2024 年 9 月 18 日

情報空間のリスク研究会
「サイバー戦・情報戦における生成 AI の脅威
―情報操作型サイバー攻撃を中心に―」
実施報告

中曽根平和研究所・情報空間のリスク研究会では 2024 年 9 月 18 日、独立行政法人情報処理推進機構の長迫智子委員からのご報告をもとに、議論を行いました。要旨は以下の通りです。

長迫智子委員より「サイバー戦・情報戦における生成 AI の脅威―情報操作型サイバー攻撃を中心に―」と題して報告が行われた。

まず、サイバー情報戦の様相として「人間の認知領域は戦闘領域である」と見なされたことにより、敵国側の影響工作が一層容易になってきたという現状を概観した。従来のサイバー攻撃に見られた「機能破壊型」や「情報窃取型」に加えて、ディスインフォメーションや特定のナラティブが用いられた「情報操作型サイバー攻撃」が登場し活発化しており、特に生成 AI（Generative AI）で作成した画像や映像の活用が進んでいると指摘された。こうした情報戦、認知戦は平時から行われるので、実際に有事を想定した当事国のみならず、第三国や国際世論においても影響が懸念されている。生成 AI の普及は「情報操作型」の脅威を高度化かつ容易化した点で、このような類型のサイバー攻撃拡大への懸念がより高まった。

さらに、「情報窃取型」と「情報操作型」を組み合わせた「ハイブリット型」のサイバー攻撃の脅威が高まっていることも指摘された。例えば、米国・ペロシ下院議長の訪台時に、中国からと見られるデジタルサイネージのハッキングが起こり、サイネージには米国の台湾関与を牽制したり「一つの中国」を強調したりするメッセージが流れた。いまや、サイバーセキュリティと情報戦・認知戦の一体的な備えが必要となっていると捉えることができる。加えて、アジアにおける情報戦・認知戦の動向として、2020 年代以前はロシアが積極的に攻撃を行なっていたが、近年では中国の活動が活発化しており、ソーシャルネットワーキングサービス（SNS）上のみならず、フェイクメディアが作られ、そのフォーマットと SNS を組み合わせてディスインフォメーションを拡散する事例も確認されている。

次に、生成 AI の脅威について説明があった。まず、生成 AI サービスの増加によって、高度なスキルなしに様々なコンテンツが容易に大量生成できる状況が指摘された。さらに、よりリアリティのあるコンテンツの作成も可能となっており、短時間に大量にディスインフォメーションを含む説得力のある記事を作成して拡散することが可能となった。。一方で、ハルシネーションといった生成 AI の技術的な限界も存在しており、正確無比なレベルには達していないと考えられることから、対抗策の一環として人々のリテラシーを向上させる必要性が述べられた。現時点におけるディープ

フェイクの判別ポイントとしては、風景画像における境界線の曖昧さや複数人が映っている場面における顔の描写の曖昧さ、人物画像における手指や髪等細部の描写の曖昧さ等が挙げられた。

さらに生成 AI を用いた攻撃事例として、各国の選挙に関連した事例が挙げられた。例えば、米国のバイデン大統領が徴兵への協力を呼びかけたというディスインフォメーション動画について、視聴者が判別ポイントを意識せずに視聴してしまうと違和感なく本当にバイデン氏がそう話した（＝真実）と受け入れてしまうクオリティになっており、一見してディープフェイクと見抜くのは難しいレベルになっている、と指摘された。

台湾・民進党の頼清徳総統を中傷する工作事例では、生成 AI を用いたバリエーションのあるミーム画像の大量生成や、SNS 上で拡散されやすいセンセーショナルな要素を満たした動画の作成が容易になっている。民主主義に対する脅威としてこのような外国からの影響工作や選挙干渉だけでなく、生成 AI による動画や静止画が政党間の争い等内国上の問題で特定政党や候補者への批判に利用される例も増加している。生成 AI を用いた選挙キャンペーンは有権者の選挙に対する不信感を高める可能性や、中国、ロシア等諸外国の影響工作に利用される可能性があるので、自律的・自制的であるべきだと述べられた。

加えて、有事の事例として、ウクライナのゼレンスキー大統領が降伏声明を出したと偽る動画の事例等に触れつつ、有事の際にはディスインフォメーションが格段に増加するだけではなく、生成 AI の悪用が社会の混乱を一層深める懸念が指摘された。そのほか、日本の事例として、加藤官房長官や岸田首相のディープフェイク画像・動画等が挙げられた。ウクライナ支援に係る岸田首相の事例を除き、これらはいたずら目的であったと考えられているが、その目的から離れて影響工作キャンペーンに悪用される可能性もあるため、ディープフェイクが蔓延しない仕組みづくりが必要であると指摘された。

こうした攻撃を行うアクターとして、生成 AI を用いて多言語の画像を大量生成しているとされるロシアのグループ「ドッペルゲンガー」や、ディープフェイク動画・画像、ミーム画像の作成、拡散を任務とする中国のグループ「Storm-1376」やSpamouflageキャンペーンの存在が挙げられた。また、生成AIを用いた攻撃が懸念される事例として、「オペレーション・オーバーロード」(Operation Overlord) についても説明があった。「オペレーション・オーバーロード」とは、フェイクメディアとディスインフォメーションの大量の通報を組み合わせたロシアによる影響工作キャンペーン手法の一つで、ロシア側のアクターが投稿したディスインフォメーションを自ら拡散しつつ、同時にその情報のファクトチェックを依頼するというものである。この手法は、（１）サイバー空間に大量にディスインフォメーションが拡散されていると印象付けて社会の信頼を毀損する、（２）ファクトチェック結果の公開によって親ロシア的なナラティブの拡散を助長する可能性がある、（３）ファクトチェック側に負荷をかけてディスインフォメーションの検証機能が阻害される可能性があることから、問題視されている。

このような AI 時代の情報戦への対策として、Ｇ７サミットにおける AI プロセスの採択や米英に

おけるAIセーフティ・インスティテュートの設置を含むAIガバナンスに関する各国の動向が挙げられたが、これらは現状発展途上であるとも指摘された。また、生成AIの悪用に対して、電子透かしやAI画像判定、オリジネーター・プロファイル（Originator Profile）、サイバーワクチンといった技術も紹介された。さらに、事後のファクトチェックのみならず、デバンキング（Debunking）を事前に行う「プレバンキング」（Prebunking）を行うことで、国民のレジリエンスを高めることが重要であるとも指摘された。

一方で、ディスインフォメーションや影響工作についての警告が過剰になると、人々が選挙や報道に不信感を抱くようになり結果として民主主義の衰退を招く、という指摘もある。ファクトチェックを利用した工作も存在するため、バランスのとれた警戒体制・情報発信とそれを受け止める国民のリテラシー涵養が両輪として必須であると言及された。

以上の報告を受けて、質疑応答では「生成AI動画に関する判別ツールの有無」、「拡散経路に関する研究の必要性」、「AI以外でディープフェイクを見抜く技術」、「生成AIサービスプロバイダーによるガードレール」「ディスインフォメーションに対する過大評価の問題」、また、「生成AIコンテンツの影響力」などに関する質問、コメントが交わされた。