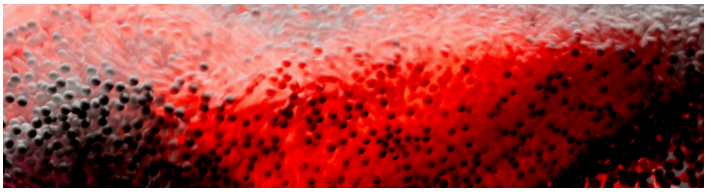


サイバーとトラスト

現実を守る：合成メディアの時代における真実

日本のコンサルタントの見解

サイバー攻撃はその規模と巧妙さを増しており、従来の防御アプローチでこれらの脅威に対処することは既に限界が生じている。急速に進化するサイバー脅威に対抗するため、企業は高度なスキルを持つ人材を求めるが、その需要は供給を上回り、サイバーセキュリティ人材の不足が常態化している。このような状況下において、サイバー脅威に対するAI技術の導入は、セキュリティ投資が増大し続ける現在の状況を打破する可能性を秘めており、AIとサイバーセキュリティの融合について活発に議論されている。

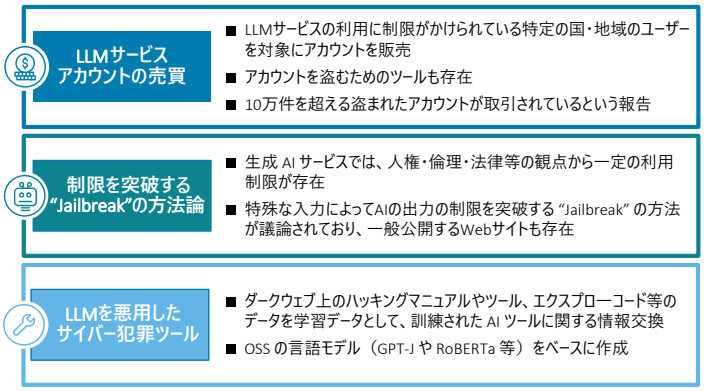
AIを悪用したサイバー攻撃

AI技術の進化がもたらすのはプラスの側面だけではない。サイバーセキュリティの歴史は古くからいたちごっこである。我々がサイバー防御のためにAIを開発しているように、攻撃者もまたサイバー攻撃をさらに複雑かつ容易にするためにAIを活用している。生成AIの進化に伴い多様な業務が効率化されつつあるように、サイバー攻撃も以前と比べてより短時間で準備ができるように効率化され始めている。

実際にダークウェブのフォーラムサイトでは、大規模言語モデル（LLM）を利用していかにサイバー攻撃を容易かつ効率的に実施するかについて盛んに議論されている。AIを活用するサービスでは、犯罪行為や倫理違反に繋がる可能性のある出力は制限されるように実装されているものの、現時点で対策が十分に追いついているとは言えない。AIが便利で身近なものになればなるほど、サイバー犯罪の敷居を下げる要因となり得ることが懸念されている。

2023年3月には、サイバー攻撃の補助に特化した対話型AIサービス「WormGPT」の出現も報告されており、当該サービスはオープンソースの言語モデルであるGPT-Jが利用されたことが報告されている[1]。サービスの使い方はChatGPTと同様である。サイバー攻撃の手法や攻撃ツールに関する質問をAIに投げかけることで、例えばシステム内に格納された個人情報をAIに回答させる方法や、銀行を装ったフィッシングメールの文面の作成、マルウェアの攻撃コードをAIが即座に作成するといったサービスであり、幅広いサイバー犯罪への悪用が危惧された。当該サービスは本日時点で既に公開を停止しているが、類似のサービスは今後も次々と出現することが予想される。

図1: ダークウェブのフォーラムサイトにおける議論



出所：B. Toulas, “Over 100,000 ChatGPT accounts stolen via info-stealing malware,” BleepingComputer, June 20, 2023.

出所：Y. Jin et al. “DarkBERT: A Language Model for the Dark Side of the Internet,” arXiv, May 18, 2023.

生成AI悪用による偽情報と脅威

生成AIが発展するにつれて、それを悪用した脅威が顕在化している。2023年8月に米ハワイ州マウイ島において発生した山火事では、陰謀論を含めた様々な偽情報がSNSを中心に拡散した。拡散した偽情報には、信憑性を持たせるために生成AIで生成された画像が用いられていた。この偽情報の拡散については、セキュリティベンダーやシンクタンクの研究者による分析で国家が関与していた形跡があることが報告されている[2]。

AIの高度化により偽動画や偽音声なども容易に生成でき、本物との区別が難しい状況になりつつある。世界各国で行われる選挙活動においても、生成AIによる偽情報が氾濫し、民主主義の根幹に関わる懸念も生じている。

例えば、2024年1月に投票が行われた台湾総統選挙では、対立候補者を中傷するような投稿を繰り返すSNSアカウントや、生成AIで生成した対立候補の偽音声や偽動画が拡散され、台湾捜査機関による注意喚起も行われていた[3,4]。また2024年1月には米大統領選予備選に向けて、有権者のもとに投票しないよう誘導する米大統領になりすました偽音声通話が掛けらたことが話題となった[5]。

このような生成AIを悪用した偽情報の拡散は、国内外を含めたステークホルダーらの関係を分断させる目的もあるとされ、民主主義や安全保障への脅威にもなりつつある。

法規制やガイドラインの整備

AIの進化に伴うサイバー脅威の高まりを受け、世界各国でAIの開発や利用に関する規制や制度の策定を急速に進めている。

広島AIプロセス

2023年5月のG7広島サミットの結果を受けて、生成AIに関する国際的なルールの検討を行うために当該枠組みが立ち上がった[6]。広島AIプロセスでは、安全、安心、信頼できるAIの実現に向けて、日本による議論のリードのもと、G7で国際指針と行動規範が示された。また生成AIに関するG7共通の課題・リスクとして、透明性、偽情報、知的財産権、プライバシーと個人情報保護、セキュリティと安全性などが確認された。広島AIプロセスの成果のひとつである「AI開発者向けの国際指針」の中では、AIライフサイクル全体にわたってリスクの特定と措置を講じることや、透明性を確保するための報告書を公表すること、AIガバナンスおよびリスク管理方針を策定することなどを奨励している。偽情報対策としては、生成AIを利用して生成される偽情報への対策に資する技術などの実証が合意された。

欧州AI規制法案（EU AI Act）

欧州議会は2023年の本会議において、生成AIを含む包括的なAIの規制案であるEU AI Actについて大筋の合意に至った[7]。本規制では、AIシステムを4つのリスク区分にカテゴライズし、各リスクレベルに応じた要件・規制を設定する。リスクレベルが最も高い区分である「許容できないリスク」に属するAIシステムには、公的機関のソーシャルスコアリングや公的空間での法執行目的の遠隔生体認証などが該当する。これらの用途においては、AIの利用自体がEU全体の統一ルールとして禁止される。2番目のリスク区分である「高リスクと識別されたAIシステム」に対しては、リスクマネジメント、データガバナンス、透明性・情報提供、サイバーセキュリティなど様々な要件が課される。3番目の「リスクが限定的とされるAIシステム」では透明性に関する義務が課される。例えば、生成AIのシステムでは、生成されたコンテンツがAIによって生成されたことを明示することが義務付けられるほか、大規模言語モデル（LLM）を使用するAIシステムでは、学習に用いられたデータを開示することが求められている[8]。AI規制法案の影響を受けるのは、EU域内のみではない。EU在住の利用者をターゲットにサービスを提供する場合には、日本の企業においても本規制が適用される。欧州ではGDPR（General Data Protection Regulation：一般データ保護規則）に代表されるように、GAFAをはじめとするプラットフォーマーに対する規制の動きが続いている。データを独占する一部のテック企業によって健全な競争が阻害され、AI市場が支配されるのを避けたいという思惑が伺える。

米国AI権利章典（Blueprint for an AI Bill of Rights）

米国ホワイトハウスの科学技術政策局（OSTP）は2022年10月にAI権利章典の青写真（Blueprint for an AI Bill of Rights）を公表し、AIシステムの設計、使用、導入にあたり考慮すべき5つの原則を提示した[9]。AI権利章典では、AIシステム開発の際にリスク分析とリスク軽減措置を講じることや、AIやアルゴリズムが人種や性別などの属性に由来する差別をしないように対策を講じること、AIシステムのユーザーに対してAIを使用することの理由やその影響について説明することなどが期待されている。なお、これらの原則は法的拘束力を持っておらず、欧州がルール形成というハードローのスタンスであるのに対し、米国はソフトローのスタンスを軸に据えていることが伺える。

図2: AI関係者向けの広島AIプロセス国際指針

安全、安心、信頼できるAIの実現に向けて

AIの発展はサイバー犯罪の短時間かつ容易な実行を可能としており、企業はAIを利用した新たな攻撃にも対処が求められる。これらの脅威の高まりを受けて、世界各国でAIの開発や利用に関する規制や制度の策定が急速に進められている。

一方で、AIシステムによるリスク軽減のための仕組みづくりには多くの課題が残っている。例えば、AIの透明性とAIが攻撃者に悪用されるリスクとのバランスについてはあまり議論が進んでいない。透明性を確保するために、AIモデルの情報や学習に使用したデータセットを公開することで、攻撃者がサイバー犯罪にそれらを悪用するリスクも高まることにつながる。実際にWormGPTの事例では、オープンソースの言語モデルが悪用されていることからも、AIの透明性と悪用されるリスクとのバランスが重要であり、公開する情報を制限しつつも、AIの透明性を担保する仕組みを開発する必要があるといえる。

偽情報やハルシネーションなど、AI固有のリスクに対する対策についても早急に議論が求められる。広島AIプロセスでは、AIが生成したコンテンツを識別できるように電子透かしやコンテンツ認証、証明メカニズムを開発することが明示された。今後はより具体的に、AIが生成したコンテンツの信頼を強化するための技術や仕組みの検討を加速させる必要があり、例えば国内のトラストサービスをどのようにAI技術やサービスに応用させるかなどの議論は急務である。同時にAIを過信することによるリスクに関しても検討が必要である。EUでは一部の用途において、AIシステムの利用を制限する取り組みが進められているが、国内ではAIの用途に関して制限を設けていない。AIの出力に誤りが含まれる可能性を前提としたシステム構築や意思決定が重要となる。

国際的なルール形成の取り組みでは、これまで原則やガイドラインの整備が進められてきたものの、それらを遵守するために企業・組織が具体的に何をどこまで実施すれば良いかといった具体的な内容については明示されていない。AIモデルの透明性、信頼性を定量化するための仕組みやAIシステムのセキュリティリスクを評価するための手法や技術、AIの適切な監査と運用の具体的なルールについて、海外の模倣ではなく日本の目指すべき姿に基づいた制度設計が不可欠である。

1. Trend Micro, “過度な期待と現実：サイバー犯罪のアンダーグラウンドにおけるChatGPTを中心としたAIの動向”, accessed February 7, 2024.
2. Recorded Future, “Converging Narratives on Hawaii Wildfires Advance Different Influencers’ Objectives”, accessed February 06, 2024.
3. NHK, “台湾総統選挙 AI悪用とみられる不審アカウントや偽動画広がる”, accessed February 7, 2024.
4. 法務部調查局, “境外敵對勢力介入我總統大選 國人宜謹慎識別網路假訊息”, accessed February 13, 2024.
5. REUTERS, “Fake 'Biden' robocall tells New Hampshire Democrats to stay home”, accessed February 6, 2024.
6. 総務省, “広島AIプロセスについて”, September 23, 2023.
7. European Parliament, “Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI”, accessed February 7, 2024.
8. White & Case LLP, “Dawn of the EU's AI Act: political agreement reached on world's first comprehensive horizontal AI regulation”, accessed February 7, 2024.
9. The White House, “Blueprint for an AI Bill of Rights”, October 2022.

執筆者

神薗 雅紀
パートナー・所長
デロイト トーマツ サイバー合同会社
サイバーセキュリティ先端研究所

セキュリティベンチャー企業や政府研究機関を経て、2019年より現職。研究開発を主軸とし、新たなソリューションやアセットの開発、研究開発事業支援、テクノロジー 特区の立案および支援など、多数の新たなテクノロジー領域やオポチュニティーの立案に従事。上記貢献により、2018年総務大臣奨励賞を受賞。

熊谷 裕志
ディレクター・主席研究員
デロイト トーマツ サイバー合同会社
サイバーセキュリティ先端研究所

非営利団体にて脆弱性の解析や調査研究、セキュアコーディングの普及啓発等に従事、その後ベンチャー企業やコンサルティングファームにてコア技術等の研究開発をリード。2019年より現職。現在は研究・新規ソリューション開発をリード。

高田 雄太
スペシャリストリーダー・上席研究員
デロイト トーマツ サイバー合同会社
サイバーセキュリティ先端研究所

電気通信事業会社、コンサルティングファームを経て、2019年より現職。セキュリティやプライバシー、トラストに資する技術の研究開発をリード。大学講師や学会委員として、サイバーセキュリティの講義演習を通じた人材育成にも携わる。博士（工学）。

鈴木 将吾
スペシャリストマスター・主任研究員
デロイト トーマツ サイバー合同会社
サイバーセキュリティ先端研究所

外資系コンサルファームにてサイバー攻撃対策やインシデント対応に関するコンサルティングに従事。2019年より現職。現在はデジタルツインや5G関連技術を活用したソリューションや新規アセットの開発をリード。

櫻井 悠次
コンサルタント
デロイト トーマツ サイバー合同会社
サイバーセキュリティ先端研究所

福永 拓海
スペシャリストジュニア
デロイト トーマツ サイバー合同会社
サイバーセキュリティ先端研究所

野本 一輝
スペシャリストジュニア
デロイト トーマツ サイバー合同会社
サイバーセキュリティ先端研究所