AI 時代のサイバー攻撃と防御

― 生成 AI が変える脅威と組織に求められる新常識 ―

1. AI 普及がもたらす新しい脅威環境

生成 AI の急速な普及により、企業や自治体が扱うデータ量は飛躍的に増加し、業務効率化が進む一方で、データは複数のクラウドやサービスを横断して流通するようになった。これにより、従来の境界型セキュリティでは把握しきれない領域が拡大し、情報管理の難易度は大きく上昇している。

警察庁は「サイバー空間をめぐる脅威の情勢」において、不審アクセス件数が依然として高水準であること、そして生成 AI を悪用した攻撃事案が確認されていることを報告している。ランサムウェア被害も最多級で、攻撃の自動化・高速化が進んでいる点が指摘されている。

IPA の「情報セキュリティ 10 大脅威 2026」(*1) では、「AI の利用をめぐるサイバーリスク」が組織向け脅威の上位に位置づけられた。また、AI 利用リスク調査では、多くの企業が AI 利用に伴う脅威を十分に把握できていない実態が示されている。

内閣サイバーセキュリティセンター（NISC）も国家戦略の中で、AI が脅威環境を変化させていると明記している。AI が攻撃者の能力を底上げする一方、防御側にも AI を活用した新たな対策が求められている。

本レポートでは、AI 時代に特徴的な脅威を整理し、組織が取るべき防御の方向性を考察する。

(*1)情報セキュリティ 10 大脅威 2026｜情報セキュリティ ｜IPA 独立行政法人 情報処理推進機構

2. AI 時代のサイバー攻撃の特徴

2-1. 攻撃の高度化と自動化

生成 AI は、攻撃コード作成や脆弱性探索といった専門的作業を自動化し、攻撃者の技術的ハードルを大幅に下げている。AI エージェントがネットワーク構成を分析し、最適な侵入経路を自律的に選択するなど、攻撃シナリオの自動生成も現実化している。マルウェアの変種生成も容易となり、シグネチャ型防御では検知が難しくなっている。

2-2. ソーシャルエンジニアリングの精密化

生成 AI は、ターゲットの属性や過去の文面を模倣した自然なフィッシングメールを大量に生成できる。ディープフェイク技術（*2）の進展により、経営者の声を模倣した送金指示詐欺など、内部統制を揺るがす攻撃も増加している。SNS 情報を AI が分析し、個人の行動パターンに合わせて攻撃を仕掛けるケースも確認されている。

2-3. データ流通の増加が生む新リスク

AI サービス間の API 連携が一般化し、データが複数クラウドを横断して移動することで、データの所在管理が難しくなっている。生成 AI サービスへの入力データからの情報漏洩も懸念され、業務文書や顧客情報が誤って外部に流出するリスクが高まっている。また、AI モデル自体が攻撃対象となるケー

スも増え、プロンプトインジェクション（*3）による情報漏洩の可能性が指摘されている。

2-4. 誰でも攻撃者になれる構造変化

AI により、従来は高度な技術が必要だった攻撃が初心者でも実行可能になりつつある。攻撃ツール生成や脆弱性悪用方法の取得が容易になり、攻撃者の参入障壁が低下している。その結果、攻撃の総量が増加し、組織が直面する脅威の幅が広がる可能性が高い。

2-5. 攻撃のスピードと規模の拡大

AI は大量のフィッシングメール生成や複数経路からの同時攻撃を短時間で実行できる。自動化された攻撃は人手による対応を上回る速度で進行し、被害拡大前の検知・遮断が難しくなる。攻撃規模も拡大し、防御体制への負荷が増している。

(*2) ディープフェイク技術は、深層学習（AI）が人物の顔や声を学習し、本物そっくりの偽映像・偽音声を自動生成する技術である。従来の手作業によるフェイクより精度と再現性が圧倒的に高く、見破りが難しい。結果として、本人確認や情報の信頼性を揺るがす新たなサイバーリスクとなっている。

(*3) プロンプトインジェクションとは、生成 AI に与える指示文を悪意ある形で操作し、AI に意図しない動作をさせる攻撃である。直接不正な指示を送る方法に加え、外部データ内に指示を埋め込む「間接型」が特に危険とされる。これにより、AI が内部情報を漏洩したり、業務プロセスが誤動作する可能性がある。

3. 攻撃事例

3-1. AI 生成メールによるフィッシング

企業の文体や担当者の書き方を模倣した自然なフィッシングメールが大量に作成されている。SNS 情報を分析し、受信者の役職や関心に合わせて文面を最適化する手法も増加している。こうした手口により、利用者の注意力や教育に依存した対策では対応しきれず、人の判断を前提とした防御の見直しが必要となっている。

3-2. AI チャットボットの悪用

正規企業の問い合わせ窓口を模倣した偽チャットボットが登場し、自然な対話を通じて個人情報や認証情報を入力させる手口が確認されている。対話型であるため警戒心が薄れやすく、被害が拡大している。このような攻撃により、正規サイトかどうかを人が見分ける前提の本人確認や注意喚起が機能しにくくなっている。

3-3. プロンプトインジェクション攻撃

AI に与える指示文に不正な命令を埋め込み、意図しない動作をさせる攻撃が増加している。特に Web ページや画像内に隠された命令を AI が読み取り実行する「間接型」が深刻で、誤案内や内部情報漏洩につながる事例が報告されている。

その性質上、従来の入力チェックやアクセス制御では防ぎにくく、AI 連携業務そのものがリスク源となる。

3-4. ディープフェイク音声による送金指示詐欺

経営者や CFO の声を AI で模倣し、経理担当者に緊急送金を指示する詐欺が世界的に増加している。数十秒の音声から声質を再現できるため、本人確認が困難になっている。結果として、音声による指示

や上司確認を前提とした業務フロー自体の見直しが求められている。

3-5. AI エージェントによる脆弱性探索の自動化

AI エージェントがネットワーク構成を解析し、攻撃可能なポイントを自律的に探索する手法が広がっている。公開情報や設定情報を統合分析し、最適な侵入経路を提示するケースも確認されている。これにより、脆弱性が発見されてから攻撃までの時間が短縮され、対応の遅れが即被害につながるリスクが高まっている。

4. AI 時代に露呈した従来防御の限界

4-1. シグネチャ型防御の限界

既知マルウェアの照合に依存するシグネチャ型防御は、AI が自動生成する変種マルウェアに追いつけない。基礎的防御としては有効だが、単独では不十分である。

4-2. 境界型セキュリティの弱体化

クラウド利用や SaaS 連携が進む現在、境界型セキュリティの前提は崩れつつある。データが社内外を横断するため、「社内＝安全」という考え方では防御が成立しない。この構造変化により、境界の内外を分けるだけの対策では侵入後の被害拡大を防げなくなっている。

4-3. パスワード依存の脆弱性

AI 生成フィッシングや偽チャットボットにより、パスワード窃取が容易になっている。ID・パスワード単独の認証は限界に達しており、多要素認証やデバイス認証が不可欠である。

4-4. ログ監視・手動対応の限界

AI による攻撃は高速かつ大量で、人手によるログ監視やアラート対応では初動に追いつかない。異常検知 AI や自動レスポンスを組み合わせた体制が求められる。

4-5. ゼロトラストの必要性

AI 時代の防御では「何も信頼しない」を前提とするゼロトラストが中心となる。ユーザー・デバイス・通信ごとに継続的な検証を行い、侵入後の横移動を防ぐ仕組みが重要である。

4-6. AI を活用した防御への転換

攻撃者が AI を使う以上、防御側も AI を積極的に取り入れる必要がある。異常検知 AI は振る舞いの違和感を捉え、従来手法では見逃す攻撃を早期に検知できる。ログ分析や初動対応の自動化により、SOC の負荷を軽減できる。

これらの限界と、AI 時代の代替アプローチを以下の表に示す。

表. 従来手法の限界と AI 時代の代替アプローチ

従来手法の限界	AI 時代の代替アプローチ
シグネチャ型防御の限界	振る舞い検知・AI 分析
境界型の弱体化	ゼロトラスト
パスワード依存	MFA・デバイス認証
手動 SOC の限界	SOAR・自動レスポンス

・シグネチャ（signature）：マルウェアや攻撃の“特徴パターン”をデータ化したもの

・MFA（Multi-Factor Authentication）：多要素認証
・SOC（Security Operation Center）：組織内のセキュリティ監視・分析・対応を担う「運用の司令塔」
・SOAR（Security Orchestration, Automation and Response）：SOC の作業を自動化するための仕組み

5. 公共機関・自治体が直面する AI 時代の現実

5-1. レガシー環境と複雑なシステム構成

自治体はレガシーシステムと新規クラウドが混在し、統一的なセキュリティ管理が難しい。古い基盤が全体の弱点となりやすく、脆弱性が長期間放置されるリスクが高い。

5-2. 人材不足と運用体制の脆弱さ

多くの自治体ではセキュリティ専任者が不足し、高度化する攻撃に対応できる体制が整っていない。地方ほど専門人材の確保が難しく、インシデント対応が属人的になりがちである。

5-3. データ連携拡大によるリスク増大

住民情報が複数組織・クラウドを横断して流通する中、AI サービスへの入力ミスや外部サービスの設定不備が情報漏洩につながるリスクが増加している。

5-4. サプライチェーンと住民サービスの脆弱性

自治体システムは多くの委託事業者に依存しており、サプライチェーン全体のセキュリティが弱点となる。偽チャットボットやディープフェイクによる住民詐欺も増え、利用者保護と利便性の両立が課題となっている。

6. AI が加速させる攻撃と防御の時代をどう生き抜くか

AI の進化はサイバー攻撃と防御の両面で決定的な影響を及ぼす。

攻撃側では、生成 AI や自律型エージェントの高度化により、脆弱性探索・攻撃コード生成・攻撃シナリオ構築が自動化され、攻撃準備の高速化と成功率の向上が進む。ディープフェイク技術の精度向上により、音声・映像を用いたなりすましが常態化し、組織内部の意思決定プロセスを揺るがすリスクも増大する。

防御側では、従来手法の限界が明確となり、AI を活用した振る舞い検知やゼロトラストへの移行が不可避となる。SOAR による自動レスポンスが普及し、初動対応は AI が担う構造へと変化していく。防御の中心は「AI 対 AI」へと移行し、組織は AI 前提のセキュリティ運用体制を整備する必要がある。

公共機関・自治体も例外ではなく、住民サービスの利便性向上と安全性確保の両立が求められる。重要なのは、組織の種類を問わず、すべての主体が AI を前提としたリスク管理へ転換できるかどうかである。

今まさに「AI が攻撃と防御の両方を加速させる時代」である。技術・人材・制度を一体で整備し、AI を脅威としてだけでなく防御の基盤として活用できる組織こそが、次の時代の安全性を確保できる。私たちが問われているのは、「AI をどう恐れるか」ではなく、「AI をどう使いこなし、未来の安全を築くか」という姿勢である。

以上