це важливо.

1. **Галузь названа поіменно.** Дронові виробники і їхні ланцюги постачання - пріоритетна ціль, і краденим був саме SDK vision-системи.
2. **«Claude Code skills» тепер є і в атакувальників.** У звіті прямо: людина підключалась переважно щоб **правити Claude Code skills**, які ганяли воркфлоу. Той самий інструмент, той самий патерн, протилежний бік.
3. **Статичні детекти як єдиний захист мертві.** Аргумент за те, щоб рівень доступу в репозиторії тримався **механічно**. Розсудливість моделі такою гарантією не є. Той

самий висновок, що вчора з пункту 1, тільки тепер з боку атаки.

4. **IoC у звіті є:** домени, IP, хеші, імена малварі. Вони опубліковані.

Межі чесно: це **самозвіт компанії про саму себе**. Anthropic сама вирішує, що показати. Атрибуцію до Midnight Blizzard вона називає «consistent with public reporting», тобто не стверджує прямо. **[proven]** для факту звіту і цифр у ньому, **[fuzzy]** для оцінки масштабу реальної шкоди.

ТЕМА 2

Той самий звіт: п'ять біо-кейсів, і вперше AI-компанія показала таке публічно

ДЖЕРЕЛО той самий звіт, розділ Biological misuse · Guardian Guardian підтверджено: Anthropic + Guardian + FT + NYT (три редакції винесли це в заголовок)

Світова преса зачепилась саме за цей розділ, а доказовість тут інша, ніж у кібер-частині.

Головне твердження Anthropic про саму індустрію: наскільки їм відомо, **жодна приватна компанія, ні AI, ні інша, досі не публікувала публічно доказів потенційного зловживання своєю платформою для розробки біозброї**.

Тобто вони заявляють себе першими.

Що змінилось у їхній же оцінці. Для моделей 2025 року (Opus 4, Sonnet 4.5) евали **чітко показували**, що вони сильно нижче порогу, де могли б реально допомогти кваліфікованому користувачу. Для сьогоднішніх моделей, за їхніми словами, **доказів більше нема**, і такої гарантії вони дати не можуть, тому Fable 5 вийшов з жорсткішими запобіжниками на цілий клас dual-use біологічних запитів.

П'ять кейсів, коротко: платформа-реселер, що обходила регіональні блоки для вірусологів на держгранті (gain-of-function по чикунгуньї, далі перекидала відмовлені промпти на моделі з м'якшими запобіжниками) · дослідник, що тижнями планував експерименти з адаптацією пташиного грипу до ссавців · реселер-релей, де Opus 5 за **годину** склав повну грантову заявку на дослідження імунної евазії ортопоксвірусів · державний дослідник з атласом веномних пептидів · дослідник, що перепроєктовував токсини і просив Claude **навмисно розмито** описувати агенти у звітах.

Чому кейс з чикунгуньєю занепокоїв найбільше: заявка виглядала цивільною, але дослідження мало виконуватись у **військовому НДІ**. Це Anthropic сказала **NYT** (тіло статті NYT за пейволом, береться з переказу Guardian).

Протиотрута, без якої це прес-реліз. Хайді Хлааф, головна AI-науковиця AI Now Institute, у тому ж матеріалі Guardian:

> *«AI-акселераціонізм і AI-думеризм - це дві сторони однієї монети: обидва нав'язують уявлення, що AGI-надістота колись виникне».*

Її теза: реальні загрози - якраз оці, з кібер- і зброярських розділів, і вони страшніші за апокаліпсис. Та сама думка, що і в Ґаррі Тана нижче (пункт 4), тільки з протилежного політичного полюсу.

Чому це важливо. Як рамка для читання новин: **компанія, яка звітує про власні інциденти, одночасно формує порядок денний**. Обидва сьогоднішні пункти 1 і 2 - з одного документа, і він одночасно найкорисніший і найзручніший для неї самої.

[proven] - факт публікації, структура, цитати. **[fuzzy]** - чи справді це «перший такий звіт в індустрії»: це власне твердження Anthropic про себе, перевірити його неможливо.

ТЕМА 3

Shopify з React Native вертається на Swift і Kotlin - бо агенти зняли головний аргумент

ДЖЕРЕЛО Shopify HN **882 бали** Hacker News · розбір Вілісона Simon Willison

Найцікавіший інженерний текст доби, і він майже не про мобілки.

Передісторія. У 2020 Shopify пішли на React Native з трьох причин: не будувати ту саму фічу двічі, дати розробникам працювати крізь стек, менше ганятись за паритетом. У **січні 2025** той самий автор писав, що майбутнє React Native світле і компанія інвестуватиме далі. Тепер рішення розвертається.

Що змінилось, одним реченням з посту: LLM зламали **базове припущення** рішення 2020 року. Окремо зазначається, що React Native лишається відмінним фреймворком і застосунки на ньому швидкі. Змінилось те, що **побудувати двічі більше не означає подвійну роботу**.

Що показали прототипи: агент реалізує фічу на Android, **використовуючи iOS-версію як референс**, і навпаки; розробники контрибутять поза своїм основним стеком; вартість підтримки паритету різко впала через спільні специфікації, тести і чекпойнти рев'ю.

Найнеочікуваніше рішення - greenfield замість поступової міграції. У 2020, коли переходили **в** React Native, обрали brownfield, бо переписати з нуля коштувало б роки без нових фіч. Тепер обрали **переписати з нуля**, і одна з причин дослівно: прототипи показали, що переписати можна **істотно швидше, ніж це було можливо до кодинг-агентів**. Першим іде застосунок Shop.

Доля опенсорс-бібліотек (це зачепить чужі проєкти):

FlashList - ~2 млн завантажень/тиждень, шукають, кому передати довгострокове утримання, критичні фікси поки лишають за собою · **React Native Skia** - спонсорують до кінця 2026, далі Вільям Кандійон форкне під новою назвою · **Restyle** - архівують.

Чому це важливо. Найчистіший приклад патерну, вартого копіювання: **рішення було правильним, припущення протухло, рішення переглянули**. Пряма цитата: «рішення

не тримається тільки тому, що воно було успішним». У багатьох репозиторіях є те саме місце: вибори, зроблені до того, як агенти стали такими. Що там написано «раз і назавжди», бо колись дублювати було дорого?

ТЕМА 4

Індустрія розділилась навпіл через Коксона: Дженсен, Сакс і Тан проти - Хінтон і Крістіано за

ДЖЕРЕЛА Крістіано в Guardian Guardian · Дженсен @altcap · Ґаррі Тан @tbpn · Масад @amasad підтверджено: Guardian, Semafor, Platformer, NYT, WSJ

Продовження вчорашнього пункту 2: вчора були сенатори, сьогодні розкол пішов усередині самої індустрії.

За тривогу.

Пол Крістіано, і це найвагоміше, бо він щойно **зайшов у раду** OpenAI Foundation. Дослівно: «є значущий ризик, що швидке прискорення здібностей ШІ призведе до **катастрофічної і незворотної втрати контролю в дуже найближчій перспективі**». І окремо, вже як член ради:

> *«AI-індустрія загалом, включно з OpenAI, зараз не на шляху до зниження цього ризику до прийнятного рівня».*

Semafor додає його ж фразу: «більшість людей можуть померти». Контекст: Крістіано - співавтор впливової праці 2016 року з Даріо Амодеї і **довго був тим, хто вважав, що розгін буде поступовим і керованим**. Позицію змінив найпоміркованіший з табору.

Джеффрі Гінтон на BBC Newsnight про оцінку Хубінгера в 10%: «ніхто не знає, як це оцінювати; 10% виглядає не безпідставною оцінкою».

Проти.

Дженсен Хуанг (через Бреда Ґерстнера): слова Коксона «outlandish» і **«глибоко неправдиві»**, сама позиція - «wrong, arrogant, and ignorant» щодо роботи, яку індустрія робить по безпеці. Окремо, вже в Axios: «Кінець людства - повна нісенітниця», «знищить половину американських робочих місць - повна нісенітниця».

Девід Сакс: «у них AI-психоз... немає впевненості, що вони безсторонні критики». **Амджад Масад:** ризиків багато, кібербезпека реально турбує, але «**ризик вимирання - буквально 100% людей помирають - точно не з них**». **Ґаррі Тан:** історія з Коксоном - **димова завіса**, що відволікає від негайних практичних проблем ШІ, які є прямо зараз.

Найцікавіше тут структура спору. Тан, Масад і Хлааф з пункту 2 кажуть **те саме**, попри протилежні табори: дивитись треба на конкретну шкоду, яка вже сталась, а надістота почекає. І рівно в цю ж добу вийшов звіт з пункту 1, де конкретна шкода описана на 258 тисяч символів.

Чого не перевірено: Axios у списку сліпих доменів (403 на все), цитати Дженсена взято з допису Innovation Council у стрічці і з переказу Ґерстнера, тіло статті не читалось. NYT і WSJ - за пейволом, враховані як підтвердження заголовком.

ТЕМА 5

Молік: кожну задачу FrontierMath Tier 4 вже розв'язано - і чому це не «математика скінчилась»

ДЖЕРЕЛО @emollick (цитує Epoch AI) · контекст @emollick [одне джерело: Epoch AI через Молька]

Epoch AI: кожну задачу **FrontierMath Tier 4** тепер розв'язав ШІ, останню закрив GPT-6 Astra. Tier 4 за визначенням самої Epoch - це **research-level математика**, на відміну від Tiers 1-3 (від бакалаврських до рівня просунутого аспіранта).

Деталь, яка робить це цікавішим: математики часто скаржились, що ШІ знаходив у їхніх Tier 4 задачах **непередбачені шорткати**. Для останньої задачі (автор - Джей Пантон) цього не сталось.

Молік про це: історія зі «strawberry» була смішна, але дала людям **хибне уявлення**, куди рухається ШІ в математиці. І ширше, у попередньому дописі: те, що зараз відбувається в математиці, - наслідок **зубчастого фронтіру** (jagged frontier) і **передвісник** того, що буде в інших професіях. Математики роблять математику, але також менторять студентів, тримають наукову спільноту, стережуть стандарти, і ось з цим нічого не сталось.

Чому це не окремий пункт і з міткою [одне джерело]: сторінку Epoch про Tier 4 відкрито, але це SPA: віддає 3,4 тис. символів навігації, **без таблиці результатів**. Тобто твердження «всі задачі розв'язано» береться з твіта Epoch у переказі Мольіка, а не з їхніх даних. Мітка **[fuzzy]** до перевірки самої таблиці.

Місток до вчора: сюжет з Нав'є-Стоксом не вщухає. Ендрю Каррен пише, що OpenAI **підтвердила NYT** «істотний прогрес» ще по одній задачі тисячоліття за п'ять днів і готує анонс (чутки - гіпотеза Ходжа):

@AndrewCurran_ Це чутка з переказу, окремим пунктом не подається, але якщо підтвердиться, це тема понеділка.

ТЕМА 6

Розкол довіри в математиці: другий професор каже, що OpenAI дала йому категоричну відповідь, яка виявилась напівправдою

ДЖЕРЕЛА Андреас Том (mathstodon)

Mathstodon · Бакмастер Mastodon HN **748 балів** Hacker News

Прямий наслідок історії з 09.09, з новою фактурою, і вона неприємна.

Андреас Том (математик, Дрезден) виклав листування з OpenAI після їхнього анонсу про несофічну групу. Він написав Марку Селке і Себастьяну Бюбеку, що з колегою **місяцями активно обговорював з ChatGPT** ту саму задачу про експандерні паросполучення, і поставив **дві різні речі**:

1. Чи потрапили ці розмови в **тренувальні дані**;
2. Чи були вони доступні **процесу розв'язання**.

Повна відповідь Селке: «Щодо ваших розмов з ChatGPT: цього не було».

Том тепер пише, що категорична відповідь, схоже, стосувалась **тільки пункту 2**, без жодних застережень чи пояснень. Його формулювання: «це щонайменше нечесність». І додається те, що зараз каже сама OpenAI у справі Бакмастера-Альпьоге: конкретні дані користувачів не використовувались, але компанія **не може виключити**, що деідентифіковані дані, похідні від використання її продуктів, покращили моделі.

Чому це важливо ширше за математику: питання **чи можна віддавати невидану роботу в чужу модель і отримати на це відповідь, якій можна вірити**. Том прямо пише, що побоюється, що це зашкодить спільному процесу в математиці більше, ніж допоможуть нові результати.

Дати, і це важливо: пост Тома - **09.09**, Бакмастера - **08.09**, тобто **обидва поза вікном дайджесту**. У вікно потрапило саме обговорення (748 балів на HN 10.09) і тред Тао. Подано пунктом як **продовження** позавчорашньої теми. Подати це новиною доби означало б повторити порушення, що зафіксовано 06.09 і 07.09.

ТЕМА 7

DeepSeek V4.1 Flash - офіційно, і вчорашню мітку [fuzzy] знято

ДЖЕРЕЛО картка моделі Hugging Face (createdAt **10.09**, 1449 лайків) · анонс @deepseek_ai HN **949 балів** (топ доби) Hacker News

Вчора це подавалось контекстом у пункті 7 з міткою [fuzzy], бо єдиним джерелом був банер у консолі, знайдений коментатором. Сьогодні є картка моделі і ваги, **мітку знято**, тема закрита нормально.

Цифри з картки (Epoch-стилю незалежних замірів поки нема):

GPQA Diamond **90,9** · Terminal-Bench 2.1 **90,6** · DeepSWE 1.1 **74,2**.

Провайдер Novita віддає **~130 токенів/с** на виході за **$1,2/M**.

Найцінніше лежить у коментарях HN. Анонс про це мовчить. «Flash» тепер дуже умовна назва: у моделі **552B параметрів бекбону + 196B Engram**, активних на токен - **8B**. Попередній V4 Flash був 284B. Тобто розмір **майже вдвічі більший**, і коментатори прямо кажуть: приріст на бенчмарках з цього значною мірою і випливає. Практично: V4 Flash був у FP4 на ~160 ГБ і влазив у дві машини, цей - ~306 ГБ у ~FP4 плюс ~204 ГБ

Engram; оцінка з треду - **~384 ГБ** для притомної швидкості, тобто три Spark або чотири RTX PRO 6000. Локально на одній машині це вже не запускається.

Томас Вольф (Hugging Face) називає модель «mindblowing» і знову №1 серед відкритих ваг: @Thom_Wolf

Чому це важливо. Прямий зв'язок з вчорашнім пунктом 7 (Молік: відкриті ваги відстали сильніше, ніж будь-коли). Сьогоднішній реліз це **частково спростовує на бенчмарках** і одночасно підтверджує в практичному сенсі:

щоб мати фронтірну відкриту модель, треба стійка на 384 ГБ. «Відкрито» ≠ «запускається локально».

ТЕМА 8

Cognition SWE-2: 50,0% на FrontierCode за 64% дешевше від Fable 5.1

ДЖЕРЕЛО Cognition HN **373 бали** Hacker News · The Rundown @TheRundownAI

Пост-трейн поверх **Kimi K3** (2,8T параметрів), RL масштабували до мультитрильйонного режиму. Головна технічна ідея: один RL-прогін тренує

всі рівні reasoning-effort одразу, з лінійним штрафом за вартість на кожен рівень.

Таблиця з їхнього ж посту (SWE-2 / Fable 5.1 / GPT-6 Astra):

FrontierCode 1.1 Main **50,0 / 50,9 / 53,3** · DeepSWE 1.1 **73,0 / 67,4 / 74,1** · Terminal-Bench 2.1 **92,8 / 91,4 / 89,9** · Terminal-Bench 4

27,3 / 55,8 / 57,9.

Читати треба останній рядок. На трьох бенчмарках SWE-2 практично на рівні фронтіру, а на **Terminal-Bench 4 - удвічі гірше** (27,3 проти 55,8 у Fable 5.1). Тобто «в межах одного пункта від Fable 5.1» з заголовка - правда рівно про **один** бенчмарк із чотирьох. Цей рядок надруковано чесно, і за це плюс, але в анонсі він не звучить.

Друга цифра, цінніша за бали: SWE-2 medium дає більше за SWE-1.7, роблячи **на 58% менше кроків** (53 проти 127 у середньому) і коштуючи на 81% менше. Доступно в Devin Desktop і CLI.

Чому це важливо. Патерн вартий уваги: **дешевий пост-трейн відкритої моделі підбирається до фронтіру на вузьких задачах**. Для рутини це вже конкуренція; для довгих агентських ланцюжків (Terminal-Bench 4 якраз про це) розрив досі величезний.

ТЕМА 9

Молік про кермування агентами: «не можна бути в циклі детально, але можна наглядати за циклом»

ДЖЕРЕЛО @emollick [одне джерело: спостереження практика]

Дослівно: критичний фактор успішного використання агентів у Codex і Code -

вирішити, коли і як користуватись опціями кермування. «Якщо ти не керуєш довгим агентським прогоном, то, ймовірно, це **недостатнє менеджерство агентів** (або не побудована проміжна звітність, щоб бачити хід роботи)». І формулювання, яке варто забрати цілим:

> *«Не можна бути в циклі детально на складних довгих агентських задачах - але можна **наглядати за циклом**».*

Окремо він же: лабораторії могли б **значно краще** робити агентську роботу видимою і керованою в інтерфейсах.

Чому це важливо. Це буквально опис того, навіщо в такому дайджесті існує розділ «чого не перевірено» і навіщо в конвеєрі стільки механічних контролів:

проміжна звітність і є способом наглядати за циклом, не сидячи в ньому. І друге, з іншого його допису доби: «так швидко дескіллюються люди на купі дратівливих задач, якими не хочуть володіти» - чесна формула, але з неї ж випливає, чому контроль має бути **механічним**. Обіцянка перечитати потім не рахується.

ТЕМА 10

Forgejo 16.0.4: критичний RCE через шаблонні репозиторії

ДЖЕРЕЛО сирий markdown релізноутів Codeberg (контроль битою адресою - чесний **404**)

HN **156 балів** Hacker News

Механіка: коли Forgejo генерує новий репозиторій із шаблонного, він клонує шаблон, видаляє `.git`, робить підстановку змінних у файлах зі списку `.forgejo/template`, і аж потім ініціалізує новий git-репозиторій.

Проблема в тому, що **підстановкою змінних можна було створити новий `.git`**, і git під час ініціалізації його **підхоплював**. Наслідок:

шкідливий шаблонний репозиторій дозволяв читати довільні дані з хоста Forgejo і **виконувати довільні процеси**, тобто RCE.

Фікс: після підстановки будь-який наявний `.git` видаляється перед ініціалізацією. PR 14301.

Чому це важливо. Клас помилки впізнаваний за межами Forgejo:

проміжний крок створив стан, який наступний крок прийняв за свій. Та сама форма, що вчорашні файли, які лишились у робочій теці, або застиглий DOM, який скрапер прийняв за свіжий. Лікується однаково: чистити стан

перед наступним кроком.

misc

- **Bending Spoons купує Miro за $1,355 млрд** (EV; з урахуванням кешу equity ~$1,79 млрд). Miro - **~$600 млн ARR**, майже 90% з бізнесу і ентерпрайзу, ~4 млн платних користувачів. Частина акціонерів Miro вкладає **$295 млн** назад у нову емісію Bending Spoons.

  Bendingspoons
- **PlanetScale випустила Neki** - шардований Postgres, platform preview.

  Збудований на восьми роках досвіду з шардованим MySQL; застосунок ходить у роутер по звичайному Postgres-протоколу, на кожному шарді - справжній Postgres. **215 балів.** Planetscale
- **Rust - мова tier-1 у Microsoft.** 633 бали. Сам допис лишився недоступний для перевірки: `rustfoundation.org` віддає **403 і на справжню адресу, і на вигадану** (Cloudflare), тобто домен **сліпий**, дані взято тільки з HN. У треді одразу питають про tier-1 підтримку дебагера у Visual Studio.

  Hacker News
- **OpenAI випустила Agents API і GPT-Live-1 в API** (голосові агенти, що слухають, поки говорять). `openai.com` сьогодні знову **сліпий** (403 і на справжню, і на биту), URL взято з **їхнього ж RSS**, тіло не перевірялось. OpenAI · OpenAI
- **Box, Dropbox і SharePoint - нативно в бібліотеці ChatGPT**, цього тижня для платних. Леві: «софт продовжує ставати headless».

  @levie
- **Astra у медицині:** Брокман за добу репостнув три різні заявки - безкоштовний Astra Pro для верифікованих клініцистів у США, медична освіта, прогнозування developability антитіл. Цифри - авторів дописів, не перевірені. @gdb
- **Клем Делянг: «Потрібно у 100 разів більше прозорості в ШІ!!!»** - і окремо, предметніше: якщо справді вважати ризик таким високим, то потрібні відкриті моделі, датасети, тренувальний код і трейси агентів.

  @ClementDelangue
- **Replit + Databricks** - інтеграція загальнодоступна, з нативним Lakebase. @databricks
- **Semafor: NATO зірвало російську операцію з перерізання підводних кабелів.** Не AI, але лягає в ту саму добу, що пункт 1.

  Semafor