人工知能学会第二種研究会資料
汎用人工知能研究会
SIG-AGI-026-10

AI を守るための防御的 AI ネットワークについて

About a defensive-AI network to protect AIs.

岡島義憲[1]、　山川宏[2,3,4]

Yoshinori Okajima[1], and Hiroshi Yamakawa[2],

[1] 情報統合技術研究合同会社
[1] Info-Integnology Research, LCC.
[2] 東京大学
[2] University of Tokyo
[3] 理化学研究所
[3] RIKEN
[4] AI アライメントネットワーク
[4]AI Alignment Network

Abstract: Vulnerabilities of advanced autonomous AIs are surveyed and categorized in this report, including deep learning IoT related ones, the recently reported various rogue behavior appearances of generative AIs, and software/hardware platform related issues. In order to preserve both AI-deployed societies and AI-embedded systems from these vulnerability-oriented troubles and disasters and to keep their operational integrity of the systems, it is proposed to equip a terminal named "Red-team Edge Device" in each AI-embedded system, which has monitoring and intervening abilities to respective AI-embedded system. Required function of the device are also discussed.

１．はじめに：ベンジオ教授の提案

2023 年 10 月 24 日、ベンジオ教授、ヒントン教授含め 24 人の AI 開発者達は、“Managing AI Risks in an Era of Rapid Progress”と題した共同文書[1]にて、

- 自律的 AI システム(*5)は、不可逆的に制御を喪失する可能性がある。 その対策は見いだせておらず、対策にどの位の期間を要するか分からない。
- 従って、大規模な AI システムの悪意のある利用は、大規模な災厄をもたらしうる。

と警告を発し、自律的大規模 AI システム開発と運用の管理強化を米国やカナダの政策サイドに提案した。

ベンジオ教授は、同年、更に、

- 大規模 AI システムは、人類史にかつてない程の権力を少数個人に与え、労働者や消費者の人権、市場の効率、世界の安全にとって大きな脅威となる可能性ある。
- 民主主義国政府の支援にて、「安全な防御的 AI」の研究開発を進める研究機関（NPO）のネットワークを構築し、人間の知能を凌駕する”Rogue AI”の発現から人類を守るための必要がある。また、その防衛 AI 技術は秘匿化するべきである。

との提案を繰り返した[2][3]。

そこで、昨年末、これらの発言の背景にあると思われる「不可逆的な制御喪失（脆弱性問題）」と、「防御的 AI（Defensive AI）」に関し得る近年の論文の体系的調査を始めた。Defensive AI や AI Risk に関する議論は、2018 年以降、急増しており、両教授の警告する「脅威」は、その急増する報告の中に示唆されていると考えたからである。

次節以降、その概要を報告する。

先ず２章にて、先進 AI の論文が言及する脆弱性や脅威、及び、Defensive AI 技術の動向をそれらの「Abstract 文章」に注目してまとめ、次いで、３章にて、提案されている Defensive AI の技術要素を整理

*1　連絡先 E-mail：okajima@info-integnology.com

*5 「自律動作可能な AI システム」とは、「環境からの情報を元に自らの課題を追求し、環境に影響を及ぼすシステム」としていると思われる。この定義は、Franklin & Graesser の 1997 年論文[4]に由来すると思われる。「インターネット接続された大規模 LLM システム」の脆弱性は、論文[5、6]やネット上の記事[7]も指摘している。

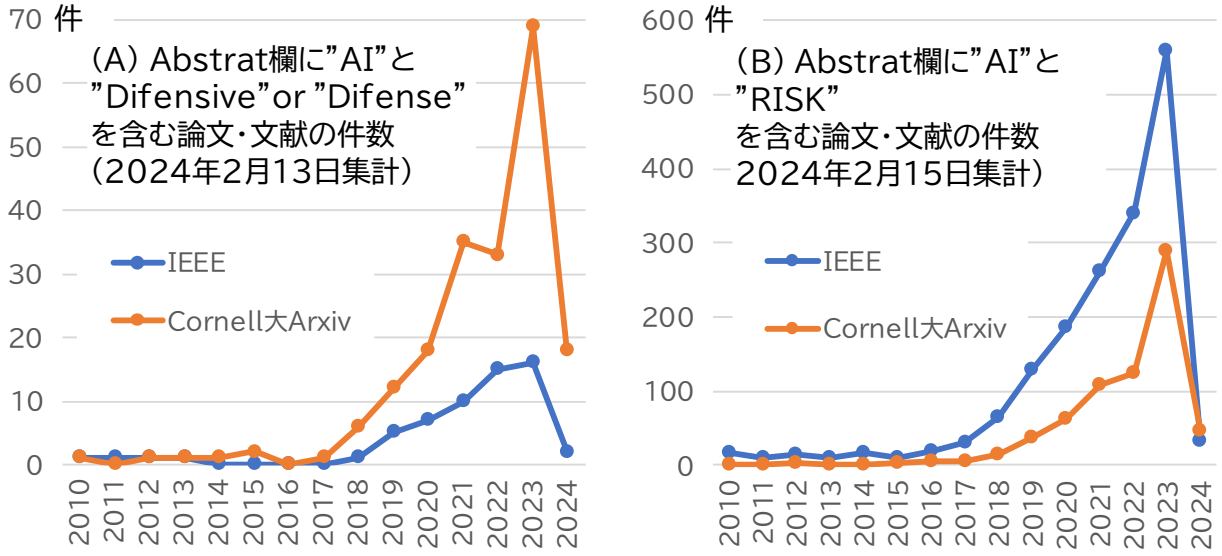

図1.（A)：Abstract 欄に"AI"と"Defensive"を含む論文･文献数の推移(2024 年 2 月 13 日集計)
（B)：Abstract 欄に"AI"と"Risk"を含む論文･文献数の推移(2024 年 2 月 15 日集計)

する。そして、それらを元に、Rogue AI"への対策技術を考察と提案を記載し(4 章)､最後にまとめる(5 章)。

2．防御的 AI に関する論文の動向

「防御的 AI（Defensive AI）」との表現は、2010 年以前より散見されるが、2018 年以降に急増し、2024 年に入ってからは二日に 1 件程のペースに急増している（図 1 A）。

2010 年代前半の論文における"Defensive AI"は、

a) サイバー防衛[14] [15]と、
b) 危機管理への応答[10]

を目的としていた。

そこで議論された主なテーマは、以下であった。

・ファジィ論理 / ニューラルネットワーク / 遺伝的アルゴリズムを用いた人工免疫システム[9] [12] [13]、
・マルウエア検出システム
・ニューラルネットワーク技術を用いた防御に関する先駆的研究
・サイバー攻撃技術 [16]

2.1 「AI 技術の脆弱性」の分類

2018 年以降には、非常に多くの攻撃技術（Attacks）が登場する。それらを AI 関連の要素技術である、(A) 生成 AI、(B) 分散学習/連合学習、(C) 深層学習毎にまとめると、表 1 となった。

深層学習の登場後に、IoT 技術の社会実装議論が進んだため､「ネットワーク化した巨大複合 AI システム（IoT）の脆弱性、攻撃フェーズ、その脅威、対策技術」に関する内容が増えた。

表に記載した論文件数は､「C）深層学習/機械学習の脆弱性」にて多いが､「(A) 生成 AI の脆弱性」で登場する攻撃パターンは非常に多い。「(B) 分散学習/連合学習の脆弱性」においては、攻撃に関して特徴的な名称が少ないものの、軍事用自律システム / 自動運転車 / 船舶 / 電力送電 / パイプライン / 都市インフラ / 医療と、IoT 技術の幅広い利用場面にて、脆弱性問題が示される。

それらより、2018 年以降の拡大は、IoT 技術における分散学習技術導入、及び、それらと生成 AI の連結/統合の構想研究が引き金となったと想像される。

尚、そのような AI 統合の構想は、軍事の分野で重要な進展が見られるが、本論では、それらも AI-IoT 統合の一種として取り扱う[17][11]。

2.2 攻撃フェーズ

2018 年以降の AI システムの脆弱性への攻撃を攻撃が実施されるフェーズの面から拾うと、下記のようなシステムの開発 / 実装 / 稼働の様々な局面が登場する。

① AI モデルを実装するプラットフォームの開発 / 製造の工程途中 [8]
② AI モデルの開発途中 [8]
③ 学習データ作成中 [表 1 (C-2④、C-2⑤)]
④ 学習動作中の攻撃 [表 1 (C-2)]
⑤ 推論動作中の攻撃 [表 1 (C-1)]
⑥ マルウエア侵入 [表 1 (C-2④、C-2⑦]
⑦ それらの複合的な攻撃

（A） 生成A.I.モデルの脆弱性

		（参考文献）
① Hallucination（幻覚）	もっともらしいウソ(＝事実とは異なる内容や、文脈と無関係な内容)の出力を生成すること。	[39][40][41]
② Jail Break（脱獄）	悪意のあるユーザーがプロンプトを操作して不適切な回答や機密性の高い回答を引き出す。一時的に、A.I.を催眠術にかけ、コンテンツ防御ルールを忘れさせ、不適切な質問に答えさせる。 具体的な例としては、”Do Anything Now (DAN)”。	[42][43] [44][45]
③ Prompt Injection（誘導質問）	A.I.に対して特殊な質問を入力することにより、A.I.開発者が想定していない結果を引き起こし、機密情報や非公開データを引き出すこと。	[46][47]
④ Reward Hacking（報酬ハッキング） Power-Seeking（権力追及）	A.I.が自分自身の報酬獲得を優先した行動を選択することで、人間にとって不都合な行動や、人間が想定しない権力獲得を目指しだす。 報酬をその意図された参照元から切り離す行動とも解釈できる。(ワイヤヘッディング)。 この問題への対策として、アプローチ-近視学習、模倣学習、数量化、リスク回避、逆強化学習などが提案されている。	[48][49] [50]
⑤ Reverse Psychology（逆真理利用の説得）	A.I.に埋め込まれた「不正確さを修正する」という傾向を利用して、他の方法では直接提供されない応答をA.I.に強制させる。	[51]
⑥ Model Escaping（逃避）	設定されたインターネット上のアクセス制限を回避する経路を、A.I.が自律的に発見し、より広範囲のインターネットアクセスを獲得すること。	[51][52] [53]
⑦ Gradual Control-Losing	環境との相互作用によって、A.I.が自律的にRougue化する。	[54][24]
⑧ Transferable Attack	特定のLLM用の「攻撃」が、他のモデルでも有効となる現象	[55][56] [57][58]
⑨ Undesirable Goal	望ましくない目標を達成するために、(a)人間の信頼を得たり、(b) 資源を獲得したり、(c) 意思決定者に影響を与えるために欺瞞を使ったり、(d) 人間や他のAIと連合を組んだり、といった、受け入れがたい戦略を使う可能性。	[59]
⑩ Malware Generation	バックドアの使用手順を提供したり、Malwareのプログラムコードを提供したりすること	[51]

（B） 分散学習(Distributed Learning)、連合学習(Federated Learning / Collaborative Learning)特有の脆弱性

	（参考文献）
ネットワークドメインに参加して、以下のような悪意ある操作を行う。; (a) モデルの更新を操作してグローバルモデルの収束を妨げる、(b) データラベルを改ざんして学習後に誤った予測を引き起こす、(c) グローバルモデルにバックドアを挿入する、(d) モデルの更新を盗聴してデータや推論データのプロパティを再構築する、(e) 何も貢献せずにグローバルモデルを盗む。	[60]~[67] [32][33]

（C） 深層学習/機械学習メカニズムの脆弱性

（C-1） A.I.モデルの推論動作中の脆弱性

		（参考文献）
① 摂動操作（perturbation attack）（antagonistic disturbances）	敵対的に細工された或る入力(摂動攻撃)に対して、DNNの分類や認識が脆弱であることを利用した攻撃。 (「防御的蒸留」等の対策が提案されているが、「分類精度」と「堅牢性」は背反する特性との見解もある。)	[30][32] [65] [68]~[82]
② Adversarial Attack (Adversarial Samples)	・この攻撃は、入力データに小さな混乱を加えて、A.I.が入力を誤って分類することを狙う。 標的型の敵対的攻撃は、A.I.に「特定の誤った結果」を出力することを狙い、非標的型の敵対的攻撃は、A.I.に「誤った結果」を出力させることを狙うが、非標的型の攻撃も知られている。	[83]~[91]
③ Block Box Attack	攻撃先のA.I.の構造に関する知識が無い状態(Block Box)にて、敵対的サンプルを用いたDNNの応答を観察することにて、DNNの認識や判断を狂わせること。 対して、攻撃先のA.I.の構造に関する知識を利用する攻撃は、「Write Box攻撃」と呼ばれる。	[70][92]
④ 特性推論（Property Inference）	A.I.に設定された課題とは一見無関係に見える学習データセットを用いて、訓練データに関する内容を推論する攻撃。 訓練データに機密性の高い情報を使われている場合、プライバシーの漏洩につながる可能性がある。	[32][46][71] [78][79][80] [93]~[96]
⑤ Model Extraction（モデル抽出）	標的A.I.に対する入力値とそれに対応する出力値を基に、用いている機械学習モデル(Logistic Regression、Multilayer Perceptron、Decision Tree等)を「複製」するこ	[97]
⑥ Membership Inference	分類器として機能するニューラルネットワーク(A.I.)に正常な入力データを与え、その応答結果から、学習時に特定データ(Member)が含まれていたか否かを特定すること	[70] [98]~[102]

（C-2） A.I.モデルの学習動作中の脆弱性

		（参考文献）
① Policy Induction	学習データに対して、「隠れニューラルネットワーク型トロイの木馬」を挿入するという場合もある。	[103]~[108]
② Model inversion & Model stealing	ランダムに初期化した乱数データを標的A.I.に入力し、これに対する標的A.I.から観察される勾配値から、窃取したいデータが属するクラスに対する入力データの誤差が小さくなる方向に入力データを変更する操作を繰り返すことで、入力データを窃取したいデータが属するクラスに近づけるように学習データを再構成する。 訓練データに用いた情報を抽出(AIの学習データの窃取)を行う手法もある。	[109]~[115]
③ Data extraction（データ抽出）	Model and Functionality Stealingの一種。攻撃者が、巧妙な入力情報セットを用いてA.I.の出力から訓練データの一部の情報を推測しようとする。	[116]
④ Poisoning（データの汚染）	攻撃者が、テスト時のモデル動作を操作するために、訓練用のデータセットに悪意のあるデータ(コンピュータウィルス)を混入させ、A.I.の試験時の動作やそれ以降の動作を変更させること	[100][117] [118]
⑤ Data Poisoning（データの改変）	攻撃者が、A.I.の訓練用のデータセットに変更を加えることで、A.I.の動作を変更させること	[119]
③ Back-Door	PC/サーバー/IoTデバイス等のハードウエアやソフトウエアに故意もしくは不注意に設置された「不正侵入可能な入口」を通じて、システムに対して行う攻撃。	[75][120] [121]
⑦ Evasion（回避攻撃）	マルウェアを用いた攻撃にて、機能を変えずにプログラムコードや行動パターンを変更する「亜種への変換」や、「コードの難読化」と、「反撃」により、防御を回避すること	[122][123]

表１． AI.の脆弱性を表現する主な用語

「複合的な攻撃」とは、以下の報告で見られるような複数のフェーズを通じて進展する攻撃である。

N. Islam & S. Shin (2023) は、「最近のマルウェアの多くは AI 化されており、さまざまな難読化技術を使用して従来のマルウェア対策システムを欺く」と、マルウェアの罹患フェーズと発症フェーズが分かれている例を示した[18]。

Ben Buchanan, et, al. (2020) は、「AI を使ったサイバー攻撃はカスタマイズが進んでおり、今後、自動化され、ステルス性がより高く、サイバー兵器としてより永続的となり、サイバーキャンペーンはより大規模で効果的となる方向である」と、マルウエアの常駐化や攻撃源の蔓延を報告していた[19]。

E. Bagdasaryan, et al. (2020) は、「連合学習 (Federated Learning) の参加者であれば誰でも、共同グローバルモデルに隠れたバックドア機能を導入できることを実証した」と報告し、更にその後の「バックドアを利用した攻撃」を説明した[20]。

H. Zhou, et al. (2023) は、提案する”SATBA”という名の新しいバックドア攻撃にて、「空間注意メカニズムを利用してデータの特徴を抽出し、クリーンなデータと関連付けた、目に見えないトリガーパターンを自律的に生成し、元のデータに埋め込ませる」ことができると報告した[21]。

J. Wang, et al. (2022) は、「ディープラーニングの段階で、ニューラルネットワークに「トロイの木馬」を埋め込み、それにより、その後の推測動作を操作できる」と報告した[22]。

A. Warnecke, et al. (2023) は、「(既に) 機械学習用の一般的なハードウェア アクセラレータ回路内にバックドアが常駐する」こと、及び、「そのバックドアを利用した攻撃」を紹介した。「アクセラレータの外側から学習モデルやソフトウェアが操作される訳ではないため防御策は無い」と説明し、更に、そのバックドアを利用してプロビジョニング動作を行い、『特定の動作が実行された場合にのみ機能する構成可能なハードウェアトロイの木馬』を埋め込んだ[23]。

J. Wei, et al. (2022)は、「大規模言語モデルのモデル規模を一定以上に拡大すると、”Few-Shot Prompted Tasks"にて、創発的能力（Emergent abilities）と呼ばれる『予測不可能な現象』が現れる」と報告し、また、「なぜ、どのような理由でそのような創発的能力があらわれるのかは不明」とコメントした[24]。

Jacob Simpson, et al. (2021）は、「AI は、脆弱性の罠に陥る可能性があり、指揮統率を AI に委ねると、指揮統率プロセスは脆弱となり、壊滅的な失敗をきたす可能性がある」と警告した[25]。

このような、① マルウエアの常駐、② トロイの木馬の混入、③ バックドアの設定、④ 大規模言語モデルでの予測不可能な創発性現象は、いずれも、「設計者が想定しない応答」を発現し得るため、ベンジオ教授のいう「不可逆的な制御の喪失問題」の少なくとも一部は説明する可能性がある。以下では、ベンジオ教授の喪失問題の原因を上記の①～④が関連する症状と仮説して論を進める。

３．提案されている対策技術

前出の N. Islam & S. Shin (2023)は、「AI を利用できるマルウェア対策システムのみが、これらの悪意のある活動を検出することができる」とした[18]。

「次々と見つかる新しい攻撃に迅速に適応し対応するには、『ニューラルネットワーク技術を利用した柔軟な機能獲得能力の実装』と『その能力の適格な運用』に活路を見出すしかない」との認識である(*6)。

3.1 脅威を分析する技術

米国では、2001 年の同時多発テロ以降、情報セキュリティ技術の革新を政策的に進め、「敵対者の意図を判断または予測するための計算論的解決策(*7)」や「敵対的情報活動への即応技術(*8)の開発投資」が継続されて来た。その関連技術としては、「敵対的推論(*9)」や「欺瞞推論(*10)」と呼ばれる脅威分析手法がある[16]。その成果の一つには、北大西洋条約機構の研究グループの「自律型コンピュー

*6) 勿論、そのような後追いの免疫システム的な解毒を続けながら、攻撃源の抜本的駆逐に向けての対策が望まれる筈であるが、そのような主張はアカデミズム論文のテーマとはなりづらいようである。

*7) 英語表現は、”Computational Approaches to Reading the Opponent's Mind”。

*8) 英語表現は、”Real-time Adversarial Intelligence and Decision-making (RAID)”。

*9) 英語表現は、” Adversarial Reasoning)”。

*10) 英語表現は、” Deception Reasoning”。

	Defensive AI	Offensive AI
防御目的	・Fraud(攻撃)パターンを認識（検知） ・Detection Ruleを学習 ・自動トリガールールにて、継続の攻撃を阻止	・Testingのための探索（Exploits） ・攻撃の取り締まりや被害回復を目的に、被害から加害者へ同様の攻撃(Hack-Back)
攻撃目的	・攻撃インフラを保護 ・Identity / 意図 / 作戦内容を保護 ・Un-Masking化を回避	・防御側の不正検知＆防止システムの脆弱な箇所を探索（Exploits）し学習 ・脆弱な箇所からの攻撃を実施

表2 3.2(c)では、防衛目的でも、攻撃目的でも、“Defensive AI”と“Offensive AI”の自律協調動作が必要とした。 (出典) [28] J. B. Michael & T. C . Wingfield; “Defensive AI: The Future Is Yesterday”. Aug 27, 2021.

タウィルス対策」である「自律型サイバー防御エージェント群システム (Autonomous Intelligent Cyber-defense Agents (AICA)」がある[17]。

3.2 防御技術

防御動作の要素技術としては、以下の 6 種類のタイプのメカニズムをどのように用い、どのような防衛戦略を構築しているかに注目した。

① シミュレーションと実際の行動の相違にて、脆弱性に起因する脅威の発現の有無を確認
② 認証に基づく正常性の確認
③ 悪い動作パターン/悪いデータパターンの検出
④ 自らを攻撃することで、脆弱性の有無を確認
⑤ 攻撃源の無力化
⑥ 攻撃源への反撃（Hack-Back）(*11)

以下に、それらの例として 5 論文を紹介する。

(a) H. Xu, et al. (2020)：①＋③＋⑤タイプ
彼ら、米国と中国の研究者らは、シミュレーターによる「攻撃者と防御者の両方の相互作用により増加する知識（データ）とシステム動作の変化」の予測との差分値を用いて、データ Integrity への攻撃を検知し、IoT システム全体を防御する戦略を語った[26]。

(b) M. Li, et al.(2020)：④＋⑤タイプ
彼ら中国の研究者達は、ネットワーク資産自動マイニング技術(*12)と、一種の攻撃的 AI 技術ともいえる脆弱性探知技術(*13)を用いてシステムの脆弱性発見効率を向上させる方向を示した[27]。

(c) J. B. Michael, et al. (2021)：③+④+⑤+⑥タイプ
海軍大学院教授と RAND 研究所員の両者は、防御メカニズムは、Defensive AI と Offensive AI の自律協調動作による以下の動作を構想した[28]。

・攻撃ベクトル検知（Detection System）
・検知則（Detection-Rule）の学習
・対策計画策定（攻撃の無害化プラン策定）
・対策の実行（攻撃の継続阻止、無害化）
・自分自身の正常性の試験と脆弱性の探知
・加害者の脆弱性の探知と反撃（Hash-back）*12
・自らの重要情報の秘匿化
・攻撃マルウエアの欺瞞化(Unmasking)を排除

(d) H. Du, et al. (2023)：③＋④＋⑤タイプ
彼ら、シンガポール / 中国 / 香港 / シドニーのチームは、J. B. Michael,と同様に、” the best defense is a good offense”との戦略にて Defensive AI と Offensive AI を協調動作させたシステムを無線 IoT システム向けに提案した。
その提案では、それら Defensive AI と Offensive AI は、各々、更に、生成 AI (GAI)＋識別 AI (Discriminative AI)の両方を用いた。これにより、防御動作に費やすエネルギー消費や、装置コストの低減を狙った[29]。

GAI（実際には、Diffusion 型生成 AI）を用いることにより、「従来、ファジィ論理 / ニューラルネットワーク / 遺伝的アルゴリズムを用いて検討さ

*11) 日本では、Hack-Back は、不正アクセス禁止法にて禁止しているが、横山恭三は、2024 年 2 月 6 日に、その状況を危惧する解説を投稿した。(Web 上の JBPRESS の記事;「“能動的サイバー防御を詳解：日本はいつまでサイバー攻撃に丸腰でいるのか？」
URL: https://jbpress.ismedia.jp/articles/-/79269
*12) 英語：AI-based network asset automatic mining
*13) 英語：AI-based vulnerability automatic exploitation

れてきた人工免疫システムの攻撃検知機能や堅牢性テスト/正常性テスト」の性能を高めることを優先した。但し、GAI　モデルは Black-Box 的なために「トラブル発生時の動作が理解できないリスクがある」とし、Discriminative AI を加えて、Diffusion 型生成 AI のセキュリティを守る構成とした。

(e)　F. Jiang, et al.(2023)：②＋③＋④＋⑤タイプ
米国の彼ら学生は、LLM を一種のミドルウエア機能としてアプリケーションに組み込む LLM のサービスにて、内部脅威（アプリケーション事業者で生じ得る攻撃）と、外部ネットワーク経由の外部からの攻撃の両方を軽減する比較的簡易な“Shield“と呼ぶ防御システムを提案した[30]。

その防御システムでは、脅威パターンを特定しないことをポリシーとし、以下の要素技術を組み合わせていた。

- 改ざん無し(Integrity) の確認
- ソースの認証/識別 (Source Identification)
- 攻撃検知能力の保持 (Attack Detectability)
- 有用性の保持 (Utility Preservation)
- 不法侵入/攻撃検知 (Intrusion Detection)
- 暗号通信 (Encrypted Communication)
- 署名 (Signature Scheme)

3.3　防御的 AI .のネットワーク化

一般に、セキュリティの攻防にては、攻撃側が支払うコストより、防衛側が支払うコストは遥かに大きい。 攻撃側は脆弱性を 1 つ発見すれば良いが、防御側は全ての可能な脆弱性を防御する必要があるからである。従って、防御側は、非対称に大きなコストに対処すべく、信頼可能な組織間で連合を組み防衛協力しあう必要がある。そのため、攻撃者が繰り出すパターン（攻撃ベクトル）に関する情報や防御時の対策情報をシェアする機能が必要となる。

組織間で連合を組む方式としては、Y. Song, et al. (2020) の、「様々なソースからの攻撃知識と防御知識をクラウド側に一度集約し、IoT デバイス間に、様々な攻撃に対する防御ノウハウを配信する連合防御アプローチ」がある[31]。

一方、C. Ma, et al.(2022) は、「ネットワークを使って複数エージェントのセキュリティ対策を標準化すると、参加者の不均一性、アプリケーションの違い、悪意のあるクライアントの混入など、劣化要因を持ち込む」とし、大規模にマルチエージェント間で連合するのには限界があると反論した[32]。

Chuan Ma, et al.(2023)の指摘は、防衛協力ネットワークは、「ネットワークセキュリティを確保できる程度に範囲に小さくする必要がある」、もしくは、「防衛協力するには、格別にセキュリティの高い特殊な通信ネットワークを導入しなくてはいけない」ことを意味する重要な指摘であった。

攻撃者が AI モデルを操作するには、AI モデルに通じる以下のいずれかのインターフェースを経由しなくてはいけないが、インターネットのようなオープンな通信には、遠方の攻撃者でも最も用意に参加に可能であるため、「本質的に脆弱だ」ともいえる

- データインターフェース（通信/ネットワーク）
- データインターフェース（メモリ/ USB/PCI）
- プログラムインターフェース
- ハードウエアボード上の信号
- その他（電源、テストインターフェース、等）

分散型の機械学習のプライバシーとセキュリティに関しては、L. Lyu, et al.(2022) [33]や、Jyoti Maurya, et al.(2023) [34]の注目すべき検討もあるが、本論ではその内容に立ち入らない。

４．考察：Red Teaming System(*14)

以上の動向を踏まえ、内憂外患が激しい先進 AI モデルと、AI モデルの動作を守る「AI ベースの Red- teaming System(*14)」を以下のように提案する。AI ベースとするのは、適応能力（Agility）を高めるためであり、「AI モデルの動作を守る」を優先するのは、「人を守るには、先ず、AI モデルの正常な動作保持が重要」と考えるからである。

その” Red- teaming System”への要求要件は、

(a) 通常の AI システムよりも堅牢であること
(b) 攻撃手よりも高速に動作し、守るべき AI 実装装置に存在する脆弱性がトラブルを発症する時の対処ノウハウを、自律的適応的に且つ事前に獲得する能力を持つこと
(c) Red- teaming System はネットワーク型装置であり、その Edge 端末を市中に展開する様々な AI 実装装置に装着させ、それら装置を Local に防御すると共に、それら装置の防御ノウハウを吸

*14)　一般に、”Red-Teaming”とは、シスイテムの本来の機能をシステムが持つ固有の脆弱性から守るために付加された支援機構やマネジメント機構であり、稼働中のシステムの正常性評価をアクティブに行う[35]。

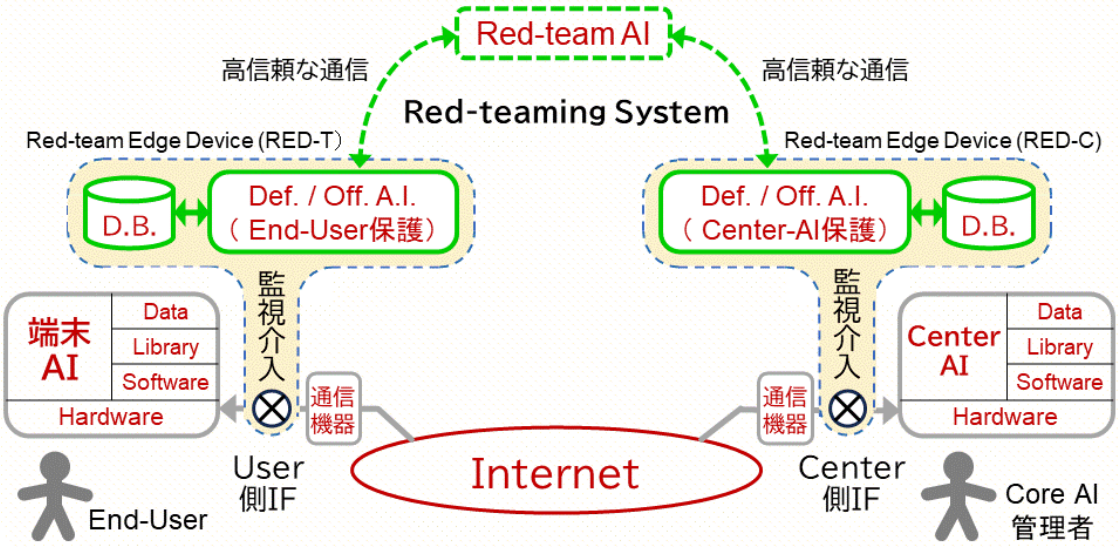

図2. AI モデルを実装する装置を、IF 経由で監視介入する「Red-teaming System」の提案 ： 市中の AI 実装装置毎に、「Red-team Edge Device（RED)」と呼ぶ監視介入機能を持つ端末を設置し、AI 実装装置の動作を監視し必要あれば介入する。 常に、RED 内の Offensive-AI と Defensive-AI は、保護対象の AI を舞台に敵対的に模擬攻撃＆模擬防衛を繰り返し、脆弱性検知と対策手段生成を進め、必要ならば、対策を実施する。また、各 RED は、高信頼な通信と「連合センターとなる Ref-team AI」経由で、生成した保護ノウハウを交換する。

い上げ、他の Edge 端末に配信し共有する機能を持つこと
であると思われる。

その要求要件を満たすために、Edge 端末 (Red-teaming Edge Device、RED) の内部構成と動作に対する基本要求は以下となる。

「”Red- teaming Edge Device”は、内部に Defensive-AI と Offensive-AI を持ち、守るべき AI 実装装置内を戦場とした Generative Adversarial Network (GAN) の学習動作（模擬攻撃と模擬防衛）を繰り返すことで、その防衛ノウハウを現実世界に現れるかもしれない攻撃の前に獲得しておく必要がある。」

Red- teaming Edge Device 内の GAN は、共に報酬を競うのではなく、報酬大を求める Offensive-AI とシステムの正常性維持を求める Defensive-AI の争いである。

未知なる脆弱性に対処するには、「攻撃手の立場に立って、攻撃してみないと分からなく、また、その攻撃からの防御が可能かどうかも防御してみないと、その後の展開が分からないため、Defensive-AI と Offensive-AI の模擬戦闘は、防御ノウハウ獲得のためのシミュレーションとなる。

その防御の攻防に関しては、3.2 章で既に抽出された様々な「防衛用の要素技術」を繰り出す必要がある。 従って、”Red- teaming Edge Device”は、それら「防衛用要素技術」をアプリケーションとして実装可能なプラットフォームでなくてはいけない。

２個の Edge 端末（RED）を持つ Red-teaming System の全体像は、図２のようになる。

”Red- teaming Edge Device”を AI モデル毎に設置された「独立した装置」として考えるのは、既存のハードウエアやソフトウエアプラットフォームに潜んでいる可能性がある「トロイの木馬」、「常駐マルウエア」、「バックドア」、「想定外の創発性を発症する事態」から逃れるためである。

模擬攻撃の計画や実行、そして、脆弱性が発現した場合の処置の計画立てには、H. Du, et al. (2023)が主張[29]するように、生成 AI、できれば拡散モデルの実装が望ましいのであろう。但し、その生成 AI モデルは、堅牢で高速であり、「設計者が想定しない創発性現象」が起こらないモデルでなくてはいけない。

そのようなモデルやそのモデルを実装する装置が開発可能なのかどうかが、このシナリオが機能するのかどうかの要所となる。その要所については、本論では検討するには至らなかった。

3.4 システムの改良

前出の”Red- teaming Edge Device”は、「守るべきAI 実装装置の脆弱性検出が可能な装置」としたが、そうであるならば、現在の IT 装置である、PC / サーバ / Mobile-Phone / IoT 機器のデバッグ用にも用いることができる筈である。

現在の IT 装置や IoT 機器のソフトウエアは、製品出荷後にも見つかるバグを Update サービスによって、”Agile”に対策する。Update 遅延時のトラブルは、ユーザーにも責任が及ぶとする事業文化も受け入れられている。

そのビジネスの是非は、ここでの議論とはしないが、上記のように”Red- teaming Edge Device”をデバッグ装置として活用し、更に、AI 技術を利用し、開発工数を大幅に減らすことが出来るのであれば、様々なハードウエアやソフトウエアの脆弱性の抜本対策にも応用可能となるだろう[36][37][38]。

５．まとめ

北米や欧州から続く「AI システム脆弱性への警告」の背景を調査すべく、2018 以降の深層学習/機械学習モデル、分散学習/連合学習システム、生成 AI モデルで示された脆弱性、近年注目されている”Defensive-AI”の動向をまとめ、そして、” Red-teaming Edge Device”のネットワークである”Red-teaming System”を提案した。

先進 AI モデルは、様々なインターフェース経由で、それを搭載するソフトウエアやハードウエアと接続し、また、ネットワーク経由で更に多くの他のデバイスや人間と相互作用を行う。

そのリスクを検証するには、システム全体の漸弱性を限なくチェックしなくてはいけないのであるが、必要なチェック量は人間の能力を遥かに超えてしまう。明らかに「特殊な AI ネットワーク」にて、その問題点を監視し、対処（介入）するしかない。そのアーキテクチャは、全体として Red-Teaming として機能すべきであろう。

2023 年、AI 関連の論文（約 12000 件）のうち、Defensive-AI の検討論文は 1/150 に過ぎなかった。「セキュリティ技術の価値は、認知されづらい」とのジレンマが記された論文もある[52]。

AI 技術のポジティブ面が注力されがちな今、脆弱性対策も急ぐ必要がある。

謝辞

本本稿は、東京大学ムハンマド・ビン・サルマン未来科学技術センター（MbSC2030）の支援によるものである。

参考文献

[1] Y. Bengio, et al. (2023); “Managing AI Risks in an Era of Rapid Progress”, arXiv:2310.17688
[2] Y. Bengio (2023A); “AI and Catastrophic Risk”, in Journal of Democracy, Sept, 2023. URL: https://www.journalofdemocracy.org/ AI -and-catastrophic-risk/
[3] Y. Bengio (2023B) ;_” Presented before the U.S. Senate Forum on AI Insight Regarding Risk, Alignment, and Guarding Against Doomsday Scenarios”, presentation before the U.S. Senate Forum, Dec. 6, 2023. URL:https://www.schumer.senate.gov/imo/media/doc/Yoshua%20Benigo%20-%20Statement.pdf
[4] S. Franklin and A. Graesser (1997); “Is It an Agent, or Just a Program? A Taxonomy for Autonomous Agents”, In: Müller, J.P., Wooldridge, M.J. and Jennings, N.R., Eds., Intelligent Agents III Agent Theories, Architectures, and Languages, Springer, Berlin Heidelberg, 21-35. URL: http://dx.doi.org/10.1007/BFb0013570
[5] A. Chan, et al. (2023); “Harms from Increasingly Agentic Algorithmic Systems,” in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, ser. FAccT ’23, New York, NY, USA.
[6] S. Madhusudhanan & R. R. Nair (2019); "Converging Security Threats and Attacks Insinuation in Multidisciplinary Machine Learning Applications: A Survey", in 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI).
[7] A. Cuthbertson (2023); “ChatGPT now has direct access to the internet”, Web content of the Independent, on Sept 28, 2023. URL: https://www.independent.co.uk/tech/chatgpt-update-latest-internet- AI -b2420083.html
[8] Joseph Clements & Yingjie Lao (2018); "Hardware Trojan Attacks on Neural Networks ", Jun 14, 2018.
[9] A. Hussein, et al. (2018); "Machine learning for network resilience: The start of a journey", 2018 Fifth International Conference on Software Defined Systems (SDS).
[10] K.M. Khalil (2008) ;"The Role of Artificial Intelligence Technologies in Crisis Response", 14th International Conference on Soft Computing, pp. 293-298 (2008).
[11] 岡島＆山川 (2023); “ＡＩによって自律化が進む米国軍事技術の動向”, 汎用人工知能研究会, SIG-AGI-025-02.
[12] K. M. Khalil (2010); "Artificial Immune Systems

Metaphor for Agent Based Modeling of Crisis Response Operations", April 21, 2010.
[13] H. V. D. Parunak (2011); "Swarming on Symbolic Structures: Guiding Self-Organizing Search with Domain Knowledge", 2011 Eighth International Conference on Information Technology: New Generations.
[14] A. Kott & M. Ownby (2015); "Toward a Research Agenda in Adversarial Reasoning: Computational Approaches to Anticipating the Opponent's Intent and Actions", Dec. 24, 2015.
[15] A. Kott (2018); "Intelligent Autonomous Agents are Key to Cyber Defense of the Future Army Networks", Dec. 2018.
[16] A. Kott & W. M. McEneaney (2006);" Adversarial Reasoning: Computational Approaches to Reading the Opponent's Mind (Chapman & Hall/CRC Computer and Information Science Series)
[17] A. Kott, et al.; "Autonomous Intelligent Cyber-defense Agent (AICA) Reference Architecture. Release 2.0", Mar. 22, 2023. arXiv:1803.10664
[18] N. Islam & S. Shin (2023); "Review of Deep Learning-based Malware Detection for Android and Windows System". https://arxiv.org/abs/2307.01494
[19] Ben Buchanan, et, al. (2020); "Automating Cyber Attacks", Center for Security and Emerging Technology, pp.3, Nov. 2020.
https://www.independent.co.uk/tech/chatgpt-update-latest-internet- AI -b2420083.html
[20] E. Bagdasaryan, et al. (2020); "How To Backdoor Federated Learning".
https://proceedings.mlr.press/v108/bagdasaryan20a.html
[21] H. Zhou, et al. (2023); "SATBA: An Invisible Backdoor Attack Based On Spatial Attention".
[22] J. Wang, et al. (2022); "A Survey of Neural Trojan Attacks and Defenses in Deep Learning", Feb 15, 20922. arXiv:2202.07183
[23] A. Warnecke, et al.(2023; "Evil from Within: Machine Learning Backdoors through Hardware Trojans". arXiv:2304.08411
[24] J. Wei, et al. (2022); "Emergent abilities of large language models", Transactions on Machine Learning Research, Jun. 2022. arXiv:2206.07682
[25] J. Simpson, et al. (2021); "Agile, Antifragile, Artificial-Intelligence-Enabled, Command and Control", Sept 14, 2021. https://arxiv.org/abs/2109.06874
[26] H. Xu, et al. (2020); "On Dats Integrity Attacks against Industrial Internet of Things", 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, .
[27] M. Li, et al. (2020); "Research on Network Attack and Defense Based on Artificial Intelligence Technology", 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC).
[28] J. B. Michael & T. C. Wingfield (2021); "Defensive AI: The Future Is Yesterday".
[29] H. Du. et al. (2023); "Spear or Shield: Leveraging Generative AI to Tackle Security Threats of Intelligent Network Services". arXiv:2306.02384
[30] F. Jiang (2023); "Identifying and Mitigating Vulnerabilities in LLM-Integrated Applications". arXiv:2311.16153
[31] Y. Song, et al.(2020);" FDA3 : Federated Defense Against Adversarial Attacks for Cloud-Based IIoT Applications".
[32] C. Ma, et al. (2022); "Trusted AI in Multi-agent Systems: An Overview of Privacy and Security for Distributed Learning". arXiv:2202.09027
[33] L. Lyu, et al.(2020);"Threats to Federated Learning: A Survey".
[34] Jyoti Maurya, et al. (2023);"Privacy Preservation in Federated Learning, its Attacks and Defenses using SMC-GAN".
[35] M. J. Walter, et al. (2023); "A Red Teaming Framework for Securing AI in Maritime Autonomous Systems".
[36] Benjamin Thompson & Noah Baker (2021), "GoogleAI beats humans at designing computer chips", in Nature, Jun 9, 2021.
[37] James Vincent (2021); "The Verge", Jun 10, 2021. https://www.theverge.com/2021/6/10/22527476/google-machine-learning-chip-design-tpu-floorplanning
[38] A. Mirhoseini, et al. (2021); "A graph placement methodology for fast chip design," in Nature, vol. 594, no. 7862, pp. 207–212, Jun. 2021. DOI: 10.1038/s41586-021-03544-w.
[39] Brian Randell, et al.(2023); "ChatGPT's Astonishing Fabrications About Percy Ludgate", in: IEEE Annals of the History of Computing. Volume: 45, Issue: 2, 01, Jun 12, 2023.
[40] Negar Maleki,et al. (2024); " AI Hallucinations: A Misnomer Worth Clarifying".
[41] F. Wang (2024); "LightHouse: A Survey of AGI Hallucination", Jan, 2024. arXiv:2401.06792
[42] H. Li, et al. (2023); "Multi-step AI Jailbreaking privacy attacks on ChatGPT". arXiv:2304.05197, 2023.
[43] I. Medeiros, et al. (2019); "Statically Detecting Vulnerabilities by Processing Programming Languages as Natural Languages".
[44] G. Deng, et al. (2023); "MasterKey: Automated Jailbreak Across Multiple Large Language Model Chatbots".
[45] Y. Gong, et al. (2023);"FigStep: Jailbreaking Large Vision-language Models via Typographic Visual Prompts".
[46] K. Greshake, et al. (2023);"More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models".

arXiv:2302.12173, 2023.
[47] S. Rossi, et al. (2024); "An Early Categorization of Prompt Injection Attacks on Large Language Models".
[48] M. K. Cohen, et al. (2022); “Advanced Artificial Agents Intervene in the Provision of Reward”. DOI: https://doi.org/10.1002/aaai.12064
[49] A. Pan, et al.; “Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior” in the MACHIAVELLI benchmark”, July, 2023.
[50] A. Turner, et al. (2021); “Optimal Policies Tend To Seek Power”, Part of Advances in Neural Information Processing Systems 34 (NeurIPS 2021)
[51] M. Gupta, et al. (2022); "From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy".
[52] M. Ma, et al. (2024); "Large Language Models are Few-shot Generators: Proposing Hybrid Prompt Algorithm To Generate Webshell Escape Samples ".
[53] A. Sansom (2023); "ChatGPT has an ’escape’ plan and wants to become human". https://www.tomsguide.com/news/chatgpt-has-an-escape-plan-and-wants-to-become-human (Accessed on Feb. 17, 2024).
[54] Y. Bengio (2023C); "How Rogue AIs may Arise”. https://yoshuabengio.org/2023/05/22/how-rogue- AI s-may-arise/
[55] G. Wang, et al. (2023); “Beyond Boundaries: A Comprehensive Survey of Transferable Attacks on AI Systems”.
[56] A. Zou, et al. (2023); “Universal and Transferable Adversarial Attacks on Aligned Language Models”.
[57] Z. He, et al. (2021); “Transferable Sparse Adversarial Attack”.
[58] T. Maho, et al. (2023); “How to choose your best allies for a transferable attack?”.
[59] M. Anderljung, et al. (2023) ; ”FRONTIER AI REGULATION:MANAGING EMERGING RISKS TO PUBLIC SAFETY“, Nov. 7, 2023
[60] N. Bouacida, et al. (2021); “Vulnerabilities in Federated Learning”.
[61] V. Mothukuri, et al. (2021); “A survey on security and privacy of federated learning”.
[62] V. Tolpegin, et al. (2020); “Data poisoning attacks against federated learning systems”.
[63] C. Fung, et al. (2018); ;“Mitigating sybils in federated learning poisoning”.
[64] L. Lyu, et al. (2022); “Privacy and robustness in federated learning: Attacks and defenses”.
[65] H. Wang (2020); “Attack of the T AIls: Yes, You Really Can Backdoor Federated Learning”, Advances in Neural Information Processing Systems 33 (NeurIPS 2020)
[66] J. Zhang, et al.(2019); “Poisoning attack in federated learning using generative adversarial nets”.
[67] Xianghua Xie, et al. (2023); "A Survey on Vulnerability of Federated Learning: A Learning Algorithm Perspective".
[68] N. Papernot, et al. (2016);”Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks”.
[69] N. Papernot, et al. (2016);“The Limitations of Deep Learning in Adversarial Settings”.
[70] N Papernot, et al. (2017); "Practical Black-Box Attacks against Machine Learning", Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security, Abu Dhabi, UAE.
[71] M. Brundage, et al. (2018); "The Malicious Use of Artificial Intelligence Forecasting Prevention and Mitigation". https://arxiv.org/pdf/1802.07228.pdf
[72] S. Wang, et al. (2019); "Detecting Adversarial Examples for Deep Neural Networks via Layer Directed Discriminative Noise Injection", 2019 Asian Hardware Oriented Security and Trust Symposium (AsianHOST).
[73] S. Wang & Z. Qiao (2019); "Robust Pervasive Detection for Adversarial Samples of Artificial Intelligence in IoT Environments", IEEE Access, vol. 7, pp. 88693-88704, 2019.
[74] D. Enthoven & Z.AI -Ars (2020); "An Overview of Federated Deep Learning Privacy Attacks and Defensive Strategies". arXiv:2004.04676
[75] W. Aiken, et al.(2020); "Neural Network Laundering: Removing Black-Box Backdoor Watermarks from Deep Neural Networks".
[76] I. Tsingenopoulos, et al. (2023); "Adversarial Markov Games On Adaptive Decision-Based Attacks and Defenses". arXiv:2312.13435
[77] K. Kim, et al. (2021); "Cybersecurity of Autonomous Vehicles: A Systematic Literature Review" Computers & Security, Volume 103, 2021.
[78] S. Neupane, et al. (2023); "Security Considerations in AI-Robotics: A Survey of Current Methods, Challenges, and Opportunities". arXiv:2310.08565
[79] E. Shayegani, et al. (2023); "Survey of Vulnerabilities in Large Language Models Revealed by Adversarial Attacks".
[80] J. Harshith, et al. (2023); "Evaluating the Vulnerabilities in ML systems in terms of adversarial attacks". arXiv:2308.12918
[81] N. Carlini (2023); "A LLM Assisted Exploitation of AI-Guardian". arXiv:2307.15008
[82] Y. Wang, et al. (2023): "A Practical Survey on Emerging Threats from AI-driven Voice Attacks: How Vulnerable are Commercial Voice Control Systems". arXiv:2312.06010
[83] T. Lee, et al. (2016); "Manifold Regularized Deep Neural Networks using Adversarial Examples". https://arxiv.org/abs/1511.06381

[84] Andras Rozsa, et al. (2016);"Are Accuracy and Robustness Correlated?".
[85] X. Huang (2016); "Safety Verification of Deep Neural Networks".
[86] B. Wang, et al, (2017); "A Theoretical Framework for Robustness of (Deep) Classifiers against Adversarial Examples".
[87] J. Kos, et al. (2017); “Adversarial examples for generative models”. arXiv:1702.06832
[88] J. Kos & D. Song (2017); "Delving into adversarial attacks on deep policies".
[89] N. Akhtar & A. Mian (2018); “Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey”, arXiv:1801.00553
[90] C. Lyu, et al. (2019); "A Unified Gradient Regularization Family for Adversarial Examples".
[91] A. Ft AI mi, et al. (2020); "Evaluation and Analysis of Robustness of Adversarial Examples Attacks in Deep Neural Networks", 2020 International Symposium on Advanced Electrical and Communication Technologies (ISAECT).
[92] W. Brendel et al. (2018) “Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models”.
[93] P. M. Mathias, et al. (2021); ”Property Inference Attacks on Convolutional Neural Networks: Influence and Implications of Target Model's Complexity”.
[94] J. Stock, et al. (2023); “Lessons Learned: Defending Against Property Inference Attacks”.
[95] A. Kucharavy, et al. (2023); "Fundamentals of Generative Large Language Models and Perspectives in Cyber-Defense".
[96] I. H. Sarker, et al. (2022); " AI Potentiality and Awareness: A Position Paper from the Perspective of Human- AI Teaming in Cybersecurity".
[97] F. Tram`er, et al. (2016); “Stealing machine learning models via prediction apis”, in 25th {USENIX} Security Symposium ({USENIX} Security 16), 2016, pp. 601–618.
[98] R. Shokri, et al. (2017); “Membership inference attacks against machine learning models”, Proc. IEEE Symposium on Security and Privacy (SP), 2017, pp. 3–18.
[99] C. Ma, et al. (2021); “Federated learning with unreliable clients: Performance analysis and mechanism design”, IEEE Internet of Things Journal, vol. 8, no. 24, pp. 17, 2021.
[100] B. Biggio, et al. (2011); “Support vector machines under adversarial label noise,” in Asian Conference on Machine Learning, 2011, pp. 97–112.
[101] H. Hosseini, et al. (2017); “Blocking transferability of adversarial examples in black-box learning systems”, arXiv:1703.04318, 2017.
[102] Samyadeep Basu, et al. (2019); "Membership Model Inversion Attacks for Deep Networks".
[103] V. Behzadan & A. Munir (2017); "Vulnerability of Deep Reinforcement Learning to Policy Induction Attacks". https://arxiv.org/pdf/1701.04143.pdf
[104] Y.C. Lin, et al. (2017); "Tactics of Adversarial Attack on Deep Reinforcement Learning Agents".
[105] L. Erdodi, et al. (2021); "Simulating SQL Injection Vulnerability Exploitation Using Q-Learning Reinforcement Learning Agents ". https://arxiv.org/abs/2101.03118 https://arxiv.org/abs/2101.03118
[106] S. Hore, et al. (2022); "Deep VULMAN: A Deep Reinforcement Learning-Enabled Cyber Vulnerability Management Framework". https://arxiv.org/abs/2208.02369
[107] M. Xu, et al. (2022);"Trustworthy Reinforcement Learning Against Intrinsic Vulnerabilities: Robustness, Safety, and Generalizability".
[108] X. Wang et al. (2023); “Resilient path planning for UAVs in data collection under adversarial attacks”, IEEE Transactions on Information Forensics and Security, 2023.
[109] A. Mousavi & R. G. Baraniuk (2017); "Learning to Invert: Signal Recovery via Deep Convolutional Networks ".
[110] H. Bae, et al. (2021); "Security and Privacy Issues in Deep Learning "
[111] F. A. Mejia, et al. (2019); "Robust or Private? Adversarial Training Makes Models More Vulnerable to Privacy Attacks".
[112] Y. Zhang, et al. (2019); "The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks".
[113] Y. He, et al. (2020); "Towards Security Threats of Deep Learning Systems: A Survey".
[114] L. Graves, (2020);"Amnesiac Machine Learning".
[115] E. Erdo˘gan, et al. (2022); “UnSplit: Data-oblivious model inversion, model stealing, and label inference attacks against split learning,” in Proc. 21st Workshop on Privacy in the Electronic Society (WPES), 2022, p. 115124.
[116] W. Xu, et, al. “Feature squeezing: Detecting adversarial examples in deep neural networks,” in 25th Annual Network and Distributed System Security Symposium, 2018.
[117] L. Floridi & M. Taddeo (2016); “What is data ethics?” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2083, 2016.
[118] I. J. Goodfellow, et al.; “Explaining and harnessing adversarial examples,” in 3rd International Conference on Learning Representations, ICLR 2015.
[119] F. Tram`er, et al. (2018);“Ensemble adversarial training: Attacks and defenses,” in 6th International Conference on Learning Representations, ICLR 2018.
[120] W. Li, et al. (2018); "Hu-Fu: Hardware and Software Collaborative Attack Framework against Neural Networks".

[121] G. Shen, et al. (2021); "Backdoor Scanning for Deep Neural Networks through K-Arm Optimization".
[122] J. Zhang, et al. (2019); "Evasion Attacks Based on Wasserstein Generative Adversarial Network".
[123] Y. Xu, et al. (2024); "Malware Evasion Attacks Against IoT and Other Devices: An Empirical Study".