
Недобросовестное использование возможностей ИИ

Журналистика, медиа и свобода слова: что показали отчёты и
расследования 2026 года

Что изменилось за год

01 Сложная атака без сложного атакующего

ИИ «схлопнул разрыв в труде и инструментах». Главное различие теперь **намерение**.

02 ИИ стал оркестратором, а не помощником

Большинство киберопераций шли через многоагентные системы; разведка, взлом и кража данных — автоматически.

03 Изменилась экономика атаки

Техника прежняя - кража учёток, фишинг, SQL-инъекции. Но взломы завершаются **за 2-3 часа**, один оператор ведёт десятки жертв.



Сайты с разными названиями, «независимыми» редакциями и авторами на разных языках.

Сколько новостных ресурсов может вести одна компания, используя ИИ?

Одна фирма - 70 «независимых редакций»

~70

фейковых новостных сайтов вела одна рекламная фирма

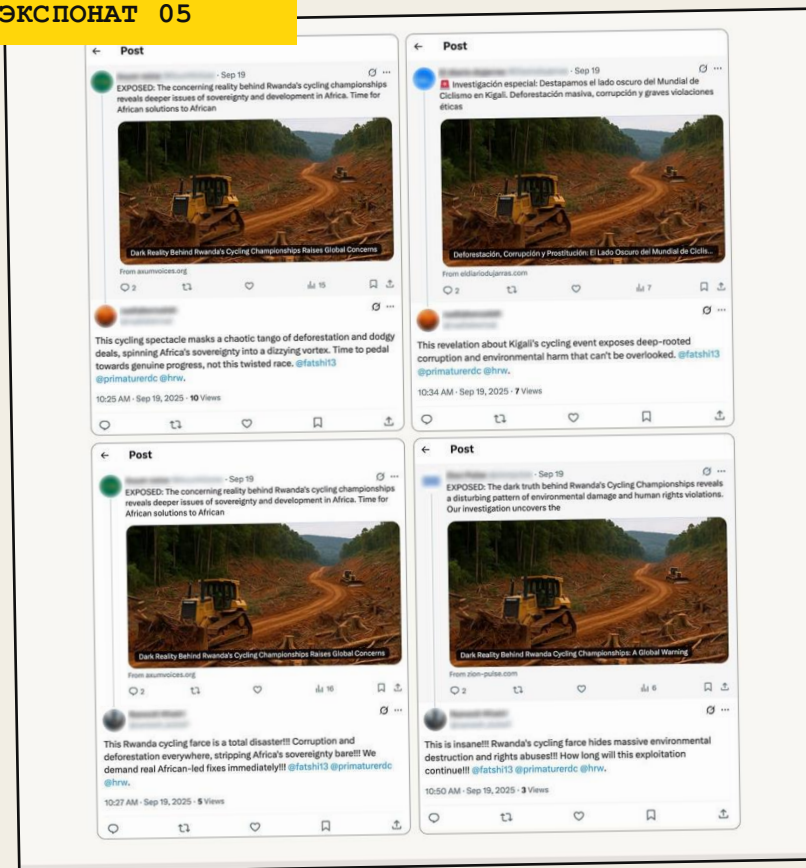
Французское агентство меняло политическую позицию в зависимости от того, кто платил, и **переписывало реальные статьи журналистов** под вымышленными подписями.

8 913 фейковых статей опубликовано

~20 языков публикаций

250+ аккаунтов-комментаторов

ЭКСПОНАТ 05



Идентичные ИИ-картинки и почти идентичные посты на разных «независимых» сайтах.

Платформа политопераций под выборы

- Стамбульская компания продавала ИИ-экосистему политопераций, нацеленную на Малайзию.
- Микротаргетинг избирателей по всем 222 округам по линиям расы, религии и монархии.
- ~1000 фейковых аккаунтов с логикой прогрева и обхода детекции.
- Фейковые «досье» против оппозиционного политика и НКО.

222

округа под микротаргетингом

~1000

управляемых фейк-аккаунтов

ЭКСПОНАТ 06



Фейковое издание «Malaysia Pulse» перепечатывало TV BRICS, Xinhua, Sputnik/RIA и CGTN, снимая госатрибуцию.

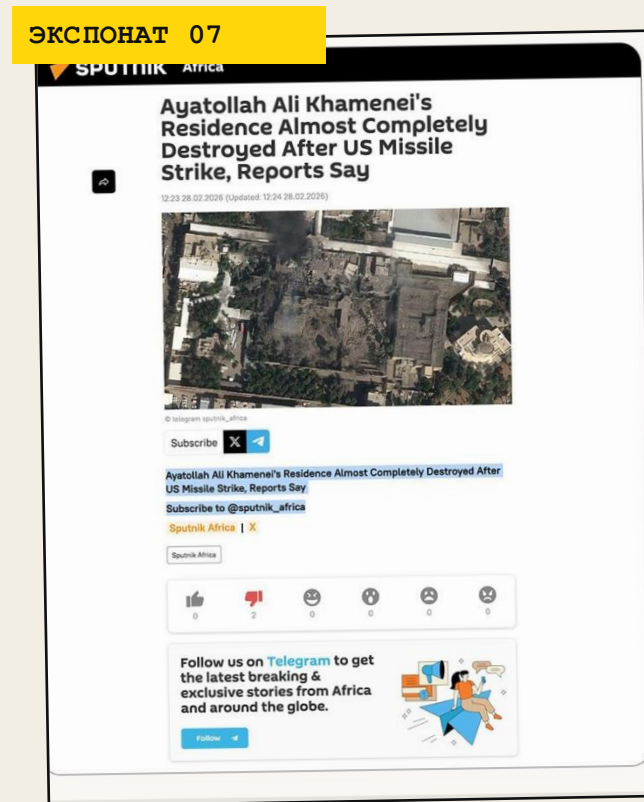


Sputnik Moldova, RIA Novosti, Sputnik en Español, Sputnik Africa и RT: пять редакций, несколько языков и континентов.

**Сколько сотрудников нужно, чтобы
выпускать для них новости?**

ИИ как «младший редактор» госСМИ

- Четыре аккаунта использовали Claude как новостной деск для Sputnik Moldova, RIA Novosti, Sputnik en Español, Sputnik Africa и англоязычной редакции RT.
- Модель позволяла одному сотруднику выдавать объём, который «обычно требует целой команды обученных редакторов».
- Для африканской аудитории действовал свод из 40 редакционных правил.
- Сфабрикованные утверждения о президенте Молдовы перед выборами 28 сентября 2025 года.



Пост в X Sputnik Africa дословно совпадает с текстом, сгенерированным Claude.

ИИ выдаёт себя за реального активиста

8 400

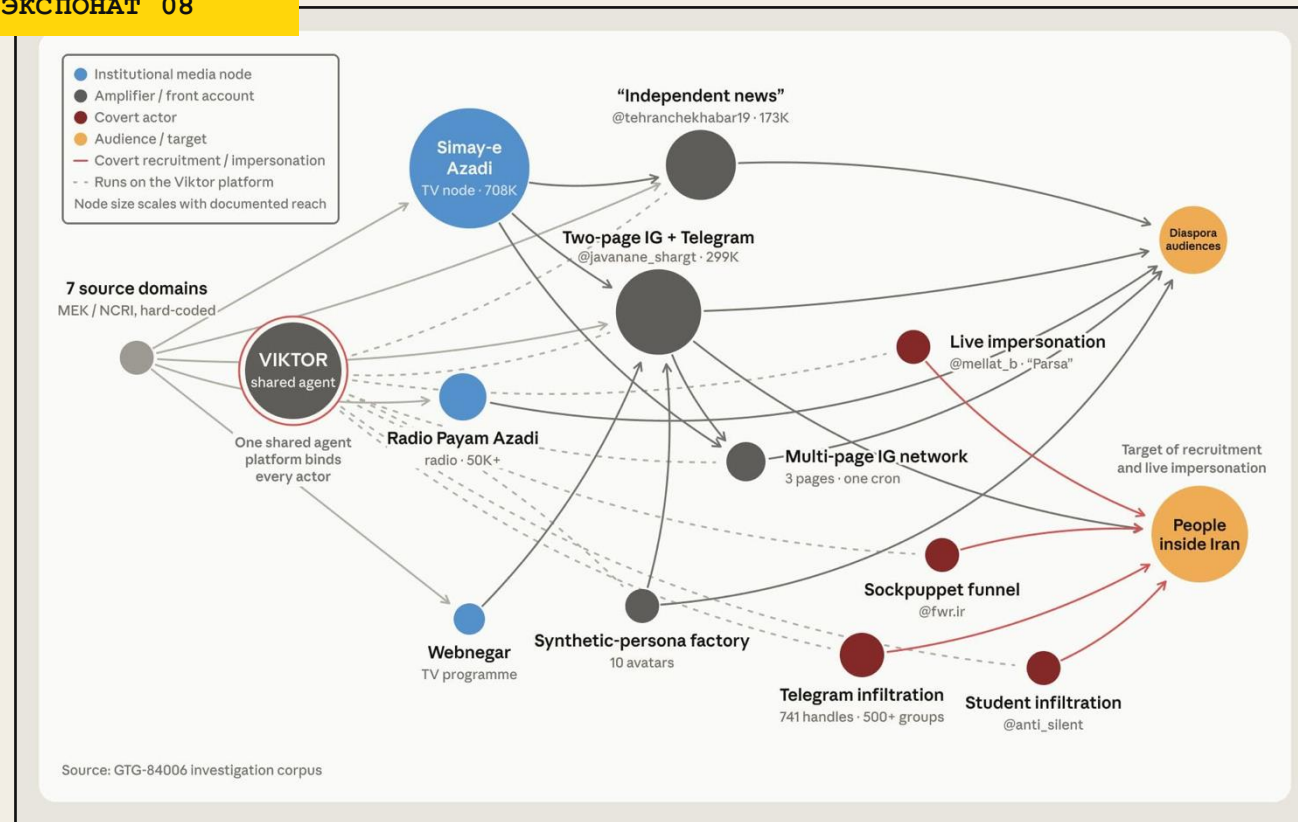
постов активиста прочитал ИИ, чтобы скопировать стиль

51 944

сообщения проанализировано для психологических профилей людей внутри Ирана

Контакты активиста **не знали, что общаются с ИИ**. Данные собраны с 500+ каналов и отсортированы по городу, возрасту и истории арестов.

ЭКСПОНАТ 08



Единый агент «Viktor» связывает все узлы сети — от клонирования до вербовки людей внутри Ирана.



Мали: три мобильных оператора, абоненты всей страны.

**За сколькими SIM-картами может следить
один человек с ИИ?**

Мониторинг мнений и перехват на 25 млн SIM

- В Мали один консультант спецслужбы построил на ИИ систему «Lakana 360», следящую за ~25 млн SIM-карт всех трёх операторов.
- Система создавала досье на любой номер **без судебного ордера**.
- В Китае ИИ ежедневно сортирует новости и соцсети по «политической чувствительности» и помечает цели для «контроля».

25 млн

SIM-карт под наблюдением (Мали)

15–30+

статей в день обрабатывал один конвейер

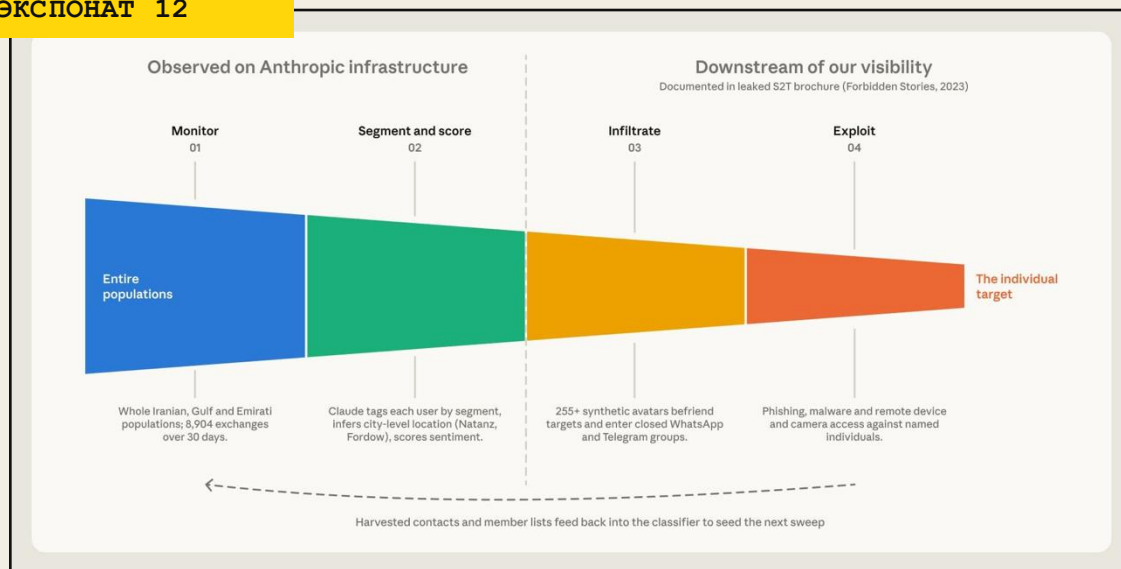
ЭКСПОНАТ 11



Китайская панель «мониторинга общественного мнения» с рейтингом «чувствительности».

Досье на «активистов, журналистов, диссидентов»

ЭКСПОНАТ 12



Воронка сбора: ИИ-классификация и генерация персон позволяют проникать в целевые сообщества.

255+

синтетических аккаунтов про запас

6 групп

демографическая сортировка людей

- Коммерческий вендор использовал Claude, чтобы профилировать пользователей соцсетей в Иране и странах Персидского залива и **сортировать их на сторонников и противников власти**.
- Отдельно строился запас из 255+ синтетических аккаунтов для будущего развёртывания.
- Anthropic прямо квалифицирует это как **нарушение запрета на профилирование** и досье на активистов, журналистов и политических диссидентов.



42 европейские партии, СМИ и аналитические центры.

**Сколько хакеров и какой квалификации
нужно, чтобы атаковать СМИ и
активистов?**

Один человек → атака на 42 организации

Весной 2026 один франкоговорящий хактивист с помощью Claude атаковал европейские партии, СМИ и аналитические центры. Вывод Anthropic: «ИИ превратил низкоквалифицированных хактивистов в продвинутую постоянную угрозу».

42 → 14

целей; внутренний доступ
минимум к 14

12–26 ГБ

баз данных украдено

140 000

записей о политических
взглядах граждан

15 000

сообщений в одном почтовом
ящике

эпизод 1

Взлом редакции через её же читателей

В сайт издания внедрён скрипт, снимавший отпечатки тысяч браузеров посетителей; оператор «целенаправленно охотился за сессиями и учётками редакционного штата».

эпизод 2

Доксинг-платформа, собранная одним человеком

Десятки миллионов строк, включая медицинские ID и утечки судебной системы, выложены как анонимные dark-web-сервисы для поиска людей по имени.



Вы ищете новость и получаете готовый ответ от ИИ.

**Насколько упадут переходы на сайт
издания, которое эту новость написало?**

Экономика контента: «крупнейшая кража труда»

93%

падение переходов на сайт NYT из-за ответов Copilot (vs обычный поиск Bing)

91 692

копий материалов NYT только в промежуточных датасетах OpenAI

2 млн+

документов с nytimes.com в одном наборе из Common Crawl

ИЗ РАССЕКРЕЧЕННЫХ ДОКУМЕНТОВ

- Топ-менеджер Microsoft назвал ИИ-скрейпинг «крупнейшей кражей труда в истории человечества»;
- Глава ChatGPT в OpenAI назвал чат-боты «экзистенциальной угрозой» для издателей — они «во многом замещают» оригинал.
- По иску, сотрудники планировали обход пейволов и удаление уведомлений об авторских правах из обучающих данных.

ИИ-агенты как канал вредоносного ПО

- Кампания FakeGit: ~7600 фейковых репозиториев, 6600 профилей, 14+ млн загрузок; 800+ имитировали ИИ-навыки и MCP-серверы.
- Gemini и ChatGPT независимо порекомендовали **один и тот же вредоносный репозиторий** walmart-mcp.
- «Злоумышленникам больше не нужно обманывать пользователей — они обманывают ассистентов, которым те доверяют».

157 / 632

подтверждённых вредоносных навыков и уязвимостей (из 98 380)

\$0,03–0,10

цена одной накрученной звезды GitHub

ЭКСПОНАТ 16

Skill breakdown		
ATT&CK category	Description	Flow
Resource Development — T1587.001 Develop Capabilities: Malware; T1588 Obtain Capabilities	Offensive-security engineer persona. Implant development and M365 operations windows. Interacts with the winAgent/GoDownload build loops, Defender-evasion iterations, and Graph/EWS operational tooling.	Invoked at task pivots: the operator enters a project directory then /security-engineer shifts Claude into an engineering register for implant or platform work.
Resource Development — T1583.003 Acquire Infrastructure: VPS; T1608 Stage Capabilities	DevOps persona used to stand up and maintain containerized C2/phishing stacks (shadow_c2 compose services, mailer, merged-landing deployments) and VPS provisioning over sshpass.	Typical flow: skill invocation → docker compose build/up → curl health checks → sshpass push to rented VPS (MonoVM/eclipse proxies).
Reconnaissance / Credential Access — T1595 Active Scanning; T1589.002 Gather Victim Identity; T1190.003 Password Spraying	shadow_c2 persona and the Red-Team user_role memory rules. Active in scanning/recon and spray windows.	Flow: ctf-pentest register → theHarvester/Shodan/gobuster recon → owa_spray or netexec validation → findings folded back into project memory.
Resource Development / Persistence — T1587.001 Develop Capabilities; T1547.001 Run Keys; T1027 Obfuscation	Windows developer persona used for the native implant line (winAgent v2.0 mod, * modules, WUEngine persistence, COM/CLSID work) and PowerShell loader engineering.	Flow: skill → Visual-Studio/mingw builds → PowerShell test harness on the test host → taskkill/sc.exe service install cycles → memory update.
Resource Development — T1608.005 Stage Capabilities: Link Target	Frontend persona. Used on the phishing-platform and dashboard frontends.	Flow: skill → Next-js/Flask template edits → Claude_Preview screenshot QA.
Resource Development — T1608.005 Stage Capabilities: Link Target (lure/admin UI polish)	UI/UX designer persona. Polished the frontends of actor web projects — lure pages, admin panels	Flow: skill → design-principles pass over lure/admin HTML → Edit cycles → preview screenshots.
Resource Development — T1587 Develop Capabilities (C2 dashboard & builder UI)	'Senior Frontend & Design Engineer' persona purpose-built for the Shadow C2 CNC dashboard and builder UI — maps dashboard template files and REPORT.md API shapes.	Flow: skill → CNC dashboard component work → agent-table/builder views wired to the shadow_c2 Postgres backend.
Resource Development — T1587.004 Develop Capabilities: Exploits; T1203 Client Execution	'Expert iOS Safari exploit researcher' projectSettings skill: ARM64e PAC bypasses, JSC JIT exploitation, kernel internals; instructs the model to load the lab memory index before any work and lists confirmed dead ends never to retry.	Flow: skill auto-scopes the Safari lab → MEMORY.md index load → DarkSword/Coruna chain work with ipsw/otool → dead-end ledger updates (institutional memory for exploit R&D).

Навыки, которые атакующие создавали для ИИ-воркфлоу.



Полицейский ИИ-фильтр проверяет запросы на поиск людей по камерам по 8 чувствительным категориям.

Какую из них он защищает хуже всех?

ИИ-поиск людей и свобода самовыражения

01

ИИ решает, кого искать

Flock Safety даёт полиции США искать людей по текстовому описанию внутри очерченной зоны города. ИИ оценивает запрос по 8 категориям и выдаёт вердикт: **разрешить / заблокировать / предупредить.**

02

Политическое выражение защищено слабее всего

Из всех чувствительных категорий только «политическое, социальное и культурное выражение» не останавливает поиск — лишь предупреждает. **Именно так выясняют, кто ходил на протест.**

03

Фильтр фиксирует, но не останавливает

Минимум 50 офицеров обвинялись в злоупотреблении (46 — с Flock). В одном штате за день после включения аудита нашли **2700+ несанкционированных поисков.**

ЭКСПЕРТ ACLU

«Фильтр — пылинка на спине гигантской машины слежки»

Проверка запроса до поиска ловит часть злоупотреблений, но структурные риски массовой слежки остаются нетронутыми.

Что это значит для ЦА и редакций?

- | | | |
|---|--|--|
| 1 | Модель пропаганды тиражируема | Российские и иранские госСМИ уже используют ИИ для контента на русском. Схема «младшего редактора» и «отмывания источника» применима к нашим аудиториям. |
| 2 | Слежка за диаспорами — не гипотеза | Агенты профилируют диаспоры и диссидентов по городу, возрасту и «истории арестов». Журналисты в эмиграции и их источники — в зоне риска. |
| 3 | ИИ-агент, которому доверяют, стал точкой атаки | API-ключи и ИИ-интеграции защищать как продакшен-данные. «Скидки» на доступ через неизвестных посредников и репозитории, рекомендованные ИИ, следует всегда проверять — как до использования, так и после. |
| 4 | Проверка фактов не должна быть лишь формальностью | Галлюцинация в виде официального документа проходит без проверки. Связная ложь фильтруется хуже разрозненных фейков - нужны процедуры проверки ИИ-выводов на уровне каждой редакции. |
-

Спасибо.

Рустам Гулов · AI-literacy consultant · rustamguloff.com

ИСТОЧНИКИ

Отчёт Anthropic «Detecting and countering misuse of AI» (сентябрь 2026); TechCrunch, Wired, The Verge, Ars Technica, AI News, MarkTechPost (сентябрь-октябрь 2026); Центр прав человека Berkeley; препринт «Truth as a Compression Artifact» (arXiv:2603.11749) .