実践的なサイバーセキュリティの研究分野　研究構想の具体例

（注）作業途上のサンプルであり取扱注意

人工知能セキュリティ

目標　人工知能（機械学習）が浸透する社会において機械学習とセキュリティに関する基盤技術の確立を目指す。

概要　機械学習の重要性の高まりを受け、機械学習に立脚したシステムのセキュリティ、および機械学習を活用したセキュリティに係る研究分野を開拓し、理論から応用に至る包括的な研究により基盤技術の確立を目指す。機械学習に対する情報セキュリティの重要3要素(機密性、完全性、可用性（CIA)) をいかに確立するかを根源的な狙いの1つとする。

1. 研究内容と背景

- 将来の機械学習技術を高度に適用する社会においては、機械学習を活用したシステム全般に、より高度なレベルのセキュリティと信頼性が求められる。それらの確立を狙いとした研究が意欲的に研究され始めているが、包括的なフレームワークの確立には至っていない。
- 機械学習を応用したシステムを対象とし、機械学習に対する機密性 (Confidentiality)、完全性 (Integrity)、可用性 (availability) を確立することを狙いとし、理論から応用に至る包括的な基盤技術に関する研究を行う。
- 機械学習を従来のセキュリティ対策技術に高度に応用した飛躍的な技術の開発研究、およびいわゆるオフェンシブセキュリティ（攻撃者視点の研究）のアプローチにより、攻撃者による機械学習の悪用がもたらす脅威と対策技術に関する基盤的研究を行う。

2. 具体的な研究例

機械学習とセキュリティの研究者のコラボレーションによる研究実施が期待され、機械学習のCIA 確立は基礎研究を中心とし、機械学習のセキュリティ技術への応用は、応用研究を中心とする。

A　機械学習のCIA確立

＜機密性＞

機械学習が扱うデータ、および訓練済みのモデルを保護することを狙いとした研究

例・ モデル抽出攻撃 (model extraction)と対策 （理論、応用）

・ データ再構築攻撃 (model inversion)と対策 （理論、応用）

・ プライバシー保護データマイニング技術（理論、応用） 基礎研究として差分プライバシーを含む秘密計算の機械学習への適用（理論、応用） 基礎研究として秘密計算関連研究全般を含む

＜完全性＞

機械学習を応用したシステムに対する悪意がある入力に対する保護を狙いとした研究

例・ adversarial machine learning の生成および検出方法に関する研究 （理論、応用）

・ adversarial example (AE) の生成および検出方法に関する研究（理論、応用）

＜可用性＞

MLaaS（machine learning as a service）のような機械学習アルゴリズムの出力を提供するクリティカルなシステムにおいて、不正なクエリの発行やデータ汚染が行われた際もシステムが影響を受けずに利用可能な状態を維持する技術の研究

例・ 機械学習アルゴリズムに対する不正な入力パターンの発見、検出技術（応用）

B　機械学習のセキュリティ技術への応用

＜防御技術＞

本テーマは機械学習を高度に応用し、一般的なセキュリティ対策技術の飛躍的な性能向上を図る狙いとした研究である。スパムフィルタやマルウェア検出においてMLの適用が進むが、ML技術の進展に応じ、更なる発展の見込みがある。また、攻撃技術として機械学習が使われた場合、いかに機械学習で対抗できるかに関する基礎検討を進める。

＜攻撃技術＞

攻撃者が高度に機械学習を利用することで生じる脅威に関する研究

例・ 機械学習を用い、実際には存在しない画像、動画、音、テキストを巧妙に生成する技術と対策方法の研究

・ 本来秘匿されるべきデータを機械学習を適用することにより、高精度で推定する技術と対策方法の研究

3. 想定する研究の進め方（PI人数規模のイメージなど）

AはCIA確立の元となる基礎研究を含め理論を中心に進める。B は応用を中心に進める。

A PI 5名 （理論4、応用1） B PI 5名（理論1、応用4） 程度あるいはそれ以上