「AI とセキュリティ」の論点とリスクシナリオの整理

2024 年 4 月

JCIC 客員研究員　古澤一憲

はじめに

AI を巡るセキュリティの議論への関心が増している。生成 AI 技術の進化により、対話型 AI をはじめとする生成 AI アプリケーションの普及が急速に進んでいることが大きなきっかけだろう。実社会での利用が増えるに伴い、リスクにさらされる対象もまた増していく。各国政府や国際機関も法規制やガイドライン整備などを急いでいる。

こうした過渡期の状況において、AI とセキュリティに関係する論点に混乱が生じていると懸念される場面を見かける。AI は IT 技術が登場した際と同様に、これからより一層多くの場面での活用が広がる技術だ。IT とセキュリティの関係を議論が多くの視点と分析を必要としてきたのと同様に、AI とセキュリティの関係もまた多角的で重層的なものとなるだろう。あらゆる資産がサイバー脅威にさらされる状況にある以上、これは必然的な構図といえる。

AI とセキュリティの関係は、例えば「セキュリティへの AI の活用（AI for Security）」「AI 自身のセキュリティ対策（Security for AI）」といった主客の視点から分類をされることがある。直感的な理解をしやすい簡潔な整理であり、説明を伴わず暗黙の前提として扱われることも少なくない。一方で、各論について高度な研究や議論が進み、複雑な成果や具体的な事例も増えたことで、これだけで正確に文脈を理解することは難しくなってきている。

セキュリティの議論について、何から何を守るのかという前提を共有することは極めて重要である。本稿では、情報セキュリティの基本的枠組みであるリスク分析の手法を活用し、AI とセキュリティの論点とリスクシナリオを整理する。具体的には、リスク分析の主要な観点である **脅威・脆弱性・資産** の観点から、AI によって従来のリスクシナリオにどのような変化が生じようとしているのか、要素別の影響を踏まえたリスクシナリオを検証する枠組みを提示する。

これにより、対処すべき脅威と脆弱性の所在を明らかにし、必要な対策の枠組みを精緻に検討することを可能にする。また、想定される AI セキュリティリスクシナリオへの対応戦略について、短期リスク・中長期リスクの視点を踏まえながら分析した結果を示す。

AI とセキュリティはそれぞれが高度に専門的な領域であり、両者に精通した専門家の数は世界的に見てもまだ少ない。しかし、生成 AI の急速な普及により、AI セキュリティの検討もまた喫緊の課題となってしまった。AI とセキュリティの論点の解像度を上げることは、AI にルーツを持つ専門家とセキュリティにルーツを持つ専門家が共通認識に基づいた議論を交わし、それを効果的に進展させることにつながるだろう。本稿がその一助となることを望む。

1. AI とセキュリティの論点の基礎的分類の振り返り

AI とセキュリティはそれぞれ多くの論点を含む極めて専門的な領域として発展してきた。それらの複合領域となる AI セキュリティを単純な掛け合わせによって論じようとすれば、複雑で高度な論点を膨大な分量で扱うこととなり、議論の発散は避けられない。

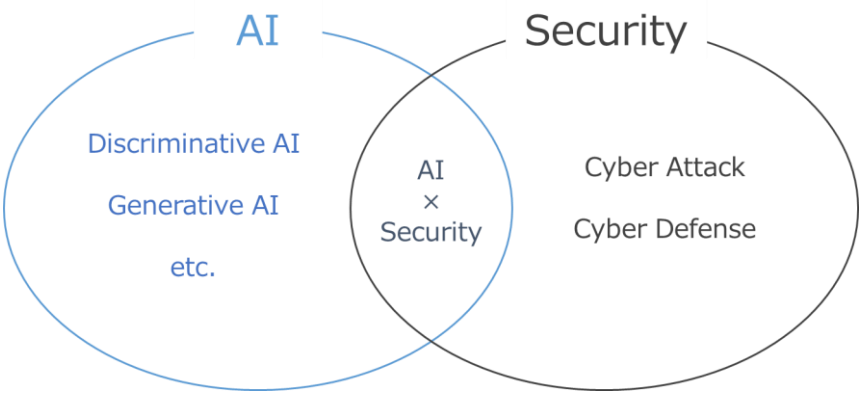


図 1 AI とセキュリティの交差する領域

これに対し、従来からの分類として「AI for Security」と「Security for AI」を区別する考え方がある。

「AI for Security」は、AI をサイバーセキュリティに応用する分野を意味する。EDR や次世代 Firewall での機械学習ベースの脅威検知への応用などが代表的である。これらの AI を活用したサイバーセキュリティ対策は「Measures by AI」とも呼ばれる。逆に、サイバー攻撃者によって AI 技術がマルウェアの開発やセキュリティ対策の迂回などに悪用される場合もあり、これらは「Attack using AI」と分類される。

もう一方の「Security for AI」は、AI 自身のセキュリティを確保するための分野である。AI への悪意ある攻撃や機微情報の漏えい等を防ぐため、AI 特有の脆弱性とその対策を明らかにすることを目的としている。AI 特有の脆弱性としては、敵対的サンプル攻撃やデータポイニング攻撃、大規模言語モデルにおけるジェイルブレイクなどがよく知られている。これらは「Attack to AI」とも呼ばれ攻撃の対象が AI であるということが強調されている。

この分類はシンプルで明快であり、研究者が自身の専門領域にもとづいてテーマを識別しやすくするという文脈でも機能してきた。少々乱暴に俯瞰をすれば、「AI for Security」は主にセキュリティ研究者やセキュリティベンダーによる研究開発の文脈で新しいメソッドとして取り込まれてきた経緯があり、「Security for AI」は AI 技術者がアルゴリズム研究の延長線上で、AI の安全性を担保するための品質保証に近いテーマとして発展してきた。

しかし、AI の社会への急激な普及を迎えた現在、「Attack using AI」も「Attack to AI」も、研究段階を超え今そこにある脅威として複雑に絡み合っている。AI は単に AI モデルではなく、AI システムとして社会の中に実装されはじめている。AI チャットツールへの機密情報の入力が漏えいする懸念や、生成 AI を悪用した精巧なフィッシングメールによる被害、またはディープフェイクによる偽情報の拡散まで、「AI に関係する脅威」は多種多様な様相をみせている。発生している事象に対してどのように手を打てばよいか、これを正確に検討するためには、原因の所在と被害発生のメカニズムを解き明かすことが必須である。

2. AI とセキュリティに関係するリスクシナリオの整理

本稿では、AI をシステムの一部として捉えたうえで情報セキュリティリスクアセスメントの手法を適用することで AI セキュリティに係るリスクシナリオを包括的に整理する。

まず、ISO/IEC 27000 ファミリー規格[1] をはじめ、情報セキュリティの「リスク」を構成する概念として広く用いられている「脅威」「脆弱性」「資産の価値」および「管理策」のフレームワークを導入する。この際、リスクは以下の通り表すことができる。

リスク = 脅威 × 脆弱性 × 資産（の価値）

ここで、「脅威」は、資産に望ましくない影響を与えるリスク源となる要素を意味し、「脆弱性」は脅威によって付け込まれる資産や管理策の弱点、「資産の価値」は資産が侵害された際に事業にあたえる影響度をそれぞれ意味する。また、「管理策」は、組織がリスクを低減や移転などによって管理するための対策とする。

情報セキュリティリスクマネジメントのフレームワークを基に AI セキュリティのリスクを考慮するにあたり、これらの構成要素すべてが AI に関係する性質を帯びる可能性がある。一方で、すべての要素に AI が関係せずとも、いずれかの要素に AI が関係した際には、最終的なリスクは AI に関係したリスクと捉えるべきものになる。例えば、AI を悪用したサイバー攻撃者（脅威）が従来の IT システムを攻撃した場合、これは AI に関連するセキュリティリスクといえる。逆に、データアクセス権限の不備など、AI とは直接的な関係のない人的な不注意によって AI システム内の情報資産（資産）が漏えいした場合、これもまた AI に関連するセキュリティリスクとみることができるだろう。しかし、両者において求められるリスク管理策は根本的に異なるものになる。本節の目標は、これら要素別の AI への関係をリスクシナリオとして整理することである。

図 2 個別のリスクファクターへの AI の関与有無

図 2 では、脅威・脆弱性・資産のそれぞれの要素について、AI の関与によって生じる新たな要素を プラス AI として個別に扱っている。この整理方法のメリットは、総合的なセキュリティ対策の検討を行う者にとって、AI によって生じる従来と異なるリスクについて、その原因となる要素を特定し、直接的効果を生む対策を講じるため必要な可視性

[1] https://www.iso.org/standard/73906.html

を提供する点にある。この点についてより具体的な理解を進めるための助けとして、表 1 に現在すでに顕在化している、または今後課題となることが懸念されている代表的 AI セキュリティリスクシナリオを、脅威・脆弱性・資産への AI の関与有無を識別した形で分類した例を示す（色付き項目がプラス AI 要素を示す）。

表 1 代表的 AI リスクシナリオのリスクファクターにおける AI 要素の関与分析

	脅威	脆弱性	資産・被害	AI セキュリティリスクシナリオの主要な例 （参照文献の詳細は巻末に記載）
	（意図的脅威）			
1	AI の仕組みを熟知したサイバー攻撃者	AI 特有の脆弱性 (1	AI システムの機微情報 AI サービスの安全性 AI サービスの正確性	– 産業スパイが生成 AI へ作為的な入力を行い、出力からモデルやデータの情報を窃取する (2 (3 – テロ行為者が作為的な標識を道路に設置し、自動運転車の AI を誤作動させる (4 – 犯罪組織が対象 AI の学習データを汚染し、本来と異なる応答をさせる (5
2	AI の仕組みを熟知したサイバー攻撃者	公開 AI サービスの脆弱性または悪用可能性	悪用可能な技術情報 標的組織の情報	– 標的型攻撃グループが生成 AI をマルウェア開発の支援ツールとして悪用する (6 – サイバー犯罪組織が生成 AI で標的組織の情報を収集する
3	AI を攻撃ツールに利用するサイバー攻撃者	従来のシステム脆弱性	従来システムの機微情報 従来サービスの正常稼働	– 犯罪組織が闇市場の生成 AI で作成した高精度フィッシングメールでマルウェアを拡散する (7 – 産業スパイが AI を悪用したツールでユーザー認証や不正検知メカニズムを回避する (8 (9
4	AI を攻撃ツールに利用するサイバー攻撃者（生成 AI を不正な目的で利用する組織等）	ヒトの認知機構の弱点（錯覚、思い込み）	ヒトの認識、世論 プライバシーやその他権利	– 対立する国や組織の情報機関が偽情報の拡散によって自身に有利な世論操作を行う (10 – 闇ビジネスとしてフェイクポルノや他者の著作権を侵害した商品が流通する (11
5	従来のサイバー攻撃者	従来のシステム脆弱性	AI システムの機微情報 AI サービスの正常稼働	– 犯罪組織がメール詐欺で AI システム管理者のアクセス権限を不正に取得し、データを盗む – テロ行為者が AI システムへ DDoS 攻撃をしかけ、サービスの正常稼働を妨害する
6	従来の内部犯行犯	組織の情報管理不備	AI システム内の個人情報 AI システム内の営業秘密	– 内部犯行者が組織内に保存されている AI システム関係の社外秘データを持ち出す
	（偶発的脅威・環境的脅威）			
7	AI サービスの利用者の不注意	運用ルールの不備、手順の誤り等	利用者の個人情報 組織の営業秘密	– 社員が入力データを学習に利用する生成 AI 対話サービスに自社の機密情報を入力してしまう – 生成 AI の対話履歴を公開設定にしてしまう
8	AI システムの不具合や設定ミス	AI 特有の脆弱性	利用者の個人情報 利用者資産への不利益	– AI システムの不具合で情報漏えい等がおこる – 組み込んだ基盤モデルの不具合によって自社サービス利用者の情報が他者に開示されてしまう
9	災害や地政学的リスクなど	組織体制や統制の未整備	AI システム内の機微情報 AI サービスの正常稼働	– 他国の規制によりデータの扱いが変わる (12 – 他国や地域への災害により基盤 AI サービスが利用できなくなる

リスクシナリオの整理にあたっては、脅威の特性から「意図的脅威」と「偶発的脅威・環境的脅威」への分類を行っている。意図的脅威に分類されるシナリオは、悪意のある意図をもった行為者により生じるリスクであり、近年特にサイバーセキュリティリスクとして重要性が増している。サイバー攻撃者と資産オーナーの間で行われる攻防の各要素にAI の要素が様々な形で加わってきている。偶発的脅威・環境的脅威については、より情報セキュリティリスク管理の文脈に馴染むもので、組織が AI を安全に利用するための統制のあり方が主なテーマとなる。

①シナリオ 1-2：意図的脅威の最上段に示したシナリオは、AI を熟知した攻撃者が、AI 特有の脆弱性を悪用する攻撃を行うによって生じるリスクである。主に「Security for AI」の文脈で原理と対策の研究が進んでいる領域に該当し、特に技術面の対応において全く新しいセキュリティ対応が今後求められることになる。既に顕在化した脅威としては、生成 AI から本来想定されたものと異なる応答やセーフガードで禁止された応答を引き出すプロンプトインジェクション（ジェイルブレイク）と呼ばれる攻撃手法が有名である。この他にも回避攻撃やデータポイズニング攻撃といった機械学習アルゴリズム特有の脆弱性をつく、これまでとは異なるメカニズムで発生する脅威もこのシナリオに該当する。現時点では研究段階や一部の限られたケースでの実証に留まるシナリオが多いが、将来的な AI の適用範囲は広く、例に挙げたような自動運転車を誤作動されるような人体の安全にかかわる重大なインシデントが発生する可能性を含んでいる。

②シナリオ 3-4：脅威に AI が関連するその他のシナリオは、従来のサイバー攻撃が AI 技術によって強化されるシナリオにあたる。サイバー攻撃者がその戦術や技術の一部として、AI を攻撃ツール開発に活用したり、既に開発された AI 攻撃ツールを利用したサイバー犯罪を展開したりするものである。標的システムへの侵入やデータの窃取・破壊といった目的や基本的な戦術はこれまでと同様であったとしても、AI を技術として利用することでその成功確率が高まる。実際に、ダークウェブで流通している非正規の生成 AI を使用して他言語のフィッシングメールをより自然な文章にしたり、マルウェア開発に AI を応用し不正な処理や通信を検知するセキュリティ保護技術を回避したりするという形で既に AI 技術が悪用されている。また、従来は難しかったヒトの認知プロセスに影響を与えるに足る品質の誤情報の拡散が懸念すべき脅威シナリオとなった点も生成 AI 技術の進展が影響しているといえる。

③シナリオ 5-6：資産に AI が関係するシナリオは、既存の手法を用いた脅威のターゲットとして AI システムを想定するものである。AI の資産としての価値が、新たな攻撃のインセンティブを与える可能性がある。企業の独自モデルが窃取の対象となったり、重要な AI システムを停止させるための妨害攻撃なども考えられる。ここで重要なのは、AI 資産へ被害を与える攻撃手法は、AI を利用したものでも、AI の脆弱性をつくものである必要もない点である。例にもあげたソーシャルエンジニアリング攻撃や DDoS 攻撃は、標的が AI 資産であること以外は従来のサイバー攻撃と変わるものではない。また、内部犯行の観点においても、先端通信技術などの例と同様に競合他社へリークなどの観点で AI 技術情報も資産価値が高いと考えられる。

④シナリオ 7-9：偶発的脅威・環境的脅威に関連するシナリオについては、AI サービスの利用者の不注意や AI システムの不具合、災害や地政学リスクなど意図的な主体による行動に基づかないセキュリティ脅威を挙げている。従来の情報セキュリティの枠組みにおいても、メールの誤送信や資料の誤廃棄といった運用のミスや IT ツールを安全でない設定のまま利用してしまったことによる情報セキュリティ事故が存在している。AI システムについても同様に、誤った使い方や不具合による情報漏えいが起こり得る。既に生じている事象としては、入力データを学習するタイプの生成 AI サービスに個人情報や営業秘密を入力してしまうケースや、生成 AI サービスの利用履歴を外部公開してしまうことによる漏えいなどが挙げられる。また、AI は各国で制度検討や規制が急速に進んでおり、国ごとの方針の差異もみられるため、地政学リスクも主要リスクとして含めた。

3. AI リスクシナリオへの対応戦略

ここまで、情報セキュリティリスクアセスメントの手法を適用して AI のセキュリティシナリオの分類とその分析を示してきた。本章では、これらの AI セキュリティリスクシナリオの特徴を考慮した上で、企業がセキュリティガバナンスの一環としてどのような対応していくべきかについて検討する。2 章で整理したシナリオ類型ごとに対応戦略をまとめる。

AI リスクシナリオへの対応の方針は、企業が AI にどのような立場で関わるかにより大きく異なる。総務省および経済産業省がとりまとめた「AI 事業者ガイドライン案[2]」では、様々な事業活動において AI を活用する事業者の類型を「AI 開発者」「AI 提供者」「AI 利用者」の 3 つに整理している。加えて、本稿で見てきた通り AI セキュリティを包括的に考えた際、事業者自身の AI の活用有無に限らず脅威主体による AI の悪用によるリスクにさらされることになる。これを「すべての事業者」として扱うことで計 4 つの主体を想定することとする。

シナリオ類型①「AI を熟知した攻撃者による AI 特有の脆弱性を悪用したサイバー攻撃」への対応

AI 特有の脆弱性を悪用するサイバー攻撃については、既に一部で実際の攻撃事例が見られる一方で、全体としては研究段階に留まる脅威シナリオの比率が高い。AI 自身がいまだ普及期にあり、攻撃者の視点からも AI のメカニズムを熟知し、新たな攻撃の戦術・技術・手順を確立するまでには、資金や人材などのリソースの問題や、その必要性（攻撃する価値のある標的としての AI システム数）の視点からも一定程度の時間が見込まれることが背景にあるだろう。そこで、AI を熟知した攻撃者による AI 特有の脆弱性を悪用したサイバー攻撃のシナリオに関して、既に顕在化した短期の脅威シナリオと将来的な AI の適用範囲を見据えた中長期のシナリオを区別して考える。

それぞれについての対応シナリオを以下に示す。

- **短期シナリオへの対応**

  現時点において AI 特有の脆弱性を悪用するサイバー攻撃については、生成 AI へのプロンプトエンジニアリング等が中心である。これらは既に顕在化した脅威であり、特に主要な生成 AI チャットなどの公開サービスで正規の応答では制限された内容を出力させようとするケースが主で、資産への影響は AI モデルの内部情報を取得しようとする直接的なケース、不正コードの作成やサイバー犯罪の標的のリサーチなどの AI サービスと無関係の被害者の資産への間接的な被害を生ずるケースがある。

  次の展開としては、主要サービス以外にも多くの生成 AI サービスが市場に増加していることを踏まえ、大小様々なサービスがこれらの攻撃の標的となる可能性がある。大きな資本をもつ主要サービスに比べてセキュリティ対策が不十分となりえるため、これまで以上に深刻な被害が個別のサービスの単位で生じることも考えられる。

  対策の方針について、「AI 開発者」の視点からは AI 開発をセキュリティ・バイ・デザインの原則にもとづいて実行することと、新しい脅威の情報を収集し、適時必要な更新を実施していくことが基本となる。「AI 提供者」は、これに加え外部モデルを採用する場合にセキュリティアセスメントを行うこと、AI モデルの直接的な対応以外にもシステム全体としてのセキュリティ対策による保護を行うことなどが求められる。例えば、既知の脆弱性が認められるモデルに対して入力の安全性検証を行うことや AI ファイアウォールによるフィルタリングなどが考えられる。より詳細な対策の実践については、前出の AI 事業者ガイドライン案における該当部分の記述も参考になる。

  「AI 利用者」と「すべての事業者」は、AI モデルもしくはシステムを直接的に保有しないため AI システムの脆弱性を有さず、このシナリオでは間接的な被害のみが懸念事項となる。利用中のサービスのセキュリティ対策や

[2] https://www.soumu.go.jp/menu_news/s-news/01ryutsu20_02000001_00009.html

AI を利用した新しい攻撃に関する脅威インテリジェンスを収集して自社システムの対策に反映するといった対応が考えられる。

- **中長期シナリオへの対応**

中長期の視点にたつと、文章生成系 AI 関係のシナリオ以外にも、現在は多くが研究段階にある敵対的 AI 技術やデータポイズニングによって AI のモデル自体に作為的な影響を及ぼそうとする攻撃が実環境で蔓延する可能性を考慮する必要がある。サイバー攻撃のリスクは、脅威主体である攻撃者にとって、それを行う合理性のある領域で最も深刻化する。社会における AI の普及に比例する形で攻撃者にとっての標的としての魅力は増す。例えば、ランサムウェア犯罪グループがシステム停止による社会影響がより大きい標的から多くの身代金を得やすいことを認識し、重要インフラ組織が標的になるケースが徐々に増えてきた歴史を考えると分かりやすい。

現在の AI システムの普及を概観すると、ウェブサービスでの推薦システムや企業用の業務アプリにおける需要予測システムや画像解析システムの不良品検出への利用などが先行し、直近では生成 AI を組み込んだサポートセンターアプリやオフィス業務支援アプリなどの利用が急速に拡大している。一方で、よりフィジカルなシステムや生活に直結するサービスへの AI 実装（たとえば製造業での AI ロボティクス・金融審査業務・行政サービス・インフラメンテナンス・自動運転車など）の本格化はこれからという見通しだが、様々なステークホルダーによる実証実験やスタートアップ企業のサービスは増加してきている。

攻撃者の目線からは、AI がより重要な社会サービスで使われるようになるにつれ、より標的とするインセンティブが増すことになる。現在は、技術的な難易度が高く、攻撃ツールとしての整備もさほど進んでいない AI 固有の脆弱性を悪用する攻撃も、メリットがコストを上回るタイミングで急速に環境が変化する可能性がある。

対策の方針としては、「AI 開発者」「AI 提供者」の視点からは短期のシナリオと同様にセキュリティ・バイ・デザインなどの普遍的な原則と既にガイドライン化されているベストプラクティスの実装に取り組むことが第一に考えられる。加えて、今後新しい攻撃の戦術・技術・手順が多く出現することが予想される領域であることから、積極的な情報収集が重要となるだろう。現状における生成 AI セキュリティの技術体系整理や脅威の類型の整理については、米国国立標準技術研究所（NIST）による NIST.AI.100-2[3] や、米 MITER 社による ATLAS フレームワーク[4] 、OWASP による機械学習セキュリティ TOP10[5] などが参考になる。国内でも情報処理推進機構（IPA）が、AI セキュリティに関する取り込みをまとめたページを公開している[6]。

また、セキュア開発サイクルを確立しておくことで、新しい脅威情報を既存のプロダクトによりスムーズに反映することが可能になる。「AI 利用者」と「すべての事業者」においても、間接的なリスク保有者の立場からも大きな影響を生じる可能性のある情報を収集し、利用するサービスの選定やサイバー攻撃環境の変化について把握していることが望ましい。

シナリオ類型② 「AI をツールとして利用し、AI 以外の脆弱性を突くサイバー攻撃」への対応

2 番目に挙げるシナリオは、主に標的型攻撃グループやサイバー犯罪グループが、AI をツールとして利用することで AI 以外の脆弱性を突く攻撃シナリオをより効果的・効率的に実施するタイプのシナリオである。前章のシナリオ整理結果からも分かる通り、この類型のシナリオは既に脅威として顕在化した例が多い。これは、シナリオ類型①にあるよ

[3] https://csrc.nist.gov/pubs/ai/100/2/e2023/final
[4] https://atlas.mitre.org/
[5] https://owasp.org/www-project-machine-learning-security-top-10/
[6] https://www.ipa.go.jp/digital/ai/security.html

うな AI のメカニズムを熟知する必要がある攻撃が現状ではコスト高であることに対し、既存の攻撃シナリオの延長線上での AI の活用は、戦術的な変更が小さく、新たな技術や手順の採用の一環とみることができるため、低リスクで短期の投資対効果を見込めるアプローチであるが背景にある。ここでは、従来の IT システムの脆弱性をより効率的に悪用する攻撃のシナリオと、生成 AI ツールによってこれまでは異なる水準でヒトの認知プロセスの脆弱性が突かれるシナリオについて分析する。

それぞれについての対応の方向性を以下に示す。

- **従来の IT システムの脆弱性をより効率的に悪用するサイバー攻撃シナリオへの対応**

  このシナリオで特徴的な点は、AI があくまで攻撃者のツールとして利用されているため、対象となる脆弱性や資産について AI の関与有無は問わず、「全ての事業者」に対する脅威として存在することである。これまでのサイバー攻撃でも用いられた不正アクセスやマルウェア攻撃、サーバ侵害等のための技術・手順の一部に AI を利用することで、より防ぐことが難しい攻撃手法が出現する可能性がある。

  一方で、事業者の対策の方針も脅威インテリジェンスに基づいたリスク適応的な対応の延長線上で考えることができるだろう。情報収集の対象としている脅威アクターが AI 攻撃ツールを用いるようになった場合、脅威の分析と対応の検討のサイクルの中に AI をツールとして利用した攻撃も自然と含まれることになる。

  また、ツールとしての AI 活用は防御側でも進んでおり、「AI for Security」の取り組みの中で実用化された技術的対策も多い。高精度でのマルウェア検出や振る舞い検知、生成 AI を活用した生産性向上など様々な種類の活用が見られる。攻撃側の AI 利用によって従来の枠組みでの対策が難しい攻撃が発生した場合に備え、AI を利用したセキュリティ対策ツールの活用も選択肢のひとつとして考えられる。

- **生成 AI ツールによる偽情報の拡散シナリオへの対応**

  ヒトの認知の脆弱性を突いた偽情報拡散のリスクは、生成 AI の登場により大きく侵攻した脅威シナリオである。生成 AI が現実には存在しない文章や画像、映像等を本物と見分けがつかないような精度で作成可能であることから、ツールとして悪用することで、偽情報を参照にした者を不正な意図に基づいた認知へ誘導することができる。国家間で自国に有利な世論形成のために行われる認知戦や愉快犯による事件などが主だった事例として挙げられるが、公的な主体のみならず、重要政策の実行に関係する企業や認知の高い企業なども巻き込まれることを想定すべきである。

  この時、ツールとしての生成 AI の役割は、不正なコンテンツの作成であり、不正侵入のサイバー攻撃との対比でいえばマルウェア製造を補助しているような関係になる。配布の手段としては、水飲み場攻撃に近く、直接被害者の環境を攻撃するのではなく、多くの人が訪れるウェブサイトや SNS などが配布対象となる。しかし、脆弱性が情報システムの脆弱性ではなく、ヒトの認知であるため、マルウェアのダウンロードやインストールのような課程は必要なく、閲覧者が内容を正しいものと信じさえすれば攻撃者の目的が達成される点に注意が必要である。

  対策として最も直接的なものは不正なコンテンツの流通を抑制することであり、捜査当局による摘発やプラットフォーマーによる不正コンテンツの削除などが挙げられる。しかし、多くの企業からすればこれらは攻撃な規制や他の企業への期待であり、偽情報の脅威に自身で事前の抑止対策を講ずることの難しさを示している。そこで、多くの企業では、偽情報シナリオの発生を想定した有事対応シナリオを用意しておくことが最も現実的な策となるだろう。これに関連して、米国サイバーセキュリティ・インフラセキュリティ庁（CISA）が公表しているディープフェイク

への対策をまとめた資料[7] や総務省で開催されているプラットフォームサービスに関する検討会資料[8] などが参考になる。

シナリオ類型③「AI 資産を保有する者に対する従来のサイバー脅威」への対応

3 つ目のシナリオは、AI 資産を保有するものに対する従来の手法を用いた攻撃による脅威についてである。2 章では、AI 資産の価値は高く、従来型の脅威アクターにとっての標的として優先度が高いものになる見方を示した。このシナリオでは、脅威主体や悪用される脆弱性は AI 特有の要素が関係しない従来のものである点に特徴がある。

AI システムも、その他の情報資産と同様に IT インフラ上で稼働し、ヒトの手によって運用されることから、従来の IT システムや情報資産管理の運用と同様の脆弱性が存在することとなる。同時に、脆弱性の悪用のシナリオも従来のものと基本的な性質が同質のものとなることから、対応の方針も従来の情報セキュリティ管理運用の延長線で考えることが出来る。つまり、AI 資産を保有する「AI 開発者」「AI 提供者」においても、AI セキュリティ対策のみならず、情報セキュリティ対策のベストプラクティスに従った対策が求められることになる。これには、情報セキュリティマネジメントシステムや NIST サイバーセキュリティフレームワークのようなベストプラクティスの活用に効果が期待できるだろう。そのうえで、AI 資産保護に必要な対策を調整することで対応することが望ましい。

ここで特に考慮すべき点として、AI 資産のリスクアセスメントが考えられる。自社の AI 資産の価値を AI の特性を考慮した形で評価し、必要なレベルの保護を与えることが重要となる。AI の学習に用いるデータの性質や独自のノウハウの資産としての重要性を評価し、AI システムを開発・提供する環境や保存場所等を加味したリスク評価を行い、必要な対策の水準を決定する。AI システムの開発や提供は先端研究開発や新規事業の枠組みでの取り組みとされ通常と異なる業務規定で扱う企業も多いが、その中であっても必要な水準の対策が確実に実装されることが望ましい。

シナリオ類型④「AI サービスの利用者において意図せず発生するミスや環境変化」への対応

AI サービスの利用者において意図せず発生するミスや環境変化（偶発的脅威・環境的脅威）に関連するシナリオについては、AI の適切な利用方法が整理されていないことによって生じるリスク、国内外の規制によって技術の利用環境が変わる可能性などを挙げている。新しい技術の普及期には常に考慮されるべきリスクシナリオであるが、AI の場合はデータの学習により成立する技術であるという特性が重要となる。

情報セキュリティの観点からは、利用者のデータが AI サービスの学習に利用されるか否かが大きな論点であり、AI に入力するデータの質に応じたルール設計が目下の大きな課題となっている。この他にも、AI の類を見ないほどの進化と普及の速度、国際的な規制の傾向といった要素が大きく影響する。企業としては、変化の速い環境の情報を収集し、自社のポリシーや運用ルールを適切に追従させる形で対応することが求められる。

具体的には AI の特性を踏まえた上で、AI 利用ガイドラインにセキュリティ上の遵守事項をふまえたルールを作成するなど、企業のガバナンスのなかで AI セキュリティリスクを扱っていくことが望ましい。この点については、既刊の JCIC レポート「企業が生成 AI の奔流を乗り越えるためのアジャイルリスク管理」[9]にも解説しており、是非参照いただきたい。

[7] https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF
[8] https://www.soumu.go.jp/main_sosiki/kenkyu/platform_service/02kiban18_02000283.html
[9] https://www.j-cic.com/pdf/report/Generative-AI.pdf

4. まとめ

本レポートでは、AI とセキュリティに関係する論点をセキュリティリスクアセスメントの方法論を用いて、体系的に整理した。本編の要旨を提言としてまとめると次の 3 点となる。

提言 1.　AI セキュリティをリスクの枠組みで捉えなおし、脅威と脆弱性を明らかにすべきである

提言 2.　新しい領域と従来の延長線上にある領域を見極め、適材適所の対応を設計すべきである

提言 3.　中長期の視点で AI とセキュリティのデュアル・メジャー人材を育成すべきである

提言 1 について、AI セキュリティは AI とセキュリティが相互に関係する複雑な領域であり、一貫したものさしで全体像をとらえないことには適切な対策を形にすることはできない。セキュリティリスクアセスメントは、脅威・脆弱性・資産の視点からリスクシナリオとその対策を検討するための普遍的な枠組みであり、AI とセキュリティの論点とリスクシナリオを俯瞰的に論じるために適している。本レポートでは、AI が脅威に与える影響、AI 特有の脆弱性、AI の資産としての価値という個別の要素ごとに AI の影響を分析することで、より具体的な対応の方針を示した。

提言 2 について、本レポートでは脅威の展望として、AI 特有の脆弱性を悪用するサイバー攻撃のような高度な脅威は現時点で頻繁にみられるものではないが、中長期では AI の普及発展にともない攻撃者のインセンティブは確実に増してくると予想した。一方で、AI システムに対する攻撃は AI の脆弱性をつく攻撃に限らず AI が IT システム基盤や運用や管理に携わるヒトの脆弱性を狙うものがあること、脅威主体による AI の悪用は標的組織の AI を利用に関わらず脅威となることを確認した。これらは既に顕在化した脅威であり対処の必要があるが、脅威インテリジェンスの活用や組織のガバナンスといった既存の対策の拡張による対応の可能性を紹介した。脅威の質と現況を理解することで、限られた対策リソースを効率的に割り当てることが出来るだろう。

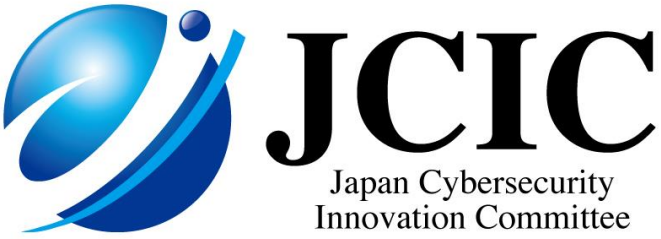
提言 3 について、AI セキュリティはまだ新しいテーマであり、本レポートで触れた以上の新しい検討事項が今後も次々と現れてくるだろう。セキュリティ人材や AI 人材が不足している現状にあって AI セキュリティ人材を育成することはさらに厳しい取り組みとはなるが、だからこそ限られた人材が力を発揮できるよう環境が整備されている必要がある。AI セキュリティ人材への最も大きな期待は、AI とセキュリティの双方に精通し、優先して対処すべき脅威の特定や次に必要になる対応への準備といった見通しを示すことになるだろう。提言 1 および提言 2 において、AI とセキュリティを俯瞰するリスクアセスメントの視点や適材適所の配置についての考えを述べた。AI セキュリティ人材が方向性を示すことで、AI 人材やセキュリティ人材がそれぞれの領域で AI セキュリティ対策を実現する方向づけが重要になる。

最後に、セキュリティの本質を脅威からの保護と考えれば、保護すべき対象に対する深い理解は不可欠である。AI 原則の視点からも、倫理や透明性といった適正なあり方を論ずる他の原則とは異なり、セキュリティの原則は悪意のある行為やその他の脅威を議論の前提とする点にその特徴がある。極論すれば、AI の利用環境が侵害されれば他の原則は基盤そのものが失われることになり、その点において特異である。本レポートが、AI の専門家とセキュリティの専門家の双方にとって AI セキュリティの全体像を共有し、相互の交流や分野の発展の一助となれば幸いである。

以　上

[表 1 の参照文献]

1. 総務省 X MBSD, AI セキュリティ 情報発信ポータル, https://www.mbsd.jp/aisec_portal/
2. SOMPO CYBER SECURITY, プロンプトインジェクションとは, https://www.sompocybersecurity.com/column/glossary/prompt-injection
3. Reza Shokri, et al., Membership Inference Attacks against Machine Learning Models, https://arxiv.org/abs/1610.05820
4. Chawin Sitawarin, et al., DARTS: Deceiving Autonomous Cars with Toxic Signs, https://arxiv.org/abs/1802.06430
5. Chen Zhu, et al., Transferable Clean-Label Poisoning Attacks on Deep Neural Nets, https://arxiv.org/abs/1905.05897
6. OpenAI, Disrupting malicious uses of AI by state-affiliated threat actors, https://openai.com/blog/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors
7. Trend Micro, 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, https://www.trendmicro.com/ja_jp/research/23/i/hype-vs-reality-ai-in-the-cybercriminal-underground.html
8. Mahmood Sharif, et al., Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition, https://users.ece.cmu.edu/~lbauer/proj/advml.php
9. Bojan Kolosnjaji, et al., Adversarial Malware Binaries: Evading Deep Learning for Malware Detection in Executables, https://arxiv.org/abs/1803.04173
10. NSA FBI CISA, Contextualizing Deepfake Threats to Organizations, https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF
11. みずほリサーチ＆テクノロジーズ株式会社, ディープフェイクについて（総務省 プラットフォームサービスに関する研究会）, https://www.soumu.go.jp/main_content/000749422.pdf
12. JETRO, EU、AI を包括的に規制する法案で政治合意、生成型 AI も規制対象に https://www.jetro.go.jp/biznews/2023/12/8a6cd52f78d376b1.html



[本調査に関する照会先]

客員研究員　古澤一憲　furusawa@j-cic.com

JCIC 事務局　info@j-cic.com