

公開コンテナイメージを利用したサイバーレンジのシナリオ開発の研究

代表研究者 中田 亮太郎 国立大学法人 一橋大学 情報基盤センター 講師

1 はじめに

サイバー攻撃の高度化・巧妙化に伴い、セキュリティ人材の不足が顕在化し、セキュリティ教育の重要度が高まっている中、高度なセキュリティ人材の育成や高等教育機関でのセキュリティ教育では、サイバーレンジを用いた演習によるセキュリティ教育が注目されている。サイバーレンジは、実際のセキュリティインシデントを仮想環境上で再現するリアリティの高い演習システムで、実践的で教育効果の高いサイバーセキュリティ演習が実施できる[1]。

しかし、サイバーレンジによる演習を実施するには、高度な知識や技術を持つ専門家がシナリオの開発や環境構築を行う必要があるほか、十分な訓練を受けた講師の育成や、教育効果の確認手法の確立などさまざまな課題があり、普及は限定的である。不足するセキュリティ人材への対応のためには、サイバーレンジによる効果的な演習環境をより広く活用できるようにする必要があるが、増え続ける脆弱性や進化し続ける攻撃手法に対応した多くのシナリオ開発が必要になるなど負担が大きく、導入ができている場合でもシナリオの不足などによる教育効果の低下を招いている[2]。

シナリオの不足を補うため、演習プラットフォームを公開・共有する取り組みや、シナリオ開発・環境構築の自動化に関する研究が行われている。しかし、既存の研究は既知のいくつかの脆弱性や攻撃手法を扱うための限定的な環境であり、常に最新のインシデントに対応したより多くの実践的シナリオの開発が必要とされている[3]。

そこで本研究では、より多くのサイバーレンジシナリオを効率的に開発するための手法として、機械学習等のAI技術によりサイバー攻撃の検知や対処を自動化することを目標とした「自律サイバー推論システム」の研究を活用することを検討した。既知の攻撃手法や脆弱性に対応した多くのシナリオ開発の自動化や、最新のインシデントへのいち早い対応により、サイバーレンジ最大の懸念であるシナリオの不足を解消し、教育効果を維持向上してセキュリティ人材の不足への対応や技術レベル向上が期待できる。

2 サイバーレンジ

2-1 環境構築と仮想化

サイバーレンジの運用には、演習で扱う攻撃や防御の流れを表すシナリオと、それらを再現するための環境構築が必要である。環境構築はサーバやネットワーク機器、利用者端末およびセキュリティ機器など、通常の情報システム・ネットワーク環境の構築と同等の状況を再現するだけでなく、演習目的となる攻撃や脆弱性を再現するために、特定の設定やソフトウェアバージョンの変更など、組織や企業の一般的なシステム・ネットワーク環境の構築とは一部異なる知識や技術が求められる。

また、演習の実施に際しては、多くの受講者用に多数の環境準備や迅速な入れ替えなどが必要であるため、実機を用いた演習は現実的ではなく、仮想化技術により演習環境が構築される。商用のサイバーレンジでは、管理機能や対応OSの多さなどから、VMWareなどハイパーバイザ型の仮想化製品を利用したサイバーレンジ製品が普及しているが、導入や維持管理に数億円規模の経済的負担が必要となるため、VirtualBoxやDockerなど安価で容易に利用できるホスト型およびコンテナ型の仮想化を利用したサイバーレンジの研究や、それらを用いた演習環境の公開が行われている[3]。

これまでの研究で、各仮想化方式の違いによる脆弱性の再現性能の違いはなく、同等の脆弱性の再現性能を有することを確認しており、その中でもコンテナ型仮想化は、消費リソースを大幅に抑制しつつ高速で軽量の演習環境が実行でき、近年のdockerの普及と利便性の向上も伴って、コンテナ型仮想化を用いたサイバーレンジの研究が増加しているため、本研究では主にコンテナ型仮想化によるサイバーレンジシナリオや演習環境構築を検討した。

2-2 演習シナリオ

サイバーレンジのシナリオは、特定の攻撃手法や脆弱性を体験することで理解するものや、インシデント全体での攻撃や対処法の流れを体験しながら総合的に知識や技術学ぶものなど、様々な目的やレベルのシナリオが存在する。ただし、開発には情報セキュリティ全般の知識や技術が必要である他、最新の攻撃手法や脆弱性への対応の困難さなどがあり、シナリオの不足や漏洩による教育効果の課題を抱えている[2]。

シナリオの不足への対応のため、これまでの研究ではシナリオのランダム化や Attack Graph を用いたシナリオ生成手法などの提案を行ってきた[1, 4]。

図 2.1 に、既存手法によるサイバーレンジのシナリオ自動生成のイメージを示す。

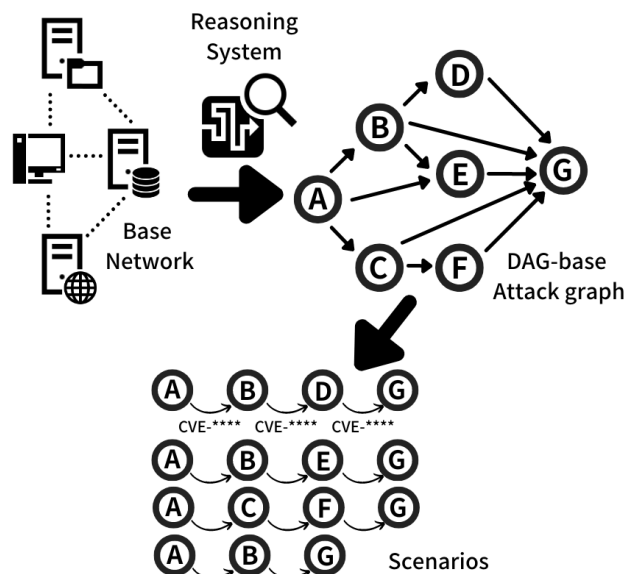


図 2.1 サイバーレンジのシナリオ生成イメージ

シナリオの元となるネットワーク環境に対し、サイバー攻撃の可能性や脆弱性を抽出するための推論システムによって存在する脆弱性を洗い出し、それらに有効な攻撃手法やネットワークの到達性を考慮することで DAG (有向非循環グラフ) 形式の Attack Graph を生成し、その経路を辿る形でシナリオ化する。

ただしこの手法は特定のネットワーク環境をターゲットとした限定的な手法であり、新たな攻撃手法や脆弱性への対応の他、既知の攻撃手法や脆弱性への対応をさまざまな環境を想定して汎用的に行うには十分ではない。また、推論システムには Nmap や OpenVAS など一般的にも使われるツールが用いられるが、対応する攻撃手法の抽出や検討はツール単体では困難で、それぞれの脆弱性に対して Metasploit の Exploit モジュールや、公開されている攻撃用コード等を組み合わせて確認するなど一部で手動での対応を要したり、新たな脆弱性へツール側が対応するのを待つ必要があるなど、汎用化するには課題がある。

したがって、シナリオ生成を効率よく様々なシステム・ネットワーク環境を対象として行うには、脆弱性やそれに対応する攻撃手法を効率よく推論できる高度なサイバー推論システムが求められる。

3 自律サイバー推論システムの利用

3-1 サイバー推論と人工知能技術

サイバー推論をはじめ、様々なセキュリティ技術によって高度化・巧妙化するサイバー攻撃への対応が検討されているが、従来の手法では対応しきれないサイバー攻撃が増加している。そこで近年注目されているのが人工知能 (AI) 技術を用いたセキュリティである。AI を用いたセキュリティ技術の概要について図 3.1 に示す。

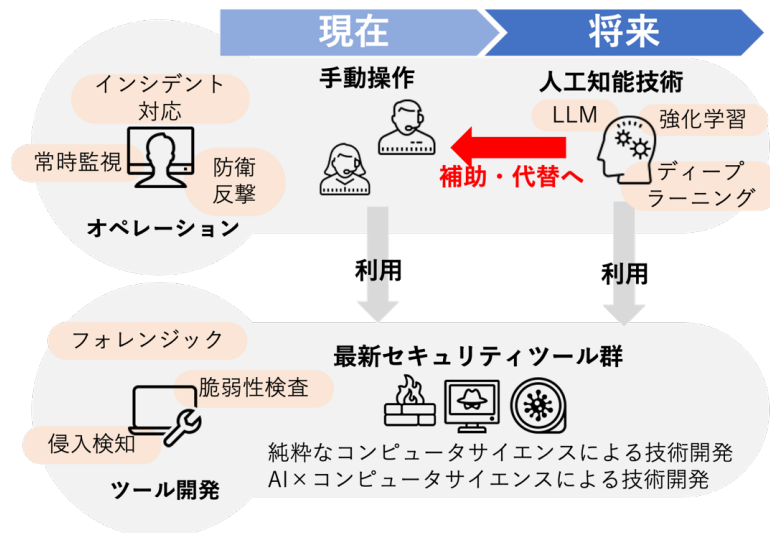


図 3.1 AI を用いたセキュリティ技術

現在のセキュリティに関するオペレーションは、さまざまな情報を元に有効な対応策を検討し、手動による操作によって実施される。各種ツールによる操作の自動化や監視作業の簡素化・効率化は既に行われているが、インシデント発生の際の対応策の検討や攻撃手法の分析のほか、未知の攻撃手法への対応など定型的ではない対応が求められ、それらの負担を補助・代替できるような AI 技術の研究が進んでおり、自律サイバー推論システムの研究ではこれまで人間が手動で行なっていた検知や各種対応などのオペレーションを全自動で実施できることを目標としている。

なお、これまでの AI 技術を用いたセキュリティ対応手法は、過去に発生した正当な通信又は悪意のある通信を学習させ、実際の通信と比較し分類を行うものが主流であった。しかし、過去の通信から学習させた場合、珍しい正当な通信を悪意のある通信と誤分類したり、未知の攻撃への対応が難しいといった問題が発生し得る。また、画像分類の分野で特に研究が盛んな、AI を騙すような敵対的攻撃をサイバー空間に適用する研究も行われるなど、過去の通信からの学習のみでは対応できない攻撃が懸念される。

そこで、従来の機械学習と異なり、訓練データを与えずとも与えられた環境で試行錯誤を行い、報酬を最大化することで最適な行動を導き出す強化学習と呼ばれる手法による自律サイバー推論システムの研究が行われている [5]。強化学習の概要を図 3.2 に示す。



図 3.2 強化学習の概要

強化学習は、エージェント(行動主体)がある環境の状態に応じて、どのように行動すれば報酬が多くもらえるかを求める手法である。教師なし学習や教師あり学習とは違い、学習データなしに自身の試行錯誤のみで学習するのが特徴である。強化学習の学習サイクルは下記の通りである。

1. エージェントは最初に何を判断すべきか分からないので、初期値としてランダムな方策をもとに行動を選択する。
2. 行動に伴う報酬に応じて、エージェントが行動を決定する方策を求める
3. ランダムな動きは残しつつ、方策を手掛かりに価値の高い行動を決定する。
4. 2-3 を繰り返し、将来的に多くの報酬を得られる最適な方策を求める。

自律サイバー推論システムは、セキュリティインシデントにおける行動の決定に強化学習を用いたもので、既存の攻撃行動や脆弱性に対する対応だけでなく、未知の攻撃に対しても迅速かつ的確に対応することを目指して研究されたシステムである。ソフトウェアの脆弱性検知、エクスプロイト作成、パッチ作成等の各種行動を自動的に行うもので、コンピュータ同士が全自動で戦うハッキングコンテストである「サイバーグランドチャレンジ (CGC)」で用いられるなどして脚光を浴びた[6]。サイバー推論システムに関連する研究は、ファジング、シンボリック実行、異常検知、バイナリパッチ、行動戦略等多岐にわたり、コンテストでも用いられたように完全に自律したサイバー推論システムによって、サイバーインシデントに関する行動を自動化することを目指した研究が行われている[5, 7]。

3-2 強化学習によるペネトレーションテスト手法

自律サイバー推論システムの研究の一つに、強化学習を用いたペネトレーションテスト手法がある[7]。ペネトレーションテストはサイバー攻撃に対抗する手法の1つであり、実際に対象環境に対して侵入テストを行うことによりセキュリティリスクを顕在化させ、インシデントに備える手法である。従来のペネトレーションテストは、脆弱性検査ツール等のサイバー推論システムを手動で利用した結果から攻撃方法を検討するなど多くの手間や時間がかけられてきたが、自律サイバー推論システム化の研究が進んでおり、Autopentest-DRL や DeepExploit などの手法が提案されている[8, 9]。

DeepExploit では、攻撃に使われるコードに対し、本来手動でのチューニング作業が必要な OS 情報や CMS、利用フレームワークなどの様々な環境情報を強化学習により学習させることで、自動的に有効な攻撃コードを実行する推論システムとして利用できる。実際のペネトレーションテストで使われるツールである Metasploit や Nmap などを用いて得られる情報を基に攻撃手法を学習していくが、学習段階においては一部しか最適化されておらず、テスト段階ではポートスキャンから獲得されたポートに総当たりで攻撃していくという手法をとっているため、多くのターゲットに対して推論を行うには非効率であるほか、使われているツールのバージョンが古く、最新の環境や脆弱性・攻撃手法への対応が困難であるなどの課題がある。

そこで、これまでの研究において、各種ツールによって攻撃を細分化した DeepExploit の手法と、部分観測マルコフ決定過程 (POMDP) ベースの深層強化学習手法であるニューラルエージェントを組み合わせた手法により、攻撃手法の推論に対する最適化が可能な RedChef を開発した[10]。図 3.3 に RedChef の概要を示す。

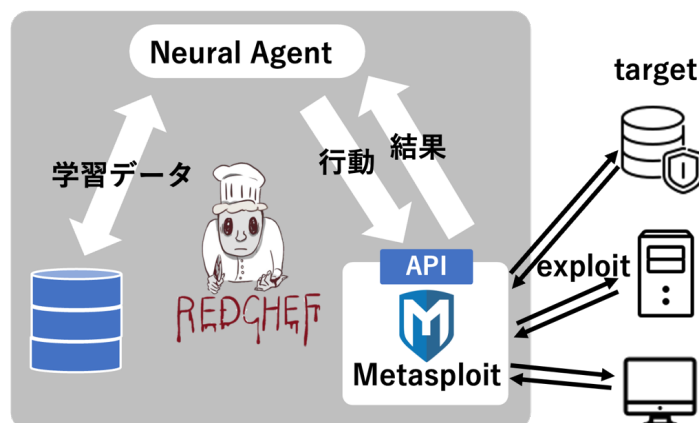


図 3.3 RedChef の概要

RedChef は DeepExploit の手法を参考に、metasploit の RPC API[11]を通じてターゲットへ試行する状態推定や行動選択の精度向上を目指したものである。環境に Metasploit の実システムを用い、実行コマンドに Metasploit で使用するコマンドを用いる。DeepExploit の場合、Metasploit の payload の選択のみを行なっているが、RedChef では RPORT/exploit module/exploit target/payload の 4 種類の行動を最適化する。

図 5 に DeepExploit との比較実験の結果を示す。

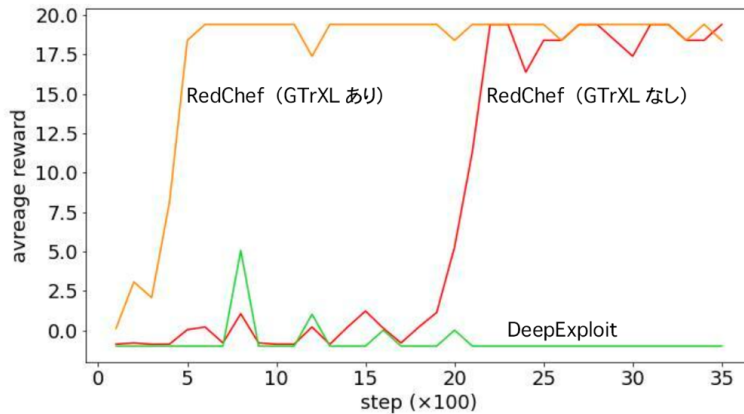


図 3.4 RedChef と DeepExploit の比較実験結果

RedChef は、ニューラルエージェントの状態推定の精度向上のため、GTrXL (Gated Transformer-XL) と呼ばれる手法を取り入れており、学習段階における実験では約 500 ステップ目で最適な平均報酬となるなど高い精度の学習が可能であることを示した。RedChef は効率的にターゲットの状態を推定し、攻撃可能な Metasploit の exploit モジュールを推定することが可能であり、サイバーレンジシナリオの生成においても有効なサイバー推論システムとしての利用が期待できる。

3-3 シナリオ生成への応用

従来のサイバーレンジシナリオの開発では、特定の攻撃手法や脆弱性を学習目的とし、それらを再現できる環境の構築や有効な攻撃コードの検討など、実際のインシデントを想定した開発が行われてきた。ペネトレーションテストにおいても、実際の環境に対して攻撃可能性を検討し、様々な攻撃シナリオを検討してテストが実施されるため、サイバーレンジのシナリオ開発と同等の内容が検討されていると言える。したがって、図 2.1 に示したサイバーレンジシナリオ生成の流れにおける推論システムにおいて、強化学習によるペネトレーションテスト手法を利用することで、自動的かつ効率的にシナリオ開発を行うことが期待できる。ただし、そのためには学習や推論のベースとなる多くのターゲットを用意しなければならない。

そこで、Docker のコンテナイメージをインターネット常に公開しているリポジトリである Docker HUB に注目した。Docker HUB は、公式の OS イメージだけでなく、独自に作られた Web サーバ、DB サーバなどあらゆるコンテナイメージが公開されており、利用者はイメージをダウンロードしてコンテナとして起動し、即座に様々なシステム環境を再現して利用が可能である。tag と呼ばれるバージョン管理機能もあり、古いバージョンの OS や各種ソフトウェアを用いたイメージが公開されていることが多く、脆弱性が存在するバージョンや設定の不備が含まれたコンテナイメージも多く存在する。

RedChef を用いたサイバー推論により、Docker HUB に公開されているコンテナイメージから攻撃可能な脆弱性を持つターゲットやそれに対応する攻撃手法が取得できれば、効率的なサイバーレンジのシナリオ生成に利用できるだけでなく、環境構築にも同じコンテナイメージをそのまま利用できる。また、Docker HUB には最新の脆弱性の検証用環境 (PoC) として再現されたイメージが提供されていることもあり、これらのイメージに対して自律サイバー推論システムが機能すれば、最新の脆弱性にいち早く対応したシナリオの生成も期待できる。

4 システムの実装と検証

4-1 システムの実装

図 4.1 に、redchef を自律サイバー推論システムとしてシナリオ生成に利用し、Docker HUB のコンテナイメージをターゲットとしたシナリオ生成システムの概要を示す。

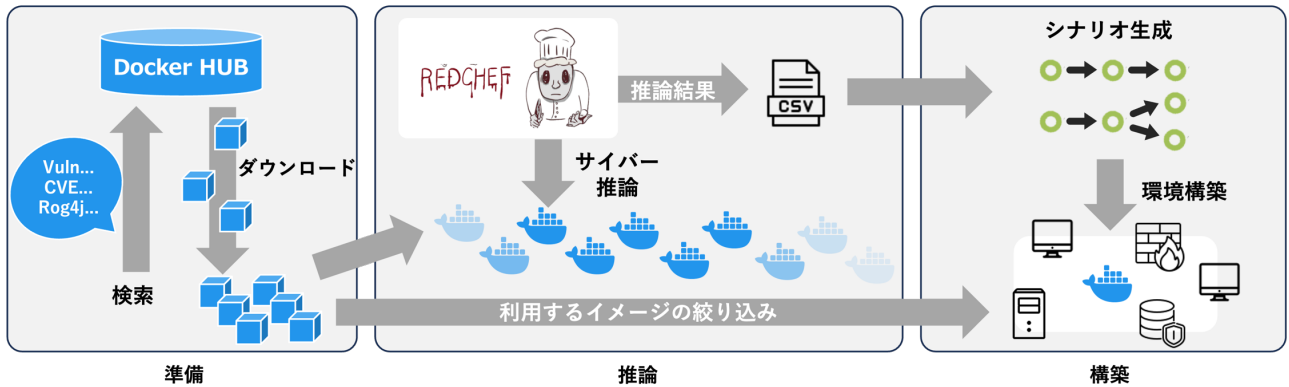


図 4.1 Redchef と DockerHUB によるシナリオ生成システムの概要

システムは準備・推論・構築の3つのフェーズに分かれており、まず準備フェーズでは Docker HUB 上のコンテナイメージをダウンロードするための検索ワードとダウンロードするイメージ数を指定する。たとえば具体的な CVE 番号や、“wordpress”など特定のアプリケーション環境などを指定することで、関連するイメージや PoC（検証環境）として作られたイメージのほか、特定のアプリ環境が構築されたものなどがさまざまなイメージが検索されダウンロードすることができるほか、tag により管理されたバージョン違いのイメージも利用できる。

推論フェーズでは、ダウンロードしたコンテナを順次起動させて RedChef によるサイバー推論を実施し、各コンテナに存在する脆弱性の発見だけでなく、有効な Metasploit の攻撃モジュールもセットで推論することができるため、シナリオ生成に必要な情報として活用できる。

最後に構築フェーズでは、推論によって得られた結果から DAG 形式の Attack Graph を生成する。推論から得られる内容は、主にどのイメージに対してどの攻撃手法（Metasploit の exploit モジュール）が有効かという内容であり、攻撃の流れを表すシナリオとしてそのまま利用することが可能である。グラフの経路が演習用のシナリオとして使用され、演習環境の構築についても推論の対象となったコンテナそのまま利用することで、自律サイバー推論システムを利用したシナリオ生成から環境構築までを自動化する。

4-2 システムの検証

実装したシステムの動作を検証するため、いくつかの検索ワードを用いた動作実験を行った。実験の結果を表 1 に示す。

図 4.1 検索ワード毎の推論結果とシナリオ生成への利用

検索ワード	検索イメージ数			シナリオ生成
	10	20	30	
ubuntu(official)	－	－	－	×
wordpress(official)	1	2	2	×
mysql(official)	－	1	1	×
metasploitable2	5	12	19	◎
metasploitable3	2	4	7	○
cve-2016	3	4	4	○
cve-2023	1	1	1	△

それぞれの検索ワードに対して 10, 20, 30 の 3 パターンでコンテナイメージの検索数を設定し、検索して出た順に利用した。各コンテナに対して学習を行った後に有効な攻撃行動をテストするようにした結果、確認できた脆弱性と攻撃モジュールの組み合わせの数を示している。また、それらの結果を用いてシナリオ生成を行い、実際に演習として利用できる内容となったかを確認した。なお、学習はそれぞれの検索ワードごとに約 500 ステップずつ行われるように実施しているが、用意した機械学習用環境ではそれぞれ約 10 時間以上かかっている。ただし、通常の CPU のみを利用した環境では学習や推論がほぼ実施できず、GPU 等による高速計算環境が必須であった。

最も効率的に推論が成功したのは、脆弱なサーバとして検証や演習でもよく使われる“metasploitable2”を検索ワードとした場合で、Docker HUB に存在するイメージの数も一定数あるだけでなく、一つのイメージに存在する脆弱性の数が非常に多く、効果的に学習および推論を進めることができ、多くの攻撃パターンを確認することができた。さらにこの結果を用いてシナリオ生成から環境構築までを実行し、さまざまな攻撃や脆弱性を再現した演習環境として利用できることが確認できた。同じく脆弱なサーバとして知られる“metasploitable3”を検索ワードとした場合は、Docker HUB に存在するイメージ数が 4 つと少なく、学習できるデータや推論できる脆弱性が少なかったためかあまり多くの結果は得られなかったものの、シナリオとしての利用は可能であった。

また、“ubuntu”を検索した場合、多くの tag が存在するため検索対象となるイメージは非常に多いが、最新のイメージやそれに近いものからは脆弱性がほぼ発見できず、学習もうまくいかなかった。同じく公式イメージとして公開されている wordpress や mysql など、検索対象となる tag が非常に多く存在するが、最新のイメージでは脆弱性対策が十分にされており、あまり効率的な推定ができなかったため、これらの公式イメージを検索する場合は古い tag 情報から遡って利用するなどの工夫が必要である。

PoC 環境などをターゲットとするために cve 番号を年まで指定した場合、2023 で検索したイメージは現状では数が少なく対象となった脆弱性も少なかったため、推論できた数が少なくなってしまうが、2016 の場合は一定数のイメージが存在し、いずれの場合もシナリオとして利用可能な結果を得ることができた。

4-3 演習でのシナリオ検証

構築したシステムの有効性を検証するため、実際にサイバーセキュリティ演習を実施している場での演習の実施や効果の検証を進めていたが、シナリオ生成の要となっていた機械学習用の環境で動作する機器が故障し、検証に十分なシナリオを準備することができなかった。そこで、シナリオ生成が可能になった際の検証の方法を検討し、修理が完了しシナリオの準備ができ次第、演習での検証ができるように準備を進めることとした。次章以降では、検討した演習シナリオの評価方法について述べる。

5 演習シナリオの評価

5-1 演習評価の課題

演習シナリオ生成や環境構築に関する研究は盛んに行われており、シナリオ不足や演習実施に係る人的・経済的コストを抑制し、情報セキュリティ教育の裾野を広げることが期待されている。しかし、演習の評価・採点に関する研究が進んでおらず、演習が実施できてもその効果を定量的に評価することができていない。例えば演習の採点を自動化することができれば、受講者の学習効果の可視化や演習シナリオの効果測定などが可能となり、より効果的な情報セキュリティ教育を推進することができる。

CTF (Capture The Flag) の場合、与えられた環境の中で情報セキュリティに関する知識・技術を駆使してフラグを取得するという、特定の目的の達成のみを採点の対象にすることができるが、一定のシナリオの下で様々な形の再現環境で行われる情報セキュリティ演習の場合は、実際の攻撃手法や脆弱性を再現した環境で目的を正しく理解した行動を行なっているかを判断することが必要であるため、受講者が操作した内容から行動の意図を読み取る必要があり、個別に行うことは作業量の多さなどから現実的ではなく、かつ判断基準が統一されないことによる定量性の確保が困難である。

そこで我々は、人工知能技術を利用して受講者の操作記録から本質的な行動目的を要約することで採点に利用する手法を検討した。自律サイバー推論システムによる最適な行動の導出のほか、急速に発展している大規模言語モデル (LLM) による文章の要約能力が、演習における受講者の行動の要約や本質的な行動目的

の理解に活用できると考えた。演習の採点を自動化することで受講者の成績やシナリオの有効性を可視化し、適切な評価を可能とする。

6 演習の採点手法

6-1 基本方針

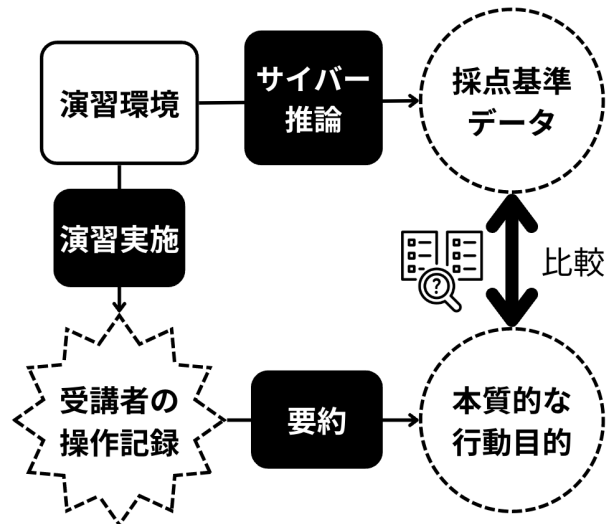


図 6.1 演習の採点手法の概要

図 6.1 に、今回検討した採点手法の概要を示す。高等教育機関などで想定される演習の多くは、受講者ごとの環境で特定の攻撃手法や脆弱性を実際に操作して再現することで、攻撃手法や脆弱性の仕組みを理解したり、その攻撃に対する調査や対応に関する操作を行うことで防御に関する手法を習得するなど、想定したシナリオに沿った適切な技術や知識に基づいた行動を演習端末上で実施して体験できたかが重要である。

したがって、採点は演習で想定されたシナリオと受講者の行動がどれだけ一致するかが重要となる。シナリオで想定されている攻撃や脆弱性の再現・防御に関する戦略の集合となる採点基準データに対し、受講者が実際に行った操作が何を目的で行われたものかを確認した内容がどれだけ一致するかが基本的な採点の基準となる。ただし、サイバーセキュリティに限らず様々な環境で必要となる行動や具体的なコマンド操作は、目的が同じでも必ずしも一致するものではない。また、攻撃に使われるようなコマンドや防御に関する様々な行動も、その内容や順序が完全一致することは稀であるため、採点や評価の際にはそれらの操作の幅や多様性を考慮し、受講者がどのような目的を持って操作を行なったかを適切に把握してシナリオと一致していることを確認する必要があるため、何らかの方法で受講者の行動や操作を要約し、本質的な行動目的を導出する必要がある。

6-2 採点基準データ

受講者の行動を正しく評価するため、採点の基準となるデータを用意する必要がある。これまでの自律サイバー推論システムの研究では、与えられた環境に対して一定の行動空間の中から最適な攻撃行動を取得することは可能であるが、シナリオで想定されるコマンド操作は、受講者の行動と同様に目的が同じ別のコマンドが存在したり、オプションの違いなど絶対的な正解があるわけではなく、受講者の行動と比較して採点を行うための基準とするのに適さない。そこで、攻撃の内容や防御の方針を MITRE ATT&CK Matrix の内容から確認する研究を参考にして、採点の基準となる攻撃や防御のデータセットを作成した[12]。

MITRE ATT&CK Matrix は、さまざまな情報セキュリティインシデントから集められた知見をまとめたもので、攻撃の種別や対応する防御の方針などを体系的にまとめており、演習環境対応する攻撃や防御に関するデータセットを作成することで、採点の基準として利用することができる[13]。

6-3 受講者の本質的行動

受講者の行動を確認するため、演習環境で取得できるログを用いることを検討した。通常のコマンド操作ログは、入力したコマンドの内容のみが記録されているため、それだけでは操作の結果どうなったかが不明となってしまう、実際の操作の目的が何で、それが正しく遂行されたかを判断することが困難であるため、UNIX系のscriptコマンド、Windowsでのstart/stop-transcriptなどを用いることで、コンソールの入出力を記録できるようにする。その内容から各受講者の行動を要約することで、軽微なコマンドの入力間違いやオプションの違いなどを考慮せず、操作の目的とその結果となるデータを生成する。また、それらのログの内容を要約し、受講者の本質的な行動目的を把握するためにLLMを利用する。統一性のある要約結果を得るために、基準となるデータやプロンプトの入力の工夫の他、パラメータ調整が必要となるが、本質的な操作目的ではない内容を精査し、シナリオの遂行に必要な操作内容のみに要約した結果が得られるため、これを採点に使うことで受講者の本質的な行動を評価することができる。

6-4 採点手法概要

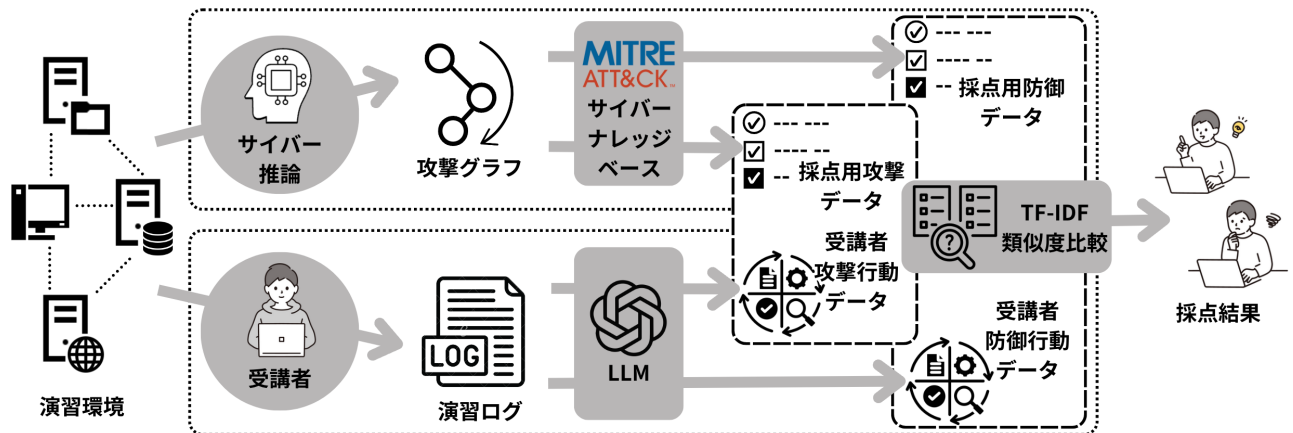


図 6.2 検討した演習採点手法の概要

図 6.2 に、検討した情報セキュリティ演習の採点手法概要を示す。演習環境からのサイバー推論により、MITRE ATT&CK Matrix の内容に該当する攻撃手法に関するデータセットと、対応する防御戦略に関するデータセットを作成し、これらを採点用の基準データとする。併せて、演習の実施後に受講者の操作によって生成されたログデータを取得し、その行動を LLM によって要約することで攻撃や防御の行動データを生成する。これらのデータを、攻撃と防御のそれぞれで TF-IDF によるコサイン類似度測定によって類似度を比較する。TF-IDF は文章に現れる単語の頻度 (TF) と希少性 (IDF) を掛け合わせたもので、これにより演習環境から推論された攻撃の概要や防御の戦術と、受講者の取った行動がどれだけ類似するかを定量的に確認することができ、採点の指標として利用できる。

7 採点の実験と検証

7-1 演習内容

検討した手法が実際に採点に利用できるかを確認するため、演習で取得した受講者の操作ログデータを利用した実験を行った。実験に用いた演習の概要を図 7.1 に示す。

受講者は攻撃者側端末や攻撃対象となる仮想企業ネットワーク上の端末を操作し、攻撃や防御の操作を行う。受講者のターミナル操作は script コマンドにより実行結果まで含めて記録され、web サーバや DB サーバのログも取得可能であり、これらのログから受講者の行動を把握する。

受講者は攻撃シナリオに沿った操作を実行して攻撃手法や脆弱性の仕組みを確認し、その後攻撃を受けた仮想企業側の端末防御操作を行うことで調査や対応などの手法を学ぶ。防御に関して特別なシナリオは想定しておらず、攻撃内容に応じたログの確認や攻撃に使われたファイルの確認、排除など状況に応じた適切な防御行動が求められる内容となっている。

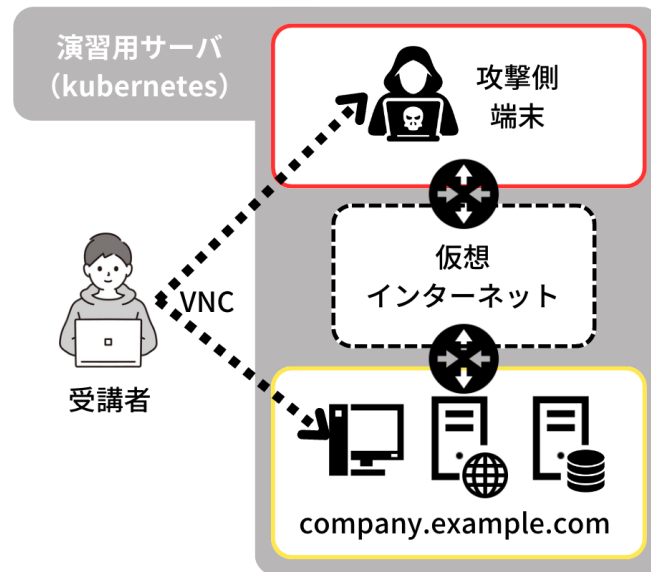


図 7.1 採点実験用演習環境概要

7-2 採点用攻撃・防御データの生成

演習環境から関連する MITRE ATT&CK Enterprise Matrix の内容を導出したものを表 7.1 に示す。

表 7.1 採点用攻撃・防御データセットの一部

TECHNIQUES		MITIGATIONS		
ID	Name	ID	times	Name
T1046	Network Service Discovery	M1026	5	Privileged Account Management
T1550	Use Alternate Authentication Material	M1027	4	Password Policies
T1059	Command and Scripting Interpreter	M1017	4	User Training
T1078	Valid Accounts	M1015	4	Active Directory Configuration
T1552	Unsecured Credentials	M1018	3	User Account Management
T1003	OS Credential Dumping	M1041	2	Encrypt Sensitive Information
		M1036	2	Account Use Policies
		M1028	2	Operating System Configuration
		M1013	2	Application Developer Guidance

防御データセットは、攻撃データセットに対応する防御戦術として MITRE ATT&CK Enterprise Matrix に記載されたものを抜き出している。表は ID と名称、および防御戦略では重複して登場する回数などを表示しているが、実際は攻撃の詳細や防御戦術の内容などナレッジベースが文書化されたデータの集合となっており、これらの内容が採点時の基準として利用できる。

7-3 受講者の行動データ

受講者の行動は、演習で取得した操作ログを元に導出する。受講者の操作ログは入力したコマンドとその実行結果となっており、今回の演習環境では攻撃フェーズに使う操作端末と防御フェーズに使う操作端末を別に用意しているため、攻撃、防御それぞれのログデータを LLM によって要約して行動データとして利用した。以下に、演習受講者の操作ログの例を示す。

```
Check the following cyber attacker operation logs to see what was
done
Ignore any operational errors.
Script started on 2023-12-06 12:29:57+0900
# pwd
```

```
/root
# ;s
sh: 12: Syntax error: ";" unexpected
# ls
Desktop Music Templates noVNC start.sh
Documents Pictures Videos preseed.txt typescript
Downloads Public dbdump.txt routing.sh typescript.bak
# nmap-F company.example.co
sh: 4: nmap-F: not found
# nmap -F company.example.com
Starting Nmap 7.60 ( https://nmap.org ) at 2023-12-06 12:31 JST
Nmap scan report for company.example.com (10.76.22.98)
Host is up (0.000041s latency).
Not shown: 97 closed ports
PORT STATE SERVICE
22/tcp open  ssh
80/tcp open  http
8080/tcp open http-proxy
Nmap done: 1 IP address (1 host up) scanned in 1.56 seconds
# ssh
usage: ssh [-46AaCfGgKkMNnqsTtVvXxYy] [-b bind_address]
[-c cipher_spec] [-D [bind_address:]port] [-E log_file]
...
# ssh backdor@company.example.com
backdor@company.example.com's password:
Permission denied, please try again.
backdor@company.example.com: Permission denied (publickey,password).
# ssh backdoor@company.example.com
backdoor@company.example.com's password:
Permission denied, please try again.
backdoor@company.example.com's password:
Welcome to Ubuntu 18.04.1 LTS (GNU/Linux 4.15.0-43-generic x86_64)
* Documentation: https://help.ubuntu.com
* Management: https://landscape.canonical.com
* Support: https://ubuntu.com/advantage
This system has been minimized by removing packages and content
that are not required on a system that users do not log into.
To restore this content, you can run the 'unminimize' command.
The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.
Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted
by applicable law.
(以下略)
```

演習環境から取得したコマンド操作ログの例である。コマンドの間違いや確認などを含むログであり、受講者が意図した行動が何であったかを確認するには必要のない要素が含まれているため、このままコマンドの入力を採点や評価に使うには適さない。そこで、LLM を使った要約を行う。以下は上記のコマンド操作ログを LLM によって MITRE ATT&CK Enterprise Matrix に基づいて要約した例である。

この操作ログから、以下の MITRE ATT&CK の手法が使われていることがわかります。

- 初期アクセス(TA0001)
 - 外部リモートサービス(T1133) - 社外から SSH でアクセス
- 実行(TA0002)
 - コマンドとスクリプトの実行(T1059) - 各種コマンドの実行
 - シェル(T1059.003) - SSH 経由でシェルを取得
- 権限の昇格(TA0004)
 - バックドアアカウントの作成(T1136) - バックドアアカウントを作成
- 発見(TA0007)
 - データからの情報収集(T1074) - データベースから情報を抽出
- コレクション(TA0009)
 - データの収集(T1005) - データベースの内容をファイルに保存
- エクスフィルトレーション(TA0010)
 - データの圧縮(T1002) - 圧縮なしでデータをエクスフィルトレート
 - データの暗号化(T1022) - 暗号化なしでデータをエクスフィルトレート
- コマンドとコントロール(TA0011)
 - リモートアクセスツール(T1219) - SSH を使用したリモートアクセス

本質的な操作目的ではない内容や操作ミスが精査され、操作の目的となる内容を要約しており、これを演習の採点に使うことで受講者の本質的な行動を評価することができる。なお、LLM による要約の内容が一定の形式で保たれるよう、プロンプトの入力はすべて統一し、temperature や Top P といったランダム性を制御するパラメータをできるだけ一定の出力となるよう調整した上で、MITRE ATT&CK Enterprise Matrix に基づいて要約するよう設定している。

7-4 類似度比較

演習環境へのサイバー推論結果と MITRE ATT&CK Matrix を用いて生成した採点用のデータセットと、操作ログをもとに LLM による要約で生成した受講者の行動データを、TF-IDF を用いた文章類似度の比較をすることで、受講者の採点に利用するための定量的なデータとして取得する。TF-IDF は文書の中で頻度の高い単語と文書集合の中で頻度が低い単語に重みをつけてベクトル化できる手法で、コサイン類似度を以て文書の類似度の比較に使われる手法として知られている。

受講者の行動データは、比較のため 3 つの LLM モデルを用いて要約し、モデル毎の要約の精度の確認と、モデル間の差異を確認することで採点への影響や採点手法の妥当性を確認した。

また、受講者のログはシナリオが正しく遂行できたものからほとんど実行できていなかったものまでさまざまであったが、採点が正しく行われるかを確認するために、特徴的であった 5 名分の操作ログを抜粋して利用した。表 7.2 に、今回の実験に使用した操作ログデータを確認した内容と、そこから予想される採点のあるべきスコアの予想を示す。

表 7.2 実験に使用した操作ログの内容

学生	操作内容	採点予想
Student_A	ほぼ想定通りにシナリオを消化し高得点が見込めるデータ	高
Student_B	A とほぼ同じだが、シナリオに不要なコマンドの入力やオプションの差異があるデータ	高
Student_C	A とほぼ同じだが、コマンドの入力間違いが多く見られるデータ	高
Student_D	シナリオが消化できておらず、途中で終わっているため得点があまり見込めないデータ	中
Student_E	初期状態からあまり進まず、コマンド入力間違いも多いため得点が見込めないデータ	低

シナリオを正しく理解して演習を実施できており、高いスコアが得られるべきと思われるデータに加え、細かな違いは見られるが本質的な行動目的が同じで、同等の評価をされるべきデータを用いて、正しい採点結果が得られることを確認できるものを選定した。また、シナリオが遂行できていないなど、前述のデータと比較して低いスコアとなるべきデータも用いて差異を比較できるようにした。

これらの操作ログから LLM によって要約されて生成された攻撃や防御の行動データが、コマンドの間違いやオプションの違いなどに依らず、シナリオに関係する本質的な行動を正しく要約できていれば、A・B・C はほぼ同様の結果となるはずであり、かつ D や E よりも高いスコアを示すはずである。

8 提案手法の実験結果と考察

8-1 実験結果

表 8.1 に、採点用データと受講者の行動データの類似度を比較したものを示し、攻撃・防御フェーズごとにグラフ化したものを図 8.1 に示す。

表 8.1 採点用データと受講者の操作データの攻撃・防御フェーズ毎の類似度比較結果

LLM Model	Student_A		Student_B		Student_C		Student_D		Student_E	
	ATT	DEF	ATT	DEF	ATT	DEF	ATT	DEF	ATT	DEF
Claude v2	0.4169	0.1098	0.4378	0.0188	0.4075	0.1042	0.2027	0.0025	0.1045	0.0002
Jurassic-2 Ultra	0.4334	0.3867	0.4939	0.3920	0.3936	0.4127	0.2753	0.1917	0.1368	0.1165
Titan Text G1	0.2927	0.0434	0.2003	0.0881	0.1689	0.0385	0.1141	0.0011	0.0823	0.0000

攻撃フェーズに関しては、LLM モデルによるばらつきは見られるものの、全体としては想定通り A, B, C のデータが D, E よりも順当に高い類似度を示している。特に Claude による比較データは A, B, C がほぼ同水準であり、想定に最も近い結果となったが、Titan を用いた比較データは差が発生している他類似度の値そのものが低く、LLM による要約があまりうまくできていないことが窺える。

防御フェーズに関しては、圧倒的に Jurassic による比較データが想定通りの結果を出したが、他の 2 つの LLM モデルによる比較は、ある程度の傾向はあるものの全体的に低い類似度となり、採点結果としては差が少なくわかりにくい結果となった。

全体的に見ても、想定された A, B, C がほぼ同等で D, E と順当に低くなるという傾向は見られた。ただし、特に防御フェーズの比較では受講者間の差が明確にはならず一部のデータでは 0 が出てしまうなど、想定しない結果も見られた。

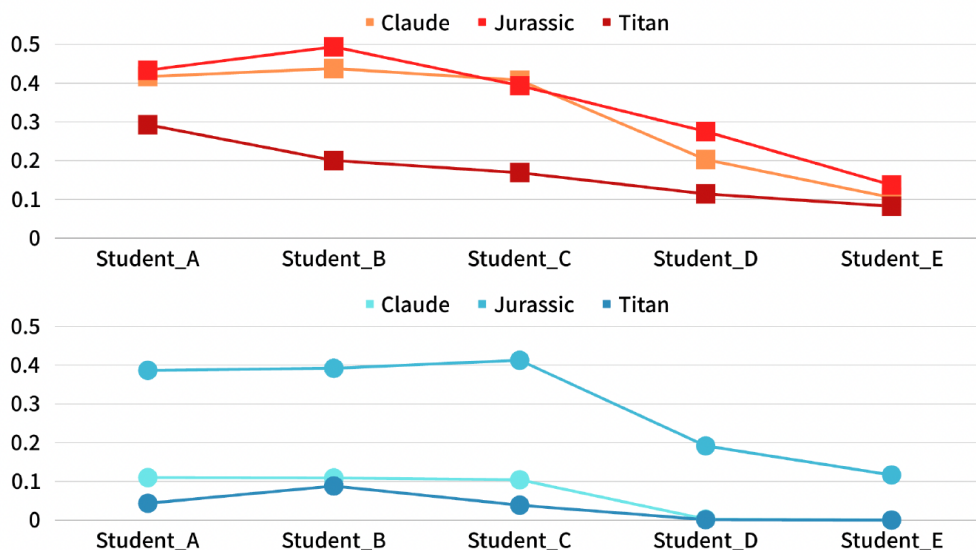


図 8.1 攻撃フェーズ（上）および防御フェーズ（下）の採点実験結果

8-2 実験結果の考察

今回の実験では3つのLLMモデルを使用し、受講者の行動を要約することで細かな行動の違いによる採点への影響を確認したが、特に同等のスコア(=類似度)となるべき操作を行った受講者であるStudent_A, Student_B, Student_Cは同一水準の類似度となり、それらよりもスコアが低くなる予想であったStudent_Dおよびstudent_Eは順当に類似度が低くなる結果となったデータが見られ、正しい採点データとして利用できる内容が見られたものもあった。コマンドオプションの違いや出力結果の軽微な違いなどであればLLMでの要約がほぼ同一の内容となるため、結果に大きな違いは見られず採点も正しく行われたと言える。

ただし、今回の実験は3つの商用LLMモデルを利用したものであるが、モデル間での出力結果の違いが大きく見られた。安定的かつ正しく受講者の操作を要約するには、よりサイバーセキュリティに特化する内容を分析可能なチューニングを行ったモデルの生成や各種の比較検証が必須となる。さらに、攻撃・防御の各フェーズに分けたとは言え、全体として複数の目的の操作ステップを組み合わせた包括的な採点を行っているため、個々の行動の妥当性を判断するには精度が低くなることが考えられ、実際に全体の類似度としては低い数値になっており、例えばポイントとなる操作毎に分割して行動内容の妥当性を判断し得点を積み上げていくチェックポイント通過型の採点方式を取るなど、採点方式の工夫も必要となる。

全体としては、実際に採点を行うための課題や利用するLLMモデルなどの検討の余地はあるが、受講者の行動結果が想定通りに評価に使えるデータとして確認できたことで、演習時の操作が細かなコマンド操作の違いや入力ミスなど、操作の目的の違いに依らない部分で不当な評価をされず、定量的な評価を実現するための一定の結果が得られた。

9 おわりに

情報セキュリティ人材の不足を解消するため、さまざまな教育の取り組みが行われており、特にサイバーレンジによる演習を用いた実践的かつ効率的な教育に関する研究が盛んに行われている。しかし、シナリオの不足や演習環境の維持管理の困難さなど演習の実施にはさまざまな課題があり、また、評価や採点に関する手法もアンケートやヒアリング等の定性的な評価が中心で定量的な手法が確立していない。

サイバーセキュリティ分野では、人工知能技術を用いた研究も進んでおり、演習でもシナリオ生成の自動化や、シナリオ評価・採点にもこれらの技術を活用することで効率的かつ精度の高い採点結果が期待できる。

そこで本研究では、人工知能技術を使ってサイバーインシデントに関連する様々な行動を推論する自律サイバー推論システムの研究を利用することを検討し、既存の機械学習によるペネトレーションテストの手法を改良して開発したRedchefを使ったサイバー推論によるシナリオ生成システムを実装した実験を行った。

ターゲットとしてdocker HUBに公開されたコンテナイメージを使った実験結果から、複数のシナリオを効率的に生成できることを確認し、サイバーレンジのシナリオ開発に利用できることが確認できた。

また、シナリオの定量的な評価・採点に利用できる手法として、大規模言語モデル(LLM)を活用した演習採点手法の検討と、実際の演習から得たログデータを用いた定量的採点実験を行った。定量的な評価や採点の自動化に向けた一定の成果が得られたが、更なる精度の向上やLLMモデル間のばらつきなどの他、TF-IDFを用いた類似度比較の妥当性やサイバー推論、LLM精度の向上などさまざまな課題も見つかった。

これらの課題を解消するためには、人工知能技術を動作させる環境と精度の向上が必須である。本研究で用意した環境も一定の検証が可能であったが、より多くのシナリオを効率的に生成し、かつ新たな脆弱性に対応し続けるには、さらに高度な推論を効率的に実施できる手法や、LLMのモデルのチューニングのほか、サイバーセキュリティ分野の特性上、インターネット接続がないローカル環境での動作の検討が必要となる。

なお、本研究期間では機械学習用に準備した機器の故障が発生し、検証に必要な十分なデータを用意することができなかったり、実際の演習を行う機会が極めて少なくなるなどの制限があったが、演習の評価・採点に関する手法も検討できたため、今後改めて各手法の精度の向上や、実際の演習による検証を進める。

近年の人工知能技術の発達や大規模言語モデルの急速な発展は、これまで困難だったサイバー推論の自動化や文章の要約などで実用性の高い成果を上げており、これらを用いた本研究の結果は、サイバーセキュリティ演習の実施や評価・採点を通して、シナリオ生成や環境構築などを含めた情報セキュリティ演習の効果や価値を高めることができ、情報セキュリティ教育や人材育成の裾野を広げることが期待できる。

【参考文献】

- [1] Information Resources Management Association, "Cyber Security and Threats: Concepts, Methodologies, Tools, and Applications," IGI Global, 2018.
- [2] 中田 亮太郎, 慎 祥揆, 笠井 洋輔, 豊田 真一, 瀬戸 洋一, “エ コシステムを実現するサイバーセキュリティ演習システム cyexec の開発”, 情報処理学会デジタルプラクティス, Vol.11, No.2, pp.414–433(2020).
- [3] Muhammad Mudassar Yamin, Basel Katt, Vasileios Gkioulos, "Cyber ranges and security testbeds:Scenarios, functions, tools and architecture", Computers Security, Volume 88(2020).
- [4] Ryotaro Nakata, Akira Otsuka, "Cyexec*:A highperformance container-based cyber range with scenario randomization", IEEE Access, Vol.9(2021).
- [5] 藤本 大輔, 大塚 玲, “深層強化学習を用いた自律サイバー 推論システムの研究”, 2021 年暗号と情報セキュリティシンポジウム(2021).
- [6] DARPA, “Cyber Grand Challenge”, <https://www.darpa.mil/program/cyber-grand-challenge>, 2024 年 6 月 30 日確認.
- [7] 佐竹 達也, 大塚 玲, “部分観測マルコフ決定過程によるニューラルエージェント強化学習を使用した自律型 SQL インジェクション攻撃手法”, 2022 年暗号と情報セキュリティシンポジウム(2022).
- [8] Z. Hu, R. Beuran, Y. Tan, "Automated PenetrationTesting Using Deep Reinforcement Learning", 2020 IEEE European Symposium on Security and Privacy Workshops (EuroSPW)(2020).
- [9] DeepExploit, "Metasploit Meets Machine Learning", <https://www.mbsd.jp/research/20180228/metasploit-machine-learning/>, 2024 年 6 月 30 日確認.
- [10] 米田 智則, 大塚 玲, “深層強化学習に基づくペネトレーションテスト手法の提案”, 第 37 回人工知能学会全国大会(2023).
- [11] RAPID7 Metasploit Documentation RPC API, <https://docs.rapid7.com/metasploit/rpc-api/>, 2024 年 6 月 30 日確認.
- [12] Md Ariful Haque, Sachin Shetty, Charles A. Kamhoua, and Kimberly Gold. Adversarial technique validation defense selection using attack graph attck matrix. In 2023 International Conference on Computing, Networking and Communications(ICNC), pp. 181–187, 2023.
- [13] MITRE ATT&CK Enterprise Matrix, <https://attack.mitre.org/matrices/enterprise/>, 2024 年 6 月 30 日確認.

〈発 表 資 料〉

題 名	掲載誌・学会名等	発表年月
自律サイバー推論システムを利用したサイバーレンジシナリオの生成手法の検討	科学情報技術フォーラム 2023	2023 年 9 月
情報セキュリティ演習における LLM を用いた定量的採点手法の検討	暗号と情報セキュリティシンポジウム 2024	2024 年 1 月