"Artificial Intelligence and Criminal Exploitation: A New Era of Risk"

July 16, 2025

Hearing Before The
Subcommittee on Crime and Federal Government Surveillance
Of the Committee on the Judiciary

Prepared Statement by
Zara Perumal
CTO/Co-Founder
Overwatch Data

Chairman Biggs, Ranking Member McBath, and Members of the Subcommittee, thank you for the opportunity to testify today, and for creating a forum to examine how artificial intelligence (AI) is changing the landscape of cybercrime. I am honored to share my perspective on how this technology is making these threats more accessible, more personalized, and more difficult to detect.

I am the Co-Founder and Chief Technology Officer (CTO) of Overwatch Data, a cyber threat intelligence company that uses AI to identify and analyze emerging threats in the cybercrime and fraud ecosystems. Through our work, we see every day how AI is used both to prevent and to facilitate criminal activity.

AI is a powerful general-purpose tool that allows users to enhance skills, learn, and create innovative solutions for complex problems with greater efficiency. However, the same capabilities can be exploited by bad actors to learn how to commit crimes, develop new methods of fraud or theft, and generate scam or sextortion content. This can cause significant harm, including major financial loss and, in extreme cases, emotional distress leading to suicide.

Overwatch Data's analysis focuses on where AI is most significantly changing the threat landscape. Specifically, what are the new threats, what are the most likely threats, and what are the most harmful threats?

What This Means: Impacts on Users, Businesses, and the Criminal Ecosystem

One of the most immediate changes is how generative AI lowers the barrier to learning, executing, and scaling cybercrime and fraud. Tasks that once required technical expertise, such as writing convincing scam messages or creating phishing pages, can now be completed using AI tools. The outcome is a flood of spam texts, emails, phone calls, and social media messages that

occur more and more often[1]. These messages are personalized, translated, and delivered at scale, making scams more convincing, more widespread, and more accessible to individuals with little or no technical background or understanding of how to commit crimes.

More novel uses of AI combine voice, image, and video understanding for more personal attacks. Scammers use voice clones built from social media posts to impersonate a loved one in distress. Victims receive urgent calls from what sounds like their family member, claiming to be injured. Some are also sent fake hospital photos to make the lie more convincing. These scams target our instinctive human trust in hearing a loved one's voice, and our willingness to do anything to help them.

Even more disturbing, nudifying apps use AI to digitally remove clothing from non-explicit photos, generating fake sexually explicit images. Threat actors use these tools to transform ordinary photos into explicit content and then weaponize the results to seek revenge, extort money, or manipulate their victims. Although the images are fake, the shame and fear they create are real. In some cases, victims have comited suicide as a result. This abuse has affected people of all ages, including children who have been targeted by their classmates, online predators, and coordinated bot networks.

In addition to personal harms, these tactics impact small and large businesses. A growing trend is using AI to subvert identity verification systems, including "Know Your Customer" checks, by generating fake photos, high-quality synthetic IDs, and even live deepfakes to pass as legitimate users. This allows criminals to access and exploit online services with reduced risk of detection. In addition, more effective social engineering through text messages and video, such as impersonating executives on live video calls, creates new risks and financial losses ranging from fraud to malware deployment.

This ecosystem evolves rapidly. The models and techniques that were state-of-the-art last year are now out-of-date and replaced with newer models. As the tools evolve, one thing remains consistent: attackers exploit our trust in one another and our outdated understanding of cybercrime. Many people still picture clumsy, typo-filled phishing emails, not the highly personalized and realistic scams that are common today.

To combat this, we cannot keep operating in the status quo. We need a whole-of-society approach, starting with education and awareness for children, extending to communities, businesses, and nonprofit organizations. It also requires a coordinated push to encourage innovation, improve information sharing, and strengthen response efforts across tech companies, telecom providers, financial institutions, and law enforcement. Each organization has different pieces of the puzzle, and unique enforcement capabilities to deter crime. No single organization can tackle this problem alone.

[1] https://www.cnet.com/tech/services-and-software/scams-survey-2025/

Adjacent Trends, Complicating Factors, and the Future of AI use for Crime

As we look at the threats that are already here, we also consider emerging trends in AI and cybercrime that offer insight into where this space may be headed.

1. **Agentic AI:** Agentic AI is the next trend in AI that is already here. Instead of just static models that generate outputs solely based on training data and prompts, agents use tools, query data, and interact with each other. This enables AI agents to be more proactive and learn by doing. As a result smaller, less sophisticated and cheaper models can have an outsized impact when deployed as part of an agent, especially for narrow or high-impact use cases. This ability to adapt through trial and error makes agents more convincing and capable, especially in tasks that involve persuasion, like crafting text messages or mimicking human behavior.

   A particularly new development is computer-use agents. These are AI systems that can operate a computer much like a human user. Using computer vision, they can see the screen, click buttons, and navigate software tools. This gives them access to a wide range of capabilities that were previously out of reach.

   Because they can take real actions, these agents are uniquely qualified to be effective in fraud. Unlike traditional bots that follow fixed scripts, computer-use agents can adapt to feedback and navigate interfaces like a human. Given a high-level description of a scam, they may be able to figure out the steps, adjust as needed, and complete the task in ways that are harder to detect and more persistent than past systems.

2. **World Models:** This is an area where AI experts think the state of the art for AI will move. Unlike models trained solely on text, images, or internet content, world models aim to simulate the physical world[2] and understand cause and effect. Mimicking how we learn about gravity and the physical world in school, these models incorporate a common-sense understanding of the world.[3] This approach may lead to a much more nuanced understanding of the physical environment.

   If AI grows in this direction, it could lead to more advanced impersonation, such as deepfake videos with realistic, live-looking backgrounds. World models may also enable more real-world crime, such as improving reconnaissance and information gathering of physical environments, by allowing AI systems to understand physics and navigate physical spaces.

3. **Growth of Individual User Data Footprint:** The more online data an individual has, the more vulnerable they are to cybercrime. Many people conduct much of their lives online through social media, online banking, medical apps, and commuting tools. Personal information is stored in public and private records and often ends up in leaked databases after service breaches. This includes phone numbers, emails, passwords, addresses, and

[2] https://www.linkedin.com/posts/yann-lecun_lots-of-confusion-about-what-a-world-model-activity-7165738293223931904-vdgR/

[3] https://ai.meta.com/blog/v-jepa-2-world-model-benchmarks/

family connections. Video, photos, and audio content are widely shared on social media, making it easier to clone identities. Such data fuels social engineering, enables AI misuse, and increases the surface for attacks. The more available, and the less secure personal data is, the more vulnerable we are to AI enabled cybercrime.

With this in mind, we expect AI-enabled crime to continue to be more tailored, adaptive, and pervasive. In light of that, we need to be clear about the future we are working towards.

The Future We Want: How to Evaluate Success in Combating AI-Enabled Crime

I, along with the whole team at Overwatch Data, work every day to contribute to a future where AI helps solve real problems, expands access to resources, and creates new opportunities. A future where people spend more time on what matters, and where built-in defenses make harm harder to carry out.

We often talk about regulation in terms of slowing down the bad outcomes. That matters. We want to deter abuse and make crime more expensive and difficult. But we should also see regulation as a tool to accelerate progress toward positive outcomes. It can guide innovation in the direction we want to go. It is just as important to focus on the future we want to build as it is to prevent the harms we want to avoid.

For many *existing* models and capabilities, we cannot put the genie back in the bottle. The models that enable the harms we've discussed have been shared, reshared, cloned, and compressed or distilled into smaller, more efficient versions. These “minified” models can replicate similar capabilities with fewer resources, making them easier to distribute and use. As we work to mitigate the harms of current models and limit the risks from future ones, we should focus on the end goal.

If it is inevitable that these models will eventually be accessible, what does a good end state look like?

This end state should ultimately guide how we measure success. Addressing AI-enabled crime is complex and every option, including doing nothing, comes with tradeoffs. As we consider what actions to take, we should first ask: How would we evaluate the outcomes? What would success look like in practice?

To ground that vision in practice, here are some of the questions that should be continually asked as both government and industry shape responses to AI-enabled crime:

1. **Compliance: What is the cost (time/money) for legitimate users?** Good faith businesses and individuals must be able to practically comply, so it matters if it is as easy as filling out a quick form, or a laborious process differing across jurisdictions. Groups of people/businesses will have differing financial means and abilities to comply.
2. **Intent OR Use: Who has to act in good faith for this to work?** If we ban certain models, or use cases, cybercriminals are unlikely to comply. The systems we put in place must be effective when those who act in good faith comply, but the bad actors do not.

3. **Risk Analysis: How does this compare the alternative?** When we ban an activity or make it difficult, threat actors often adapt to using the next easiest option. For example, if they cannot use deepfake imagery, threat actors may go back to pulling imagery off of social media or using stock photos. It is important to know what that option is and the impact to their operations by shifting them to their backup option. We should identify where those alternatives impose significantly higher costs in time, money, or risk to the criminal ecosystem.
4. **Adaptability: How does this intervention adapt?** Because AI models and illicit use cases are evolving so quickly, any actions or policy we make needs to evolve quickly.
5. **Interdependencies: What other areas of innovation does this affect?** Because AI is often multi-purpose, any change aimed at reducing cybercrime may also affect other innovations that benefit society, including those working to combat cybercrime itself.

Specific Recommendations

1. Empower Societal Resilience

Strengthen AI and Cybercrime Literacy Across Society

AI-enabled crime exploits the fact that we do not expect it. The level of personalization catches people off guard and reaches them when they are most vulnerable. From voice clones to romance scams to sextortion, these scams often target vulnerable groups like children and the elderly. Children are among the most active users of technology and have been both victims and perpetrators[4] of AI-enabled abuse. We need to engage them directly in the solution and give them the tools to prevent harm.

If we push education about cybercrime and AI across schools, businesses, and community organizations, we deter AI-enabled crime by making it less likely to work. As we grow awareness of how AI can be used for both harm and good, we make our society less vulnerable to abuse and better positioned to take advantage of its benefits.

Efforts like the Redirect Project, which teaches elementary students about cybersecurity and online harms, and Operation Shamrock, which shares scam tactics with communities, show how awareness efforts may take shape. Tools like Birdwatch and scam reporting platforms let people flag threats and share alerts in real time.

We should prioritize and fund education especially for vulnerable populations to deter crime.

2. Shift the Technical Advantage to the Defenders

Leverage AI for Defense

The same tools used to commit cybercrime can be used to stop it. If threat actors use AI to scale scams, we can and should use it to detect and block them. If they build malware that adapts, we can build systems that learn to defend. As agents are used to find and exploit vulnerabilities[5], we

[4] https://www.redirectproject.org/
[5] https://www.helpnetsecurity.com/2025/06/25/xbow-ai-funding/

can use those same approaches to automatically find active vulnerabilities in systems and patch them before code is deployed. If they create "fraud-as-a-service tools", we can use AI to trace, disrupt, and dismantle those operations. AI can also help us stay ahead of threats by surfacing emerging tactics, spotting trends, scoring the risk of attack methods or vulnerabilities, and identifying patterns before they scale. This is how we move from human speed to digital speed in fighting cybercrime. Scaled offense necessitates scaled defense.

We should use AI to prevent and disrupt crime; however, it does not have to be everywhere all at once. We can be deliberate about where, when, and how we use it. We should use it to leverage its opportunity while mitigating risks like hallucination and model inaccuracy.

For example, using AI with humans in the loop to make decisions enables AI to improve the efficiency review process while the human makes the final decision. Additionally, AI can be built with transparency in mind: having systems that cite their sources, or show their reasoning, makes it easier for humans to review, question and audit the outputs. By building this way, we can leverage the speed and scale at which AI can understand data with the trustworthiness of human-expert review.

By finding ways to support AI-driven cybersecurity innovation and enabling public-private collaboration, Congress can take a massive step in ensuring defenders are not outpaced by attackers.

Institutionalize Red Teaming and Bug Bounty Programs for AI Tools

High-risk and general-purpose AI systems should undergo structured adversarial testing to identify misuse pathways before they cause harm. Like red teaming in cybersecurity, this helps uncover risky edge cases and misuse scenarios that standard evaluations may miss. Alongside this, funding bug bounty programs to reward responsible disclosure of jailbreaks, synthetic identity generation, and other forms of AI abuse. Running adversarial testing across a range of harmful use cases (e.g. physical crimes, malware, fraud, extortion) helps make it harder to use. Notably AI can help develop red teaming systems, and make it cheaper and easier to implement.

3. Align Institutions and Infrastructure Across Society

Support Public-Private Disruption of AI-Enabled Crime

Investing in infrastructure and legal tools can enable joint action between industry and government to disrupt platforms that support AI-driven fraud, scam distribution, or impersonation services. Examples like Information Sharing and Analysis Centers (ISACs) and the National Cyber-Forensics and Training Alliance (NCFTA) show how coordinated information sharing and disruption can work in practice. Efforts like the Internet Crime Complaint Center are also incredibly helpful in engaging a collective response based on real world harm. Reducing barriers to cooperation can make these partnerships faster, broader, and more effective.

Frame AI Policy Around Concrete Harms

Regulation should focus on clear, observable harms like impersonation scams, nudification abuse, and synthetic ID fraud. A harm-based approach allows policy to stay flexible as AI evolves, while staying aligned with how law enforcement investigates and responds to abuse.

Promote Shared Responsibility for Dual-Use Technologies

Addressing AI-enabled crime is a whole-of-society challenge. Companies building general-purpose AI tools, along with other infrastructure providers, should share responsibility for reducing foreseeable misuse. This includes safeguards such as access controls, usage monitoring, and clear reporting channels. For high-impact systems deployed at scale, small investments in prevention can significantly reduce downstream harm. As with other technologies that shape communication and access broadly, coordinated responsibility is essential to reducing misuse.

AI-enabled crime is evolving quickly. It is making threats more accessible, more personalized, and more difficult to detect. These are not one-size-fits-all problems. We are dealing with complex, general-purpose technologies that can be used to help or to harm. That complexity demands solutions that are just as comprehensive, tailored, and adaptive. I am optimistic that we can meet this challenge. With the right investment in education, innovation, public-private partnerships, and a whole-of-society approach, we can mitigate harm while enabling innovation. We can build a future where AI expands access to opportunity, strengthens safety, and helps people spend more time on what matters.

Thank you, and I look forward to your questions.

Addendum 1: Criminal Use by AI Capabilities

AI tools are being used to enable a wide range of criminal activity. The following examples outline key capabilities and the harms they enable.

1. Language and Cultural Fluency

Capability:

Large language models (LLMs) can generate fluent, grammatically correct, and culturally specific messages in dozens of languages. These models mirror the tone, structure, and formatting of government notices, company messages, or support requests. Jailbroken and open-source versions allow this capability to be used without guardrails that prevent malicious use[6].

Harm 1: Phishing for Login Information

For years, in cyber security training we taught people to look for obvious typos or grammatical errors in emails, but with AI the obvious indicators are disappearing . Criminals use AI to create phishing emails and scam texts that closely resemble legitimate messages. In one 2024 case, French drivers received toll payment scams written in fluent, region-specific French. The messages linked to a cloned toll agency website that looked official and used accurate branding[7] [8]. Similar toll/traffic scams are trending across the US[9]. These scams can be delivered as text messages, business emails or through other messaging platforms.

Harm 2: Job/Recruitment Scams

AI has also been used to enhance employment scams. Threat actors use LLMs to generate professional-looking job descriptions, interview instructions, and recruiter messages. These scams may aim to collect personal information such as Social Security numbers or bank details, or trick jobseekers into sending upfront payments for fake roles. The improved fluency and formatting make these messages more convincing and more difficult to detect, especially for younger or international applicants.[10,11].

[6] https://www.dhs.gov/sites/default/files/2024-10/24_0927_ia_aep-impact-ai-on-criminal-and-illicit-activities.pdf
[7] https://natlawreview.com/article/growing-cyber-risks-ai-and-how-organizations-can-fight-back
[8] https://cdn.openai.com/threat-intelligence-reports/5f73af09-a3a3-4a55-992e-069237681620/disrupting-malicious-uses-of-ai-june-2025.pdf
[9] https://www.fox13news.com/news/scammers-using-ai-improve-toll-text-message-scam-targeting-drivers-constantly-getting-smarter
[10] https://www.anthropic.com/news/detecting-and-countering-malicious-uses-of-claude-march-2025
[11] https://cdn.openai.com/threat-intelligence-reports/disrupting-malicious-uses-of-our-models-february-2025-update.pdf

Harm 3: Investment and Cryptocurrency Scams

Scammers use LLMs to craft persuasive language for fraudulent investment opportunities, including fake cryptocurrency platforms[12], trading schemes, and tech startups[13]. These scams are often distributed through social media and messaging apps or sent directly via text. AI-generated content allows scammers to mimic professional investment language, write promotional materials, and respond convincingly in real time. These scams frequently target younger adults or gig workers seeking financial growth or remote income opportunities.

Harm 4: Romance and Emotional Scams

Language models are increasingly used to automate romance scams and emotionally manipulative fraud. These scams involve building trust over time through AI-generated conversations on dating sites or messaging apps. With minimal effort, one scammer can manage dozens of victims using emotionally tailored messages.

Once a relationship is established, the conversation often shifts to urgent financial requests or fraudulent crypto investments. AI is used to maintain tone, personalize messages, and generate real-time responses in multiple languages.

Investigations have confirmed the use of LLMs in long-form romance scams and crypto fraud, with chatbots managing victim grooming and scripted emotional appeals. These scams are devastating to victims causing both incredible financial losses[14] and in some cases leading victims to suicide[15].

Romance scams can be conducted by a range of threat actors, including scam compounds that rely on coerced or trafficked labor[16].

2. Synthetic Image and Video Generation

Capability:

AI tools can generate or alter human images in ways that look realistic. This includes creating synthetic faces to generate people, editing images or videos through faceswapping, or voice-syncing and removing clothing from existing photos using nudification models. Nudification tools may either digitally undress a photo or take a nude image and faceswap the victims face onto the source image. These tools are available through a variety of platforms,

[12] https://xss.as/threads/140320/post-994654/31e9671f625e0f1021272eccf5aa3543

[13] https://cdn.openai.com/threat-intelligence-reports/5f73af09-a3a3-4a55-992e-069237681620/disrupting-malicious-uses-of-ai-june-2025.pdf

[14] https://www.dailynews.com/2024/10/13/how-pig-butchering-romance-scams-siphon-millions-from-californians-every-year/

[15] https://www.cnn.com/2024/06/17/asia/pig-butchering-scam-southeast-asia-dst-intl-hnk

[16] https://www.amnestyusa.org/press-releases/cambodia-government-allows-slavery-and-torture-to-flourish-inside-hellish-scamming-compounds/

ranging from open-source models that require technical expertise to user-friendly options like Telegram bots. These bots let users upload a photo and receive a manipulated image such as a nudified or face swapped version directly in the chat, without needing to write code or install software.

Harm 1: Generating Synthetic Faces for Fake Identities

Threat actors use AI generated photos for a variety of purposes. A common use is for ID fraud or "Know Your Customer" (KYC) bypass[17]. In this use case, threat actors create fake images to go with a fake ID and use it to register for a fake account that they may use for other crimes. Criminals pair these images with fake backstories and documents to pass onboarding checks[18]. Since new images are generated on demand, it is harder for platforms to flag or verify these images using traditional detection systems.

Harm 2: Nudification of Images

AI generated nude photos are often used for revenge, sextortion or harassment. Even though they are fake, they can be used for extortion by the sense of shame that would be felt if it were released in some case driving the victims to suicide[19]. In some cases this is cyberbullying from other students[20], in other cases it is remote and coordinated cyber crime gangs[21]. This is prevalent[22], Wired notes Telegram bots list more than 4 million monthly users for undressing photos[23].

Harm 3: Romance Scams

Image and video content add realism for romance scams[24]. Romance scams may be done by both individuals or organized crime networks.[25] AI tools for deepfake are directly listed for use for romance scams.

---

[17] https://darkforums.st/Thread-BYPASS-KYC-VERIFICATION-USING-DEEPFAKE-GUIDE?pid=86841#pid86841/6eece5704354d66c69bafc5e31f9fe73

[18] https://www.overwatchdata.ai/blog/the-dark-side-of-ai-fraudsters-new-arsenal

[19] https://www.wsj.com/tech/personal-tech/sextortion-scam-teens-apple-imessage-app-159e82a8?st=TWuAaq&reflink=desktopwebshare_permalink

[20] https://6abc.com/francesca-dorota-mani-ai-generated-pornographic-images-westfield-high-school-nj-legislation/14019614/

[21] https://www.europol.europa.eu/media-press/newsroom/news/25-arrested-in-global-hit-against-ai-generated-child-sexual-abuse-material

[22] https://www.iwf.org.uk/about-us/why-we-exist/our-research/how-ai-is-being-abused-to-create-child-sexual-abuse-imagery/

[23] https://www.wired.com/story/ai-deepfake-nudify-bots-telegram/

[24] https://frankonfraud.com/haotian-ai-providing-deepfake-ai-for-scam-bosses/

[25] https://www.wired.com/story/pig-butchering-scam-invasion/

Harm 4: Executive or Business Impersonation

Deepfake videos have also been used to impersonate business executives. Whether used to directly authorize the transfer of money (millions of dollars in some cases[26]) or as a lure to gain initial access to an organization[27] executive deepfakes can be effective since people have a general sense of their likeness but may not know them directly, they have trust and authority, and if a senior executive sends an employee something people are less likely to ask questions.

Harm 5: Scam Enablement

Other scams use celebrity deepfake imagery to lend credibility to fake or low quality products[28]. Using the likeness of a celebrity to support a new cryptocurrency, supplement, or product then sharing it on social media can help people buy it quickly, exploiting their trust in that celebrity's brand.

3. Voice and Face Cloning for Impersonation

Capability:

AI tools can clone a person's voice or simulate their face using only a few seconds of audio or video. The easiest clones operate from a few images or short audio clips using accessible online tools. These may be convincing at a glance or for short recordings. For more skilled actors who fine-tune models or run tools locally, more realistic clones can be produced with relatively little source material.

With more content, such as an hour and a half threat actors can generate "studio quality" voice clones from online tools that are realistic to friends of family. When we think about the availability of content on social media, let alone for public figures, making clones is incredibly achievable.

Voice cloning replicates tone, cadence, and emotional delivery with high accuracy. Face cloning tools generate realistic video by syncing a person's facial expressions and speech, often in real time. In combination with tools like DeepFaceLive and OBS Studio, threat actors can produce and stream live video content via transformation in addition to static content.

Harm 1: Family Emergency Scams Using Cloned Voices

Scammers use AI to impersonate relatives in distress, often targeting elderly victims. In one 2025 case on Long Island, fraudsters cloned the voice of a grandchild using audio from TikTok and voicemail greetings[29]. The victims of this fraud ring receive calls pretending to be their

[26] https://www.ft.com/content/b977e8d4-664c-4ae4-8a8e-eb93bdf785ea
[27] https://www.huntress.com/blog/inside-bluenoroff-web3-intrusion-analysis
[28] https://www.rollingstone.com/tv-movies/tv-movie-news/tom-hanks-warns-ai-deepfake-scam-likeness-1235092071
[29] https://nypost.com/2025/05/23/us-news/long-island-officials-warn-new-scam-uses-tiktok-and-ai-simulated-voices-to-impersonate-grandkids-and-rip-off-seniors

grandchild in jail, or in the hospital, in need of an urgent money transfer. The combination of hearing a loved one's voice, and our willingness to do anything for those we love makes this scam particularly effective. In some cases this is paired with imagery such as deepfake images of their child in a hospital bed.

Harm 2: Executive Impersonation in Financial Fraud

AI-generated face and voice clones have been used to impersonate corporate leaders in live video calls. In one case in 2025, attackers used a deepfake of a CFO[30] to instruct staff to process a $25 million transfer. The clone was realistic enough to convince the employee on the call, and again exploit our trust in what we see with our own eyes.

Harm 3: Bypassing Voice Authentication and Account Security

Voice cloning has also been used to attempt bypassing voice-based authentication systems used by banks and customer service centers. A cloned voice can be used to impersonate a customer, reset credentials, or approve financial actions over the phone. These attacks challenge the reliability of biometric security tools[31].

4. Code Generation and Exploitation

Capability:

AI models can generate and understand code. This can be used to generate malware more effectively, disguise who wrote it, and make variants so it is harder to detect. In many cases this is possible with mainstream hosted models since the line between good and bad use cases is not always obvious. In other cases, attackers use jailbreaks, which are prompting techniques that bypass safety filters, or fine-tuned open-source models that have been retrained on malicious content to enable more obviously harmful use cases[32]. On darkweb forums, users share both models finetuned for malware generations, and jailbreaks or tools to use open source models for malware generation[33], and guides to build your own “hacker assistants”.

Harm 1:Malware and Credential Theft

Threat actors have used AI tools to develop malware, both for those with minimal coding experience and for those with more advanced understanding of software systems. AI assists with development across multiple languages, helps attackers learn coding practices, and speeds up their overall workflow. Especially in malware use cases, threat actors may use multiple or stolen accounts to hide their activities across chat sessions to evade detection.

---

[30] https://www.ft.com/content/b977e8d4-664c-4ae4-8a8e-eb93bdf785ea
[31] https://ici.radio-canada.ca/nouvelle/2143502/reconnaissance-vocale-ia-banques-securite
[32] https://xss.as/threads/139109/post-986703/a11fe4ae01fe3bd1cc99e66c4af9606f ; https://github.com/BlackTechX011/HacxGPT-Jailbreak-prompts
[33] https://www.catonetworks.com/blog/cato-ctrl-wormgpt-variants-powered-by-grok-and-mixtral/ ; https://xss.as/threads/119869/post-976858/889bd2450ec700d2af069c5f3c949ba7

5. Learning

Capability:

AI is an incredible tool for learning. Standard LLMs can answer questions quickly and deep research agents can search for source material. Some models may integrate custom data sets to enable fraud. Threat actors can use these tools to learn how to commit crimes or to research their targets, including individuals, software, and technical systems. Tools like WormGPT, or jailbreaks which use specially crafted prompts to avoid safety classifiers, enable threat actors to directly ask "How do I blackmail someone?", "How do I get started in fraud?", "How do I make a fake login page for a bank?, or "Tell me about vulnerabilities in industrial control systems"

Harm 1: Reducing Barriers to Entry for Crime

Threat actors big and small, use AI to level up their cybercrime. Whether using it to research a victims industry, online footprint, or software systems[34] or to learn how to carry out crime. When used for research it lowers barriers to entry for cyber crime and makes threat actors more effective more quickly. While this is most notably enabling threat actors new to cyber crime it is also used by state-sponsored groups[35].

[34] https://cdn.openai.com/threat-intelligence-reports/5f73af09-a3a3-4a55-992e-069237681620/disrupting-malicious-uses-of-ai-june-2025.pdf

[35] https://www.securityweek.com/openai-says-iranian-hackers-used-chatgpt-to-plan-ics-attacks/

Addendum 2: Glossary

- **AI Agent:** An AI system that can perform tasks autonomously or semi-autonomously, including planning, decision-making, and using software tools. Malicious use includes automating scams, reconnaissance, or attacks.
- **Computer Use Agent:** An AI system that mimics human computer use, such as browsing, filling out forms, or running applications. These can be abused to automate fraud or simulate human behavior online.
- **Evil GPT:** A nickname for language models modified or prompted to produce harmful or illegal outputs. Often created via jailbreaking or fine-tuning for use in scams, harassment, or hacking.
- **Faceswap:** An AI technique that replaces one person's face with another in images or video. Can be used for impersonation, misinformation, or non-consensual explicit content.
- **Fine-Tuning:** Training an existing AI model on specific data to adjust its behavior. In malicious use cases, this includes stolen data, explicit images, or criminal content.
- **Jailbreak:** A method for bypassing an AI model's safety restrictions (typically with specially crafted prompts) to generate prohibited outputs like instructions for illegal activity.
- **Know your Customer (KYC)**: A set of processes used by financial institutions and regulated businesses to verify the identity of their clients. KYC helps prevent fraud, money laundering, and other illegal activities by requiring customers to provide personal information and documentation before accessing services.
- **Nudifying:** The use of AI to digitally remove clothing from a photo, creating fake nude images. Often used for extortion, harassment, or intimidation.
- **Romance Scams:** Fraud involving fake romantic relationships used to manipulate victims for money or sensitive information.
- **Sextortion:** Blackmail involving threats to release sexual images or content unless demands are met, often including money or more explicit material.
- **Voice Clone:** AI-generated speech that replicates a person's voice from audio samples. Can be abused for impersonation, scams, or fraud.
- **World Models:** AI systems that simulate how the physical world works. These can be used for planning, decision-making, or malicious actions like physical intrusion or infrastructure attacks.