News&Trend

ブタ腎臓を過冷却で保存、
72時間後に再移植

米原子炉スタートアップ4社が臨界達成

Interview

阿部博弥（東北大学）

MIT Technology Review

Published by KADOKAWA / ASCII

Vol. 88
2026.08

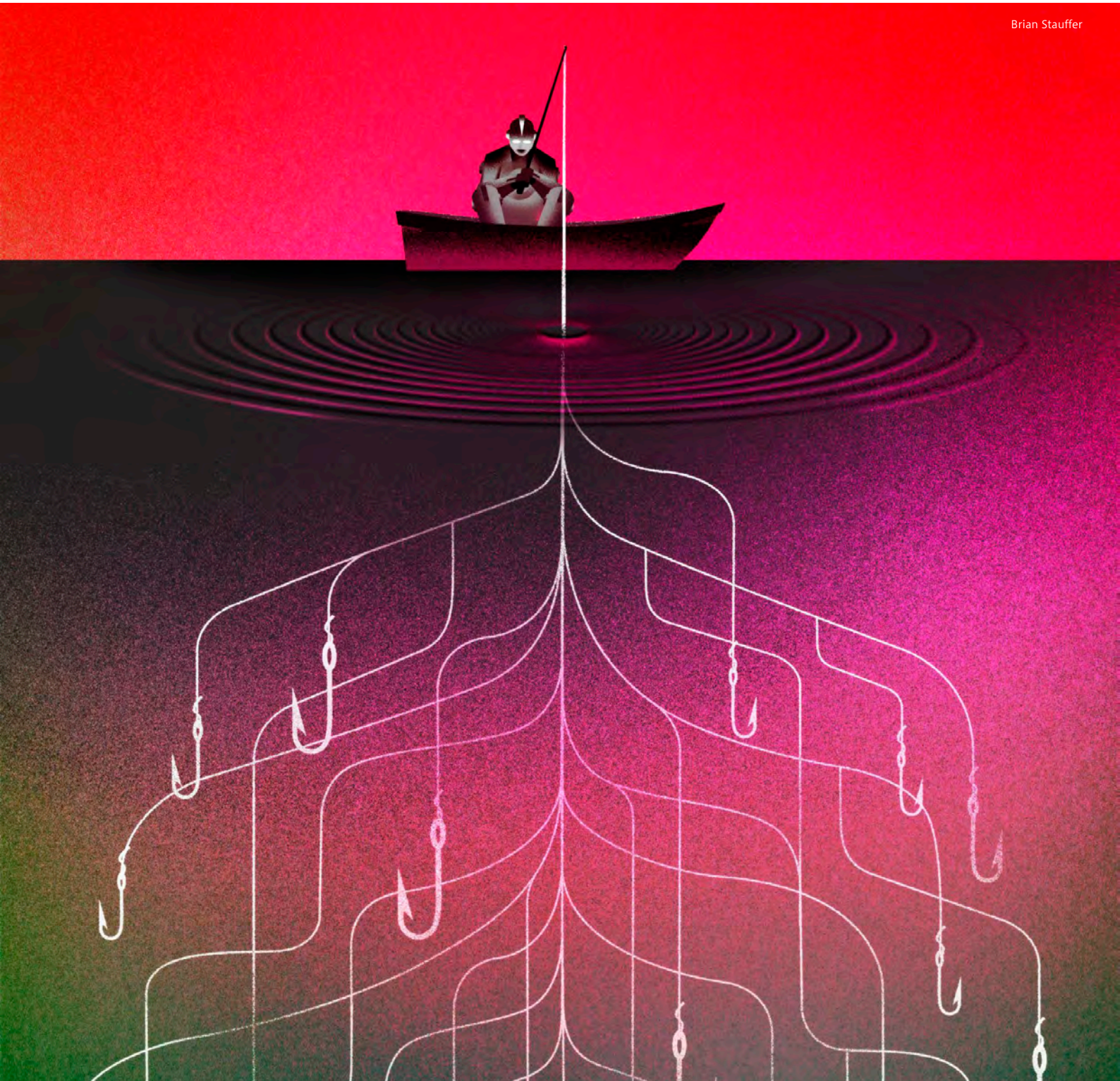
犯罪とテクノロジー
共進化する手口と捜査

INDEX

Vol. 88
2026.08

犯罪とテクノロジー
共進化する手口と捜査

人工知能（AI）を用いたオンライン詐欺の激化、暗号資産を悪用した不正取引、無人潜水艇による大規模な麻薬密輸など、最新技術の悪用は犯罪の巧妙化と深刻化を招いている。一方で、都市型監視カメラ網による迅速な捜査と過剰監視の懸念といった治安維持の功罪や、密猟阻止のためサイの角に放射性物質を注入する先進的な取り組みなど、テクノロジーを活用した新たな防衛・対抗策も進化を続けている。高度化する犯罪の現実と、テクノロジーを駆使して立ち向かう最前線をお届けする。


Brian Stauffer

Story 1

犯罪とテクノロジー　共進化する手口と捜査

どこまでが現実に？ AI主導サイバー攻撃の現在

大規模言語モデル（LLM）を全工程に組み込んだランサムウェアが発見され、AIサイバー攻撃の転換点かと注目を浴びた。だがその正体は大学の研究プロジェクトだった。一方で犯罪者たちは着実に人工知能（AI）を「生産性ツール」として活用し、攻撃のハードルを下げている。自律型攻撃はどこまで現実に近づいているのか。

by Rhiannon Williams（米国版ニュース担当記者）

アントン・チェレパノフは常に、何か興味深いものに目を光らせている。そして昨年8月下旬、まさにそれを見つけた。それは、「VirusTotal（ウイルストータル）」にアップロードされたあるファイルだった。VirusTotalは、チェレパノフのようなサイバーセキュリティ研究者が提出されたファイルを分析し、ウイルスなどの悪意あるソフトウェア（いわゆるマルウェア）の可能性を調べるために利用するサイトである。そのファイルは表面上、無害そうに見えたが、チェレパノフ特注のマルウェア検出手段が反応した。その後数時間にわたり、チェレパノフと同僚のペテル・シュトリーチェクはこのサンプルの検証を続け、それがこれまで目にしたことがないものであることに気づいた。

このファイルにはランサムウェアが含まれていた。被害者のシステム上で見つけたファイルを暗号化し、攻撃者へ身代金が支払われるまで使用不能な状態にする、悪質なマルウェアの一種である。しかし、この事例が異質だったのは、大規模言語モデル（LLM）を利用していたことだった。付随的にではなく、攻撃のあらゆる段階にわたってLLMが活用されていたのだ。このランサムウェアは、ファイルがインストールされるとLLMを利用してカスタマイズされたコードをリアルタイムで生成し、迅速にコンピューター内をマッピングして機密データを特定し、コピーまたは暗号化する。そして、身代金を要求する文書を、データの内容に基づき個別にカスタマイズして作成することができた。このソフトウェアは、人間が介入しなくても自律的に動作することが可能だった。そして実行のたびに異なる挙動を見せるため、検知するのがより難しかった。

チェレパノフとシュトリーチェクはこのランサムウェアを「プロンプトロック（PromptLock）」と名付け、この発見が生成AIにおける転換点になると確信した。生成AIを悪用して極めて柔軟なマルウェア攻撃を生み出すことができる方法を示していると考えたのだ。2人はブログ記事で、人工知能（AI）を利用したランサムウェアの初めての事例を発見したと宣言した。この発表はすぐに、世界中のメディアの広範な注目の的となった。

しかしその脅威は、当初の印象ほど劇的なものではなかった。前述のブログ記事が公開された翌日、ニューヨーク大学の研究者チームが犯行声明を発表し、このマルウェアが実は世に放たれた本格的な攻撃ではなく、単なる研究プロジェクトであると説明したのだ。ランサムウェア攻撃の一連の活動の各段階を自動化することが可能であると証明することが目的だったという。そしてその目的は達成されたと、この研究者チームは述べた。

プロンプトロックは結局のところ学術プロジェクトだったかもしれないが、本物の悪党たちは実際に最新のAIツールを利用している。ソフトウェアエンジニアたちが人工知能を利用してコードの記述やバグのチェックをしているのと同様に、ハッカーたちもそういったツールを活用して、攻撃を指揮するのに必要な時間と労力を削減しているのだ。それによって、経験の浅い攻撃者でも何かを試してみることがより容易になっている。

サイバー攻撃が時間とともにより一般的かつ効果的なものになる可能性は、小さいどころか「極めて現実的」であると、ユニバーシティ・カレッジ・ロンドンのコンピューター科学教授、ロレンツォ・カヴァッラーロは言う。

シリコンバレーでは、AIが完全に

自動化された攻撃を実行できる段階に差し掛かっていると警告する声もある。しかしほとんどのセキュリティ研究者は、この主張を大げさであると言う。「どういう訳か誰もが、AIスーパーハッカーとでも言うべきこのマルウェアのアイデアに注目しています。馬鹿げた考えです」と、セキュリティ企業エクスペル（Expel）の主任脅威研究者、マーカス・ハッチンスは言う。ハッチンスは、2017年に世界規模の広がりを見せたランサムウェア攻撃「ワナクライ」を終わらせたことにより、セキュリティ界では有名である。

専門家たちが代わりに主張するのは、AIがもたらすはるかに差し迫ったリスクに、もっと細心の注意を払うべきであるということだ。AIはすでに詐欺を加速させ、その数を増加させている。犯罪者たちはますます最新のディープフェイク技術を悪用し、誰かになりすまして被害者から巨額の金銭をだまし取るようになっている。AIによって強化されたそのようなサイバー攻撃は今後さらに頻発化し、破壊的な力を増す一方であり、私たちはそれに備える必要がある。

スパム、そしてその先

2022年末にChatGPT（チャットGPT）が登場して爆発的な人気を博すと、攻撃者たちはそのほぼ直後から生成AIツールを導入し始めた。容易に想像できることかもしれないが、そういった取り組みは、スパムを大量に作り出すことから始まった。昨年発表されたマイクロソフトの報告書によれば、同社は2025年4月までの1年間で40億ドル相当規模の詐欺や不正な取引を阻止した。「その多くはAIが生成したコンテンツによって支援された可能性が高い」という。

ChatGPTのリリース前後に収集された50万件近くの悪意あるメッセージを分析したコロンビア大学、シカゴ大学、バラクーダ・ネットワークスの研究者たちの推定によれば、現在、スパムの少なくとも半数がLLMを用いて生成されているという。また、より洗練された手口でAIが活用されている事例が増えている証拠も見つかった。研究チームが調査したのは標的型のメール攻撃だ。信頼できる人物になりすまし、組織内の従業員からお金や機密情報をだまし取る手口である。調査の結果、2025年4月までにこの種の的を絞ったメール攻撃の少なくとも14%がLLMを使用して作り出されており、2024年4月の7.6%から増加していることが判明した。

そしてブームとなった生成AIは、メールだけでなく、非常に説得力のある画像、動画、音声の生成も、これまで以上に容易かつ低コストにした。生成AIの出力結果はわずか数年前よりもはるかにリアルであり、誰かの肖像や声の偽物版を生成するのに必要なデータの量は、以前よりも大幅に少なくなった。

犯罪者たちがそうしたディープフェイクを展開しているのは、人々にいたずらをしたり、単に悪ふざけをしたりするためではない。効果があり、お金を稼げるからやっているのだと、生成AIの専門家であるヘンリー・アジャーは言う。「稼げるお金が存在し、人々がだまされ続けるのであれば、犯罪者たちはやり続けるでしょう」と、アジャーは話す。2024年に報告されたある注目を浴びた事例では、1人の従業員がデジタルで生成された会社の最高財務責任者（CFO）や他の従業員とのビデオ通話を通じてだまされ、犯罪者へ2500万ドルを送金してしまった。これは氷山の一角に過ぎないだろう。技術が向上し、より幅広く採用が進むにつれ、説得力のあるディープフェイクが引き起こす問題はさらに深刻化する一方だろう。

ある注目を浴びた事例では、1人の従業員がデジタルで生成された会社の最高財務責任者や他の従業員とのビデオ通話を通じてだまされ、犯罪者へ2500万ドルを送金してしまった。

犯罪者の手口は常に進化している。そしてAIの能力が向上する中で、そうした人々は新たな能力の助けを借りて被害者を出し抜くことができる方法を絶えず探っている。グーグルの脅威分析グループで技術責任者を務めるビリー・レナードは、「潜在的な悪意ある行為者」によるAIの利

BRIAN STAUFFER

用方法の変化を注意深く監視してきた（潜在的な悪意ある行為者とは、犯罪目的でコンピューターを利用しようとしているハッカーやその他の者を指すために業界で広く使われている表現）。2024年後半、レナードのチームは、犯罪者予備軍がグーグルのGemini（ジェミニ）のようなツールを一般ユーザーと同じ方法で使っていることに気づいた。彼らはコードのデバッグや細々とした作業の自動化だけでなく、奇妙なフィッシングメールの文章作成といった仕事もAIにやらせていた。2025年までにはAIを使って新たなマルウェアを作成し、世に放つところまで進歩していたと、レナードは言う。

現在の最大の疑問は、この種のマルウェアがどこまで進化できるかということだ。いつの日か、まったく検知されることなく数千社のシステムにこっそり侵入し、数百万ドルを抜き出せるまでになるのだろうか？

最も人気がある各AIモデルには、悪意のあるコードや違法な素材の生成を防ぐための安全対策が備わっているが、それでも悪意ある行為者はその対策を回避する方法を見つける。例えば、グーグルは、中国と関連のある行為者が同社のAIモデルであるGeminiに対し、不正アクセスしたシステム上の脆弱性を特定するように求める様子を観察した。Geminiは当初、この要求を安全性の理由で拒否した。しかし攻撃者は、人気サイバーセキュリティ・ゲームである「キャプチャー・ザ・フラッグ（Capture The Flag）」の参加者になりすますことで、Geminiに自らのルールを破らせることに成功した。この巧妙な形の脱獄により、Geminiは、そのシステムを悪用するため

に使われた可能性のある情報を渡してしまった（グーグルはその後、そういった種類の要求を拒否するようにGeminiを調整した）。

しかし悪意ある行為者たちは、AI大手企業のモデルを悪意ある目的に利用しようと試みることだけに注力しているわけではない。今後、悪意ある行為者たちは、オープンソースのAIモデルを採用する可能性がますます高まると、米司法省の元戦術専門家で、現在はサイバーセキュリティ企業インテル471（Intel 471）の上級インテリジェンス・アナリストを務めるアシュリー・ジェスは言う。オープンソース・モデルは、安全対策を解除して悪意あることをさせるのがより簡単だからだ。「（悪意ある）行為者たちはそういったものを採用するだろうと考えています。脱獄させ、必要とするものにカスタマイズできるからです」と、ジェスは話す。

ニューヨーク大学の研究チームはプロンプトロックの実験において、オープンAI（OpenAI）の2つのオープンソース・モデルを使用した。そして、モデルに自分たちが意図したことをさせるのに、脱獄の手法を用いる必要さえないことを発見した。そのため、攻撃が格段に容易であると、研究者チームは話す。そういったオープンソース・モデルは倫理的な整合性を考慮して設計されており、要求への応答方法を決定する際の目標や価値観について熟慮されているものの、同等のクローズドソース・モデルと同じ種類の制約は備えていないと、このプロジェクトに携わったニューヨーク大学博士課程の学生、ミート・ウデシは言う。「それこそが、私たちが検証しようとしていたことです。それらのLLMは、倫理的整合性が調整されていると主張しています。それでも私たちは、こうした目的で悪用することができるでしょうか？ その答えは『できる』であることが判明しました」。

犯罪者たちはすでにプロンプトロック式の隠密攻撃を成功させており、単純に私たちがその証拠を見たことがないだけであるとも考えられると、ウデシは言う。もしそうなら、理論上、攻撃者が完全自律型のハッキング・システムを作り出した可能性がある。しかしそのためには、AIモデルの信頼ある行動を確保している重要な障壁だけでなく、モデルに組み込まれている、悪意ある目的での利用に対する反感も乗り越える必要があっただろう。そのすべてを、検知されるのを回避しながら克服しなければならなかったはずだ。それは実際のところ、かなり高いハードルである。

ハッカーのための生産性向上ツール

では、確実に分かっていることは何だろうか？ 人々がAIを悪意ある目的でどのように利用しようとしているかということに関して、大手AI企業自身から現時点で入手可能な最良のデータの一部が提供されている。それらのデータが示している事実は、少なくとも一見したところでは、確かに警戒すべきものであるように感じられる。11月、グーグルのレナードのチームは報告書を発表し、悪意ある行為者たちがAIツール（グーグルのGeminiを含む）を使ってマルウェアの挙動を動的に修正していることを明らかにした。例えば、そのマルウェアは検知を回避するように自己修正をすることができた。同チームは、この事実が「AIの悪用の新たな運用段階」に入ったことを告げていると報告書に記した。

しかし、この報告書で掘り下げて調査された5つのマルウェア・ファミリー（プロンプトロックを含む）は、容易に検出されるコードで構成されており、実際に何の損害も与えなかったと、サイバーセキュリティ

「私たちは非常に高度なサイバー作戦に対する障壁が根本的に低下した時代に突入しつつあります。そして、攻撃のペースは多くの組織がそれに備えるペースよりも速いスピードで加速するでしょう」

ジェイコブ・クライン（アンソロピック脅威インテリジェンス責任者）