日立ソリューションズ・データ科学センター共催セミナー
AI とセキュリティの交差点
～技術革新の中の安全確保～

1. 飛躍的な進化を遂げる AI 技術 セキュリティの確保は重要な課題に

2024 年 3 月 6 日、早稲田キャンパスにおいて、株式会社日立ソリューションズ・データ科学センター共催のセミナー「AI とセキュリティの交差点 ～技術革新の中の安全確保～」を開催した。日立ソリューションズと早稲田大学は 2020 年 2 月に学術交流協定を締結。同社の寄附により、グローバルエデュケーションセンター設置の提携講座「未来社会を創るセキュリティ最前線」を実施した。

本セミナーには、日立ソリューションズからセキュリティエバンジェリストの扇健一氏とホワイトハッカーの米光一也氏、データ科学センターからは小林学教授と中原悠太講師が登壇。セミナーの前半は小林教授と中原講師による AI の技術動向やセキュリティリスクの事例紹介があり、後半はその内容を受けて 4 名によるパネルディスカッションが行われた。

開会に際し、司会を務めるデータ科学センターの須子統太教務主任から本セミナーの趣旨説明があり、「データ科学とセキュリティは切っても切り離せない関係。データをビジネスの中で活用し、実装していく中で、データ科学とセキュリティは両輪のような役割になっていく」と今回のテーマの重要性が述べられた。また、扇氏は日立ソリューションズの事業紹介とともに、巧妙化するサイバー攻撃に対して早期回復を目指す「サイバーレジリエンス」が重要になっていると現状を説明した。

2. ニューラルネットワークを誰でも手軽に活用できる時代 そこに潜むさまざまな攻撃リスクも顕在化

AI の技術動向とセキュリティリスクに関するセッションは、中原講師がニューラルネットワークや Generative Adversarial Network（GAN）について解説を行い、それぞれに対して小林教授が関連するサイバー攻撃のリスクや事例について紹介する形で行われた。また、会場では実際にセッションの内容を体験できるよう、QR コードを通じてプログラムファイルを配布した。

ニューラルネットワークについて、「平たくいえば ChatGPT などの基礎になっている技術」と中原講師は紹介。特に画像処理の分野における有用性に触れ、Google Vision AI のデモンストレーションを行った。画像に写っているものの判別や認識がニューラルネットワークの得意とする領域であり、医療現場でのガン発見や車の自動運転など、その可能性を説明した。

配布した Python プログラムのファイルをもとに、中原講師は「ニューラルネットワークでは大量の画像の学習、ディープラーニングが重要な技術」と話を進める。求める出力や結果を得るためには、入力に対するパラメータ（関数）の設定が不可欠であり、良いニュートラルネットワークをつくるためには、パラメータを少しずつ動かしながらチューニング（学習）していくことが必要、と分かりやすく説明を行った。

中原講師のニューラルネットワークに関する説明を受け、小林教授は「敵対的サンプル」を取り上げ、そのリスクを参加者に提示。入力となる画像への攻撃を行うことで、人間の目には一切変化がなくても AI による出力が変わる、つまり人間が感知できないところで攻撃が行われるリスクがあるとした。「敵対的パッチ」という手法もあり、特定の画像を持ったり、T シャツにプリントして着用したりすることで、AI から認識されなくなるといった悪用法も紹介した。

続いて中原講師は「GAN」の技術について説明。GAN はニューラルネットワークを活用し、画像など人工的にデータを生成するための技術であり、現実と見分けがつかないほどの画像が、現在生成可能になっている。GAN もニューラルネットワークの一種で、入力に対してあるパラメータを通して出力を返す仕組みである。GAN における入力はノイズおよび実際の画像であり、入力されたノイズを画像として人工的に生成するネットワークと、それを実際の画像か人工の画像かを判別するネット

ワークからなる。この双方のネットワークを組み合わせて繰り返し学習することで、パラメータを調整・学習していく、と中原講師は紹介した。

小林教授は、そうしたGANの技術に対して、人間の顔を別人と誤認させる「敵対的眼鏡」の存在を例示。さらに、その技術を発展させ、ウイルス検知を通過してしまう「敵対的マルウェア」もすでに現れていることを紹介した。「攻撃方法はすでに色々と出てきており、対策はイタチごっこになってしまっている」と現状について述べ、前半のセッションをまとめた。

3. AI技術を悪用した攻撃を防ぐためには、研究開発側とセキュリティ側の協力が重要

後半のパネルディスカッションでは、小林教授がモデラー、扇氏、米光氏、中原講師がパネリストとして登壇し、AIとセキュリティに関する議論を行った。

小林教授はまず、AIへの攻撃に関する問題提起を行った。攻撃には、学習データを歪ませる「データ汚染」、ニューラルネットワークのモデルを変え、出力を誤らせる「モデル汚染」、そして非公開のニューラルネットワークを模倣し取得する「モデル搾取」があり、すでに技術が公開されていると紹介。「Hugging Faceをはじめとしたオープンなプラットフォームで改ざん・攻撃されたニューラルネットワークが公開された場合、利用者がそれに気づくことは難しい」と警鐘を鳴らした。

米光氏はそうした攻撃の事例に対して、危機感を感じた部分と安心した部分があると応答。AIの技術の進歩には驚いたものの、AIに対する攻撃は、現状では針の穴に糸を通すものであり、システム全体で対処していくセキュリティの考えの中で十分対応できる余地があるとした。

ホワイトハッカーの扇氏は、現在も起こっているDDoSアタックを例に挙げ、「どれが攻撃でどれがそうではないか、その境界を我々も常にチューニングしている。AIでもそうした積み重ねが必要になるだろう」と述べ、セキュリティ側だけではなく、AIの開発者によるセキュリティデザインも必要との見方を示した。

データ科学の研究者である中原講師は、見た目にはほとんど変わらないのにAIが別のものだと認識してしまうことが問題であり、その本質はわずかな入力の変化に対してニューラルネットワークの出力を大きく変えてしまう過学習（オーバーフィッティング）の問題であると言及。また、目的に合わせたAI技術を利用するべきであり、それを判断できるリテラシーを持った人材の育成が必要と提言した。

質疑応答のパートでは、会場から「音声分野での悪用」や「システム運用全体で見た際のセキュリテイ」について質問が寄せられ、扇氏からは、「現在も金融や防衛といった重要なシステムはマルウェア検知製品が複数使われている。AI はフリーで提供されているものも多くあり、複数組み込むことで比較的安価にセキュリティを担保できるのでは」と指摘があった。

4. 人工知能の悪用による攻撃はより身近に
セキュリティの担保が社会実装の決め手になっていく

パネルディスカッションの 2 つ目のテーマは、「AI を悪用した攻撃」。小林教授からは AI の悪用による「標的型攻撃」と「情報抽出」の解説が行われた。

米光氏はこうしたAIを使用した攻撃の増加に対し、「便利なツールができたら、皆それを使う。それを活用するか、悪用するかは人間側の問題。総じて抽象的な判断にならざるを得ず、イタチごっこもある程度はやむをえないのではないか」と私見を述べた。

扇氏は、サプライチェーンの脆弱な部分への攻撃リスクを指摘。従来、高度なハッカーでしか見つけられなかった脆弱性を誰でも突けるようになると、サイバーリスクのさらなる深刻化につながることを憂慮している、と述べた。

一方で中原講師は、AI 研究の自由度という観点から、最初から様々な悪用のリスクを全て考慮した上で研究を行うというのはなかなか難しいと説明。現状では人間の詐欺と同じく、リスクを把握したうえで行動することが必要なのではと語った。

ディスカッションの最後のテーマとして、小林教授から、世間でも注目されているディープフェイクの話題が提起された。音声や動画を加工し、別人になりすますことができるディープフェイクでは、すでに詐欺被害も発生している。

米光氏はこうした状況に対して「かなり深刻な状態。テクノロジーを使って人間の心理を攻撃している。攻撃者にとって、AI のコストパフォーマンスが上がってきていて、悪用のハードルが下がっている」と述べ、今後のさらなるリスク増大に危機感を表した。

パネルディスカッションの終わりにあたり、小林教授からは、「AI についてもセキュリティについて考える必要性が生まれてきている。進化のスピードが非常に早いので、人材育成も含めて、セキュリティ側と AI（データ科学）側がタッグを組んで対策を進めていかなければならない」と総括があり、本セミナーは盛況のうちに幕を閉じた。