

米国国土安全保障省

重要インフラにおける人工知能の役割と責任の枠組み

人工知能安全安心委員会との協議について

2024年11月14日

内容

長官からの手紙.....	4
人工知能安全・セキュリティ委員会.....	6
エグゼクティブサマリー.....	7
I. 序文.....	9
II. 範囲.....	10
III. 重要インフラに対する AI のリスク.....	11
IV. 重要インフラにおける AI の安全かつ確実な展開のための役割と責任.....	13
A. 概要.....	13
B. 主要用語と構成.....	14
C. 枠組み.....	15
I. クラウドおよびコンピューター・インフラストラクチャ・プロバイダ.....	16
II. AI 開発者.....	19
III. 重要インフラ所有者および運営者.....	23
IV. 市民社会.....	26
V. 公共部門.....	28
V. 結論.....	30
A. 枠組みの望ましい成果.....	30
附属書 A : AI の役割と責任マトリックス.....	32
附属書 B : 用語集.....	34
附属書 C : 注意事項.....	36

長官からの手紙

米国の重要なインフラ・システム（家庭や企業に電力を供給し、きれいな水を供給し、私たちをつなぐデジタル・ネットワークを促進するシステムなど）は、国内および世界の安全と安定にとって不可欠である。米国の病院や救急サービス、変電所、パイプライン、水処理施設、その他の重要なシステムの円滑な運用が妨げられると、安全保障に壊滅的な影響を及ぼしかねない。

人工知能（AI）はすでに、アメリカ人が重要インフラと接する方法を変えつつある。例えば、新しいテクノロジーは、アメリカの一般家庭への郵便物の仕分けや配達、地震の迅速な検知と余震の予測、停電やその他の電気サービスの停止を防ぐのに役立っている。地震の誤報はパニックを引き起こしかねないし、新技術によってもたらされた脆弱性は重要なシステムを悪意のある行為者にさらすリスクになりかねない。

一方、敵対勢力は AI を駆使して、米国の重要インフラに対して複雑で巧妙かつ頻繁なサイバー攻撃を仕掛けており、今後も仕掛けてくるだろう。こうした脅威に対抗するため、重要インフラ事業者は AI を活用することで、悪質な攻撃をより効果的に防御し、重要インフラ・サービス全体のレジリエンスを改善することができる。

アメリカの継続的な安全保障と繁栄は、重要インフラ関係者がいかに AI を開発・展開するかにかかっている。

国土安全保障省（DHS）は、進化し拡大する脅威環境において、米国の重要インフラを守る責務を担っている。この重要な任務に対する我々のアプローチを示すため、私は今年初め、技術、産業、公民権、学界、政府における全米屈指のリーダーたちを招集し、他に類を見ない AI 安全・セキュリティ委員会を設立した。この委員会と緊密に協議しながら、DHS は「重要インフラにおける人工知能の役割と責任の枠組み」を策定した。

この枠組みは、AI エコシステム全体の事業者が、最終的に米国民にサービスを提供する重要なサービスとの役割や関係に基づいて検討、適応、実施できる具体的な行動を提案している。このフレームワークは、製品開発、調達決定、情報交換を推進する新規および既存のプロセスに組み込まれるように設計されている。また、AI の安全性とセキュリティに関する基礎研究の支援や、米国民に利益をもたらす AI のユースケースの探求など、完全に完了することはないだろうが、それでも継続的な進歩を確実にするためには、国を挙げての協調的な取り組みが必要な責務も捉えている。

米国の重要インフラにおける AI の異常な規模と影響に対処するには、国家を挙げてのアプローチが必要である。本省は、ホワイトハウス、AI 安全研究所、サイバーセキュリティ・インフラセキュリティ庁の重要な取り組みから生まれたこの枠組みの中核的要素を社会化し、調和させる上で重要な役割を担っている。今後、私たちは評論家であるインフラ・パートナーと協力し、この枠組みをどのように分野固有のニーズに適応させることができるかを理解していく。また、重要インフラ全体における AI の安全性とセキ

セキュリティへのアプローチを世界的に調和させる方法について、国際的なパートナーとの対話を開催する予定である。国土安全保障長官として、また AI 安全・セキュリティ委員会の委員長として、私たちが共に成し遂げたことを誇りに思います。我々は、これらの重要な問題に関する継続的な対話を歓迎し、この驚異的な技術が成長し進化し続けるにつれて、この枠組みを更新することを楽しみにしている。

アレハンドロ・N・マヨルカス

米国国土安全保障省長官

人工知能安全・セキュリティ委員会委員長

人工知能安全・セキュリティ委員会

人工知能安全・セキュリティ委員会（Artificial Intelligence Safety and Security Board、以下「委員会」）は、我が国の重要インフラにおける AI 技術の安全・安心で責任ある開発と展開について、長官、重要インフラ・コミュニティ、その他の民間セクターの利害関係者、およびより広範な一般市民に対して助言を行う。理事会の任務は、あくまで諮問的な性格を持つ。理事会のメンバーは、各セクターの無報酬の代表者として活動する。理事会は、国土安全保障省による重要インフラにおける AI の役割と責任の枠組み（「人工知能の役割と責任の枠組み」）の策定において、助言、情報、勧告を提供した。

アレハンドロ・N・マヨルカス（米国国土安全保障省長官、人工知能安全・セキュリティ委員会委員長）

サム・アルトマン（OpenAI CEO）

ダリオ・アモデイ、アンソロピック CEO 兼共同創設者

エド・バスティアン、デルタ航空 CEO

マーク・ベニオフ、セールスフォース会長兼 CEO

ヒューメイン・インテリジェンス CEO、ルーマン・チョードリー博士

マット・ガーマン、アマゾン ウェブ サービス CEO

アレクサンドラ・リーブ・ギブンス（民主主義とテクノロジーのためのセンター代表兼 CEO）

ブルース・ハレル（ワシントン州シアトル市長、米国市長会議テクノロジー・イノベーション委員会委員長）

デモン・T・ヒューイット「法の下での市民権」弁護士委員会会長兼事務局長

ヴィッキー・ホルブ、オクシデンタル・ペトロリアム社長兼 CEO

ジェンセン・フアン、エヌビディア社長兼 CEO

アルヴィンド・クリシュナ、IBM 会長兼 CEO

フェイフェイ・リー博士、スタンフォード人間中心人工知能機構共同ディレクター

ウェス・ムーア（メリーランド州知事）

サティア・ナデラ（マイクロソフト会長兼 CEO）

シャンタヌ・ナライエン、アドビ会長兼 CEO

スندانル・ピチャイ、アルファベット CEO

アラティ・プラバカー博士、大統領補佐官（科学技術担当）、ホワイトハウス米国科学技術政策局局長

チャック・ロビンス（シスコ会長兼 CEO、ビジネス・ラウンドテーブル議長）

リサ・スー、アドバンスド・マイクロ・デバイス（AMD）会長兼 CEO

ニコル・ターナー・リー博士（ブルッキングス研究所シニアフェロー兼技術革新センターディレクター）

キャシー・ウォーデン（ノースロップ・グラマン会長兼 CEO 兼社長）

マヤ・ワイリー（市民と人権に関するリーダーシップ会議会長兼 CEO）



エグゼクティブサマリー

人工知能（AI）システムは、米国の重要インフラを変貌させ、新たな機会を引き出し、重要なシステムやサービスに新たなリスクをもたらしている。AI システムの開発方法、アクセス方法、より大規模なシステム内での機能に関して、組織や個人がどのような選択をするかによって、AI が米国の重要インフラの幅広い分野に展開されたときに与える影響が決まる。こうした選択に資するため、国土安全保障省（DHS）は、人工知能安全・セキュリティ委員会（Artificial Intelligence Safety and Security Board、以下「委員会」）との緊密な協議の下、「重要インフラにおける AI の役割と責任に関する枠組み」（以下「枠組み」）を策定した。この枠組みは、米国の重要インフラにおける AI の安全かつ確実な開発・展開のための主要な役割と責任を推奨している。

この枠組みは、ホワイトハウスの「自主的約束」、「AI 権利章典のための青写真」、「人工知能の安全、安心、信頼できる開発と使用に関する大統領令 14110」、「ガバナンスの促進に関する OMB M-24-10 覚書」、「人工知能の使用における米国のリーダーシップの促進に関する覚書」、「AI 安全研究所の活動」、「DHS の人工知能の安全およびセキュリティに関するガイドライン」によって確立された AI の安全およびセキュリティのベストプラクティスを補完し、前進させることを目的としている、人工知能の政府機関利用のためのガバナンス、イノベーション、リスクマネジメントの促進に関する OMB M-24-10 覚書、人工知能における米国のリーダーシップの促進に関する覚書、AI 安全研究所の活動、重要インフラ所有者および運営者のための DHS 安全・セキュリティガイドラインなど。

重要インフラにおける AI に関連する役割、責任、ユースケース、リスクは複雑であり、相互に依存し合っており、技術の進化とともに時間とともに変化する。こうしたことを考慮し、本枠組みを策定した：

- 米国の重要インフラにおける AI の安全・安心な利用について、クラウドやコンピュート・インフラのプロバイダ、AI 開発者、重要インフラの所有者・運営者、市民社会、公共部門という 5 つの重要な役割に分けた自主的な責任を提案する。
- 環境のセキュリティ確保、責任あるモデルとシステム設計の推進、データガバナンスの実施、安全でセキュアな展開の確保、重要インフラのパフォーマンスと影響の監視という 5 つの責任分野にわたって、これらの役割を評価する。
- 国内の 16 の重要インフラ部門に展開される AI システムの安全性、セキュリティ、信頼性を高めるための技術的、プロセス的な提案を行う。

AI のエコシステム全体で採用され、実施されれば、この枠組みは、安全性とセキュリティの実践の調和を含む、重要インフラにおける AI の安全性とセキュリティを促進し、重要なサービスの提供を改善し、事業体間の信頼と透明性を高め、市民権と市民的自由を保護し、重要インフラが責任を持って AI を運用し、展開することを可能にする AI の安全性とセキュリティの研究を促進することを意図している。

この枠組みは、事業者が特定のシステムやアプリケーションに AI を使用することで、重要なインフラ資産、部門、国家的に重要なシステム、またはそのようなシステムによってサービスを受ける個人に危害が及ぶ可能性があるかどうかを評価することを可能にする既存のリスク枠組みに基づいている。この枠組みにおける責務は、技術的リスク低減、説明責任の仕組み、日常的なテストの実施、インシデント対応計画の実施を通じて、これらの潜在的な危害に対処するように調整されている。重要なことは、この枠組みは、AI の安全性とセキュリティの重要な要素として、透明性、コミュニケーション、情報共有の役割を優先していることである。この枠組みは、AI の安全・安心に関する関係事業者の責任の完全なリストを示すものではなく、安全・安心な利用を進めるための活動に焦点を当てたものである。

AI の中でも特に重要インフラに関連するもの。クラウドやコンピューティング・インフラのプロバイダから AI 開発者、重要インフラの所有者や運営者に至るまで、すべての関連事業者は、この枠組みが提示する基本的責任に加えて、分野別および文脈別の AI リスク低減を検討すべきである。

AI は依然として新たな技術であり、AI の安全性とセキュリティの実践は並行して発展し続けている。DHS と AI 安全・セキュリティ委員会は、わが国の重要インフラにおける米国のイノベーションの未来を守り、形成するための共通のコミットメントとして、本枠組みの中核概念を支援、強化、前進させるため、率先垂範して行動する。

I. 序文

アメリカの重要インフラとは、アメリカ社会における 16 のセクターを指し、そのシステムは非常に重要であるため、その機能停止は国家生活に壊滅的な影響を与える。¹これには、国防、エネルギー、輸送、情報技術、金融サービス、食料・農業、コミュニケーションなどの米国のシステムが含まれる。

米国国土安全保障省（DHS）は、そのパートナーとともに、米国人の家庭やビジネスへの電力供給、金融取引、情報共有、医療へのアクセスと提供、食卓への食料供給など、多くの日常生活を支える手段の保護に貢献している。AI は、重要インフラが提供するサービスを改善し、レジリエンスを構築し、脅威を検知し、災害復旧を支援する強力な力となり得る。こうした重要インフラ・システムを所有し、運用する事業者が AI をますます採用するようになる中、こうしたシステムや、それらがサービスを提供する消費者に悪影響を及ぼしかねないリスクを理解し、予測し、対処することは、同省の義務である。この義務には、重要インフラ・システムが米国のすべての人々のために機能し、米国に貢献することを保証することも含まれる。重要インフラにおける AI システムも例外ではなく、効果的で、プライバシー、市民権、市民的自由を尊重し保全する方法で意図的に設計、開発、展開されなければならない。²同省は、重要インフラにおける安全で確実な成果を推進するための必要条件として、これらの権利を保護するためにその役割を果たす。

米国の 16 の重要インフラ部門

化学	商業施設	通信	重要な製造事業者
ダム	防衛産業基盤	緊急サービス	エネルギー
金融サービス	食品・農業	政府施設	医療と公衆衛生
情報技術	原子炉	輸送システム	上下水道

これらのシステムの重要な性質と、米国民にサービスを提供するために AI を使用する役割を認識し、2023 年 10 月に大統領が出した「人工知能の安全、安心、信頼できる開発と使用に関する大統領令」（大統領令 14110）は、国土安全保障長官に AI 安全・セキュリティ委員会の設置を指示し、AI 使用に関する「助言、情報、勧告」を長官と重要インフラ・コミュニティに提供することを任務とした。³その第一歩として、DHS は同委員会と緊密に協議しながら、重要インフラにおける AI の役割と責任に関する自主的な枠組み（以下、「枠組み」）を策定した。

* 重要インフラへの AI の安全、安心、レジリエンスな開発と展開のための AI エコシステム。*

II. 範囲

この枠組みは、実践の標準を示すものであり、いかなる法的権利、特権、要件をも確立するものではない。

この枠組みは、重要インフラにおける AI の安全・安心な利用について、共有と分離の責任モデルを提案するものである。この目的のため、枠組みは以下の通りである：

- AI を開発・展開する際に、**重要なインフラシステム**およびそれによってサービスを受ける人々に対する危害のリスクを低減するためのリスク低減およびユースケースに基づく低減を推奨する。
- 国の重要インフラの多くが現在依存している、あるいは近い将来依存することになる AI を活用したサービスの開発・展開において、**クラウドコンピューティング・サービスのプロバイダ、AI モデル開発者、重要インフラ所有者・運営者**の役割にまたがる一連の自主的責任を提案する。
- これらの重要なシステムを利用したり、その影響を受けたりする人々のためのアドボカシー活動、新技術の様々な側面を改善するための研究支援、強力なリスクマネジメントの推進において、**市民社会と公共部門**が果たすべき一連の自主的責任を提案する。
- 事業者が、特定のシステムやアプリケーションに AI を使用することで、重要インフラ資産やセクター、あるいは米国民にサービスを提供するその他の国家的に重要なシステムに危害を及ぼしかねない深刻なリスクがあるかどうかを評価できるようにするため、既存のリスク枠組みに依拠する。これらのリスクカテゴリー間の関係やその低減に関するさらなる研究は、事業者がユースケース・ベースでこの評価を実施するのに役立つだろう。

さらに、この枠組みは、大統領令第 14110 号に基づいて策定された重要インフラの年次 AI セクター別リスクアセスメントプロセスや、近々予定されている国家インフラリスク管理計画など、DHS が調整する AI および重要インフラセキュリティプログラムから収集された情報を補完し、活用するものである。

III. 重要インフラに対する AI のリスク

この枠組みでは、重要インフラに対する潜在的な脅威、危険、脆弱性、影響という文脈で安全・セキュリティリスクを参照し、AI の開発・展開のライフサイクル全体で実施可能なリスク低減策を提示する。これらの要因に関する我々のアセスメントは、過去に発表された米国政府の覚書や報告書、特に「重要インフラの安全・セキュリティガイドライン」、「重要インフラのセキュリティとレジリエンスに関する国家安全保障覚書」、OMB 覚書 M-24-10 や「AI 権利章典のためのホワイトハウスの青写真」による市民権と市民的自由に対するリスクに関するガイダンスなどを参考にしている。⁴。

DHS は、サイバーセキュリティ・インフラセキュリティ庁（CISA）を通じて、また他のセクター・リスクマネジメント機関（SRMA）と連携して、⁵、重要インフラ全体における AI の安全性とセキュリティの攻撃ベクトルと脆弱性を、AI を利用した攻撃、AI システムを標的とした攻撃、設計と実装の失敗という 3 つのカテゴリーに識別した。国民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、これらの脆弱性の性質を理解し、それらに適切に対処することは、単なる運用上の要件ではなく、国家的な急務である。

重要インフラのセキュリティとレジリエンスに関する国家安全保障覚書」（NSM22）は、潜在的な被害の規模と重大性に基づいて重要インフラに対するリスクを分類するアプローチを明確にしており、これによりリスクマネジメントの取り組みの優先順位付けが可能になる。以下では、重要インフラの AI 利用に関連する具体的なリスクについて、これらのカテゴリーをマッピングして説明する。リスク・カテゴリーは有害な影響をアセスメントし、規模の大きい順に記載している：

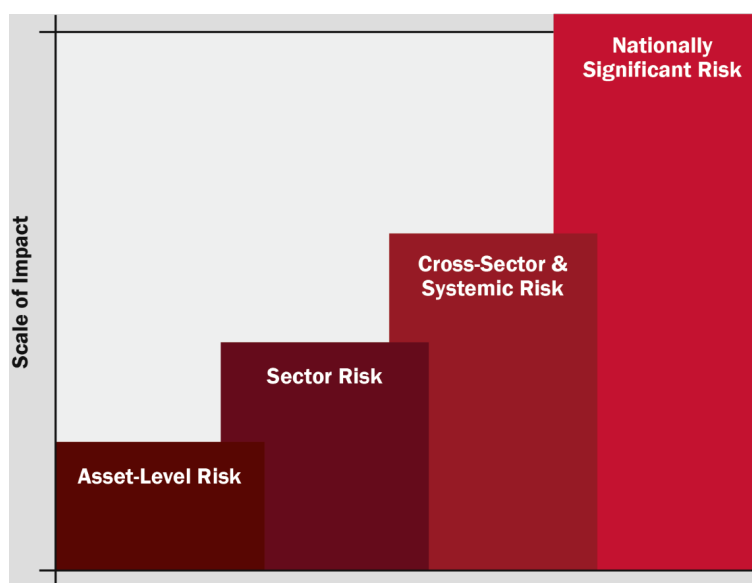
1. アセットレベルのリスク：重要イン

フラの運用、資産、システムに対する混乱や物理的損害、あるいはリスクの高い AI の使用から生じる直接的な供給者を含むが、これらに限定されない。これらのリスクには、住民や地域へのサービスを損なう設計や展開の欠陥が含まれる。

2. セクター・リスク：（個々の資産に

とどまらず、セクターの運営に影響を及ぼす一連の資産に対するリスクを含むが、これに限定されない）。

これらのリスクは、特に AI の運用障害から構成される 図 1. 図 1. 枠組みのリスクカテゴリーは、エネルギーや潜在的な被害の規模に応じて展開されるシステムを調整する。水道事業や、（病院が提供する医



療サービスや銀行やその他の金融機構が提供する金融サービスなどの) 重要なサービスの提供の中断、AI の悪用によって可能になる選挙インフラ・サブセクターの標的化。

3. システミックリスクとクロスセクターリスク：これには、重要なサービスへのアクセスを制限する AI を活用したサイバー攻撃やインシデントによる情報技術部門の混乱、AI 導入の拡大に伴う環境への悪影響、重大な財務的損失をもたらす AI 事故、重要インフラ事業者のサービスへのアクセスを大幅に妨げる AI を活用したセンシング技術のエラー、AI を活用したプロセスの失敗による物流サプライチェーンの混乱、重要インフラの相互依存性の高まりに起因するその他のリスクなどが含まれるが、これらに限定されるものではない。

4. 国家的に重大なリスク：(AI が安全または権利⁽⁶⁾に広範な影響を及ぼすリスク、あるいは通常兵器、化学兵器、生物兵器、放射性兵器、核兵器(CBRN)の開発、製造、または展開を著しく支援するリスクを含むが、これらに限定されない)。

この枠組みは、関連する活動を行う事業者が実施すれば、各リスクカテゴリーに関連する結果の可能性と重大性を低減できる低減策を提案している。さらに、このリスクの枠組みは、これらのカテゴリーが相互に依存しあっていることを明らかにするものであり、資産レベルのリスクを放置しておく、セクター全体またはセクター横断的なリスクへと複合化する可能性がある。逆に、重要資産の安全性またはセキュリティを改善するために設計された低減策は、国家的に重大な結果の可能性を防止または低減する可能性がある。また、AI モデルがどのように開発され、どのようにアクセスされ、より大きなシステムの中でどのように機能するかという様々な選択が、米国の重要インフラの幅広い分野に展開されたときに与える影響にとって重要であることも認識している。公共部門と市民社会は、この影響を理解し形成する上で極めて重要な役割を担っており、その結果、利益を部門間で共有し、危害を予防、低減し、必要に応じて是正することができる。

IV. 重要インフラにおける AI の安全かつ確実な展開のための役割と責任

A. 概要

したがって、AI の安全性とセキュリティリスクをマネジメントする責任は、必然的に複数の組織に分散することになる。この枠組みでは、クラウドや計算インフラのプロバイダ、AI モデル開発者、重要インフラの所有者や運営者、市民社会、公共部門が、「環境の安全確保」、「責任あるモデルやシステム設計の推進」、「データガバナンスの実施」、「安全でセキュアな展開の確保」、「パフォーマンスと影響のモニタリング」という 5 つの基本カテゴリーにおいて、共有する責任と個別に負う責任を具体的に評価する。

前述の事業体間の相互依存は、信頼と共通理解を促進するためのコミュニケーション・チャンネルとデータ共有を活用し、透明性を確保するための協調的努力によって対処されなければならない。例えば、重要インフラ事業体は、AI の開発者が、モデルの設計とテストを通じて、関連する安全性とセキュリティのリスクがどのように対処されたかについての情報を顧客に提供すれば、モデルが安全でセキュアであるかどうかをより適切に評価できるようになる。⁷同様に、モデル開発者やサービス・プロバイダは、クラウドやコンピューティング・インフラのプロバイダがハードウェアやソフトウェアのコンポーネントやソース・ベンダーに関する文書を共有すれば、IT 環境に対するリスクをより効率的に評価し、十分な情報に基づいた調達決定を行うことができる。⁸さらに、重要なインフラ事業体は、AI の展開について、関連する背景やそのプロセスも含めて透明性を提供することで、AI 開発プロセスの改善と情報提供において中心的な役割を果たすことができる。透明性は、この枠組みで特定された事業体をつなぎ、それぞれが安全とセキュリティに対するそれぞれの責任を果たすことを可能にするという重要な目的を果たす。

さらに、事業体は AI のライフサイクルに関連する複数の役割を果たすこともある。最近の AI の発展により、モデルのコンフィギュレーションやツール化がますます可能になり、テクノロジープロバイダーと顧客との距離が縮まっている。AI アプリケーションを構築する企業は、それに依存するソフトウェアを所有することもある。重要なインフラ事業体は、特定の特注サービス用に既製の AI モデルを微調整することができる。AI のサプライチェーン全体でサービスと能力の重複が増加しているため、AI の安全性とセキュリティの中核的責任がどこにあり、誰にあるのかに関しても不確実性が生じている。⁹この枠組みは、各事業体が AI の安全・安心に関する自らの義務を評価するのに役立つ勧告を提供する一方で、すべての事業体が互いに協力してコミュニケーションを図り、共有された責任を確実に果たすことを奨励するものである。

B. 主要用語と構成

AI エコシステム全体の事業者は、重要インフラにおける AI の安全かつ確実な開発・展開のために、共有された明確な責任を有する。本枠組みでは、役割と責任を以下のように定義する：

- ▶ **事業者**とは、クラウドやコンピュータ・インフラのプロバイダ、AI モデルの開発者、重要インフラの所有者や運営者、市民社会組織、公共部門の政府などを指す。
- ▶ **役割**は広範に定義され、事業者または事業者のグループが、重要インフラのユースケースを含むさまざまな用途の AI の開発と展開をサポートするために提供する中核的なサービスを包含する。クラウドや計算インフラのプロバイダ、AI 開発者、重要インフラの所有者や運営者、市民社会、公共部門などが含まれる。それぞれの役割は複数の事業者を包含する場合があるが（例えば、AI 開発者にはアプリケーション開発者だけでなくモデル開発者も含まれる）、これらの事業者は、重要インフラにおける AI の安全性とセキュリティに対する責任の類似性に基づいてグループ化されている。
- ▶ **防御**には、技術的リスク低減、説明責任と透明性の仕組み、権利関連の保護、テストベンチマークが含まれる。この枠組みで説明されているように、責任は事業者が行うタスクのようなものであり、以下の共有目標にグループ化されている：環境の安全確保、責任あるモデルとシステム設計の推進、データガバナンスの実施、安全でセキュアな展開の確保、パフォーマンスと影響の監視。
- ▶ 事業者は、重要インフラにおける AI の安全かつセキュアな開発・展開に関連する具体的な活動だけでなく、この枠組みにおけるどの責務が自らの役割全体に関連するかを評価する必要がある。
- ▶ 事業者は、技術開発、調達、及びコンプライアンスプロセスに組み込まれたものを含め、これらの責任を現在又は開発中の AI ガバナンスプログラムに採用し、組み込むべきである。本枠組みは、事業者が他に従う既存の法的又は規制的要件に取って代わるものではない。

C. 枠組み

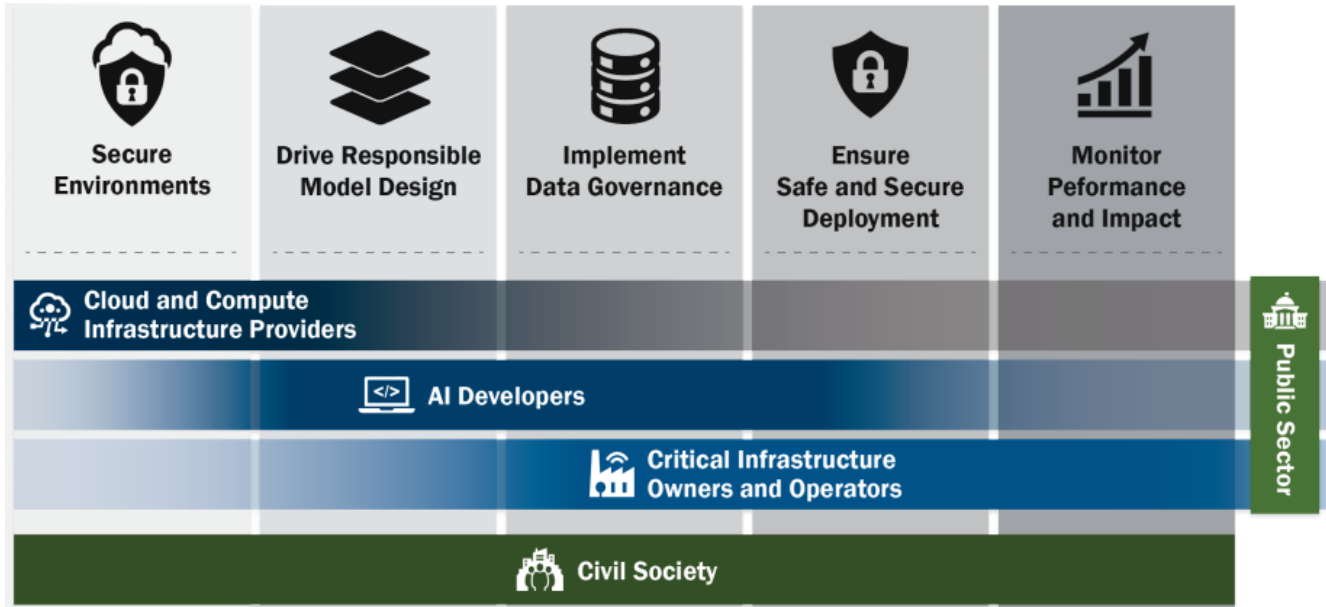


図 2.この枠組みは、責任の共有と分離のモデルの中でベストプラクティスを推奨している。クラウド・コンピューティング・インフラのプロバイダ、AI 開発者、重要インフラの所有者・運営者は、この図の上部にある AI の安全・安心に関する 5 つの実質的な各分野において責任を持つが、その主要な責任は、この図中のラベルの位置によって示されている。枠組みは、公共部門が、すべての部門にわたる関連する民間事業者が個人とコミュニティの権利を適切に保護していることを確保する責任を負い、市民社会がすべての実質的な AI の安全・安心分野にわたる研究と個人の権利を促進するのを支援することを推奨している。

I. クラウドおよびコンピューター・インフラストラクチャ・プロバイダ

クラウドおよびコンピューター・インフラストラクチャ・プロバイダは、AI モデルの構築、調整、実行に必要なオンデマンドでスケーラブルなコンピューティング・リソースへのアクセスを可能にする。また、顧客はこれらの事業者からインフラを調達し、自社内でモデル開発と展開をホストすることもできる。これらの事業者の責務はいくつかのカテゴリに分類されるが、この枠組みでは、データの完全性、可用性、機密性を確保することで、重要なインフラストラクチャに対して、信頼性が高く、レジリエンスが高く、セキュアバイデザインの AI 製品やサービスを提供する方法に関する提言に焦点を当てる。

クラウドおよびコンピューティング・インフラストラクチャ・プロバイダの責任 概要				
A.安全な環境	B.駆動責任モデルとシステム設計	C.データガバナンスの導入	D.安全で確実な展開の確保	E.パフォーマンスと影響を監視する
1.ハードウェアとソフトウェアのサプライヤーを吟味する 2.アクセス管理のベストプラクティスを導入する。 3.脆弱性管理の確立 4.物理的セキュリティの管理	1.脆弱性の報告	1.データの機密保持 2.データの可用性の確保	1.システムテストの実施	1.異常な活動を監視する 2.インシデントに備える 3.有害な活動を報告するための明確な経路を確立する。

A. 安全な環境

1. ハードウェアとソフトウェアのサプライヤーを吟味する：クラウド及びコンピューターインフラストラクチャのプロバイダは、その構成要素の信頼性とセキュリティを確保するために、サプライチェーンにおけるハードウェア及びソフトウェアをレビューすべきである¹⁰。サプライチェーンリスクマネジメントの手法は、ハードウェア部品表枠組み、ソフトウェア部品表枠組み、及びソフトウェア取得ハンドブックに示されており、状況に応じて、どのような情報を収集することが適切であるかについての指針を提供している¹¹。
2. アクセス管理のベストプラクティスを導入する：クラウドおよびコンピューターインフラストラクチャのプロバイダは、すべてのユーザー、デバイス、およびアプリケーションによるシステム、モデル、

またはデータソースへのアクセスを監視および管理するためのベストプラクティスを有効化または実装すべきである。

3. 脆弱性管理の確立：クラウド・コンピューティング・インフラのプロバイダは、脆弱性管理手法を利用して、AIに起因する脅威を含むインフラをスキャンするか、顧客がスキャンできるようにすべきである。プロバイダは、外部機関によるソフトウェア・セキュリティ・レビュー、侵入テスト、監査を実施するか、実施できるようにし、重要インフラ・アプリケーションにおける潜在的なサプライチェーン攻撃に対するレジリエンスを構築すべきである 12。
4. 物理的セキュリティの管理：クラウド及びコンピュートインフラストラクチャのプロバイダは、重層的な物理セキュリティモデルを確立すべきである。状況は様々であろうが、一般的に成功する物理的セキュリティモデルには、a) 周囲のフェンスと車両バリア、b) 電子アクセスカード、c) アクセスログ、d) 24時間365日の活動監視、e) 改ざんに対するサーバーとハードウェアの堅牢化が含まれる 13。

B. 駆動責任モデルとシステム設計

1. 脆弱性の報告：関連する場合、クラウドプロバイダおよびコンピュートインフラストラクチャプロバイダは、モデルおよびシステム設計プロセスに影響を及ぼす可能性のある脆弱性を報告する際、標準的かつ協調的な脆弱性プロセスに従うべきである。

C. データガバナンスの導入

1. データの機密保持クラウドおよびコンピュートインフラストラクチャのプロバイダは、静止時および転送時のデータを暗号化するか、暗号化方法を有効にすることによって、モデルの訓練や微調整に使用される個人データや顧客の機密データが暴露、漏えい、攻撃されるリスクを低減すべきである。そのための方法としては、顧客が暗号化を利用できるようにすることや、データ保護のためのその他のベストプラクティスを採用することが挙げられる。
2. データの可用性を確保する：クラウドやコンピューティング・インフラのプロバイダは、高可用性ネットワーク（人手を介することなく常に最適なレベルで動作するネットワーク）やバックアップ計画を顧客と緊密に連携して活用し、重要なサービスにおけるレジリエンスを確保すべきである。

D. 安全で確実な展開の確保

1. システムテストの実施クラウド及びコンピューティング・インフラストラクチャのプロバイダは、様々な停電シナリオにおいて、サービスの継続的な可用性を確保するために、システム環境のテストを行うべきである。このセキュリティテストは、自社で実施することも、信頼できるサードパーティ・プロバイダーを通じて実施することもできる。可能であれば、クラウドコンピュートインフラ

トラクチャのプロバイダは、コンピュータ環境の運用上の安全性とセキュリティテストを容易にするためのツールと診断を顧客に提供すべきである。

E.パフォーマンスと影響を監視する

1. 異常なアクティビティの監視クラウドおよびコンピュータ・インフラストラクチャ・プロバイダは、適切なツールを使用して、潜在的な脅威や不正使用について、ネットワーク・アクティビティを分析するか、顧客が分析できるようにすべきである。
2. インシデントに備える：クラウド及びコンピュータ・インフラの事業者は、プロバイダと協力して、システム又はデータへの不正アクセスを伴うインシデントが発生した場合のエスカレーション、調査、復旧及びコミュニケーションのプロセスを計画すべきである。サイバーセキュリティ、物理的セキュリティ、インサイダー脅威、その他のインシデントが原因で不正アクセスが発生した場合、事業者はそのような計画を実行すべきである。
3. 有害な活動を報告するための明確な経路を確立する：クラウドプロバイダは、その顧客と協力して、疑わしい活動や有害な活動を報告するための明確な経路を確立し、州、地域、および連邦政府の報告要件を遵守し、官民の研究者と協力して、関連する危害を低減すべきである。関連する場合、事業者は情報共有分析センター（ISAC）などの既存のインシデント報告チャネルを活用すべきである。

II. AI 開発者

ここで AI 開発者とは、自社またはサードパーティーのプラットフォームサービスやツールを通じて、AI モデルやアプリケーションを開発、訓練、および/または AI モデルやアプリケーションへのアクセスを可能にする事業体として定義される。開発者は、自らモデルを開発したり、サードパーティ製モデルを修正したり、サードパーティ製モデルへのアクセスを可能にしたり、開発者が AI モデルを使用または設定できるようにするソフトウェアツールを提供したり、顧客のためにモデルを下流のアプリケーションに展開したりすることができる。これらの開発者は、重要インフラの AI ライフサイクル全体にわたって、特に、長期にわたって責任あるモデルとアプリケーションの設計、テスト、保守を推進することに関連する安全性とセキュリティの責任を負う。一方、上流の AI モデルや下流の AI アプリケーションへのアクセスや、モデルの重みが広く利用可能かどうかによって、特定の種類の AI 開発者に特別に適用される責任もある。この枠組みでは、AI モデル、プラットフォーム、アプリケーションの開発者であるこれらの事業体は、最終的に重要インフラ事業体によって使用される AI 技術の開発において同様の役割を果たすため、単一のカテゴリーに集約している。

デュアルユースモデルが重要インフラへの攻撃を実行するために意図的に悪用されることを防止することに関する本セクションの責任については、AI モデル開発者は、追加的かつより詳細な推奨事項については、AI Safety Institute のガイドライン案「Managing Misuse Risk for Dual-use Foundation Models」を参照することが推奨される。¹⁴

AI 開発者の職務概要				
A.安全な環境	B.駆動責任モデルとシステム設計	C.データガバナンスの導入	D.安全で確実な展開の確保	E.パフォーマンスと影響を監視する
1.モデルとデータへのアクセスを管理する 2.インシデント対応計画の作成	1.セキュア・バイ・デザインの原則を取り入れる 2.モデルの危険な能力を評価する 3.人間中心の価値観との整合性の確保	1.個人の選択とプライバシーを尊重する。 2.データとアウトプットの質を高める	1.モデルへのアクセスを管理する際には、リスクベースのアプローチを用いる。 2.AI が生成したコンテンツの区別 3.AI システムの妥当性確認 4.顧客と一般大衆に有意義な透明性を提供する。	1.AI モデルに異常な動きや敵対的な動きがないか監視する 2.リスクを識別し、コミュニケーションし、対処する。 3.独自のアセスメントをサポートする。

			5.現実世界のリスクと起こりうる結果を評価する。	
			6.脆弱性の報告と低減のためのプロセスを維持する。	

A. 安全な環境

1. モデルとデータへのアクセスを管理する：AI 開発者は、モデルの重み、学習データ、ソースコードなど、AI システムの全体的なセキュリティ態勢において重要な役割を果たす AI システムのコンポーネントが、不正アクセスから確実に保護されるよう支援する必要がある。****該当する場合、開発者は、機密情報へのアクセス、変更、および流出の試みを検知し、防止するための管理を実装する必要があります。**¹⁵
2. インシデント対応計画の準備：AI 開発者は、インサイダー脅威を含むインシデントに迅速に対応するために、明確な報告と評価のプロセスを作成すべきである。計画は定期的に見直し、研修や机上演習に組み込んで、従業員が効率的にリスクをアセスメントして特定し、エスカレーションの経路をたどり、インシデントの影響を食い止められるように準備すべきである。インシデントから学んだ教訓は、将来のインシデントへの対応を改善するために計画に取り入れるべきである。¹⁶

****** この責任は、一般の人々が使用や研究のために広くアクセスできるようにする場合、これらの構成要素を確保することについては立場をとらない。

B. 駆動責任モデルとシステム設計

1. セキュア・バイ・デザインの原則を取り入れる：AI 開発者は、CISA の「Secure by Design Pledge (セキュア・バイ・デザイン誓約)」などの確立された枠組みを用いて、セキュアな AI を構築するためのアプローチを導き、公に伝えることで、製品が当初からセキュリティに重点を置いて開発されていることを確認する必要がある。開発者は、セキュリティ目標に向けた進捗を反映させるために、一般に公開するセキュリティポリシーの定期的な更新を行い、関連するセキュリティ脆弱性の特定と緩和方法について従業員を訓練し、顧客が考慮すべき最新の情報を提供し、セキュリティのベストプラクティスに関する新たな研究やガイダンスを取り入れるべきである。¹⁷
2. モデルの危険な能力を評価する：AI モデル開発者は、自律的活動、物理科学、生命科学、サイバーセキュリティ、および関連する高リスクの文脈で展開された場合に重要インフラに影響を与える可能性のあるその他の能力に関連する能力を特定するための戦略を確立し、遵守すべきである。

3. 人間中心の価値観との整合性を確保する：AI モデル開発者は、AI モデルが人間の価値観や目標を反映したものであることを、可能な限り保証すべきである。¹⁸AI アプリケーションの開発者は、関連する市民社会と連携して、市民の権利、市民の自由、適用法を尊重する価値観にユースケースを合わせるべきである。¹⁹

C. データガバナンスの導入

1. 個人の選択とプライバシーを尊重する：AI 開発者は、AI システムのための効果的なデータ管理が、個人の法的権利、表明された選択、プライバシーに対する合理的な期待を考慮したものであることを保証すべきである。AI 開発者は、個人情報収集、処理、保持、移転する際に、データの最小化を含むプライバシーのベストプラクティスを実施し、適用される規制を遵守すべきである。²⁰
2. データとアウトプットの質を高める：AI モデルを訓練するための入力として使用されるデータソースの数と多様性を考慮すると、AI 開発者は、モデルの意図しない結果や有害な結果を防ぐために、モデルの入力の質を一貫して評価し、データのフィルタリング、微調整、分類など、出力の質と信頼性を高める方法を使用すべきである。

D. 安全で確実な展開の確保

1. モデルへのアクセスを管理する際には、リスクベースのアプローチを用いる：AI モデル開発者は、モデルウェイトを広く公開する前にリスクアセスメントを行うべきである。アセスメントでは、悪用のリスクに対して、AI の安全・安心研究への影響を含め、オープン化のメリットを考慮すべきである。²¹
2. 生成的 AI コンテンツの識別：技術的に可能で商業的に合理的な場合、AI 開発者は、コード、テキスト、画像、音声、映像など、AI が生成または操作したコンテンツが、法的要件や安全性のベストプラクティスと一致するように、生成された時点と時点で明確に識別でき、したがって人間が生成したコンテンツと区別できるようにすることが推奨される。AI 開発者は、これらの分野の研究が進むにつれて、評価方法を更新し続けるべきである。
3. AI システム使用の妥当性確認：AI 開発者は、可能であれば利用可能なベンチマークを使用して、一般的な信頼性と堅牢性をテストし、AI システムが通常の条件下や想定される幅広い条件下で計画通りに動作することを確認するのに役立つべきである。²²ベンチマークがまだ存在しない場合、AI 開発者は、アプリケーション固有のベンチマークを含め、関連するテスト、評価、妥当性確認、検証 (TEVV) アプローチに関する社内外の研究を支援すべきである。
4. 顧客と一般市民に対して有意義な透明性を提供する：AI 開発者は、重要インフラの顧客が自らリスクアセスメントを行い、AI をいつ、どのように利用するかについて、十分な情報に基づいた意思決定を行うことができるような情報を提供すべきである。AI 開発者が提供する情報には、訓練データ

やモデル・アーキテクチャに関する情報、関心のあるベンチマークでのパフォーマンス、危険な能力を検知するための評価結果、安全対策やセキュリティ対策を含むリスクマネジメントの詳細などが含まれる。²³

5. 現実世界のリスクと起こりうる結果を評価する：AIの開発者は、モデルやアプリケーションがリスクの高い状況で展開されることを意図または設計されている場合、有害な結果をもたらす可能性のあるバイアス、故障モード、脆弱性をテストする評価を受けるべきである。モデル開発者は、AIのレッドチーム、²⁴、対応する安全対策がリスクを閾値以下に低減するのに有効でない場合にモデルのさらなる開発や展開を制限するために使用できるリスク閾値などのテスト手順を含むリスクマネジメントポリシーを実施すべきである。²⁵ AIアプリケーションの開発者は、意図されたユースケースに対するパフォーマンスをアセスメントし、AIアプリケーションが不注意による格差の影響を生じさせる可能性があるかどうかを慎重に検討すべきである。
6. 脆弱性の報告と低減のプロセスを維持する：公共の利益と強固なセキュリティ・プロトコルに合致するように、AI開発者は、重要インフラに危害を及ぼす可能性のある既知の脆弱性に対して、適時に修復活動を行うべきである。事業者は、実行可能であれば、ISACやその他の適切なチャネルを通じて、顧客、パートナー、規制当局に改善活動を開示することを検討すべきである。

E. パフォーマンスと影響を監視する

1. AIモデルの異常行動または敵対的行動を監視する：AI開発者は、AIを利用したサイバー攻撃や敵対的な操作、モデル・ドリフト、データ・ポイズニングなど、重要インフラにリスクをもたらす異常な行動や悪意のある活動を示唆するような、侵害の指標（²⁶）を監視する、あるいは顧客が監視できるようにする必要がある。信頼性の高い特定の展開では、データの性質、プライバシー、またはセキュリティ上の懸念のために、開発者が顧客自身がこれらの機能を実行するのを支援することが適切な場合がある。
2. リスクを識別し、コミュニケーションし、対処する：AI開発者は、新たに識別されたモデルリスクを文書化し、コミュニケーションし、影響、可能性、利用可能なリソースおよび低減策に基づいてリスクに優先順位をつけるためのプロセスを確立すべきである。AI開発者は、バグ報奨金、脆弱性の報告とパッチ適用に関する定期的なスケジュール、顧客からの報告を受け分析するプロセスなどの手段を確立すべきである。
3. 独立したアセスメントを支援する：AIモデル開発者は、CBRN関連能力（²⁷）および/または重要インフラ事業者およびその消費者に対するリスクの高まりなど、国家的に重大なリスクを提示するモデルを、資格のある信頼できるサードパーティが評価できるようにすべきである。

III. 重要インフラ所有者および運営者

重要インフラの所有者や運用者は、重要システムの安全な運用と保守を管理しており、コスト削減、信頼性向上、効率性向上のために AI への依存度が高まっている。これらの重要インフラ事業者は通常、AI アプリケーションやプラットフォームと直接やり取りし、特定のユースケース向けに AI モデルを構成できるようにしている。AI のユースケースは機能面でもリスク面でもセクターによって大きく異なるが、AI モデルやシステムの展開方法は、重要なサービスだけでなく、そのようなサービスを利用する個人にとっても重要な安全・セキュリティ上の意味を持つ 28。

重要インフラ所有者および運営者の責任 概要				
A.安全な環境	B.駆動責任モデルとシステム設計	C.データガバナンスの導入	D.安全で確実な展開の確保	E.パフォーマンスと影響を監視する
1.既存の IT インフラを保護する	1.責任ある調達ガイドラインの使用 2.AI の使用事例と関連リスクの評価 3.安全機構の導入 4.適切な人的監督の確立	1.モデルの設定や微調整に使用される顧客データの防御 2.データ収集と利用の管理	1.サイバー衛生の維持 2.透明性と消費者の権利の提供 3.AI に対する安全、安心、説明責任の文化を構築する。 4.労働力の育成	1.インシデント対応計画における AI の考慮 2.パフォーマンスデータの追跡と共有 3.定期的および事故関連の TEVV を実施する。 4.影響を測定する 5.システムの冗長性を確保する

A. 安全な環境

1. 既存の IT インフラを保護する：重要インフラ事業者は、展開環境ガバナンスの管理、堅牢なアーキテクチャの確保、構成の堅牢化、脅威からのネットワークの保護など、適用可能で AI システムが展開される場所で、国際的に認められた標準と慣行（²⁹）を適用すべきである。³⁰

B. 駆動責任モデルとシステム設計

1. 責任ある調達ガイドラインを使用する：重要インフラ事業者は、AI 開発者に対し、その AI 製品とサービスが、運用と主題の専門家によってテスト、評価、妥当性確認、検証されていることを確認すべきである。事業者は、「需要に応じた安全な」アプローチを実践し、サイバーセキュリティ、プライ

バシー、データ完全性（データの正確性と一貫性を含む）の標準を満たすようにエンタープライズと製品の両方のセキュリティを評価すべきである。³¹。

2. AI のユースケースと関連リスクを評価する：これには、a)AI アプリケーションの意図された目的と予想されるユースケース、b)AI アプリケーションを使用する、または AI アプリケーションに依存するシステムに関連する起こりうる故障モード、c)重要インフラに関連するバイアスと公平性の問題に関連する影響と低減が含まれる。
3. 安全メカニズムを導入する：重要インフラ事業者は、重要インフラ資産または従業員に深刻なリスクをもたらす自動意思決定プロセスによって引き起こされる可能性のある、安全に影響を与える結果の潜在的な重大性を防止または軽減するための制御を AI システムを導入すべきである。
4. 適切な人的監視の確立：重要インフラ事業者は、重要インフラ資産、サービス、または消費者に悪影響を及ぼす可能性のある結果的な決定を行う、またはその決定に情報を提供するために、適切な人間の関与を取り入れるべきである。そのような意思決定について、事業者は、意味のある人間の監視を確立し、所有者および運営者が AI 生成的出力、予測、および/または予報に依存すべき範囲を特定すべきである。

C. データガバナンスの導入

1. モデルの設定または微調整に使用される顧客データの防御：専有データまたは個人データの使用に関連するリスクを軽減するために、重要インフラ事業者は、特にモデルの訓練または微調整を行う際にそのようなデータが使用される場合、不適切なエクスポージャーから顧客データを保護すべきである。事業者は、関連する顧客の認可や同意と整合性を保ちながら、目前の特定業務に必要な限度の顧客データのみを収集、処理、保持、移転すべきである。
2. データの収集と使用を管理する：重要インフラ事業者は、AI モデルの微調整や AI アプリケーションの構成に使用される顧客データを追跡し、保護する必要がある。必要に応じて AI 開発者と協力し、モデルのトレーニングのために購入または調達した外部データセットの完全性（データの正確性、一貫性、セキュリティを含む）を検証し、現実的に可能であれば、所定のコンテキストでの使用の適合性を評価すべきである。³²

D. 安全で確実な展開の確保

1. サイバー衛生の維持：重要インフラ事業者は、CISA の Cyber Performance Goals に概説されているような強力なサイバーセキュリティの実践を実施し、AI システムの管理を維持すべきである。³³
2. 透明性と消費者の権利のプロバイダ：重要インフラ事業者は、国民に商品、サービス、便益を提供するための AI の利用に関して、有意義な透明性を提供すべきである。事業者は、知的財産の安全性を

確保しつつ、地方政府、地域社会、一般市民を含むステークホルダーと協力して、適切な情報開示の種類、レベル、頻度を決定すべきである。可能であれば、個人または地域社会に悪影響を及ぼす可能性のある AI システムを使用する事業者は、影響を受ける個人に対して、システムがどのように影響を及ぼすかについて説明を受ける機会を与えるべきである。

3. AI の安全、セキュリティ、説明責任の文化を構築する：重要システムに AI を使用する場合、重要インフラの所有者と運営者は、適切なガバナンス方針と手順に支えられた重要な意思決定に、経営幹部が関与するようにすべきである。
4. 従業員の訓練：重要インフラ事業者は、AI の適切な使用方法と、フィッシング攻撃、マルウェア、脆弱なデバイス、不十分なパスワード衛生、データ・ポイズニングなど、重要インフラ所有者や運営者の従業員に特に影響を与えるセキュリティ脆弱性について、従業員を訓練する必要がある。

E. パフォーマンスと影響を監視する

1. インシデント対応計画において AI を考慮する：重要インフラに影響を与えるインシデントが発生した場合、重要インフラ事業者は、最新の一般的なインシデント対応計画を準備しておくべきである。この計画では、AI システムの運用を停止する方法、重要インフラ・サービスの継続的な可用性を提供するバックアップ・システムの運用を開始する方法、適切な政府認定機関に通知する方法、影響を受ける顧客やその他の利害関係者に適宜通知する方法、AI システムへのアクセスを安全に停止する方法を指示すべきである。事業者は、問題が発生した場合に AI システムをロールバックするための計画とプロトコルを策定すべきである。
2. パフォーマンスデータの追跡と共有：該当する場合、重要インフラ事業者は、モデルまたはシステム開発者と協力して、AI システムがどのように動作したかに関する情報および／またはデータを共有するためのプロセスを定めるべきである。また、重要インフラ事業者は、モデルの挙動と現実の結果との関係を開発者がよりよく理解できるように、そのパフォーマンスデータの評価結果を共有することも検討すべきである。このデータ共有は、個人データを保護する方法で行われるべきである。
3. 定期的及びインシデント関連のテスト、評価、妥当性確認、検証の実施：継続的なシステムモニタリング／リスクマネジメントプロセスの一環として、重要インフラ事業者は、修理やアップグレードを実施するためのテストや計測の十分性と有効性を定期的にアセスメントすべきである。³⁴
4. 影響を測定する：重要インフラ事業者は、AI が、そのモデルが統合されたシステム全体に与える影響を、継続的に測定すべきである。これには、そのシステムを消費したり、その他の形で影響を受けたりする個人やコミュニティに対する格差の影響も含まれる。
5. システムの冗長性を確保する：重要インフラ事業者は、自然災害、事故、その他の事象による中断の影響を最小化するために、冗長性を組み込むべきである。

IV. 市民社会

市民社会とは、産業界や政府とは異なる組織であり、非政府組織、労働組合、慈善団体、専門家団体、財団、学術機関、研究機構などを含む。市民社会組織の役割は、民間の AI 企業と関わる非政府組織から、イノベーション文化に貢献する研究機構に至るまで多様である。セクターとして、市民社会は、政府や産業界とのパートナーシップの下で、個人やコミュニティに対する技術の影響を測定し、伝えるのに役立つ標準、枠組み、そしてソリューションに貢献している。市民社会は、AI のライフサイクル全体にわたる安全性、プライバシー、公平性、セキュリティ低減の支援、標準化、改善において重要な役割を果たし、公衆を保護し、信頼性を育む。

市民社会の責任 概要				
A.安全な環境	B.駆動責任モデルとシステム設計	C.データガバナンスの導入	D.安全で確実な展開の確保	E.パフォーマンスと影響を監視する
<ol style="list-style-type: none"> 1. 政府や産業界とともに、標準、ベストプラクティス、測定基準の開発とコミュニケーションに積極的に関与する。 2. 政策立案者と国民を教育する 3. AI システム開発と展開の指針を示す 4. プライバシー強化技術の使用をサポートする。 5. レッドチーム標準の重要インフラの使用例を検討する。 6. 研究とイノベーションの推進と支援の継続 				

1. 政府や産業界とともに、標準、ベスト・プラクティス、測定基準の開発とコミュニケーションに積極的に関与する：市民社会は、重要なインフラストラクチャーによる AI の使用のための AI の安全基準と、分野別のアプリケーションの影響を測定するための具体的な測定基準の開発とコミュニケーションを支援すべきである。これらの標準や測定基準は、開発段階におけるバイアスを定義し、評価し、監視するためのツールを備えた、権利を肯定するものでなければならない。
2. 政策立案者と一般市民を教育する：市民社会事業体は、AI の用途、利益、リスクについて政策立案者や一般大衆に情報を提供するのを助けるために、AI 教育リソースを開発したり、適切な貢献を特定すべきである。
3. AI システムの開発と展開のための指針となる価値観とセーフガードを知らせる：市民社会は、透明性があり、プライバシー、公民権、人権、社会の福利を保護する AI システムを開発・展開するため

に、適切な規則、方針、手続きを含む指導的価値観とセーフガードを策定し、クラウドや計算インフラのプロバイダや AI 開発者とともに実施するために、一般市民や政府と協力すべきである。³⁵

4. プライバシー強化技術の使用を支援する：市民社会は、該当する場合には産業界と協力し、AI に使用されるデータの収集、処理、および学習のためのプライバシー強化技術（PETs）の採用を促進する機会を特定すべきである。
5. レッドチーム標準のための重要インフラのユースケースを検討する：市民社会は、様々なユースケースやリスク閾値にわたって重要インフラが容易に採用できる実用的なレッドチーム標準を開発するために、該当する場合には AI 開発者と関わるべきである。
6. 研究とイノベーションの推進と支援を継続する：市民社会は、重要なインフラにおける AI の安全な開発と展開を支援する AI アプリケーションの基礎研究と、AI における重要な社会技術的概念を推進すべきである。そのような研究は、平等、公平、民主的価値に焦点を当て、責任あるイノベーションの育成を目指すべきである。

V. 公共部門

公共部門（連邦、州、地方、部族、地域の事業者の政府機関を含む）は、米国民とその機構に奉仕し、保護する。このような重要な責務に鑑み、本枠組みは、重要インフラにおける AI の安全かつ確実な利用を促進するための取り組みと同様に、公共事業者自身の AI36 の利用という観点から、公共事業者に対して以下の実践を推奨する。

公共部門の役割と責任	
クラウドおよびコンピューティング・インフラストラクチャ・プロバイダ	1 必要不可欠なサービスと緊急対応を提供する 2 世界的なアル規範の推進 3 重要インフラの機能改善のために AI を責任を持って活用する。
AI 開発者	4 法律と規制を通じて実践の標準を推進する。
重要インフラ所有者および運営者	5 地域のリーダーを巻き込む 6 AI の安全性とセキュリティの基礎研究を可能にする 7 重要インフラの安全・安心な AI 導入をサポート 8 監督機能の開発

1. 必要不可欠なサービスと緊急対応を提供する：公共部門は、AI の利用が政府の中核的機能、すなわち必要不可欠なサービス、生命と安全、緊急対応、経済と地域社会の支援を支援し、決してそれと衝突しないことを保証すべきである。
2. グローバルな AI 規範を推進する：米国は AI の世界的リーダーであり、パートナーと協力して AI の安全・安心に関する強力な規範と標準の確立を主導する。連邦政府は AI に関して国際的なパートナーと協力し、共通の脅威を特定し、国際的な規制と標準を推進し、すべての地球市民を保護するために共有される責任を中心に収束すべきである。
3. AI を責任を持って活用し、重要インフラの機能を改善する：公共部門は、効率性を改善し、重要なサービスの価格と利用可能性を高め、透明性と国民へのコミュニケーションにおいて模範を示すべきである。政府サービスにおける責任ある AI の実践を促進するプログラムの開発とそのための資金提供を優先すべきである。公共事業者は、公共部門による AI の利用に関して市民社会と相互に関与し、差別的な結果を生み出したり、個人のプライバシーを侵害したり、その他の法的権利を侵害するような形での AI の利用を避けるべきである。公共事業者は、差別的な技術に資金を提供すべきではない。

4. 法律と規制を通じて診療の標準を推進する：連邦政府には、法律や規制を通じて標準を推進する機会がある。その際、政府は、特に AI のダイナミックで急速に進化する状況を踏まえ、イノベーションを阻害しないようにしなければならない。法規制は、個人の基本的権利を保護し、イノベーションを促進し、異なる法的要件の調和を進め、コンプライアンスを簡素化し、インシデント報告プロセスを明確にすべきである。
5. 地域社会のリーダーを巻き込む：公共部門は、地域社会の指導者と協力し、AI が住民、労働力、機構に与える影響を、特に脆弱性に重点を置いて予測・理解すべきである。連邦政府は地方自治体と連携し、影響を測定、理解し、連邦政府の政策に反映させるべきである。
6. AI の安全性とセキュリティに関する基礎研究を可能にする：公共部門は AI 開発者や研究者と協力し、AI の技術的進歩が金融、教育、医療などの重要なサービスの安全性とセキュリティにどのような影響を与えるかをテスト、分析、監視すべきである。公共部門のリーダーは、学術機構と提携し、National AI Research Resource (NAIRR) のような AI の資金調達イニシアチブを活用して、業界やセクターを横断する強力な AI の安全・安心の実践に関する標準を構築する取り組みを主導し、参加すべきである。
7. 重要インフラによる AI の安全・安心な導入を支援する：公共部門は、重要インフラによる AI の責任ある使用が有益である場合には、その使用を支援することを明確にし、AI の使用が不適切である状況を特定する。
8. 監督体制の整備：DHS および理事会は、この枠組みの継続的な妥当性を評価し、その監視およびアセスメントメカニズムを公表すべきである。

V. 結論

最近の AI の機能は、関連するリスクを効果的にマネジメントすることができれば、重要インフラの機能を向上させる並外れた可能性を提示している。この枠組みは、セクター、産業、政府を超えたリーダーが、組織内および他者との相互作用の一環として、AI の安全・セキュリティに対する責任を共有・分担することで、この分野の発展に貢献するための基盤を提供するものである。この枠組みは、とりわけ、AI の安全・安心の実践の調和をさらに強化し、AI によって実現される重要なサービスの提供を改善し、AI のエコシステム全体の信頼と透明性を高め、重要インフラ向けの安全・安心な AI の研究を進め、すべての事業体によって市民権と市民的自由が確実に保護されるようになれば、成功する。

米国の重要インフラが持つ並外れた規模と影響力は、わが国の中核的価値観に合致した、重要システム向けの AI を開発・展開するためのアプローチで足並みを揃えることの重要性を強調している。この枠組みは、ホワイトハウスの「AI 権利章典のための自主的コミットメントと青写真」、AI に関する OMB M-24-10 覚書、AI 安全研究所の活動、重要インフラ所有者および運営者のための DHS 安全・セキュリティガイドラインなどによって確立された AI の安全・セキュリティのベストプラクティスを補完し、前進させることを目指す。

A. 枠組みの望ましい成果

- 1. 基盤となる安全・安心の実践の調和強化。** 最近の AI 能力の進歩により、AI の安全・安心に関するガイダンス、標準案、原則が急増している。しかし、事業体がこれらのリソースをどのように運用するかについて合意できなければ、陳腐化するリスクがある。この枠組みは、確立されたガイダンスと AI エコシステム全体の関係者の見識を活用し、共有と分離の責任モデルを提供するものであり、採用、実施、更新を行い、変化する技術に対応できるように設計されている。具体的には、この枠組みは、他の政府発行のガイダンスと並行して、企業方針、事業体間の合意、AI 開発の技術標準、および AI がどのように開発され、米国の重要インフラに展開されるかの重要な側面を最終的に支配する法的措置や規制措置に情報を提供するために使用されるべきである。さらに、米国の事業体がこの枠組みを広く採用することは、米国のリーダーシップとイノベーションの世界的影響力を考えれば、国際的な規範や標準の形成に役立つだろう。このような国内外での調和は、今後何年にもわたって重要インフラのレジリエンスを強化・構築する AI の力を定義するイノベーションの原動力として不可欠である。
- 2. 重要なサービスの計画と提供の改善。** 石油パイプラインや航空交通システムなどの重要なシステムを運用し、その安全性とセキュリティを確保するには、高度な計画、リアルタイムの監視、高度な予測など、複数のプロセスを慎重に調整する必要がある。これらの各プロセスは、個々の実行と協調の両面で、AI から大きな恩恵を受けることができる。これらのシステムの所有者や運営者は、重要なシステムに新進技術を導入するリスクを常に慎重に考慮しなければならないが、重要インフラ事業体は

歴史的に、米国民にサービスを提供するための新技術の研究、試験運用、展開をリードしてきた。今日、AI と生成的 AI における最近の進歩は、リーダーシップとイノベーションの新たな機会を提供すると同時に、リスクをアセスメントし、消費者のための重要な保護を組み込む必要性を明確にしている。重要インフラが AI システムを構築または調達する方法を管理する政策とプロセスに、この枠組みの関連する構成要素を導入することにより、所有者と運営者は、成長する AI の安全性とセキュリティの分野を形成し、肯定的な影響の例証と促進を支援することができます **結論**

- 3. AI エコシステム全体の信頼性と透明性の向上。**信頼と透明性は、あらゆる複雑なエコシステムにおいて手を携えて機能するが、AI も例外ではない。クラウドやコンピュータインフラストラクチャのプロバイダや AI 開発者を含むテクノロジープロバイダにとって、透明性を提供することは顧客の信頼と信用を構築するために不可欠である。さらに、彼らが共有する情報には意味があり、該当するリスクレベルに応じて調整されていなければならない。特定の高リスクの状況においては、テクノロジープロバイダーはさらに、AI システムが信頼できる方法で動作するという保証を提供できなければならない。事業体間の強固な透明性の実践は、重要インフラ事業体などの下流の事業展開者が、重要システムにおいて AI と直接的または間接的に相互作用する消費者からの懸念に対処し、これを解消することを可能にする。したがって、本枠組みでは、信頼と透明性を、事業体の役割や AI エコシステムに沿った立場によって異なる形をとる共有責任と捉えている。事業体は、これらの共有責任と個別責任を採用することで、重要な調達の意思決定を促進し、有意義な情報交換を可能にし、信頼されるサードパーティの関与を支援し、エコシステム全体の協力と共通理解を促進することができる。
- 4. 重要インフラにおける AI の安全性とセキュリティの成果に関する研究の推進。**今日の AI 研究コミュニティは、ますます高度化する AI モデルの斬新な能力とリスクの開発とテストを包含し、それに歩調を合わせるために急速に拡大している。さらなる研究開発は、最先端の AI モデルを活用しないものも含め、重要インフラが AI のユースケースを運用・展開し、特定の文脈における AI の適用から生じるリスクをマネジメントするのに役立つはずである。この枠組みは、エネルギー管理からヘルスケアに至るまで、重要なサービスにおける AI の利用に焦点を当てた研究を支援し、推進することで、これらの分野における AI の潜在的なユースケースや、AI モデルのアーキテクチャと実世界の成果との関係についての理解を深めることを、すべての事業体に対して求めている。さらに、このような研究は、効果的でリスクに基づいた AI 規制を策定するための取り組みに役立つ。
- 5. 公民権の尊重**重要インフラにおける AI の使用と、それに対応するコストと便益は、特定の用途、分野や使用ケースの背景、その他多くの要因によって異なる。とはいえ、プライバシー、市民的権利、市民的自由への配慮は基礎的なものであり、すべての AI システムに貫かれなければならない。したがって、この枠組みは、重要インフラにおける AI の開発と展開を支援する AI のエコシステム全体にわたって、公民権の保護、格差のある影響の特定、被害の低減を共通の責務とする。

附属書 A : AI の役割と責任マトリックス

	セキュアな環境	駆動責任モデルとシステム設計	データガバナンスの導入	安全で確実な展開	パフォーマンスと影響を監視する	公共部門
クラウドおよびコンピュータインフラストラクチャ・プロバイダ	<ul style="list-style-type: none"> - ハードウェアとソフトウェアのサプライヤーを吟味する - アクセス管理のベストプラクティスを導入する。 - 脆弱性管理の確立 - 物理的な管理 	<ul style="list-style-type: none"> - 脆弱性の報告 	<ul style="list-style-type: none"> - データの機密保持 - データの可用性を確保する 	<ul style="list-style-type: none"> - システムテストの実施 	<ul style="list-style-type: none"> - 異常な活動を監視する - インシデントに備える - 有害な活動を報告するための明確な経路を確立する 	<ul style="list-style-type: none"> - 必要不可欠なサービスと緊急対応を提供する - グローバルな AI 規範を推進する - AI を責任を持って活用し、重要インフラの機能を改善する。 - 法律と規制を通じて実践の標準を推進する。
AI 開発者	<ul style="list-style-type: none"> - モデルとデータへのアクセスを管理する - インシデント対応計画の作成 	<ul style="list-style-type: none"> - セキュア・バイ・デザイン原則を取り入れる - モデルの危険な能力を評価する - 人間中心の価値観との整合性を確保する 	<ul style="list-style-type: none"> - 個人の選択とプライバシーを尊重する - データとアウトプットの品質を促進する 	<ul style="list-style-type: none"> - モデルへのアクセスを管理する際には、リスクベースのアプローチを使用する。 - AI 生成的コンテンツの見分け方 - AI システムの妥当性確認 - 顧客と一般大衆に有意義な透明性を提供する。 - 現実のリスクと起こりうる結果を評価する。 - 脆弱性の報告と低減のためのプロセスの維持 	<ul style="list-style-type: none"> - AI モデルに異常な動きや敵対的な動きがないか監視する - リスクの識別、コミュニケーション、対処 - 独立したアセスメントをサポートする 	<ul style="list-style-type: none"> - コミュニケーション・リーダーの参加 - AI の安全性とセキュリティの基礎研究を可能にする - 重要インフラの安全・安心な AI 導入を支援する - 監督機能の開発
重要インフラ所有者および運営者	<ul style="list-style-type: none"> - 既存の IT インフラを保護する 	<ul style="list-style-type: none"> - 責任ある調達ガイドラインの使用 - AI のユースケースと関連リスク 	<ul style="list-style-type: none"> - モデルの設定や微調整に使用される顧客データの防御 - データ 	<ul style="list-style-type: none"> - サイバー衛生の維持 - 透明性と消費者の権利の提供 	<ul style="list-style-type: none"> - インシデント対応計画における AI の考慮 	

セキュアな環境	駆動責任モデルとシステム設計	データガバナンスの導入	安全で確実な展開	パフォーマンスと影響を監視する	公共部門
	<ul style="list-style-type: none"> の評価 - 安全メカニズムの導入 - 適切な人的監視の確立 	<ul style="list-style-type: none"> の収集と使用の管理 	<ul style="list-style-type: none"> - AI に対する安全、安心、説明責任の文化を築く - 労働力の育成 	<ul style="list-style-type: none"> - パフォーマンスデータの追跡と共有 - 定期的およびインシデント関連の TEVV の実施 - 影響を測定する - システムの確保 	
市民社会	<ul style="list-style-type: none"> - 政府や産業界とともに、標準、ベストプラクティス、測定基準の開発とコミュニケーションに積極的に関与する。 - 政策立案者と国民を教育する - AI システム開発と展開の指針を示す - プライバシー強化技術の使用をサポートする。 - レッドチーム標準の重要インフラの使用例を検討する。 - 研究とイノベーションの推進と支援の継続 				

附属書 B : 用語集

本書における多くの用語は、法律やガイダンスによって様々な定義がある。分かりやすくするため、ここでは参考となる用語をいくつか記載したが、事業者には、それぞれの活動に最も適した定義を使用するよう委ねる。

人工知能 (AI) : 人間が定義した目的に対して、現実または仮想環境に影響を与える予測、推奨、決定を行うことができる機械ベースのシステム。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。合衆国法典第 15 編第 9401 条(3)を参照のこと。

AI アプリケーション : 情報技術システム上でホストされるコンピュータプログラムであり、AI に関連する特定の一連のタスクまたは要件を実行するように設計されている。AI アプリケーションは開発者によって作成され、データやアルゴリズムの作成、変更、収集、処理、および/または提示を含むことができるが、必ずしもそうである必要はない。NIST, *Common Platform Enumeration* を参照 : *Naming Specification Version 2.3*, at 3 (Aug. 2011), <https://doi.org/10.6028/NIST.IR.7695>.

AI デプロイヤー : 認可を受けて AI システムを使用する事業者。エンドユーザーに影響を与える意思決定を行うために AI システムを使用することもあれば、AI システムを使用して AI モデル、アプリケーション、プラットフォーム、またはサービスを製造することもある。AI へのアクセスを可能にしたり、自らモデルを開発したり、組織が特定の用途のために AI モデルを使用・設定できるようにするソフトウェア・ツールを作成したりする。

AI モデル : AI 技術を実装し、計算、統計、または機械学習技術を使用して、与えられた入力セットから出力を生成する情報システムのコンポーネント。大統領令第 14110 条第 3 項(c)参照。

AI プラットフォーム : AI モデルやシステムを開発、インストール、修正、統合、配布、またはエンドユーザーが実行できるデバイス、オペレーティングシステム、または仮想環境を運営する事業者。AI プラットフォームは、そのサービスを確実に提供するために複数のテクノロジー、テナント開発者、および/またはエンドユーザーに依存するダイナミックなエコシステムである。AI プラットフォームは開発者のサブカテゴリであり、展開やサービス提供など、さまざまな AI システムサービスをサポートする可能性がある。NIST, *Trustworthy Platforms*, <https://www.nist.gov/trustworthy-platforms> を参照。

AI レッドチーム : 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、多くの場合、管理された環境で、AI の開発者と協力して行われる。大統領令第 14110 条第 3 項(d)を参照。

AI システム : 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。大統領令第 14110 条第 3 項(e)参照。

市民社会：特に非政府組織、労働組合、慈善団体、専門職団体、財団、学界、研究機構を含む、産業界や政府とは異なる組織。

クラウドおよびコンピューティング・インフラストラクチャ：AI モデルの構築、調整、実行に必要な、オンデマンドでスケーラブルなコンピューティング・リソースへのアクセスを提供する事業体。顧客は、オンプレミスでモデル開発と展開をホストするために、これらの事業体からインフラを調達することもできる。

重要インフラ：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。合衆国法律集第 42 編第 5195c 条(e)参照。重要インフラ事業体は、重要システムの安全な運用と保守を管理しており、その一部は AI モデルを使用している（一部は設計している）。

エンドユーザー：完成品またはサービスの最終消費者。合衆国法律集第 22 編第 8541 条(5)を参照のこと。

基盤モデル：広範なデータで訓練され、一般に自己監視を使用し、少なくとも数百億のパラメータを含み、広範な文脈に適用可能で、タスクにおいて高レベルのパフォーマンスを示すか、または示すように容易に修正可能な AI モデル。大統領令第 14110 条第 3 項(k)参照。

モデルカード：訓練された機械学習モデルに付随する短い文書で、意図された応用領域に関連する異なる文化的、人口統計的、または表現型のグループにわたるなど、さまざまな条件でのベンチマーク評価を提供する。モデルカードは、モデルの使用目的、性能評価手順の詳細、その他の関連情報を開示する。Mitchell et al., *Model Cards for Model Reporting*, 前掲注 7.

モデルの重み：AI モデル内の数値パラメータで、入力に対するモデルの出力を決定するのに役立つ。大統領令第 14110 条第 3 項(u)参照。

公的部門：連邦政府、州政府、地方政府、部族政府、および準州政府、ディビジョン、および役人。

附属書 C : 注意事項

- 1 連邦法は、重要インフラを「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、そのようなシステムおよび資産の機能不全または破壊が、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させるような影響を及ぼすもの」と定義している。合衆国法典第 42 編第 5195c 条 (e)。
- 2 一般的には、ホワイトハウス、AI 権利章典のための青写真 (2022 年 10 月)、<https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf>; 行政管理予算局、人工知能の政府利用のためのガバナンス、イノベーション、リスクマネジメントの推進に関する行政府・省庁長向け覚書、附属書 I (2024 年 3 月 28 日)、<https://www.whitehouse.gov/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Useof-Artificial-Intelligence.28,2024>), <https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Useof-Artificial-Intelligence.pdf> (OMB, M-24-10) (権利や安全に影響を及ぼすと推定される AI の用途リスト)。
- 3 ホワイトハウス、大統領令第 14110 号 : 安全、確実、信頼できる人工知能の開発と利用、§4.3(a)(v) (2023 年 10 月 30 日)、<https://www.whitehouse.gov/briefing-room/presidentialactions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-ofartificial-intelligence/>。
- 4 The White House, *National Security Memorandum on Critical Infrastructure Security and Resilience*, NSM-22, (Apr 30, 2024), <https://www.whitehouse.gov/briefing-room/presidential-actions/2024/04/30/national-security-memorandum-on-critical-infrastructure-security-and-resilience/>; DHS, *Safety and Security Guidelines for Critical Infrastructure Owners and Operators* (Apr 26, 2024), <https://www.dhs.gov/publication/safety-and-security-guidelines-critical-infrastructure-owners-and-operators>; OMB, M-24-10, *supra* note 2, § 2; The White House, *Blueprint for an AI Bill of Rights*, *supra* note 2.
- 5 重要インフラのセキュリティとレジリエンスに関する国家安全保障覚書 (NSM-22) は、16 の重要インフラセクターと関連するセクターリスクマネジメント局 (SRMA) を特定している。SRMA は、指定された重要インフラ・セクターの日常的な連邦政府のインターフェイスとして機能し、セクター固有のリスクマネジメントとレジリエンス活動を実施する。ホワイトハウス、NSM-22 (前掲注 4) 参照。
- 6 OMB, M-24-10, 前掲注 2, § 5 参照。
- 7 この文書は、"モデルカード"または"ファクトシート"の形をとることができる。Margaret Mitchell et al., *Model Cards for Model Reporting, Conference on Fairness, Accountability, and Transparency* (Jan. 29-31, 2019), <https://arxiv.org/pdf/1810.03993>; John Richards et al., *A Methodology for Creating AI Factsheets*, IBM (June 28, 2020), <https://arxiv.org/pdf/2006.13796> を参照。
- 8 この文書は「ハードウェア部品表」の形をとることができる。CISA, *Hardware Bill of Materials (HBOM) Framework for Supply Chain Risk Management* (Sept. 25, 2023),

<https://www.cisa.gov/resourcestools/resources/hardware-bill-materials-hbom-framework-supply-chain-risk-management> を参照。

- 9 組織の AI サプライチェーンを文書化することは、安全な開発に不可欠な要素である。CISA and NCSCUK, *Guidelines for Secure AI System Development*, at 12, (Dec. 13, 2023), <https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development> を参照。
- 10 John Pendleton et al., *Cloud Reassurance : A Framework to Enhance Resilience and Trust*, Carnegie Endowment for International Peace, (Jan. 2024), <https://carnegieendowment.org/research/2024/01/cloud-reassurance-a-framework-to-enhance-resilience-and-trust>.
- 11 NIST, *Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations* (May 2022), <https://csrc.nist.gov/pubs/sp/800/161/r1/final>; CISA, 前掲注 8; CISA, *Software Acquisition Guide for Government Enterprise Consumers : CISA, Software Acquisition Guide for Government Enterprise Consumers: Software Assurance in the Cyber-Supply Chain Risk Management (C-SCRM) Lifecycle* (Aug. 1, 2024), <https://www.cisa.gov/resources-tools/resources/software-acquisition-guide-government-enterpriseconsumers-software-assurance-cyber-supply-chain>. NIST, *Security and Privacy Controls for Information Systems and Organizations, SP 800-53, Revision 5* (Sept. 2020), <https://doi.org/10.6028/NIST.SP.800-53r5> も参照のこと。
- 12 CISA, *A Guide to Critical Infrastructure Security and Resilience* (Nov. 2019), <https://www.cisa.gov/resourcestools/resources/guide-critical-infrastructure-security-and-resilience> を参照。
- 13 ペンデルトン他、前掲注 10 参照。
- 14 NIST, *Managing Misuse Risk for Dual-Use Foundation Models, Initial Public Draft* (July 2024), <https://doi.org/10.6028/NIST.AI.800-1.ipd>. を参照のこと。
- 15 CISA and NCSCUK, 前掲注 9, at 14 参照。
- 16 NIST、前掲注 14 参照。
- 17 CISA, *Secure by Design Pledge* (2023), <https://www.cisa.gov/securebydesign/pledge> を参照。
- 18 例えば、AI モデルとプラットフォームの開発者は、適切な代表者データセットを使用し、異なる集団や人口統計に関してシステムがどのように動作するかを考慮し、バイアスやその他の交絡変数を特定し緩和するために展開前テストを実施することで、この整合を達成することができる。アライメントの規範的・技術的定義については、NIST, *The Language of Trustworthy AI: An In-Depth Glossary of Terms* (Mar. 29, 2023), <https://www.nist.gov/publications/language-trustworthy-ai-depth-glossary-terms> を参照のこと。
- 19 市民権、識別的差別禁止、公正競争、消費者保護、その他の極めて重要な法的保護といった既存の法的権限は、他の慣行に適用されるのと同様に、AI モデルやシステムの設計と使用にも適用される。DOJ and DHS et al., *Joint Statement on Enforcement of Civil Rights, Fair Competition, Consumer Protection, and Equal Opportunity Laws in Automated Systems* (Apr. 4, 2024)を参照。

- 20 OECD, *Guidelines on Protection of Privacy and Trans-border Flows of Personal Data* (Sept 9, 2013), <https://doi.org/10.1787/9789264196391-ja> を参照。
- 21 NTIA, *Dual-Use Foundation Models with Widely Available Model Weights Report* (July 30, 2024), <https://www.ntia.gov/issues/artificial-intelligence/open-model-weights-report>; CISA, *With Open Source Artificial Intelligence, Don't Forget the Lessons of Open Source Software* (July 29, 2024), <https://www.cisa.gov/news-events/news/open-source-artificial-intelligence-dont-forget-lessons-open-source-software>.
- 22 CISA, *Software Must Be Secure by Design and Artificial Intelligence Is No Exception* (Aug 18, 2023), <https://www.cisa.gov/news-events/news/software-must-be-secure-design-and-artificial-intelligence-no-exception>; NIST, *Guidelines for Evaluating and Red-Teaming Generative AI Models and Systems and Dual Use Foundation Models* (coming) を参照。
- 23 Helen Toner & Timothy Fist, *Regulating the AI Frontier : Design Choices and Constraints*, Georgetown University Center for Security and Emerging Technology (Oct 26, 2023), <https://cset.georgetown.edu/article/regulating-the-ai-frontier-design-choices-and-constraints/>.
- 24 「レッドチーム」とは、「AI システムの欠陥や脆弱性を見つけるための構造化されたテスト活動で、多くの場合、管理された環境で、AI の開発者と協力して行われるもの」を意味する。ホワイトハウス、大統領令第 14110 号、前掲注 3、第 3 条(d)。
- 25 L. Koessler, *Risk Thresholds for Frontier AI*, arXiv (2024), <https://arxiv.org/pdf/2406.14713> を参照。
- 26 NIST, *Managing Misuse Risk for Dual-Use Foundation Models, Initial Public Draft, supra* note 14 を参照のこと。
- 27 DHS, *Report on Reducing Risks at Intersection of AI and Chemical, Biological, Radiological, and Nuclear Threats* (June 2024), https://www.dhs.gov/sites/default/files/2024-04/24_0429_cwmd-dhs-fact-sheet-ai-cbrn を参照。
- 28 連邦政府の政策では、AI 開発者や計算インフラプロバイダーは、IT セクターにおける重要インフラ事業者とみなされている。本枠組みでは、AI の安全・安心の問題について、これらの事業者の責任を取り上げるために、AI 開発者や計算インフラプロバイダーを重要インフラ事業者とは区別しており、AI 開発者の責任については IV.C.II 節で、重要インフラの責任については IV.C.III 節でそれぞれ取り上げている。
- 29 ISO & IEC, *ISO 27001 Information Security, Cybersecurity, and Privacy Protection* (2022), <https://www.iso.org/> を参照。
- 30 CISA & NSA, *Joint Guidance on Deploying AI Systems Securely* (Apr 2024), <https://media.defense.gov/2024/Apr/15/2003439257/-1/-1/0/CSI-DEPLOYING-AI-SYSTEMS-SECURELY.PDF> を参照。
- 31 CISA, *Secure by Demand Guide* を参照 : *How Software Customers Can Drive a Secure Technology Ecosystem* (Aug. 2024)、<https://www.cisa.gov/resources-tools/resources/secure-demand-guide>.

- 32 強固な展開環境の確保については、CISA & NSA, 前掲注 30 を参照のこと。
- 33 CISA, *Secure by Design : Shifting Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software* (Oct 25, 2023), <https://www.cisa.gov/resources-tools/resources/secure-by-design>; CISA, *Cross-Sector Cybersecurity Performance Goals* (Mar. 2023), <https://www.cisa.gov/cross-sector-cybersecurity-performancegoals>; CISA & NSA, *supra* note 30.
- 34 NIST, *Artificial Intelligence Risk Management Framework Playbook, Measure 1.2* (2022), <https://airc.nist.gov/> を参照。
- 35 OECD, *Advancing Accountability in AI* (Feb 2023), https://www.oecd.org/en/publications/2023/02/advancing-accountability-in-ai_753bf8c8.html; NIST, *The NIST Definition of Cloud Computing, SP 800-145* (Sept 2011), <https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf> を参照。
- 36 行政管理予算局(OMB)は、連邦政府機関に対し、AI の取得に関するリスクを適切に管理するためのガイダンスを提供している。OMB, M-24-10, 前掲注 2 参照。