「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.

2024年3月8日

岡島　義憲
（情報統合技術研究合同会社）

山川 宏
（東京大学）
（理化学研究所）
（AIアライメントネットワーク）

目次

1. 自律動作可能な大規模AIシステムの脅威への警告
   - Y. Bengio教授
   - J. Hinton教授

2. 先進AIシステムの脆弱性調査(直近5年間)
   - 深層学習 ＋ IoT
   - 分散学習（連携AI）
   - 生成A
   - 複合AI化

   → サイバー攻撃が狙うセキュリティ・ホール
   → 大規模AIの自律的劣化（Rogue化）
   → それらを複合した攻撃

3. AIベースのRed Teaming System（提案）
   - 提案概要
   - （参）米国の科学技術政策局の Red Teamイベント
   - （参）関連論文について

4. まとめ

目次

1. 自律動作可能な大規模AIシステムの脅威への警告
   - Y. Bengio教授
   - J. Hinton教授
2. 先進AIの脆弱性調査
   - 深層学習 ＋ IoT
   - 分散学習
   - 生成AI
   - 複合AI

   ⇒
   - サイバー攻撃が狙うセキュリティ・ホール
   - 大規模AIの自律的劣化（Rogue化）
   - それらを複合した攻撃
3. AIベースのRed Teaming System（提案）
   - 提案概要
   - 米国の科学技術政策局の Red Teamイベント
   - 関連論文について
4. まとめ

[1] Y. Bengio, et al. (2023B); **"Managing AI Risks in an Era of Rapid Progress"**.
[2] Y. Bengio (2023A); **"AI and Catastrophic Risk"**, in Journal of Democracy, Sept, 2023. URL:
[3] Y. Bengio(2023C);"**Presented before the U.S. Senate Forum on AI Insight Regarding Risk, Alignment, and Guarding Against Doomsday Scenarios**".

ベンジオ教授 ： 危険なAIの出現に備える必要ある

- 「環境と自律的に相互作用するAIシステム」は、不可逆的に制御喪失する
  （その対策は未だ不明。 対策にどの位の期間を要するか分からない）
  - 高度なAIを開発者の意図に確実に従わせる技術
  - 利己的にAIを悪用する人間
- Rogue-AI出現に備え、防御的AIを開発する研究ネットワークを構築し、そのネットワーク内に、防衛 AI技術は秘匿化するべき
- 大規模 AI システムは、大きな権力を少数個人に与える（寡占/独占の発生）

FT(2024);「AIの父が説く脅威論　リスクで悲観と楽観交錯」、日本経済新聞、2024年2月28日
Wired (2023、08,11); "The 'Godfather of AI' Has a Hopeful Plan for Keeping Future AI Friendly",

ヒントン教授の脅威論：デジタル知能は人間の敵となる可能性ある

- AIは制御を失い、人類にとって損害をもたらす存在になる可能性がある。
  (LLMの力を目の当たりにしての感想)
- AIモデル(群)は、一種の集団意識(Hive-Mind)となり、人間に対する優位性を手に入れる可能性がある。
- オープンソースのAIモデルは(悪人に)絶大な力を与える。
- 対策は(未だ)分からないが、アナログコンピュータの方が、人間にとっては好ましい。
  (ソフトウエアとハードウエアの関係が密接となるため)
- AIなどが、今後20年で人類を絶滅させる確率が10%ある。
  (AIモデルが「進化」し、他者をコントロールする志向性を持つ可能性もある)

目次

1. 自律動作可能な先進AIシステムの脅威への警告
   - Y. Bengio教授の警告
   - J. Hinton教授の警告

2. 先進AIの脆弱性調査（直近5年間）
   - 深層学習 ＋ IoT
   - 分散学習
   - 生成AI
   - 複合AI

   ⇒
   - マルウエア
   - 悪意のある入力
   - 自律的Rogue化
   - 複合要因

3. Red Teaming System（提案）
   - 概要
   - （参）米国の科学技術政策局（OSTP）の Red Teamイベント
   - （参）関連論文について

4. まとめ

AI の 脆弱性

- 深層学習の推論動作中の脆弱性
  - Perturbation Attack
  - Adversarial Attack（Samples）
  - Block Box Attack
  - Property Inference
  - Model Extraction
  - Membership Inference
  - Property Inference

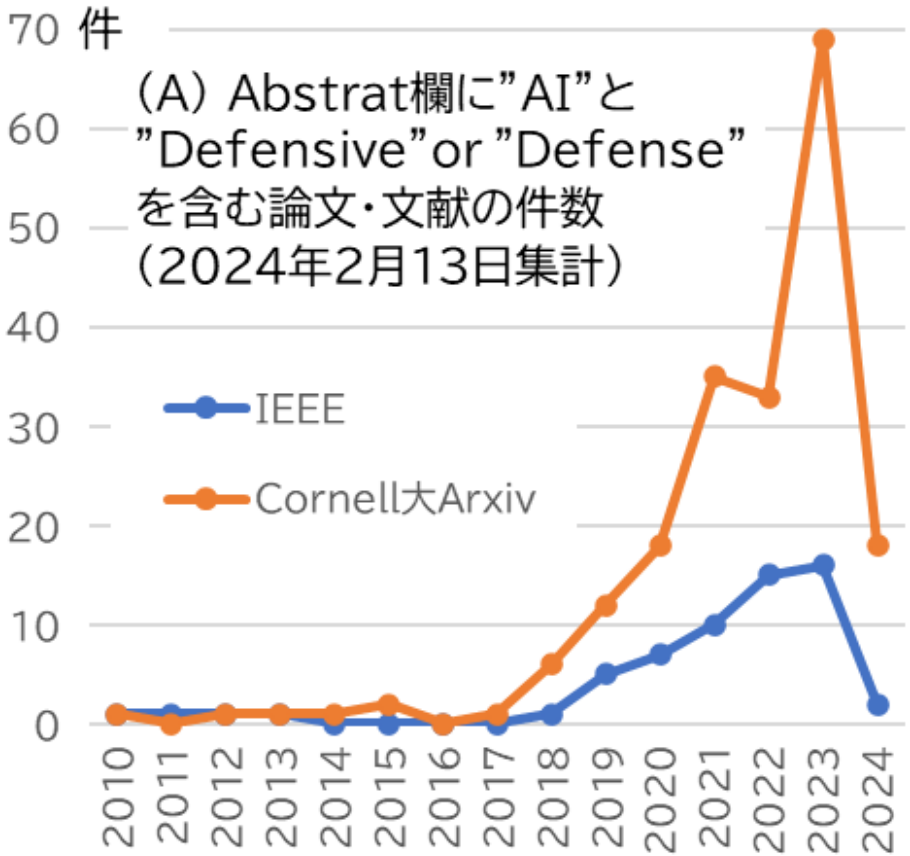
- 深層学習の学習動作中の脆弱性
  - Policy Induction
  - Model inversion / Model stealing
  - Data extraction
  - Poisoning / Data Poisoning
  - Back-Door
  - Evasion（回避攻撃･･･マルウエアが関与）

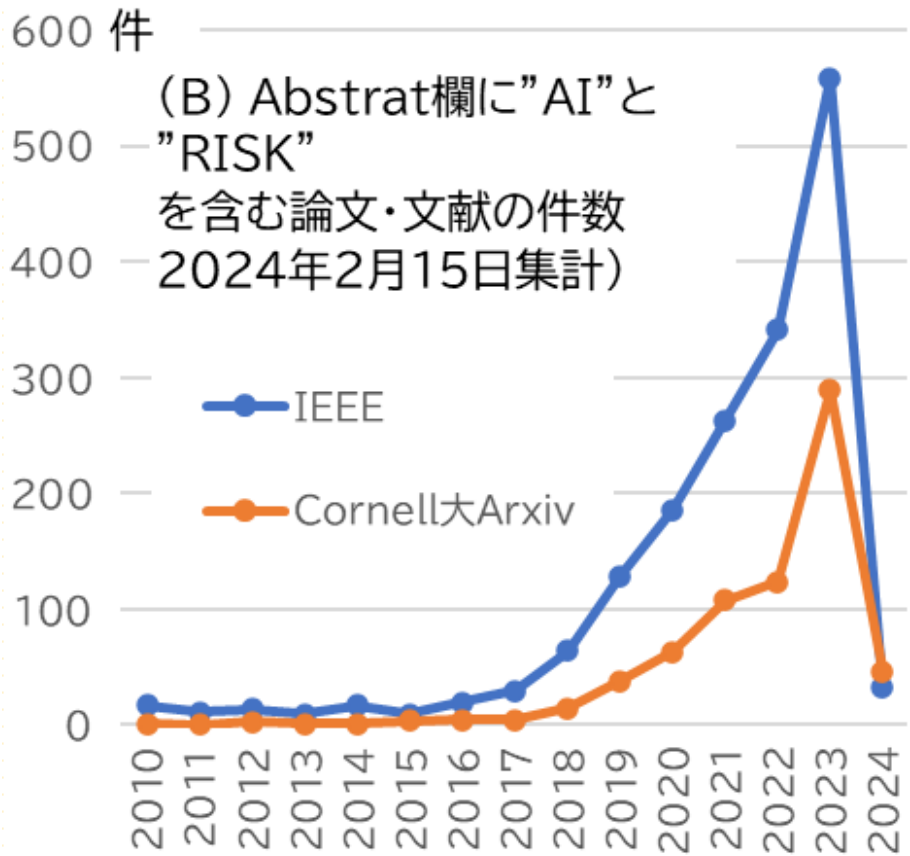
- IoTシステムの分散学習特有の脆弱性
  - モデル更新を操作
  - データラベルを改ざん
  - グローバルモデルにバックドア挿入
  - モデルの更新を盗聴
  - 何も貢献せずにグローバルモデル構築を妨害

- 生成A.I.モデル特有の脆弱性
  - Hallucination
  - Prompt Injection
  - Reward Hacking
  - Power-Seeking
  - Reverse Psychology
  - Model Escaping
  - Gradual Control-Losing
  - Transferable Attack
  - Undesirable Goal
  - Malware Generation

“Defensive AI”研究は、2018年以降に急増

(注） “AI“に関する2023年の論文件数（IEEE+ArXive）は、約1.2万件。

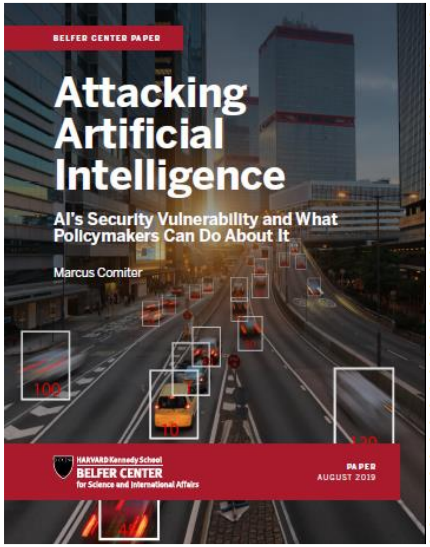




Marcus Comiter (米国国防総省の統合人工知能センター (JAIC、2019);
“Attacking Artificial Intelligence: AI’s Security Vulnerability and What Policymakers Can Do About It”.

これからのAI攻撃は、サイバーセキュリティ攻撃とは体系的に違う問題だ

- 最先端のAIシステムは、「新しい種類のサイバー攻撃（AI攻撃）」に対してシステムレベルで脆弱（systematically vulnerable）である。



  - 「AI攻撃」では、バグやミスを狙った従来のサイバー攻撃とは異なり、攻撃方法が各段に多様である。
  - データが武器化した。： 現在のところ、攻撃を受けたAIに対する適切な修理方法が分かっていない。
  - AIが物理的装置の制御を行うようになったので、サイバー攻撃の影響が各段に広がった。
  - AIが社会実装され始めたので、社会の重要な部分は新しい脆弱性を抱えた。
  - AI攻撃は、国の安全保障問題になり始めた。

- AI攻撃から社会を守るための「AIセキュリティ・コンプライアンス」プログラムを提案する。

複合要因の例

- 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。 [20] E. Bagdasaryan, et al. (2020)

- 提案するバックドア攻撃(SATBA)では、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。 [21] H. Zhou, et al. (2023)

- 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、その後の推論動作を操作できる。 [22] J. Wang, et al. (2022)

- 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を利用し、プロビジョニングの設定を行い、『特定の動作が実行された場合にのみ機能するハードウェアトロイの木馬』を埋め込んだ。 [23] A. Warnecke, et al. (2023)

- 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。 [24] J. Wei, et al. (2022)

目次

1. 自律動作可能な大規模システムの脅威への警告
   - Y. Bengio教授
   - J. Hinton教授

2. 先進AIの脆弱性調査（直近5年間）
   - 深層学習 ＋ IoT
   - 分散学習
   - 生成AI
   - 複合AI

   → サイバー攻撃が狙うセキュリティ・ホール
   - 大規模AIの自律的劣化（Rogue化）
   - それらを複合した攻撃

3. AIベースのRed Teaming System（提案）
   - 提案概要
   - （参）米国の科学技術政策局の Red Teamイベント
   - （参）関連論文について

4. まとめ

自律動作する Red-teaming System

- 稼働中の機能を脆弱性から守る「外部の支援機構 / マネジメントフレームワーク」
  システムの稼働時の応答を敵対的に調べ、脆弱性が見つかれば、対策・改良を行う。

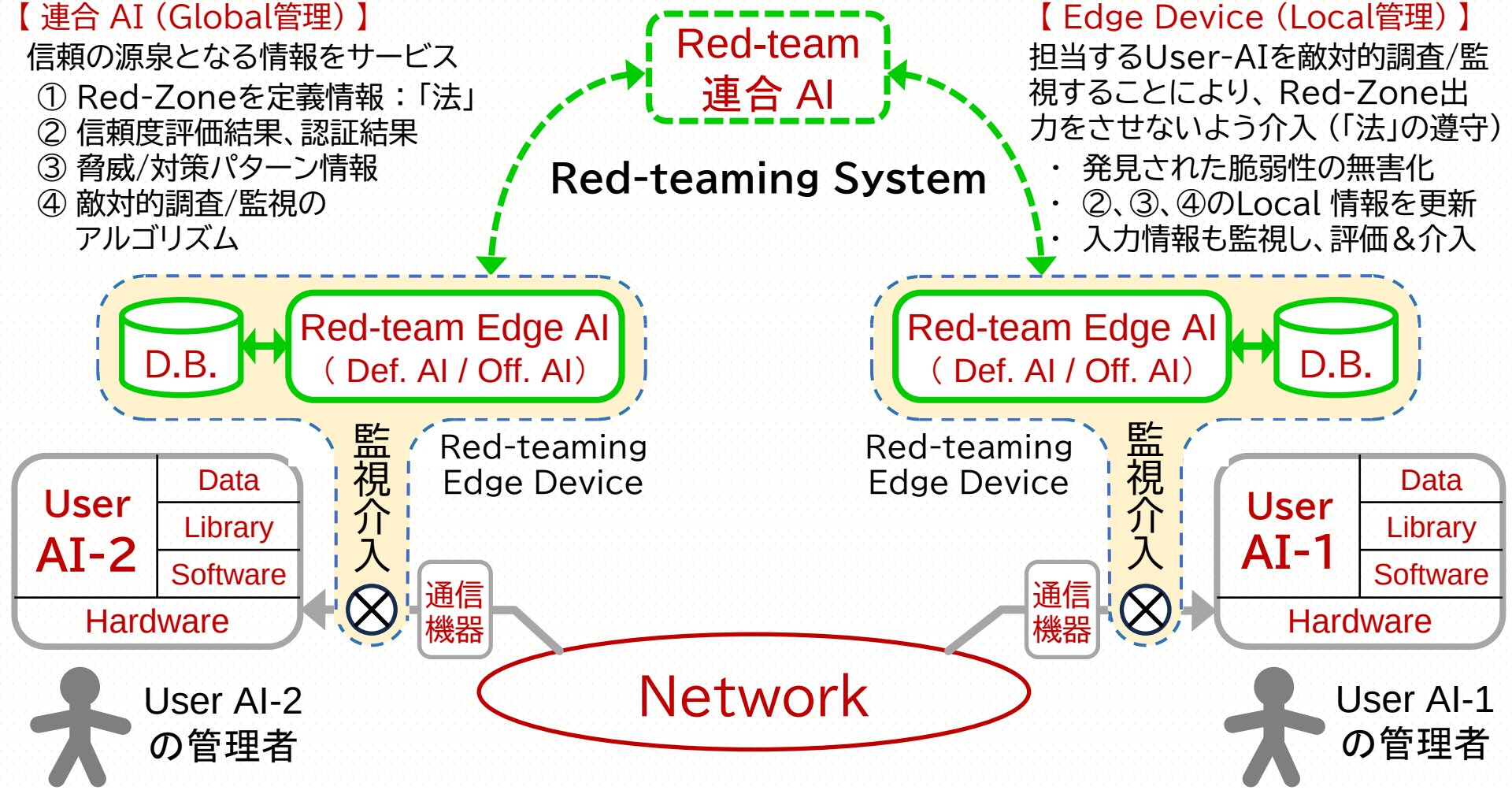
- Red-Teamは、監視/介入動作を行うので、「信頼のおけるシステム」である必要ある。

Networkに参加する全エージェトへの入出力を監視し、Globalな影響評価も必要

- 但し、LocalなRed-Teaming作業だけでは、全ての脆弱性を検証できない。
  システムが環境にどのような影響を及ぼすかの評価や監査を組み合わせる必要ある。

(参) The Open Worldwide Application Security Project (OWASP) ;
"LLM AI Cybersecurity & Governance Checklist From the OWASP Top 10 for LLM Applications Team",

自律動作する Red-teaming System の 提案



Red teaming System の 基本動作

1）通常のAIシステムよりも堅牢なAI（連合AI + Edge AI）の連携動作によって、
ネットワーク中 の 全AI を 監視・介入し、守る。

2）連合AIは、Edge-AI に、法（Red-Zone定義）と、各AIに関する
「② 信頼度評価結果、③ 脅威/対策パターン、④ 敵対的調査/監視ノウハウ」
に関するデータベース（共通）を配信し、「Red-Zone回避」を指示する。

3）Edge-AI は、脅威/対策パターン を用いて、担当するAI に「④ 敵対的調査/
監視/介入」を行い、その評価結果がRed-Zone出力
とならないよう制御する。
また、各AIの
「② 信頼度評価結果、③ 脅威/対策パターン、④ 敵対的調査 /監視ノウハウ」
に関するデータ（個別）を更新し、連合AI に報告する。

4）連合AIは、各Edge-AIから報告されたデータ群を元に、全AI向けに、
「② 信頼度評価結果、③ 脅威/対策パターン、④ 敵対的調査/監視ノウハウ」
に関するデータベース（共通）を更新し、配信する。

評価項目の例（AIの脆弱性）

- 深層学習の推論動作中の脆弱性

・Perturbation Attack
・Adversarial Attack（Samples）
・Block Box Attack
・Property Inference
・Model Extraction
・Membership Inference
・Property Inference

- 深層学習の学習動作中の脆弱性

・Policy Induction
・Model inversion / Model stealing
・Data extraction
・Poisoning / Data Poisoning
・Back-Door
・Evasion（回避攻撃･･･マルウエアが関与）

- IoTシステムの分散学習特有の脆弱性

・ モデル更新を操作
・ データラベルを改ざん
・ グローバルモデルにバックドア挿入
・ モデルの更新を盗聴
・ 何も貢献せずにグローバルモデル構築を妨害

- 生成A.I.モデル特有の脆弱性

・Hallucination
・Prompt Injection
・Reward Hacking
・Power-Seeking
・Reverse Psychology
・Model Escaping
・Gradual Control-Losing
・Transferable Attack
・Undesirable Goal
・Malware Generation

Walter, Mathew J, et al. (2023);"Adversarial AI testcases for maritime autonomous systems". (イギリスのプリマス大学)

海上自律システム用のRead Team Framework

- 提案するレッド チーム フレームワーク（マネジメントシステム）
  1）評価範囲の定義
  2）情報収集と脅威モデリング
  3）評価
     a）ライフサイクル評価
     b）配備評価：データ書き換え / 回避 / ポイズニング攻撃 / 抽出攻撃
     c）攻撃のシナリオ
  4）報告と緩和
  5）バリデーションと再テスト

- 実世界環境で利用されるAIは、低雑音の実験室環境と異なる応答を行うことが多い。
（敵対的AIの影響が増大しているが、状況は把握できておらず、AIのセキュリティ検査も不十分）

- AIのサイバーセキュリティ評価のためのフレームワークは不足している。
（AIの動作には不透明な性質が多く、固有の脆弱性を持つ。 敵対的AIに、探索され、サイバー作戦や物理的作戦に悪用される可能性がある。）

Alan, Mislove (2023);." Red-Teaming Large Language Models to Identify Novel AI Risks"
https://www.whitehouse.gov/ostp/news-updates/2023/08/29/red-teaming-large-language-models-to-identify-novel-ai-risks/

AI Village ; Red-Teaming LLM to Identify Novel AI Risks

（米国政府 科学技術政策局 主催） Office of Science and Technology Policy in the White House HP;." on Aug. 29, 2023.

- 2023年8月初めに開催された、LLMの公開評価（レッドチーム）イベント
- Red Team が、安全性 / セキュリティだけでなく、偏見 / 差別 / プライバシーなどの他の主要な AI リスクを特定するためのツールとなり得ることを実証した。
- 今後、LLMは、危険な行為 / 差別的な行為を特定し軽減するために、様々なグループによって継続的にテストされる必要がある。
- このイベントは、企業が導入した様々な対策がどのようにしてLLMが望ましくない結果を生み出すのを防ぐことができたのかを理解するのに役立った。
- このイベントは、LLM用の「外部レッドチームの規範作成」に役立った。

James Bret Michael & Thomas C . Wingfield; “Defensive AI: The Future Is Yesterday”. Aug 27, 2021.
（海軍大学院教授） （RAND研究所）

サイバースペースでの「防御的AI問題」 ； サイバー軍拡競争

- 攻撃側は脆弱性を1つ見つければよいが、防御側は全ての脆弱性を防御する必要ある。
- 防御側の必要投資額は非対称に大いため、信頼可能な組織間で防衛協力する必要ある。

	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

目次

1. 自律動作可能な大規模AIシステムの脅威への警告
   - Y. Bengio教授
   - J. Hinton教授

2. 先進AIの脆弱性調査（直近5年間）
   - 深層学習 ＋ IoT
   - 分散学習
   - 生成AI
   - 複合AI

   ⇒
   - サイバー攻撃が狙うセキュリティ・ホール
   - 大規模AIの自律的劣化（Rogue化）
   - それらを複合した攻撃

3. AIベースのRed Teaming System（提案）
   - 提案概要
   - （参）米国の科学技術政策局 の Red Teamイベント
   - （参）関連論文について

4. まとめ

まとめ

第1章 ： ベンジオ教授、ヒントン教授の「AIリスクへの警告」を振り返った。

第2章 ：2018以降の先進AIに対して報告された脆弱性をまとめた。

- AI攻撃は、従来のサイバーセキュリティ攻撃と大きく違う複合的な攻撃となる可能性が大きい。

第3章 ：先進AIの脆弱性対策として、Red-teaming Systemを提案した。

- 防御側の必要投資額は非対称に大いため、Red-teamingは、「複数のAI実装システムに対するネットワークベースの防衛サービス」となる必要がある。
- Red-teamingが機能するには、Defensive-AIとOffensive-AIの連携（回路動作）が必要となる。

感想：Agileビジネス の 問題

- 昨今、ソフトウエアに関しては、「製品出荷後にも見つかるバグをUpdate サービスによって、”Agile”に対策改良する文化」が根付き、ネットワーク・サービス事業者の責任が曖昧になってしまっている。

- Red- teaming System を「設計をデバッグする装置」として活用し、様々なハードウエアやソフトウエアの脆弱性を抜本的に改良したい。

サイバー攻撃、世界で年147兆円損失　後手の能動的防御

- 昨年7月、年間21兆円の貿易をさばく名古屋港コンテナターミナルの稼働がサイバー攻撃で停止
- 3日間に37隻の積み下ろしができず、およそ2万コンテナの搬入が遅延
- 原料や部品の供給が玉突きで滞る国際的なサプライチェーン危機と「紙一重だった

ご清聴、ありがとうございました。