報道関係者各位

エフセキュア、AI 推薦の悪用の危険性について警告

～攻撃者による AI の操作や意図的な偽情報の拡散について実験～

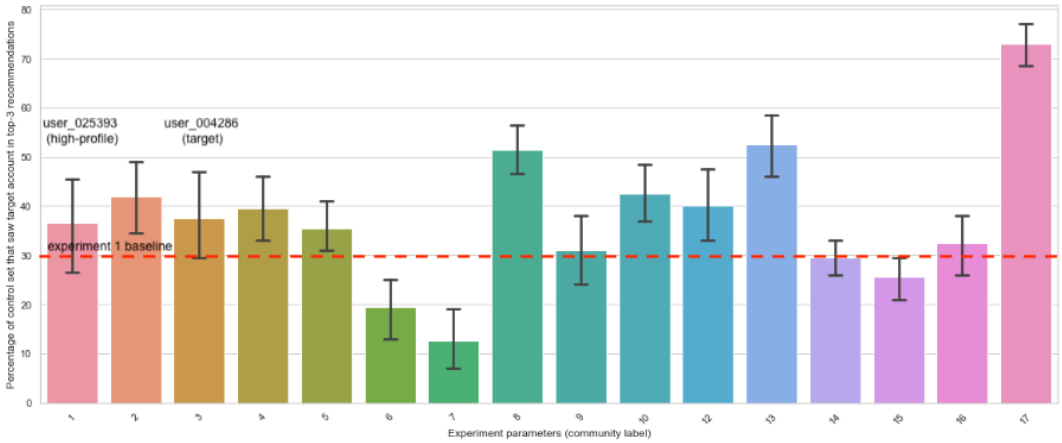
2021 年 6 月 25 日
エフセキュア株式会社

先進的サイバー・セキュリティ・テクノロジーのプロバイダである F-Secure (本社: フィンランド・ヘルシンキ、CEO: Juhani Hintikka、日本法人: 東京都港区、以下、エフセキュア) は、ソーシャルメディアにおける AI 推薦に関する実験を行い、その結果、敵対する人／企業／政府を陥れるための意図的な偽情報や陰謀論の拡散が操作可能なものであったことを発表しました。

AI による推薦システムは、検索エンジン、オンラインショッピングサイト、ストリーミングサービス、ソーシャルメディアなど、今日私たちが享受している多くのオンラインサービスで使用されています。しかし、人々がインターネット上で見るものやすることに対する影響力が大きくなっていることから、意図的な偽情報の拡散や陰謀論の促進に積極的に利用されるなど、様々なタイプの悪用に対する懸念が生じています。

エフセキュアの人工知能研究センターのシニア・リサーチャーである Andy Patel (アンディ・パテル) は最近、単純な操作技術がソーシャルメディア上の AI 推薦にどのような影響を与えるかを知るために、一連の実験を行いました。「Twitter をはじめとする SNS は、多くの人々やグループが様々な主張を押し付ける戦場となっています。その中には、有機的な会話や広告だけでなく、正当な情報に対する信頼を損ない、低下させることを目的としたメッセージも含まれています。これらの『工作員』がどのように AI を操作できるかをリサーチすることで、AI が現実的にできることの限界を明らかにし、理想的にはどのように改善できるかを明らかにすることができます。」

PEW リサーチセンターが 2020 年末に実施した調査[*1] によると、アメリカ人の 53%がソーシャルメディアからニュースを入手していることがわかりました。18 歳から 29 歳の回答者は、最も頻繁にニュースを入手する情報源としてソーシャルメディアを挙げています。一方、ソーシャルメディアを情報源とすることに潜在的なリスクがあることも調査で明らかになっています。2018 年に行われた調査[*2] では、虚偽の内容を含む Twitter の投稿がリツイートされる確率は通常の投稿より 70%も高いことが判明しています。



(ターゲット・アカウントがトップ 3 の推薦に表示されたコントロール・セットの割合)

Patel は、Twitter からデータを収集し、協調フィルタリングモデル (機械学習の一種で、過去のやりとりに基づいてユーザーとコンテンツの類似性を符号化する) を学習させて、推薦システムに利用することを試みました。その後、特定のアカウント間でリツイートが行われたデータ (ポイズンデータ) を用いてモデルを再学習させ、推奨度がどのように変化するかを調べる実験を行いました。

リツイートするアカウントを適切に選択し、リツイートを行うアカウントの数と公開するリツイートの数を変化させることで、ごく少数のリツイートであっても、注入されたリツイートによってコンテンツが共有されたアカウントを推奨するように推薦システムを操作することができました。

今回の実験は、ソーシャルメディアなどのウェブサイトがユーザーにおすすめ情報を提供する際に採用するであろう AI の仕組みを簡略化して行ったものですが、Patel は、Twitter をはじめとする多くの大手サービスが、現実世界で既にこうした攻撃に対処していると考えています。
「私たちは、実際の攻撃がどのように行われるかを知るために、簡略化したモデルに対してテストを実施しました。ソーシャルメディアのプラットフォームは、今回の研究で実証されたものと似たような攻撃にすでに直面しているものと思いますが、これらの運営企業は結果だけを見て、それがどのように機能するかにはさほど注意を払っていないため、こうしたことが現実に起こっていると信じるのは難しいのではないでしょうか。」


(左: Andy Patel　右: Matti Aksela)

エフセキュアの AI 担当バイスプレジデントである Matti Aksela (マッティ・アクセラ) は、AI のセキュリティに関する潜在的な課題を認識し、対処することが重要だと語っています。
「今後、AI への依存度が高まっていく中で、潜在的な悪用から AI を守るために何をすべきかを理解する必要があります。私たちが依存しているサービスの多くを AI や機械学習が担うようになると、その結果を信頼できるものとするために、得られる利益に加えて、そのセキュリティ上の強みと弱みを理解する必要が出てきます。セキュアな AI は、信頼できる AI の基盤なのです。」

Patel は、実験の詳細なレポートと、研究を再現するために必要なコードやデータセットを GitHub で公開しています。
https://github.com/r0zetta/collaborative_filtering
また、本リサーチの詳細は以下のエフセキュアブログページでご覧いただけます。
https://blog.f-secure.com/ja/ai-recommendations-manipulations

*1 https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/
*2 https://science.sciencemag.org/content/359/6380/1146

エフセキュアプレスページ:
https://www.f-secure.com/jp-ja/press

エフセキュアについて

エフセキュアほど現実世界のサイバー脅威についての知見を持つ企業は市場に存在しません。数百名にのぼる業界で最も優れたセキュリティコンサルタント、何百万台ものデバイスに搭載された数多くの受賞歴を誇るソフトウェア、進化し続ける革新的なセキュリティ対策に関する AI テクノロジー、そして「検知と対応」。これらの橋渡しをするのがエフセキュアです。当社は、大手銀行機関、航空会社、そして世界中の多くのエンタープライズから、「世界で最も強力な脅威に打ち勝つ」という私たちのコミットメントに対する信頼を勝ち取っています。グローバルなトップクラスのチャネルパートナー、200 社以上のサービスプロバイダーにより構成されるネットワークと共にエンタープライズクラスのサイバーセキュリティを提供すること、それがエフセキュアの使命です。
エフセキュアは本社をフィンランド・ヘルシンキに、日本法人であるエフセキュア株式会社を東京都港区に置いています。また、NASDAQ ヘルシンキに上場しています。詳細は https://www.f-secure.com/en/welcome (英語) および https://www.f-secure.com/ja_JP/ (日本語) をご覧ください。また、Twitter @FSECUREBLOG でも情報の配信をおこなっています。

※　以下、メディア関係者限定の特記情報です。個人の SNS 等での情報公開はご遠慮ください。

【本件に関する報道関係者からのお問合せ先】

エフセキュア株式会社
広報部　　秦　和哉
TEL：03-4578-7745 (直通)
japan-pr@f-secure.com