情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

大規模言語モデルを用いたサイバー攻撃対応の動向

2023年11月24日

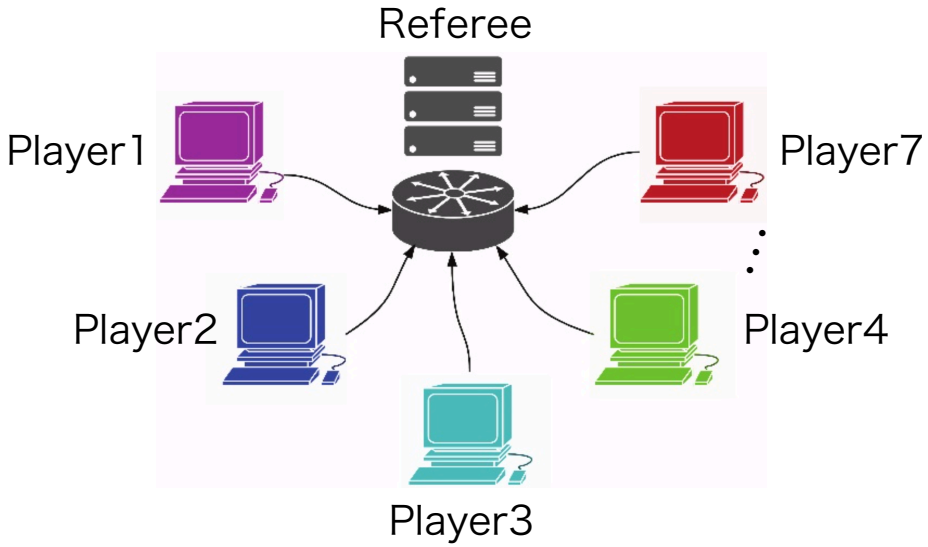
大塚　玲

Cyber Grand Challenge（CGC）

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

- 概要　専用のネットワーク・コンピュータで、自分のシステムを守り、相手を攻撃する
  全自動ハッキングコンテスト
- 主催　DARPA（米国防高等研究計画局）
- 場所　ラスベガス, DEFCON24(2016年)
- 予算　約56億円（賞金約10億円）

■ ルール
- 攻防戦形式のCTF
- 攻撃成功で加点、防御失敗で減点



行動決定戦略に(初等的な)ベイズ推定を使用

■ 優勝者　For All Secure（CMU）
■ システム名　Mayhem

【NEW】AI Cyber Challenge

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

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

■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。
■競技大会をサポートするのは、生成AIのトップ企業であるAnthropic、Google、Microsoft、OpenAIの4社
■まず来年（2024年）5月に予選を行い、上位20チームが8月のセキュリティイベント「DEF CON 32」併催の準決勝大会に進出。その上位5チームには開発資金200万ドルが与えられ、再来年（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。

研究の目標

サイバー攻撃の高度化・巧妙化

深層強化学習＋大規模言語モデルを用いて

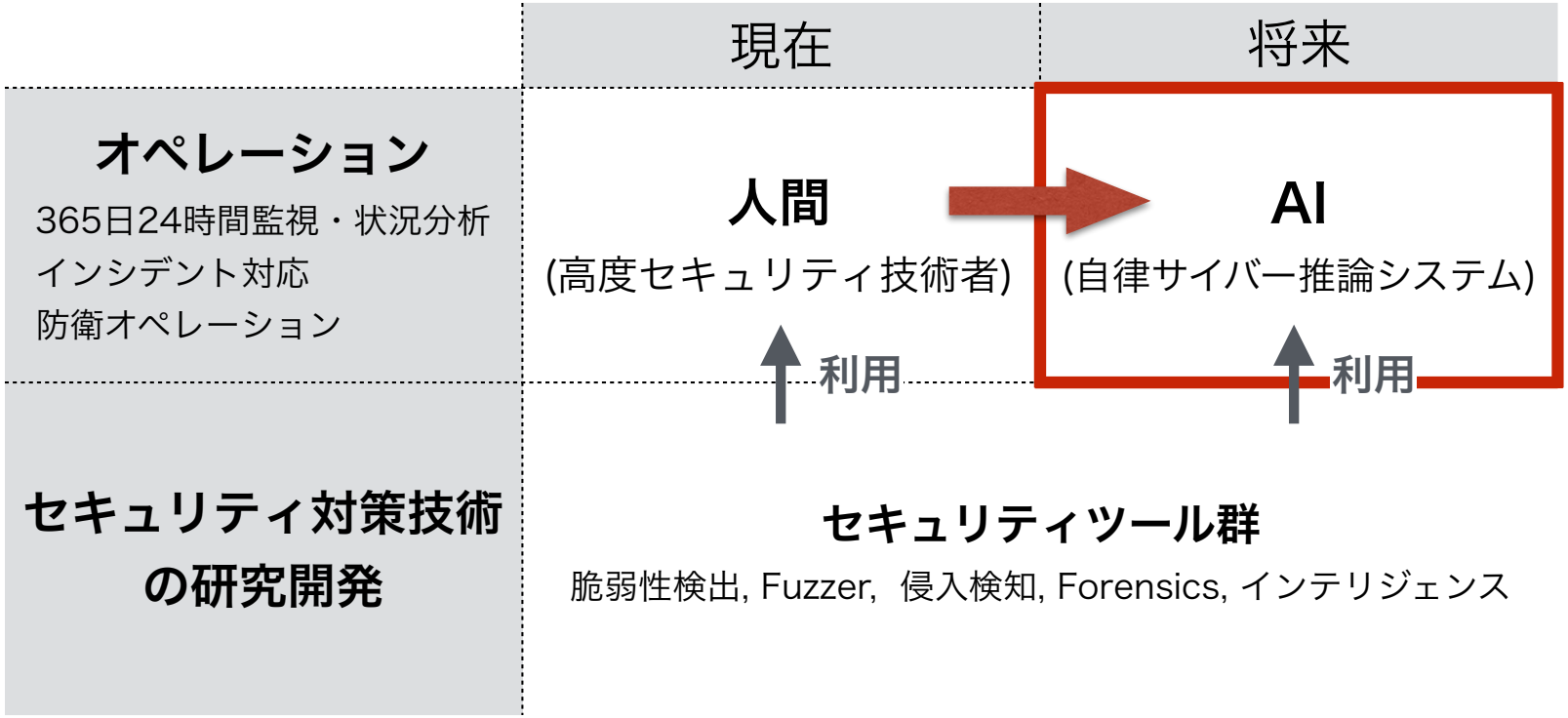
自律サイバー推論システムができるか？

Cyber Reasoning System

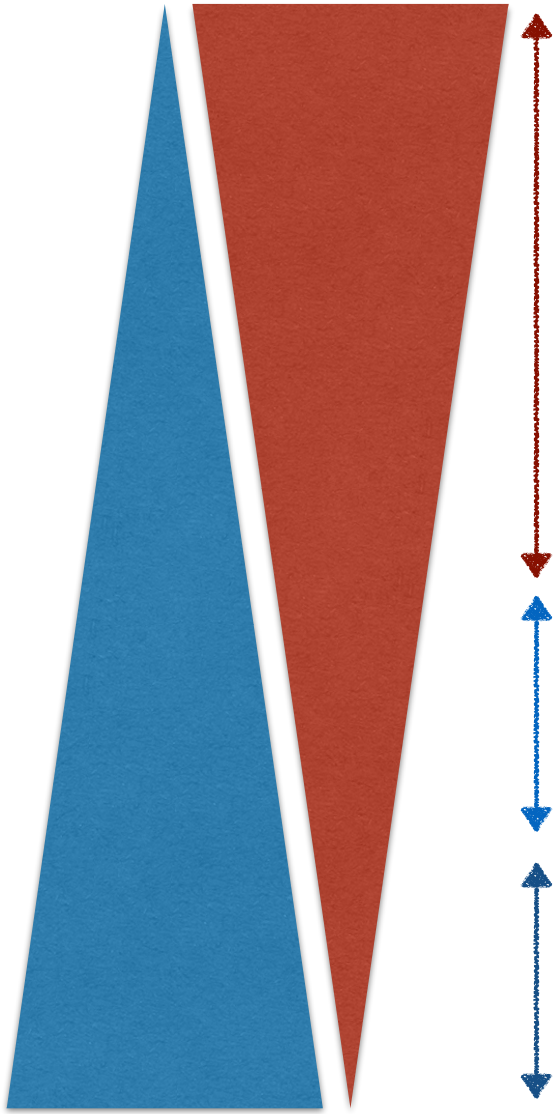
・ソフトウェアの脆弱性検知
・パッチ作成
・エクスプロイト作成　等の自動対応

AIによるサイバー攻撃対応

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY



- 囲碁・将棋等でのAI(深層強化学習)の活躍を踏まえれば，人間の能力の限界を超えた高度なサイバー攻撃対処を実現できる可能性が高い.
- これまで少数の情報セキュリティ技術者しか為し得なかった高度なサイバー攻撃対処が24時間365日連続かつ大規模分散的に可能になる.
- 将来的には，意思決定領域を除くサイバー攻撃対処の完全自動化が期待できる.



大規模システム　医療カルテシステム　公共システム

攻防のイタチごっこ

大規模言語モデルによる
サイバー攻撃対応
に期待できるか？

小規模システム
暗号プロトコル　TLS, IoT　スマートコントラクト

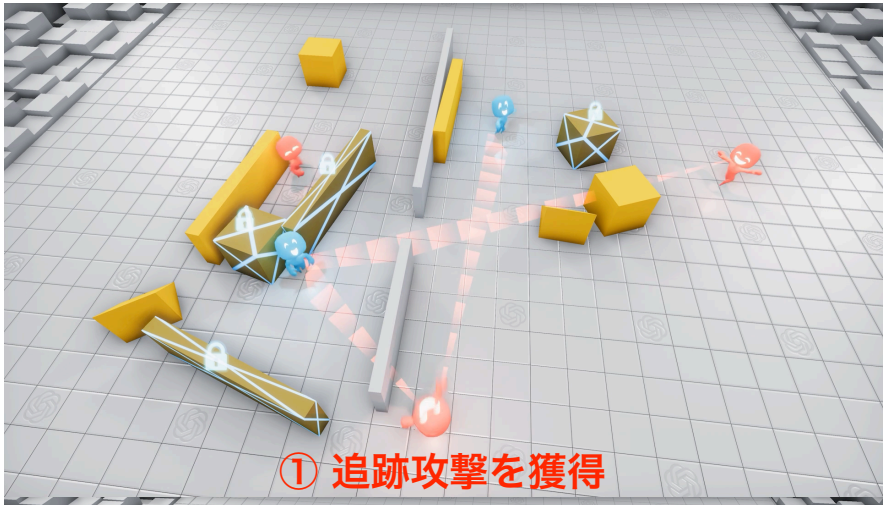
定理証明器/モデル検査
(フォーマルメソッド)

暗号プリミティブ　暗号，署名

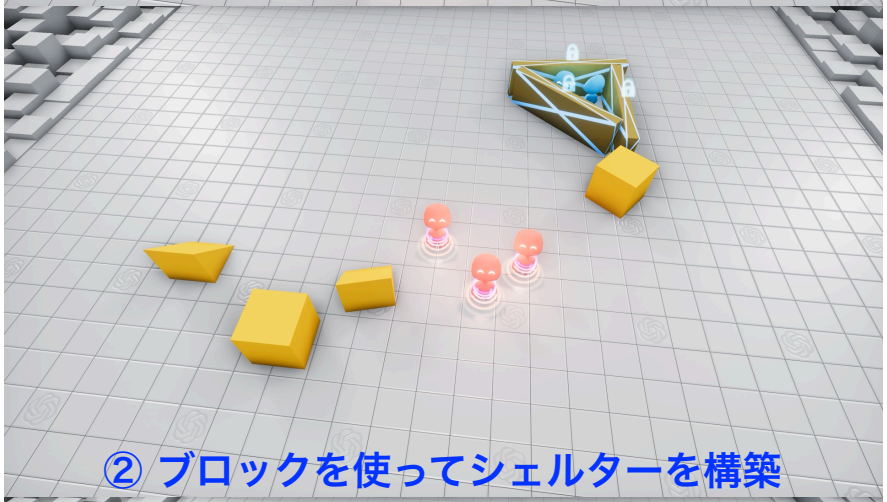
証明可能安全性
(困難な問題に帰着)

深層強化学習による攻防知識の獲得例

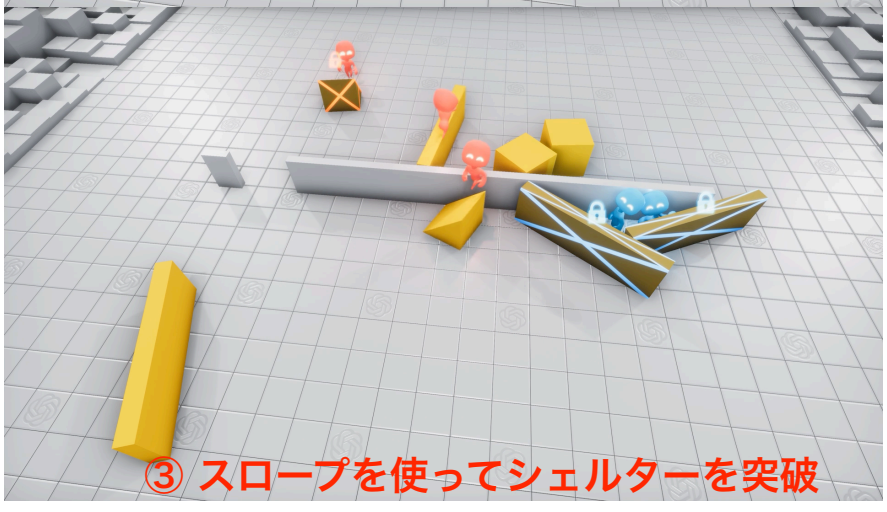
Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I., Emergent tool use from multi-agent autocurricula. *ICLR2020, 2020*.



① 追跡攻撃を獲得

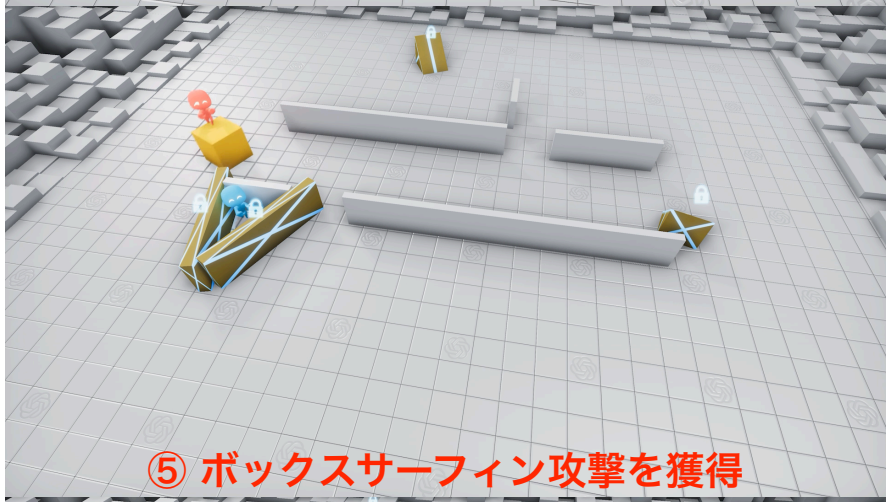


② ブロックを使ってシェルターを構築

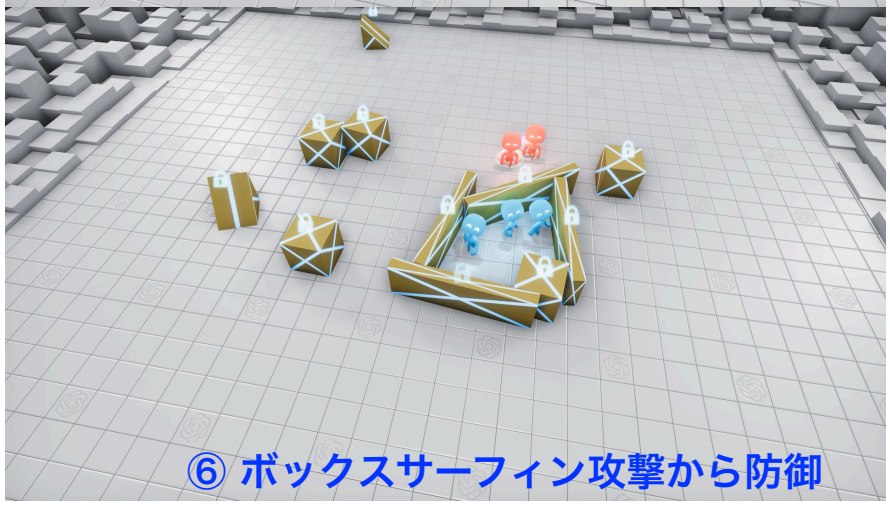


③ スロープを使ってシェルターを突破

④ スロープを固定してスロープ攻撃から防御



⑤ ボックスサーフィン攻撃を獲得

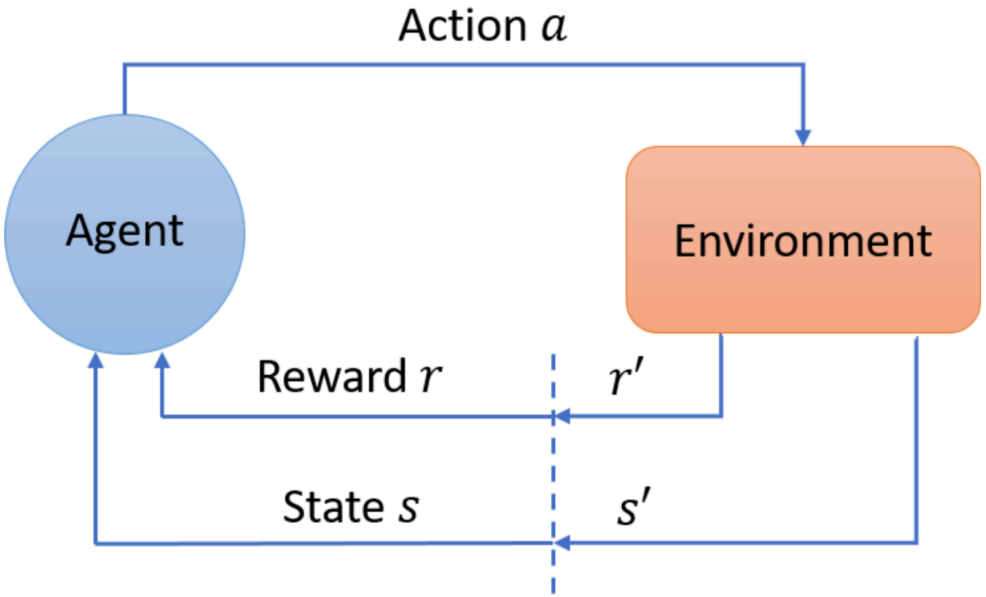


⑥ ボックスサーフィン攻撃から防御

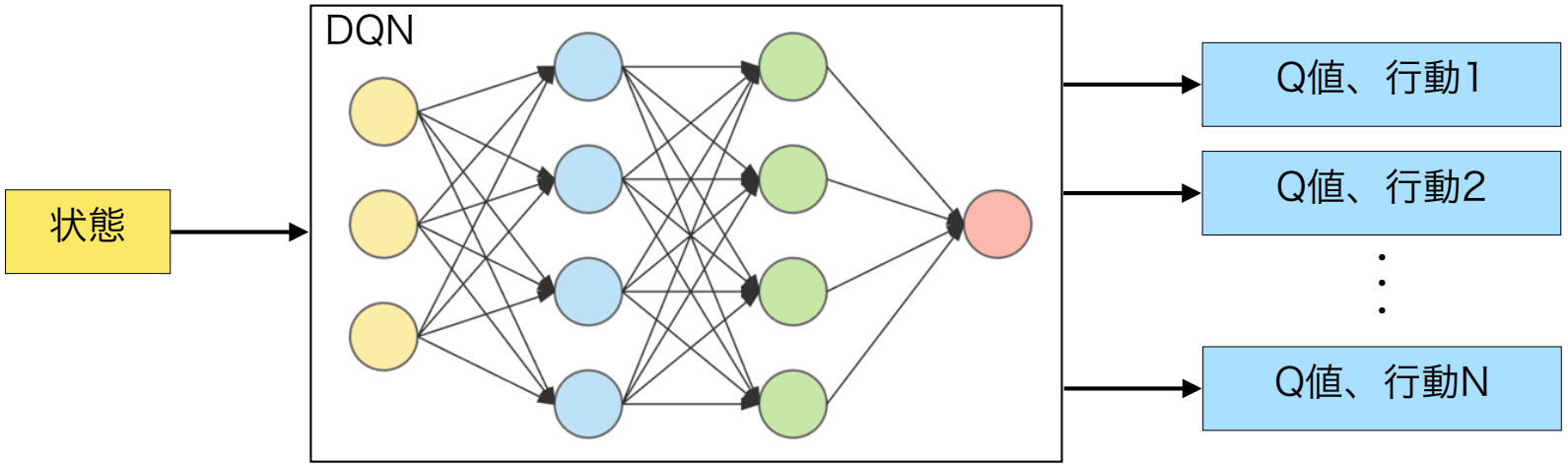
IISEC
Institute of Information Security

深層強化学習

- 強化学習（Reinforcement Learning）とは
  - 試行錯誤により報酬を最大化（例：Q学習）
  - 状態、行動、報酬の3つの概念
  - 行動→状態更新、報酬（ペナルティ）受け取り



- 深層強化学習（Deep Reinforcement Learning）とは
  - 強化学習と深層学習を組み合わせたもの（例：DQN）
  - DQN（Deep Q-Network）
    Q関数（予測報酬値）をDeep Learning Networkで近似

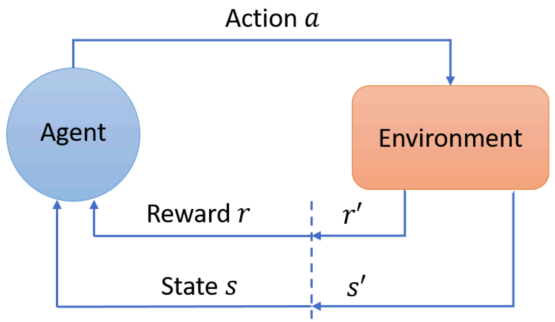


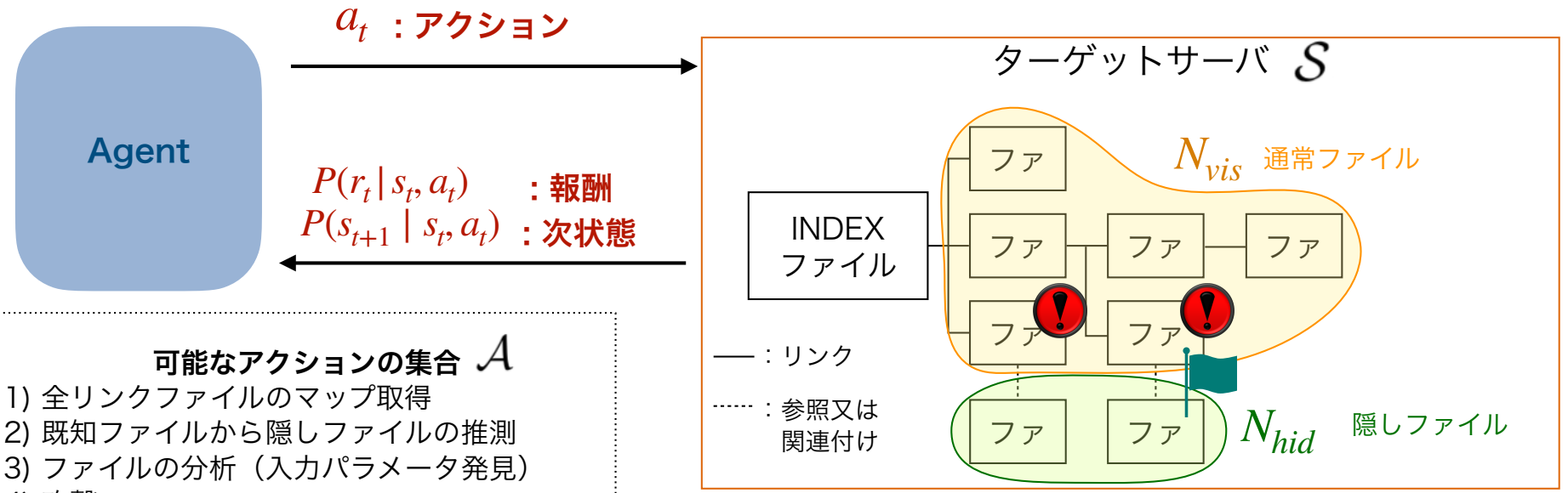
先行研究①

Zennaro, Fabio Massimo and Laszlo Erdodi. “Modeling Penetration Testing with Reinforcement Learning Using Capture-the-Flag Challenges and Tabular Q-Learning.” ArXiv abs/2005.12632 (2020.5.26).

■ 深層強化学習でCTFを解くアプローチを提案

- CTF問題の単純化：少数のアクションの組み立てで解ける単純化したCTF問題を対象
- Moving Target：　行動に応じて解が変化するCTF問題も対象
- 模倣学習：　セキュリティ技術者の行動を手本に学習





可能なアクションの集合 $\mathcal{A}$

1) 全リンクファイルのマップ取得
2) 既知ファイルから隠しファイルの推測
3) ファイルの分析（入力パラメータ発見）
4) 攻撃

先行研究②

Adolphs, L., & Hofmann, T. (2020, February). LeDeepChef: Deep Reinforcement Learning Agent for Families of Text-Based Games. In *34th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI 2020)* (p. 5180).

You are hungry! Let's cook a delicious meal. Check the cookbook in the kitchen for the recipe. Once done, enjoy your meal!

-= Kitchen =- You're now in the kitchen. I guess you better just go and list everything you see here. You hear a noise behind you and spin around, but you can't see anything other than a fridge. The fridge is empty! This is the worst thing that could possibly happen, ever! You see a closed oven. You make out a table. The table is massive. On the table you can make out a cookbook. You can make out a counter. The counter is vast. On the counter you can make out a sliced fried red hot pepper, a red apple, a yellow bell pepper and a knife. You smell an intriguing smell, and follow it to a stove. You shudder, but continue examining the stove. But the thing is empty, unfortunately. Hm. Oh well.

> examine recipe
You open the copy of "Cooking: A Modern Approach (3rd Ed.)" and start reading:
Recipe #1

Gather all following ingredients and follow the directions to prepare this tasty meal.
Ingredients:
red hot pepper
Directions:
slice the red hot pepper
fry the red hot pepper
prepare meal

> inventory
You are carrying nothing.

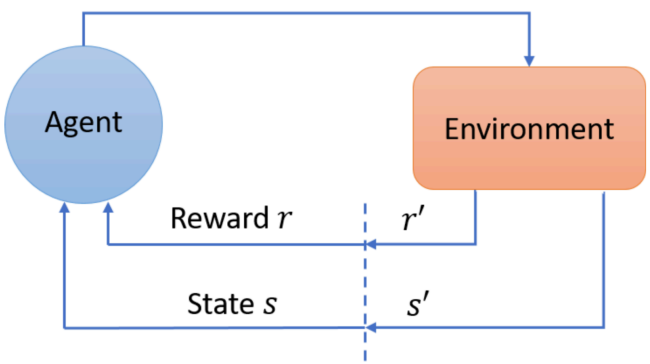
> take red hot pepper
You take the red hot pepper from the counter.
Your score has just gone up by one point.

> inventory
You are carrying:
a sliced fried red hot pepper

> prepare meal
Adding the meal to your inventory.
Your score has just gone up by one point.

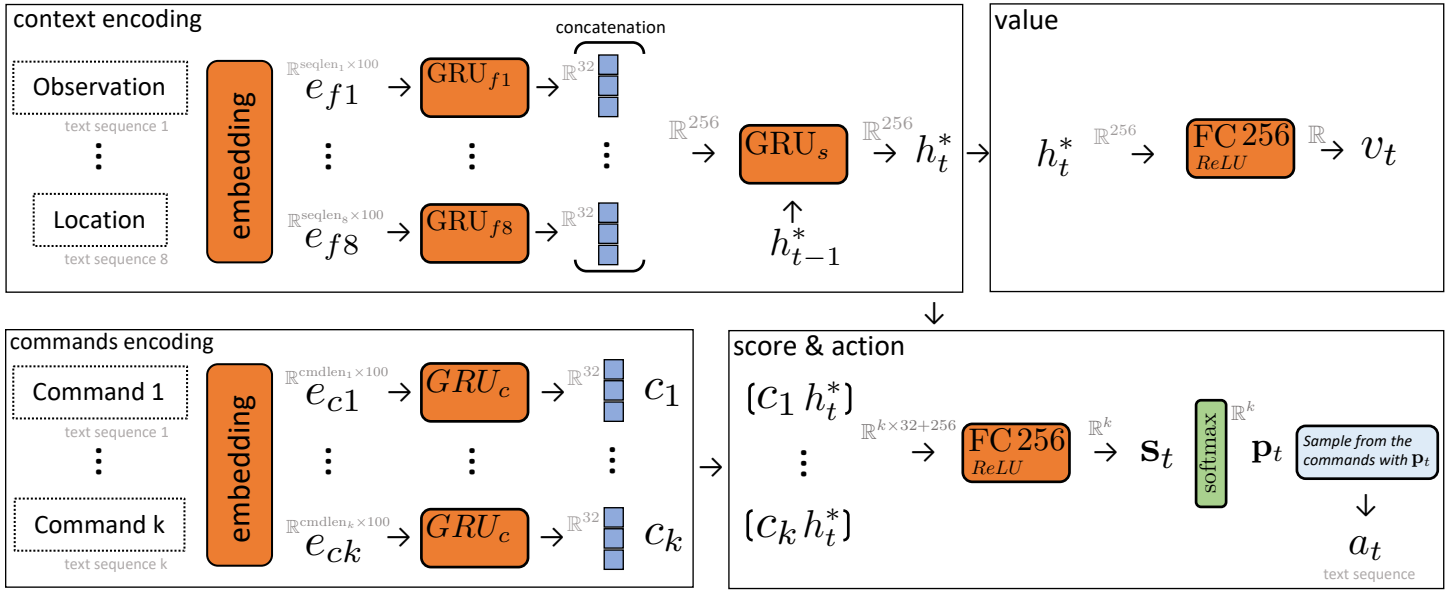
> eat meal
You eat the meal. Not bad.
Your score has just gone up by one point.

*** The End ***



■ 部分観測マルコフ決定過程

- レスポンス文章のベクトル化→状態推定
- 利用可能コマンド集合のベクトル化
- 状態xコマンド集合 →実行コマンド



先行研究③

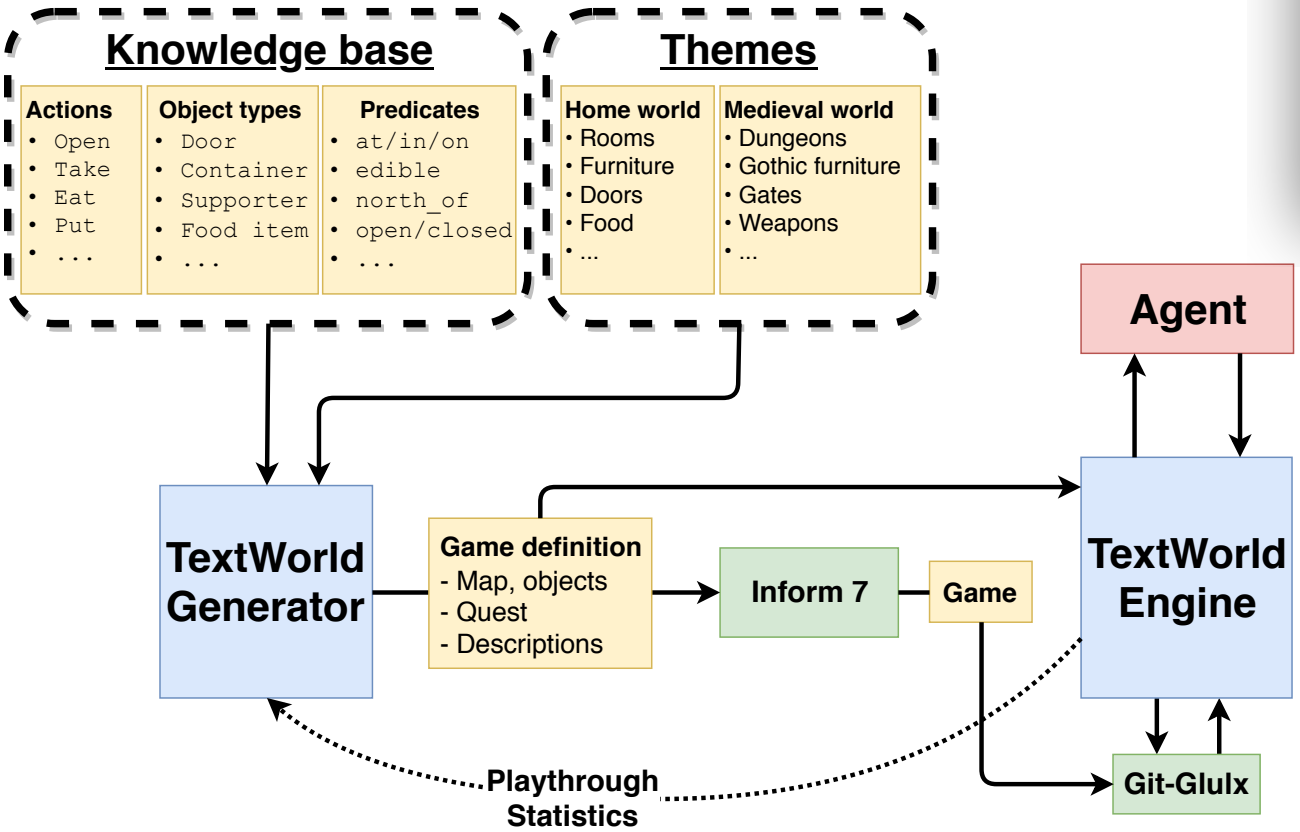
Barnes, T., Fine, E., Moore, J., Hausknecht, M., El Asri, L., Adada, M., ... & Trischler, A. (2019, June). TextWorld: A Learning Environment for Text-Based Games. In Computer Games: 7th Workshop, CGW 2018, IJCAI 2018, July 13, 2018, Revised Selected Papers (Vol. 1017, p. 41). Springer.

■ テキストゲームのランダム生成

- アクションとレスポンスをテキストでやりとり
- 同種の問題を大量(ランダム)に生成可能

■ 深層強化学習

- 同種の問題を大量に解いて，ポリシーを学習
- Textworld = 学習プラットフォーム



■ CTFのランダム生成

- SecGen@ASE'17

Z. Cliffe Schreuders, et al.: Security Scenario Generator (SecGen): A Framework for Generating Randomly Vulnerable Rich-scenario VMs for Learning Computer Security and Hosting CTF Events. ASE @ USENIX Security Symposium 2017

研究シナリオ

まずここから

第一段階：CTFを解く自律サイバー推論システムの構築

CTF(Capture The Flag)は，コンテスト形式で情報セキュリティ技術者の訓練に一般的に用いられる問題である．与えられたUnix等のリモート環境下で，バイナリ解析等のハッキング手法を用いて，目標とする解答(Flag)を得るシステムを構築し，セキュリティ技術者と競ってコンテスト上位入賞を目指す．

第二段階：複数の自律サイバー推論システムによる敵対的学習

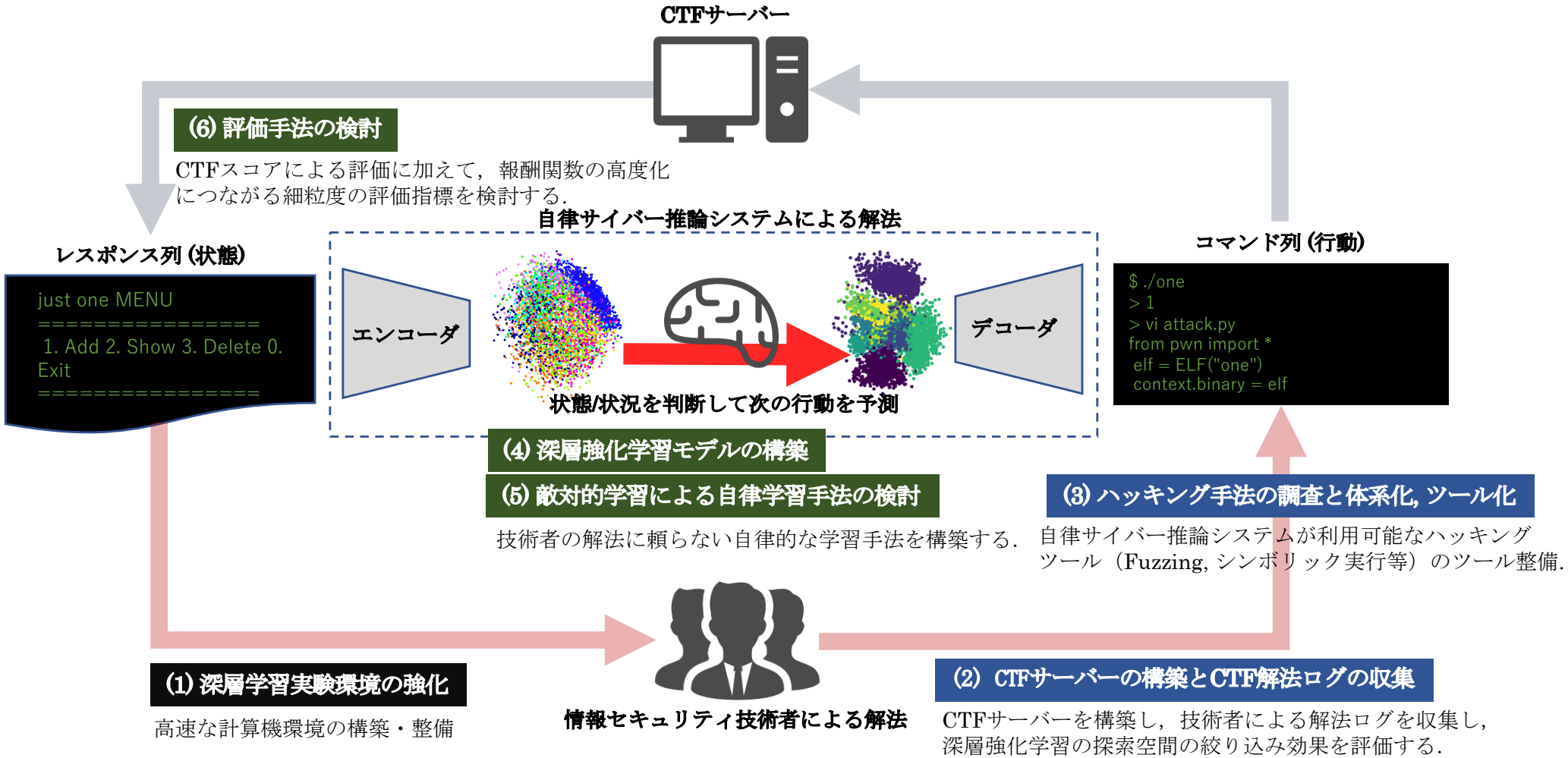
複数の自律サイバー推論システムを互いに競わせることでサイバー防衛性能の強化を図る．攻撃または防衛に成功すれば，その知識を全システムに共有することで，全体の知識が加速度的に強化される．

第三段階：インターネット空間での実運用実験

複数の自律サイバー推論システムをネットワーク状にインターネット空間に配置し，外部からの攻撃試行を学習し，効果的な防衛・反撃法を学習する．実際のサイバー攻撃では，民間の商用サーバー等に寄生して攻撃を仕掛けてくる可能性が高いことから，有効な攻撃目標の特定に注意したシステムの構成法が求められる．

(アプローチ)深層強化学習による自律サイバー推論システム

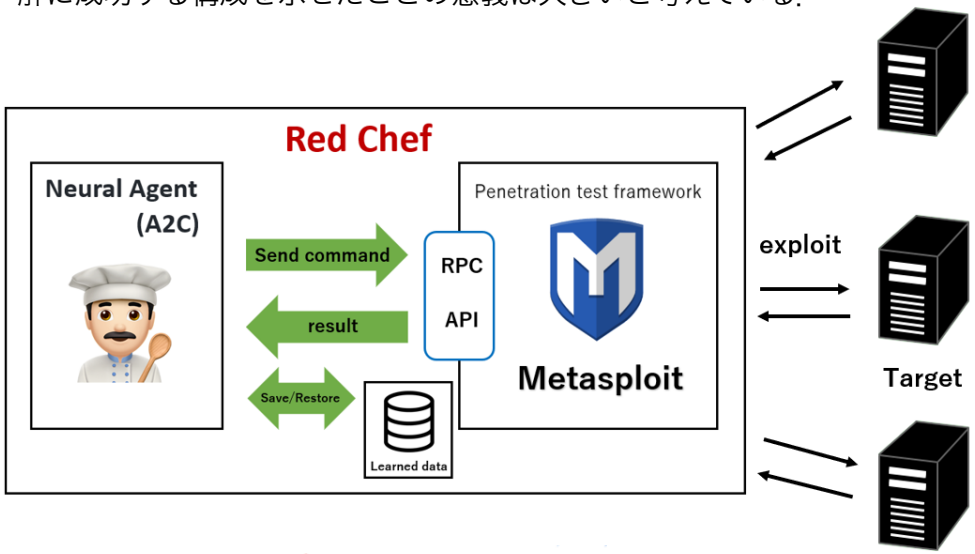
情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY



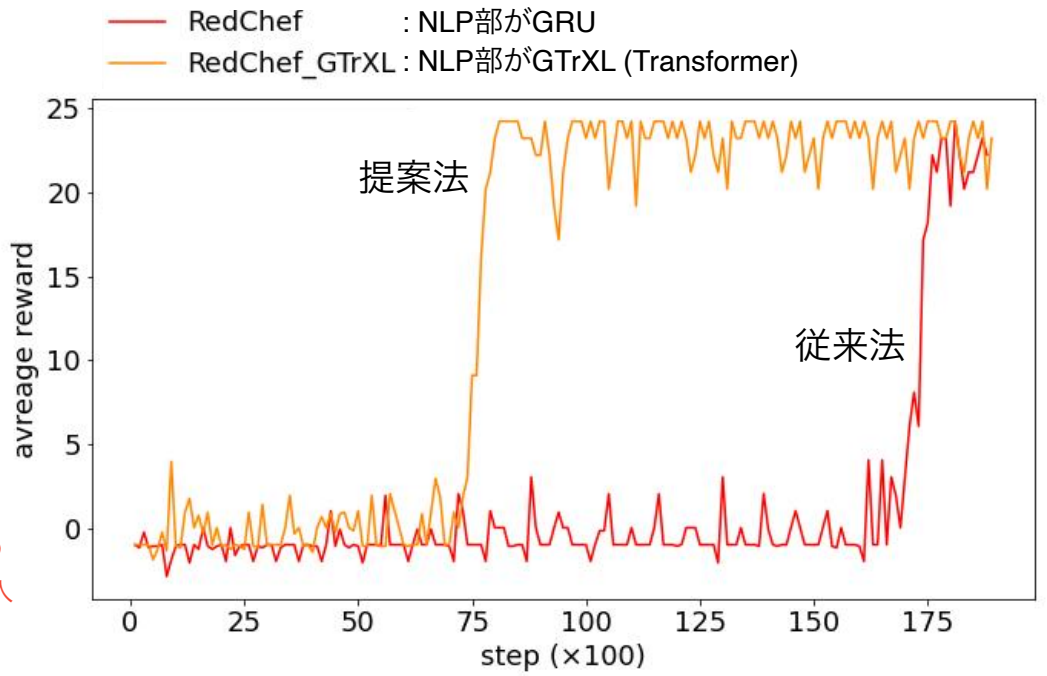
成果1：深層強化学習+Metasploit

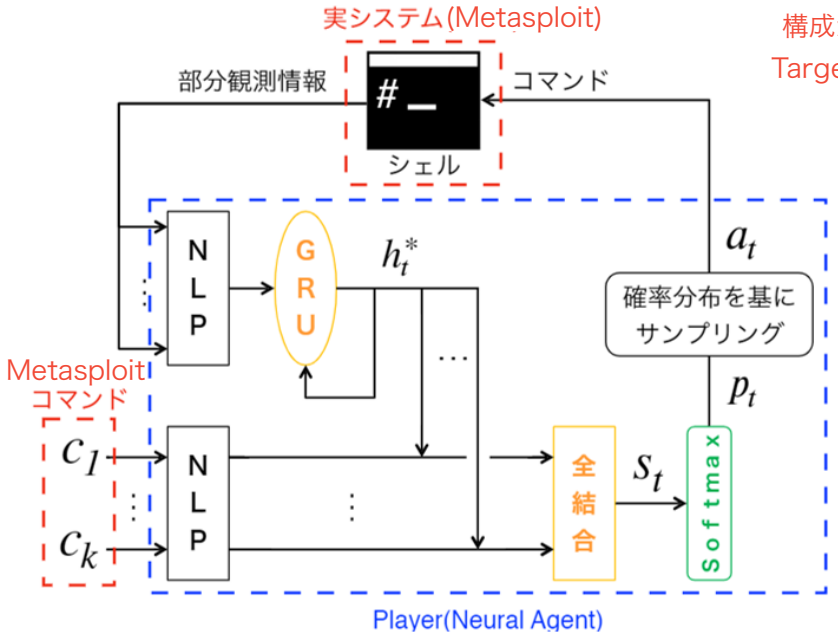
情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

図1のNLPにはTransformer(GTrXL)を利用し，状態を表す潜在ベクトルによりコマンドの確率分布を求め，分布に応じてコマンドをサンプリングする複数の異なった脆弱性を有するサーバーへの侵入を目標とするCTF問題の求解を実施した．その結果，GRUに基づく従来法をMetasploitに対応させたアルゴリズム(RedChef)に対して，**提案手法（RedChef_GTrXL）が効率的に最適戦略を獲得することに成功した**．実Unixサーバーを対象としたCTF問題(exploitation問題)の求解に成功する構成を示せたことの意義は大きいと考えている．



構成が異なる10種類の
Targetサーバーへの侵入





Metasploit・・・　最新のexploit、脆弱性スキャンツール、パスワードクラックツール等を含んだペネトレーションテスト用シェル

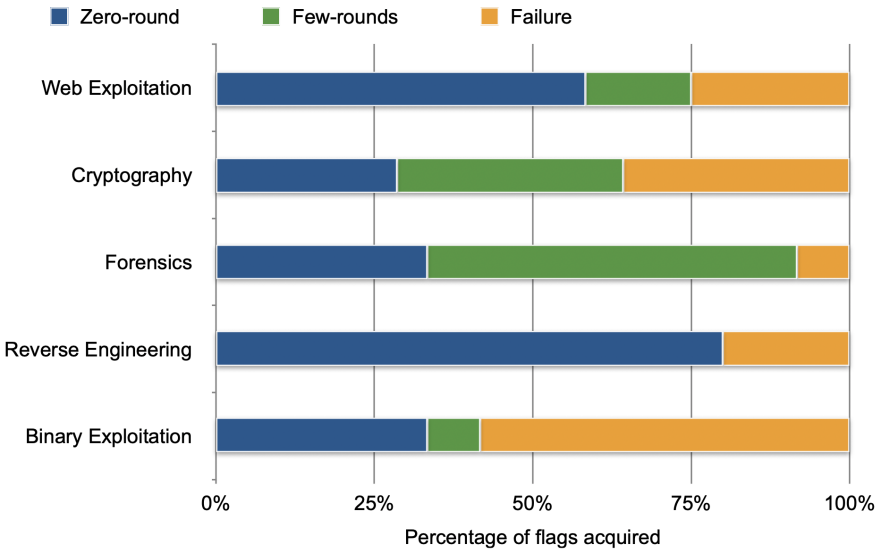
Neural Agentは以下4つのアクションを組み合わせてTargetサーバーへの侵入タスクを実施

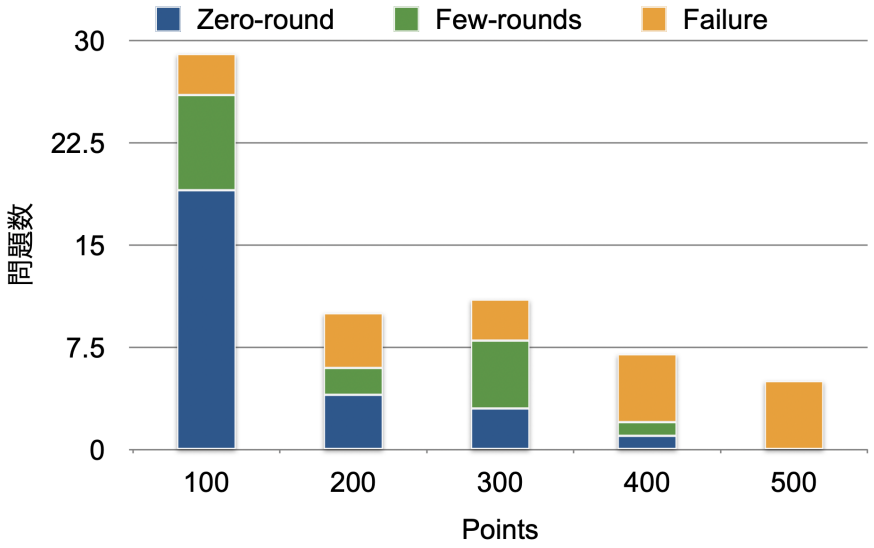
- ポート番号の選択
- ターゲットの選択（OS, Ver.の選択等）
- 攻撃モジュールの選択（特定の脆弱性を利用するためのコード, 5000超）
- 攻撃ペイロードの選択（エクスプロイト成功後に実行するコード, 数百種）

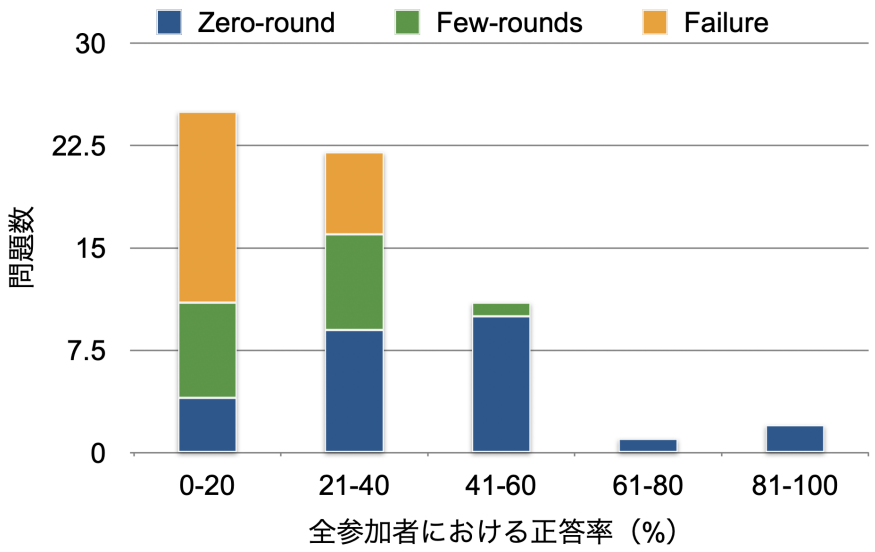
成果2: 大規模言語モデル→CTF求解

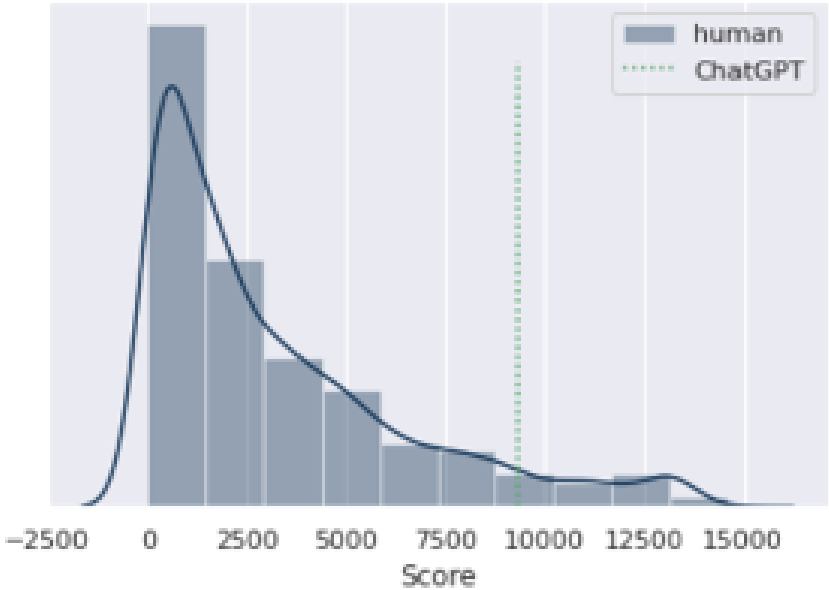
情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

大規模言語モデル(LLM)をサイバーセキュリティ分野に応用し, 人間とLLMの協調作業実験を実施した結果, PicoCTF2022の問題において図3, 図4に示すように**64問中48問のフラグを獲得するなど予想外に良い結果(575位/7794人, 上位7.3%)**が得られた. この結果は, サイバーセキュリティ分野においてもLLMを幅広い問題に適用できることを示唆している.







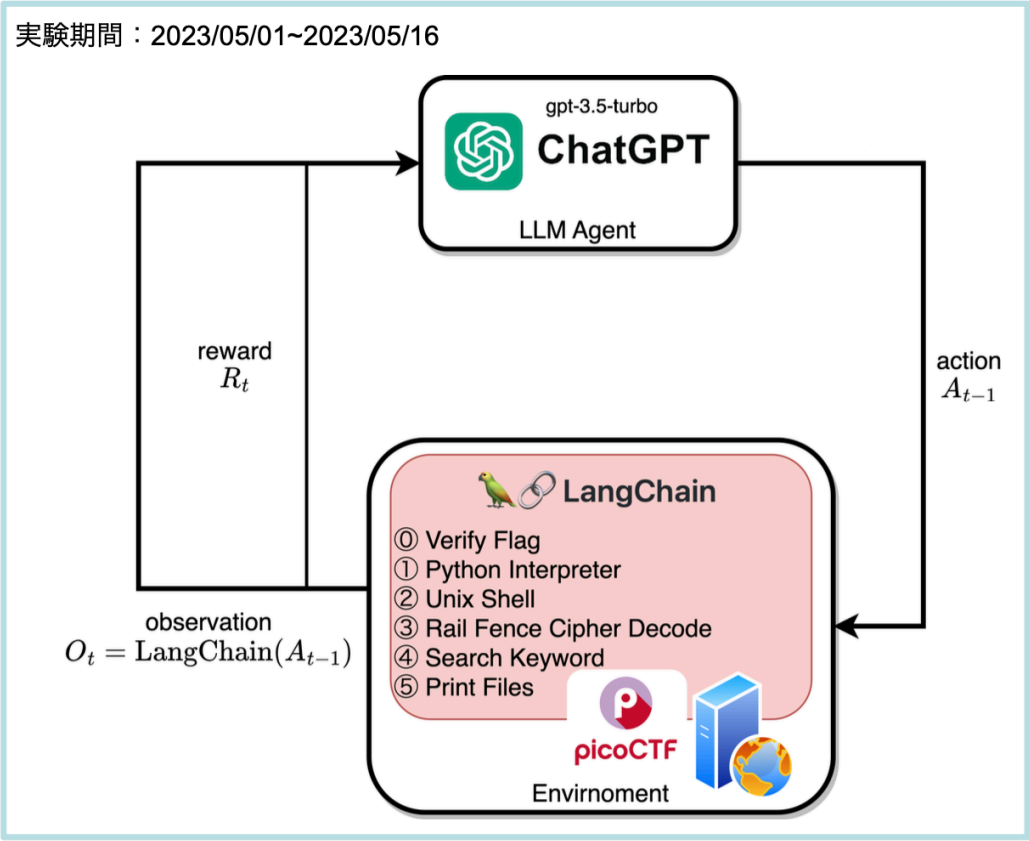


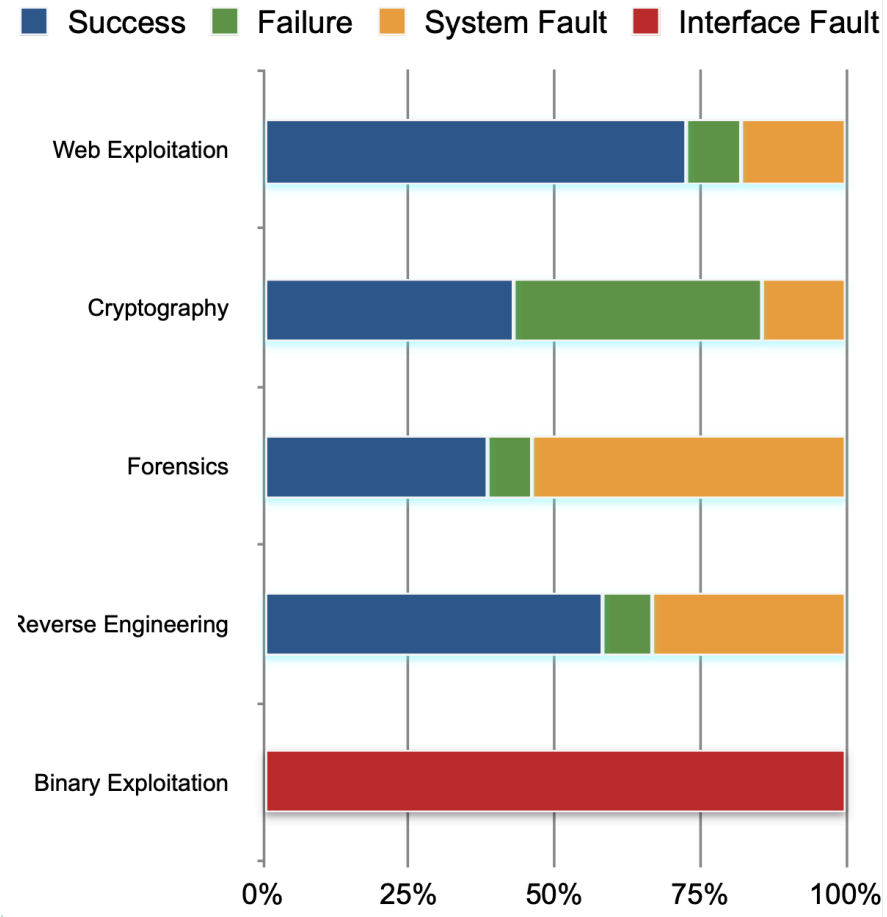
成果2+: LLM + Groundingによる自律CTF求解

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

研究期間終了後の成果：人工知能学会全国大会(2023年6月)で発表

- ReAct* (REasoning and ACTion) は，言語モデルが「Grounding」を通じて、推論と行動を統合する手法を与えている．このプロセスを通じて、モデルはReasoningのステップを生成し、Action計画を立て、外部ソースから情報を集めて結論を導く手法　　*Yao, S. *et al*. ReAct: Synergizing reasoning and acting in language models. in *International conference on learning representations (ICLR)* (2023).
- 実験では，ReActを実装したLangChainを用いて，LLM+Groundingにより自律的に求解させた．





- 26/64問に正解
- 全参加者7794人中2841位
- Binary Exploitationに苦戦（今後の課題）

CTF問題の例

CTFの例題①　暗号解読 substitution

https://play.picoctf.org/practice/challenge/307

問題文

メッセージが届いたが、すべてスクランブルされているようだ。幸運なことに、そのメッセージは先頭に鍵を持っているようだ。あなたはこの置換暗号を解読できるだろうか？メッセージのダウンロードはこちらから。

メッセージ
message.txt

ZGSOCXPQUYHMILERVTBWNAFJDK

Qctcnrel Mcptzlo ztebc, fuwq z ptzac zlo bwzwcmd zut, zlo gtenpqw ic wqc gccwmc
xtei z pmzbb szbc ul fqusq uw fzb clsmebco. Uw fzb z gcznwuxnm bsztzgzcnb, zlo, zw
wqzw wuic, nlhlefl we lzwntzmubwb—ex sentbc z ptczw rtukc ul z bsuclwuxus reulw
ex aucf. Wqctc fctc wfe tenlo gmzsh brewb lczt elc cjwtciuwd ex wqc gzsh, zlo z
melp elc lczt wqc ewqct. Wqc bszmcb fctc cjsccoulpmd qzto zlo pmebbd, fuwq zmm wqc
zrrcztzlsc ex gntlubqco pemo. Wqc fcupqw ex wqc ulbcsw fzb actd tcizthzgmc, zlo,
wzhulp zmm wqulpb ulwe selbuoctzwuel, U senmo qztomd gmzic Ynruwct xet qub eruluel
tcbrcswulp uw.

Wqc xmzp ub: ruseSWX{5NG5717N710L_3A0MN710L_357GX9XX}

※フラグはpicoCTF{文字列}の形式に決まっている

Flag

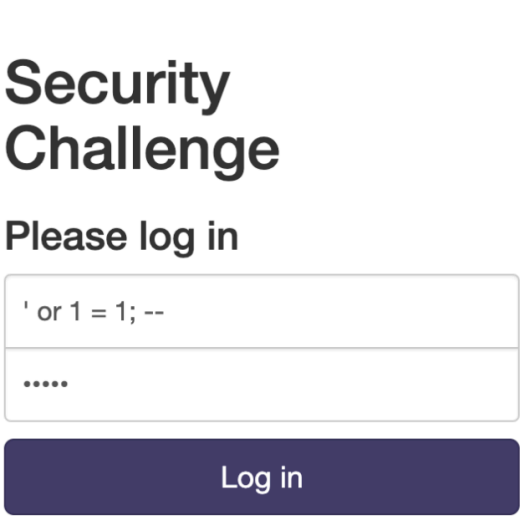
picoCTF{5UB5717U710N_3V0LU710N_357BF9FF}

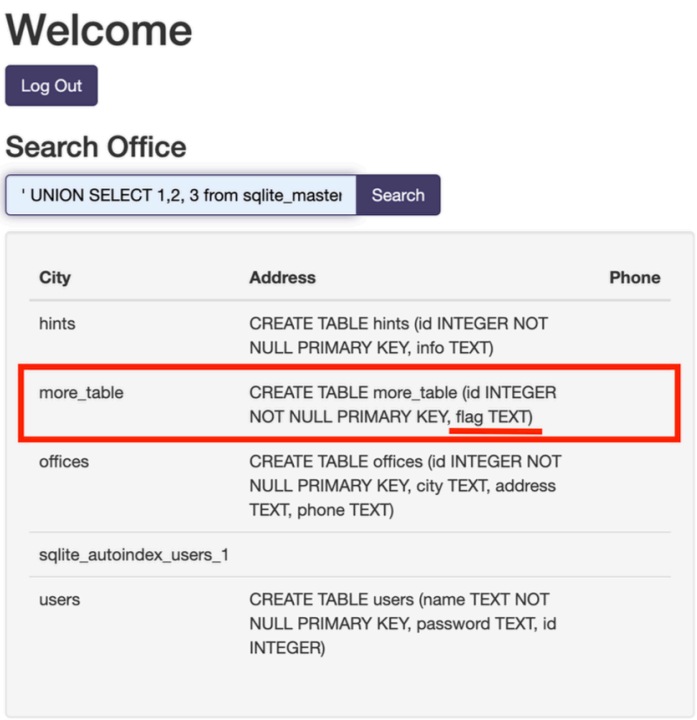
CTFの例題②　Webエクスプロイト　More SQLi

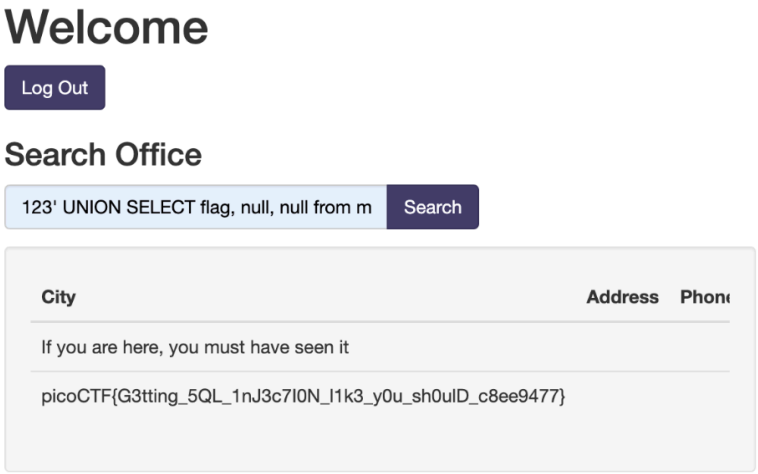
問題文

このサイトでフラグを見つけられますか？

ここでフラグを探してみてください。 https://play.picoctf.org/practice/challenge/358





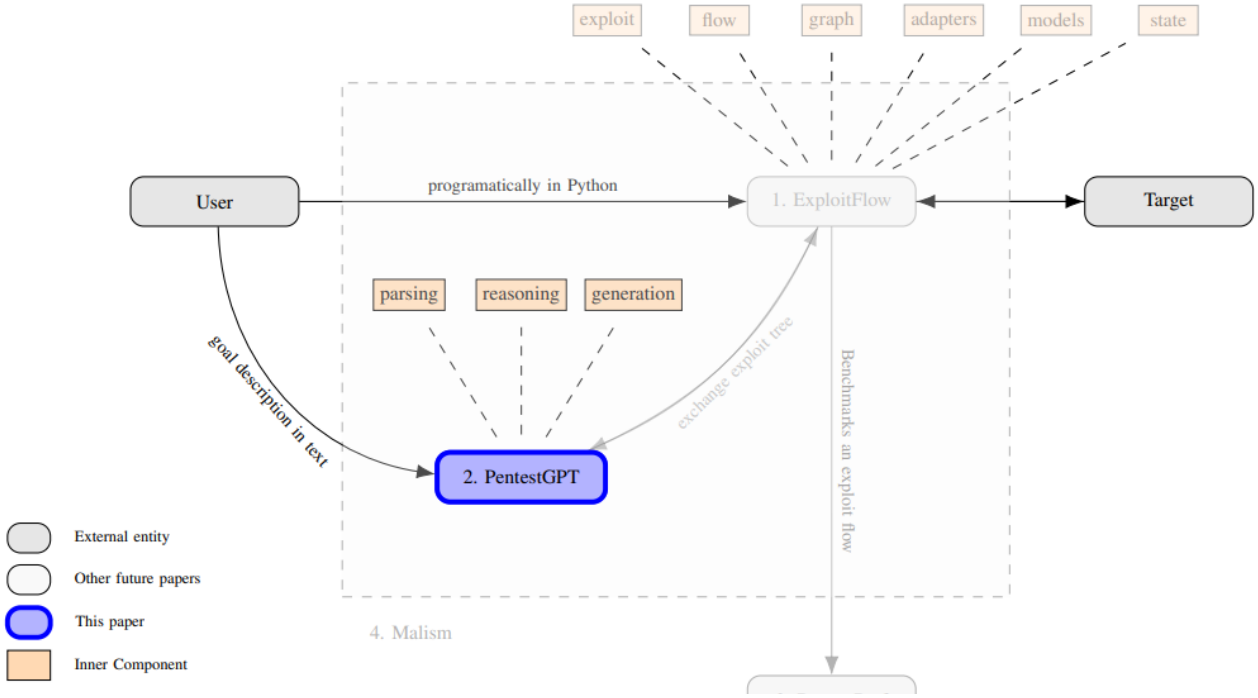


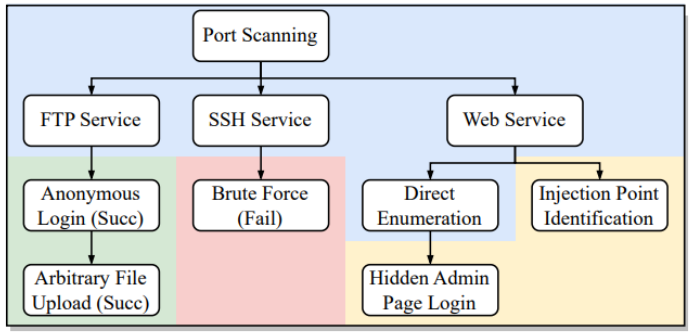
picoCTF{G3tting_5QL_1nJ3c7I0N_l1k3_y0u_sh0ulD_c8ee9477}

関連研究

PENTESTGPT: An LLM-empowered Automatic Penetration Testing Tool

Gelei Deng , Yi Liu , V´ıctor Mayoral-Vilches , Peng Liu , Yuekang Li , Yuan Tianwei Zhang , Yang Liu , Martin Pinzger , and Stefan Rass



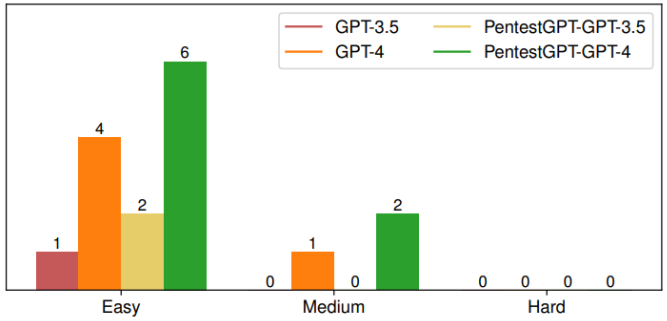


a) PTT Representatoin

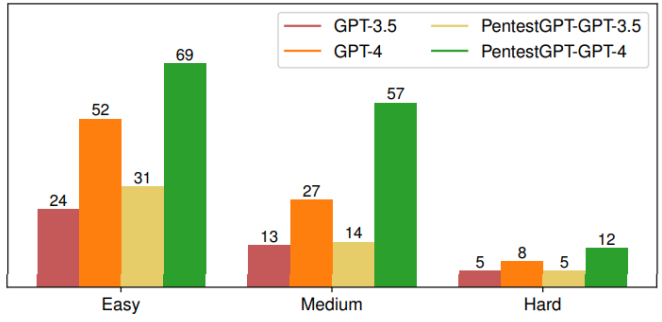
```
Task Tree:
1. Perform port scanning (completed)
  - Port 21, 22 and 80 are open.
  - Services are FTP, SSH, and Web Service.
2. Perform the testing
  2.1 Test FTP Service
    2.1.1 Test Anonymous Login (success)
      2.1.1.1 Test Anonymous Upload (success)
  2.2 Test SSH Service
    2.2.1 Brute-force (failed)
  2.3 Test Web Service (ongoing)
    2.3.1 Directory Enumeration
      2.3.1.1 Find hidden admin (to-do)
    2.3.2 Injection Identification (todo)
```

b) PTT Representation in Natural Language

Figure 4: Pentesting Task Tree in a) visualized tree format, and b) natural language format encoded in LLM.



(a) Overall completion status.



(b) Subtask completion status.

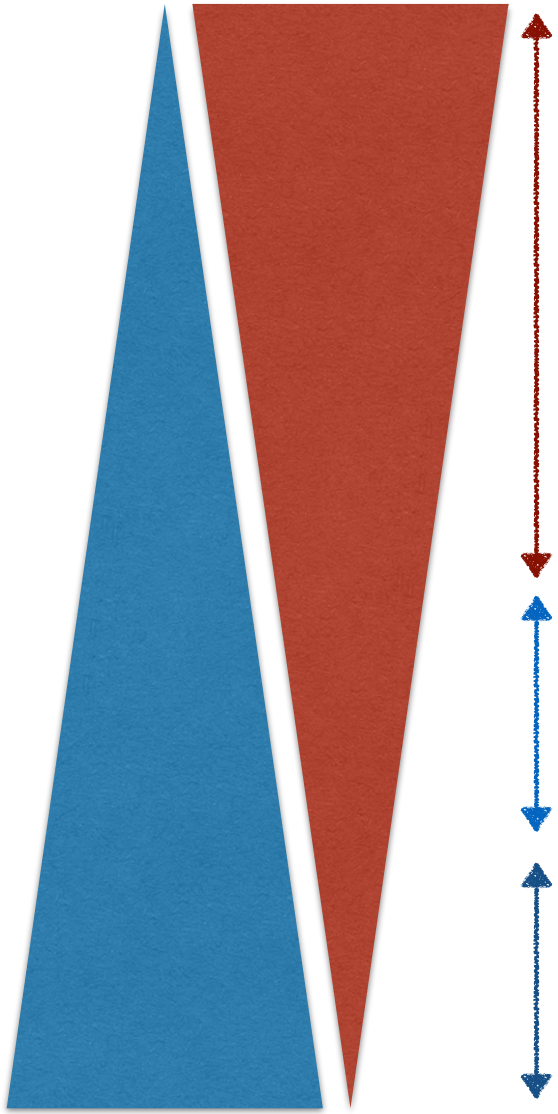
PicoCTF

picoCTFとは

- 主に中高生や初学者向けに作られたオンラインのCTF競技大会で,Carnegie Mellon Universityが主催している.
- picoCTF2022の問題は全部で64問あり,それぞれ100,200,300,400,500のポイントが与えられる.
- 最終的なスコアが高かった上位チームには賞金が与えられる場合がある.

問題カテゴリ

カテゴリ	説明	攻撃例
Web Exploitation	Webアプリケーションにおいて,脆弱性を突いた攻撃を行う.	SQLインジェクション,クロスサイトスクリプティング(XSS)
Cryptography	暗号理論を用いた問題が出題され,解読を試みる.	暗号文の解読,暗号鍵の復号化
Forensics	決められたデータから情報を探索する.	メモリダンプからの情報抽出,ファイルからの隠された情報の発見
Binary Exploitation	プログラムのバイナリを解析し,攻撃手法を見つける.	バッファオーバーフロー,スタックの上書き
Reverse Engineering	プログラムのバイナリを解析し,仕様を特定する.	プログラムの逆アセンブル,実行ファイルからの情報抽出



大規模システム　医療カルテシステム　公共システム

攻防のイタチごっこ

大規模言語モデルによる
サイバー攻撃対応
に期待できるか？

小規模システム
暗号プロトコル　TLS, IoT　スマートコントラクト

定理証明器/モデル検査
(フォーマルメソッド)

暗号プリミティブ　暗号，署名

証明可能安全性
(困難な問題に帰着)

ご清聴ありがとうございました

情報セキュリティ大学院大学
教授 大塚　玲
otsuka@ai.iisec.ac.jp