資料2-2

AIセキュリティに関する研究動向について

令和2年11月25日
サイバーセキュリティ戦略本部　研究開発戦略専門調査会

内閣サイバーセキュリティセンター（NISC）
基本戦略第1グループ

AIセキュリティに関する研究動向

事務局注: ここでのAIは機械学習のことを指している。

a）AIを活用したサイバーセキュリティ対策（AI for Security）

現状の取組・動向

- ✓ あるルーチンの仕事の自動化や、人的に行っている監視、分析、対応の支援を行うことにより、**AI技術がサイバーセキュリティを強化すると期待**される。※1
- ✓ サイバーセキュリティの脅威を特定し対応するための**鍵となるツール**としてAIを使用する機会が増えよう。※2
- ✓ AIを利用したセキュリティ製品やサービスは**既に商用化が進んでいる**。【有識者ヒアリング結果】

今後の取組・動向

- ✓ （研究開発目標として）マルウェア及び侵入の検知等以外に、**新しいAIベースの技術**を開拓・研究する。セキュリティ能力の**AIによる自動的な管理**を開発する。※1
- ✓ （研究開発目標として）AIを活用したセキュリティシステムやAIベースのセキュリティ制御の**セキュリティや信頼を評価するためのモデル、定義、評価手法**を開発する。※1
- ✓ 人間自身が脆弱性になりうるため、AIを用いて、問題となる人間の行動を検知できる技術の研究が重要ではないか。【有識者ヒアリング結果】
- ✓ 将来、様々なシステムにAI機能が組み込まれるため、AI for SecurityとSecurity for AIはサイバーセキュリティ分野では最終的に同一視されるようになるのではないか。【有識者ヒアリング結果】

b）AIを使ったサイバー攻撃（Attack using AI）

現状の取組・動向

- ✓ AIシステムは**人間の能力や現在の技術的能力を超える速度と規模で動作**する。**AIのサイバー攻撃への悪用が懸念**され、**AIがサイバー防御にも同様に使用**されない限り、ますます非対称な戦いになる。※1
- ✓ **サイバー防御を担うAIシステム**が、適切な制御ができるよう実装されていなければ、**サイバー攻撃に悪用され得る**。※1

今後の取組・動向

- ✓ **攻撃の視点から知見を得ることにより、先手を打ってセキュリティ対策を高度化する**プロアクティブな研究が、サイバー防御を担うAIシステムにおいても重要と考えられる。【有識者ヒアリング結果】

AIとセキュリティに関する観点には、概して、これらの3観点以外に「AI自身による自律的な攻撃（Attack by AI）」があると考えられる。しかしながら、AIの指す内容が異なり得るとともに、現時点では現実的ではないと考えられるため、ここでは注釈に留める。

c）AIそのものを守るセキュリティ（Security for AI）

現状の取組・動向

- ✓ 多くの機械学習アルゴリズムは、**ライフサイクルを通じて攻撃を受ける可能性**がある。その**脆弱性がどんなものかまだ十分に理解されていない**。※1
- ✓ （説明可能で堅牢で安全なAIのための研究が望まれるが）セキュリティの観点として、権限のない者による意図的または意図的でない改ざんをどう防止するか。また、敵対的機械学習は研究が必要な更なる領域である。※3
- ✓ 2018年くらいまでは敵対的サンプル（AE）が研究としてホットトピックであったが、**最近はAEに対する防御研究が多く**なっている。【有識者ヒアリング結果】
- ✓ AEに対する防御研究には、敵対的学習*ⅰと、Certified Defenses*ⅱがある。また、一画像だけで結果を判断するのではなく、様々な情報を基に結果を判断することも対策になり得る。【有識者ヒアリング結果】

*ⅰ：AEを学習データとして用いることにより、機械学習モデルのAEに対する耐性を上げる研究
*ⅱ：「保証された防御」とも呼ばれ、AEを多少変化させるだけでは誤認識されないように、機械学習モデルが作成されることを保証する防御手法に関する研究

今後の取組・動向

- ✓ （研究開発目標として）機械学習システムに対する**攻撃と防御を理解するためのツールと技術を開発**する。機械学習アルゴリズムのセキュリティと堅牢性を検証するフォーマルメソッド技術を向上する。※1
- ✓ AIに関する研究構想として、機械学習システム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）の確立を目的とする例が考えられる。例えば、機密性に関しては、Model Extraction攻撃*ⅲやModel Inversion攻撃*ⅳに関連し、完全性に関しては、AEに関連する。【有識者ヒアリング結果】

*ⅲ：「モデル抽出攻撃」とも呼ばれ、攻撃対象の機械学習モデルから取得した入出力値を基に、攻撃対象と同等の偽の機械学習モデルを抽出する攻撃
*ⅳ：「モデル逆推定攻撃」とも呼ばれ、機械学習モデルに対して、出力データから、学習データに使用した画像などの具体的な入力データを逆推定する攻撃

【出典】（翻訳と強調は事務局にて付記したもの）

- ※1: 米国 FEDERAL CYBERSECURITY RESEARCH AND DEVELOPMENT STRATEGIC PLAN（連邦サイバーセキュリティ研究開発戦略計画）［2019年12月 米国国家科学技術評議会（NSTC）］を参考に記載。
- ※2: 英国 Interim Cyber Security Science & Technology Strategy（暫定サイバーセキュリティ科学技術戦略）［2017年11月 英国政府内閣府］を参考に記載。
- ※3: 欧州 Analysis of the European R&D priorities in cybersecurity（サイバーセキュリティにおける欧州の研究開発優先事項）［2018年12月 欧州ネットワーク情報セキュリティ庁（ENISA）］を参考に記載。