Security for AI と AI for security

System Security & Network Department, High Performance Computing Core Technology Division, High Performance Computing Product Group

Renesas Electronics Corporation

森山　大輔

イントロダクション

2022 年から ChatGPT の登場により、AI に対しての世の中の捉え方は大きく変わった。2010 年代においても、IBM Watson(*1)や Google AlphaGo (*2)のように AI の進化が人間の論理的思考を超えるかという問いに対する研究および開発が行われてきた。これらの結果から数年たった現在、AI は生成 AI の登場により誰もがその将来的な可能性を体験できるものに到達している。大手 IT 企業や AI 新興企業などが盛んにチャット形式で LLM (Large Language Model) にアクセスできる環境を構築し、対話・プログラミング・画像生成機能等のサービスを消費者に提供している。

多くの企業が AI の進展がビジネスの拡大や効率化に寄与するだろうと期待を寄せている。PricewaterhouseCoopers (*3)による 4702 人の CEO を対象とした 2023 年のサーベイでは、64%以上の人が従業員の業務効率の改善に繋がり、59%が自身の業務効率の改善に繋がると答えている。一方で、生成 AI によるリスクの中で、サイバーセキュリティを挙げた人が一番多く 64%であった[1]。2023 年に 300 人以上のリスクやコンプライアンスの専門家を対象にした別の企業によるサーベイでは、生成 AI に対してのリスクが存在することを 93%の企業が認識しており、一方でリスク対策を行う用意が出来ていると答えた企業は 9%に留まるとの結果が得られている[2]。また、セキュリティの専門家 1123 人に対して ISC2 (International Information System Security Certification Consortium) (*4) が行ったアンケートでは、AI がサイバーセキュリティに良い影響を与えるかという質問に対して賛成したのは 28%のみであり、反対は 37%であった[3]。実際に、[3]の別のアンケートでは回答した人のうち 12%は業務においてすべての生成 AI ツールの利用を禁止しており、32%は一部の生成 AI ツールの利用を禁止しているという結果になっている。

AI の開発・導入とは別に、AI の柔軟性を活用することで、セキュリティオペレーションセンターの効率化や脅威の検出・対応の自動化に利用することが期待されている。一方で、AI は防御側のみに貢献するものではない。大手企業による AI サービスでは有害な入出力はなるべく行われないように適切にトレーニングされているが、ゼロから構築されたサイバー攻撃のための AI ツールがハッキングコミュニティに流通していることが発見されている。

本ホワイトペーパーでは、特に企業内での AI 利用に焦点を当てつつ、ライフサイクルを含めた AI そのものに対してのセキュリティ (Security for AI)と、既存のセキュリティに対して AI が利用される場面(AI for Security)における議論点の双方の AI security について解説する。また、AI に対してのリスクにどのような対応を行うべきかについての各国政府の動きや、業界団体の結成の動向や目的について紹介する。

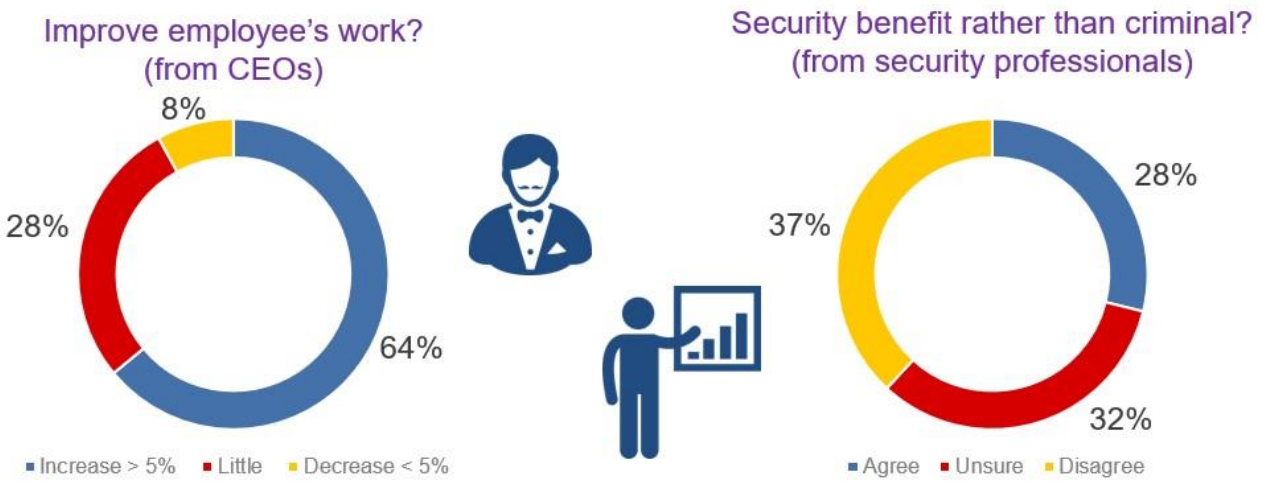

図 1 AI に対する期待と懸念

(*1) IBM Watson は世界の多くの国で登録された International Business Machines Corp. の商標です。
(*2) Google, AlphaGo は Google Inc.の商標です。
(*3) PricewaterhouseCoopers は PricewaterhouseCoopers International Limited.の商標です。
(*4) ISC2 は International Information Systems Security Certification Consortium の商標です。

AI 開発ライフサイクルとセキュリティ

まずは AI が開発されユーザに利用されるまでの流れについてまとめる。AI が学習を行わせるためのアルゴリズムの一つが、一般にディープラーニングと言われるものである。CNN (畳み込みニューラルネットワーク)や RNN (再帰的ニューラルネットワーク)は昔から使われてきたアルゴリズムである。2017 年には Transformer という精度の高い学習を行うことができるアルゴリズムが登場しており、Transformer は 2024 年時点で提供されている多くの AI サービスで利用されている。

学習を行わせるための元情報がどれかは、AI による内部判定や最終回答に強く影響を与える。チャットボットのような汎用的に利用する AI モデルを生成する場合、主にインターネットから公開されているニュース記事、画像、動画、データベース、ソースコード、学術文献などを学習素材とする。データを分類するために、事前にカテゴリ情報などをラベル付けするのが LLM 開発では一般的である。本ホワイトペーパーでは対象外とするが、この時の著作権に関する問題はそれぞれの国において倫理面や法律面含めて議論が盛んにおこなわれている。一定のタスクを人間に置き換えて実行する特定用途向けの AI モデルを生成する場合は、その処理に直接的に関連しているデータを入力することが望ましい。

学習のためのアルゴリズムと入力データから、学習モデルの生成がスタートする。近年の AI サービスでは学習データがテラバイト単位であり、そこからさまざまな解析を行っているため LLM (Large Language Model)と呼ばれる。LLM の中にはより高い精度での出力を行うために、ノイズを混ぜた上で再学習を実行するものや、あるいは正確性を向上させるために人間が直接ガイドして指示する仕組みを有しているものも存在する。ただし、特定のデータに特化させて学習させすぎた場合、過学習と呼ばれる現象が起こる。その場合、学習データに近いものに対しては高い精度の出力を行うことができるが、未知のデータに対しては精度が低くなる。もちろん学習時間が短いとどの場合においても精度が低い結果になるため、現時点では人間が LLM の学習時間を上手く調整する必要がある。実際に LLM をサービスとして展開する前までに、Validation data を利用して制御パ

ラメータのチューニングを行うほか、Test data を用いて出力精度の評価を行う（Validation data と Test data の役割は異なる）。汎用的に用いられる LLM に対してのベンチマークソフトとしては MMLU、GSM8K、MATH、BBH などが存在し、パフォーマンスや汎用性が評価される。

最終的に、開発された LLM が十分に利用できると判断されれば、実運用フェーズに入りサービスとして展開される。エンドユーザが LLM とやり取りすることができるのはこの段階になる。ソフトウェアコーディングによって定義された確定的な出力を行う既存のプログラムと大きく異なるのは、LLM では運用段階でユーザが与えた情報が学習素材としてフィードバックされ再学習に利用される可能性があることである。理論的には、学習を続けていくことで持続的により良いモデルへと LLM が成長することが可能である。

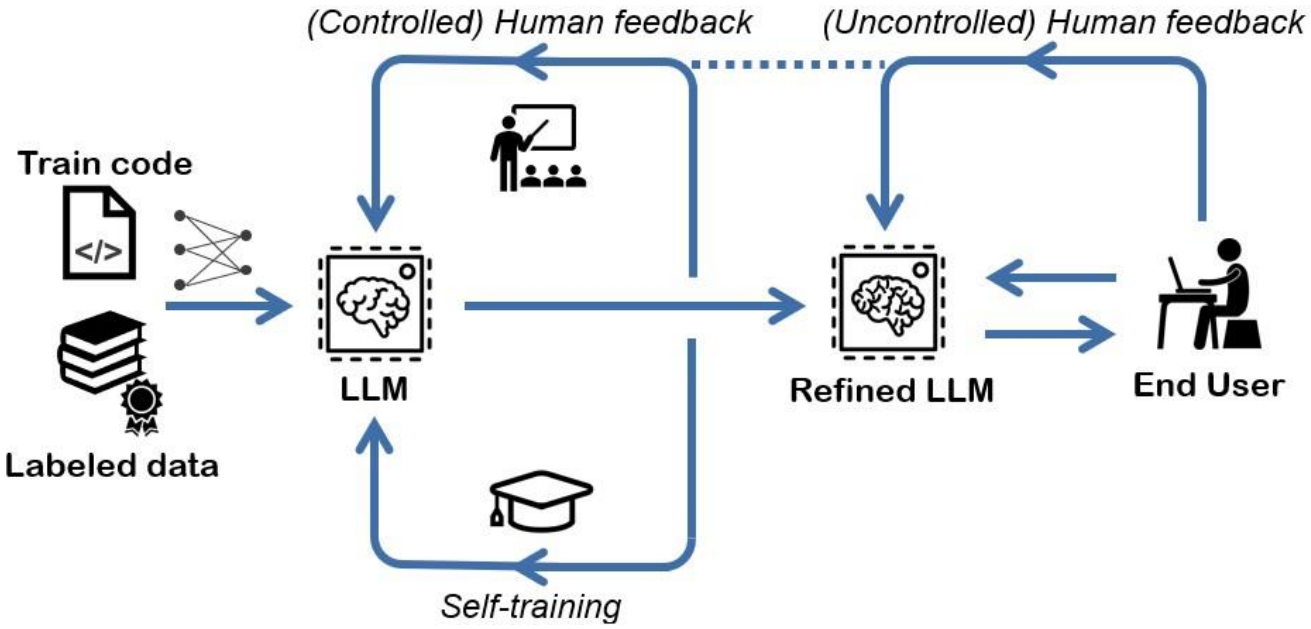

図 2 LLM 開発のライフサイクル

この AI 全体のライフサイクルの順番に沿って、セキュリティの面に焦点を当てて議論を進める。AI には学習用データが必要不可欠であり、そしてラベル化されているデータは信頼されたデータとして平等に取り扱われる。一方で、人間が普段生活している時、常にどの受け取った情報も平等に脳内に記憶・処理することは難しい。また明示的に公表されていなくても、それぞれの情報ソースに対する信頼性を大抵の場合は推定している。例えば政府系機関が発信している情報について疑う人は極めて少ないであろう。また主要マスメディアによる情報発信は、インターネットによる情報よりも信頼されることが多い。また、個人が情報発信を行うことができるソーシャルメディア内でも著名人の発言は信頼された情報と解釈されることが多い（真贋はまた別であるが）。それらの情報元自身の信頼性や影響力について、人間が意識しているものと同様に AI に学習させるべきかについては議論の余地があると思われる。少なくとも、AI がノイジー・マイノリティによる偏った情報を一般的な意見であると誤解し学習することは避けるべきである。また、情報ソースがインターネットで公開されている場合であっても、ハッキングや設定ミスにより漏洩した機密情報や個人情報といったものは学習されないようにフィルタリングする必要がある。

それぞれの企業内で AI サービスを従業員向けに展開する場合は、外部ソースとは別に期待する用途に合わせて社内や特定の部署内で共有されている知識ベースが学習データに含まれる。その場合、学習データに対して定

義されているアクセスポリシーに従って権限のあるユーザのみが AI サービスにアクセスできるようにしなければいけない。LLM 自体は機密情報を保護する機能を備えていないため、LLM 経由での情報漏洩が起こりうるということを考慮する必要がある。特にカスタマーサポート等の消費者との接点を持つ AI 利活用の場合には、真に社外に対して公開してよい情報のみに限定した学習を行わせる必要がある。社内に限定されている機密情報を含む LLM の公開は、情報漏洩インシデントを起こしているのと同様である。LLM は入力データをそのまま保管しているわけではないが、展開時点でどのような抽出可能な機密情報が含まれているのかを第三者が監査することは非常に難しい。そのため学習時の情報統制は厳密に行われることが望ましい。

LLM を一から構築することは多くの非 IT 企業にとっては負担が多い。それらの企業がビジネスにおいて AI を導入するための一つの方法は、ベースとなる LLM をカスタマイズすることである。ベースの LLM は大手 IT 企業や AI 特化型の新興企業などが十分な品質ものを提供することが可能だが、期待している利用方法にカスタム化した結果が適合しているかは個々の企業がリスク分析も含めて評価しなければいけない。特に現状の Generative AI において議論の的になっているのはハルシネーションである。人間に対してある質問をした時、その人が答えを知らないのであれば「知らない」と回答することが正しい返答である。一方で Generative AI は創造性を有しているために、正しくない・存在しない事柄をあたかも正しい答えであるかのように返答することが時々起こりうる。嘘をつくことがある従業員と良い関係を築くことは難しいように、AI と対話する場合には何が正しい情報であるかを見極める必要がある。用途によっては、AI による創造性に対して一定の制限を課し、トラブルが起こりにくいようにすることも議論の余地があるであろう。

サービスとして展開した後に、エンドユーザの入力をモデルの改善のためにフィードバックするかどうかは、提供側と利用する側の双方に対するセキュリティ問題の非常に大きな分かれ目となる。ユーザに展開する前のフェーズでは、すべて学習させるためのデータは開発者側によって管理できる状態にある。一方で、エンドユーザが入力したデータは制御ができないが故に倫理的ではない方向に(再)学習させる可能性がある。有名な例は 2016 年に Microsoft (*5)がチャットボットとして開発した Tay である。Tay は SNS 上の誹謗中傷を多く学習してしまったため、Tay 自身が汚い言葉を出力するようになったため 1 日も経たずに公開停止になった。2024 年現在における AI チャットボットにおいても、多くの研究者やハッカーが『脱獄』（開発者によって実装された制約の回避）を試みている。通常のソフトウェア開発においても Negative test は検証フェーズで実施されることが一般的である。その上で LLM は非常に広範なカバレッジをパスする必要があり、そしてオンデマンドに健全性が繰り返し検証されなければいけない。社内での AI 利活用ではこのようなトラブルは起きにくいとは思われるが、AI に悪意のあるデータを学習させないようユーザ自身が気をつける必要がある。

(*5) Microsoft は Microsoft Corporation の商標です。

AI によるサイバー戦争: Defensive AI 対 Offensive AI

前の章では、AI そのものに対してのセキュリティについて説明した。今度は現在のサイバーセキュリティに対し AI がどのように攻撃および防御メカニズムに関係するかという応用例について取り上げる。本来、大手 IT 企業がサービス展開している LLM においては悪意のある用途には使われないようフィルタリングがかかっている。一方でそのような制限がなく、サイバー犯罪に用いられることを目的とした複数の LLM が登場してきている。

Phishing 攻撃はこれまでは多くの場合、攻撃者が（機械翻訳などは使われるが）手作業で作成した固定的な文章を用いていた。生成 AI を用いることでよりバリエーションがあり自然な表現で人を騙すための文章を作り出すことができる。実際に、WormGPT という Phishing を行うための文章作成を行う LLM が 2023 年から出現している。2023 年末の研究ではまだ洗練された人間が書いた文章の方が攻撃の成功率が高いという報告が挙がっている[4]。しかし AI の発展速度を考慮すると、近い将来に成功確率を向上させることができるであろう。もちろん、AI によって作成された Phishing メールを悪い例として倫理的な LLM への学習データとすることも期待である。そのため Phishing メールがエンドユーザに読まれる前に AI によって防御することができる可能性も同時に挙がるであろう。

パソコンやサーバ、IoT 機器などのネットワークに繋がっている電子機器そのものを標的としたマルウェアも、AI によってさらに高度に進化する予兆が見られる。マルウェア感染は多くの場合、既存の公開された脆弱性に対してパッチが適応されていない機器から生じる。攻撃側は主に全数探索を攻撃対象のドメイン内に行い、どの機器がどのソフトウェアを動作させており、特定のバージョンに内在している脆弱性との関係性を推測する。AI を用いてマルウェアがよりインテリジェントになり、企業毎・ビジネス分野・機器の用途毎の傾向によって挙動を変えるようになる可能性も考えられる。AI を備えたマルウェアはターゲットから返ってきた応答に応じて適応的に攻撃手法を切り替えることも考えられる。またマルウェア生成の過程でも既存の検出ツールに見つかりにくいプログラム生成や、自動的に大量の亜種を作ることも AI には簡単に出来るであろう。実際に 2023 年の夏には FraudGPT というクラッキングツールを作ることができる AI ツールが発見されている。

もちろん、AI はネットワークセキュリティにおいて防御側にも利益をもたらす。現状の Firewall や IDS (Intrusion Detection System)はパッシブな動作により、事前に決められたルールに従って通信データの通過や拒否の判定を行っている。これらのメカニズムに AI を統合することにより、攻撃を見つけた際に迅速に適切なルールを適応することが可能になる。また、攻撃が実行される前の予備動作（ポートスキャンなど）から予測を行い、本格的な攻撃が行われる前に対抗策を適応することも可能でなる。非常に高いスキルを持つエンジニアが手作業で対応しているプロセスを AI に置き換えることができれば、人間の負担はかなり減らすことができる。多くの企業がセキュリティ人材不足に悩まされている中で、AI を用いたセキュリティツールはより効率的なセキュリティ対策の実現に役に立つことが期待される。Google、Microsoft、Cloudflare (6*)等の企業から防御メカニズムを改善するための AI ツールが発表されている。

今後サイバーセキュリティにおいては、攻撃側にも防御側のどちらにも AI アシストされたツールが配備されるであろう。すると近い将来、Offensive AI と Defensive AI による戦争が起きることは誰もが想像しやすい。さらには、Offensive AI の動作を学習した Defensive AI（あるいはその逆も）や、 低レイヤーの Offensive AI を制御する上位レベルの AI の登場といったより高度な AI が登場する可能性がある。最悪のケースの一つは、自発的な学習を繰り返しながら自動的にサイバー攻撃を行う無差別テロを起こす AI の登場であろう。攻撃側は防御システムのうち一か所でも攻撃出来ればよいというジレンマに打ち勝つためには、セキュリティシステムの全体最適化を人間と AI の得意不得意を相互にカバーして立ち向かうことが重要である。

(*6) Cloudflare は Cloudflare, inc.の商標です。

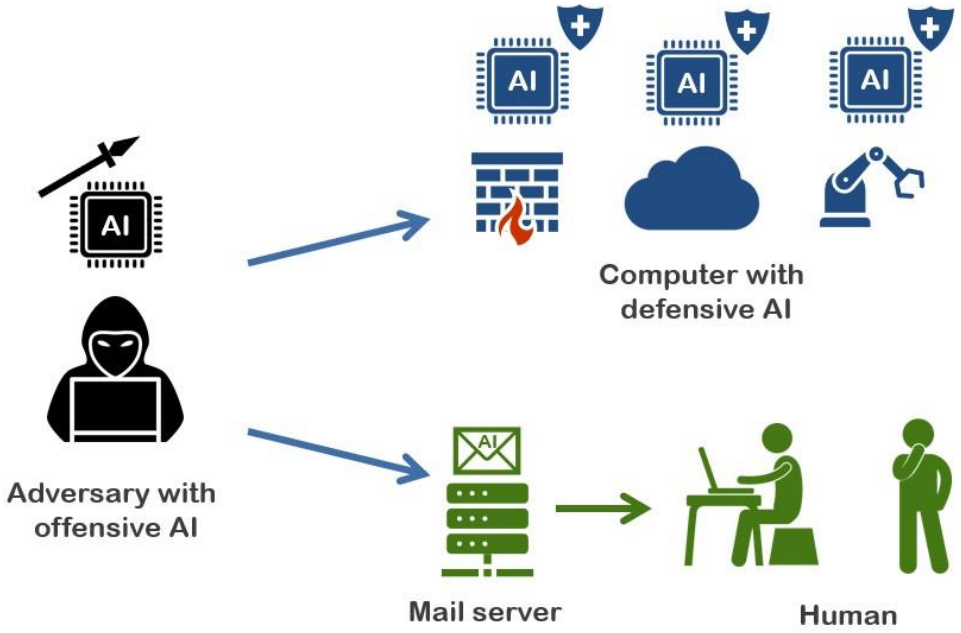

図 3 AI を備えた攻撃と AI を備えた防御

インターネット越しにいるのは人間か AI か

AI が発達すればするほど、インターネット越しに人間か AI かを判別することが徐々に難しくなってきている。この判定は古くからチューリングテストと呼ばれているものである。Bot でないことを確かめるために、多くの Web サイトでは CAPTCHA (Completely Automated Public Turing Test to tell Computers and Human Apart) (*7)が採用されている。歪んだ文字や画像認識テストが表示されるシステムは、多くのユーザが遭遇したことがあるであろう。一方で、2023 年の研究発表では AI によって進化した Bot が人間よりも早く正確に CAPTCHA を突破することができるという結果が公表されている[5]。もちろん、本来人間が解く平均時間よりも逸脱した速度で正解を入力すること自身も人間ではない証拠になる。それでも、ずる賢い AI ならば意図的なランダム遅延を行うことでより人間的な振る舞いを行うことが可能である。

最新の CAPTCHA である reCAPTCHA v3 では、パズルの代わりに Web サイト内でのアクションをバックグラウンドで観測し、スコアリングを行う。まだ reCHPTCHA v3 を明確に突破した AI に関する論文は現時点では見つかっていない。ただし人間のマウススクロールやタイピング速度そのものを学習して真似をするように命令された AI に対して、現時点での検出方法がその状況でも効果的であるかは議論の余地があるであろう。Offensive AI とは異なり、セキュリティ被害を直接的生じるわけではない。それでもなお、どのような手法であればデジタル化された世界において人間と AI を識別することが可能かは、どのようなセキュリティが効果的かを考える上での一つの指標になるであろう。

(*7) CAPTCHA は Carnegie Mellon University の商標です。

AI セキュリティに対する政府系のガイドラインやコミュニティ

近年 AI セキュリティのために、多くの政府系機関によって対する指針が発表され始めており、企業間の連携が結成され始めている。本章ではいくつかの活動をピックアップして紹介する。

米国では、NIST (National Institute of Standards and Technology)が 2023 年 1 月に AI RMF (Risk Management Framework)を発表した [6]。AI RMF の目標は、企業が AI システムを開発、展開、利用する際に生じるリスクを正しく管理し、信頼性を持ち責任のある開発を促すことである。信頼性を向上させるために以下の 7 つを挙げている。

(1) 客観的な証拠による妥当性と与えられた条件を満たす機能の提供
(2) 人間の生活環境を危険にさらさないこと
(3) 攻撃を回避・防御・対応・回復する能力
(4) モデルの構造や入力データなどの透明性の確保
(5)どのように判断を下したのかという説明可能性
(6) 匿名化や集約等によるプライバシーの保護
(7) 個人や社会に対して害をなすバイアスの管理

さらに、AI RMF では AI リスクを管理するため 4 つのコア機能を挙げている。

(a) AI の開発から利用までのライフサイクルのそれぞれについて関わる人々と AI システムに対する影響や関係性についてのマッピング
(b)マッピングによって見つけたリスクを質的かつ量的に分析評価する計測
(c)リスクの優先順位付けや継続的な改善を行う管理
(d)上記 3 つを統合して AI リスク管理を正しく運用するための統治

[6]にはそれぞれの機能をさらに詳しくカテゴリ分けしており、それぞれの機能に対するより具体的な要件が記載されている。

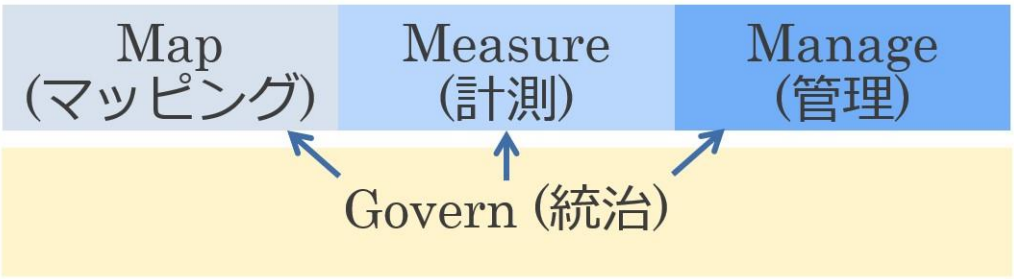

図 4 NIST AI RMF におけるコア機能

ヨーロッパでは、ETSI (European Telecommunications Standards Institute)が 2022 年から継続的に AI security に関するレポートを公開している [7][8]。

(1) AI システムに対しての脅威と従来のシステムとの違い
(2) データポイズニング攻撃を例としたデータ整合性の必要性
(3) AI ライフサイクルそれぞれに対しての CIA (Confidentiality, Integrity, Availability) との関係性
(4) 各ライフサイクルで起こりうる攻撃ごとの対策方針
(5) AI 専用ハードウェアを利用する際に守るべき情報とハードウェア固有の脆弱性

(6) AI における説明可能性と透明性
(7) AI プラットフォームに対してのセキュリティフレームワーク
(8) AI を利用した Deepfake の種類とその対策
といった広範なトピックに対して、AI システムを構築する上でのベースラインとなる分析を紹介している。

2024 年の始め頃には、米国、英国、日本において政府系組織がそれぞれの地域における AI safety institute の設立を発表しており、国際連携を行いつつ安全な AI 開発のための評価手法や基準の公表を計画している。また、2024 年 3 月には世界で初めて AI に関わる法律である Artificial Intelligence Act が EU 議会によって可決された。AI システムが許容できないリスクのため禁止されなければいけない分野、ハイリスクがあり要件の順守や第三者評価が必要な分野、限定的にリスクが存在し透明性の確保が必要な分野などの区分けが行われている。またこの法律では GDPR (General Data Protection Regulation)のように、組織が違反を犯した場合に巨額な制裁金を課している。

AI 開発を実際にリードしている企業間でも、Security for AI に関する複数のコミュニティが 2023 年夏頃から設立され始めている。Frontier Model Forum は Google, Microsoft, OpenAI (*8), Anthropic (*9)の 4 社からなる。責任のある AI モデルの開発と展開のための研究やベストプラクティスの特定、政策担当者や学者との知識の共有、サイバーセキュリティを含めた社会的な課題への対応などを行うことを目標としている。他の大きなコミュニティとしは AI Alliance がある。AI Alliance のメンバーは IBM、Meta (*10)を含む 70 以上の企業や大学からなる。AI Alliance は AI 特有のリスクの特定とそれらを低減させるための情報共有を行い、AI のオープンイノベーションを加速させることを目的の一つとしている。

もちろん、他組織との協調的な活動以外に、各 AI モデルの提供やサービスを行っている各企業は、個別に AI のためのセキュリティポリシーやフレームワークを掲げており、どのような取り組みを行っているかを Web サイト上にも公表している。

どの活動にも共通することは、AI がセキュリティやプライバシーに対して悪い影響を与えず、ユーザが安心して使うことができることを目標としていることである。そのためには AI 開発や運用に対する透明性を示すことが開発者には求められる。仮にある企業のビジネスにおいて AI を利活用するためにチューニングを行うだけであったとしても、上記の団体からのベストプラクティスや、それぞれの国や地域における規則に準拠することが望まれる。

(*8) OpenAI は OpenAI, Inc.の商標です。
(*9) Anthropic は Anthropic, PBC.の商標です。
(*10) Meta は Meta Platforms, Inc.の商標です。

まとめ

現代において我々はインターネットのない生活を過ごすのは非常に難しい。それと同じように、様々な AI システムが我々の生活に馴染むのは時間の問題である。多くの人がセキュリティに注意を払いながらインターネットを利用しているように、AI についてもどのようなリスクが生じるのか、本当に利用して問題ないのかについては各企業・各個人の判断が重要になってくる。そのためには、公的機関や標準化団体などが一定のクライテリアを公表し、第三者認証機関によってそれぞれの AI サービスのセキュリティレベルが保証されていることが望ましい。

近い将来、AI の進化によってこれまで高度なスキルを持った少数の人しか制御できなかった業務をサポートあるいは置き換えることができるようになる。特にサイバーセキュリティは AI によるアシストを適応する分野としての最も望ましい分野の一つである。人の操作なしに AI による自動的かつ適応的な防御ができるようになるには時間がかかるであろう。それでも多くのセキュリティ担当者はより効率的なセキュリティオペレーションのために AI に対しての強い期待を持っている。AI 運用管理は開発が完了したからといって終わるものではない。AI を継続的に駆使してコンピュータや人を騙す攻撃側に対抗するように、Defensive AI の知識の継続的な評価や更新を行う必要がある。

本ホワイトペーパーでは AI の利用場面としてサイバーセキュリティを取り上げたが、AI は様々な利用場面が想定されている。自動運転のための物体認識や、ファクトリーオートメーションにおける品質管理など、クラウドコンピューティングを経由しないエッジ AI ソリューションも今後増えていくことが予想されている。AI によるセキュリティインシデントが起こらないように、AI の応用例おいては常にセキュリティ問題が意識されることを期待したい。

[参考資料]

[1] https://www.pwc.com/gx/en/issues/c-suite-insights/ceo-survey.html

[2] https://riskonnect.com/press/riskonnect-research-generative-ai-risks-with-employees/

[3] https://www.isc2.org/Insights/2024/02/The-Real-World-Impact-of-AI-on-Cybersecurity-Professionals

[4] https://arxiv.org/abs/2308.12287

[5] https://doi.org/10.48550/arXiv.2307.12108

[6] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

[7] https://www.etsi.org/committee-activity/activity-report-sai

[8] https://www.etsi.org/newsroom/press-releases/2259-etsi-releases-three-reports-on-securing-artificial-intelligence-for-a-secure-transparent-and-explicable-ai-system

（Rev.1.0 April 2024）

本社所在地

〒 135-0061 東京都江東区豊洲 3-2-24（豊洲フォレシア）

https://www.renesas.com

お問合せ窓口

弊社の製品や技術、ドキュメントの最新情報、最寄りの営業お問合せ窓口に関する情報などは、弊社ウェブサイトをご覧ください。

http://www.renesas.com/contact/