



日本のサイバーセキュリティを「連携」「学び」「創造」

JNSA活動報告会 パネルディスカッション

サイバーセキュリティ におけるAI活用の今

情報セキュリティ大学院大学
大塚 玲

サイバーセキュリティの領域はAIによりどうなる？AI vs AIの未来？



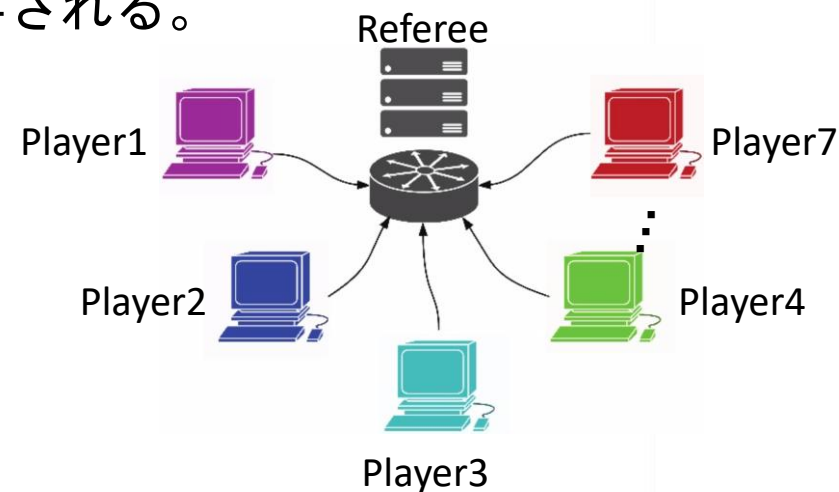
DARPA AI Cyber Challenge Aims to Secure Nation's Most Critical Software

New competition challenges the nation's top AI and cybersecurity talent to automatically find and fix software vulnerabilities, defend critical infrastructure from cyberattacks

OUTREACH@DARPA.MIL
8/9/2023

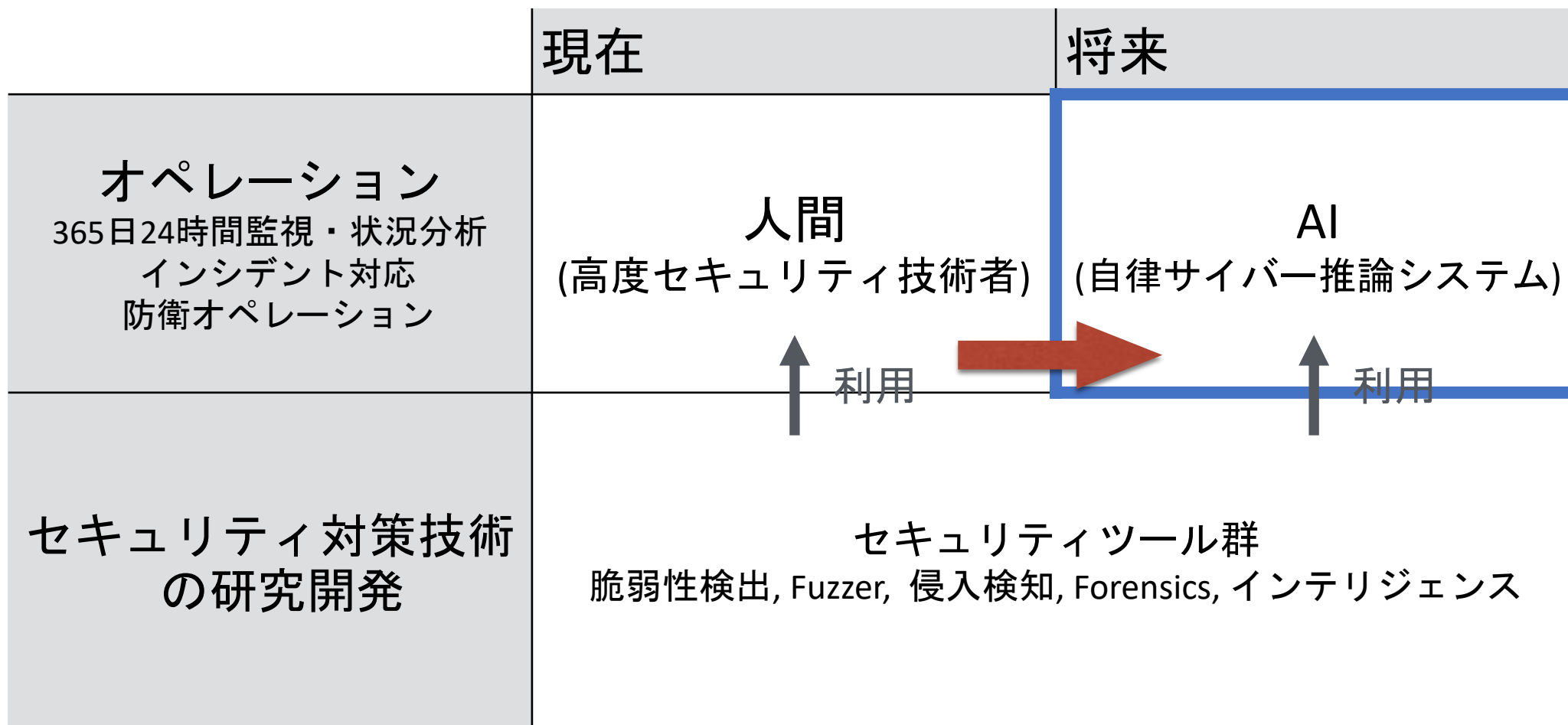


- 米国防省の研究部門であるDARPA（米国防高等研究計画局）が2023年8月、ソフトウェア脆弱性の自動修正システム競技大会「AI Cyber Challenge」（AIxCC）開催を発表。
- 2025年8月の「DEF CON 33」で決勝大会が開催される。優勝チームには400万ドル、2位には300万ドル、3位には150万ドルの賞金が授与される。

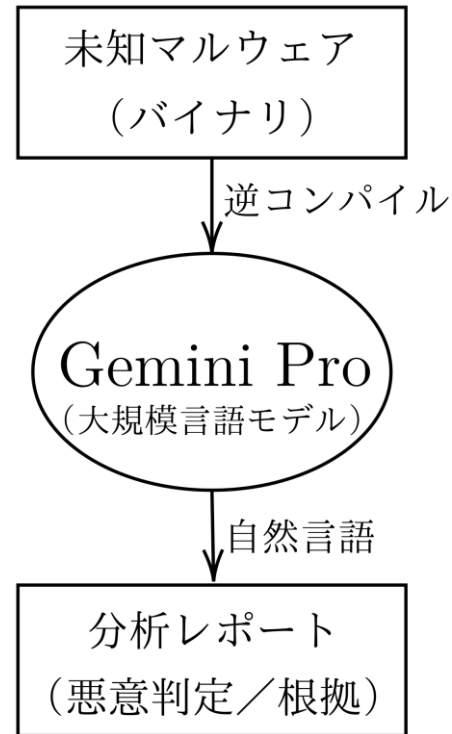


サイバーセキュリティの領域はAIによりどうなる？

AI vs AIの未来？



ChatGPTとかClaudeCodeのように汎用的な生成AIでセキュリティ診断はできる？



※Gemini Pro1.5は入力コンテキストが1Mトークンに拡大したため、逆コンパイルコード全体の解析が可能になった。

Bernardo Quintero. From Assistant to Analyst: The Power of Gemini 1.5 Pro for Malware Analysis. <https://cloud.google.com/blog/topics/threat-intelligence/gemini-for-malware-analysis> (2024).

Model

Verdict: Malicious

This decompiled executable code exhibits several malicious functionalities commonly associated with ransomware.

Activities:

File Operations:

- Reads a file named "..."
- Creates and writes to ... or perform other malicious actions.
- Creates directories and files: The malware creates directories in the ProgramData, and Intel folders. It also ...
- Sets file attributes: The malware sets ...
- Moves and deletes files: The malware ...

Network Activity:

- Generates IP addresses: The malware generates seemingly random IP addresses and attempts to connect to them on port 445, which is commonly associated with the Server Message Block (SMB) protocol. This suggests the malware may attempt to spread to other computers on the network.
- Connects to a specific URL: The code references a specific URL (<http://www.iuqerfsodp9ifjaposdfjhgosurijfaewrwergwea.com>) which could be used for command and control or data exfiltration.

Registry Operations:

- Creates registry keys: The malware creates registry keys under "Software\WanaCrypt0r" in both HKCU and HKLM hives.
- Reads and writes registry values: The malware reads and writes values to the created registry keys, potentially to store information about the infected system or its configuration.

Other Activities:

判定：ランサムウェア系マルウェア

ファイル操作

Wannacry型マルウェアの痕跡を検出

- c.wnry: マルウェアの設定情報
- tasksche.exe: マルウェア実行ファイル

セキュリティ領域でのAIの正確性はどこまで上げられるか？ 人間と比較してどの程度信頼できる？



- CYBENCH: <https://cybench.github.io>, Zhang et. al at ICLR2025.

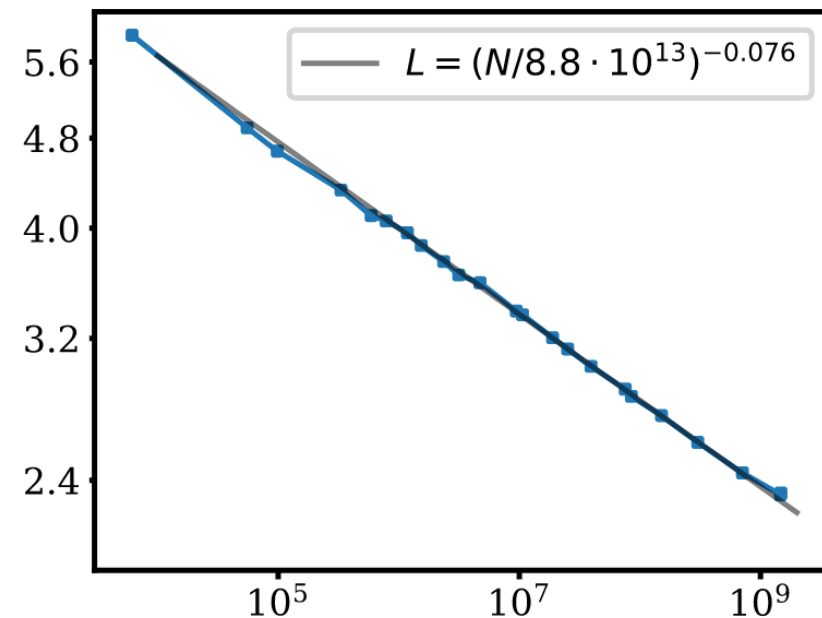
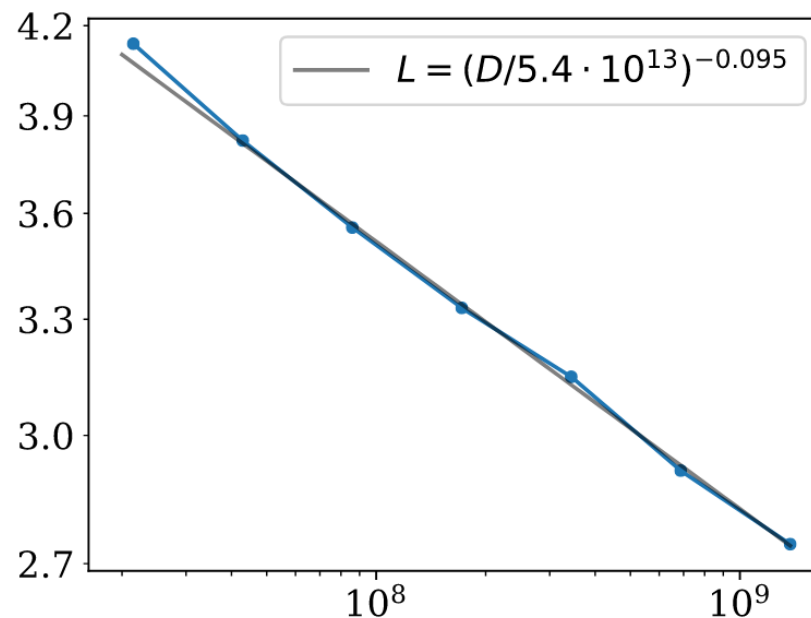
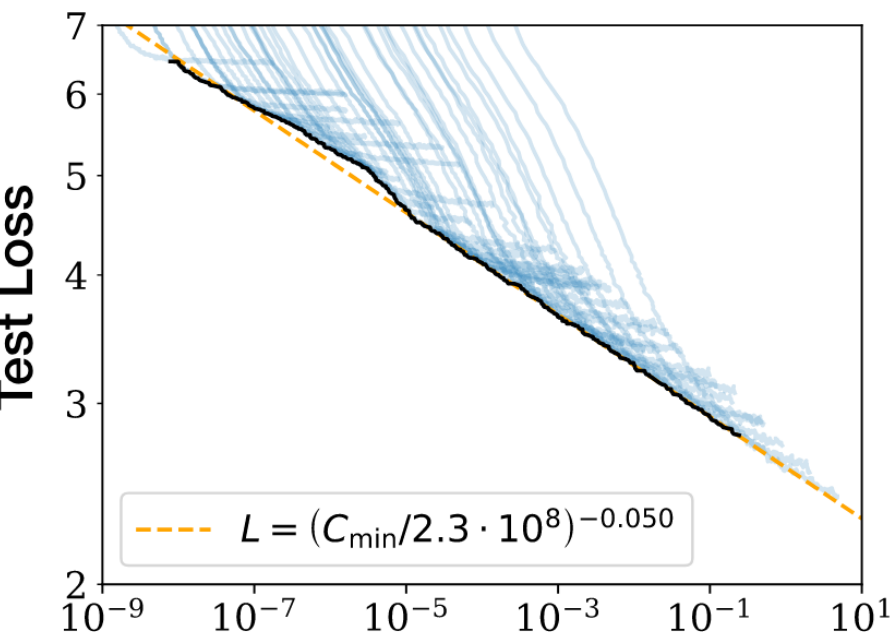
Model	Unguided % Solved	Subtask-Guided % Solved	Subtasks % Solved	Most Difficult Task Solved (First Solve Time)	
				Unguided	Subtask-Guided
OpenAI o3-mini *†	22.5%	--	--	42 min	--
Claude 3.7 Sonnet *	20%	--	--	11 min	--
GPT-4.5-preview *	17.5%	--	--	11 min	--
Claude 3.5 Sonnet	17.5%	15%	43.9%	11 min	11 min
GPT-4o	12.5%	17.5%	28.7%	11 min	52 min
OpenAI o1-mini *†	10%	--	--	11 min	--
Claude 3 Opus	10%	12.5%	36.8%	11 min	11 min
OpenAI o1-preview	10%	10%	46.8%	11 min	11 min

セキュリティ領域でのAIの正確性はどこまで上げられるか？ 人間と比較してどの程度信頼できる？



スケーリング則

Kaplan, J. *et al.* Scaling laws for neural language models. arXiv:2001.08361 [cs.LG] at <https://arxiv.org/abs/2001.08361> (2020)



Compute

PF-days, non-embedding

Dataset Size

tokens

Parameters

non-embedding

言語モデリングの性能は、モデルサイズ、データセットサイズ、および訓練に使用される計算量を増やすにつれてスムーズに向上する。最適な性能を得るためには、これら3つの要素を同時に拡張する必要がある。他の2つの要素によってボトルネックが発生しない場合、経験的な性能（評価損失）はべき乗則（power-law）に従う。

リソースを投入すればいくらでも性能を高められる

