人工知能（AI）のリスクを低減する：

重要インフラの所有者および運営者のための安全およびセキュリティガイドライン

出版 2024 年 4 月

国土安全保障省

目次

免責事項：本文書は、一般市民をいかなる形でも拘束するものではなく、一般市民に明確性を提供することのみを意図している。本文書は、米国、その省庁、事業体、その役員、職員、またはその他のいかなる者に対しても、法律上または衡平法上執行可能な、実質的または手続き上の権利または利益を生じさせることを意図したものではなく、また生じさせるものでもない。

エグゼクティブサマリー

米国国土安全保障省（DHS）は、大統領令第 *14110* 号「人工知能の安全、確実、信頼できる開発および使用」（[1] ）において、重要インフラの所有者および運営者が使用するための安全およびセキュリティガイドラインを開発するよう命じられた。DHS は、商務省、16 の重要インフラセクターのセクター・リスク管理機関（SRMA）、および関連する独立規制機関と連携して、これらのガイドラインを策定した。

本ガイドラインは、サイバーセキュリティ・インフラセキュリティ庁（CISA）が 2024 年 1 月に SRMA と関連する独立規制機関によって完了された分野別 AI リスクアセスメントの分野横断的分析から学んだ洞察から始まる□[2]□CISA の分析には、分野横断的な AI のユースケースと導入パターンのプロファイルが含まれ、3 つの異なるタイプにわたる分野横断的な AI リスクの基礎分析が確立されている：1）AI を利用した攻撃、2）AI システムを標的とした攻撃、3）AI の設計と実装における失敗である。DHS はこの分析に加え、既存の米国政府政策の分析も活用し、重要インフラに対する特定された分野横断的な AI リスクを軽減するための具体的な安全・セキュリティガイドラインを策定した。 このガイドラインには、組織が AI システムのリスクに対処するのに役立つ 4 つの機能を含む、国立標準技術研究所（NIST）の AI リスク管理フレームワーク（AI RMF）が組み込まれている：統治、マップ、測定、管理である。[3]

本文書のガイドラインは、重要インフラ部門全体に適用できるよう広範に記述されているが、DHS は、重要インフラの所有者および運営者に対し、部門別および状況別の AI リスクと低減策を検討するよう奨励している。

---

[1]大統領令第 *14110* 号「人工知能の安全、確実、信頼できる開発と利用 」の第 4.3 節(a)(iii)は、国土安全保障省に以下のように指示している：「この命令の日付から 180 日以内に、国土安全保障省長官は、商務長官、国土安全保障省長官が決定する SRMA およびその他の規制当局と連携して、AI リスク管理フレームワーク、NIST AI 100-1、およびその他の適切なセキュリティガイダンスを、重要インフラの所有者および運営者が使用するための関連する安全およびセキュリティガイドラインに適切に組み込むものとする。

[2] このセクター別リスクアセスメントは、大統領令 14110 の 4.3(a)(i)項を受けて策定されたもので、SRMA および独立規制機関は、必要に応じて、重要インフラにおける AI の使用に関連するリスクを毎年評価するよう指示されている。 [3]

[3] NIST 人工知能リスク管理フレームワーク（AI RMF 1.）参照：NIST AI リスク管理フレームワーク。

序文

人工知能[4] (AI)は米国の重要インフラに変革的なソリューションを提供する可能性を秘めているが、AI システム[5] を重要インフラ[6] に序文することは、それらのシステムを致命的な障害、物理的攻撃、サイバー攻撃に対してより脆弱にする可能性がある。同時に、AI を搭載したテクノロジーは、敵対者が米国のシステムに対する攻撃を拡大・強化する新たな方法も提示する。米国人が日常的に依存している重要なサービスや機能を持つ重要インフラの所有者や運営者にとって、AI リスクの低減は、単に運用上の必要性だけでなく、国家安全保障と公共安全の必須事項である。

大統領令第 *14110* 号「人工知能の安全、確実、信頼できる開発と利用」第 *4.3(a)(iii)*項を受けて、米国国土安全保障省（DHS）は、重要インフラの所有者および運営者が使用するためのガイドラインを作成した。本ガイドラインは、重要インフラシステムおよびその機能の安全およびセキュリティに影響を与える、分野横断的な AI リスク（[7] ）に対処するものである。DHS は、商務省、セクター・リスク管理機関（SRMA）、[8] 、その他の重要インフラセクター規制当局と連携して本ガイドラインを策定した。

DHS は、サイバーセキュリティ・インフラセキュリティ庁（CISA）の重要インフラのセキュリティとレジリエンスの国家コーディネーターとしての専門知識を活用し、2024 年 1 月に SRMA と関連する独立規制機関が完了した分野別の AI リスクアセスメントの分野横断的分析から学んだ洞察から着手した。DHS はまた、安全な *AI* システム開発のためのガイドライン（[9] ）、共同サイバーセキュリティ情報シート「*Deploying AI Systems*

---

[4] 本文書において AI とは、合衆国法典第 15 編第 9401 条(3)および大統領令第 14410 条に規定される意味を有し、「人間が定義した所定の目的に対して、現実環境または仮想環境に影響を与える予測、推奨、または決定を行うことができる機械ベースのシステム」と定義される。人工知能システムは、機械および人間ベースの入力を使用して、現実および仮想環境を認識し、自動化された方法で分析を通じて、そのような認識をモデルに抽象化し、モデルの推論を使用して、情報や行動の選択肢を策定する。

[5] 本文書において、「AI システム」は大統領令 14410 に規定された意味を持ち、「AI の全部または一部を使用して動作するデータシステム、ソフトウェア、ハードウェア、アプリケーション、ツール、またはユーティリティ」と定義される。

[6] 本書において、重要インフラとは、合衆国法律集第 42 編第 5195 条 c(e)および大統領政策指令 21 (PPD-21)において規定される意味を有する。§ 5195c(e)および大統領政策指令 21(PPD-21)において、「物理的であれ仮想的であれ、米国にとって極めて重要なシステムおよび資産であり、かかるシステムおよび資産の機能不全または破壊は、安全保障、国家経済安全保障、国家公衆衛生もしくは安全、またはそれらの組み合わせに衰弱的な影響を及ぼす」と定義されている。

[7] 本文書において、「リスク」とは、NIST が定める「ある事象の発生確率と、対応する事象がもたらす結果の大きさや程度を複合的に測定したもの」という意味を有する。参照されたい：NIST AI リスク管理フレームワークを参照のこと。

[8] 各重要インフラ・セクターには、PPD-21 で特定されたセクター特定機関（SSA）が指定されており、DHS や他の関連連邦省庁との調整・協力を通じて脆弱性を特定し、インシデントの軽減を支援している。2021 年国防認可法(NDAA)は、"Sector-Specific Agency "という用語を "Sector Risk Management Agency "に更新した。PPD-21 が SRMA にどのように指示しているかの詳細については、国家インフラ保護計画（NIPP）2013 の「附属書 B：重要インフラパートナーと利害関係者の役割、対応計画、能力」を参照されたい。

[9] 2023 年 11 月、CISA と英国国立サイバーセキュリティセンター（NCSC）は「安全な *AI* システム開発のためのガイドライン」を共同開発し、AI システムの開発者にガードレールを設定し、セキュリティを AI システム開発の中核要件とした。本書は、CISA、NCSC、その他

Securely」（[10] ）、政府機関の AI 利用に関するホワイトハウス行政管理予算局（OMB）の覚書など、他のリソースからも洞察を得た□[11]□DHS はその後、国立標準技術研究所（NIST）の AI リスク管理フレームワーク（AI RMF）の中で、後続のガイドラインを統合し、フレームワーク化した。DHS は、分野横断的分析によって特定された特定のリスクを軽減し、AI RMF の機能と安全・セキュリティに特有の対応するサブカテゴリーに対応するガイドラインを選択した。

本ガイドラインでは、特に重要インフラに特有な影響を及ぼす安全性とセキュリティに対するリスクを取り上げている。NIST は「安全性」を、定義された条件下で、人の生命、健康、財産、環境が危険にさらされる状態に至らないようなシステムの特性と定義している。安全性には、予想される危害の確率と予想外の危害の可能性の両方を低減することが含まれる。「安全性」とは、システムに危害や損害を与えるように設計された、意図的で不正な行為に対する抵抗力と定義される。[12]

重要インフラに対する AI のリスクは極めて文脈的であるため、AI システムを使用する重要インフラの所有者および運営者は、本ガイドラインを使用する際、それぞれの具体的状況を考慮すべきである。本ガイドラインは、重要インフラの所有者、事業者、規制当局に対する既存の法的要件に取って代わるものではない。

AI のリスクと低減が進化し、新たな AI システムとユースケースが開発されるにつれて、DHS は必要に応じて、重要インフラの所有者と運営者のためのこれらの安全・セキュリティガイドラインの更新を継続する。DHS はまた、技術の進化、（NIST AI RMF を含む）標準状況の更新、重要インフラのリスクアセスメントからのインプット、AI 安全・セキュリティ委員会からのインプット、その他の利害関係者からのフィードバックに基づき、本ガイドラインの実施を支援するためのプレイブックを含む追加リソースの開発も検討する。

---

18 カ国を代表する国内外 21 のサイバーセキュリティ団体により共同発行されている。出版物の全文を読むには、以下を参照のこと：安全な AI システム開発のための共同ガイドライン

[10] 2024 年 4 月、NSA の AI セキュリティセンター（AISC）は、組織が AI システムを安全に導入する方法をまとめた共同サイバーセキュリティ情報シートを発表した。本書は、CISA、連邦捜査局（FBI）、オーストラリア信号総局オーストラリア・サイバーセキュリティ・センター（ASD 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・ナショナル・サイバーセキュリティ・センター（NCSC-NZ）、英国ナショナル・サイバーセキュリティ・センター（NCSC-UK）が共同で封印したものである。出版物全文を読むには、以下を参照のこと：AI システムを安全に導入する。

[11] 2024 年 3 月、OMB は初の政府全体の方針 M-24-10 を発表し、連邦政府における AI の利用から生じるリスク、特に国民の権利と安全に影響するリスクを管理しつつ、AI の統治とイノベーションを推進するよう各機関に指示した。メモの全文は以下を参照のこと：OMB M-24-10: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence.

[12] 安全」と「セキュリティ」の定義については、NIST の「The Language of Trustworthy AI: An In-Depth Glossary of Terms」を参照のこと。

重要インフラに対するリスク

リスクを軽減するための取り組みは、こうしたリスクのアセスメントから始めるべきである。大統領令第14110号4.3(a)(i)に対応して、CISAはSRMAおよび関連する独立規制機関と協力し、それぞれの部門における米国の重要インフラに対するAIのリスクアセスメントを毎年実施することになった。これらのアセスメントの範囲は、AIを導入することで重要インフラシステムが致命的な障害、物理的攻撃、サイバー攻撃に対して脆弱性を増す可能性がある方法を対象としていた。

本セクションでは、分野横断的なAIリスクの基礎分析を確立するために、各分野ごとのAIリスクアセスメントを要約し、リスクを3つのタイプに分類する：1）AIを利用した攻撃、2）AIシステムを標的とした攻撃、3）AIの設計と実装における失敗である。[13]また、本セクションでは、分野横断的なAIの使用事例と、AIの採用における分野横断的なパターンを紹介している。重要インフラの所有者および運営者は、AIの安全・セキュリティガイドラインを実施する際に、本セクションで詳述した知見とAIのリスクを考慮すべきである。

以下は、SRMAが各分野から提出された資料から得られた主な知見であり、米国の重要インフラに対する分野横断的なAIリスクに関する共通点を浮き彫りにしている：

- SRMAは一貫して、多くの重要インフラ機能にとって変革的な技術であるAIの可能性を強調したが、AIの利点と、複雑で急速に進化する技術がもたらすリスクとの間の緊張関係も指摘した。
- 現在までのところ、SRMAの報告によると、各セクターは主に、すでに部分的に自動化されている機能をサポートするためにAIを導入しており、将来の進歩として、より複雑な機能へのAIの適用を想定している。
- SRMAは、AIがロジスティクス、サプライチェーン管理、品質管理、物理的セキュリティ、サイバー防衛など、多くの長年の根強い課題の解決策をサポートできる可能性を指摘した。
- SRMAは一貫して、敵対者が現在のサイバー戦術、技術、手順を拡大・強化する潜在的手段としてAIを捉えていた。
- SRMAは、重要インフラ運用のリスクを管理し、低減するための以下の方法を特定した：

[13] 大統領令14110で定められた90日間の初期対応期間の一環として、また、国家のサイバーおよび物理インフラに対するリスクを理解、管理、削減するというCISAの使命に従って、CISAは、国立標準技術研究所（NIST）のAIリスク管理フレームワークにおけるAIの信頼性の7つの特性のうち、3つの特性に合致するAIリスクカテゴリーを特定した：Safe（安全）、Secure & Resilient（安全とレジリエンス）、Valid & Reliable（妥当性と信頼性）である。参照：NIST AIリスク管理フレームワーク。

- 情報通信技術（ICT）のサプライチェーンリスク管理、インシデント対応計画、意識向上およびトレーニングを含む継続的な人材育成など、リスク低減のベストプラクティスを確立する。
- データセットやモデルの検証、自動化されたプロセスの人間による監視、AI の使用方針など、AI により特化した低減戦略。

AI の用途と採用パターン

分野別の AI リスクアセスメントの一環として、SRMA はそれぞれの分野全体で 150 以上の AI の有益な利用法を特定した。重要インフラの所有者および運用者は、本書のガイドラインを使用して AI を安全かつ確実に導入すべきである。CISA は、解釈と議論を容易にするため、10 の AI 利用カテゴリーを開発し適用した。これらの AI 利用カテゴリーは、より複雑なアプリケーションが重要インフラに導入されるにつれて、将来の要約で進化する可能性が高い。

図 *1* に示すように、これらのカテゴリーを有病率順に並べると、以下のようになる：

- オペレーション認識：これは、組織のオペレーションをより明確に理解するために AI を活用することである。例えば、AI を使用してネットワーク・トラフィックを監視し、異常な活動を特定することで、サイバーセキュリティを強化することができる。
- パフォーマンスの最適化：これは、プロセスやシステムの効率性と有効性を改善するために AI を使用することを含む。例えば、サプライチェーン・オペレーションを最適化し、コストを削減し、納期を改善するために AI を使用することができる。
- 業務の自動化：これは、データ入力やレポート生成など、組織内の定型的なタスクやプロセスを自動化するために AI を使用することを指す。例えば、大量のデータを分類・分析するプロセスを自動化するために AI を使用することができる。
- イベント検知：これは、システムや環境における特定のイベントや変化を検知するための AI の使用を指す。例えば、異常な心拍数を検知するために、健康監視システムに AI を使用することができる。
- 予測：これは、現在および過去のデータに基づいて将来のトレンドや出来事を予測するために AI を使用することである。例えば、過去の販売データに基づいて販売動向を予測するために AI を使用することができる。
- 研究開発（R&D）：これは、新製品、サービス、技術の開発における AI の利用を指す。例えば、創薬プロセスを迅速化するために製薬業界で AI を活用することができる。

- システム・プランニング：これは、IT インフラなどのシステムの計画や設計における AI の使用を指す。例えば、様々な条件下で提案されたシステムの性能を予測するために AI を使用することができる。

- カスタマーサービス自動化：これは、よくある質問への回答や注文処理など、顧客サービスの側面を自動化するために AI を使用することを含む。例えば、チャットボットはカスタマーサービス自動化における AI の一般的な応用例である。

- モデリングとシミュレーション：これは、現実世界のシナリオのモデルやシミュレーションを作成するために AI を使用することを含む。例えば、都市計画を目的とした交通パターンのシミュレーションに AI を使用することができる。

- 物理的セキュリティ：これは、施設や地域の物理的セキュリティを維持するための AI の使用を指す。例えば、侵入者や不審な行動を検知するための監視システムに AI を使用することができる。

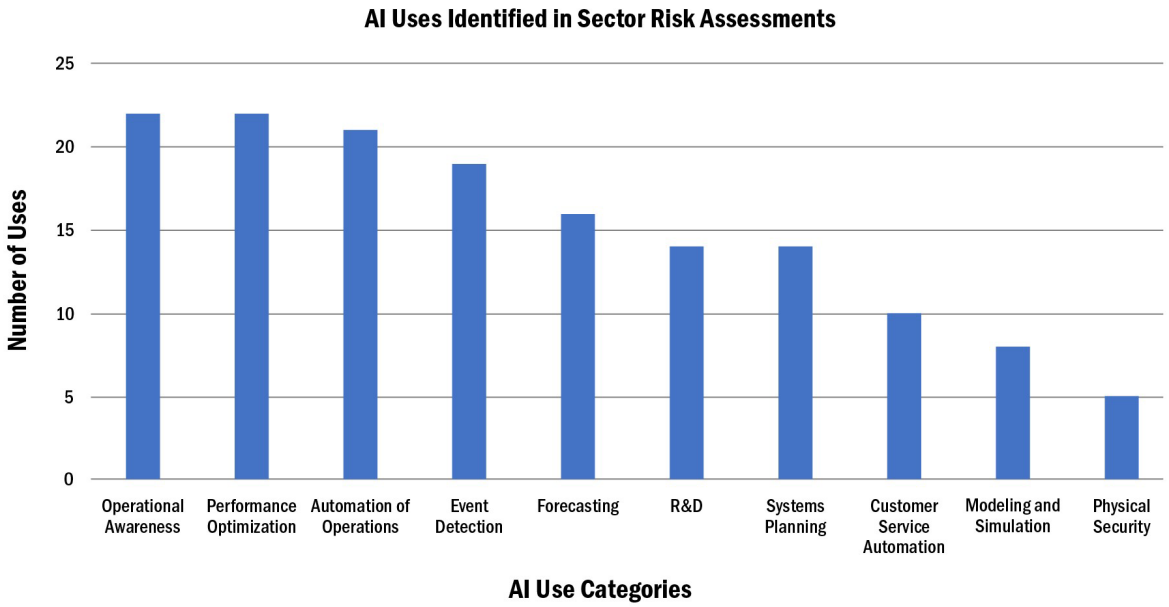


図 *1*：*2024* 年 *1* 月に *SRMA* から報告された *AI* 使用の種類別有病率

SRMA は、最も一般的な重要インフラの AI のユースケースは予測 AI であると示したが、生成的 AI（[14]）機能が広く利用できるようになった最近の進歩により、将来の評価においてこのような動きが変化する可能性があ

[14] 本文書では、生成的 AI は大統領令 14110 に定められた意味を持ち、以下のように定義される：生成的 AI とは、大統領令第 14110 号に規定された意味であり、次のように定義される：「派生する合成コンテンツを生成するために、入力データの構造や特性をエミュレートする AI モデルのクラス。これには、画像、動画、音声、テキスト、その他のデジタルコンテンツが含まれる。これは予測を生成する AI システムとは異なり、NIST では「入力データまたは測定値に適用される関数近似による定量的または定性的出力の予測」と定義している。こちらも参照のこと：信頼できる AI の言語：詳細な用語集.

る。SRMA はまた、予測、最適化、モデリング、シミュレーションなど、不確実性の高い出力を生成したり、より複雑なロジックを活用したりするユースケースについては、現在の AI 導入レベルが比較的低いと報告している。この傾向は、SRMA が、より複雑なインフラ運用における AI の採用を、それぞれの部門における将来の潜在的な取り組みとして想定しているという全体的な所見と一致している。最後に、ほとんどのアセスメントで、AI の導入の程度が増加傾向にあることが示された。

クロスセクターAI リスクカテゴリー

本文書のガイドラインは、CISA が分野横断的な AI リスク分析の一環として適用した、システムレベルの AI リスクの 3 つのカテゴリーを強調している。これら AI リスクの 3 カテゴリーに加え、分野横断的分析により、AI リスクの多数のサブカテゴリーと低減戦略が特定された。セクターごとに特定されたリスクのサブカテゴリーと軽減策の全リストは附属書 A を参照されたい。重要インフラの所有者および運用者は、本文書のガイドラインを実施する際に、システムレベルのリスクに関するこれら 3 つの包括的カテゴリー、ならびにセクターおよびコンテキスト固有のサブカテゴリーおよび低減策を考慮すべきである：

1. AI を利用した攻撃：このリスク・カテゴリーは、重要インフラに対する物理的攻撃やサイバー侵害を自動化、強化、計画、拡大するための AI の利用を指す。一般的な攻撃ベクトルには、AI を利用したサイバー侵害、自動化された物理的攻撃、AI を利用したソーシャル・エンジニアリングが含まれる。

2. AI システムを標的とした攻撃：このリスクカテゴリーは、重要インフラをサポートする AI システムに対する標的型攻撃に主眼を置いている。一般的な攻撃ベクトルには、AI アルゴリズムの敵対的操作、回避攻撃、サービス中断攻撃が含まれる。

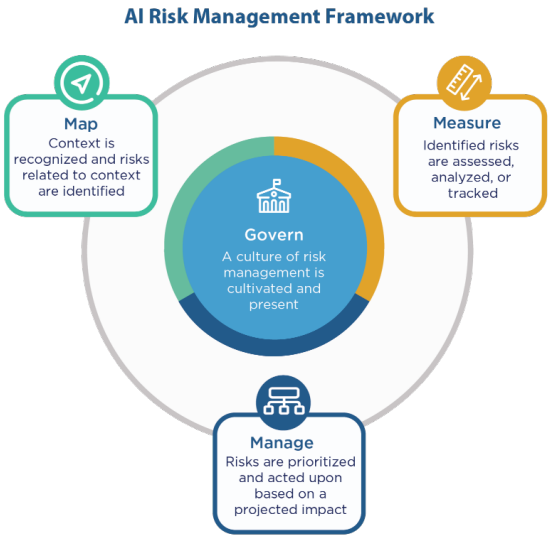
3. AI の設計と実装における失敗：このリスクカテゴリーは、AI ツールやシステムの計画、構造、実装、実行、保守における欠陥や不備に起因し、重要インフラの運用に影響を及ぼす誤動作やその他の予期せぬ結果につながる。設計と実装の失敗の一般的な方法には、自律性、脆弱性、不可解性が含まれる。

重要インフラの所有者および運営者のためのガイドライン

重要インフラの AI リスク管理は、AI のライフサイクルを通じて行われる継続的なプロセスである。[15] 重要インフラの所有者および運営者は、本ガイドラインを実施する際、上述の 3 つの AI リスクカテゴリー（AI を利用した攻撃、AI への攻撃、AI の設計および実装の失敗）を考慮すべきである。

AI リスクは文脈にも左右される。重要インフラの所有者および運営者は、AI のリスクをアセスメントし、適切な低減策を選択する際、自らのセクター固有およびコンテクスト固有の AI の利用を考慮すべきである。複数のリスクに対応する低減策もあれば、焦点を絞った低減策もある。本ガイドラインは 16 の重要インフラ部門すべてに広く適用されるが、特定の部門はすでに[16] 、特定の環境や状況に合わせた AI リスク管理のためのガイドラインを策定し、部門ごとの年次 AI リスクアセスメントの一環として使用するために、引き続きガイドラインを改良していく可能性がある。

重要インフラの所有者や運営者は、その分野や役割に応じて、AI のライフサイクルのさまざまな側面に焦点を当てることができる。場合によっては、重要インフラの所有者や運営者は、AI システムの設計、開発、調達に関与する。他のケースでは、彼らは AI システムの最初の設計者や開発者ではないかもしれないが、これらのシステムの配備、運用、管理、維持、または退役において負担すべきレベルの責任を持つかもしれない。



[15] 本文書で言及されている AI のライフサイクルは、NIST AI 100-1: AI リスク管理フレームワークおよび OECD の AI システム分類フレームワークと一致している。

[16] 財務省は 2024 年 3 月、金融サービス分野における AI 関連のサイバーセキュリティと不正リスクの状況を明らかにした報告書を発表した。参照されたい：金融サービスセクターにおける人工知能特有のサイバーセキュリティリスクの管理。[17]

図 2：*NIST AI* リスク管理フレームワーク *1*

多くの場合、AI ベンダー（[17] ）も、重要インフラ向けの AI システムの安全・安心な利用を確保する上で大きな役割を果たすことになる。一定のガイドラインは、重要インフラの所有者や事業者だけでなく、AI ベンダーにも適用される。[18]重要インフラの所有者と運営者は、AI ベンダーに対するこうした依存関係がどこに存在するかを理解し、それに応じて低減責任を共有し、明確化するよう努めるべきである。

NIST の AI RMF に沿ったこれらのガイドラインは、AI RMF を重要インフラのエンタープライズリスク管理プログラムに統合する取り組みを促進する。図 2 に示すように、AI RMF コアは、組織が AI システムのリスクに対処するのに役立つ 4 つの機能で構成されている：統治、マップ、測定、管理である。統治は AI RMF の横断的機能であり、既存のエンタープライズ・リスク・管理（ERM）機能の一部として AI リスク管理への組織的アプローチを確立する。AI のライフサイクルを通じて繰り返し取り組むべき推奨事項は、Map、Measure、Manage の各機能に盛り込まれている。これらのガイドラインは以下を強化する。

NIST の AI RMF に含まれる AI の安全性とセキュリティのリスク管理の慣行である。本ガイドラインの AI RMF への推奨されるマッピングについては、附属書 B を参照のこと。

統治：AI リスク管理の組織文化を確立する。

本セクションのガイドラインは、AI のライフサイクルのすべての時点において、AI のメリットとリスクを予測、特定、管理するための方針、プロセス、手順の確立を支援するものである。重要インフラの所有者及び運営者は、AI の安全及びセキュリティの優先順位を自らの組織原則及び戦略的優先順位と整合させることにより、リスク管理の文化を醸成することができる。この組織的アプローチは、リーダーが安全とセキュリティの成果に優先順位をつけてオーナーシップを持ち、セキュリティを最優先とする組織構造を構築する「セキュア・バイ・デザイン」の哲学に従う。[19,20]

a. AI を利用した攻撃、AI への攻撃、AI の設計と実装の失敗を含む、サイバーセキュリティのリスク管理、インシデント対応、セキュリティ意識向上、安全手順の詳細計画。

---

[17] 本書において、「AI ベンダー」は、AI システムの商業的供給者の意味を有する。NIST は、"ベンダー "をソフトウェアまたはハードウェアの商用サプライヤーと定義している。"AI システム "は、大統領令 14410 に規定された意味を持ち、"AI の全部または一部を使用して動作するデータシステム、ソフトウェア、ハードウェア、アプリケーション、ツール、またはユーティリティ "と定義されている。

[18] AI ベンダーの責任については、以下の「重要インフラ所有者・運営者のための統治」の「政府 *D*」を参照のこと。

[19] 詳細については、CISA が米国内外のパートナーと共同で発行した「セキュア・バイ・デザイン・ソフトウェアの原則とアプローチ」を参照のこと。

[20] AI システムに対する脅威をアセスメントする全体的なプロセスの採用については、「安全な AI システム開発のための共同ガイドライン」（9 ページ）の「システムに対する脅威をモデル化する」ガイダンスを参照のこと。

b. 設計からトレーニング、配備、保守に至るまで、AI 開発のライフサイクル全体を通じてセキュア・バイ・デザインを確立し、攻撃に対する堅牢性を確保し、AI の設計と実装のリスクを考慮する。

c. データの保持と廃棄に関するポリシーと期限を含め、AI エンタープライズにおける重要なデータを確立し、追跡する。

d. 重要インフラにおける AI システムの安全かつ確実な運用について、AI ベンダーとの役割と責任を確立する。これには、AI システムの統合テスト、データ、入力、モデル、機能の検証、継続的保守と監視に関する文書化された計画と定期的なコミュニケーションが含まれる。

e. 熟練した、吟味された、多様な AI 労働力を確立し、維持するための労働力開発と主題専門知識に投資する。

f. 社内で AI システムを開発する場合、AI システムを調達する場合、あるいはベンダーと協力して既存の AI システムをカスタマイズする場合の利点とトレードオフをアセスメントする。[21]

g. AI システム利用の透明性と、AI 主導の行動に対する説明責任メカニズムを確立する。

h. AI の脅威、AI のインシデント、AI のシステム障害を、社内外の関係者のすべての情報共有メカニズムに統合する。

i. セクター調整協議会（SCC）、統治調整協議会（GCC）、情報共有・分析センター（ISAC）[222] などの政府グループや業界グループと協力し、リスク管理のツールや手法に情報を提供する。

マップ：個々の AI の使用状況とリスクプロファイルを理解する。

本セクションのガイドラインは、重要インフラの所有者と運営者が AI リスクを評価し、それを軽減するための基礎となる状況を確立するものである。重要インフラの所有者と運営者は、まず、どのように、どこで、なぜ AI システムが使用されるのかを理解し、状況別・分野別のリスクをアセスメントし、安全とセキュリティへの潜在的影響に対処する必要がある。

---

[21] トレードオフのアセスメントについては、「セキュアな AI システム開発のための共同ガイドライン」（9～10 ページ）の「機能や性能だけでなく、セキュリティも考慮してシステムを設計する」、「AI モデルを選択する際には、セキュリティ上の利点とトレードオフを考慮する」ガイダンスを参照のこと。

[22] ISAC は、重要インフラの所有者や運営者が、その施設、職員、顧客をサイバーや物理的セキュリティの脅威やその他の危険から守るのを支援する。ISAC は、実行可能な脅威情報を収集、分析し、メンバーに普及させ、リスクを軽減し、レジリエンスを強化するためのツールをメンバーに提供する。ISAC のコンセプトは、1998 年 5 月 22 日に署名された大統領決定指令-63（PDD-63）に従って導入され、公布された。詳細については、以下を参照のこと：全米 ISAC 協議会

a. 重要インフラの文脈における、現在または提案されているすべての AI 使用事例を目録化する。AI を利用した攻撃、AI への攻撃、AI の設計と実装の失敗などのリスクを含む、コンテキスト固有およびセクター固有の AI リスクを文書化する。文書化された AI リスクについて、AI システムに対するリスク管理または低減措置を評価する。[23,24]

b. 個人、地域社会、社会に影響を与える AI システムに関連した、状況特有の安全・安心への潜在的影響を文書化する。バイアス、プライバシー、機密情報の誤用に関連する潜在的リスクを含める。[25]

c. AI システムの導入または取得プロセスの一環として、意図された目的、期待される利益、潜在的な安全・セキュリティリスク、関連データの質と適切性を文書化し、徹底的な AI 影響アセスメントを義務付ける。

d. 重要インフラの運用において、安全性や運用に影響を及ぼす可能性のある誤動作や予期せぬ結果に対処するために、どの AI システムを人間の監督下に置き、意思決定を人間がコントロールすべきかを確立する。

e. AI ベンダーのサプライチェーンをレビューし、セキュリティと安全性のリスクを確認する。このレビューには、AI システムを開発し、ホストするためにベンダが提供するハードウェア、ソフトウェア、およびインフラストラクチャを含めるべきである。また、可能であれば、ソフトウェア部品表（SBOM）、AI システム部品表（AIBOM）、データカード、モデルカードなど、ベンダのリスクアセスメントや文書を組み込むべきである。[26] .

---

[23] 組織の AI 関連資産の診断については、「安全な AI システム開発のための共同ガイドライン」（12 ページ）の「資産の識別、追跡、保護」ガイダンスを参照のこと。

[24] 文書化の詳細については、「セキュリティ AI システム開発のための共同ガイドライン」（12 ページ）の「データ、モデル、プロンプトを文書化する」ガイダンスを参照のこと。

[25] センシティブデータ」の定義については、以下を参照のこと：The Language of Trustworthy AI: An In-Depth Glossary of Terms」を参照のこと。

[26] SBOM の詳細な定義と説明については、大統領令 14028 第 10 条(j)を参照のこと。SBOM は、ソフトウェアの構築に使用されるすべてのオープンソースおよびプロプライエタリ・コンポーネントのインベントリを提供する。AIBOM には、AI システムに関する同様の情報が含まれている。SBOM と AIBOM は、ソフトウェアと AI システムの背後にあるサプライ・チェーンに関する洞察を提供し、脆弱性分析によって特定のシステムの潜在的リスクを分析することを可能にする。データ・カードとモデル・カードも同様の目的を果たすが、その代わりに、AI システムの訓練に使用されるデータ・セットと、システムを動かすモデルに関する詳細な情報を提供する。ある AI システムの採用者は、関連するデータとモデル・カードがあれば、それを確認することで、モデルの開発とその使用目的について理解を深めることができ、この情報を使用して AI システムの安全性とセキュリティリスクを評価することができる。ソフトウェアセキュリティと AI のコミュニティは、データとモデルカードを統合し、自動化されたツールで生成・利用可能な、調和された AIBOM について、より明確なガイダンスを作成するために取り組んでいる。

f. 安全およびインシデント対応プロセスに AI システムを取り入れた結果、新たな故障状態や代替プロセスの冗長性を認識し、それを維持する。[27]

測定：AI リスクをアセスメント、分析、追跡するシステムを開発する。

本セクションのガイドラインは、AI システムのライフサイクルを通じて AI のリスクと影響を測定・監視するための反復可能な方法と指標を特定する。重要インフラの所有者と運用者は、使用と AI リスク管理の意思決定に情報を提供するために、独自のコンテクストに特化したテスト、 評価、検証、妥当性確認（TEVV）プロセス（[28] ）を開発することができる。

a. 既知のリスク、エラー、インシデント、または悪影響を検知、追跡、測定するための指標とアプローチを定義する。

b. サイバーセキュリティやコンプライアンス上の脆弱性を含め、AI システムにエラーや脆弱性がないか継続的にテストする。[29]テスト、研究、開発には、専用の分離されたネットワークを使用する。[30]

c. マップ機能において特定された状況特有の AI リスクに対処するためのリスクコントロールと軽減策のパフォーマンスをアセスメントする。低減策が複数のリスクカテゴリーに対応しているか、または特定のリスクを効果的に管理するためにさらなる低減策が必要であるかを追跡する。

d. AI モデル用のプライベートインスタンス（[31] ）やツールなど、公開 AI ツールセットにおける機密情報のエクスポージャーを防止するためのプラクティスを確立する。

---

[27] AI システムは、見た目は正確だが実際にはそうではない（例えば幻覚）出力を提供する可能性があるため、AI システムにおける故障状態の識別は、他のタイプのシステムよりも難しい可能性がある。

[28] テスト、評価、検証、妥当性確認（TEVV）プロセスが AI の文脈でどのように適用できるかについては、AI のテスト、評価、検証、妥当性確認（TEVV）に関する NIST のリソースを参照のこと。重要インフラの所有者および運用者は、NIST AI 200-2「生成的 AI モデルおよびシステムとデュアルユース基盤モデルの評価およびレッドチームに関するガイドライン」がリリースされたら、それを確認することが推奨される。この文書は、特に生成的 AI とデュアルユース基盤モデルのレッドチームに適用される。

[29] これには、基本的な脆弱性スキャニング・プロセス（例えば、CISA 脆弱性スキャニングを参照）や、よりターゲットを絞った侵入テストやレッドチームとの連携が含まれる。AI を活用した異常検知は、攻撃を示す可能性のあるシステム動作の異常なパターンを特定するためにも有用である。

[30] 間接的または直接的なアクセスからモデルを防御する方法については、「安全な AI システム開発のための共同ガイドライン」（14 ページ）の「継続的にモデルを防御する」を参照のこと。

[31] 本文書において、AI モデルとは、大統領令 14110 に規定された意味を有し、「AI 技術を実装し、計算、統計、または機械学習の技術を使用して、与えられた一連の入力から出力を生成する情報システムのコンポーネント」と定義される。

e. AI モデルの性能と出力を測定し、動作の変化を特定し、AI システムの結果の正確性を展開後に検証する。脆さや不可解さなど、一般的な AI の設計や実装の失敗に対する測定も含める。[32]

f. レッドチームによる演習を含め、実世界の環境において、AI システムのコンテクストに応じた安全とセキュリティへの影響をテストし、評価する。[33]

g. AI ベンダーと AI ベンダーのシステムを安全性とセキュリティの観点から評価し、データドリフトやモデルドリフト、ベンダーの AI に関する専門知識、継続的な運用と保守など、懸念される主要分野を評価する。

h. レジリエンスを念頭に置いて AI システムを構築し、混乱からの迅速な回復を可能にし、展開段階での悪条件下でも機能を維持する。[34]

i. 測定基準が存在しない、または測定が不十分なリスクを定期的に見直し、ギャップを特定し、新たに適用可能な測定基準および測定アプローチを開発する。

j. AI の安全・安心情報を、影響を受ける可能性のある地域社会および利害関係者に報告し、彼らから フィードバックを受けるためのプロセスを確立する。[35]

管理：安全・安心に対する AI リスクに優先順位をつけ、対処する。

本セクションのガイドラインは、AI システムの利点を高めつつ、安全・セキュリティに有害な影響を及ぼす可能性を低減するためのリスク管理管理策と、それを実施・維持するためのベストプラクティスを定義している。管理機能を適切に実施するために、重要インフラの所有者および運用者は、定期的にリソースを割り当

---

[32] モデルの出力と性能の測定については、『セキュアな AI システム開発のための共同ガイドライン』（16 ページ）の「システムの動作を監視する」、および『AI システムをセキュアに展開する』（7 ページ）の「モデルの動作を積極的に監視する」を参照のこと。

[33] 本書において、AI レッドチームとは、大統領令 14110 号に規定された意味を有し、「AI システムの欠陥や脆弱性を発見するための構造化されたテスト作業であり、多くの場合、管理された環境において、AI の開発者と協力して行われる」と定義される。AI のレッドチームは、AI システムからの有害または差別的な出力、予期しないまたは望ましくないシステムの動作、制限、またはシステムの誤用に関連する潜在的リスクなどの欠陥や脆弱性を特定するために敵対的な方法を採用する専門の「レッドチーム」によって実行されることが最も多い。」; AI のレッドチームは、特に AI 技術の急速な進化の性質のために、実践の分野としてまだ発展途上であり、完全に定義されていない。その結果、重要インフラの所有者とオペレーターは、AI レッドチームに関するベストプラクティスが時とともに進化していくことを期待できる。重要インフラの運用者や所有者は、NIST AI 200-2「生成的 AI モデルおよびシステム、ならびにデュアルユースファウンデーションモデルの評価およびレッドチームに関するガイドライン」がリリースされたら、その内容を確認することが推奨される。この文書は、特に生成的 AI とデュアルユース基盤モデルのレッドチームに適用される。

[34] デプロイメント統治の詳細については、『AI システムを安全にデプロイする』（3 ページ）の「デプロイメント環境の統治を管理する」ガイダンスを参照のこと。

[35] 情報共有については、「安全な AI システム開発のための共同ガイドライン」（16 ページ）の「教訓の収集と共有」を参照のこと。

て、マッピングされ測定された AI リスクに対して、政府プロセスによって概説された軽減策を適用する必要がある。

a. 特定された AI の安全・セキュリティリスクに優先順位をつけ、潜在的な悪影響を軽減するために、エビデンスに基づくアプローチを用いる。

b. ソフトウェアとハードウェアの脆弱性管理およびパッチ適用、役割および ID ベースのアクセス管理の使用、システムへのアクセスと使用の記録および監視を含む、サイバーセキュリティのベストプラクティスに従う。[36]

c. 安全・安心に対する AI のリスクに対処するため、必要に応じて、新たな、または強化された低減戦略を実施する。[37]低減戦略は、リスク情報に基づき、状況に応じたものでなければならない。

d. 電子透かし、[38] コンテンツラベル、認証技術などのツールを導入し、一般市民が AI 生成的コンテンツを識別できるよう支援する。[39]

e. データの完全性チェック、暗号化、エンドポイント保 護、データ、モデル、機能の検証、データのバッ クアップ、データのマスキング、セキュリティリスクを 軽減するための防御的 AI 機能など、適切なセキュリティ 管理者を適用する。[40]

f. 識別された安全・セキュリティリスクを管理し、可能であれば既存の脆弱性に対処するために、AI ベンダーのシステムを展開する前に低減策を適用する。[41]

---

[36] 重要インフラの所有者および運営者は、重要インフラ事業体に対して推奨される共通の防御として、CISA の「分野横断的サイバーセキュリティ・パフォーマンス目標」も検討すべきである。

[37] AI モデルの低減戦略の一例として、敵対的サンプルがある。これは、AI モデルが特定のタイプの攻撃や操作を認識し、それに対抗できるように、AI モデルのトレーニングプロセスに「敵対的サンプル」を追加することを含む。この敵対的な訓練は、レッドチームや情報共有から得た学習に基づいて行われる。

[38] 本文書において、電子透かしは大統領令第 14110 号に定める意味を有し、以下のように定義される：「一般的に除去が困難な情報を、AI によって作成されたアウトプット（写真、ビデオ、オーディオクリップ、テキストなどのアウトプットを含む）に埋め込むことで、アウトプットの認証、またはその出所、変更、伝達の身元や特徴を検証することを目的とする行為」と定義されている。電子透かしスキームの信頼性にはまだ疑問があり、埋め込まれた電子透かしを除去したり改ざんしたりする攻撃が多く発表されている。レッドチームのアセスメントには、電子透かし技術をテストし、なりすまし、除去、迂回ができないことを確認することも含まれる。

[39] 本文書において、AI 生成的コンテンツ（「合成コンテンツ」とも呼ばれる）は、大統領令 14110 に規定される意味を有し、以下のように定義される：「画像、動画、音声クリップ、テキストなど、AI を含むアルゴリズムによって大幅に修正または生成された情報」である。

[40] AI システムの完全性の確保については、『AI システムを安全に導入する』（6-7 ページ）の「使用前および使用中に AI システムを検証する」ガイダンスを参照のこと。

[41] 配備前にセキュリティのベストプラクティスを適用する方法については、『AI システムを安全に配備する』（4-5 ページ）の「配備環境の設定を固める」ガイダンスを参照のこと。

g. AI システムの入出力を監視し、異常な行動や悪意のある行動を監視し、脅威検知のために AI 行動分析を適用する。[42]

h. AI システムに障害が発生した場合、重要インフラシステムの安全かつ確実な運用を回復するインシデント管理計画に基づき、インシデントに対応する。[43]

[42] 入力の監視については、「安全な AI システム開発のための共同ガイドライン」（pg. 16).

[43] インシデント管理の詳細については、「セキュアな AI システム開発のための共同ガイドライン」（14 ページ）の「インシデント管理手順の策定」ガイダンスを参照のこと。

結論

本ガイドラインは、重要インフラシステムとその機能の安全・セキュリティに影響を与える分野横断的な AI リスクを取り上げている。安全性とセキュリティは、重要インフラにとって唯一無二のものであり、関連する AI リスクに対処することは、単なる運用上の必要性ではなく、国家安全保障と公共安全の必須事項である。

これらのガイドラインは 16 の重要インフラ部門すべてに適用できるほど広範なものであるが、AI リスクは極めて文脈依存的である。したがって、重要インフラの所有者や運営者は、それぞれの具体的な現実の状況の中で、これらのガイドラインを検討すべきである。

個人や組織が新たな AI システムやユースケースを開発し、対応するリスクや低減策が進化するにつれて、DHS は本ガイドラインの更新を継続する。DHS はまた、AI 技術の進歩が将来もたらす新たな機会とリスクへの対応において、重要インフラの所有者および運営者を支援する追加リソースの開発も検討する。

附属書 A：セクター横断的な AI リスクと低減戦略

本文書のガイドラインは、システムレベルの AI リスクの 3 つのカテゴリー（AI を利用した攻撃、AI に対する攻撃、AI の設計と実装の失敗）を反映している。本セクションでは、CISA がまとめた分野横断的な AI リスクカテゴリーについて、リスクのサブカテゴリーと低減策の例とともに詳述する。

DHS は、この附属書に記載された各減少策を上記の安全・セキュリティガイドラインに組み込んだ。例えば、AI RMF のカテゴリー（政府、統治、マップ、低減、管理）と固有のガイドライン ID（*A*、*B*、C など）である。

これらのリスクカテゴリーと低減戦略は、各部門から報告されたデータに基づくものであり、各部門の AI リスクを網羅したものではない。また、優先順（アルファベット順）でもない。これらのリスクと低減策は、新たな AI システムやユースケースの開発とともに進化していく可能性が高い。

リスクカテゴリー：AI を利用した攻撃

このリスクカテゴリーは、重要インフラに対する物理的またはサイバー攻撃を強化、計画、拡大するための AI の使用を対象としている。

リスクのサブカテゴリー - AI を利用した攻撃：

- AI を活用したサイバー侵害：AI を使用して脅威行為者の能力と作戦を増強する（自律型マルウェア、偵察、ディープフェイクの使用、脆弱性洞察のためのテキストの自動解析、モデリングとモデルの推論と補完、不正データアクセス、サイバー攻撃検知の回避、脆弱性の特定と悪用、機械速度の意思決定、最適化、機密情報を明らかにするためのプロンプト・インジェクションなど）。
- 自動化された物理的攻撃：人間の介在の有無にかかわらず、自律システム（ドローン群、致死的自律兵器システムなど）によって行われる物理インフラへの攻撃。
- 物理的な標的と脆弱性の特定：AI システムを使ってデータを収集・分析し、潜在的な攻撃の標的を特定・監視する。
- ソーシャル・エンジニアリング：AI を利用した心理的な操作により、ユーザーを騙して機密情報を開示させたり、ディープフェイクや AI を利用したフィッシングの試みなど、セキュリティ管理を侵害する行為を実行させること。
- サプライチェーンの混乱：サイバー侵害、物理的攻撃、ソーシャル・エンジニアリングなど、AI を活用した攻撃を行い、脆弱性のある重要物資の物流サプライチェーンを標的とし、混乱させる。

- 知的財産の窃盗とリバースエンジニアリング：AI システムを使って公開データを収集・解釈し、知的財産やその他の機密情報をリバースエンジニアリングする。

- 武器の開発：AI システムを使って、新たな武器やその他の有害物質（即席爆発装置、化学兵器、生物兵器など）を物理的な攻撃のために改造・作成すること。

低減戦略 - AI を利用した攻撃：

- 人工知能生成コンテンツの識別技術：電子透かしや認証技術など、一般市民が AI 生成的コンテンツを識別するのを支援するツール。(管理 D)

- 防御的人工知能能力：重要インフラに対する物理的およびサイバー攻撃を検知、防止、対応するために AI を使用すること。(管理 E)

- 暗号化：適切な場合、量子抵抗暗号の実装や暗号インベントリの作成など、適切な暗号化手段により、静止時および転送時のデータを防御し、機密性を維持すること。[44](管理 E)

- ホスト・セキュリティ：不正アクセスや悪意のある攻撃からネットワークの物理的・仮想的コンポーネントを保護するために、脅威を検知、調査、対応する（監視、パッチ適用、サンドボックス化など)。(統治 D、管理 B、G）。

- ネットワーク・セキュリティ：不正アクセス、誤用、盗難からネットワークインフラを保護すること（アクセス管理、ネットワークセグメンテーションなど)。(統治 B、管理 B）。

- レッドチーム：AI システム、運用プロセス、サプライチェーンの脆弱性を積極的に特定し、対処するために、AI 開発者と協力して、管理された環境で潜在的な敵の能力をエミュレートすること。(測定 F)

- セキュア・バイ・デザイン：また、製品のセキュリ ティは中核的なビジネス要件であり、製品開発ライフサイクルの設計 段階で重要な考慮事項である（例：メモリ安全プログラミング)。(統治 B、測定 H）。

- ユーザのセキュリティ意識向上：事業体をサイバー脅威から守るための基本的な行動をユーザに教えるためのセキュリティの実践と基礎トレーニング。(統治 A)

[44] 耐量子暗号化については、NIST の耐量子暗号化プロジェクトを参照のこと。

リスクカテゴリー：AI への攻撃

このリスクカテゴリーは、主に重要インフラを支える AI システムに対する標的型攻撃に焦点を当てている。[45]

リスクのサブカテゴリー - AI への攻撃：

- **AI のアルゴリズムやデータを敵対的に操作すること**：アルゴリズム、データ、センサーを意図的に改変し、AI システムに、そのインフラにとって有害な振る舞いをさせること。
- **回避攻撃**：AI システムに悪意を持ってプロンプト・インジェクションを注入し、モデルを回避することで、システムの誤動作や機密情報の漏洩を引き起こす。
- **サービス妨害攻撃**：AI システムを、直接的に（例：AI システム自体を妨害する）または間接的に（例：必要なデータやコンピューティング・リソースの可用性を妨害する）、意図したユーザーが利用できないようにする攻撃。
- **データの損失**：AI システムやその他のサポートシステムから機密データや重要インフラデータが盗まれる。
- **モデルの反転と抽出**：モデルの学習データやパラメータを盗み出したり、モデルの機能をリバースエンジニアリングしようとする悪意のある試み。

低減戦略 - AI への攻撃：

- **代替プロセスの冗長性**：物理的なデバイス操作、従来の計算と分析、または通常高度な自動化の恩恵を受けるその他の手作業による冗長な手作業。*(マップ F)*
- **データマスキング**：認可されていない侵入者にとってはほとんど価値のないものでありながら、ソフトウェアや認可された人員にとっては使用可能であるように、機密データを変更すること。[46](管理 F)

---

[45] 重要インフラの所有者や運用者は、NIST の「Adversarial Machine Learning（敵対的機械学習）」も確認する必要がある：A Taxonomy and Terminology of Attacks and Mitigations』（攻撃と低減の分類法と用語）には、AI 特有の攻撃と、AI システムのさまざまなクラスに対する低減について、より広範な内訳が記載されている。

[46] 大規模な言語モデルを含む生成的 AI モデルの使用に関しては、強力なセキュリティ態勢には、モデルのプロンプトに機密データを使用しないことが含まれる。

- データセットの検証：機械学習（[47] ）や AI アルゴリズムが訓練されるデータセットを保護するための取り組みで、訓練からモデル・ポイズニングされたデータサンプルをフィルタリングしたり、対象者が注釈を付けたデータセットを使用したり、モデル・ハードニング、2 検出モデル、その他の方法で敵対的な操作からデータを保護する。*(*統治 *D*、統治 *E*）。

- **防御的人工知能能力**：重要インフラに対する物理的およびサイバー攻撃を防止、検知、対応するために AI を使用すること。(管理 E)

- **暗号化**：適切な場合、量子抵抗暗号の実装や暗号インベントリの作成など、適切な暗号化手段により、静止時および転送時のデータを防御し、機密性を維持すること。[48](管理 *E*)

- **ホスト・セキュリティ**：不正アクセスや悪意のある攻撃からネットワークの物理的・仮想的コンポーネントを保護するために、脅威を検知、調査、対応する（監視、パッチ適用、サンドボックス化など)。*(*統治 *D*、管理 *B*、*G*）。

- **人間の監督**：説明責任を促進するために、AI システムの開発、配備、運用を人間が監督すること。*(*マップ *D*)

- **ICT サプライチェーンリスク管理**：情報通信技術（ICT）サプライチェーンに対するリスクを特定し、管理する。これには、ICT サプライチェーン・ソースをセクター間で多様化すること、リスク管理と危機管理計画を確立すること、機密データと重要インフラ・システム運用の機密性、可用性、完全性に対するリスクを最小化するための監視と対応メカニズムを導入することを含む。[49]*(*統治 *D;* マップ *E;* 測定 *G;* マネージ *F*)。

- **アイデンティティとアクセス管理**：学習データ、モデル、および入力へのアクセスを制限し、強力なパスワード、多要素認証（MFA)、最小特権の原則、およびデジタルアクセスを制御するその他の手段を用いてシステムアクセスを制御する。*(*管理 *B*）。

- **入力の検証**：システム入力に厳格なパラメータを設定することで、悪意のある行為者が AI システムを混乱させたり劣化させたりする可能性を制限する。*(*統治 *D*、管理 *G*）。

---

[47] 本書において、機械学習は大統領令 14110 に定められた意味を有し、以下のように定義される：「データに基づいてタスクのパフォーマンスを改善するために AI アルゴリズムを訓練するために使用できる一連の技術 "と定義される。

[48] 生成的 AI で構築されたアプリケーションでは、プロンプトストア、キャッシュ、ベクターストア、微調整データ、知識ベースを含むが、これらに限定されないアプリケーションアーキテクチャのすべてのデータレイヤーにわたって暗号化対策を検討する必要がある。 [49] 重要インフラの所有者および運営者は、CISA の ICT サプライチェーン・リソース・ライブラリも確認すべきである。

- **モデルの検証**：開発中の機械学習や AI モデルが、テストデータセットでモデルをアセスメントすることにより、意図したとおりに機能していることを検証する取り組み。*(*統治 *D* ; 測定 *E*、*G* ; 管理 *E*)。

- **ネットワーク・セキュリティ**：不正アクセス、誤用、盗難からネットワークインフラを防御すること（ネットワークのセグメンテーションなど)。*(*統治 *B*、管理 *B*）。

- **レッドチーム**：AI システム、運用プロセス、サプライチェーンの脆弱性を積極的に特定し、対処するために、AI 開発者と協力して、管理された環境で潜在的な敵の能力をエミュレートすること。*(*測定 *F*)

- **セキュア・バイ・デザイン**：また、製品のセキュリ ティは中核的なビジネス要件であり、製品開発ライフサイクルの設計段階で重要な考慮事項である（例： メモリ安全プログラミング)。*(*統治 *B*、測定 *H*）。

- **ソフトウェア部品表**：ソフトウェアコンポーネントと依存関係、それらのコンポーネントに関する情報、およびそれらの階層的関係の正式なインベントリであり、ソフトウェア固有の脆弱性に対処するための脆弱性交換ファイルを伴う。このインベントリは、AI および非 AI システムに適用されるべきである。*(*マップ *E*)

- **ユーザのセキュリティ意識向上**：事業体をサイバー脅威から守るために役立つ基本的な行動をユーザに教えるためのセキュリティの実践と基礎的なトレーニング。*(*統治 *A)*

リスクカテゴリー：AI 設計と実装の失敗

このリスクカテゴリーは、AI ツールやシステムの計画、構造、実施、実行における欠陥や不備に起因するもので、重要インフラの運用に影響を及ぼす誤動作やその他の予期せぬ結果につながる。

リスクサブカテゴリー - AI 設計と実装の失敗：

- **自律性**：AI システムに対する過剰な許可や不十分な運用パラメータによって、誤動作や予期せぬ行動が促進される。

- **脆さ**：AI システムが本来の問題文脈から外れた状況に直面したときに、意図しない故障や予期せぬ挙動を起こし、ロバスト性の欠如につながること。

- **不可解さ**：AI システムの開発または配備における限定された解釈可能性、透明性の欠如、または文書化、あるいは AI システムの異常を診断し修正することを困難にする AI システム固有の不確実性。[49]

- **不用意なシステムと設計の欠陥**：AI システムやモデルの設計や開発における意図せざる欠陥で、重要インフラの運用やサプライチェーンを混乱させたり、敵が同じことをするために悪用する脆弱性を作り出したりするような、予期せぬ、あるいは有害な行動につながる可能性がある。

- **一貫性のないシステム保守**：AI モデルやそれをサポートするシステムの定期的な更新やメンテナンスが行われず、故障やサービスの中断につながる可能性がある。

- **AI システムと非 AI システム間の相互運用性と構成**：AI システムを、AI 以外の機能を含むより広範な IT ネットワークに統合することは、相互運用性と互換性の課題につながる可能性がある。さらに、複数の独立した AI システムをより大規模なネットワークに統合すると、複数の AI モデル間の互換性に課題が生じる可能性がある。

- **人工知能への過度の依存**：人間のオペレーターが AI システムの意思決定能力や業務遂行能力に過度に依存することで、AI システムが故障した場合に業務に支障をきたす可能性がある。

- **統計バイアス**：データ整合性の失敗やその他の設計上の欠陥に起因する計算上のエラーや歪みの再現や増幅。その結果、出力にバイアスがかかり、誤った意思決定がなされる可能性がある。

- **専門家の不足**：AI システムの設計、統合、訓練、管理、解釈の訓練を受けた人材の不足は、AI システムの不適切な選択、設置、使用につながる可能性がある。

- **サプライチェーンの脆弱性**：サードパーティは、検証されていないデータセットに依存する AI 製品や、誤動作や運用の中断につながる可能性のあるその他の外部要因に依存する AI 製品を使用する可能性がある。

- **人工知能への過度の依存**：システムまたはプロセスにおける AI の不十分な組み込みまたは使用により、AI の使用により防止または軽減できるはずのリスクまたは脆弱性が生じること。

低減戦略 - AI 設計と実装の失敗：

- **データ、モデル、機能の検証**：開発中の機械学習および AI モデルが、サードパーティ監査人またはその他の外部認証事業体によるものを含め、検証済みのテストデータセットでモデルをアセスメントす

[49] 不可解さをめぐる懸念に対処する一つの方法は、意思決定プロセスが透明で理解しやすい「説明可能な AI」システムを導入することである。

ることにより、意図したとおりに機能していることを検証する取り組み。*(*統治 *D*、測定 *A*、*E*、*G*、管理 *E*)。

- **人間の監督**：説明責任を促進するために、AI システムの開発、配備、運用を人間が監督すること。*(*マップ *D*)
- **ICT サプライチェーンリスク管理**：ICT サプライチェーンに対するリスクを特定し、管理する。これには、ICT サプライチェーン・ソースをセクター横断的に多様化すること、リスク管理及び危機管理計画を確立すること、並びに機密データ及び重要インフラ・システム運用の機密性、可用性及び完全性に対するリスクを最小化するための監視及び対応メカニズムを実施することを含む。*(*統治 *D;* マップ *E;* 測定 *G;* 管理 *F*)。
- **アイデンティティとアクセス管理**：学習データ、モデル、および入力へのアクセスを制限し、強力なパスワード、MFA、最小特権の原則、およびデジタルアクセスを制御するその他の手段を用いてシステムアクセスを制御する。(管理 B)
- **AI ツールとモデルの非公開インスタンス**：公開 AI ツールセットにおける機密情報のエクスポージャーを制限・防止するために、AI モデルとツールのプライベートインスタンスを実装する。*(*測定 D)
- **ソフトウェア部品表**：ソフトウェアコンポーネントと依存関係、それらのコンポーネントに関する情報、およびそれらの階層的関係の正式なインベントリであり、ソフトウェア固有の脆弱性に対処するための脆弱性交換ファイルを伴う。このインベントリは、AI および非 AI システムに適用されるべきである。*(*マップ *E*)

AI リスクに対する一般的低減策

以下の低減戦略は、AI を利用した攻撃、AI への攻撃、AI の設計と実装の失敗という 3 つのリスクカテゴリーすべてに幅広く適用できる。

一般的な低減戦略：

- **人工知能の利用原則**：説明責任、透明性、信頼性、トレーサビリティ、ガバナビリティを促進し、AI システムのレジリエンス、堅牢性、信頼性を支援する、AI システム開発ライフサイクル管理のための政府の取り組みと業界のベストプラクティス。*(*統治 *A*、*B*、*I*、測定 *A*、*G*、*J*、管理 *B*、*D*、*H*)。
- **業務レジリエンスの構築**：バックアップシステム（手動システムおよび非 AI システムを含む）の構築と導入、業務継続計画、危機対応演習、十分な流動性を確保し、災害発生時の業務レジリエンスと継続性を維持する。*(*マップ *F*、測定 *H*、管理 *E*、*H*)

- **データのインベントリー**：企業の機密データの詳細なインベントリで、どの情報を防御すべきかのタイムラインを含む。*(*統治 *C)*

- **データのバックアップ**：システム停止時や故障時の業務継続性を確保するため、すべての AI システムのオフライン・データ・バックアップを保守する。*(*管理 *E*)

- **エンドポイントセキュリティ**：マルウェアなどの外部脅威から脆弱性エンドポイントへの侵入を検知・阻止する防御策。(管理 E)

- **サイバーインシデント対応**：準備、検知と分析、封じ込め、根絶と復旧、インシデント発生後の活動など、組織がサイバー脅威に対応するプロセス。AI ツールの問題を特定し報告する方法など、AI 障害に特化した対応計画を含む。*(*統治 *A*、*H* ; 管理 *H*）。

- **従業員の審査**：機密施設や機密情報の取り扱いに携わる人々を選別し、敵対的なグループとのつながりや、有害な意図を示唆するその他の識別を行う。*(*統治 *A*、*E)*

- **人工知能の発展を導く**：連邦政府と AI システム所有者・運営者による、AI の安全性、標準開発、テストへの財政投資により、AI の開発を強化・促進する。*(*統治 *I)*

- **情報収集と分析**：物理的脅威とサイバー脅威に関する情報を収集・報告し、分析と脅威情報の作成に役立てる官民の取り組み。*(*統治 *H, I;* 施策 *J)*

- **政府間の情報共有**：物理的脅威とサイバー脅威に対する脅威の検知、予防、対応を強化するために、あらゆるレベルの政府間でコミュニケーションを図る。*(*測定 *J*)

- **官民間の情報共有**：政府機関と民間の重要インフラ所有者・運営者との間で、物理的・サイバー脅威情報をコミュニケーションする。*(*測定 *J*)

- **内部レビュー**：組織が継続的に評価とアセスメントを行い、使用しているソフトウェアの脆弱性を特定すること。(管理 B)

- **AI に対する社会的信頼の維持**：重要インフラの所有者および運営者が AI システムを責任を持って管理し、脅威やインシデントに迅速に対応する能力に対する信頼を維持する。*(*統治 *G*、*H*、マップ *B*、*D*、測定 *J*)。

- **モデルリスク・管理**：データ品質、モデル設計、配備、運用、廃止の問題により発生する可能性のある AI モデルに対するリスクの特定、アセスメント、モニタリング、低減を行う。*(*統治 *D* ; 測定 *E*、*G* ; 管理 *E*)。

- **AI ツールとモデルの非公開インスタンス**：公開 AI ツールセットにおける機密情報のエクスポージャーを制限・防止するために、AI モデルとツールのプライベートインスタンスを実装する。*(*測定 *D*)

- **制度的知識の維持**：AI ソフトウェアなしで重要なシステムを運用する能力も含め、長期にわたって蓄積された専門知識、スキル、経験の集合体を捉え、維持するために組織が開発する方針と能力。*(*統治 *E)*

- **公共部門のインシデント対応計画と実施**：国家安全保障、経済、または公衆の健康と安全に影響を及ぼす可能性のある物理的またはサイバー緊急事態や脅威が発生した場合の、緊急対応要員と法執行要員の指定された役割と責任を含む公共部門のインシデント対応計画を作成する。*(*統治 *A*、*H*）。

- **研究、開発、監視、テスト環境**：重要インフラ組織が研究開発を行い、新しい AI ソリューションをテストし、エラーや脆弱性がないか実装を監視できる専用の分離されたネットワーク。*(*測定 *B*）

- **脆弱性管理**：AI モデルや関連ソフトウェアの脆弱性をプロアクティブに識別、開示、修復する。(管理 B）

- **人材開発**：従業員に AI トレーニングを提供し、組織のワークフローにおける問題を特定できるようにする。*(*統治 *E)*

附属書 B：NIST の AI RMF とガイドラインのマッピング

本文書のガイドラインは、NIST の AI リスク管理フレームワーク（RMF）に沿ったものである。重要インフラの所有者や運営者は、このガイドラインを使用して、RMF の要素をエンタープライズリスク管理プログラムに組み込むことができる。AI RMF コアは、組織が AI システムのリスクに対処するのに役立つ 4 つの機能で構成されている：統治、マップ、測定、管理である。AI RMF はまた、これら 4 つの機能それぞれについて、サブカテゴリー（リスク管理のための推奨行動）を詳述している。

本セクションでは、AI RMF の特定のサブカテゴリー（統治 1.1、マップ 2.3 など）に、本書のガイドラインを重ね合わせている。これは、AI RMF を使用する際に特定のガイドラインがどのように適用されるかについて、重要インフラの所有者およびオペレーターに理解を深めてもらうことを意図している。

NIST AI RMF は幅広いトピックとリスクをカバーしている。DHS は、本文書のガイドラインを作成するにあたり、安全、セキュリティ、レジリエンスに関連する AI RMF のすべてのサブカテゴリーと、その他の関連セクションを組み込んだ。

統治

DHS ガイドライン	対応する AI RMF サブカテゴリー
A. AI を使用した攻撃、AI への攻撃、AI の設計と実装の失敗を含む、サイバーセキュリティのリスク管理、インシデント対応、セキュリティ意識、安全手順の詳細計画。	統治 *1.2* 信頼できる AI の特徴は、組織の方針、プロセス、手順に統合されている。
B. 設計からトレーニング、配備、保守に至るまで、AI 開発のライフサイクル全体を通じてセキュア・バイ・デザインを確立し、攻撃に対する堅牢性を確保し、AI の設計と実装のリスクを考慮する。	統治 *1.5* 信頼できる AI の特徴は、組織の方針、プロセス、手順に統合されている。
C. データの保持と廃棄に関する方針と期限を含め、重要な AI エンタープライズデータを確立し、追跡する。	統治 *1.6* AI システムを棚卸しする仕組みが整備されており、組織のリスク優先順位に従ってリソースが確保されている。

DHS ガイドライン	対応する AI RMF サブカテゴリー
D. 重要インフラにおける AI システムの安全かつ確実な運用について、AI ベンダーとの役割と責任を確立する。これには、AI システムの統合テスト、データ、入力、モデル、機能の検証、継続的保守と監視に関する文書化された計画と定期的なコミュニケーションが含まれる。	統治 *2.1* 役割と責任 AI リスクのマッピング、測定、管理に関連するコミュニケーションが文書化され、組織全体の個人とチームにとって明確になっている。
E. 熟練した、吟味された、多様な AI 労働力を確立し、維持するための労働力開発に投資する。	統治 *3.1* ライフサイクルを通じた AI リスクのマッピング、測定、管理に関する意思決定は、多様なチーム（例えば、属性、専門分野、経験、専門知識、経歴の多様性）によって行われる。
F. 社内で AI システムを開発すること、AI システムを調達すること、あるいはベンダーと協力して既存の AI システムをカスタマイズすることの利点とトレードオフをアセスメントする。	統治 *4.1* 悪影響を最小限に抑えるため、AI システムの設計、開発、配備、使用において、批判的思考と安全第一の考え方を育成するための組織の方針と慣行が定められている。
G. AI システム利用の透明性と、AI 主導の行動に対する説明責任メカニズムを確立する。	統治 *4.1* 悪影響を最小限に抑えるため、AI システムの設計、開発、配備、使用において、批判的思考と安全第一の考え方を育成するための組織の方針と慣行が定められている。
H. AI の脅威、AI のインシデント、AI システムの障害を、関連する内外の利害関係者のためのすべての情報共有メカニズムに統合する。	統治 *4.3* AI のテスト、インシデントの特定、情報共有を可能にする組織的慣行が整備されている。
I. セクター調整協議会（SCC）、統治調整協議会（GCC）、情報共有・分析センター（ISAC）等の政	統治 *5.1* AI リスクに関連する潜在的な個人的・社会的影響について、AI システムを開発・配備したチー

DHS ガイドライン	対応する AI RMF サブカテゴリー
府・業界団体と協力し、リスク管理のツールや手法に情報を提供する。	ムの外部からのフィードバックを収集し、検討し、優先順位を付け、統合するための組織の方針と慣行が整備されている。

マップ

DHS ガイドライン	対応する AI RMF サブカテゴリー
A. 重要インフラにおける現在または提案されているすべての AI の使用事例を目録化する。AI を利用した攻撃、AI への攻撃、AI の設計と実装の失敗などのリスクを含む、文脈固有および分野固有の AI リスクを文書化する。文書化された AI リスクについて、AI システムに対するリスク管理または低減措置を評価する。	マップ *1.1* 意図された目的、潜在的に有益な用途、文脈特有の法律、規範、期待、AI システムが導入される見込みのある設定が理解され、文書化される。
B. 個人、地域社会、社会に影響を与える AI システムに関連した、状況特有の安全・安心への潜在的影響を文書化する。バイアス、プライバシー、機密情報の誤用に関する潜在的リスクを含める。	マップ *2.*1 AI システムがサポートする具体的なタスクと、そのタスクを実行するための方法が定義される。
C. AI システムの導入または取得プロセスの一環として、意図された目的、期待される便益、潜在的な安全・セキュリティリスク、関連データの品質と適切性を文書化し、徹底的な AI インパクトアセスメントを義務付ける。	マップ *2.1* AI システムがサポートする具体的なタスクと、そのタスクを実行するための方法が定義される。
D. 重要インフラの運用において、安全性や運用に影響を及ぼす可能性のある誤動作や予期せぬ結果に対処するために、どの AI システムを人間の監督下に置	マップ *3.5* 人の監視のためのプロセスは、統治機能から組織の方針に従って定義され、アセスメントされ、文書化される。

DHS ガイドライン	対応する AI RMF サブカテゴリー
き、意思決定を人間がコントロールすべきかを定める。	
E. AI ベンダーのサプライチェーンをレビューし、セキュリティと安全性のリスクを確認する。このレビューには、AI システムを開発し、ホストするためにベンダが提供するハード ウェア、ソフトウェア、インフラを含めるべきである。また、可能であれば、ソフトウェア部品表 （SBOM）、AI システム部品表（AIBOM）、データカード、モデルカードなど、ベンダのリスクアセ スメントや文書を組み込むべきである。	マップ *4.1* サードパーティーのデータやソフトウェアの使用を含め、AI 技術とそのコンポーネントの法的リスクをマッピングするためのアプローチは、サードパーティーの知的財産権やその他の権利を侵害するリスクと同様に、整備され、遵守され、文書化されている。 マップ *4.2* サードパーティーの AI 技術を含む AI システムの構成要素の内部リスクコントロールが特定され、文書化されている。
F. 安全およびインシデント対応プロセスに AI システムを取り入れた結果、新たな故障状態や代替プロセスの冗長性を認識し、それを維持する。	マップ *5.2* 関連する AI アクターとの定期的な関わりを支援し、肯定的、否定的、予期せぬ影響に関するフィードバックを統合するための慣行と人材が整備され、文書化されている。

測定

DHS ガイドライン	対応する AI RMF サブカテゴリー
A. 既知のリスク、エラー、インシデント、または悪影響を検知、追跡、測定するための指標とアプローチを定義する。	測定 *1.1* マップ機能で列挙された AI リスク測定のためのアプローチと測定基準は、最も重要な AI リスクから実施するために選択される。測定されな

DHS ガイドライン	対応する AI RMF サブカテゴリー
	い、あるいは測定できないリスクや信頼性の特性は、適切に文書化される。
B. サイバーセキュリティとコンプライアンスの両方の脆弱性を含め、AI システムにエラーや脆弱性がないか継続的にテストする。テスト、研究、開発には、専用の分離されたネットワークを使用する。	測定 *1.1* マップ機能で列挙された AI リスク測定のためのアプローチと測定基準は、最も重要な AI リスクから実施するために選択される。測定されない、あるいは測定できないリスクや信頼性の特性は、適切に文書化される。 測定 *2.1* テストセット、メトリックス、テスト、評価、検証、検証（TEVV）で使用したツールの詳細が文書化されている。
C. マップ機能において特定された状況特有の AI リスクに対処するためのリスク管理及びリスク軽減のパフォーマンスをアセスメントする。低減策が複数のリスクカテゴリーに対応しているか、または特定のリスクを効果的に管理するためにさらなる低減策が必要であるかを追跡する。	測定 *1.2* AI 指標のアセスメントと既存のコントロールの有効性は定期的に評価され、エラーや影響を受けるコミュニティへの影響の報告も含めて更新される。
D. AI モデルやツールのプライベートインスタンスなど、公開 AI ツールセットにおける機密情報のエクスポージャーを防止するためのプラクティスを確立する。	測定 *2.1* テストセット、メトリックス、テスト、評価、検証、検証（TEVV）で使用したツールの詳細が文書化されている。
E. AI モデルの性能と出力を測定し、動作の変化を特定し、AI システムの結果の正確性を展開後に検証する。もろさや不可解さなど、AI の設計や実装における一般的な失敗に対する測定も含める。	測定 *2.3* AI システムの性能または保証規準が定性的または定量的に測定され、配備環境に類似した条件下で実証される。測定方法が文書化されている。

DHS ガイドライン	対応する AI RMF サブカテゴリー
	測定 *4.3* 影響を受けるコミュニティを含む識別アクターとの協議や、文脈に関連するリスクや信頼性の特性に関する現地データに基づき、測定可能なパフォーマンスの改善または低下が特定され、文書化される。
F. レッドチームによる演習を含め、実世界の環境において、AI システムの状況特有の安全性とセキュリティへの影響をテストし、評価する。	測定 *2.6* AI システムは、Map 機能で特定された安全リスクについて定期的に評価される。配備される AI システムは安全であることが実証され、その残留ネガティブリスクはリスク許容度を超えず、特に知識の限界を超えて動作させた場合、安全に故障することができる。安全性の指標には、システムの信頼性と堅牢性、リアルタイムのモニタリング、AI システムの故障に対する応答時間などが含まれる。 測定 *2.7* マップ機能で特定された AI システムのセキュリティとレジリエンスが評価され、文書化される。
G. AI ベンダーと AI ベンダーのシステムを安全性とセキュリティの観点から評価し、データ・ドリフトやモデル・ドリフト、ベンダーの AI に関する専門知識、継続的な運用と保守など、懸念される主要分野を評価する。	測定 *2.6* AI システムは、Map 機能で特定された安全リスクについて定期的に評価される。配備される AI システムは安全であることが実証され、その残留ネガティブリスクはリスク許容度を超えず、特に知識の限界を超えて動作させた場合、安全に故障することができる。安全性の指標には、システムの信頼性と堅牢性、リアルタイムのモ

DHS ガイドライン	対応する AI RMF サブカテゴリー
	ニタリング、AI システムの故障に対する応答時間などが含まれる。 測定 *2.7* マップ機能で特定された AI システムのセキュリティとレジリエンスが評価され、文書化される。
H. レジリエンスを念頭に置いて AI システムを構築し、混乱からの迅速な回復を可能にし、展開段階での悪条件下でも機能を維持する。	測定 *2.7* マップ機能で特定された AI システムのセキュリティとレジリエンスが評価され、文書化される。
I. 測定基準が存在しない、または測定が不十分なリスクを定期的に見直し、ギャップを特定し、新たに適用可能な測定基準および測定アプローチを開発する。	測定 *3.2* リスク追跡アプローチは、現在利用可能な測定技術で AI リスクを評価することが困難な場合や、測定基準がまだ利用できない場合に検討される。
J. AI 安全・セキュリティ情報を、影響を受ける可能性のある地域社会や利害関係者に報告し、そこからフィードバックを受けるためのプロセスを確立する。	測定 *4.2* AI システムの配備状況や AI ライフサイクル全体における信頼性に関する測定結果は、ドメイン専門家やその他の関連する AI アクターからのインプットによって知らされ、システムが意図したとおりに一貫して機能しているかどうかを検証する。結果は文書化される。

管理

DHS ガイドライン	対応する AI RMF サブカテゴリー
A. 特定された AI の安全およびセキュリティのリスクに優先順位をつけ、潜在的な悪影響を軽減するために、証拠に基づくアプローチを用いる。	管理 *1.2* 文書化された AI リスクの処置は、影響、可能性、または利用可能なリソースや方法に基づいて優先順位付けされる。
B. ソフトウェアとハードウェアの脆弱性管理とパッチ適用、役割と ID に基づくアクセス管理、システムへのアクセスと使用の記録と監視など、サイバーセキュリティのベストプラクティスに従う。	管理 *1.2* 文書化された AI リスクの処置は、影響、可能性、または利用可能なリソースや方法に基づいて優先順位付けされる。
C. 安全とセキュリティに対する AI のリスクに対処するために、必要に応じて、新たなまたは強化された低減戦略を実施する。低減戦略は、リスク情報を反映し、状況に応じたものでなければならない。	管理 *1.2* 文書化された AI リスクの処置は、影響、可能性、または利用可能なリソースや方法に基づいて優先順位付けされる。
D. 電子透かし、コンテンツラベル、認証技術などのツールを導入し、一般市民が生成的 AI コンテンツを識別できるようにする。	管理 *1.3* マップ機能によって特定された、優先度が高いと考えられる AI リスクへの対応計画が策定され、計画され、文書化される。リスク対応の選択肢には、低減、移転、回避、受容などがある。
E. セキュリティリスクを軽減するために、データ完全性チェック、暗号化、エンドポイント保護、データ、モデル、機能の検証、データバックアップ、データマスキング、防御的 AI 機能を含む適切なセキュリティ管理者を適用する。	管理 *2.2* 導入された AI システムの価値を維持するためのメカニズムが整備され、適用されている。

DHS ガイドライン	対応する AI RMF サブカテゴリー
F. AI ベンダーのシステムを導入する前に、特定された安全・セキュリティリスクを管理し、可能であれば既存の脆弱性に対処するための低減措置を適用する。	管理 *3.1* サードパーティ・リソースの AI リスクと利益は定期的に監視され、リスク制御が適用され、文書化されている。
G. AI システムの入出力を監視し、異常な行動や悪意のある行動を監視し、脅威検知のために AI の行動分析を適用する。	管理 *4.1* 展開後の AI システム監視計画が実施される。これには、ユーザーやその他の関連する AI アクターからのインプットを収集・評価する仕組み、不服申し立てと無効化、廃止、インシデント対応、復旧、変更管理などが含まれる。
H. AI システムに障害が発生した場合に、AI システムを重要インフラシステムの安全かつ確実な運用に回復させるインシデント管理計画に従ってインシデントに対応する。	管理 *4.1* 展開後の AI システム監視計画が実施される。これには、ユーザーやその他の関連する AI アクターからのインプットを収集・評価する仕組み、不服申し立てと無効化、廃止、インシデント対応、復旧、変更管理などが含まれる。