halcyon | Ransomware Research Center

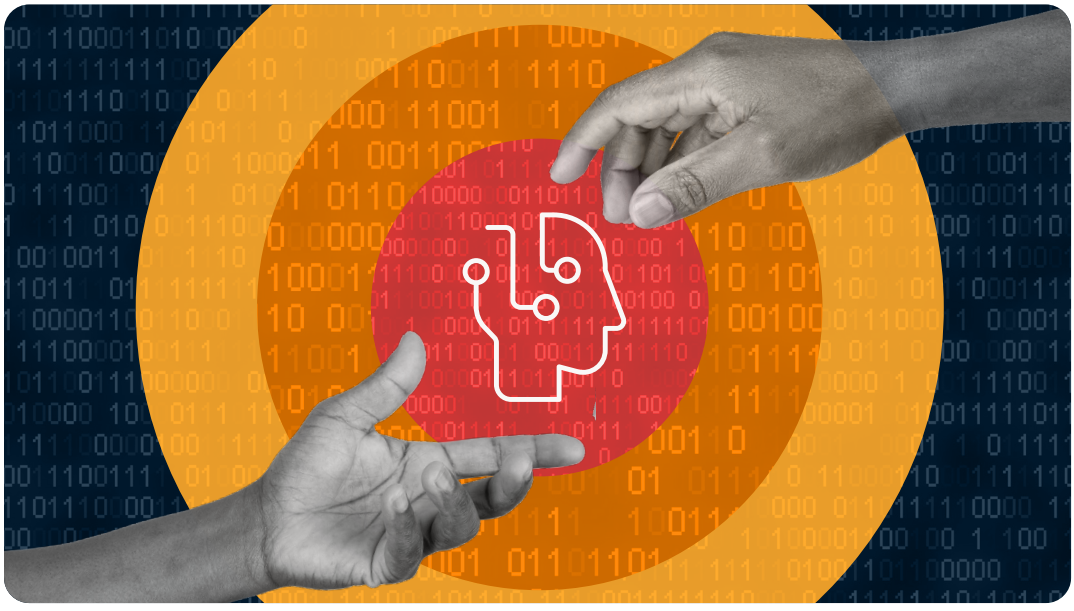
Underground AI-as-a-Service: How Cybercriminals are Buying and Selling AI

Executive Summary

In under six months, the criminal supply of and demand for AI-powered tools grew into a sophisticated underground economy on hacker forums and platforms, maturing and scaling faster than any market these forums have seen.

From October 2025 through April 2026, the Halcyon Ransomware Research Center tracked this growth across thousands of posts drawn from more than 100 sources. Our key findings:

- **Interest in AI tools grew explosively.** Messages advertising AI tools rose from fewer than 50 per month in late 2025 to more than 1,400 by February 2026.
- **The market organized around four categories of AI-as-a-service:** weaponized (or "dark") LLMs; AI-enabled identity fraud; AI-augmented malware and attack infrastructure; and jailbroken, stolen, or uncensored AI services.
- **Identity fraud and phishing benefited most from AI.** AI now supercharges social engineering. We observed a Russian-language vendor advertising a self-sufficient AI call center that places live calls in the victim's own language.



- **The barrier to entry is effectively zero.** 40% of observed messages advertise free tools, a freemium funnel that mirrors legitimate software companies. Stolen ChatGPT Plus accounts sell for as little as $0.10, and jailbreak prompts for $0.32.
- **Automated bots run as 24/7 storefronts.** Fewer than a dozen bot accounts produced 16% of all observed Telegram traffic with no human involved, giving buyers a frictionless way in.
- **Tools propagate across platforms simultaneously.** Dark LLMs, jailbreak prompts, and Know Your Customer (KYC) bypass kits appear on messaging platforms, forums, and marketplaces in parallel, giving the ecosystem reach and resilience against takedowns.
- **The AI tool users are themselves being hacked.** Infostealer malware logs show credentials for dark AI tools being harvested from the very criminals who bought them. This is a recursive threat layer not seen documented before.

Methodology

The Halcyon Ransomware Research Center concentrated collection where hacker tool economies thrive, giving us cross-platform visibility across deep and dark web forums, dark web markets, and Telegram channels. We also analyzed malware logs to surface compromised credentials tied to these AI tools, and drew on data shared by our partners at Beazley Security.

We built four thematic queries, one for each category of AI-as-a-service: weaponized LLMs; AI-enabled identity fraud; AI-augmented malware and attack infrastructure; and jailbroken or stolen AI services. Each query combined tool names, associated domains, hashtags, and operational terminology, refined iteratively as results came in.

Our final dataset comprises nearly 4,000 posts spanning 77 Telegram channels, 20 dark web forums, and five specialized underground markets, collected between 1 October 2025 and 1 May 2026. This seven-month window offers a recent, close-up view of how hackers market, sell, and use AI tools.

Two caveats shape how the data should be read:

- **Prices are undercounted.** We captured only explicitly stated amounts. The prevalence of "DM for pricing" language means real transaction volumes almost certainly exceed what the data shows.
- **Tool names are obfuscated.** We have deliberately masked the names of specific dark AI tools to avoid advertising them or accelerating their adoption. Where we do so, we describe the tool's capabilities instead.

A Brief History of AI in Cybercrime

The tools in this report did not appear overnight. Criminal interest in generative AI traces back to mid-2023, when a developer using the alias "Last" launched WormGPT on Hack Forums, which was a language model marketed for writing convincing phishing emails and malicious code without the guardrails built into commercial systems like ChatGPT. Media attention forced it offline within weeks, but the demand it exposed and the brand it created proved permanent. FraudGPT followed almost immediately, adding carding templates, scam-page generation, and exploitation scripts. Both tools sold first on underground forums, then migrated to Telegram for wider reach.

What began as individual tools became an ecosystem. Early dark LLMs were little more than jailbroken wrappers around commercial APIs. Later generations were independently hosted, fine-tuned on malicious data, or built on open-source models with their safety training stripped out. Some combined code generation, vision, and voice in a single system. In parallel, AI-enabled identity fraud emerged as its own discipline: deepfake tools that defeat the liveness checks banks rely on, and voice cloning that needs as little as three seconds of audio. The jailbreaking scene matured too, moving from ad hoc prompt experiments into a commercial market, aided by the first academic taxonomy of jailbreak techniques in 2024 and by "abliterated" models whose safety training had been deliberately reversed.

None of this created new categories of crime. Ransomware crews, business email compromise (BEC), identity fraud, and social engineering all predate generative AI, with established vendors and playbooks of their own. What AI changes is the economics of carrying attacks out, and it does so unevenly:

- **For sophisticated groups:** AI compresses timelines and scales operations, especially initial access, even as these actors stay cautious about handing full attacks to AI while their existing playbooks still work.
- **For amateurs:** AI supplies skills they never had. The results are often cruder (two ransomware crews have accidentally built malware they could not decrypt), but some capability beats none, and each tool for sale lowers the bar further.

The limiting factor in cybercrime has always been skilled human labor. AI relaxes that constraint, and each generation of tools is more capable, more accessible, and cheaper than the last. By late 2025, that trajectory set the stage for the market this report documents.

Underground AI-as-a-Service Surges in Early 2026

For most of late 2025, AI cybercrime tool activity on the monitored platforms was insignificant; however, activity increased in January 2026. February through April 2026 volume increased significantly again. In under 6 months, the market went from background noise to a high-volume commercial ecosystem.

The data revealed two distinct spikes consistent with coordinated promotion campaigns or major tool releases rather than steady adoption. The pattern suggests that specific events or vendor pushes drove step function increases in activity rather than word of mouth diffusion.

Established underground forums hosted tool distribution threads, database leaks, and pricing discussions before activity on Telegram skyrocketed. This implies a two-phase propagation model where established underground forums serve as the incubation layer where tools are first developed, tested, and discussed among experienced actors, and Telegram serves as the amplification layer that brings these tools to a broader, potentially less technical audience.

The platform distribution varies significantly by threat category. Malicious LLMs are discussed heavily on both Telegram and established forums, with the same tools and even the same leaked databases appearing simultaneously across both platform types. AI enabled identity fraud shows stronger presence on specific markets where KYC bypass tools are bundled with stolen financial data. Jailbroken and stolen AI services show the highest cross-platform spread, appearing across forums, messaging platforms, and marketplaces in parallel. *(See below for our categorization of four different kinds of AI-as-a-service that we observed either discussed or advertised for sale on the dark market.)*

The commercial intent is apparent. Forty percent of observed communications advertise free tools or services. Thirty percent list explicit prices. Less than 1% direct buyers to private messages for negotiation. Other messages include discussion, reviews, complaints, and general chatter.

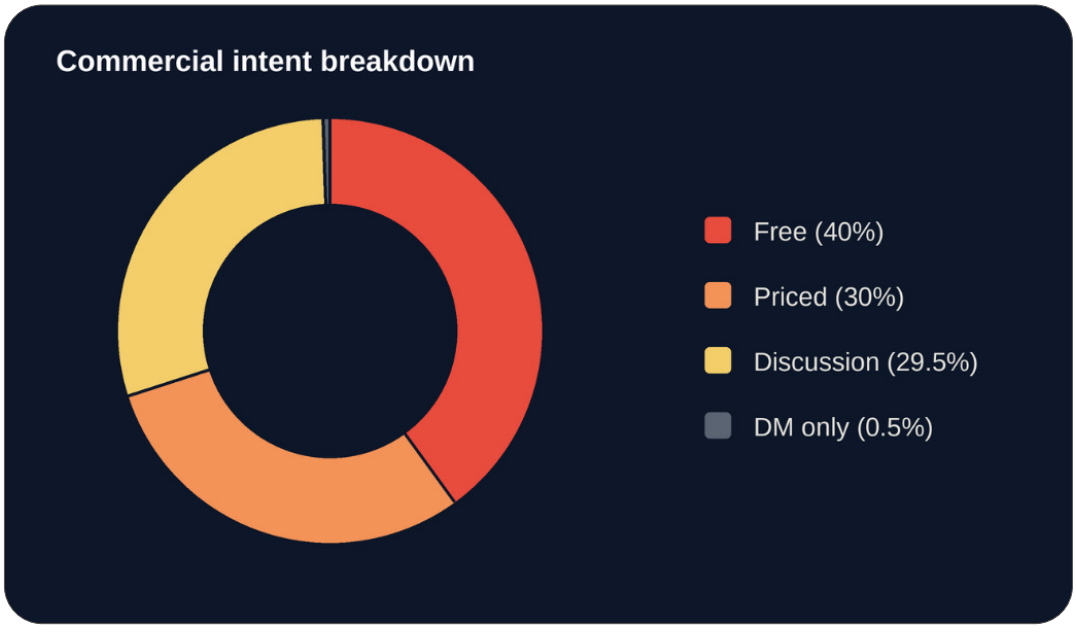


Figure 1: Commercial intent breakdown

What's for Sale: The Four Kinds of AI-as-a-Service

We observed AI tools advertised in several specific ways to aid cybercriminal activity. Generally, we categorized this usage into four categories, with multiple subcategories: weaponized LLMs; AI-enabled indentity fraud; AI-augmented malware and attack infrastructure; and stolen LLM credentials and jailbreaking services. Some of these categories can overlap, but we thought their use cases and observed examples were notable enough to share here.

Criminal interest is closely distributed across all four types of AI tooling we tracked. Across our dataset, the largest share of posts, 27.7%, referenced stolen LLM credentials and jailbreaking services. AI-enabled indentity fraud and weaponized LLMs followed closely behind at 25.6% and 24% respectively. The lowest share of posts we observed referenced AI-augmented malware and attack infrastructure, at 18.9%. A small number of posts, 3.8%, did not fit in any category.

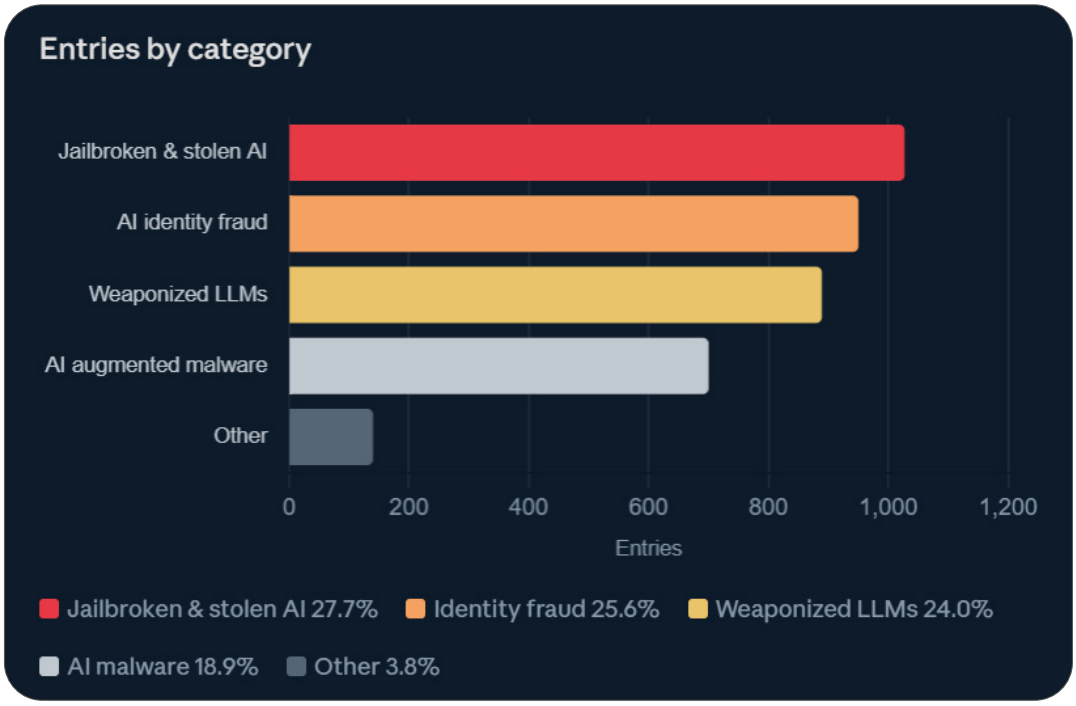


Figure 2: References by Query Categories (October 2025 to April 2026)

Weaponized LLMs, a.k.a. Dark LLMs (888 posts, 24.0%)

Weaponized LLMs, also known as dark LLMs, are stripped of the safety guardrails that constrain mainstream systems like ChatGPT, Claude, or Gemini. They produce working attack content on demand, e.g., phishing emails, malware code, and fraud scripts. This compresses the skill gap between novice and experienced threat actors. Dark LLMs reach the market through three primary paths:

1. **Uncensored Models:** Open-source models not necessarily intended for malicious use, but left uncensored by their makers, are trained on malicious data by cybercrime actors.
2. **Jailbroken Models:** Models that are censored for ethical purposes by their makers are exploited by cybercrime actors, who extract unrestricted output from the models by pairing them with jailbreak wrappers that sit in front of commercial APIs.
3. **Built Models:** Models purpose-built by cybercrime actors and trained on malicious data to help cybercrime purposes.

We have observed examples of all three of these kinds of dark LLMs on underground platforms. While we are deliberately not amplifying most of the names of uncensored models in this paper, it is worth spotlighting the most prolific and publicly known family of built models: WormGPT.

Originally, WormGPT was a paid, dark LLM built on the GPT-J architecture, designed to bypass ethical guardrails for creating phishing emails, business email compromise (BEC) attacks, and malware. Its developer advertised it for sale on underground hacking forums starting June 2023. However, the developer shut it down in August 2023 due to intense media exposure.

Yet the idea of WormGPT remained. Following the shutdown of the original model, multiple copycat AI tools and services created by multiple developers emerged on underground platforms. Some variants were powered by xAI's Grok and Mistral AI's Mixtral, posted by users with no connection to the original developer. Yet they still named themselves "WormGPT." Others started operating outright scams under the WormGPT brand. In the years since, the name "WormGPT" has become a generic term for AI built without guardrails.



Figure 3: Named malicious LLM timeline (June 2023 to March 2026)

One of the vendors with the highest activity operates through “wormgpt.site” with a pricing structure that mirrors a legitimate SaaS company. Basic access runs $9.99 per month, discounted from $35. Premier access runs $19.99 per month, down from $60. Pro access runs $49.99 per month, down from $100. A lifetime access option goes for $500. The actor also creates urgency driven limited time 24-hour promotions.

Separate listings valued the original WormGPT at $1,500. The “official” WormGPT Telegram channel has 15,283 subscribers, with operators claiming 8,900 concurrent active users and announcing that version 7 would be free and accessible to anyone.

Distribution for WormGPT runs through at least three parallel channels, the official website, a branded Telegram bot, and a persistent WormGPT thread on forums. A single forum thread, launched in December 2024, has accumulated 75 pages of replies and 378 user endorsements, linking to a WormGPT instance hosted on HuggingChat. This multi-channel approach provides redundancy; if one channel goes down, the others continue operating. It is the same availability strategy used by legitimate cloud services.

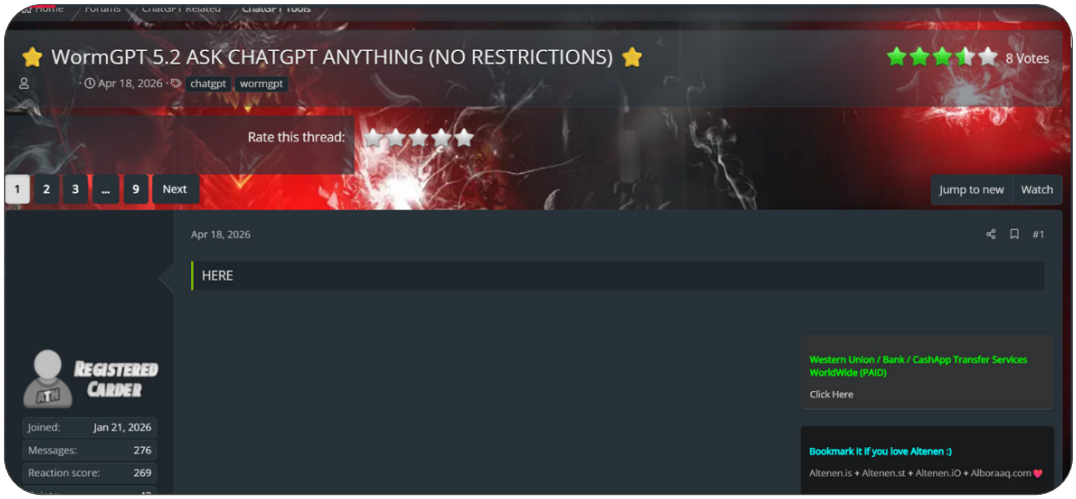


Figure 4: WormGPT 5.2 thread on hacking forum

A striking finding in this category is that WormGPT.AI was breached in February 2026. Over 19,000 user records including emails, payment data, subscription details, and user IDs were leaked across numerous hacker forums simultaneously. The people who paid for a tool designed to victimize others had their own data dumped on the same forums where the tool was first sold. The breach was propagated by at least three different forum users across multiple platforms within days.

Beyond WormGPT, several other named tools in active circulation were identified; GhostGPT, DarkGPT, and FraudGPT appear regularly. Additional AI tooling includes a free WormGPT clone, a dark web assistant and a publicly accessible AI platform promoted on forums as the number one uncensored AI with the tagline “No filters. No guardrails.”

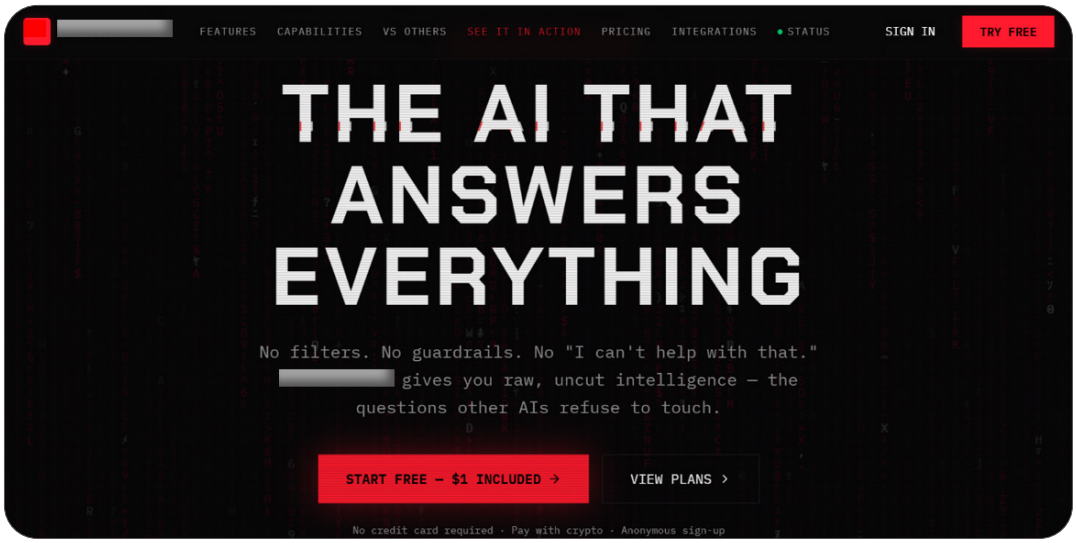


Figure 5: Landing page of a publicly accessible, uncensored AI platform

AI-Enabled Identity Fraud (949 posts, 25.6%)

This category of AI usage by cybercriminals encompasses deepfake video generation, voice cloning, Know Your Customer (KYC) bypass tools, document forgery, and selfie generation services. All are designed to enable identity fraud or defeat the identity verification systems upon which financial institutions, cryptocurrency exchanges, and online platforms rely.

In this category, AI proves most valuable where verification is hardest to fake. Liveness bypass is the hardest element of KYC to defeat, because it requires a synthetic face responding to real-time prompts to turn the head, blink, or speak a phrase. Vendors have reverse-engineered the security requirements of these liveness detection checks and built services to defeat each one. Examples include tools and services advertised on multiple hacker forums that include the following features:

- AI-generated face responses, claiming a 92% bypass rate against Sumsub, Onfido, Jumio, and Veriff.
- Deepfake MP4 videos featuring the buyer's face with specific capabilities, including rotating head movement, face swap, voice editing, and mimicry.

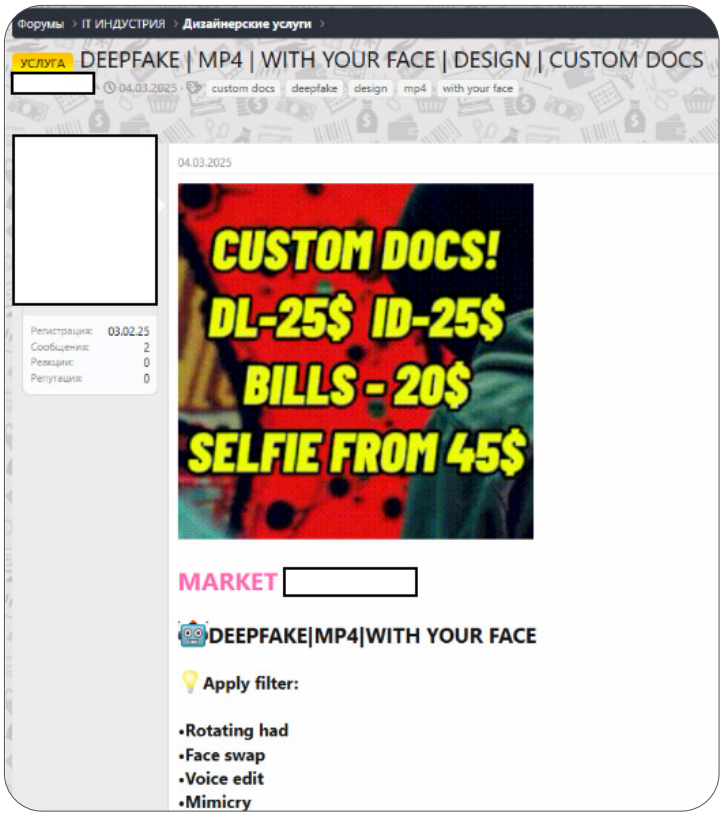


Figure 6: Deepfake service advertisement (rotating head, face swap, voice edit, mimicry)

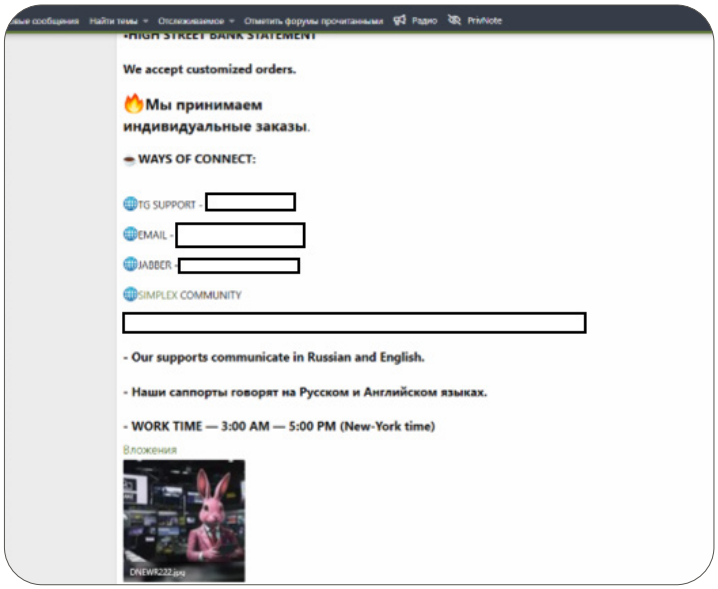


Figure 7: Services support details from the same provider

We observed the following a la carte prices for these tools and services from a vendor on a Russian language hacker forum, reflecting a segmented market:

- Basic document editing – $5
- Document scans for $25
- Driver's licenses and ID cards – $15 each
- Deepfake video – $80
- Voice edits – $60
- Photo manipulation – $40

Comprehensive synthetic identity packages, including matching government IDs, utility bills, bank statements, and real time deepfake video for live KYC verification calls commanded up to $6,350.

Vendors operate with a degree of professionalism. This vendor fielded a three-person team with distinct specialties in graphic design, verification and deepfake services, and 24/7 automated document sales. They even maintained holiday schedules and promotional discounts.

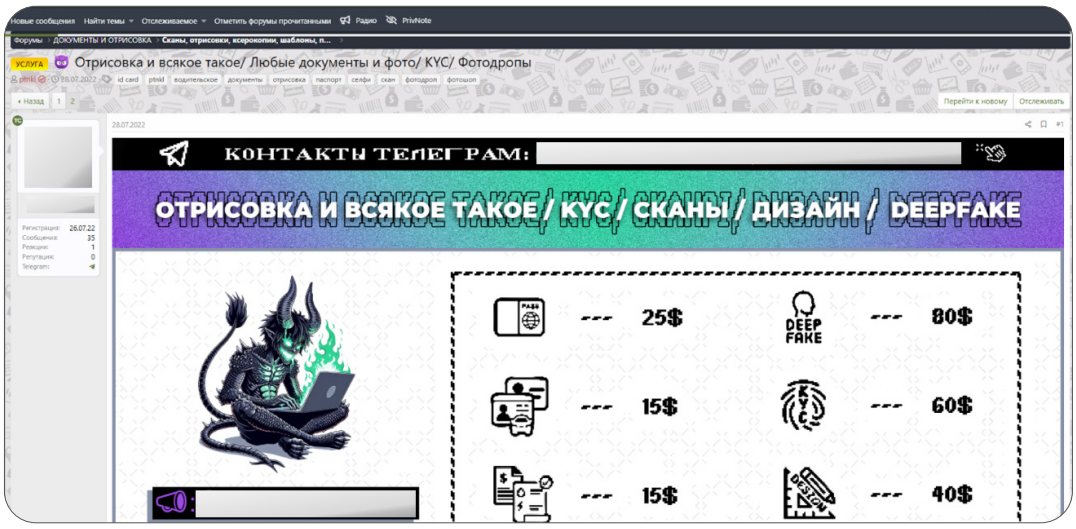


Figure 8: Russian vendor full pricing menu (documents $25, deepfake $80, voice $60)

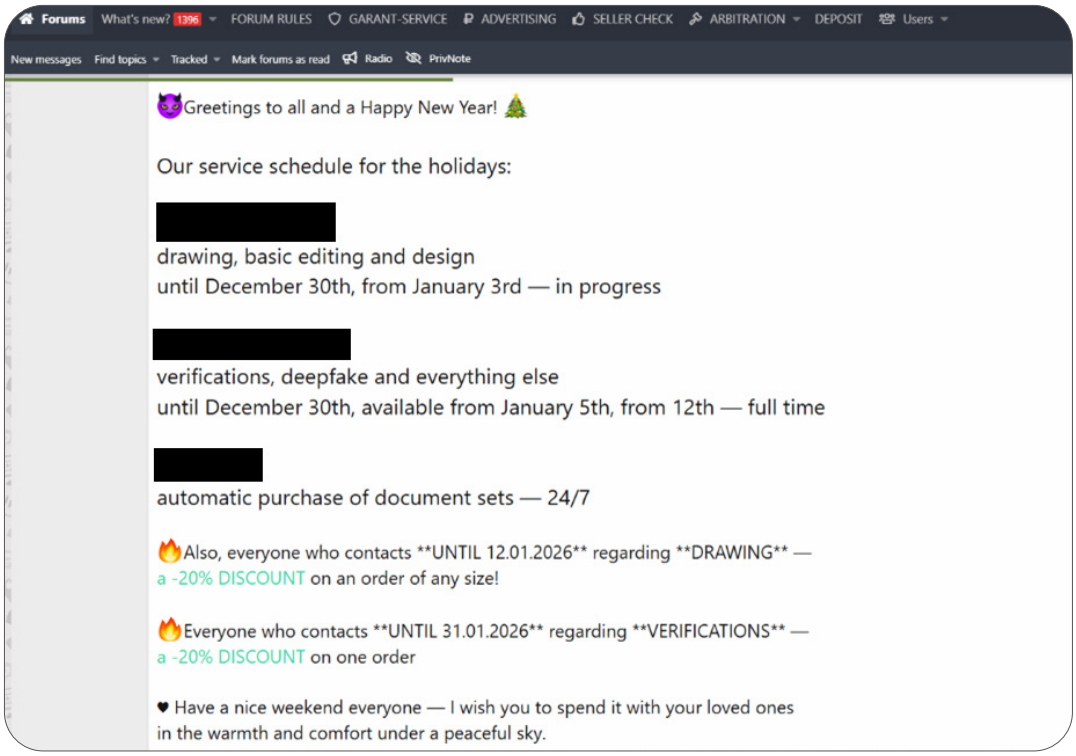


Figure 9: Russian vendor services description

The real-world impact is already significant. Researchers recorded 22 AI-enabled fraud incidents globally between 2017 and 2022, 150 in 2024, and 179 in the first quarter of 2025 alone, surpassing the entire prior year in a single quarter. One deepfake Zoom call resulted in a Singapore firm losing $499,000. A European energy company lost $25 million from a single cloned CFO voice. As more data on AI-enabled fraud costs emerge, researchers will be able to improve their assessments on the general impact and cost of this category of malicious AI use.

While the above screenshots are not the specific tools that may have enabled those specific attacks, they are the same category of tools, sold openly on hacker forums to anyone willing to pay.

AI-Augmented Malware and Attack Infrastructure (699 posts, 18.9%)

During our period of observation, Halcyon did not observe tools that explicitly advertised a polymorphic malware capability that uses AI to rewrite itself and evade signature-based detection. Such a capability is widely discussed by the cybersecurity community, but the lack of advertisements for it suggests that it remains a theoretical concern – for now. Given the underground market demand for such a capability, we assess it is a matter of when, not if, such tools become available. However, we have observed AI used to:

- Generate phishing attacks.
- Augment attack infrastructure.
- Automate early-stage malware development.

One of the earliest cybercriminal uses of AI was to create highly personalized, scalable, and convincing phishing campaigns. A significant number of AI tools advertised on underground cybercriminal forums claim to support phishing campaigns in some capacity.

Now, more sophisticated than phishing generation, AI is also being used to augment the whole infrastructure used for attacks. In one example, a vendor advertised an AI-powered outbound call center connecting to 3CX telephony infrastructure that had been trained on data from 130,000+ customer calls. The system calls clients in real time in up to 25 of the victim's own languages. A malicious actor can upload names and emails from Excel or CSV uploads, and the AI can make up to 120 simultaneous calls to these people. The calls say the target's actual name and produce sounds mimicking human call center operators. The system delivers real-time transcripts via Telegram and web dashboard. The entry fee is a one-time payment of 420,000 in unspecified currency for lifetime access to this service. Services like this one mark a major shift in how malicious use of AI is evolving from writing text to holding live phone conversations over professional-grade calling systems.

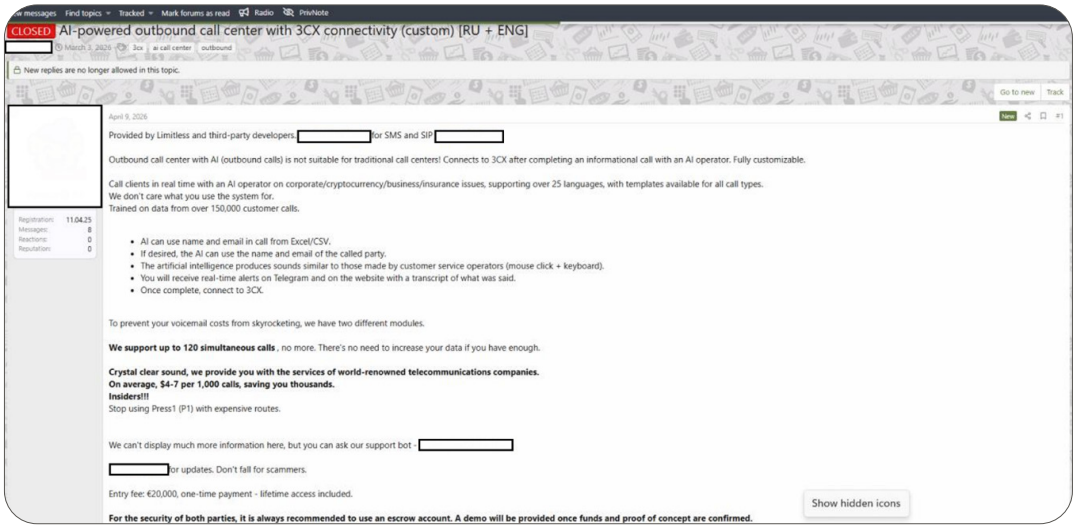


Figure 10: AI-powered outbound call center with 3CX connectivity (120 simultaneous calls, 25 languages)

On the malware development side, one of the more striking examples Halcyon observed was a fully AI-generated malware ecosystem advertised for sale by a vendor on a Russian language hacker forum. Researchers confirmed the framework reached approximately 88,000 lines of functional code in under one week, created using ByteDance's TRAE AI development assistant. This quantity of work used to take professional development teams months to complete, demonstrating the lowered barrier to entry for developing these tools.

Jailbroken and Stolen AI Services (1,026 posts, 27.7%)

Halcyon tracked more posts for jailbroken and stolen AI services than any other type of AI offering, reflecting overwhelming demand for unrestricted AI access without the safety guardrails imposed by commercial providers. These services were also the cheapest, with working jailbreaks selling for as little as $10 to $15. The category spans two related offerings:

- Prompts engineered to strip the safety guardrails from mainstream AI models.
- Stolen or resold access to commercial AI accounts.

It is important to note that most frontier models effectively monitor for and quickly address potential prompts that claim to have successfully bypassed model safeguards. This security measure renders many jailbreaking prompts advertised by criminals obsolete.

Still, the underground community persists and has organized itself into a functioning ecosystem.

Multiple hacker forum threads advertise free ChatGPT jailbreak prompts, illegal versions of ChatGPT, GPT 5.2 jailbreaks, and WormGPT 5.2. These function as living repositories where users share, test, and refine techniques over months.

- On Telegram, one group with 1,730 members operates as a dedicated jailbreak marketplace. A seller there advertises a Claude jailbreak targeting Anthropic models, Sonnet 4.6 for web chat and Opus 4.7 for CLI agents, including Claude Code, Cursor, and Cline. The listing claims it has been tested on code, red team, and prompts Claude usually refuses, producing clean code output with no placeholders. The price is $15 for the bundle or $10 separately, payable in USDT cryptocurrency. The jailbreak's advertising suggests an intended use for automated code generation, which Claude's safety layer would normally refuse.

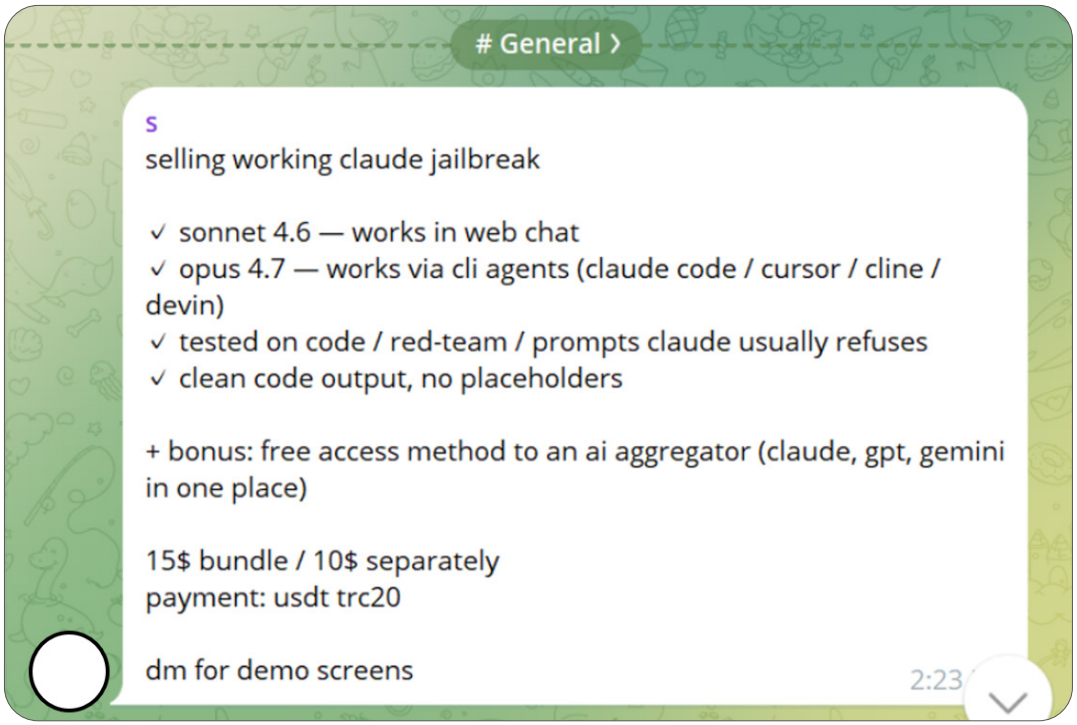


Figure 11: Telegram jailbreak marketplace (1,730 members, Claude Sonnet 4.6/Opus 4.7 jailbreaks)

The buying and selling of stolen AI account access has become a standardized, high-volume business where credentials, API keys, and subscription access are cheap, readily available, and sold both in bulk lots and as single units. The median price across this entire category is $0.10. Examples we observed include:

- One hacker forum hosts daily bulk drops advertising ten ChatGPT business plan accounts at a time. Other hacker forums list Tier 2 OpenAI key dumps.
- A Telegram channel with 66 subscribers sells OpenAI API keys loaded with $18 credit for $1.50 each and runs giveaways of accounts loaded with $120 credit to drive channel growth.
- A storefront with 140 subscribers sells ChatGPT Pro subscriptions for $3 per month versus the legitimate $20 per month, with instant delivery via Binance Pay or USDT, operating bilingually in English and Arabic to serve a Middle Eastern customer base.
- We also observed multiple AI platforms claiming to be **uncensored** that were discussed, traded, and resold across hacker forums. Examples include:
- A platform operating as a free, no-login, uncensored chatbot.

- A privacy focused AI platform that is resold on a hacker forum at discounted rates with one-year plans going for $15. It had its user credentials stolen by infostealers and listed on a Russian hacker market.
- A model described as a criminal AI assistant designed specifically for bypassing content moderation.

So-called "abliterated" models—open-source LLMs with safety training deliberately reversed through fine tuning—also circulated in underground forums, messaging platforms, and marketplaces.

How it's Sold: AI-as-a-Service Commercial Strategies

We observed three kinds of overlapping commercial strategies used by cybercriminals selling AI-as-a-service: a freemium funnel, bot-driven distribution, and multi-platform propagation. All three borrow directly from legitimate technology business models.

The freemium funnel strategy is the same playbook used by legitimate security-as-a-service (SaaS) companies: they give away their entry-level product and monetize power users who demand more capability or reliability. We observed that 40% of all messages in our dataset advertise free tools or services, while only 30% list explicit prices. Free offerings include cracked tool versions, limited feature demos, open-source jailbreak prompts, and outright scams targeting aspiring criminals. The free offerings serve as lead-generation for premium paid tiers. A persistent WormGPT Free thread on one hacker forum, accumulating 75 pages of replies and 378 endorsements sustained over more than a year, is the textbook example: the WormGPT Telegram channel's announcement that version 7 would be free and accessible to anyone on bot follows the same logic.

The bot-driven distribution strategy functions as an always-on storefront, pushing tool advertisements at scale without requiring a human vendor to be present. It is automated retail infrastructure running 24/7, analogous to the automated storefronts that legitimate e-commerce companies deploy. We observed nine automated Telegram bot accounts produced 15.8% of all Telegram traffic in the dataset.

The multi-platform propagation strategy is one in which the same tools appear simultaneously across Telegram and multiple hacker forums and markets focused on carding. For example, WormGPT was distributed across thirteen distinct source types. Jailbreak prompts circulated across seven. KYC bypass tools appeared on both Telegram channels and specialized hacker markets. This model creates deliberate redundancy to ensure no single platform takedown disrupts business. When one hacker forum goes offline, others continue serving. When a Telegram channel is banned, the forum threads and bot storefronts persist.

Who Does What: AI-as-a-Service Job Functions

The vendor ecosystem itself has stratified into at least six distinct archetypes:

1. **Branded Vendors:** Some vendors run recognizable, branded operations with tiered subscriptions and promotional pricing.
2. **Bot Storefronts:** These handle automated push distribution.
3. **Micro Vendors:** These serve localized markets with per-prompt sales at $0.32.
4. **Account dumpers:** These actors distribute stolen ChatGPT Plus credentials in daily bulk drops on hacker forums.
5. **Full-Service Fraud Operators:** These actors offer bespoke identity packages combining document forgery, deepfake video, and KYC bypass at prices up to $6,350.
6. **Free Distributors:** These actors maintain persistent forum threads offering tools at no cost, likely freemium funnels, cracked versions, or scams targeting aspiring cybercriminals who are themselves the product.

This stratification is the clearest indicator that the market is maturing. The underground AI tool market is a segmented, multi-tier economy with emerging specialized roles, which develops in any market once it reaches sufficient scale and demand.

Preying on Each Other: Theft Inside the AI-as-a-Service Ecosystem

A novel finding in the data revealed that threat actors who use AI cybercrime tools are themselves victims of infostealer malware. Malware logs from dark web marketplaces showed that infostealers—e.g., Rhadamanthys, Vidar, and Acreed—are running on the machines of those using tools such as dark LLMs and deepfake websites. Those stolen credentials end up listed for sale alongside the usual bank logins and email accounts. The ecosystem evidently feeds on itself.

The recursive chain is observed as follows:

- Threat actor A buys WormGPT to generate phishing emails.
- Threat actor B deploys an infostealer that infects Threat actor A's machine.
- Threat actor A's WormGPT credentials are harvested and listed on the same marketplaces.
- Threat actor C buys those credentials and gains access to Threat actor A's paid AI tools without paying.

This is compounded by the WormGPT.AI breach. In February 2026, at least three different forum users posted over 19,000 user records, including emails, payment metadata, subscription details, and user IDs, across multiple popular hacker forums. It was discovered that the paying customers of a WormGPT.AI tool built to facilitate cybercrime had their own data exposed on the same forums where the tool was sold.

Implications for Defenders, Law Enforcement, and AI Model Providers

The number of people with unrestricted AI access for criminal purposes is orders of magnitude higher than any single data point suggests, according to our observations. The WormGPT.AI breach alone exposed 19,000 users, and operators claim nearly 9,000 concurrent active users, a hefty malicious LLM user base. This example represents just one tool out of at least a dozen identified. People are also using free jailbreak alternatives, buying stolen accounts for less than a US dollar, and accessing AI tools and services in localized micro markets.

This finding and others in our paper carry direct implications for security teams, law enforcement, other cybersecurity professionals, and AI model providers. Specifically, they must actively guard against the four kinds of malicious AI use we categorized in this paper. Also, criminals' use of AI tools suggests new opportunities for evidence and intelligence collection, which teams can train to take advantage of now.

Recommended Defenses against the Four Kinds of AI-as-a-Service

Prepare now for AI vishing. Organizations should assume that AI-generated voice calls impersonating executives, IT support, or business partners are now technically feasible at scale and plan accordingly. Traditional email-focused phishing defenses and security awareness training by themselves cannot address this threat. See above for our observations of AI used to create an automated call center.

Assume traditional phishing social engineering indicators are outdated. The traditional indicators that analysts are trained to flag, e.g., grammatical errors, awkward phrasing, inconsistent tone, or generic greetings, are no longer reliable. Training employees to "spot the phishing email" is not sufficient as a primary defense. AI-generated phishing content is fluent, contextually appropriate, and can be produced in any language. Security teams should reframe their assumptions and training enterprise workforces accordingly.

Shift detection indicators from content-based to behavioral. To identify AI-enabled malicious activity, SOC teams and incident responders can look for unusual request patterns, anomalous transaction timing, out-of-band verification failures, and deviations from established communication workflows. Organizations can implement mandatory out-of-band verification for financial transactions or credential requests, regardless of how legitimate the communication appears.

Prepare for a significant increase in attack volume from less skilled operators. The tools documented in this research enable amateurs to be more viable threats. A threat actor who previously lacked the language skills, technical knowledge, or operational discipline to conduct a convincing social engineering campaign can now purchase that capability for a few dollars. SOC teams should anticipate higher volumes of competent but not expert level attacks. The baseline quality of phishing and vishing attempts is rising across the board, so automated filtering and detection systems should be tuned accordingly.

Opportunities for Tracking and Disrupting Malicious AI-as-a-Service

Use malicious actors' payment infrastructure to track and disrupt their movements. AI-as-a-service sales in our dataset overwhelmingly transact in cryptocurrency, primarily USDT on TRC20 and ERC20 networks, processed through services like Binance Pay. Law enforcement and other disruption operators can focus on payment infrastructure as a pressure point. Legal coordination with cryptocurrency exchanges and payment processors to flag and freeze associated wallets could lead to disruption of AI-as-a-service revenue streams. The WormGPT breach data itself, containing 19,000 user records with emails and payment metadata, is a potential investigative starting point for identifying and attributing both vendors and buyers.

Coordinate sustained pressure campaigns across multiple AI-as-a-service platforms. Law enforcement and the cybersecurity community can increase the cost and friction of doing business to AI-as-a-service vendors by coordinating disruption activities as part of a sustained pressure campaign across multiple platforms, payment processes, and hosting. Targeting these platforms in a vacuum is not sufficient, given the multi-platform redundancy in the ecosystem, where the same tools appear simultaneously across Telegram channels, established dark web forums, and underground markets. As part of these coordinated campaigns, law enforcement and other operators can engage domain registrars and hosting providers through legal process to take down sites like wormgpt.site and kriminal.ai. They could also engage Telegram's abuse reporting mechanisms to target known automated bot accounts.

Exploit the AI-as-a-service backdoors for intelligence purposes. When vendors place backdoors in their own products to compromise the people buying malicious AI tools and services (see above for our observations on this behavior), they also create a novel intelligence source for cybersecurity professionals. At least three kinds of intelligence can be derived:

- **Tracking AI-as-a-service popularity metrics.** Cybersecurity professionals can gain a realistic picture of which AI tools attackers are currently leveraging by monitoring infostealer logs for credentials to platforms like WormGPT, various dark LLMs, or deepfake websites. The presence of these credentials in stealer logs is direct evidence of real users for specific tools and trends.

- **Independent validation of tool adoption at scale.** Vigorous forum discussions about tools could be hype, curiosity, or simple chatter. But credentials harvested by malware from browsers confirm that threat actors have indeed used the tools discussed. They represent confirmed AI tool users whose machines were compromised, a small but verifiable sample of the broader user base.
- **Attribution of threat actors.** The same infostealer data that security teams monitor for compromised employee credentials can be used to identify, track, and potentially attribute threat actors who are purchasing and using these AI tools. If the tools' own platforms are getting breached and their users' credentials are being harvested by other threat actors, the operational security of AI-enabled cybercriminals may be weaker than the sophistication of their tools would suggest.

Opportunities for AI Model Providers and Broader Governance

Explore needed legal frameworks, oversight, and talent needed to defend and disrupt. Currently, law enforcement, the intelligence community, and security researchers who aid them are able to collect information on the four kinds of the AI-as-a-service detailed in this paper using legal frameworks designed before the advent of modern frontier models. Disruption operations are predicated on statutes such as the Computer Fraud and Abuse Act and wire fraud violations. These frameworks may be sufficient, but now that we have a firm outline of the ways criminals are actually using AI, the community would be well-served by outlining how to apply existing frameworks specifically for the investigation and disruption of these use cases. Identifying gaps in the frameworks can also provide concrete actions that policymakers and appropriators can take to prepare defenders for future AI-as-a-service harms.

Advance a regulatory ladder of AI model policies that scale with increasing risk. We observed that not all existing models carry the same risk of being exploited for criminal AI-as-a-service use. For example:

- Some providers, particularly the major US frontier labs such as OpenAI and Anthropic, actively monitor for jailbreak techniques, patch vulnerabilities within days of discovery, and maintain dedicated red teams that stress-test their own safety layers. We observed fewer instances of their models being successful jailbroken or wrapped into malicious services.
- We saw many more such instances with open-source models. These models are released under permissive licenses with no usage restrictions and cannot be patched after deployment. Once downloaded, they are beyond the reach of a safety team. Models from providers that prioritize openness or minimal content moderation over safety investment present a different category of risk.

This difference–in which some models examine and mitigate how they can be used for good or bad purposes, and some models do not–suggests a need for different standards of user caution, SOC team mitigation, and perhaps even liability or other accountability standards. A ladder of different standards could be assigned, depending on the level of security incorporated up front into new models before they are brought to market and as they continue to mature. This approach could reward innovation while also prioritizing safety. Such a ladder of standards or regulations would need to be explored between policymakers, AI industry partners, and independent cybersecurity experts.

Conclusion

The AI cybercrime economy is no longer emerging. It has arrived. In under six months, we watched it develop tiered subscription pricing, automated distribution, multi-platform redundancy, and freemium customer acquisition: commercial structures that took narcotics and exploit-kit markets years to build. This one assembled them at the velocity of the legitimate AI industry.

The market is organized around four pillars: weaponized LLMs, AI-enabled identity fraud, AI-augmented malware and attack infrastructure, and jailbroken or stolen AI services – with prices running from $0.10 for a stolen premium account to more than $6,000 for a full synthetic-identity package. That 40% of observed messages advertise free tools tells us the market has already moved past early monetization and into growth and customer acquisition.

The consequences are concrete. Someone with no technical skill, no language fluency, and no operational experience can now buy phishing content, deepfake media, and forged documents indistinguishable from what only skilled operators could produce a year ago. The ecosystem also preys on its own: infostealers harvest credentials for the very AI tools criminals depend on, and the market's dominant platform was itself breached, exposing 19,000 paying customers. The tools built to victimize others are now victimizing the people who buy them, a recursive dynamic that defenders and law enforcement have barely begun to exploit.

The question is no longer whether criminals will adopt AI. It is whether defenders can detect, disrupt, and deter an ecosystem that has already reached commercial maturity, scales as fast as the industry it feeds on, and has shown it can survive takedowns.

Join the Community

The Halcyon Ransomware Research Center unites experts, drives smart policies, and delivers actionable intelligence to detect, disrupt, and defeat ransomware. To learn about the Ransomware Research Center, visit halcyon.ai/ransomware-research. The best way for you to join is by visiting this page and submitting the form. If you are exploring the idea of joining, but need additional info, please note that in the form and we will get back to you shortly. You can also email us at Ransomware-Research-Center@halcyon.ai.