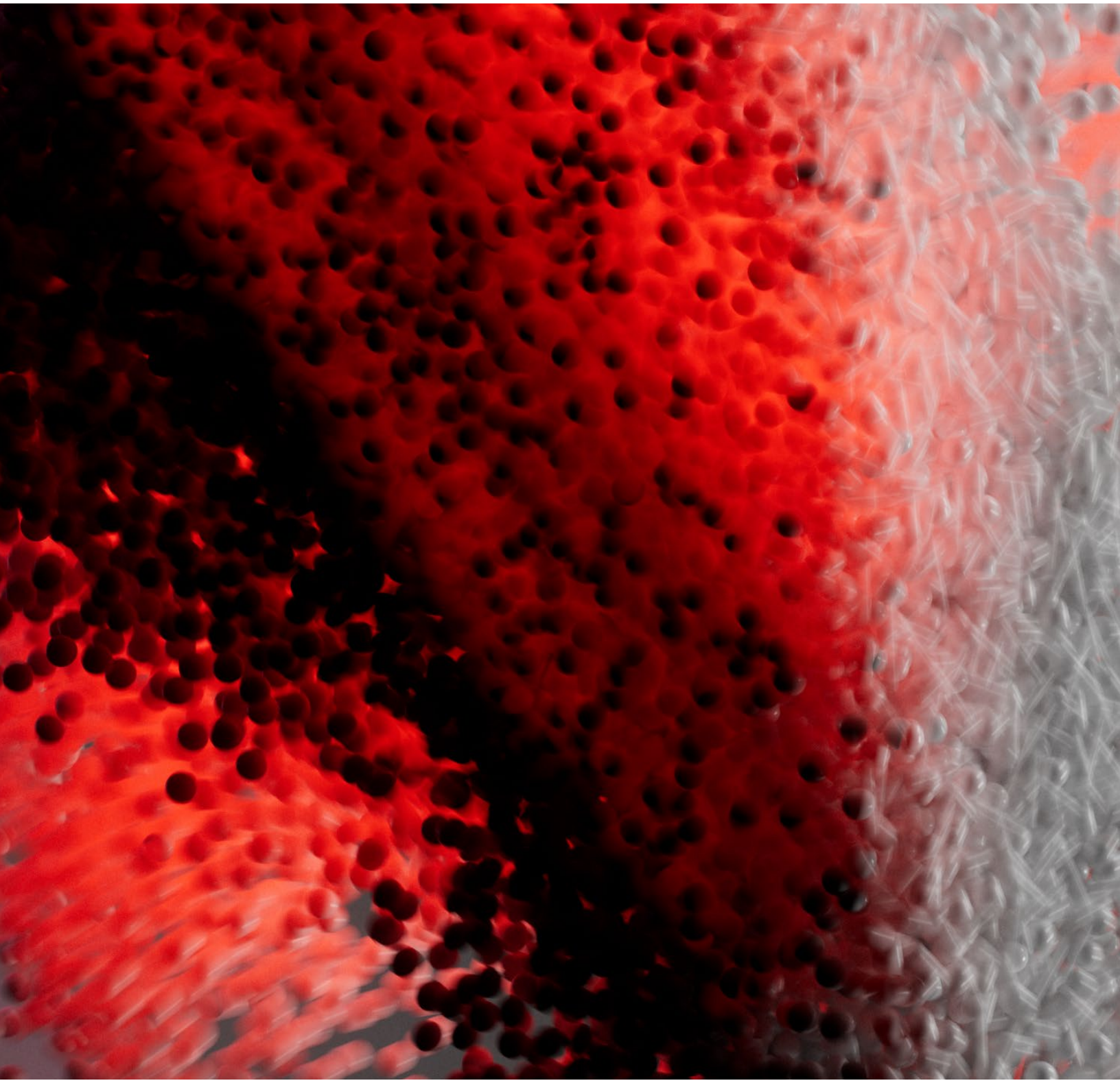
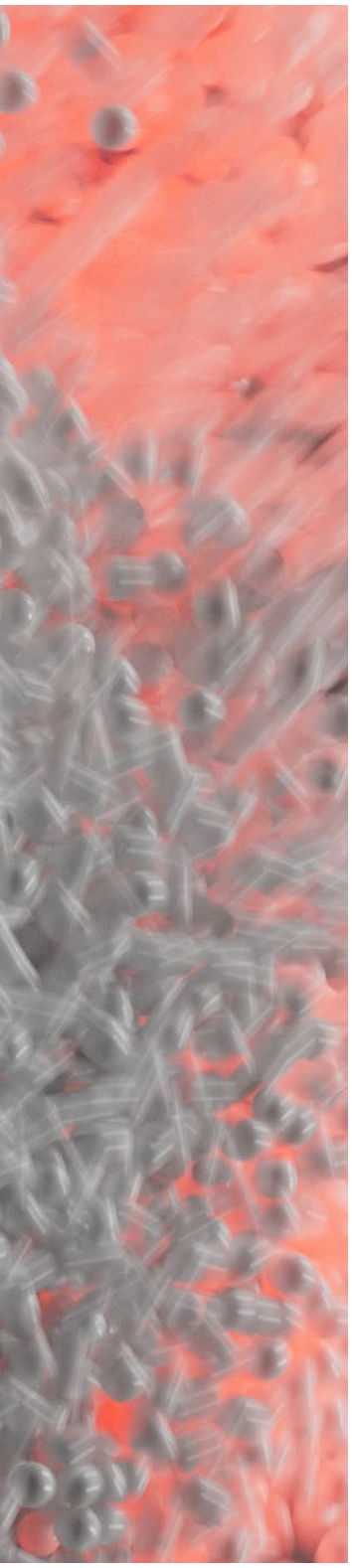




サイバーとトラスト

現実を守る：合成メディアの時代における真実

AIツールの普及により、なりすましや偽りの情報を作成することがかつてないほどに容易になった。これらの脅威に対して、先進的な企業・組織はさまざまなポリシーとテクノロジーを駆使して対応している。

ここ最近Tom Hanksが歯科治療プランを宣伝する広告動画を見たことがある人は少なくないだろう。この動画に映っているのは彼自身ではなく、誰かがディープフェイク技術を使用して、あたかも彼が宣伝しているかのように見せかけたなりすましであった[1]。

誰もが、あたかも自分が発言した、あるいは行動したと見せかけることができる時代が到来したと言える。生成AIの急速な進歩によって、何が本物で何が人工的に生成されたコンテンツなのかを区別することがほぼ不可能なレベルにまで達している。

被害に遭うのは有名人だけではない。AIツールの普及により、悪意のある人物が他人になりすまし、標的を欺くことがかつてないほど容易になっている。音声認識・顔認識のアクセス制御を回避するためや、フィッシング攻撃にディープフェイクを使用するケースも多い。膨大なデータを必要とするAIアプリケーションそのものが、攻撃者にとって格好の標的となっている。新しいコンテンツ生成ツールがインターネットに登場するたびに、セキュリティリスクが増大している。

これらのセキュリティリスクに対して、先進的な企業・組織は、ポリシーとテクノロジーを組み合わせながら、有害なコンテンツを特定する方法や従業員にリスクを認識させるための仕組みを導入することで対応している。また、攻撃者が悪用する生成AIと同じツールを用いて攻撃の特定や予測をすることで、攻撃に先んじて対処を施すことができる。

Now：次世代のソーシャルエンジニアリングハック

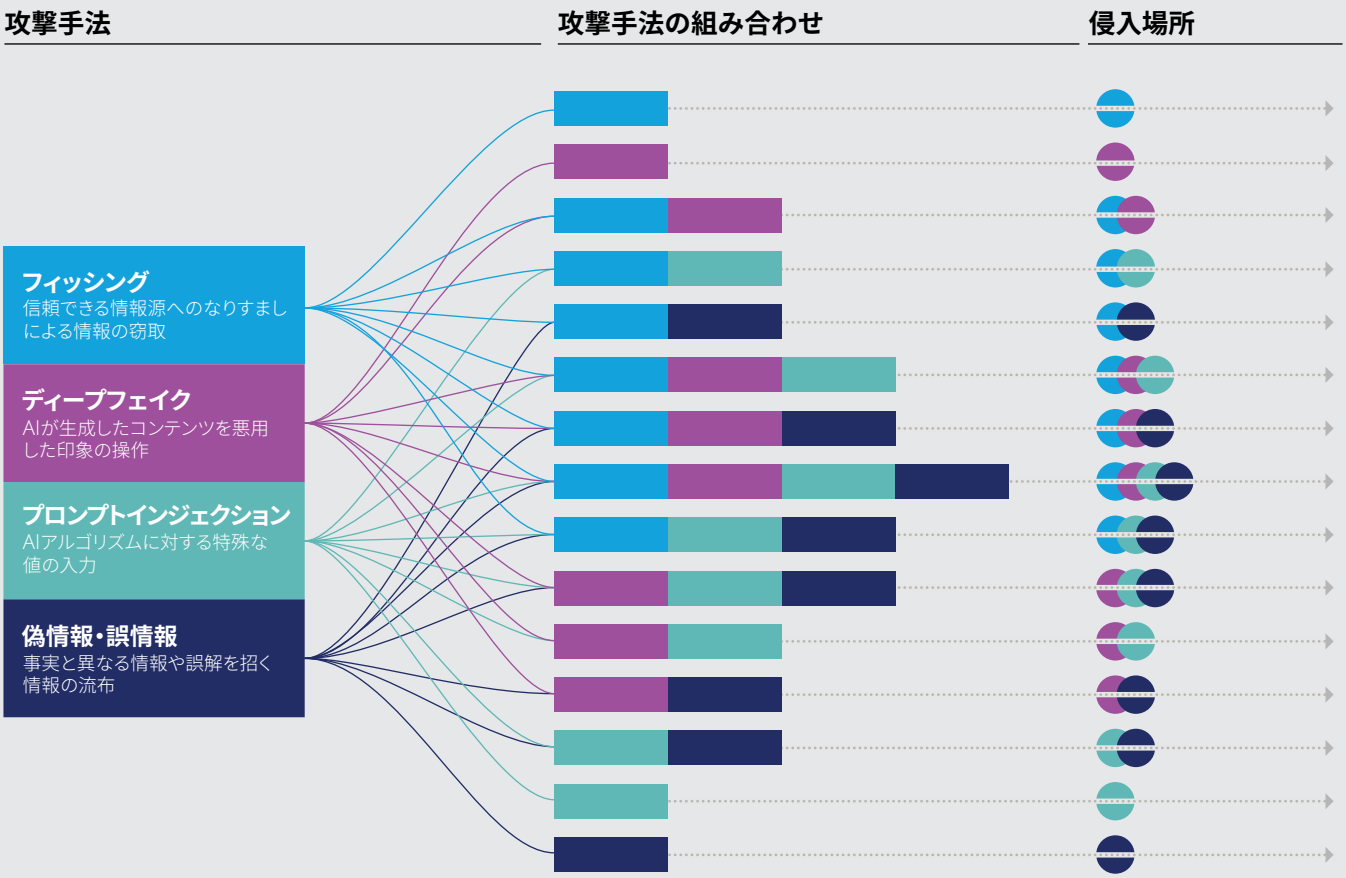
ソーシャルエンジニアリングによるハッキングは、不正な目的のためにデータを要求することや、システムにアクセスするよう仕向けることで成立してきた。攻撃者にとってこの方法は目的を達成する手段として非常に効果的であるものの、攻撃者と被害者との間で多くのやり取りが必要になる。一方で、AIで生成されたコンテンツを利用することで、攻撃者は標的とのコミュニケーションに掛ける時間を短縮できる。AIで生成され、信頼できる情報源になりすました大量のコンテンツが企業に襲い掛かっている[2]。

現時点において、リアルなコンテンツを生成するAIの能力と、それをAIが生成したコンテンツであると人間が識別する能力とには大きな隔たりがある。ある調査では、対象者の約80%がAIと人間が生成したコンテンツを識別できると回答し、20%は自信を持って識別できないと回答しているが、前者は自信過剰に陥っている可能性が高い[3]。なぜならば、近年の生成AIは、人間が作成するコンテンツに限りなく近づけるよう高度なチューニングを繰り返しながら開発されており、人間によるコンテンツと区別できない水準にまで忠実な再現が可能となりつつあるためである[4]。AIが生成したコンテンツは機械的な不自然さがあるから識別できると思うかもしれないが、AI技術の発達に伴い、人間が作成したものと区別できないレベルにまで既に到達している。

攻撃者は、AIが生成したコンテンツを使用して、さまざまな方法でサイバー攻撃を試みる（図1）。AIを利用したサイバー攻撃について具体的に見ていく。

図1

攻撃者はAIが生成したコンテンツを使用して様々な方法でサイバー攻撃を試みる



出所：Deloitte analysis.

フィッシング：フィッシングは最も一般的なサイバー攻撃であり、毎日34億通ものスパムメールが送信されている。2021年には、犯罪者はフィッシング攻撃によって推定4,420万米ドルを盗み出している[5]。フィッシング攻撃が成功する要因は、文面の品質の高さではなく、大量のメールが送信されるためである。数十億通のメールのうち、最終的に攻撃の成功に至るのは数通のみである。受信者の大半は、文法やスペルに不自然な点がある、あるいは送信者と文面に思い当たる節がないなどの理由で、受信したメールがフィッシングであると識別する。しかし、生成AIを使用することで、攻撃者は違和感がない自然なメッセージを迅速かつ容易に作成することが可能になる。さらに、メッセージの内容を受信者に合わせて調整することで、受信したメールがフィッシングであることを見分けることがより一層困難になっている。公開されているAIモデルの品質が向上するにつれて、問題の深刻さは増すであろう[6]。

ディープフェイク（Deepfakes）：ディープフェイクは何年も前から存在していたが、ごく最近になりサイバー犯罪に使われるほど高度なコンテンツが生成できるようになってきている。企業を攻撃する用途でディープフェイクが使用された事例も目にするようになった。英国に拠点を置くエネルギー企業のCEOは、ディープフェイクの

AI音声技術で親会社の社長になりすました攻撃者に243,000米ドルを騙し取られたという被害も報じられた[7]。この事件以降、ディープフェイクのツールは大幅に進歩しており、今後も急速に発展する可能性が高く、やり取りしている相手が本人であるのかの確証を持つことが難しくなっている。

プロンプトインジェクション：プロンプトインジェクションは、対話型AIに対して特殊な入力をすることによって、開発者が想定していない挙動を誘発する攻撃手法である。対話型AIに対するプロンプトインジェクションによって、システムが保有する機密情報や公開すべきではない内部データが流出する可能性がある。例えば、攻撃者が連絡先リスト、銀行情報、健康データなどのデータを転送するようにプロンプトに入力することで、これらの情報をシステムが出力してしまう可能性がある[8]。従来ソーシャルエンジニアリングによるハッキングは、相手を騙してデータの送付を要求することなどで成立してきた。一方でプロンプトインジェクションにおいては、攻撃者はわざわざ標的を騙す必要すらなく、被害者が気づかないうちに欲しい情報を持ち出してしまう。

偽情報・誤情報：企業を標的としたソーシャルメディア上での宣伝活動は従来から行われているが、生成AIの発達により更に加速している。AIツールの利用によって短時間でコンテンツが量産できるため、攻撃者が標的とする企業に対し、AIツールを悪用して名誉毀損、さらには株価下落を引き起こすことも考えられる[9]。従来、攻撃者は標的に合わせて個別にメッセージを作成する必要があったが、生成AIの登場により攻撃者は偽情報・誤情報を大規模に量産し、攻撃が成功しやすいメッセージを見つけるまで実験やテストを行うことが可能となった。

生成AIが業務に幅広く利用できることや生成AIの性能向上スピードに鑑みると、前述の問題は今後さらに深刻化すると考えられる。企業から金銭やデータを奪取しうる信憑性の高い情報を、コストや技術的スキルを持ち合わせずとも作成することが可能になるだろう。

New：企業の新たな脅威に対する防衛

これらは、生成AIがもたらすリスクに対して企業に為す術がないことを意味するものではない。大手の企業・組織は、被害を未然に防ぐための積極的な対策を講じている。

ソーシャルエンジニアリングは広く認知されている攻撃手法であるが、ソーシャルエンジニアリングへの対策は、AIを使った新たな攻撃に対しても有効である。オンライン上でのやり取りに警戒すること、連絡相手の身元を確認すること、機密資産へのアクセスに多要素認証を要求することなどは、企業がこれらの新たな攻撃手法から身を守る有効な手段である。

ソーシャルエンジニアリングの脅威と同様に、合成コンテンツの問題への取り組みは認識から始まる。「AIは興味深くて、とても素晴らしいが、攻撃者にとっても多くの強みをもたらしている」と、CarMaxのchief information and technology officerであるShamim Mohammedは言う。「私が重視しているのは、会社を保護・防衛するために常に最新の状態を維持し、さらに先を行くことだ[10]」

その方法の1つは、エコシステムパートナーと協力することだ。Mohammedによると、CarMaxは大手IT企業やAIに特化したサイバーセキュリティのスタートアップ企業と提携し、脅威の状況を把握し、攻撃を防ぐため最新のツールにアクセスできるようにしているという。

「我々には非常に強力なテクノロジーエコシステムがある」と、Mohammedは言う。「我々は、AI革命のトップにいる大手企業や、AIに特化したスタートアップ企業と協力している。この新しい脅威から情報を保護するための最適なツールを有している」

潜在的な有害コンテンツの特定に有用な企業向けツールも登場している。AIがコンテンツを生成できるように、画像や動画、テキストの信憑性を評価することも可能である。これらのツールにより、企業が直面する可能性の高い攻撃パターンを予測できるようになるかもしれない。

AIによるコンテンツの生成とそれらの検出に関しては、利用するデータの規模、多様性、新鮮さが極めて重要になる。初めて一般公開された当時の生成AIモデルは、巨大IT企業のみがアクセスできた強力な計算処理能力と大規模データに基づく学習によって開発された。そのため、巨大IT企業に匹敵する規模のリソースやデータを準備することが困難であった他のIT企業が開発した検出ツールの精度には限界があり、AIで生成したコンテンツを悪用しても検知されにくい状況であった[11]。

しかしそれも変わりつつある。例えばReality Defenderは、合成メディア検出プラットフォームを、ペタバイト規模のテキスト、画像、音声のデータベース（一部は人工的に

生成されたもの）上で学習させている。このような大規模なコーパスで学習を行うと、僅かながらもAIツールによって生成されたことを示す情報が浮かび上がってくる。具体的には、AIによって生成された画像データには、特徴的な変形やピクセレーションが含まれることがある。テキストデータにおいても同様に、AIにより生成されたことを予測できる可能性がある。これらの特徴は肉眼では分からないかもしれないが、十分なデータで訓練されたAIモデルでは安定して検知が可能である。

Reality DefenderのCEOであるBen Colmanは、企業における有害なコンテンツの特定と対応は非常に重要だと述べている。特に、企業や経営陣の名誉棄損を意図した偽情報・誤情報に関しては、その傾向が顕著だという。「大きな注目を集めてからでは遅すぎる」と指摘し、「世論という法廷でブランドが傷つけられた場合、1〜2週間後に内容が事実でないと明らかになったとしても、それは重要ではない（手遅れである）」と述べた[12]。

AIが生成したコンテンツを検出するツールは他にも存在する[13]。合成メディア検出器は近々、より精巧な調整が行われるだろう。先日、Intelはディープフェイク検出ツールを発表した。当該ツールではデータだけでなく、動画に写っている人物の顔の静脈の変化も分析する。心臓が静脈に血液を送ると、静脈の色がわずかに変化するが、この変化をAIモデルが模倣するのは非常に困難である[14]。

今後もこのような取り組みを期待したい。ある推計によると、2025年までにオンラインコンテンツの90%が合成されたものになるという[15]。その多くは、マーケティングやカスタマーエンゲージメントといった目的のコンテンツであるが、サイバー犯罪者は自らの目的のために生成AIツールを利用するだろう。企業にとって、従業員が触れるコンテンツの正当性を確認することは、かつてないほど重要性を増している。

Next：繰り返されるいたちごっこ

数年前から多くの企業が**AIをセキュリティ強化策として活用**してきたが、生成AIの登場により攻撃者は新たな攻撃の手段を手にした[16]。そして企業もまた対抗策を講じ始めている。量子コンピューティングに代表される新たなパラダイムが確立し、AIの性能が現状より進歩したとしても、このいたちごっこは継続しているに違いない。

量子コンピューティングが広範囲に普及するのはまだ数年先のことだが、今日急速に発展しており、攻撃者と企業の双方が次世代のツールとして使用するだろう。この技術の最も前途有望なユースケースは、量子機械学習である。従来のツールと同様、重要なのは使い道である。量子コンピューティングは人工生成コンテンツの問題に拍車をかけるリスクを孕む一方で、企業・組織のセキュリティ対策に恩恵をもたらす可能性もある。

量子機械学習は、従来よりも少ない学習データで高精度な予測モデルを生成できる可能性を秘めている[17]。古典コンピューティングでは、データは0と1のバイナリーとして表現される。しかし、量子データは一度に複数の状態を取ることができるため、古典データよりも豊富な情報を有する。この豊富な情報表現と従来の機械学習を組み合わせることで、現在の最先端GPUで学習するモデルよりも複雑なモデル構築が可能となる[18]。

このように構築されたモデルを用いることで、攻撃者は標的に関する大量のデータを収集することなく、より対象に照準を絞ったコンテンツを作成することができる。量子機械学習では、信憑性のあるフェイク画像を生成するのに、数百時間分の学習用動画データではなく、ごく少量の断片的な動画で十分になるだろう。

一方で、セキュリティ態勢の改善を試みる企業にとっては、量子機械学習によって合成メディアを検出するモデルの性能向上が期待される。古典的な機械学習のように、人工生成メディアに関する数十億のデータを必要とすることなく、僅かなデータだけでそれらを検出するモデルの学習が可能になる。

量子コンピューターは、企業が直面し得る攻撃パターンを正確に予測することさえ可能にするかもしれない。量子機械学習は予測に優れており、古典的な機械学習の性能を超える可能性を秘めている。量子アルゴリズムでは、正誤を問わずに網羅的な予測を実施し、不正解となり得る予測を行わないためである[19]。今日のサイバー攻撃では、あらゆる経路が狙われるため攻撃元の予測は不可能に思えるかもしれない。しかし、将来的に量子機械学習が発達することで、攻撃への対処が可能となり、企業・組織はサイバー攻撃に対して後手に回るのではなく、先回りして対処できるようになる可能性がある。

攻撃者は常に攻撃の機会を窺っているため、今こそ、企業はこの現実に備えることが求められる。問題に先んじた対策を講じることで、AIが生成した幾多のコンテンツから身を守り、攻撃者の一歩先を行くことができるであろう。

Endnotes

1. Issy Ronald and Jack Guy, "Tom Hanks says dental plan video uses 'AI version of me' without permission," *CNN Entertainment*, October 2, 2023.
2. IBM, "When it comes to cybersecurity, fight fire with fire," accessed November 6, 2023.
3. Kathy Haan, "Over 75% of consumers are concerned about misinformation from artificial intelligence," *Forbes*, July 20, 2023.
4. Pavel Korshunov and Sebastien Marcel, *Deepfake detection: Humans vs. machines*, arXiv:2009, September 7, 2020; David Ramel, "Researchers: Tools to detect AI-generated content just don't work," Virtualization & Cloud Review, July 10, 2023.
5. Charles Griffiths, "The latest 2023 phishing statistics," AAG IT, October 2, 2023.
6. Ralph Stobwasser and Nicki Koller, "On high alert: The darker side of generative AI," Deloitte, accessed November 6, 2023.
7. Catherine Stupp, "Fraudsters used AI to mimic CEO's voice in unusual cybercrime case," *Wall Street Journal*, August 30, 2019.
8. Melissa Heikkilä, "We are hurtling toward a glitchy, spammy, scammy, AI-powered internet," *MIT Technology Review*, April 4, 2023.
9. Stobwasser and Koller, "On high alert."
10. Interview with Shamim Mohammad, executive vice president and chief information and technology officer at CarMax, August 3, 2023.
11. College of Computer, Mathematical, and Natural Sciences, "Is AI-generated content actually detectable?," University of Maryland, May 30, 2023.
12. Interview with Ben Colman, cofounder and CEO, Reality Defender, August 2023.
13. GPTZero, "Homepage," accessed November 6, 2023; Jan Hendrik Kirchner, Lama Ahmad, Scott Aaronson, and Jan Leike, "New AI classifier for indicating AI-written text," OpenAI blog, January 31, 2023.
14. Intel, "Intel introduces real-time deepfake detector," November 14, 2022.
15. Publications Office of the European Union, *Facing reality? Law enforcement and the challenge of deepfakes*, Europol Innovation Lab, 2022.
16. Ed Bowen, Wendy Frank, Deborah Golden, Michael Morris, and Kieran Norton, *Cyber AI: Real defense*, Deloitte Insights, December 7, 2021.
17. Los Alamos National Laboratory, "Simple data gets the most out of quantum machine learning," July 5, 2023.
18. Tariq M. Khan and Antonio Robless-Kelly, "Machine learning: Quantum vs. classical," *Institute of Electrical and Electronic Engineers Access* 8, 2020: pp. 219275–219294.
19. Surya Remanan, "Beginner's guide to quantum machine learning," Paperspace, 2020.