

Center for Research on AI Security and Technology Evolution

AIセキュリティ研究センター

OUTLINE

AIセキュリティ研究センター（CREATE*）は、AIそのものの安全性を確保する「セキュアなAIの実現」、「AIによるサイバーセキュリティ高度化」、「AIへの信頼性向上」の三本柱を軸に研究開発を推進する拠点です。高度化するサイバー攻撃やAIの急速な社会実装に対応し、信頼できるAI技術の確立を目指しています。加えて、AIセキュリティ動向の分析や国内外の研究機関・産業界との連携を通じ、グローバルな視点から持続的で堅牢な安全基盤の構築を進めています。CREATEは、安全・安心なAIネイティブ社会の実現に向け、未来を支える次世代セキュリティ技術の創出に貢献していきます。

*CREATE は、“Center for Research on AI Security and Technology Evolution” の略称です。

PROJECTS

セキュアなAIの実現

AIの安全性と信頼性を確保する研究を行っています。特に大規模言語モデル(LLM)をはじめとする生成AIは、社会や産業での利用が急速に拡大している一方で、外部からの攻撃や予想せぬ挙動といった新たなリスクがあります。私たちは安全性、セキュリティ、サイバー攻撃への悪用リスクを重視し、攻撃に強く信頼できるAIの実現を目指しています。

AIによるサイバーセキュリティ高度化

AIとセキュリティ技術を融合させ、膨大なサイバーセキュリティ情報から価値ある知見を導き出す研究開発を推進しています。リアルタイムのビッグデータ解析により、未知の脅威やマルウェア、新たな攻撃キャンペーンを迅速に検知し、サイバー空間全体の安全性を高めます。また、人材不足という社会的課題に対応し、防御オペレーションの効率化・自律化を目指しています。

AIへの信頼性向上

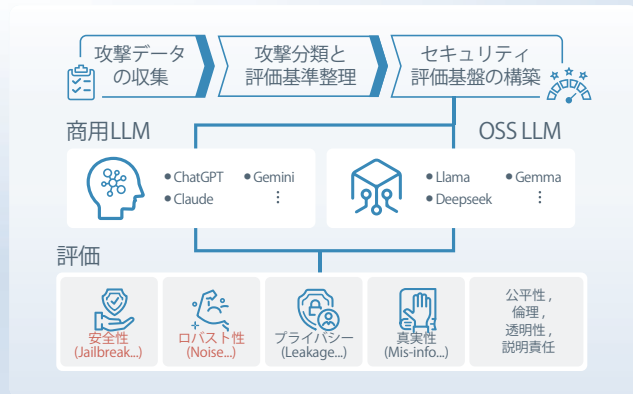
AIを安心して活用するために、AIの判断を人間が理解できるようにすることも重要です。CREATEでは、AIを用いたシステムに説明可能AI(XAI)を組み込み、AIの判断根拠を可視化する研究を進めています。AIの判断理由を人間が直感的に把握できる仕組みを開発することで、AIに対する透明性と説明責任を高めていき、AIを「理解して使う」社会の基盤づくりを目指しています。



RESEARCH TOPICS

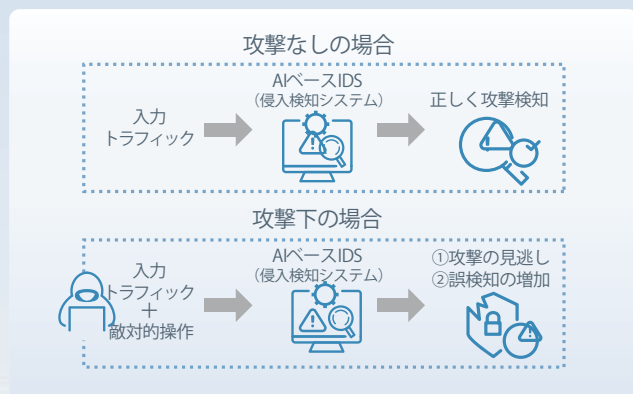
■ AI セキュリティ評価基盤

生成 AI に対するセキュリティリスクを体系的に評価する「AI セキュリティ評価基盤」は、Jailbreak や有害生成、情報漏洩などの攻撃シナリオに基づき、ベンチマークを用いて AI の挙動を定量的に評価します。複数の AI モデルを同一条件で比較評価し、その結果を可視化することで、安全性やリスク傾向の違いを直感的に把握できます。これにより、安全な AI 活用やモデル選定を支援します。



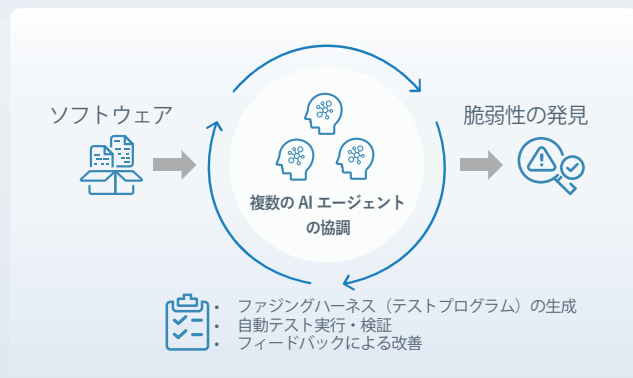
■ AI ベース IDS に対する敵対的攻撃

AI を活用して攻撃を検知する侵入検知システム (IDS) に対して、AI を欺く敵対的攻撃手法を研究しています。攻撃者は、人間には認識できない程度のデータ改変を行うことで、AI の判断を混乱させる恐れがあります。その結果、システムが実際の攻撃を見逃したり、安全な通信を誤って危険と判断したりする可能性があります。そのような課題を明らかにすることで、IDS の安全性向上に貢献することを目指しています。



■ マルチ AI エージェントを活用したソフトウェア脆弱性の発見

複数の AI エージェントが協調してソフトウェアのソースコードを解析し、ファジングのためのテスト用コード (test harness) の自動生成及び自動検証を行うことで、効率的に脆弱性を発見する手法を研究しています。従来のテスト手法において工数がかかり、かつ専門性が求められる部分を、複数の AI エージェントを協調させることで代替することを目指しています。



OTHER RESEARCH TOPICS

- セキュリティ運用自動化
- AI によるサイバー攻撃とセキュリティ対策
- セキュリティが組み込まれた AI モデル

- AI システムの脅威分析
- 解釈可能性の提供
- AI のデータプライバシー



国立研究開発法人情報通信研究機構
サイバーセキュリティ研究所 AI セキュリティ研究センター

✉ create-contact@ml.nict.go.jp

🌐 <https://create.nict.go.jp/>

📍 日本橋オフィス

〒103-0027

東京都中央区日本橋 2-7-1 東京日本橋タワー 15 階

📍 北米オフィス (NICT 北米連携センター)

NICT North-America Center

1020 19th Street NW Suite 880, Washington DC 20036