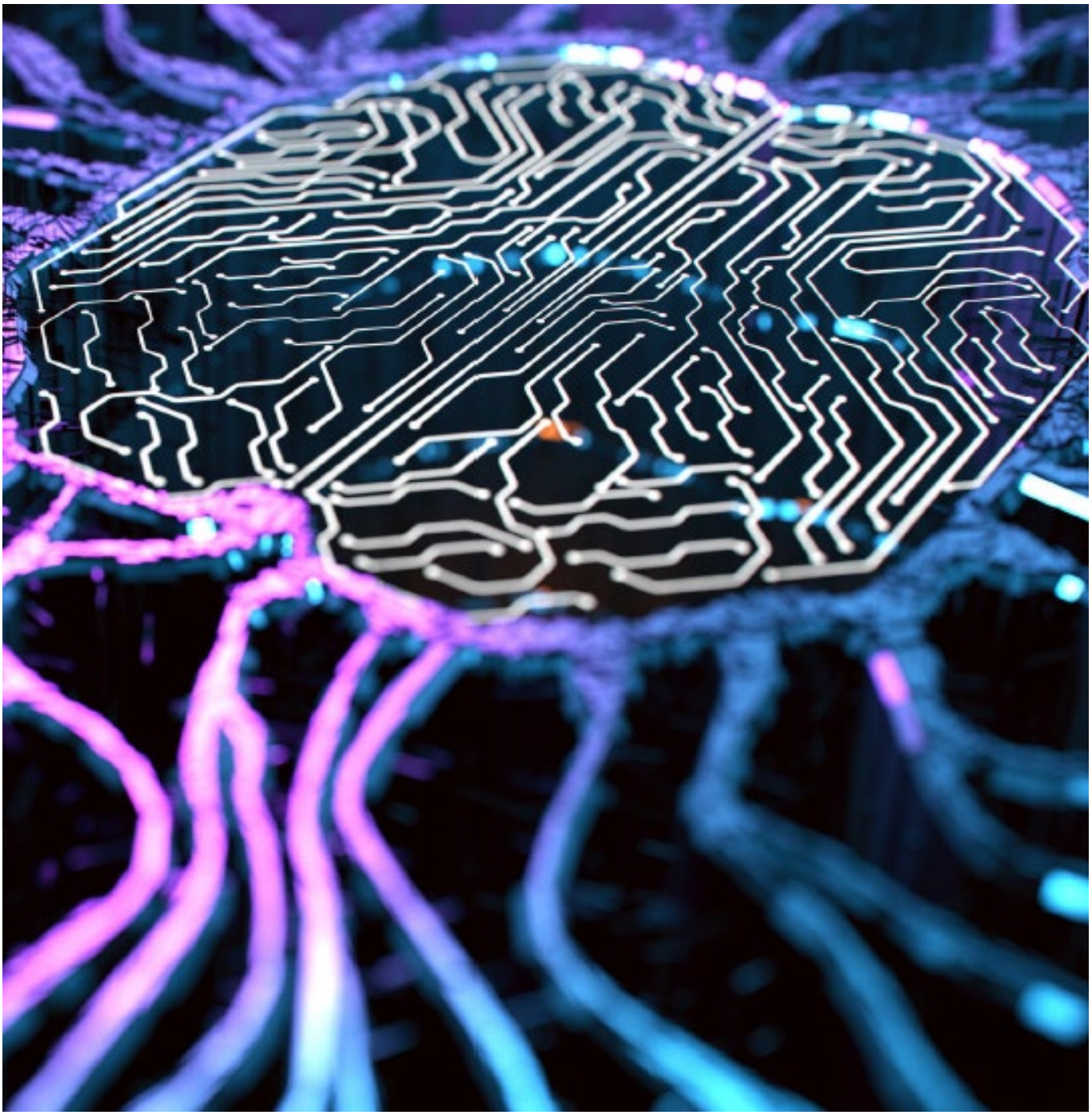
＊本資料は予告なく変更される場合があります。

仮訳

セキュアな AI システム開発のためのガイドライン



この文書について

この文書は英国サイバーセキュリティセンター（NCSC）、米国サイバーセキュリティ・インフラストラクチャー安全保障庁（CISA）及び次の国際パートナーによって公表される。

- 米国国家安全保障局 (NSA)
- 米国連邦捜査局 (FBI)
- 豪州通信電子局 (ASD)豪州サイバーセキュリティセンター (ACSC)
- カナダサイバーセキュリティセンター (CCCS)
- ニュージーランド国家サイバーセキュリティセンター (NCSC-NZ)
- チリ政府コンピューターセキュリティインシデント対応チーム
- チェコ国家サイバー情報セキュリティ庁 (NUKIB)
- エストニア情報システム庁 (RIA)
- エストニア国家サイバーセキュリティセンター (NCSC-EE)
- フランスサイバーセキュリティ庁 (ANSSI)
- ドイツ連邦情報セキュリティ庁 (BSI)
- イスラエル国家サイバー総局 (INCD)
- イタリア国家サイバーセキュリティ庁 (ACN)
- 日本内閣サイバーセキュリティセンター (NISC)
- 日本内閣府科学技術・イノベーション推進事務局
- ナイジェリア国家情報技術開発庁 (NITDA)
- ノルウェー国家サイバーセキュリティセンター (NCSC-NO)
- ポーランドデジタル省 (MC)
- ポーランド NASK 国家研究所 (NASK)
- 韓国国家情報院 (NIS)
- シンガポールサイバーセキュリティ庁 (CSA)

謝辞

下記の組織がこのガイダンスの作成に貢献した:

- Alan Turing Institute
- Anthropic
- Databricks
- Georgetown University’s Center for Security and Emerging Technology
- Google
- Google DeepMind
- IBM
- Imbue
- Infection
- Microsoft

- OpenAI
- Palantir
- RAND
- Scale AI
- Software Engineering Institute at Carnegie Mellon University
- Stanford Center for AI Safety
- Stanford Program on Geopolitics, Technology and Governance

免責

目次

要約

本文書は、人工知能（AI）を使用するシステムのプロバイダーのためのガイドラインを提言する。システムが AI を使用する限り、それがゼロから作成されたシステムであっても、他者が提供するツールやサービスの上に構築されたシステムであっても関係ない。このガイドラインを実施することは、プロバイダーが意図したとおりに機能し、必要なときに利用でき、機密データを権限のない第三者に漏らすことなく動作する AI システムを構築することを支援する。

本文書は、組織がホストするモデルを基盤とするか、外部のアプリケーション・プログラミング・インターフェース（API）を利用するかにかかわらず、主に AI システムのプロバイダーを対象としている。しかし、全てのステークホルダー（データ科学専門家、開発者、政策決定者及びリスク所有者を含む）に対し、自らの機械学習 AI システムの設計、開発、導入、運用に関し十分な情報に基づく意思決定を行うのに役立てるよう、このガイドラインを読むよう促したい。

本ガイドラインについて

AI システムは社会に多くの利益をもたらす潜在的な可能性を有している。しかし、AI の好機を十分に実現するためには、セキュアかつ責任ある方法で開発、導入、運用されなければならない。

AI システムは新たなセキュリティの脆弱性に晒され、標準的なサイバーセキュリティの脅威と並んで考慮されなければならない。開発のペースが速い時、AI はそうであるが、そのような場合、セキュリティは二次的な考慮要素となることが往々にしてある。セキュリティは、開発フェーズだけでなく、システムのライフサイクルを通じ、中核となる必須要件でなければならない。

このため、ガイドラインは AI システム開発のライフサイクルにおいて 4 つの重要分野に区分される。すなわち、**セキュアな設計、セキュアな開発、セキュアな導入、セキュアな運用とメンテナンス**である。各項目において、組織の AI 製品開発プロセスに対するリスク全般の低減に資する考察や緩和策を指摘する。

1. **セキュアな設計**：この項は、AI システムの開発ライフサイクルにおける設計段階に適用されるガイドラインを内容とする。システムとモデルデザインを考慮する際の、リスク、脅威モデル化、特定の論点やトレードオフを理解することを含む。
2. **セキュアな開発**：この項は、サプライチェーンセキュリティ、文書化、アセットと技術的負債の管理を含む AI システム開発ライフサイクルにおける開発段階に適用すべきガイドラインを内容とする。
3. **セキュアな導入**：この項は、インフラやモデルを侵害、脅威、損失から保護することや、インシデント管理プロセスの開発、責任のあるリリースなどを含め、AI システム開発ライフサイクルにおける導入段階に適用すべきガイドラインを内容とする。
4. **セキュアな運用とメンテナンス**：この項は、AI システム開発ライフサイクルにおけるセキュアな運用とメンテナンスに適用するガイドラインを内容とする。ログ記録や監視、アップデート管理、情報共有といった特にシステムが導入された際の行動に関するガイドラインを提供する。

本ガイドラインは、「セキュアバイデフォルト」アプローチに従い、NCSC の「セキュアな開発・導入ガイダンス」、NIST の「セキュア・ソフトウェア開発フレームワーク」及び CISA、NCSC、各国のサイバー当局が発表した「セキュアバイデザイン原則」で定義されているプラクティスと密接に連携している。次の諸点を優先する。

- 顧客にもたらされるセキュリティの結果に責任を負う
- 徹底的な透明性と説明責任を受け入れる
- セキュアバイデザインを経営上のトッププライオリティにするため組織機構やリーダーシップを構築する

導入

人工知能（AI）システムは社会に多くの利益をもたらす潜在的な可能性を有している。しかし、AI の好機を十分に実現するためには、セキュアかつ責任ある方法で開発、導入、運用されなければならない。サイバーセキュリティは、AI システムの安全性、強靭さ、プライバシー、公平性、効率性、信頼性のために必要な前提条件である。

しかし、AI システムは新たなセキュリティの脆弱性に晒され、標準的なサイバーセキュリティの脅威と並んで考慮されなければならない。開発のペースが速い時、AI はそうであるが、そのような場合、セキュリティは二次的な考慮要素となることが往々にしてある。セキュリティは、開発フェーズだけでなく、システムのライフサイクルを通じ、中核となる必須要件でなければならない。

本文書は、AI を使用するシステムのプロバイダー[1]のためのガイドラインを提言する。システムが AI を使用する限り、それがゼロから作成されたシステムであっても、他者が提供するツールやサービスの上に構築されたシステムであっても関係ない。このガイドラインを実施することは、プロバイダーが意図したとおりに機能し、必要なときに利用でき、機密データを権限のない第三者に漏らすことなく動作する AI システムを構築することを支援する。

本ガイドラインは、確立されたサイバーセキュリティ、リスク管理、インシデント対応のベストプラクティスと併せて考慮すべきものである。特に、我々はプロバイダーに対し、米国サイバーセキュリティ・インフラストラクチャー安全保障庁（CISA）、英国国家サイバーセキュリティセンター（NCSC）及び全ての国際パートナーが作成した「セキュアバイデザイン」原則[2]に従うことを促したい。この原則は次の点を優先する。すなわち、

- 顧客にもたらされるセキュリティの結果に責任を負う
- 徹底的な透明性と説明責任を受け入れる
- セキュアバイデザインを経営上のトッププライオリティにするため組織機構やリーダーシップを構築する

「セキュアバイデザイン」原則に従うことは、システムのライフサイクルを通じ相当のリソースを必要とする。それは、開発者が、システムの設計の各レイヤー及び開発ライフサイクルの全ての段階において、顧客を保護する**機能やメカニズム、**ツールの**実装**の優先順位を上げるための投資を行う必要があることを意味する。こうすることは、後でコストのかかる再設計を防ぎ、短期的に顧客とデータを保護する。

AI セキュリティはなぜ違うのか

本文書で「AI」との文言を、特に機械学習（ML）アプリケーションを指すものとして使用する 3。全ての

---

[1] AI システムを開発し、（もしくは、AI システムに開発させ）自らの名前又は商標で市場に出す又はサービスを導入する自然人、公的当局や機関又はその他の団体を指す。

[2] セキュアバイデザインに関する更なる情報については，CISA のセキュアバイデザインのホームページ及びガイダンス「Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software」を参照願いたい。

[3] ルールベースシステムなど非機械学習 AI とは対照的である。

ML のタイプを対象とする。機械学習アプリケーションを次のアプリケーションと定義する。すなわち、

- 人間によって明確にプログラムされる必要のあるルールを設定することなく、コンピューターにデータのパターンを認識させ、文脈を与えることができるソフトウェアコンポーネント（モデル）を含むもの
- 統計的推論に基づき予測や提言、決定を生成するもの

AI システムは、既存のサイバーセキュリティ上の脅威と同様に、新たな分類の脆弱性に晒されている。「敵対的機械学習」（AML）と言う用語は、ハードウェア、ソフトウェア、ワークフロー、サプライチェーンを含む ML コンポーネントの根本的な脆弱性の悪用を表現するために使用される。AML は攻撃者に対し、次のような ML システムにおける意図しない挙動を引き起こすことを可能とする。すなわち、

- 分類や回帰に関するモデルの性能に影響を及ぼす
- ユーザーが権限のない動作を実行できるようにする
- 機微なモデル情報を抜き取る

大規模言語モデル（LLM）領域でのプロンプトインジェクション攻撃、故意に訓練データやユーザーフィードバックを破損させる（「データ・ポイズニング」として知られている）など、シナリオ次第で、これらの効果を達成する方法は多く存在する。

誰がこの文書を読むべきか

本文書は、組織がホストするモデルを基盤とするか、外部のアプリケーション・プログラミング・インターフェース（API）を利用するかにかかわらず、主に AI システムのプロバイダーを対象としている。しかし、**全ての**ステークホルダー（データ科学専門家、開発者、政策決定者及びリスク負担者を含む）に対し、自らの機械学習 AI システムの**設計、導入、運用**に関し十分な情報に基づく意思決定を行うのに役立てるよう、このガイドラインを読むよう促したい。

すなわち、全てのガイドラインを全ての組織に直接適用できる訳ではなく、攻撃の巧妙さの度合いや方法は AI システムを標的とする敵によって異なるので、貴組織のユースケースや脅威のプロファイルに沿ってガイドラインを検討することが望ましい。

セキュアな AI を開発する責任を有するのは誰か

最新の AI のサプライチェーンには多くのアクターが存在するのが通例である。単純化したアプローチでは、2 つのエンティティを仮定する。すなわち、

- データ整理、アルゴリズム開発、設計、導入・メンテナンスに責任を有する「プロバイダー」
- インプットを提供しアウトプットを受領する「ユーザー」

である。

多くのアプリケーションでこのプロバイダー・ユーザー・アプローチが利用されているが、徐々に実際

には見られないものとなりつつある[4]。プロバイダーは、第三者が提供するソフトウェア、データ、モデル、遠隔サービスを自らのシステムに統合するようになってきているからである。このように複雑なサプライチェーンによって、エンドユーザーは、安全な AI の責任がどこにあるのか理解するのが難しくなってきている。

ユーザー（エンドユーザー、又は、外部の AI コンポーネントを統合するプロバイダー[5]）は、通常、自らが使用するシステムに関連するリスクを十分に理解、評価、対処するため必要な見通しや知見を有していない。そのため、「セキュアバイデザイン」原則に則り、**AI コンポーネントのプロバイダーがサプライチェーンの先にいるユーザーのセキュリティ結果に責任を負う**。

プロバイダーは、モデル、パイプライン、システムの中で可能な限りセキュリティ統制と緩和策を実施し、また、設定が使用される場合には、最もセキュアなオプションをデフォルトとして実施する。リスクが緩和できない場合には、プロバイダーは、

- サプライチェーンの先にいるユーザーに対し、自身（該当する場合は）及び自身のユーザーが受け入れているリスクを知らせる
- ユーザーに対し、コンポーネントを安全に利用する方法を助言する

システムの侵害が、明らかに、又は広範に、物理的な損害やイメージ低下、ビジネス運営の著しい障害、機密情報の漏えい、法的責任に繋がる場合には、AI サイバーセキュリティは**重大**と評価されることが望ましい。

---

[4] CEPS は公表資料 ‘Reconciling the AI Value Chain with the EU’ s Artificial Intelligence Act’ において７つの異なるタイプの AI 開発の相互作用を記載。

[5] ISO/IEC 22989:2022(en)はこれを「AI システムを構成する機能的な要素」と定義

セキュアな AI システム開発のためのガイドライン

本ガイドラインは AI システム開発のライフサイクルにおいて４つの重要分野に区分される。すなわち、セキュアな設計、セキュアな開発、セキュアな導入、セキュアな運用とメンテナンスである。各分野において、組織の AI 製品開発プロセスに対するリスク全般の低減に資する考察や緩和策を指摘する。

本文書に提示したガイドラインは次の文書で定義されたソフトウェア開発ライフサイクルプラクティスと緊密に連携している。すなわち、

- NCSC の「セキュアな開発・導入ガイダンス」
- 米国国立標準技術研究所（NIST）の「セキュア・ソフトウェア開発フレームワーク（SSDF）」[6]

[6] NIST は、人工知能（AI）の安全、安心、信頼できる開発と利用を促進するためのガイドラインを作成する（及びその他の措置を講じる）ことを任務としている。詳細は「NIST’s Responsibilities Under the October 30, 2023 Executive Order」を参照

1. セキュアな設計

この項は、AI システムの開発ライフサイクルにおける**設計**段階に適用されるガイドラインを内容とする。システムとモデルデザインを考慮する際の、リスク、脅威モデル化、特定の論点やトレードオフを理解することを含む。

スタッフの脅威とリスクに対する意識を高める

システム所有者や組織幹部は、セキュアな AI に対する脅威と緩和策を理解する。データ科学者と開発者は、セキュリティ脅威と故障モードに対する意識を強く保ち、リスク負担者が十分な情報に基づき決定できるようにする。（例えば、標準的な情報セキュリティのトレーニングの一部として）AI システムが直面している特有のセキュリティリスクに関する指針をユーザーに提供し、セキュアコーディング技術とセキュアで責任ある AI プラクティスで開発者を訓練する。

システムに対する脅威をモデル化する

リスク管理プロセスの一環で、システムに対する脅威を評価するために総合的なプロセスを適用する。このプロセスは、AI コンポーネントが侵害されたり予期せぬ挙動をした場合[7]に、システム、ユーザー、組織、より広い社会に与える潜在的な影響を理解することを含む。また、AI 特有の脅威[8]の影響を評価し、意思決定を文書化することを含む。

システムで使用されるデータの機密性と種類が、攻撃者の標的としての価値に影響する可能性があることを認識する。こうした評価にあたっては、AI システムが価値の高い標的とみなされるようになるにつれ、また、AI そのものが、新たな、自動化された攻撃を可能ならしめるにつれ、脅威は高まることを考慮することが望ましい。

機能性とパフォーマンスだけではなくセキュリティのためにシステムを設計する

自分の抱える業務は AI を使うことで最も適切に対処できると確信できる。これを決定してから、AI 特有の設計における選択が適当であるか評価する。機能性、ユーザー経験、導入環境、パフォーマンス、保証、監督、倫理的・法的要件などを考慮しながら、脅威モデルとそれに関連するセキュリティ上の緩和策を考慮する。例示は次のとおり。

- 自社内で開発するか、外部のコンポーネントを使用するかを選ぶにあたりサプライチェーンセキュリティを考慮する。例えば、
  - 新しいモデルの訓練、既存モデルの使用（ファインチューニングの有無にかかわらず）、外部 API を介したモデルへのアクセスのうち、要求に適した選択を行う。

---

[7] 脅威モデル化に関する更なる情報は OWASP Foundation から参照可能
[8] MITRE ATLAS の「Adversarial Machine Learning 101」参照

- 外部モデルプロバイダーとの協力を選ぶことは、当該プロバイダーのセキュリティ態勢のデューデリジェンス評価が含まれる。
- 外部のライブラリを利用する場合、例えば、恣意的なコード実行に直ちに晒されることなく、システムがライブラリから信頼できないモデルをロードすることを防ぐコントロールがあることを確保するため、デューデリジェンス評価を完了する[9]。
- 第三者モデルやシリアル化された重みは、信頼のできない第三者コードで、遠隔コード実行を可能にし得るものとして処理することが望ましく、これらをインポートする際には、スキャンニングと隔離・サンドボックス化を実施する。
- 外部 API を使用する場合には、ユーザーにログインを要求したり、潜在的な機密情報を送付する前に確認するなど、組織の統制外のサービスに送付するデータに対し適切なコントロールを適用する。
- 訓練データはシステムの挙動を定義することを認識し、ユーザーフィードバックや継続学習データをモデルに統合させる場合を含め、データとインプットのチェックとサニタイズを適切に適用する。

➢ AI ソフトウェアシステム開発を既存のセキュアな開発及び運用のベストプラクティスに統合させる。AI システムの全要素が、既知の脆弱性クラスを可能な限り低減または排除するコーディング手法と言語を使用し適切な環境で記述される。

➢ 例えばファイルの修正や外部システムへのアウトプット指示など、AI コンポーネントがアクションをトリガーする必要がある場合、可能性のあるアクションを然るべく制限する。（これには、必要に応じて外部 AI 及び非 AI のフェイルセーフも含まれることが望ましい。）

➢ ユーザーとの意思疎通に関する決定は AI 特有のリスクに関する情報に基づくべきである。例えば、
- システムはユーザーに有用なアウトプットを提供し、かつ、潜在的な攻撃者に対し不必要なレベルの詳細を漏らさない。
- 必要な場合には、システムは、モデルのアウトプットに対し効果的なガードレールを提供する。
- API を外部の顧客やパートナーに提供する場合には、API 経由での AI システムへの攻撃を緩和する適切なコントロール措置を然るべく講じる。
- 最も安全な設定をデフォルトでシステムに統合させる。
- 最低限の管理者権限という原則を適用し、システム機能へのアクセスを制限する。
- ユーザーに対しリスクの高い機能を説明しつつ、それらを使用する場合にはユーザーによるオプトインを求め、禁止されているユースケースを伝え、可能であれば、代替策をユーザーに知らせる。

AI モデルを選択する際に、セキュリティ上の利点とトレードオフを考慮する

モデルの選択には、様々な要件のバランスをとることが含まれる。これには、モデルアーキテクチャー、設定、訓練データ、訓練アルゴリズ ム、ハイパーパラメータの選択が含まれる。その決定は脅威モデルによって得られる情報に基づくものであり、AI セキュリティ研究が進み、脅威への理解が深化するに伴って、

[9] GitHub の「RCE PoC for Tensorflow using a malicious Lambda layer」参照

定期的に再評価を行うことが望ましい。

AI モデルの選択にあたり、考慮すべき点は次のとおりである。ただし、これらに限定はされない。

- 使用するモデルの複雑さ、つまり、選択したアーキテクチャとパラメータ数。モデルにおいて選択されたアーキテクチャとパラメータの数は、他の要因の中でも特に、使用する際のインプットデータの変化に対し必要となる訓練データの量や、変化に対する安定性に影響を及ぼす。
- ユースケースや特定のニーズ適応させる（例えばファインチューニングによる）フィージビリティに対し、モデルは適当か。
- モデルのアウトプットを連携させ、解釈し、説明できる能力（例：デバッグ、監査、規制コンプライアンス）。これには、解釈が難しい大規模で複雑なモデルよりも、より単純で透明性の高いモデルを使用する方が利点がある可能性がある。
- サイズ、完全性、質、感度、鮮度、妥当性、多様性など訓練データセットの特徴。
- モデルの堅牢化（敵対的訓練など）、正則化及びプライバシー強化技術を利用する価値。
- モデルや基盤モデル、訓練データ、関連ツールなどコンポーネントの来歴及びサプライチェーン

これらの要素のうち幾つがセキュリティの結果に影響を及ぼすか等については、NCSC の機械学習のセキュリティに関する原則、特に「Design for security（model architecture）」を参照願いたい。

2. セキュアな開発

この項は、サプライチェーンセキュリティ、文書化、アセットと技術的負債の管理を含む AI システム開発ライフサイクルにおける開発段階に適用すべきガイドラインを内容とする。

サプライチェーンのセキュリティ確保

AI サプライチェーンのセキュリティをシステムのライフサイクルを通じて評価・監視し、サプライヤーに対しても、自社が他のソフトウェアに適用しているのと同じ基準を遵守するよう要求する。もし、サプライヤーが組織のスタンダードを守れない場合には、既存のリスク管理ポリシーに従い対処する。

もし自社で生産できない場合には、自社システムの強固なセキュリティを確保するために、検証済みの商用、オープンソース及びその他サードパーティの開発者から、セキュリティが高く、しっかり文書化されたハードウェアとソフトウェアのコンポーネント（例：モデル、データ、ソフトウェアライブラリ、モジュール、ミドルウェア、フレームワーク、外部 API）を取得し、維持する。

セキュリティ基準を満たさない場合には、ミッションクリティカルなシステムについては代替的な解決方法に障害迂回する準備を行う。NCSC の「Supply Chain Guidance」や「Supply Chain Levels for Software Artifacts(SLSA)[10]」の枠組みなどのリソースを使用し、サプライチェーンやソフトウェア開発ライフサイクルの認証を確認する。

アセットの特定、監視、保護

モデル、データ（ユーザーからのフィードバックを含む）、プロンプト、ソフトウェア、文書化、ログ、評価（潜在的に安全でない機能や故障モードに関する情報を含む）など AI 関連アセットが持つ組織にとっての価値を理解し、それらが重大な投資の結果である場合や、それらにアクセスすることで攻撃が可能になる場合を認識する。ログを機密データとして取り扱い、その機密性、完全性、可用性を保護するコントロールを実施する。

アセットがどこに所在するかを了知し、関連するあらゆるリスクを評価し受け入れる。アセットを追跡し、認証し、バージョン管理し、セキュリティを確保するためのプロセスとツールを保有し、アセットが侵害された場合には、既知の良好な状態に修復することを可能にする。

どのデータに AI システムがアクセスできるかを管理したり、AI によって生成されたコンテンツをその機微性や、コンテンツを生成するのに使用したインプットの機微性に応じて管理するプロセスやコントロールを確立する必要がある。

データ、モデル、プロンプトを文書化する

[10] SLSA の「Safeguarding artifact integrity across any software supply chain」参照

いかなるモデル、データセット、メタプロンプトやシステムプロンプトの作成、運用、ライフサイクル管理を文書化する。こうした文書化には、訓練データの情報源（ファインチューニングデータや、人間による、またはその他運用上のフィードバック等）、対象範囲と限界、ガードレール、暗号化ハッシュ値又は署名、保持時間、推奨されるレビュー頻度、潜在的な故障モードなど、セキュリティに関連する情報が含まれる。そのために便利な仕組みとして、モデルカード、データカード、ソフトウェア部品表（SBOMs）などがある。包括的な文書化の作成は、透明性や説明責任を支える[11]。

技術的負債を管理する

あらゆるソフトウェアシステムと同様に、AI システムのライフサイクルを通じて、「技術的負債」を特定、追跡、管理する。（技術的負債は、長期的な利益を犠牲にして短期的な結果を出そうとし、ベストプラクティスに及ばないエンジニアリングの意思決定をすることから生じる。）技術的負債は、金銭的負債と同様に絶対的に悪いというものではないが、開発の最も初期段階から管理されることが適当である[12]。ただし、AI の文脈でそうすることは、標準的なソフトウェアに比べ困難であり、また、開発サイクルが速く、プロトコルやインターフェースが十分確立されていないため、技術的負債の水準は高くなる可能性が高いと認識するだろう。また、ライフサイクル計画（AI システムの廃止プロセスを含む）が、将来の同様のシステムに対するリスクを評価、認識、軽減することを確保する。

[11] METI（日本経済産業省）2023 年作成「Guide of Introduction of Software Bill of Materials (SBOM) for Software Management」参照
[12] Google research「Machine Learning: The High Interest Credit Card of Technical Debt」参照

3. セキュアな導入

この項は、インフラやモデルを侵害、脅威、損失から保護することや、インシデント管理プロセスの開発、責任のあるリリースなどを含め、AI システム開発ライフサイクルにおける導入段階に適用すべきガイドラインを内容とする。

インフラのセキュリティ確保

優れたインフラセキュリティの原則を、システムのライフサイクルのあらゆる局面で使用されるインフラに対し適用する。研究開発及び導入において、API、モデル、データ、データの訓練及び処理パイプラインに対するアクセスコントロールを然るべく適用する。これには機微なコードやデータの保持環境を適切に隔離することが含まれる。これは、モデル窃取やモデルのパフォーマンスに害を与えることを目的とした通常のサイバー攻撃を緩和するのにも有用である。

継続的にモデルを保護する

攻撃者は、モデルの重みを取得することにより直接的にモデルにアクセスしたり、アプリケーションやサービスを経由してモデルにクエリをすることで間接的にモデルにアクセスすることにより、モデルやその訓練を行ったデータ[13]の機能を再構築[14]することができるかもしれない。攻撃者は、訓練中又は訓練後にモデル、データやプロンプトを改ざんし、アウトプットを信頼できないものにするかもしれない。

次のような方法で、モデル及びデータを直接又は間接のアクセスから保護する。

- 通常のサイバーセキュリティ上のベストプラクティスを実践する。
- 機密情報へのアクセス、改ざん、外部への窃取の試みを検知し妨げるため、クエリインターフェース上でのコントロール措置を実践する。

消費するシステムがモデルを認証できるように、モデルの訓練後すぐに、モデルファイル（例：モデルの重み）及びチェックポイントを含むデータセットの暗号ハッシュ値や署名を計算し、共有する。暗号技術では常にそうであるように、適切な鍵管理が不可欠である[15]。

機密性に関するリスク緩和へのアプローチは、相当程度、ユースケースや脅威モデルに因る。例えば、極めて機微なデータを含むアプリケーションなど、適用するのが難しい、又は高価すぎる理論上の保証を必要とするアプリケーションがある。適切な場合は、消費者、ユーザー、攻撃者がモデルやアウトプットへのアクセスを有することに関連するリスクの水準を追求したり、保証するために（差分プライバシーや準同型暗号などの）プライバシー強化技術を利用することができる。

インシデント管理手順の策定

---

[13] Tramèr et al 2016「Stealing Machine Learning Models via Prediction APIs」参照
[14] Boenisch 2020 作成「Attacks against Machine Learning Privacy (Part 1): Model Inversion Attacks with the IBM-ART Framework」参照
[15] NCSC が 2020 年作成「Design and build a privately hosted Public Key Infrastructure」参照

AI システムに影響を及ぼすセキュリティインシデントが避けられないことは、インシデント対応、エスカレーション、復旧計画に反映される。計画は、異なるシナリオを反映し、システムやより広範な研究が進化するため定期的に再評価する。重要な企業のデジタルリソースはオフラインのバックアップに保存する。インシデント対応者は、AI 関連のインシデントを評価し対処するために訓練されている。インシデント対応プロセスを可能にするため、高品質の監査ログや他のセキュリティ機能や情報を追加負担なしで顧客及びユーザーに提供する。

責任を持って AI をリリースする

ベンチマークやレッドチームなどの適切かつ効果的なセキュリティ評価（安全性や公平性など、このガイドラインの対象を超えるテストと並んで）を実施した後にのみモデル、アプリケーション、システムをリリースし、既知の制約や潜在的な故障モードについてユーザーに明確に説明する。オープンソースのセキュリティテストライブラリの詳細は、この文書の末尾にある参考資料の項を参照願いたい。

ユーザーが正しいことを簡単にできるようにする

新しい設定や構成オプションは、それらが生む経営上の利益と、それらが原因となるセキュリティ上のリスクを併せて評価しなければならない。理想は、最もセキュアな設定が唯一のオプションとしてシステム統合されていることである。自身で設定する必要がある場合には、デフォルトの選択肢は一般的な脅威に対して広くセキュアであることが望ましい（それがセキュアバイデフォルト）。システムが本来設計された目的とは大きく異なる方法で使用され、導入されることを制限するコントロール措置を適用する。

ユーザーに対し、モデルやシステムの適切な使用に関するガイダンスを提供する。これには限界や潜在的な故障モードを強調することを含む。ユーザーに対し、どのセキュリティの観点に責任を持つか、ユーザーのデータがどこで（どのように）使用、アクセス、保管されるか（例えば、データがモデルの再訓練に使用される、データが従業員やパートナーによって審査される等）を明示する。

4. セキュアな運用とメンテナンス

この項は、AI システム開発ライフサイクルにおけるセキュアな運用とメンテナンスに適用するガイドラインを内容とする。ログ記録や監視、アップデート管理、情報共有といった特にシステムが導入された際の行動に関するガイドラインを提供する。

システムの挙動を監視する

セキュリティに影響を及ぼす急激又は緩やかな挙動の変化を監視するなど、モデルやシステムのアプトプットやパフォーマンスを測定する。潜在的な侵入や侵害、自然なデータドリフトを把握し確認できる。

システムへのインプットを監視する

侵害・悪用時のコンプライアンス義務、監査、調査、回復を可能とするため、プライバシーやデータ保護の必要性に沿いながら、システムへのインプット（推論要求、クエリ、又はプロンプトなど）を監視し、記録する。これには、データ準備ステップ（イメージのトリミングやリサイズなど）を悪用することを目的としたものを含む、分布外インプットや敵対的インプットを明確に検出することも含まれ得る。

アップデートにはセキュアバイデザインのアプローチに従う

すべての製品にデフォルトで自動化アップデートを実装し、安全でモジュール化したアップデート手順を使用して配布する。アップデート手順（テスト・評価体制を含む）は、データ、モデル、プロンプトの変更がシステム挙動の変化につながる可能性があることを反映する（例えば、主要なアップデートを新たなバージョンとして扱う）。また、ユーザーがモデルの変更を評価し、対応できるようにサポートする（例えば、プレビューアクセスやバージョン管理された API を提供する）。

教訓を集め共有する

情報共有コミュニティに参加し、また、経済界、学術界、政府のグローバルエコシステムで協力しベストプラクティスを共有する。セキュリティ研究者が脆弱性を研究し報告することに同意することを含め、システムのセキュリティに関するフィードバックのためのオープンなコミュニケーションラインを組織内外に維持する。必要に応じ、詳細かつ完全な共通脆弱性一覧を含む脆弱性開示に対応する報告書を公表するなど、より広範なコミュニティに問題を広める。問題を迅速かつ適切に緩和及び修復するための措置を講じる。

参考資料

AI 開発

Principles for the security of machine learning
NCSC による、機械学習コンポーネントを伴うシステムの開発、導入、運用に関する詳述**ガイダンス**

Secure by Design - Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software
CISA、NCSC、その他の機関が共同執筆。このガイダンスは、AI を含むソフトウェアシステムの製造者が、セキュリティを製品の開発段階から取り入れ、製品を搬送してすぐにセキュアに使えるために必要な方法や段取りを記述。

AI Security Concerns in a Nutshell
ドイツ BSI が作成。本文書は、機械学習システムに対する攻撃の可能性及びこうした攻撃に対する防御策を紹介。

Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems and Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems
これらの文書は、G7 広島 AI プロセスの一部として作成され、安全、安心、信頼できる AI を世界に普及させるため、最先端の基盤モデル及び生成 AI システムを含む、最も高度な AI システムを開発する組織のための指針を提供。

AI Verify
標準化された試験を通じ国際的に承認された一連の原則に照らして AI システムの性能を検証する、シンガポールの AI ガバナンス試験の枠組み及びソフトウェアツールキット。

Multilayer Framework for Good Cybersecurity Practices for AI — ENISA (europa.eu)
AI システム、運用、プロセスの安全確保のために遵守すべき措置に関し、加盟国当局及び AI ステークホルダーを手引きするフレームワーク

ISO 5338: AI system life cycle processes (Under review)
機械学習及び試行錯誤的なシステムに基づく AI システムのライフサイクルを記述するためのプロセスや関連するコンセプト。

AI Cloud Service Compliance Criteria Catalogue (AIC4)
ドイツ BSI によるカタログは AI に特化した基準を提供し、ライフサイクルにわたり AI サービスのセキュリティ評価を可能とする。

NIST IR 8269 (Draft) A Taxonomy and Terminology of Adversarial Machine Learning
機械学習及び試行錯誤的なシステムに基づく AI システムのライフサイクルを記述するプロセスや関連概念をまとめたもの。

MITRE ATLAS
MITRE ATT&CK framework をモデルにし、かつリンクさせた、敵対者の機械学習システムに対する戦術や技術、ケーススタディに関するデータベース。

An Overview of Catastrophic AI Risks (2023)

Center for AI Safety が作成した文書は、AI がもたらすリスクの領域を示している。

Large Language Models: Opportunities and Risks for Industry and Authorities
ドイツ BSI が大規模言語モデル（LLM）の開発、導入、利用の機会とリスクについて更に学習しようとする企業、政府及び開発者のために作成したもの。

Introducing Artificial Intelligence
豪州サイバーセキュリティセンター（ACSC）のブログで、人工知能とその安全な利用方法について、わかりやすいガイダンスを提供している。

ユーザーによる AI モデルのセキュリティテストを支援するオープンソースは例えば次のとおり。

- Adversarial Robustness Toolbox (IBM)
- CleverHans (University of Toronto)
- TextAttack (University of Virginia)
- Prompt Bench (Microsoft)
- Counterfit (Microsoft).
- AI Verify (Infocomm Media Development Authority, Singapore)

サイバーセキュリティ

CISA’s Cybersecurity Performance Goals
既知のリスクや敵対者の技術の可能性や影響を相当に減少させるために、全ての重要インフラ事業者が実施することが望ましいとされる共通の防護。

NCSC CAF Framework
CAF は極めて重要なサービスや活動に責任を有する組織のためのガイダンスを提供

MITRE's Supply Chain Security Framework
サプライチェーンにおいてサプライヤーやサービスプロバイダーを評価する枠組み。

リスクマネジメント

NIST AI Risk Management Framework (AI RMF)
AI-RMF は、AI に特に関連する個人、組織、社会への社会技術的リスクを管理する方法を俯瞰。

ISO 27001: Information security, cybersecurity and privacy protection
この標準は、組織に対し、情報セキュリティ管理システムの設置、実施、メンテナンスについて指針を提供。

ISO 31000: Risk management
この標準は、組織内におけるリスク管理のためのガイドライン及び原則を組織に提供。

NCSC Risk Management Guidance
このガイダンスは、サイバーセキュリティリスクの実務者が、組織に与えるサイバーセキュリティリスクを理解・管理することを改善させるのに資するもの。