Cybersecurity With Artificial Intelligence: Keeping the Internet Safe

Dorian Choy, Class of 2026

Introduction

Dorian Choy
Class of 2026 at The Bronx High School of Science

- Name: Dorian Choy
- Senior at the Bronx High School of Science
- Aiming to major in Computer Science
- My intended area of study is Cybersecurity with Artificial Intelligence as an additional focus.
- Relevant classes:
  AP Computer Science A,
  Artificial Intelligence for Game Programming,
  Multivariable Calculus,
  Ordinary Differential Equations

Why Am I Interested In Cybersecurity?:

- Previously, I had filed a ticket with Discord, which included sending my passport.
- In September-October 2025, Discord experienced a data breach, resulting in ~70,000 users having their government ID photos exposed.
- As a result, I found myself paranoid about my personal information being leaked.
- This made me interested in focusing on cybersecurity, as I previously already had an interest in Computer Science.
- As the AI industry grows, we must also focus on its effect on cybersecurity, whether as a tool or a target.

Future Goals:

- I intend to become a cybersecurity analyst in the future.
- My goal is to help prevent data breaches, both on a social level and technical level.
- Furthermore, I aim to work under companies such as Anthropic(Claude) and OpenAI(ChatGPT)
- Anthropic: "Research Engineer – Cybersecurity RL", "Privacy Research Engineer, Safeguards"
  OpenAI: "Software Engineer, Privacy", "Software Engineer, Security Observability"
- In the future, I plan to create my own company, in order to aid smaller businesses with data protection.

Thank you!

Any questions?

Viruses:

- Malicious code written to cause damage or gain unauthorized access:

AI Usage:

- Adaptive Malware(Self-Rewriting code, Code generated at Runtime)

Effects on Cybersecurity:

- Unable to scan/detect malware as easily

Examples:

- BlackMamba(Proof of Concept)

Phishing:

- A method of committing malicious actions by tricking users, mainly by impersonating someone (e.g. scam emails, fake texts, etc.)

AI Usage:

- Deepfakes(Body, Voice, Writing Style)

Effects:

- Much harder to detect imposters

Examples:

- Arup Scam(2024, $25.6 Million), Florida mother "Daughter in Distress"(July 2025, $15k lost)

SQL Injection:

- Taking advantage of unprotected user inputs to send malicious code

AI Usage:

- AI as a Target(AI Agents, RAG poisoning, MCP attacks)
- AI as a Tool(Payload refinement, Dataset generation)

Effects:

- Companies that have AI tools are more vulnerable to such attacks
- It is also easier to create SQL attacks

Examples:

- LangChain SQLDatabaseChain Vulnerability (CVE-2023-36189), Flowise AI Vector SQL Injection (CVE-2025-29189)

CSRF:

- Cross-Site Request Forgery is a web attack that tricks a logged-in browser into performing unwanted actions.

AI Usage:

- AI as a Target(Vulnerable to injection prompts, Agentic AI)

Effects:

- Vulnerability of AI allows for data to be easily leaked

Examples:

- Comet AI Browser(Perplexity)

AI as a Tool: GTG-1002 Incident

- First reported case of an AI-orchestrated cyber espionage campaign.
- A threat group (designated as GTG-1002) used Anthropic's Claude code model to perform espionage on global targets, including large tech companies, financial institutions, and government agencies.
- This was done by manipulating the AI to avoid any security filters("jailbreaking")
- With little input, Claude performed over 80-90% of operations, acting at speeds impossible to human operators.

Work Cited:

Agentic Browser Security: Indirect Prompt Injection in Perplexity Comet | Brave. (2025, August 20). Brave. https://brave.com/blog/comet-prompt-injection/

Brunner, R. (2025). Woman Conned Out of $15K After AI Cloned Her Daughter's Voice in Terrifying Scam: "I Broke Down." People.com. https://people.com/woman-conned-out-of-usd15k-after-ai-cloned-daughters-voice-terrifying-scam-11775622

CISA. (2021, February 1). What is Cybersecurity? Cybersecurity and Infrastructure Security Agency CISA; CISA. https://www.cisa.gov/news-events/news/what-cybersecurity

Disrupting the first reported AI-orchestrated cyber espionage campaign. (2025a). https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf

Disrupting the first reported AI-orchestrated cyber espionage campaign. (2025b). Anthropic.com. https://www.anthropic.com/news/disrupting-AI-espionage

Elliott, D. (2025, February 4). Cybercrime: Lessons learned from a $25m deepfake attack. World Economic Forum. https://www.weforum.org/stories/2025/02/deepfake-ai-cybercrime-arup/

Filip TRUȚĂ. (2025). Florida Woman Loses $15K to AI Voice Scam Mimicking Daughter in Distress. Bitdefender. https://www.bitdefender.com/en-us/blog/hotforsecurity/florida-woman-loses-15k-to-ai-voice-scam-mimicking-daughter-in-distress

Nolan, B. (2025, December 23). OpenAI says prompt injections that can trick AI browsers like ChatGPT Atlas may never be fully "solved"—experts say risks are "a feature not a bug." Fortune. https://fortune.com/2025/12/23/openai-ai-browser-prompt-injections-cybersecurity-hackers/

NVD - CVE-2023-36189. (2023). Nist.gov. https://nvd.nist.gov/vuln/detail/CVE-2023-36189

NVD - CVE-2025-29189. (2025). Nist.gov. https://nvd.nist.gov/vuln/detail/CVE-2025-29189

Prakash, P. (2024, May 17). A deepfake "CFO" tricked the British design firm behind the Sydney Opera House in $25 million scam. Fortune Europe; Fortune. https://fortune.com/europe/2024/05/17/arup-deepfake-fraud-scam-victim-hong-kong-25-million-cfo/

Sims, J. (2023, July 31). BlackMamba: Using AI to Generate Polymorphic Malware. Www.hyas.com. https://www.hyas.com/blog/blackmamba-using-ai-to-generate-polymorphic-malware