人工知能（AI）が浸透するデータ駆動型の経済社会に必要なAIセキュリティ技術の確立

プログラム・オフィサー（PO）

プログラム・オフィサー
松本　勉
産業技術総合研究所 フェロー

人工知能（AI）があらゆる領域で加速度的に活用されつつあるなか、その安全性・信頼性を脅かす負の側面への対応が大きな課題となっています。本構想では、AIが攻撃され、AIシステムの機能や性能が損なわれる情報システム的側面と、AIが悪用され、AIを利用する人や社会の活動が損なわれる外部作用的側面に着目し、これらAIのセキュリティリスクに対応する技術の研究開発、およびAIセキュリティの知識・技術の体系的整理に取り組んでまいります。近年、強力な生成AIが明示的になったことに代表されるように、AIセキュリティ分野は状況の変化が極めて激しいという特徴があります。このような特徴にも留意しつつ、AIセキュリティという重要な研究テーマに関して、研究実施者の皆さまと共に、Kプログラムへのご期待に添えるよう研究を進めてまいります。

副プログラム・オフィサー
高橋　克巳
NTT株式会社 社会情報研究所 主席研究員

研究開発構想概要

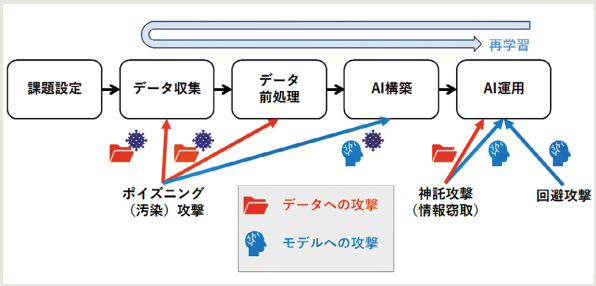
❶ Security for AI

AIを守るための機密性・完全性・可用性の確保や、AIが攻撃された際の社会的影響への対応に関する研究開発の方向性を整理し、AIが活用された具体的なシステムを対象として、防御技術のプロトタイプの開発・実証を目指す。

❷ AI for Security

具体的なシステムを対象として、最先端の攻撃技術に対する革新的なAI活用によるセキュリティ技術のプロトタイプの開発・実証を行うほか、仮想システムにおいて攻撃・防御を行う模擬対戦による技術の高度化と人材育成、コミュニティ拡大を目指す。

AIモデルへの脅威とAIライフサイクルの関係（イメージ）



出典：（独）情報処理推進機構、セキュリティ関係者のためのAIハンドブック（2022年８月）

※本研究開発構想では、公募に先立ち実施した調査研究の結果も踏まえ、以下のAIによってもたらされる負の影響に対応する研究開発課題を募集した。
(A) AI が攻撃され、AI システムの機能や性能などが損なわれる。（AIの情報システム的側面）
(B) AI が悪用され、AI を利用する人や社会の活動が損なわれる。（AIの外部作用的側面）

支援対象となる技術　▶AIセキュリティに係る知識・技術体系

予算額　最大25億円程度

研究開発構想の詳細はこちらから
https://www8.cao.go.jp/cstp/anzen_anshin/20221021_mext_3.pdf

■ 分科会委員（アドバイザー）

相澤 彰子　国立情報学研究所 コンテンツ科学研究系 教授
上田 修功　理化学研究所 革新知能統合研究センター 副センター長
大岩 寛　産業技術総合研究所 インテリジェントプラットフォーム研究部門 副研究部門長
鹿島 久嗣　京都大学 大学院情報学研究科 教授
後藤 厚宏　情報セキュリティ大学院大学 情報セキュリティ研究科 教授
櫻井 幸一　九州大学 大学院システム情報科学研究院 教授
佐藤 一郎　国立情報学研究所 情報社会相関研究系 教授
高江洲 勲　三井物産セキュアディレクション株式会社 プロダクト＆ソリューション事業部 シニアプロフェッショナル

■ 研究開発課題

公募枠 （1）一般研究開発

グラント番号 JPMJKP24C1

AIハードウェアセキュリティ基盤技術の開発

研究代表者

本間 尚文

東北大学 電気通信研究所 教授

グラント番号 JPMJKP25C5

高機密・頑健性スモールデータ連合学習の確立と医療AIでの実証

研究代表者

鈴木 賢治

東京科学大学 総合研究院 教授

グラント番号 JPMJKP25C6

異常トラフィック検知のためのセキュアな連合学習基盤の研究開発

研究代表者

塩本 公平

東京都市大学 情報工学部 教授

公募枠 （2）データ基盤構築支援型研究開発

グラント番号 JPMJKP24C2

SYNTHETIQ X：フェイク情報拡散の防御と予防を実現する研究基盤

研究代表者

越前 功

国立情報学研究所 情報社会相関研究系 教授

グラント番号 JPMJKP24C3

大規模言語モデルのミスアライメントに対するレッドチーミング基盤

研究代表者

佐久間 淳

東京科学大学 情報理工学院 教授

公募枠 （3）知識・技術の体系化研究

グラント番号 JPMJKP24C4

安心安全なAI利活用の為の知識・技術の体系化と知識集約環境構築

研究代表者

披田野 清良

株式会社KDDI総合研究所 セキュリティ部門 グループリーダー