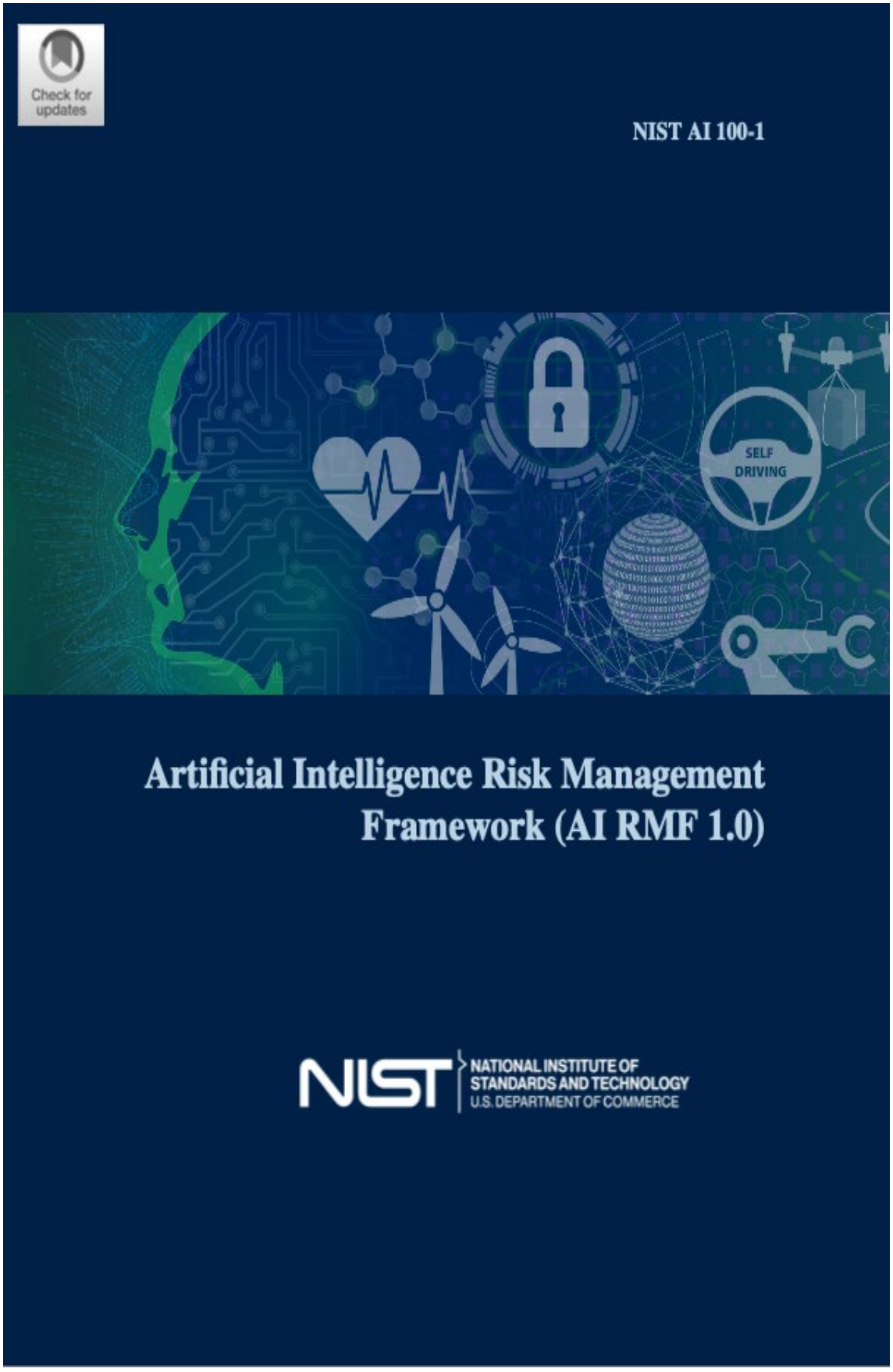
Check for updates
NIST AI 100-1
SELF DRIVING
Artificial Intelligence Risk Management Framework (AI RMF 1.0)
NIST
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Official Government Translation. Document translated by TaikaTranslation LLC.

The official English language version of this publication is available free of charge from NIST at https://doi.org/10.6028/NIST.AI.100-1.

NIST AI 100-1

NIST AI 100-1AI リスクマネジメントフレームワーク（AI RMF 1.0）

本書は https://doi.org/10.6028/NIST.AI.100-1 より無料で入手可能です。

2023 年 1 月

米国商務省
ジーナ・M・ライモンド、長官

米国国立標準技術研究所
ローリー・E・ロカシオ、NIST 所長兼標準技術担当商務次官

実験手順や概念を適切に説明するために、この文書では特定の商業団体、装置、または材料が特定される場合があります。このような識別は、米国国立標準技術研究所による推奨または承認を意味するものではなく、また、その実体、材料、または機器が必ずしもその目的に利用可能な最良のものであることを意図したものでもありません。

この出版物は、https://doi.org/10.6028/NIST.AI.100-1 から無料で入手できます。

更新スケジュールとバージョン

AI リスクマネジメントフレームワーク（AI RMF）はリビングドキュメントであることを意図しています。

NIST はフレームワークの内容と有用性を定期的にレビューし、更新が適切かどうかを判断します。AI コミュニティからの正式なインプットによるレビューは、遅くとも 2028 年までに行われる予定です。フレームワークは、主要な変更と軽微な変更を追跡し識別するために、2 つの番号によるバージョン管理システムを採用します。最初の番号は、AI RMF とその付属文書の世代を表し（例えば 1.0）、主要な改訂の場合のみ変更されます。軽微な修正は、世代番号の後に「.n」を使用して追跡されます（例：1.1）。すべての変更は、バージョン番号、変更日、変更の説明を含む履歴を特定するバージョン管理表を使用して追跡されます。NIST は AI RMF Playbook を頻繁に更新する予定です。AI RMF プレイブックに対するコメントは、いつでも電子メールで AIframework@nist.gov に送信でき、半年ごとにレビューされ、統合されます。

目次

表のリスト

図のリスト

エグゼクティブ・サマリー

人工知能（AI）テクノロジーは、商業、健康、交通、サイバーセキュリティ、環境、地球に至るまで、社会と人々の生活を変革する大きな可能性を秘めています。AI テクノロジーは包括的な経済成長を促進し、世界の状況を改善する科学の進歩を支えることができます。しかし、AI テクノロジーは、個人、グループ、組織、コミュニティ、社会、環境、地球にネガティブインパクトを及ぼすリスクももたらします。他の種類のテクノロジーのリスクと同様に、AI のリスクは様々な形で現れる可能性があり、長期的か短期的か、高確率か低確率か、系統的か局所的か、高インパクトか低インパクトか、といったものとして特徴付けることができます。

> AI RMF は、AI システムを、特定の一連の目的に対して、現実または仮想環境に影響を与える予測、推奨、意思決定などの出力を生成できる、人工的または機械ベースのシステムを指します。 AI システムは、さまざまなレベルの自律性で動作するように設計されています（引用：OECD Recommendation on AI:2019; ISO/IEC 22989:2022)。

組織が従来のソフトウェアや情報ベースのシステムのリスクを軽減するための基準やベストプラクティスは無数にありますが、AI システムがもたらすリスクは多くの点で独特です（付録 B 参照)。例えば、AI システムは、時間の経過とともに、時には予想外に大きく変化する可能性のあるデータに基づいて学習される可能性があり、理解するのが困難な形でシステムの機能やトラストワージネスにインパクトを与えます。AI システムとそれが導入されるコンテクストは複雑であることが多く、障害が発生したときにそれを検知して対応することが困難になります。AI システムは本質的に社会技術的なものであり、社会の力学や人間の行動のインパクトを受けます。AI のリスクとベネフィットは、システムの使用方法、他の AI システムとの相互作用、運用者、そしてシステムが導入される社会的背景に関連する社会的要因と組み合わされた技術的側面の相互作用から生まれる可能性があります。

このようなリスクにより、AI は、組織にとっても社会にとっても、導入して利用するのが極めて困難な技術となっています。適切な制御がなければ、AI システムは個人やコミュニティにとって不公平で望ましくない結果を増幅、永続化、悪化させる可能性があります。適切な制御を行えば、AI システムは不公平な結果を緩和し、マネジメントすることができます。

AI のリスクマネジメントは、AI システムの責任ある開発と使用の重要な要素です。責任ある AI の実践は、AI システムの設計、開発、使用に関する決定を、意図された目的と価値観に合致させるのに役立ちます。責任ある AI のコアコンセプトは、人間中心主義、社会的責任、持続可能性を強調しています。AI のリスクマネジメントは、AI の設計、開発、デプロイを行う組織とその内部チームに、コンテクストと潜在的または予期せぬネガティブインパクトとポジティブインパクトについてより批判的に考えるよう促すことで、責任ある使用と実践を推進することができる。AI システムのリスクを理解しマネジメントすることは、トラストワージネスを高め、ひいては社会からの信頼を培うことにつながります。

社会的責任とは、「透明かつ倫理的な行動を通じて、組織の意思決定や活動が社会や環境に与える影響に対する」組織が担う責任を指すことがあります（ISO26000:2010）。持続可能性とは、「将来の世代が自らのニーズを満たす能力を損なうことなく、現在のニーズが満たされる、環境、社会、経済の側面を含む地球システムの状態」を指します（ISO/IEC TR 24368:2022）。責任ある AI は、公平でアカウンタビリティを果たすテクノロジーを生み出すことを意味する。組織的な実践が「専門家としての責任」に従って行われることが期待されており、ISO では「AI システムおよびアプリケーション、または AI ベースの製品もしくはシステムを設計、開発、またはデプロイする専門家が、人々、社会および AI の未来に影響を及ぼす独自の立場を認識することを確実にすることを目指す」アプローチと定義しています（ISO/IEC TR 24368:2022）。

2020 年国家 AI イニシアチブ法（P.L. 116-283）の指示に従い、AI RMF の目標は、AI システムを設計、開発、デプロイ、または使用する組織にリソースを提供し、AI の多くのリスクをマネジメントし、トラストワージで責任ある AI システムの開発と使用を促進することです。このフレームワークは、***自発的***で、権利保護し、セクター固有ではなく、ユースケースにとらわれないことを意図しており、あらゆる規模の、あらゆるセクターの、そして社会全体の組織に、フレームワークのアプローチを実施する柔軟性を提供します。

本フレームワークは、組織や個人（ここでは AI アクターと呼ぶ）が、AI システムのトラストワージネスを高めるアプローチを提供し、長期にわたって AI システムの責任ある設計、開発、デプロイ、および使用を促進することを支援するために設計されています。AI アクターは、経済協力開発機構（OECD）によって「AI システムのライフサイクルで積極的な役割を果たす者で、AI をデプロイまたは運用する組織や個人を含むものである」と定義されています[OECD（2019）Artificial Intelligence in Society-OECD iLibrary]（付録 A 参照）。

AI RMF は、実用的であり、AI テクノロジーが発展し続ける中で AI の状況に適応し、さまざまな程度や能力の組織によって運用されて、社会が AI からベネフィットを受けながら潜在的な危害から保護されることを目的としています。

本フレームワークと支援リソースは、進化する技術、世界の標準化状況、AI コミュニティの経験とフィードバックに基づいて更新、拡張、改善される予定です。NIST は、AI RMF と関連ガイダンスを、適用可能な国際標準、ガイドライン、慣行と整合させていきます。AI RMF が使用されるにつれて、さらなる教訓が得られ、将来の更新や追加リソースに反映されるでしょう。

本フレームワークは 2 つのパートに分かれています。パート 1 では、組織が AI に関連するリスクをどのようにフレームワーク化できるかを議論し、想定される読者について説明します。次に、AI のリスクとトラストワージネスについて分析し、トラストワージな AI システムの特徴として、妥当性と信頼性、安全性、セキュリティとレジリエンス、アカウンタビリティと透明性、説明可能性と解釈可能性、プライバシーの強化、有害なバイアスを管理した公正性などを概説します。

パート 2 は、フレームワークの「コア」を構成します。組織が AI システムのリスクに実際に対処するための 4 つの具体的な機能が記述されています。これらの機能：**GOVERN**、**MAP**、**MEASURE**、**MANAGE** は、さらにカテゴリーとサブカテゴリーに分類されます。**GOVERN** は組織の AI リスクマネジメントプロセスや手順の全段階に適用される一方、**MAP**、**MEASURE**、**MANAGE** の各機能は、AI システム固有のコンテクストや AI ライフサイクルの特定の段階で適用することができます。

フレームワークに関するその他のリソースは、AI RMF Playbook に含まれており、NIST の AI RMF ウェブサイト（https://www.nist.gov/itl/ai-risk-management-framework）から入手可能です。

NIST が民間および公的セクターと協力して AI RMF を開発することは、2020 年国家 AI イニシアチブ法、人工知能に関する国家安全保障委員会の勧告および技術標準と関連ツールの開発における連邦政府の関与のための計画によって要請された、より広範な AI の取り組みと方向性が一致しています。正式な情報提供要請への回答、広く参加された 3 回のワークショップ、コンセプトペーパーとフレームワークの 2 つの草案に対するパブリックコメント、複数のパブリックフォーラムでの議論、多くの小規模グループミーティングなどを通じて、このフレームワークの策定中に AI コミュニティとの関わりを持つことで、AI RMF 1.0 の策定だけでなく、NIST やその他の機関が実施する AI の研究開発および評価にも情報を提供しました。このフレームワークを強化するための優先研究や追加ガイダンスは、NIST やより広範なコミュニティが貢献できる AI リスクマネジメントフレームワークロードマップとして取り込まれる予定です。

パート 1：基礎情報

1. リスクの枠組み

AI リスクマネジメントは、市民の自由や権利に対する脅威など、AI システムの潜在的なネガティブなインパクトを最小化する道筋を示すと同時に、ポジティブなインパクトを最大化する機会も提供します。AI のリスクと潜在的なネガティブインパクトを効果的に取り上げ、文書化し、マネジメントすることで、よりトラストワージな AI システムを実現することができます。

1.1 リスク、インパクト、危害の理解と対処

AI RMF のコンテクストでは、リスクとは、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を指します。AI システムのインパクト（結果）は、ポジティブかネガティブ、またはその両方であり、機会または脅威をもたらす可能性があります（出典：ISO 31000:2018）。潜在的な事象のネガティブなインパクトを考慮する場合、リスクは、1）その状況や事象が発生した場合に生じるネガティブなインパクト、すなわち危害の大きさと、2）発生の可能性の関数です（出典：OMB Circular A-130:2016）。ネガティブなインパクトや危害は、個人、グループ、コミュニティ、組織、社会、環境、地球に及ぶ可能性があります。

> 「リスクマネジメントとは、リスクについて、組織を指揮統制するための調整された活動を指します」（出典：ISO 31000:2018）。

リスクマネジメントプロセスは一般的にネガティブなインパクトを扱いますが、本フレームワークは AI システムの予測されるネガティブなインパクトを最小化し、ポジティブなインパクトを最大化する機会を特定するアプローチを提供します。潜在的な危害のリスクを効果的にマネジメントすることで、よりトラストワージネスの高い AI システムを実現し、人（個人、コミュニティ、社会）、組織、システム／エコシステムに潜在的なベネフィットをもたらす可能性があります。リスクマネジメントは、AI の開発者や利用者がインパクトを理解し、モデルやシステムに内在する限界や不確実性を考慮することを可能にし、ひいてはシステム全体の性能やトラストワージネスを向上させ、AI テクノロジーが有益な形で使用される可能性を高めることができます。

AI RMF は、新たなリスクが出現した場合にも対応できるように設計されています。この柔軟性は、インパクトが容易に予測できず、アプリケーションが進化している場合に特に重要です。AI のリスクやベネフィットはよく知られていますが、ネガティブなインパクトや危害の程度を評価するのは難しい場合があります。図 1 は、AI システムに関連し得る潜在的な危害の例を示しています。

AI リスクマネジメントの取り組みでは、人間は AI システムがあらゆる場面で機能し、うまく機能すると思い込んでいる可能性があることを考慮する必要があります。例えば、正しいかどうかは別として、AI システムは人間よりも客観的であるとか、一般的なソフトウェアよりも優れた能力を提供すると認識されることがよくあります。

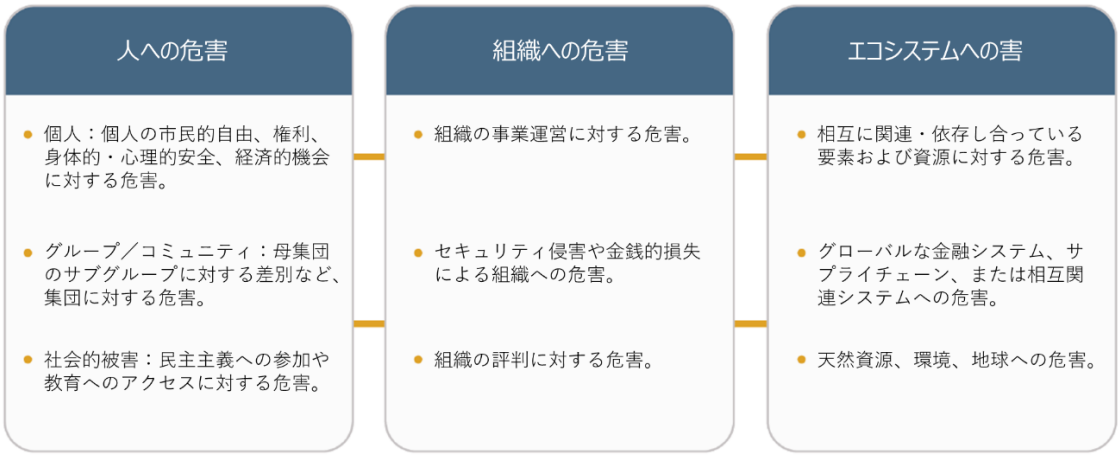

図 1. AI システムに関連する潜在的な弊害の例。トラストワージな AI システムとその責任ある使用は、ネガティブなリスクを軽減し、人、組織、エコシステムにベネフィットをもたらすことに貢献することができます。

1.2 AI リスクマネジメントの課題

いくつかの課題を以下に説明します。AI のトラストワージネスを追求してリスクをマネジメントする際には、これらを考慮する必要があります。

1.2.1 リスクの測定

不十分な定義や、十分に理解されていない AI のリスクや障害は、定量的・定性的に測定することが困難です。AI のリスクを適切に測定できないからといって、AI システムのリスクが必ずしも高い、あるいは低いことを示すわけではありません。リスク測定の課題には次のようなものがあります。

サードパーティのソフトウェア、ハードウェア、データに関するリスク：サードパーティのデータやシステムは、研究開発を加速し、技術移行を促進することができる。またリスク測定を複雑する可能性もあります。リスクは、サードパーティのデータ、ソフトウェア、ハードウェアそれ自体と、その使用方法の両方から出現する可能性があります。AI システムを開発する組織が使用するリスク測定基準や方法論は、システムをデプロイまたは運用する組織が使用するリスク測定基準や方法論と一致しない可能性があります。また、AI システムを開発する組織が、使用したリスク測定基準や方法論について透明性を持たないこともあります。リスクの測定とマネジメントは、顧客がサードパーティのデータやシステムを AI 製品やサービスにどのように使用したり統合したりするかによって、特に十分な内部ガバナンス構造や技術的な保護措置がない場合には、複雑になる可能性があります。いずれにせよ、すべての当事者と AI アクターは、開発、デプロイ、またはスタンドアロンまたは統合コンポーネントとして使用する AI システムのリスクをマネジメントする必要があります。

顕在化するリスクの追跡：新たなリスクを特定・追跡し、それを測定する手法を検討することで、組織のリスクマネジメントの取り組みは強化されます。

AI システムのインパクトアセスメントのアプローチは、AI アクターが特定のコンテクストにおける潜在的なインパクトや危害を理解するのに役立ちます。

信頼できるメトリクスの可用性：現在、リスクとトラストワージネスに関する堅牢で検証可能な測定方法、および様々な AI のユースケースへの適用に関するコンセンサスが現在のところ得られていないことが、AI のリスク測定の課題です。ネガティブなリスクや危害を測定しようとする場合、潜在的な落とし穴として、測定基準の開発はしばしば組織的な取り組みであり、根本的なインパクトとは無関係な要因を不注意に反映してしまう可能性があるという現実があります。さらに、測定アプローチが単純化されすぎたり、ゲーム化されたり、重要なニュアンスを欠いたり、予期せぬ形で信頼されるようになったり、インパクトを受ける集団やコンテクストの違いを説明できなかったりすることもあります。

集団へのインパクトを測定するためのアプローチは、コンテクストが重要であること、危害が様々な集団や小集団に異なる影響を及ぼす可能性があること、危害を受ける可能性のある地域社会やその他の小集団が必ずしもシステムの直接的な利用者ではないことを認識したうえで行うことが、最も効果的に機能します。

AI のライフサイクルの様々な段階におけるリスク：AI のライフサイクルの初期段階におけるリスクの測定は、後期段階におけるリスクの測定とは異なる結果をもたらす場合があります。ある時点では潜在的なリスクであっても、AI システムが適応し進化するにつれて増大するリスクもあります。さらに、AI のライフサイクル全体を通じて、AI アクターが異なるリスク観を持っている場合があります。例えば、事前に学習されたモデルなど、AI ソフトウェアを利用可能にする AI 開発者は、その事前に学習されたモデルを特定のユースケースにデプロイする責任を負う AI アクターとは異なるリスク視点を持つ可能性があります。そのようなデプロイヤは、その特定の用途が、最初の開発者が認識したリスクとは異なるリスクを伴う可能性があることを認識していない場合があります。すべての AI アクターは、目的に適合したトラストワージな AI システムを設計、開発、デプロイする責任を共有します。

現実世界におけるリスク：実験室やマネジメントされた環境で AI のリスクを測定すれば、デプロイ前に重要な知見が得られるかもしれませんが、このような測定は、運用中の現実世界の設定で現れるリスクとは異なる場合があります。

不可解さ：不可解な AI システムはリスク測定を複雑にする可能性があります。不可解さとは、AI システムの不透明な性質（説明可能性や解釈可能性の制限）、AI システムの開発やデプロイにおける透明性や文書化の欠如、あるいは AI システムに内在する不確実性の結果である場合があります。

人間のベースライン：例えば意思決定など、人間の活動を補強または代替することを意図した AI システムのリスクマネジメントには、比較のための何らかの基準値が必要です。AI システムは人間とは異なるタスクを実行し、異なるタスクを実行するため、これを体系化することは困難です。

1.2.2 リスク許容度

AI RMF はリスクの優先順位付けに使用できますが、リスク許容度を規定するものではありません。リスク許容度とは、組織または AI アクター（付録 A 参照）が、その目的を達成するためにリスクを負担する用意があることを指します。リスク許容度は、法律または規制の要求事項によって、影響を受ける可能性があります（出典：ISO GUIDE 73）。リスク許容度および組織や社会が許容できるリスクレベルは、極めてコンテクストに依存し、アプリケーションやユースケースによって異なります。リスク許容度は、AI システムの所有者、組織、業界、地域社会、または政策立案者によって確立された方針や規範によって影響を受ける可能性があります。リスク許容度は、AI システム、政策、規範が進化するにつれて変化する可能性が高い。異なる組織は、その組織特有の優先事項やリソースの考慮事項により、様々なリスク許容度を持つ場合があります。

有害性／コストのベネフィットのトレードオフをよりよく知らせるための新たな知識や手法は、企業、政府、学界、市民社会によって開発され、議論され続けるでしょう。AI のリスク許容度を特定するための課題が未解決のままである限り、リスクマネジメントの枠組みが AI のネガティブなリスクを軽減するためにまだ容易に適用できないコンテクストがあるかもしれません。

本フレームワークは、適用される法律、規制、規範に沿うべき既存のリスクへのプラクティスを柔軟に補強することを意図しています。組織は、組織的、ドメイン的、分野的、部門的、または専門的要件によって確立されたリスク基準、許容範囲および対応に関する既存の規制およびガイドラインに従うべきです。分野や業界によっては、危害の定義が確立し、文書化、報告、開示の要件が確立されている場合があります。セクターでは、リスクマネジメントは特定の用途やユースケースの設定に関する既存のガイドラインに依存するかもしれません。確立されたガイドラインが存在しない場合、組織は合理的なリスク許容度を定義する必要があります。許容範囲が定義されれば、この AI RMF をリスクマネジメントおよびリスクマネジメントプロセスの文書化に使用することができます。

1.2.3 リスクの優先順位付け

すべてのインシデントや障害を排除できるわけではないので、ネガティブなリスクを完全に排除しようとすることは、実際には逆効果になる可能性があります。リスクに関する非現実的な期待は、リスクのトリアージを非効率的または非現実的なものにしたり、希少なリソースを浪費するような方法でリソースを配分することに組織を導くかもしれません。リスクマネジメントの文化は、すべての AI リスクが同じではないことを組織が認識し、リソースを意図的に配分するのに役立ちます。実行可能なリスクマネジメントの取り組みは、組織が開発またはデプロイする各 AI システムのトラストワージネスを評価するための明確なガイドラインを定めています。AI システムの評価されたリスクレベルと潜在的なインパクトに基づいて、方針とリソースの優先順位を付ける必要があります。AI システムが、AI デプロイヤによる特定の使用状況に合わせてどの程度カスタマイズされるか、または調整されるかは、要因の一つとなり得ます。

AI RMF を適用する場合、特定の利用コンテクストにおける AI システムにとって最もリスクが高いと組織が判断したものは、最も緊急に優先順位を付け、最も徹底したリスクマネジメントプロセスを必要とします。例えば、重大なネガティブインパクトが差し迫っている、深刻な被害が実際に発生している、壊滅的なリスクが存在するなど、AI システムが許容できないネガティブのリスクレベルを示す場合には、リスクを十分にマネジメントできるようになるまで、安全な方法で開発とデプロイを停止する必要があります。AI システムの開発、デプロイ、使用事例が特定のコンテクストにおいて低リスクであることが判明した場合、優先順位を下げる可能性があることを示唆する可能性があります。

リスクの優先順位付けは、人間と直接相互作用するように設計またはデプロイされた AI システムと、そうでない AI システムとでは異なる場合があります。AI システムが、個人を特定できる情報のような機密または保護されたデータで構成される大規模なデータセットで学習される場合や、AI システムの出力が人間に直接的または間接的なインパクトを与える場合などには、より高い初期優先順位付けが求められる可能性があります。計算機システムとの相互作用のみを目的として設計され、非機密データセット（例えば、物理的環境から収集されたデータ）を対象として学習された AI システムは、初期の優先順位付けをより低くする必要があるかもしれません。とはいえ、人と接しない AI システムが下流の安全性や社会的影響を及ぼす可能性があるため、コンテクストに基づいて定期的にリスクを評価し、優先順位をつけることは依然として重要です。

残留リスクは、リスク対応後に残るリスクと定義され（出典：ISO GUIDE 73）、エンドユーザや影響を受ける個人およびコミュニティに直接インパクトを与えます。残留リスクを文書化することで、システム提供者は AI 製品をデプロイするリスクを十分に検討することが求められ、システムとの相互作用によるネガティブインパクトをエンドユーザに通知することができます。

1.2.4 組織統合とリスクマネジメント

AI のリスクは、単独で考慮されるべきではありません。AI アクターは、ライフサイクルにおけるそれぞれの役割に応じて、異なる責任と認識を持っています。例えば、AI システムを開発する組織は、そのシステムがどのように使用されるかについての情報を持たないことがよくあります。AI のリスクマネジメントは、より広範な企業のリスクマネジメント戦略とプロセスに統合され、組み込まれるべきである。AI リスクをサイバーセキュリティやプライバシーなどの他の重要なリスクと一緒に扱うことで、より統合された結果と組織の効率性をもたらします。

AI RMF は、AI システムリスクやより広範な組織リスクをマネジメントするための関連ガイダンスやフレームワークとともに使用することができます。AI システムに関連するいくつかのリスクは、他の種類のソフトウェア開発やデプロイに共通しています。重複するリスクの例としては、AI システムを学習するための基礎データの使用に関連するプライバシーの懸念、リソースを多用するコンピューティングに関連するエネルギーと環境へインパクト、システムおよびその学習データと出力データの機密性、完全性、可用性に関連するセキュリティの懸念、AI システムの基礎となるソフトウェアとハードウェアの一般的なセキュリティへの懸念などがあります。

組織は、リスクマネジメントを効果的に行うために、適切なアカウンタビリティ・メカニズム、役割と責任、文化、インセンティブ構造を確立し、維持する必要があります。AI RMF を使用するだけでは、このような変化をもたらすことも、適切なインセンティブの提供にはつながりません。効果的なリスクマネジメントは、上位レベルの組織的コミットメントを通じて実現されるものであり、組織や業界内の文化的変化を必要とする場合もあります。加えて、AI リスクをマネジメントする、あるいは AI RMF を実施する中小規模の組織は、その能力やリソースに応じて、大規模組織とは異なる課題に直面する可能性があります。

2. オーディエンス

AI リスクと潜在的インパクト（ポジティブなものもネガティブなものも）を特定しマネジメントするには、AI のライフサイクル全体にわたる幅広い視点とアクターが必要です。理想的には、AI のアクターは、多様な経験、専門知識、背景を持ち、人口統計学的にも専門分野的にも多様なチームが構成されます。AI RMF は、AI のライフサイクルと次元を超えた AI アクターによって使用されることを目的としています。

OECD は、AI のライフサイクル活動を 5 つの主要な社会技術的次元に従って分類するためのフレームワークを開発し、それぞれがリスクマネジメントを含む AI 政策とガバナンスに関連する特性を有しています［OECD (2022) OECD Framework for the Classification of AI systems - OECD Digital Economy Papers］。図 2 は、これらの次元を本フレームワークの目的のために NIST によって若干修正したものを示しています。NIST の修正は、AI のライフサイクル全体を通じてテスト、評価、妥当性確認および検証（Test, Evaluation, Validation and Verification: TEVV）プロセスの重要性を強調し、AI システムの運用コンテクストを一般化しています。

図 2 に示す AI の次元は、アプリケーション・コンテクスト、データと入力、AI モデル、タスクと出力です。これらの次元に関与し、AI システムの設計、開発、デプロイ、評価、使用を実行またはマネジメントし、AI リスクマネジメントの取り組みを推進する AI アクターは、AI RMF の主要な対象者です。

ライフサイクルの各次元における代表的な AI アクターを図 3 に示し、付録 A で詳述しています。AI RMF では、すべての AI アクターが協力してリスクをマネジメントし、トラストワージで責任ある AI の目標を達成します。TEVV に特化した専門知識を持つ AI アクターは、AI のライフサイクル全体を通じて統合され、特にフレームワークから恩恵を受ける可能性が高くなります。定期的に実施される TEVV タスクは、技術的、社会的、法的、倫理的な基準や規範に関連する洞察を提供し、インパクトの予測や顕在化するリスクの評価と追跡を支援するのに役立ちます。AI のライフサイクルにおける定期的なプロセスとして、TEVV は途中段階での是正と事後的なリスクマネジメントの両方を可能にします。

図 2 の中心にある「人と地球」の次元は、人権と、社会と地球のより広範なウェルビーイングを表しています。この次元の AI アクターは、主要なオーディエンスに情報を提供する別の AI RMF

オーディエンスを構成します。これらの AI アクターには、業界団体、標準化団体、研究者、アドボカシ団体などが含められます。

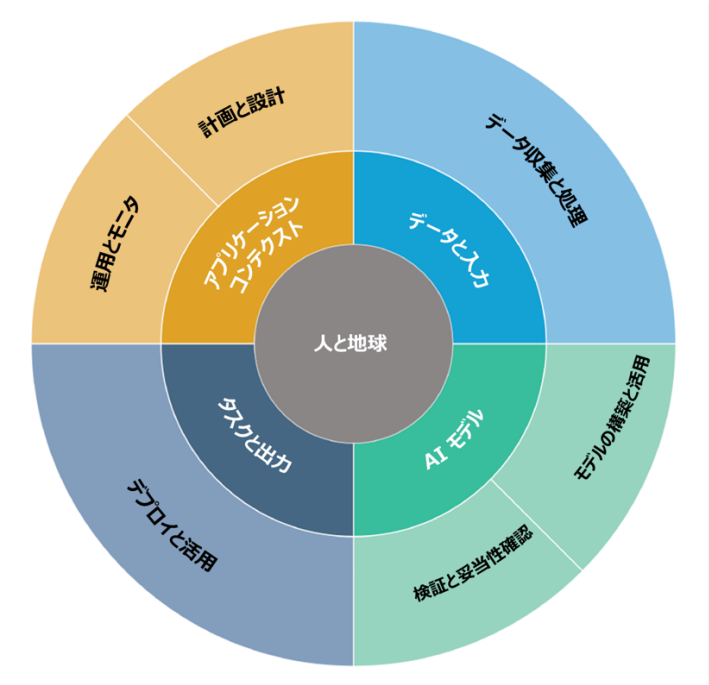

図 2. AI システムのライフサイクルと主要な次元 OECD。 (2022) OECD Framework for the Classification of AI systems - OECD Digital Economy Papers より改変。内側の 2 つの円は AI システムの主要な次元を示し、外側の円は AI のライフサイクルの段階を示しています。理想的には、リスクマネジメントの取り組みは、アプリケーションのコンテクストにおける計画と設計の機能から始まり、AI システムのライフサイクル全体を通じて実行されます。代表的な AI アクターについては、図 3 を参照してください。

環境団体、市民社会組織、エンドユーザ、そして潜在的に影響を受ける個人やコミュニティ。 これらのアクターは次のことができます：

- 潜在的および実際のインパクトについてコンテクストを提供し、理解することを支援する。
- AI のリスクマネジメントのための公式または準公式な規範およびガイダンスの情報源となる。
- AI 運用の境界（技術的、社会的、法的、倫理的）を指定する。
- 市民の自由と権利、公平性、環境と地球、経済に関する社会の価値と優先事項のバランスをとるために必要なトレードオフの議論を促進する。

リスクマネジメントの成功は、図 3 に示す AI アクターの連帯責任感にかかっています。セクション 5 で説明する AI RMF の機能には、多様な視点、専門分野、職業、経験が必要です。多様なチームは、テクノロジーの目的や機能に関するアイデアや前提をよりオープンに共有し、これらの暗黙の側面をより明確にすることに貢献します。このような広範な集団的視点は、問題を表面化させ、既存および顕在化するリスクを特定する機会を生み出します。

主要な側面	アプリケーションコンテクスト	データと入力	AIモデル	AIモデル	タスクと出力	アプリケーションコンテクスト	人と地球
ライフサイクルステージ	計画と設計	データ収集と処理	モデルの構築と活用	検証と妥当性確認	デプロイと使用	運用とモニタ	活用またはインパクト
TEVV	TEVVには監査とインパクトアセスメントが含まれる	TEVVには内部および外部の妥当性確認が含まれる	TEVVにはモデルテストを含む	TEVVにはモデルテストを含む	TEVVには、統合、コンプライアンス・テスト、妥当性確認が含まれる	TEVVには監査とインパクトアセスメントが含まれる	TEVVには監査とインパクトアセスメントが含まれる
活動	法的・規制的要件および倫理的配慮に照らして、システムのコンセプトと目的、基礎となる仮定、および背景を明確にし、文書化する。	目的、法的、倫理的配慮に照らして、データを収集、検証、クリーニングし、データセットのメタデータと特性を文書化する。	アルゴリズムを作成または選択し、モデルを訓練する。	モデル出力の検証と妥当性確認、校正、解釈。	試験運用、レガシーシステムとの互換性の確認、規制遵守の検証、組織変更の管理、ユーザーエクスペリエンスの評価。	AIシステムを運用し、目的、法規制要件、倫理的配慮に照らして、その推奨事項と影響（意図したもの、意図しなかったものの両方）を継続的に評価する。	システム／技術を利用し、影響を監視・評価し、影響の緩和を求め、権利を擁護する。
代表的な関係者	システム運用者、エンドユーザー、領域専門家、AI設計者、影響評価者、TEVV専門家、プロダクトマネージャー、コンプライアンス専門家、監査人、ガバナンス専門家、組織管理者、C-suite幹部、影響を受ける個人/コミュニティ、評価者	データサイエンティスト、データエンジニア、データプロバイダー、ドメイン専門家、社会文化アナリスト、ヒューマンファクター専門家、TEVV専門家	モデラー、モデル・エンジニア、データ・サイエンティスト、開発者、ドメイン専門家、アプリケーションのコンテキストに精通した社会文化アナリストおよびTEVV専門家の助言を得る。		システムインテグレーター、開発者、システムエンジニア、ソフトウェアエンジニア、ドメインの専門家、調達の専門家、サードパーティサプライヤー、経営幹部、ヒューマンファクターの専門家、社会文化アナリスト、ガバナンスの専門家、TEVVの専門家の助言を得る。	システム運用者、利用者、実務者、領域専門家、AI設計者、影響評価者、TEVV専門家、システム資金提供者、製品管理者、コンプライアンス専門家、監査人、ガバナンス専門家、組織管理者、影響を受ける個人／コミュニティ、評価者	エンドユーザー、事業者、実務者、影響を受ける個人／コミュニティ、一般市民、政策立案者、標準化団体、業界団体、擁護団体、環境

図 3. AI ライフサイクルの各段階における AI アクター。テスト、評価、妥当性確認、検証（TEVV）タスクの詳細を含む AI アクターのタスクの詳細については、付録 A を参照してください。なお、AI モデルの次元（図 2）における AI アクターは、ベストプラクティスとして、モデルを構築・使用活用するアクターと、モデルを検証・妥当性確認するアクターとに分離されていることに留意してください。

3. AI のリスクとトラストワージネス

AI システムがトラストワージであるためには、多くの場合、利害関係者にとって価値のある多様な基準に対応する必要があります。AI のトラストワージを高めるアプローチは、AI のネガティブなリスクを低減することができます。本フレームワークは、信頼できる AI の特徴を以下のように明確にし、それらに対処するためのガイダンスを提供します。トラストワージな AI システムの特性には、**妥当性と信頼性、安全性、セキュリティ、レジリエンス、アカウンタビリティと透明性、説明可能性と解釈可能性、プライバシー強化、有害なバイアスを管理した公正性**が含まれます。トラストワージな AI を作るには、AI システムの使用状況に基づいて、これらの各特性のバランスを取る必要があります。すべての特性は社会技術的なシステム属性ですが、アカウンタビリティと透明性は AI システム内部のプロセスと活動およびその外部設定にも関係します。これらの特性をおろそかにすると、否定的な結果が生じる確率と規模が増大する可能性があります。

図 4. トラストワージネスな AI システムの特徴。妥当性と信頼性はトラストワージの必要条件であり、他のトラストワージの特性のベースとして示されています。アカウンタビリティと透明性は、他のすべての特性に関連するため、縦長の枠で示されています。

トラストワージの特性（図 4 に示す）は、社会的・組織的行動、AI システムで使用されるデータセット、AI モデルやアルゴリズムの選択、それらを構築する人々による意思決定、そしてそのようなシステムから洞察を提供し、モニタリングする人間との相互作用と密接に結びついています。AI のトラストワージ特性に関連する具体的な指標と、それらの指標の正確な閾値を決定する際には、人間の判断が採用されるべきです。

通常、トレードオフが関係し、すべての特性がすべての設定に適用されることは稀であり、いくつかの特性はどのような状況においてもより多かれ少なかれ重要なものもあります。結局のところ、トラストワージは社会的な概念であり、その範囲は広範であり、最も弱い特性と同じくらいの強さにしかなりません。

AI のリスクをマネジメントする際、組織はこれらの特性のバランスを取る上で難しい決断に直面することがあります。例えば、特定のシナリオでは、解釈可能性の最適化とプライバシーの達成の間でトレードオフが生じる可能性があります。他のケースでは、予測精度と解釈可能性のトレードオフに直面する場合もあります。あるいは、データの希少性など特定の条件下では、プライバシーを強化させる技法は正確性の低下を招き、特定のドメインでの公平性やその他の価値に関する意思決定にインパクトを与える可能性があります。

トレードオフに対処するには、意思決定のコンテクストを考慮する必要があります。このような分析は、様々な尺度間のトレードオフの存在と程度を明らかにすることはできますが、トレードオフをどのように乗り切るかについての質問には答えてくれません。それらは、関連するコンテクストの中で作用している価値観に依存し、透明性が高く、適切に正当化できる方法で解決される必要があります。

AI のライフサイクルにおいてコンテクストの認識を強化するには、複数のアプローチがあります。例えば、対象分野の専門家は、TEVV の調査結果の評価を支援し、製品やデプロイチームと協力して、TEVV のパラメータを要件やデプロイ条件に合わせることができます。適切なリソースを確保すれば、AI のライフサイクル全体を通じて、利害関係者や関連する AI アクターからのインプットの幅と多様性を高めることで、コンテクストに配慮した評価に情報を提供し、AI システムのベネフィットやポジティブインパクトを特定する機会を高めることができます。こうした実践により、社会的コンテクストから生じるリスクが適切にマネジメントされる可能性が高まります。

トラストワージの特性の理解と扱いは、AI のライフサイクルにおける AI アクターの特定の役割によって異なります。どのような AI システムであっても、AI の設計者や開発者は、その特性についてデプロイヤと異なる認識を持っている場合があります。

本書で説明するトラストワージの特性は互いに影響し合います。安全性は高いが不公正なシステム、正確だが不透明で解釈不可能なシステム、不正確だが安全でプライバシーが強化され透明性の高いシステムは、いずれも望ましくありません。リスクマネジメントへの包括的なアプローチでは、トラストワージの特性間のトレードオフのバランスを取ることが求められます。AI テクノロジーが与えられたコンテクストや目的にとって適切なツールなのか、あるいは必要なツールなのか、そしてどのように責任を持って使用するのかを判断することは、すべての AI アクターの共同責任です。AI システムの委託やデプロイの決定は、トラストワージの特性や相対的なリスク、影響、コスト、便益の状況に応じた評価に基づき、幅広い利害関係者から情報を得て行う必要があります。

3.1 妥当性と信頼性

妥当性確認とは、「特定の意図された用途または適用に関する要求事項が満たされていることを、客観的証拠の提示することによって確認すること」です（出典：ISO 9000:2015）。不正確であったり、信頼性が低かったり、トレーニング以上のデータや設定に対する汎化が不十分であったりする AI システムのデプロイは、否定的な AI リスクを生み出し、増大させ、信頼性を低下させます。

信頼性とは、同規格において「あるアイテムが、与えられた条件下で、与えられた時間間隔の間、故障することなく、要求通りに機能する能力」と定義されています（出典：ISO/IEC TS 5723:2022）。信頼されることは、予想される使用条件下で、システムのライフタイム全体を含む所定の期間にわたって、AI システムの動作が全体的に正しいことを示すことを目標とします。

正確性と堅牢性は AI システムの妥当性と信頼性に寄与し、AI システムでは互いに緊張関係にある場合があります。

正確性は、ISO/IEC TS 5723:2022 で「観測、計算、または推定の結果が、真の値または真であると認められた値に近いこと」と定義されています。正確性の測定は、計算中心の測定（例えば、偽陽性率や偽陰性率）、人間と AI のチーム化を考慮し、外部妥当性（学習条件を超えて汎用化可能）を実証する必要があります。正確性測定は、常に、明確に定義された現実的なテストセット（想定される使用条件を代表するもの）およびテスト方法論の詳細と組み合わせるべきです。正確性測定には、異なるデータセグメントに対する結果の分解を含めることができます。

堅牢性または汎用化可能性は、「様々な状況下でシステムがその性能レベルを維持する能力」と定義されています（出典：ISO/IEC TS 5723:2022）。堅牢性とは、当初想定していなかった AI システムの使用も含め、幅広い条件や状況において適切なシステム機能を発揮するための目標です。堅牢性とは、システムが想定された用途と全く同じように動作することだけでなく、想定外の環境で動作する場合に、人々への潜在的な危害を最小限に抑えるような方法で動作することも要求されます。

デプロイされた AI システムの妥当性と信頼性は、システムが意図したとおりに動作していることを確認する継続的なテストやモニタリングによって評価されることがよくあります。妥当性、正確性、堅牢性、信頼性の測定は信頼性に寄与し、ある種の故障がより大きな危害を引き起こす可能性があることを考慮すべきです。AI のリスクマネジメントの取り組みは、潜在的なネガティブインパクトを最小限に抑えることを優先すべきであり、AI システムがエラーを検出または修正できない場合には、人間の介入を含める必要があるかもしれません。

3.2 安全

AI システムは、「定義された条件下で、人の生命、健康、財産、または環境が危険にさらされる状態に陥ってはならない」（出典：ISO/IEC TS 5723:2022）。AI システムの安全な運用は、以下の方法を通じて改善されます：

- 責任ある設計、開発、デプロイの実践
- システムの責任ある使用に関するデプロイヤへの明確な情報
- デプロイヤとエンドユーザによる責任ある意思決定
- インシデントの経験的証拠に基づくリスクの説明と文書化

さまざまなタイプの安全に関するリスクは、その状況や潜在的なリスクの重大性に基づいて、AI リスクマネジメトのアプローチを調整する必要になる場合があります。重傷を負ったり死亡したりする潜在的なリスクをもたらす安全リスクは、最も緊急に優先順位を付け、最も徹底したリスクマネジメントプロセスが必要です。

ライフサイクルの中で安全性を考慮し、計画や設計をできるだけ早い段階から始めることで、システムを危険な状態にする可能性のある故障や状態を防ぐことができます。AI の安全性に関する他の実用的なアプローチは、厳密なシミュレーションとドメイン内テスト、リアルタイムモニタリング、意図された、あるいは期待された機能から逸脱したシステムのシャットダウン、修正、あるいは人間による介入能力に関連しています。

AI の安全リスクマネジメントアプローチは、交通やヘルスケアなどの分野における安全に関する取り組みやガイドラインからヒントを得て、既存の分野やアプリケーション固有のガイドラインや標準に整合させる必要があります。

3.3 セキュリティとレジリエンス

AI システム、およびそれがデプロイされるエコシステムは、環境や使用における予期せぬ有害事象や予期せぬ変化に耐えることができる場合、または AI システムがその機能と構造を維持できる場合、レジリエンスがあると言えます。内部および外部の変化に直面し、必要に応じて安全かつ優雅に劣化します（ISO/IEC TS 5723:2022 からの引用）。一般的なセキュリティ上の懸念は、敵対的な事例、データポイズニング、AI システムのエンドポイントを介したモデル、学習データ、その他の知的財産の流出に関連しています。不正アクセスや不正使用を防止する保護メカニズムによって機密性、完全性、可用性を維持できる AI システムはセキュアであると言えるかもしれません。NIST サイバーセキュリティ・フレームワークや NIST リスクマネジメント・フレームワークのガイドラインは、ここで適用可能なものの一つです。

セキュリティとレジリエンスは関連していますが、異なる特性です。レジリエンスとは、予期しない不利な事象が発生した後に正常な機能に回復する能力のことですが、セキュリティにはレジリエンスが含まれるだけでなく、攻撃を回避、防御、対応、回復するためのプロトコルも含まれます。レジリエンスは頑健性に関連し、データの出所にとどまらず、モデルやデータの予期せぬ使用や敵対的な使用（あるいは乱用や誤用）をも包含します。

3.4 アカウンタビリティと透明性

トラストワージな AI はアカウンタビリティにかかっています。アカウンタビリティは透明性を前提としています。透明性とは、AI システムとその出力に関する情報が、そのようなシステムを操作していることを認識しているかどうかに関係なく、そのようなシステムを操作する個人が利用できる程度を反映します。有意義な透明性は、AI のライフサイクルの段階に応じた適切なレベルの情報へのアクセスを提供し、AI アクターや AI システムと相互作用したり、AI システムを利用したりする個人の役割や知識に合わせて調整されます。より高いレベルの理解を促すことで、透明性は AI システムに対する信頼を高めます。

これらの特性の範囲は、設計上の意思決定やトレーニングデータから、モデルのトレーニング、モデルの構造、意図されたユースケース、デプロイ、デプロイ後、またはエンドユーザの意思決定が、いつ、どのように、誰によって行われたかにまで及びます。AI システムの出力が不正確であったり、ネガティブインパクトをもたらしたりした場合、対処可能な救済措置を受けるためには、透明性が必要になることがよくあります。透明性は、人間と AI との相互作用を考慮すべ

きであり、例えば、AI システムによって引き起こされた潜在的または実際の不利な結果が検出された場合、人間のオペレータまたはユーザにどのように通知されるか、などです。

透明性を有するシステムが必ずしも正確で、プライバシーが強化され、安全で、公平なシステムであるとは限りません。しかし、透明性のないシステムがそのような特性を持っているかどうかを判断することは困難であり、複雑なシステムが進化していく中で、時間をかけて判断することは困難です。

AI システムの結果に対するアカウンタビリティを求める際には、AI アクターの役割を考慮すべきです。AI や技術システムに関連するリスクとアカウンタビリティの関係は、文化的、法的、分野的、社会的コンテクストによって異なります。生命や自由が危機に瀕している場合など、結果が深刻である場合、AI の開発者やデプロイヤは、その透明性とアカウンタビリティの慣行を比例的かつ積極的に調整することを検討する必要があります。リスクマネジメントのような危害軽減のための組織的慣行や統治機構を維持することは、よりアカウンタビリティを果たすシステムにつながるのに役立ちます。

透明性とアカウンタビリティを強化するための措置は、必要なリソースのレベルや専有情報を保護する必要性など、実施主体に対するこれらの取り組みのインパクトも考慮すべきです。

トレーニングデータの出所を維持し、AI システムの判断がトレーニングデータのサブセットに帰属することを裏付けることで、透明性とアカウンタビリティの両方を支援することができます。トレーニングデータは著作権の対象となる可能性もあり、適用される知的財産権法に従う必要があります。

AI システム用の透明性ツールや関連文書が進化し続ける中、AI システムの開発者は、AI システムが意図されたとおりに使用されることを保証するために、AI デプロイヤと協力してさまざまな種類の透明性ツールをテストすることをお勧めします。

3.5 説明可能性と解釈可能性

説明可能性とは、AI システムの動作の基礎となるメカニズムの表現であり、解釈可能性とは、設計された機能目的のコンテクストにおける AI システムの出力の意味を指します。説明可能性と解釈可能性はともに、AI システムを運用またはオーバーサイトする人々や AI システムの使用者が、その出力を含むシステムの機能性とトラストワージネスについてより深い洞察を得るのに役立ちます。基本的な前提は、否定的なリスクの認識は、システムの出力を適切に理解したり、コンテクスト化したりする能力の欠如に起因するという仮定です。説明可能で解釈可能な AI システムは、エンドユーザが AI システムの目的と潜在的なインパクトを理解するのに役立つ情報を提供します。

説明可能性の欠如によるリスクは、AI システムがどのように機能するかを、ユーザの役割、知識、スキルレベルなどの個人差に合わせて説明することでマネジメントできます。説明可能なシステムは、デバッグやモニタリングがより容易になり、より徹底した文書化、監査、ガバナンスに適しています。

解釈可能性に対するリスクは、多くの場合、AI システムが特定の予測や推奨を行った理由の説明を伝えることで対処できます。(ここにある「説明可能な人工知能の 4 つの原則」および「人工知能における説明可能性と解釈可能性の心理学的基礎」を参照してください。)

透明性、説明可能性、解釈可能性は、互いに支え合う明確な特性です。透明性は、システムで「何が起こったか」という疑問に答えることができます。説明可能性は、システムで「どのように」決定がなされたかという疑問に答えることができます。解釈可能性は、システムによって「なぜ」決定がなされたのか、そしてその意味やコンテクストがユーザにとってどうなのかという問いに答えることができます。

3.6 プライバシー強化

*プライバシー*強化とは、一般に、人間の自律性、アイデンティティ、尊厳を保護するのに役立つ規範と慣行を指します。これらの規範と実践は、一般に、侵入からの自由、観察の制限、または個人のアイデンティティの側面 （身体、データ、評判など）の開示またはコントロールに同意する個人の主体性に対処します。(「NIST プライバシーのフレームワーク: エンタープライズのリスクマネジメントを通じてプライバシーを改善するツール」を参照してください。)

匿名性、機密性、コントロールといったプライバシーの価値は、一般に、AI システムの設計、開発およびデプロイの選択の指針となるべきです。プライバシーに関連するリスクは、セキュリティ、バイアス、透明性に影響を与える可能性があり、これらの他の特性とのトレードオフを伴います。安全性やセキュリティと同様に、AI システムの特定の技術的特徴は、プライバシーを促進することもあれば、低下させることがあります。AI システムはまた、推論によって個人を特定したり、個人に関するこれまで非公開であった情報を特定したりすることで、プライバシーに新たなリスクをもたらす可能性があります。

AI のためのプライバシー強化技術（Privacy-Enhancing Technologies :PETs）および特定のモデル出力に対する非特定化や集約などのデータ最小化手法は、プライバシー強化 AI システムの設計をサポートすることができます。データの希少性など特定の条件下では、プライバシーを強化する技術は正確性の低下をもたらし、特定のドメインにおける公平性やその他の価値判断に影響を与える可能性があります。

3.7 公平性 - 有害なバイアスのマネジメント

AI における公平性には、有害なバイアスや差別などの問題に対処することで、平等性と公正に配慮することが含まれます。公平性さに対する認識は文化によって異なり、また用途によって変化する可能性があるため、公平さの基準は複雑で定義が難しい場合があります。このような違いを認識し、考慮することで、組織のリスクマネジメントの取り組みが強化されます。有害なバイアスが緩和されたシステムが必ずしも公平ではありません。例えば、人口統計学的なグループ間で予測値がある程度均衡しているシステムでも、障害者やデジタルデバイドの影響を受けている人々にとっては依然としてアクセスしにくいものであったり、既存の格差や体系的なバイアスを悪化させたりする可能性があります。

バイアスは、人口統計学的バランスやデータの代表性よりも幅広いものです。NIST は、考慮し、マネジメントすべき AI のバイアスを、系統的、計算および統計的、人間の認知的の 3 つの主要カテゴリーに分類しました。これらはいずれも、先入観、偏見、差別的な意図がなくても発生する可能性があります。系統的バイアスは、AI のデータセット、AI ライフサイクル全体にわたる組織の規範、慣行、プロセス、AI システムを使用する広範な社会に存在する可能性があります。計算および統計的バイアスは、AI のデータセットやアルゴリズムのプロセスに存在する可能性があり、多くの場合、代表的でないサンプルによる系統的エラーに起因します。人間の認知的バイアスは、個人や集団が AI システムの情報をどのように認識して意思決定を行うか、あるいは不足している情報を補うか、あるいは人間が AI システムの目的や機能についてどのように考えるかに関連しています。人間の認知的バイアスは、AI の設計、実装、運用、保守を含む、AI のライフサイクルとシステム使用にわたる意思決定プロセスに遍在しています。

バイアスは様々な形で存在し、私たちの生活に関する意思決定を支援する自動化されたシステムに根付いている可能性があります。バイアスは必ずしもネガティブな現象ではありませんが、AI システムはバイアスのスピードと規模を増大させ、個人、グループ、コミュニティ、組織、社会に対する危害を永続させ、増幅させる可能性があります。バイアスは、社会における透明性や公平性の概念と密接に関連しています。(3 つのカテゴリを含むバイアスの詳細については、NIST 特別出版物 1270、「人工知能におけるバイアスの識別と管理の標準に向けて」を参照してください。)

4. AI RMF の有効性

AI RMF の有効性の評価（AI システムのトラストワージネスの底上げを測定する方法を含む）は、AI コミュニティと協力して、今後の NIST の活動の一部となる予定です。

組織やその他のフレームワークの利用者は、AI RMF によって AI リスクをマネジメントする能力が向上したかどうかを、その方針、プロセス、実務、実施計画、指標、測定、期待される成果などを含めて、定期的に評価することが奨励されます。NIST は、AI RMF の有効性を評価するための指標、方法論、目標を策定し、その結果や裏付けとなる情報を広く共有するために、他組織と協力していく予定です。フレームワーク利用者は、以下のような恩恵を受けることが期待されます：

- AI リスクを GOVERN、MAP、MEASURE、および MANAGE し、結果を明確に文書化するための強化されたプロセス
- トラストワージネスの特性、社会技術的アプローチ、AI リスクの間の関係とトレードオフに関する認識の向上
- システムの試験導入およびデプロイの可否を決定するための明確なプロセス
- AI システムリスクに関連する組織のアカウンタビリティの取り組みを改善するための方針、プロセス、慣行、手順の確立
- AI システムリスクと個人、コミュニティ、組織、社会への潜在的インパクトの特定とマネジメントを優先する組織文化の強化
- リスク、意思決定プロセス、責任、陥りやすい一般的な落とし穴、テスト、評価、妥当性確認、検証（TEVV）の実践、継続的改善のためのアプローチに関する組織内および組織間でより良い横断的な情報共有の実施改善
- 下流リスクに対する認識を高めるためのコンテクスト知識の向上
- 利害関係者および関連する AI アクターとのエンゲージメントの強化
- AI システムと関連リスクの TEVV 能力の向上

パート 2：コアとプロファイル

5. AI RMF コア

AI RMF コアは、AI リスクをマネジメントし、トラストワージな AI システムを責任を持って開発するための対話、理解、活動を可能にするアウトカムと行動を提供します。図 5 に示すように、コアは **GOVERN**、**MAP**、**MEASURE**、および **MANAGE** の 4 つの機能で構成されます。これらのハイレベルな機能は、それぞれカテゴリーとサブカテゴリーに分類されています。カテゴリーとサブカテゴリーは、具体的な行動と結果に細分化されます。アクションはチェックリストを構成するものではなく、必ずしも順序付けられたステップの集合でもありません。

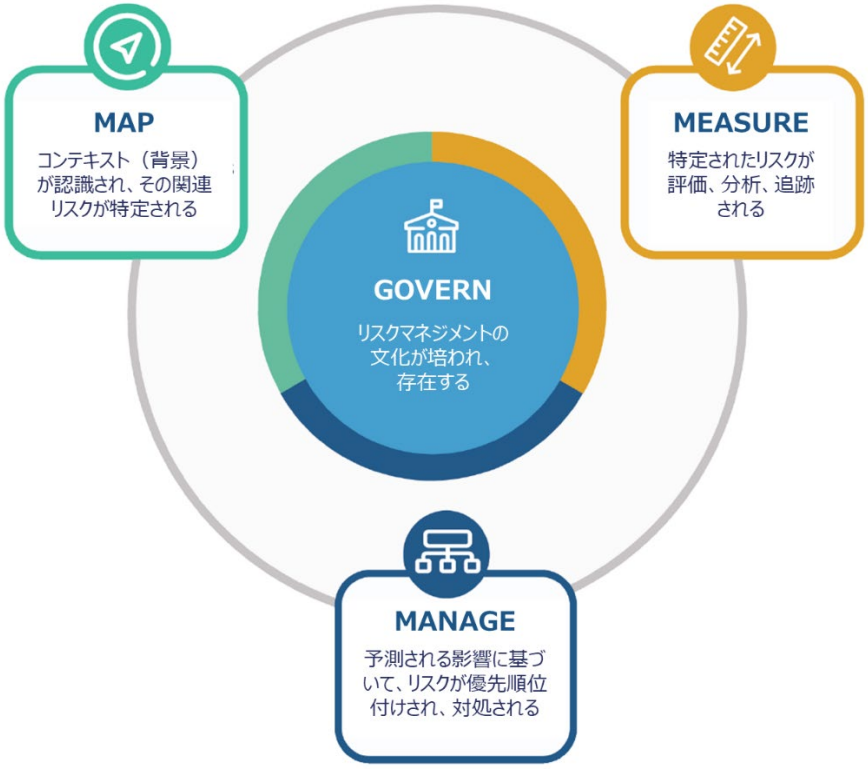

図 5. 機能は、AI リスクマネジメント活動を最高レベルで組織化し、AI リスクを GOVERN、MAP、MEASURE、および MANAGE します。GOVERN は、他の 3 つの機能に情報を提供し、浸透させるための横断的な機能として設計されています。

リスクマネジメントは、AI システムのライフサイクルの全次元にわたり、継続的かつ適時に実施されるべきです。AI RMF のコア機能は、多様かつ学際的な視点を反映した形で実施されるべきであり、組織外の AI アクターの意見も含まれる可能性があります。多様なチームを持つことは、設計、開発、デプロイ、評価される技術の目的や機能に関するアイデアや前提をよりオープンに共有することに貢献し、このことは、問題を表面化させ、既存および顕在化するリスクを特定する機会を創出することができます。

AI RMF のオンライン付属リソースである NIST AI RMF プレイブックは、組織が AI RMF をナビゲートし、それぞれのコンテクストで適用可能な戦術的アクションの提案を通じてそのアウトカムを達成するのに役立ちます。AI RMF と同様、プレイブックは任意であり、組織はそれぞれのニーズや関心に応じて提案を利用することができます。プレイブックの利用者は、提案された資料の中から自分たちで使用するために選んだオーダーメイドのガイダンスを作成したり、より広いコミュニティと共有するために提案を投稿したりすることができます。AI RMF とともに、プレイブックは NIST の「信頼できる責任ある AI リソースセンター」の一部です。

フレームワークの使用者は、自社のリソースと能力に基づいて、AI リスクマネジメントのニーズに最も適したこれらの機能を適用することができます。カテゴリーとサブカテゴリーの中から選択する組織もあれば、すべてのカテゴリーとサブカテゴリーを選択し、適用できる能力を有する組織もあります。GOVERN 体制が整っていることを前提に、フレームワークの利用者が価値を高めると判断した機能は、AI のライフサイクルを通じてどのような順序で実行してもいいでしょう。**GOVERN** でアウトカムを確立した後、AI RMF のほとんどのユーザは **MAP** 機能から開始し、**MEASURE** または **GOVERN** へと続けます。ユーザがどのように機能を統合するにしても、そのプロセスは反復的であるべきであり、必要に応じて機能間の相互参照を行うべきです。同様に、複数の機能に適用される要素を持つカテゴリーやサブカテゴリー、あるいは論理的に特定のサブカテゴリーの決定の前に行われるべきものもあります。

5.1 GOVERN

GOVERN は以下のように機能します：

- AI システムを設計、開発、デプロイ、評価、または取得する組織において、リスクマネジメントの文化を醸成し実施する。
- システムのリスクを予測、特定、マネジメントするためのプロセス、文書、組織スキーム（ユーザや社会全体に及ぶ他者を含む）およびそれらの成果を達成するための手順を概説する。潜在的なインパクトを評価するためのプロセスを組み込んでいる。
- AI のリスクマネジメント機能を組織の原則、方針、戦略的優先事項と整合させるための仕組みを提供する。
- AI システムの設計と開発の技術的側面を組織の価値観と原則に結び付け、そのようなシステムの取得、学習、デプロイ、モニタリングに携わる個人のための組織的慣行と能力を可能にする。サードパーティのソフトウェアまたはハードウェアシステムおよびデータの使用に関する法的およびその他の問題を含む、製品のライフサイクル全体と関連プロセスに対応する。

GOVERN は、AI のリスクマネジメント全体に浸透し、プロセスの他の機能を可能にする横断的な機能です。**GOVERN** の側面、特にコンプライアンスや評価に関連する側面は、他の各機能に統合されるべきです。GOVERN への留意は、AI システムの寿命および組織の階層にわたって効果的な AI リスクマネジメントを行うための継続的かつ本質的な要件です。

強力な GOVERN は、組織のリスク文化を促進するために、社内の慣行や規範を推進し、強化することができます。GOVERN 責任者は、組織の使命、ゴール、価値観、文化、リスク許容度を方向付ける包括的な方針を決定することができます。シニアリーダーシップは、組織内のリスクマネジメントの方針を打ち出し、それに伴って組織文化も作る。経営陣は、AI のリスクマネジメントの技術的側面を方針と業務に整合させる。文書化することで、透明性を高め、人的レビュープロセスを改善し、AI システムチームのアカウンタビリティを強化することができます。

GOVERN 機能で説明した構造、システム、プロセス、チームを導入した後、組織はリスクの理解とマネジメントに焦点を当てた目的主導の文化から恩恵を受けるはずです。知識、文化、AI アクターのニーズや期待が時間経過とともに変化する中で、**GOVERN** 機能を実行し続けることは、フレームワーク利用者の責務です。

AI リスクの GOVERN に関するプラクティスは、NIST AI RMF Playbook に記載されています。表 1 に **GOVERN** 機能のカテゴリーとサブカテゴリーを示しています。

表 1：**GOVERN** 機能のカテゴリーとサブカテゴリー

カテゴリー	サブカテゴリー
GOVERN 1：AI リスクの MAP、MEASURE、MANAGE に関連する組織全体の方針、プロセス、手順および実務が整備され、透明性があり、効果的に実施されている。	**GOVERN 1.1**：AI に関わる法的・規制的要件が理解され、マネジメントされ、文書化されている。 **GOVERN 1.2**：トラストワージな AI の特性が、組織の方針、プロセス、手順および実務に統合されている。 **GOVERN 1.3**：組織のリスク許容度に基づいて、必要なリスクマネジメント活動のレベルを決定するためのプロセス、手順および実務が整備されている。 **GOVERN 1.4**：組織のリスクの優先順位に基づき、透明性のある方針、手順、その他の管理を通じて、リスクマネジメントプロセスとそのアウトカムが確立されている。

[下記に続く]

表 1：**GOVERN** 機能のカテゴリーとサブカテゴリー[下記に続く]

カテゴリー	サブカテゴリー
	GOVERN 1.5：リスクマネジメントプロセスとその成果の継続的なモニタリングと定期的なレビューが計画され、定期的なレビューの頻度を決定することを含め、組織の役割と責任が明確に定義されている。 **GOVERN 1.6**：AI システムの棚卸しのための仕組みが整備され、組織のリスク優先順位に応じたリソースが確保されている。 **GOVERN 1.7**：AI システムを、リスクを増大させたり組織のトラストワージネスを低下させたりしない方法で、安全に廃止し段階的に廃止するためのプロセスと手順が整備されている。
GOVERN 2：アカウンタビリティの体制が整備され、適切なチームと個人が、AI リスクの MAP、MEASURE、MANAGE に権限を与えられ、責任を負い、学習される。	**GOVERN 2.1**：AI リスクの MAP、MEASURE、MANAGE に関連する役割と責任およびコミュニケーションラインが文書化され、組織全体の個人とチームに明確になっている。 **GOVERN 2.2**：組織の要員およびパートナーは、関連する方針、手順および協定に沿った職務と責任を果たすことができるよう、AI リスクマネジメントのトレーニングを受ける。 **GOVERN 2.3**：組織におけるエグゼクティブ・リーダーシップは、AI システムの開発・デプロイに伴うリスクに関する意思決定に責任を持つ。
GOVERN 3：従業員の多様性、公平性、インクルージョンおよびアクセシビリティのプロセスは、ライフサイクル全体を通じて、AI リスクのマッピング、測定およびマネジメントにおいて優先される。	**GOVERN 3.1**：ライフサイクル全体を通じて、AI リスクの MAP、MEASURE、MANAGE に関連する意思決定が、多様なチーム（例えば、人口統計、分野、経験、専門知識、経歴の多様性）によって行われる。 **GOVERN 3.2**：人間と AI の構成や AI システムのモニターに関する役割と責任を定義し、区別するための方針と手順が整備されている。
GOVERN 4： 組織チームは、AI リスクを考慮し、コミュニケーションする文化にコミットしている	**GOVERN 4.1**：AI システムの設計、開発、デプロイ、使用において、批判的思考と安全第一の考え方を育成し、潜在的なネガティブインパクトを最小限に抑える組織の方針と実務が整備されている。

[下記に続く]

表 1：**GOVERN** 機能のカテゴリーとサブカテゴリー[下記に続く]

カテゴリー	サブカテゴリー
	GOVERN 4.2：組織チームは、設計、開発、デプロイ、評価、使用する AI 技術のリスクと潜在的インパクトを文書化し、そのインパクトについて広く伝える。 **GOVERN 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ある。
GOVERN 5：関連する AI アクターとの堅牢な連携のためのプロセスが整備されている。	**GOVERN 5.1**：AI リスクに関連する潜在的な個人的・社会的インパクトについて、AI システムを開発・導入したチー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実務が整備されている。 **GOVERN 5.2**：AI システムを開発またはデプロイしたチームが、関連する AI アクターからの裁定されたフィードバックを定期的にシステムの設計と実施に取り入れることができるような仕組みが確立されている。
GOVERN 6：サードパーティのソフトウェアやデータ、その他のサプライチェーンの問題から生じる AI のリスクとベネフィットに対処するための方針と手順が整備されている。	**GOVERN 6.1**：サードパーティの知的財産権その他の権利の侵害リスクを含め、サードパーティ・エンティティに関連する AI リスクに対処する方針と手順が整備されている。 **GOVERN 6.2**：高リスクとみなされるサードパーティのデータまたは AI システムの障害またはインシデントに対処するためのコンティンジェンシープロセスが整備されている。

5.2 MAP

MAP 機能は、AI システムに関連するリスクを枠組みするためのコンテクストを確立します。AI のライフサイクルは、多様な関係者が関与する多くの相互依存的な活動から構成されます（図 3 参照）。実際には、プロセスの一部分を担当する AI アクターは、他の部分や関連するコンテクストを完全に可視化または制御できないことがよくあります。これらの活動間の相互依存関係や、関連する AI アクター間の相互依存関係により、AI システムのインパクトを確実に予測することが困難になる可能性があります。例えば、AI システムの目的と目標を特定する際の初期の決定が、その挙動と能力を変化させる場合があり、（エンドユーザやインパクトを受ける個人などの）デプロイ環境のダイナミクスが、AI システムの決定のインパクトを形成する可能性があります。その結果、AI のライフサイクルのある側面における最善の意図が、それ以降の他の活動における決定や条件との相互作用によって損なわれることがあります。

このような複雑さと様々なレベルの可視性は、リスクマネジメントの実践に不確実性をもたらすことがあります。潜在的なネガティブなリスク源を予測し、評価し、その他の方法で対処することにより、この不確実性を軽減し、意思決定プロセスの完全性を高めることができます。

MAP 機能の実行中に収集された情報は、ネガティブリスクの予防を可能にし、モデルのマネジメントなどのプロセスの意思決定や、AI ソリューションの適切性や必要性に関する最初の意思決定に情報を提供します。**MAP** 機能の結果は、**MEASURE** 機能と **MANAGE** 機能の基礎となります。コンテクストに関する知識および特定されたコンテクスト内のリスクに関する認識がなければ、リスクマネジメントの実行は困難です。**MAP** 機能は、リスクとより広範な要因を特定する組織の能力を強化することを目的としています。

この機能の実施は、社内の多様なチームからの視点を取り入れ、AI システムを開発またはデプロイしたチームの外部の人々との関わりを持つことによって強化されます。社外の協力者、エンドユーザ、インパクトを受ける可能性のあるコミュニティなどとの関わりは、特定の AI システムのリスクレベル、社内チームの構成、組織の方針によって異なる場合があります。このような広範な視点を集めることは、以下のような方法で、組織がネガティブなリスクを未然に防ぎ、よりトラストワージな AI システムを開発するのに役立ちます。

- コンテクストを理解する能力を向上させる。
- 使用のコンテクストに関する前提をチェックする。
- システムが意図されたコンテクスト内外で機能しない場合を認識できるようにする。
- 既存の AI システムのポジティブでベネフィットのある使用方法を特定する。
- AI および機械学習プロセスにおける限界の理解を向上させる。
- 実世界のアプリケーションにおいて、ネガティブインパクトにつながる可能性のある制約を特定する。
- AI システムの意図された使用に関連する既知および予見可能なネガティブインパクトを特定する。
- AI システムの意図された使用以外の使用によるリスクを予測する。

MAP 機能を完了した後、フレームワークの利用者は、AI システムを設計、開発、またはデプロイするかどうかに関する最初の実施／非実施の決定を下すために、AI システムのインパクトに関する十分なコンテクストに関する知識を持つべきです。続行することを決定した場合、組織は **MEASURE** 機能と **MANAGE** 機能を、**GOVERN** 機能で整備された方針と手順とともに使用し、AI リスクマネジメントの取り組みに役立てるべきです。AI システムのコンテクスト、能力、リスク、ベネフィット、潜在的なインパクトが時間経過とともに変化する中で、MAP 機能を AI システムに適用し続けることは、フレームワークの利用者の責務です。

AI リスクの **MAP** に関するプラクティスは、NIST AI RMF Playbook に記載されている。表 2 に **MAP** 機能のカテゴリーとサブカテゴリーを示しています。

表 2：**MAP** 機能のカテゴリーとサブカテゴリー

カテゴリー	サブカテゴリー
MAP 1：コンテクストが確立され、理解されている。	**MAP 1.1**：意図された目的，潜在的にベネフィットある有益な用途，コンテクスト特有の法律，規範および期待，並びに AI システムが展開デプロイされる見込みのある設定が理解され，文書化される。考慮すべき事項には、具体的なユーザ利用者の集合や種類とその期待、システムの利用が個人、地域社会、組織、社会、地球に及ぼす潜在的なポジティブインパクトとネガティブインパクトの影響、開発または製品の AI ライフサイクル全体にわたる AI システムの目的、利用、リスクに関する仮定と関連する制限、関連するテスト、評価、妥当性確認、検証 （TEVV） とシステム指標などが含まれる。 **MAP 1.2**：コンテクストを確立するための学際的な AI アクター、コンピテンシー、スキルおよび能力は、人口統計学的な多様性、幅広い領域およびユーザ経験の専門性を反映しており、その参加は文書化されている。学際的な協力の機会が優先される。 **MAP 1.3**：AI 技術に関する組織のミッションと関連する目標が理解され、文書化されている。 **MAP 1.4**：ビジネス価値またはビジネス利用の背景が明確に定義されているか、あるいは（既存の AI システムを評価する場合には）再評価されている。 **MAP 1.5**：組織のリスク許容度が決定され、文書化されている。 **MAP 1.6**：システム要件（「システムはユーザのプライバシーを尊重すること」など）が、関連する AI アクターから引き出され、理解されている。AI リスクに対処するため、社会技術的なインパクトを考慮した設計上の決定が行われる。
MAP 2：AI システムの分類が行われる。	**MAP 2.1**：AI システムがサポートする具体的なタスクとそれを実装するために使用される方法が定義されている（分類器、生成モデル、レコメンダーなど）。 **MAP 2.2**：AI システムの知識の限界およびシステムの出力がどのように利用され、人間によってオーバーサイトされる可能性があるかについての情報が文書化されている。文書化は、関連する AI アクターが意思決定を行い、その後の行動を取る際に役立つ十分な情報を提供する。

［下記に続く］

表 2：**MAP** 機能のカテゴリーとサブカテゴリー[下記に続く]

カテゴリー	サブカテゴリー
	MAP 2.3：実験計画、データ収集および選択（可用性、代表性、適合性等）、システムのトラストワージネス、 構成の妥当性確認に関するものを含め、科学的公平性および TEVV に関する考慮事項が特定され、 文書化されている。
MAP 3：AI の能力、目標とする用途、目標および適切なベンチマークと比較した期待されるベネフィットとコストが理解されている。	**MAP 3.1**：意図された AI システムの機能および性能の潜在的ベネフィットが検討され、文書化される。 **MAP 3.2**： 予期されるまたは実現される AI のエラーまたはシステムの機能性および信頼性に起因する、非金銭的コストを含む潜在的コスト（組織のリスク許容度に関連する）が検討され、文書化される。 **MAP 3.3**：システムの能力、確立されたコンテクストおよび AI システムの分類に基づいて、対象とする適用範囲が特定され、文書化される。 **MAP 3.4**：AI システムの性能と信頼性に関するオペレータと実務者の習熟のためのプロセスおよび関連する技術標準と認証が定義され、評価され、文書化されている。 **MAP 3.5**：**GOVERN** 機能の組織方針に従って、ヒューマンオーバーサイトのためのプロセスが定義され、評価され、文書化されている。
MAP 4：サードパーティのソフトウェアやデータを含む AI システムのすべての構成要素について、リスクとベネフィットがマッピングされている。	**MAP 4.1**：サードパーティの知的財産権その他の権利の侵害のリスクと同様に、サードパーティのデータやソフトウェアの使用を含む、AI 技術とその構成要素の法的リスクをマッピングするためのアプローチが整備され、遵守され、文書化されている。 **MAP 4.2**：サードパーティの AI 技術を含む AI システムの構成要素に対する内部リスクマネジメントが特定され、文書化されている。
MAP 5：個人、グループ、コミュニティ、組織、社会へのインパクトが特徴づけられている。	**MAP 5.1**：予想される使用、類似のコンテクストにおける AI システムの過去の使用、公的な事故報告、AI システムを開発またはデプロイしたチームの外部からのフィードバック、またはその他のデータに基づき、特定された各インパクトの頻度と大きさ（潜在的ベネフィットと有害なものの両方）を特定され、文書化されている。 **MAP 5.2**：関連する AI アクターとの定期的な関わりを支援し、ポジティブ、ネガティブおよび予期せぬインパクトに関するフィードバックを統合するための実務と人材が配置され、文書化されている。

5.3 MEASURE

MEASURE 機能は、AI リスクと関連するインパクトを分析、評価、ベンチマーク、モニターするために、定量的、定性的、または混合法のツール、技法、方法論を用います。**MAP** 機能で特定された AI リスクに関連する知識を使用し、**MANAGE** 機能に情報を提供します。AI システムはデプロイ前にテストされ、運用中も定期的にテストされるべきです。AI リスクの測定には、システムの機能性とトラストワージネスの側面を文書化することが含まれます。

AI リスクの測定には、トラストワージ特性、社会的インパクト、人間と AI の構成の測定基準を追跡することが含まれます。**MEASURE** 機能において開発または採用されるプロセスには、関連する不確実性の尺度、性能ベンチマークとの比較および結果の正式な報告および文書化を伴う、厳格なソフトウェアテストおよび性能評価方法論が含まれる必要があります。独立したレビューのプロセスは、テストの有効性を向上させ、内部的な偏りや潜在的な利害の対立を緩和することができます。

トラストワージの特性の間でトレードオフが生じる場合、測定は、マネジメント上の意思決定に情報を提供するための追跡可能な基礎を提供します。選択肢には、再校正、インパクトの緩和、設計・開発・生産・使用からのシステム除去のほか、補償、検知、抑止、指示、回復の各制御が含まれます。

MEASURE 機能を完了した後、測定基準、方法論、手法を含む客観的、反復可能、又はまたは拡張可能な試験テスト、評価、妥当性確認、および検証（TEVV）プロセスが実施され、それに従い、文書化されます。測定基準および測定方法は、科学的、法的、倫理的規範を遵守し、オープンで透明性のあるプロセスで実施されなければなりません。質的および量的な新しいタイプの測定を開発する必要があるかもしれません。各測定手法が、AI リスクの評価において、どの程度ユニークで意味のある情報を提供するかを検討すべきです。フレームワークの利用者は、システムのトラストワージネス信頼性を総合的に評価し、既存のリスクと顕在化するリスクを特定・追跡し、測定基準の有効性を検証する能力を高めます。測定結果は **MANAGE** 機能で活用使用され、リスクのモニタリング監視と対応への取り組み努力を支援します。知識、方法論、リスク、インパクト影響が時とともに進化するにつれて、**MEASURE** 機能を AI システムに適用し続けることは、フレームワーク・ユーザの責務です。

AI リスクの **MEASURE** に関するプラクティスは、NIST AI RMF Playbook に記載されています。表 3 に **MEASURE** 機能のカテゴリーとサブカテゴリーを示しています。

表 3：**MEASURE** 機能のカテゴリーとサブカテゴリー

カテゴリー	サブカテゴリー
MEASURE 1：適切な方法と測定基準を特定し、適用する。	**MEASURE 1.1**：**MAP** 機能において列挙された AI リスクの測定のための手法および測定基準は、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（または測定できない）リスクまたはトラストワージの特性が適切に文書化されている。 **MEASURE 1.2**：エラーの報告や影響を受けるコミュニティへの潜在的な影響を含め、AI の測定基準の適切性と既存のマネジメントの有効性が定期的に評価され、更新される。 **MEASURE 1.3**： システムの第一線の開発者ではない内部の専門家および／または独立した評価者が、定期的な評価と更新に関与している。ドメインの専門家、ユーザ、AI システムを開発またはデプロイしたチームの外部の AI アクターおよび影響を受けるコミュニティは、組織のリスク許容度に従って、必要に応じて評価のサポートに相談される。
MEASURE 2：AI システムは、トラストワージ特性について評価される。	**MEASURE 2.1**：テストセット、測定基準、TEVV で使用したツールの詳細が文書化されている。 **MEASURE 2.2**：人間を対象とする評価は、適用される要件（被験者保護を含む）を満たし、関連する集団を代表するものである。 **MEASURE 2.3**：AI システムの性能または保証基準が定性的または定量的に測定され、デプロイの設定に 類似した条件下で実証されている。測定方法は文書化されている。 **MEASURE 2.4**：AI システムおよびその構成要素（MAP 機能で特定される）の機能および挙動が、実稼働時にモニタリングされている。 **MEASURE 2.5**：デプロイされる AI システムが妥当で信頼できることが実証される。その技術が開発された条件を超える汎用化可能性の限界が文書化されている。

［下記に続く］

表 3：**MEASURE** 機能のカテゴリーとサブカテゴリー[下記に続く]

カテゴリー	サブカテゴリー
	MEASURE 2.6：**MAP** 機能で特定された安全に関するリスクについて、AI システムが定期的に評価されている。デプロイされる 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に安全側に故障することができる。安全性の指標は、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対応時間を反映されている。 **MEASURE 2.7**：AI システムのセキュリティとレジリエンス（MAP 機能で特定される）が評価され、文書化されている。 **MEASURE 2.8**：透明性とアカウンタビリティに関連するリスク（MAP 機能で特定される）が調査され、文書化されている。 **MEASURE 2.9**：AI モデルが説明され、検証され、文書化され、AI システムの出力が、**MAP** 機能で特定されているように、そのコンテクストの中で解釈され、責任ある使用とガバナンスに反映される。 **MEASURE 2.10**：AI システムのプライバシーリスク（**MAP** 機能で特定される）が調査され、文書化されている。 **MEASURE 2.11**：公平性とバイアス（**MAP** 機能で特定される）が評価され、その結果が文書化される。 **MEASURE 2.12**：AI モデルの学習およびマネジメント活動の環境へのインパクトと持続可能性（**MAP** 機能において特定される）がアセスメントされ、文書化される。 **MEASURE 2.13**：**MEASURE** 機能において採用された TEVV 指標およびプロセスの有効性が評価され、文書化される。
MEASURE 3：特定された AI リスクを長期にわたって追跡する仕組みが整備されている。	**MEASURE 3.1**：デプロイされた状況における意図された性能および実際の性能などの要因に基づき、既存、予期されないおよび出現する AI リスクを定期的に特定し、追跡するためのアプローチ、要員および文書化が実施されている。 **MEASURE 3.2**：現在利用可能な測定技術では AI リスクを評価することが困難である、または測定基準がまだ利用可能でない環境について、リスク追跡アプローチを検討する。

[下記に続く]

表 3：**MEASURE** 機能のカテゴリーとサブカテゴリー[下記に続く]

	MEASURE 3.3： エンドユーザやインパクトを受けるコミュニティが問題を報告し、システムの成果を訴えるためのフィードバックプロセスが確立され、AI システムの評価指標に統合される。
MEASURE 4：測定の有効性に関するフィードバックを収集し、評価する。	**MEASURE 4.1**：AI のリスクを特定するための測定アプローチが、配備の背景と関連付けられ、分野の専門家や他のエンドユーザとの協議を通じて知らされている。アプローチは文書化されている。 **MEASURE 4.2**： AI システムの信頼性に関する測定結果が、配備状況および AI ライフサイクル全体にわたって、ドメインの専門家および関連する AI アクターからの入力によって通知され、システムが意図されたとおりに一貫して機能しているかどうかが検証される。結果は文書化される。 **MEASURE 4.3**：インパクトを受けるコミュニティを含む関連する AI アクターとの協議や、状況に関連するリスクやトラストワージ特性に関する現地データに基づき、測定可能な性能の改善または低下が特定され、文書化される。

5.4 MANAGE

MANAGE 機能とは、**GOVERN** 機能によって定義されたとおり、定期的に、MAP され、MEASURE されたリスクにリスク・リソースを割り当てることを含みます。リスク処置は、インシデントや事象への対応、回復、コミュニケーションに関する計画で構成されます。

GOVERN で確立され、**MAP** で実施される、専門家による協議や関連する AI アクターからのインプットから得られるコンテクストに関する情報は、システム障害やネガティブインパクトの可能性を低減するために、この機能で使用されます。**GOVERN** で確立され、**MAP** と **MEASURE** で使用されている体系的な文書化手法は、AI のリスクマネジメントの取り組みを強化し、透明性とアカウンタビリティを高めます。緊急リスクを評価するプロセスは、継続的改善の仕組みとともに整備されています。

MANAGE 機能の完了後、リスクの優先順位付けと定期的なモニタリングと改善計画が実施されます。フレームワーク利用者は、デプロイされた AI システムのリスクをマネジメントし、評価され優先順位付けされたリスクに基づいてリスクマネジメントのリソースを配分する能力が強化されます。フレームワーク利用者は、手法、コンテクスト、リスク、AI アクターのニーズや期待が時とともに変化するのに応じて、デプロイされた AI システムに **MANAGE** 機能を適用し続けることが求められます。

AI リスクの MANAGE に関連するプラクティスは、NIST AI RMF Playbook に記載されています。表 4 に **MANAGE** 機能のカテゴリーとサブカテゴリーを示しています。

表 4：**MANAGE** 機能のカテゴリーとサブカテゴリー

カテゴリー	サブカテゴリー
MANAGE 1：**MAP** および **MEASURE** 機能からの評価およびその他の分析出力に基づく AI リスクを優先順位付けし、対応し、マネジメントする。	**MANAGE 1.1**：AI システムが意図された目的および明示された目標を達成しているかどうか、また、その開発または配備を進めるべきかどうかについての判断が行われる。 **MANAGE 1.2**：文書化された AI リスクの処置は、インパクト、可能性、利用可能なリソースまたは方法に基づいて優先順位付けされる。 **MANAGE 1.3**：**MAP** 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、緩和、移転、回避、受容が含まれる。 **MANAGE 1.4**：AI システムの川下にいる取得者とエンドユーザの両方に対するネガティブ残留リスク（未解決のすべてのリスクの合計として定義される）を文書化する。
MANAGE 2：AI のベネフィットを最大化し、ネガティブインパクトを最小化するための戦略が、計画、準備、実施、文書化され、関連する AI アクターからの情報に基づいている。	**MANAGE 2.1**：AI のリスクをマネジメントするために必要なリソースが、実行可能な AI 以外の代替システム、アプローチ、または方法とともに考慮され、潜在的なインパクトの大きさや可能性が低減される。 **MANAGE 2.2**：デプロイされた AI システムの価値を維持するための仕組みが整備され、適用されている。 **MANAGE 2.3**：未知のリスクが特定された場合、そのリスクに対応し、回復するための手順が守られている。 **MANAGE 2.4**：意図された使用と異なる性能やアウトカムを示す AI システムを廃止、解除、停止するための仕組みが整備され、適用され、責任が割り当てられ、理解されている。
MANAGE 3：サードパーティ・エンティティからの AI のリスクとベネフィットがマネジメントされている。	**MANAGE 3.1** ：サードパーティリソースからの AI のリスクとベネフィットが定期的にモニターされ、リスクマネジメントが適用され、文書化される。 **MANAGE 3.2**：開発に使用される事前学習済みモデルは、AI システムの定期的なモニタリングと保守の一環としてモニターされる。

表 4：**MANAGE** 機能のカテゴリーとサブカテゴリー

カテゴリー	サブカテゴリー
MANAGE 4：特定され、測定された AI リスクに対する、対応と復旧を含むリスク処置とコミュニケーション計画が文書化され、定期的にモニターされる。	**MANAGE 4.1** ：ユーザおよびその他の関連する AI アクターからのインプットを取得し評価する仕組み、不服申し立てと無効化、廃止、インシデント対応、復旧、変更管理を含む、デプロイ後の AI システムモニタリング計画が実施されている。 **MANAGE 4.2**：継続的な改善のための測定可能な活動が、AI システムの更新に組み込まれ、関連する AI アクターを含む利害関係者との定期的な関与を含む。 **MANAGE 4.3**：インシデントやエラーは、インパクトを受けるコミュニティを含む、関連する AI アクターに通知される。インシデントやエラーの追跡、対応、回復のプロセスが守られ、文書化されている。

6. AI RMF プロファイル

AI RMF ユースケースのプロファイルは、フレームワーク利用者の要件、リスク許容度、リソースに基づく特定の設定またはアプリケーションのための AI RMF 機能、カテゴリー、サブカテゴリーの実装です。たとえば、AI RMF 採用プロファイルや AI RMF 公正住宅プロファイルなどです。 プロファイルは、AI ライフサイクルのさまざまな段階、または特定のセクター、テクノロジー、または最終用途のアプリケーションでリスクをマネジメントする方法について説明し、洞察を提供します。AI RMF プロファイルは、組織が目標に沿って AI リスクをマネジメントする最適な方法を決定し、法的/規制要件とベスト プラクティスを考慮し、リスクマネジメントの優先順位を反映する方法を決定するのに役立ちます。

AI RMF のプロファイルは、組織が、その目標に合致し、法的／規制上の要件やベストプラクティスを考慮し、リスクマネジメントの優先順位を反映した AI リスクをどのようにマネジメントするのが最善かを決定する際に役立ちます。AI RMF の時系列プロファイルは、特定のセクター、業界、組織、またはアプリケーションのコンテクストにおける特定の AI リスクマネジメント活動の現状または望ましい目標状態のいずれかを記述したものです。AI RMF の現在のプロファイルは、AI が現在どのようにマネジメントされているか、また、現在の成果という観点から関連するリスクを示しています。ターゲット・プロファイルは、望ましい、あるいはターゲットとする AI リスクマネジメントのゴールを達成するために必要な結果を示します。

現状のプロファイルとターゲット・プロファイルを比較すると、AI リスクマネジメントの目標を達成するために取り組むべきギャップが明らかになる可能性が高いことがわかります。アクションプランは、所定のカテゴリーまたはサブカテゴリーのアウトカムを達成するために、これらのギャップに対処するために策定することができます。ギャップ緩和の優先順位付けは、ユーザのニーズとリスクマネジメントプロセスによって行われます。また、このリスクベースのアプローチにより、フレームワークの利用者は、自らのアプローチを他のアプローチと比較し、AI リ

スクマネジメントの目標を達成するために必要なリソース（人員、資金など）を、費用対効果の高い優先順位付けされた方法で測定することができます。

AI RMF の分野横断的プロファイルは、ユースケースや分野を超えて使用可能なモデルやアプリケーションのリスクを対象としています。分野横断的なプロファイルは、大規模な言語モデルの使用、クラウドベースのサービス、買収など、分野を超えて共通する活動やビジネスプロセスのリスクを GOVERN、MAP、MEASURE、MANAGE する方法をカバーすることもできます。

このフレームワークでは、プロファイルのテンプレートを規定していないため、実装に柔軟性があります。

付録 A：
図 2 および図 3 の AI アクターのタスクの説明

AI 設計タスクは、図 2 の AI ライフサイクルのアプリケーション・コンテクストおよびデータと入力のフェーズで実行されます。AI 設計アクターは、AI システムのコンセプトと目的を作成し、AI システムが合法的で目的に適合するように、AI システムの計画、設計、データ収集・処理タスクを担当します。タスクには、システムのコンセプトと目的、基礎となる仮定、コンテクスト、要件の明確化と文書化、データの収集とクリーニング、データセットのメタデータと特性の文書化が含まれます。このカテゴリーの AI アクターには、データサイエンティスト、ドメイン専門家、社会文化アナリスト、多様性、公平性、インクルージョン、アクセシビリティ分野の専門家、インパクトを受けるコミュニティのメンバー、ヒューマンファクターの専門家（UX/UI デザインなど）、ガバナンスの専門家、データエンジニア、データプロバイダー、システム出資者、プロダクトマネージャー、サードパーティ・エンティティ、評価者、法務・プライバシーガバナンスなどが含まれます。

AI 開発タスクは、図 2 のライフサイクルの AI モデルフェーズで実行されます。AI 開発のアクターは、AI システムの初期インフラを提供し、モデルまたはアルゴリズムの作成、選択、キャリブレーション、トレーニング、および/またはテストを含む、モデル構築および解釈タスクを担当します。このカテゴリーの AI アクターには、機械学習の専門家、データサイエンティスト、開発者、サードパーティ・エンティティ、法的およびプライバシーガバナンスの専門家、デプロイの設定に関連する社会文化的およびコンテクスト的要因の専門家が含まれます。

AI デプロイメントのタスクは、図 2 のライフサイクルのタスクとアウトプットのフェーズで実行されます。AI デプロイのアクターは、システムの実運用導入を確実にするために、AI システムをどのように使用するかに関連するコンテクスト上の決定を担当します。関連するタスクには、システムの試験運用、レガシーシステムとの互換性の確認、規制遵守の確保、組織変更の管理、ユーザーエクスペリエンスの評価などがあります。このカテゴリーの AI アクターには、システムインテグレーター、ソフトウェア開発者、エンドユーザ、オペレータや実務者、評価者、ヒューマンファクター、社会文化分析、ガバナンスの専門知識を持つドメインの専門家が含まれます。

運用とモニタリングのタスクは、図 2 のライフサイクルのアプリケーション・コンテクスト／運用とモニタリングのフェーズで実行されます。これらのタスクは、AI システムの運用を担当し、他者と協力してシステムの出力とインパクトを定期的に評価する AI アクターによって実行されます。このカテゴリーに属する AI アクターには、システム・オペレータ、ドメインの専門家、AI 設計者、AI システムの出力を解釈したり取り入れたりするユーザ、製品開発者、評価者や監査人、コンプライアンスの専門家、組織管理者、研究コミュニティのメンバーなどが含まれます。

テスト、評価、妥当性確認、検証（TEVV）タスクは、AI のライフサイクル全体を通じて実行されます。これらは、AI システムやそのコンポーネントを検査したり、問題を検出して修正したりする AI アクターによって実行されます。理想的には、検証・妥当性確認タスクを実行します。

AI アクターは、テスト・評価アクションを実行する AI アクターとは区別されます。タスクは設計の初期段階からフェーズに組み込むことができ、そこでは設計要件に従ってテストが計画されます。

- 設計、計画、データに関する TEVV タスクは、システム設計、データ収集およびデプロイや応用の意図された状況 に関連する測定に関する前提条件の内部および外部検証を中心とすることがあります。
- 開発（モデル構築）のための TEVV タスクには、モデルの妥当性確認と評価が含まれます。
- デプロイのための TEVV タスクには、システムの妥当性確認と実運用への統合が含まれ、システム とプロセスの統合、ユーザ体験、既存の法的、規制的、倫理的仕様への準拠のためのテスト と再較正が含まれます。
- 運用のための TEVV タスクには、モデルの定期的な更新、テスト、専門家（SME）による再較正のための継続的なモニタリング、報告されたインシデントやエラーの追跡とそのマネジメント、創発特性の検出と関連するインパクト、救済措置と対応のためのプロセスが含まれます。

ヒューマンファクターのタスクと活動は、AI のライフサイクルのあらゆる側面で見られる。これには、人間中心設計の実践と方法論、エンドユーザやその他の関係者、関連する AI アクターの積極的な関与の促進、システム設計におけるコンテクスト固有の規範と価値観の取り込み、エンドユーザの経験の評価と適応、AI ライフサイクルの全段階における人間とヒューマンダイナミクスの広範な統合などが含まれます。ヒューマンファクターの専門家は、使用のコンテクストを理解し、学際的かつ人口統計学的な多様性に情報を提供し、協議プロセスに関与し、ユーザーエクスペリエンスを設計および評価し、人間中心の評価およびテストを実行し、インパクト評価に情報を提供するために、学際的なスキルと視点を提供します。

ドメインの専門家のタスクには、AI システムが使用される産業部門、経済部門、コンテクスト、アプリケーションのドメインに関する知識や専門知識を提供する学際的な実務者や学者からのインプットが含まれます。ドメインの専門家である AI アクターは、AI システムの設計や開発に不可欠なガイダンスを提供し、TEVV や AI インパクト評価チームが行う作業をサポートするためにアウトプットを解釈することができます。

AI インパクトアセスメントのタスクには、AI システムのアカウンタビリティ、有害なバイアスへの取り組み、AI システムのインパクトの検討、製品の安全性、責任、セキュリティなどの要件の評価と査定が含まれます。インパクトの審査員や評価者などの AI アクターは、技術的、人的要因、社会文化的、法的な専門知識を提供します。

調達タスクは、サードパーティの開発者、ベンダー、請負業者から AI モデル、製品、またはサービスを取得するために、財務、法務、または方針に関するマネジメント権限を持つ AI アクターによって実施されます。

ガバナンスとオーバーサイトのタスクは、AI システムがデサイン、開発および/またはデプロイされる組織のマネジメント、受託および法的権限と責任を持つ AI アクターが担います。

AI ガバナンスに責任を負う主な AI アクターには、組織管理者、上位指導者、取締役会が含まれます。これらのアクターは、組織全体のインパクトと持続可能性に関わる当事者です。

その他の AI アクター

サードパーティ・エンティティには、他の組織や組織の顧客やクライアントのために、データ、アルゴリズム、モデル、システム、および／または関連サービスを提供するプロバイダー、開発者、ベンダー、評価者が含まれます。サードパーティ・エンティティは、全体的または部分的に、AI の設計および開発タスクを担当する。定義上、サードパーティは、その技術やサービスを取得する組織の設計、開発、デプロイチームの外部です。サードパーティ・エンティティから取得したテクノロジーは複雑で不透明な場合があり、リスク許容度が導入組織や運用組織と一致しない場合があります。

AI システムのエンドユーザは、特定の目的のためにシステムを使用する個人またはグループです。これらの個人またはグループは、特定のコンテクストで AI システムと相互作用します。エンドユーザの能力は、AI の専門家から初めてテクノロジーを使うエンドユーザまで多岐にわたります。

影響を受ける個人／コミュニティは、AI システムまたは AI システムの出力に基づく意思決定によって直接的または間接的に影響を受けるすべての個人、グループ、コミュニティ、または組織を包含します。これらの個人は、必ずしもデプロイされたシステムやアプリケーションと対話するとは限りません。

その他の AI アクターは、AI リスクを特定しマネジメントするための公式または準公式な規範またはガイダンスを提供することができます。このような主体には、**業界団体、標準化団体、アドボカシ団体、研究者、環境保護団体、市民社会組織**などがあります。

一般市民は、AI 技術がもたらすポジティブインパクトもネガティブインパクトも直接経験する可能性が最も高くあります。一般市民は、AI アクターがとる行動の動機付けとなる可能性があります。このグループには、AI システムが開発・デプロイされる状況に関連する個人、コミュニティ、消費者が含まれます。

付録 B：
AI のリスクは従来のソフトウェアのリスクとどのように異なるか

従来のソフトウェアと同様に、AI ベースの技術によるリスクは企業よりも大きく、組織をまたぎ、社会的インパクトにつながる可能性があります。また、AI システムは、現在のリスクフレームワークやアプローチでは包括的に対処できない一連のリスクをもたらします。リスクをもたらす AI システムの機能の中には、ベネフィットを持つものもあります。例えば、事前に学習されたモデルや転移学習は、他のモデルやアプローチと比較して、研究を進展させ、正確性とレジリエンスを高めることができます。MAP 機能におけるコンテクスト的要因を特定することは、AI アクターがリスクのレベルと潜在的なマネジメントの取り組みを決定する際に役立ちます。

従来のソフトウェアと比較して、新たに発生したり増加したりする AI 特有のリスクには以下のようなものがあります：

- AI システムを構築するために使用されるデータは、AI システムのコンテクストや意図された用途を正しくまたは適切に表現しているとは限らず、真の値が存在しないか、利用できない可能性がある。さらに、有害なバイアスやその他のデータ品質の問題は、AI システムのトラストワージネスにインパクトを及ぼし、ネガティブインパクトをもたらす可能性がある。
- AI システムの依存性と学習タスクのためのデータへの依存は、そのようなデータに典型的に関連する増大した量と複雑さと組み合わさっている。
- トレーニング中の意図的または非意図的な変更により、AI システムの性能が根本的に変化する可能性がある。
- AI システムの学習に使用されるデータセットが、本来の意図されたコンテクストから切り離されたり、デプロイのコンテクストに対して古くなったり、時代遅れになったりする可能性がある。
- AI システムの規模と複雑性（多くのシステムには数十億、あるいは数兆の意思決定ポイントが含まれている）が、より伝統的なソフトウェア・アプリケーション内に収容されている。
- 研究を進展させ、性能を向上させることができる事前学習されたモデルの使用は、統計的不確実性のレベルを増加させ、バイアス管理、科学的妥当性、再現性の問題を引き起こす可能性もある。
- 大規模な事前学習済みモデルの出現特性に対する故障モードの予測難易度が高くなる。
- AI システムのデータ集約能力の向上によるプライバシーリスク。
- AI システムは、データ、モデル、コンセプトのドリフトに起因する、より頻繁なメンテナンスと、修正メンテナンスを実施するためのトリガーを必要とする可能性がある。
- 不透明性の増大と再現性への懸念。
- ソフトウェアのテスト基準が未整備であり、最も単純なケースを除き、従来から設計されている。ソフトウェアに期待される基準で AI ベースのプラクティスを文書化することができない。
- AI システムは従来のコード開発と同じ管理対象ではないため、AI ベースのソフトウェアテストを定期的に実施すること、あるいは何をテストすべきかを決定することが困難である。

- AI システム開発のための計算コストと、環境や地球への影響。
- AI ベースのシステムの副作用を統計的な手段を超えて予測・検出することができない。

プライバシーとサイバーセキュリティのリスクマネジメントに関する考慮事項とアプローチは、AI システムの設計、開発、デプロイ、評価、使用において適用可能です。プライバシーとサイバーセキュリティのリスクは、より広範な企業リスクマネジメントの検討事項の一部として考慮され、AI リスクを組み込むこともできます。「セキュアでレジリエント」や「プライバシー強化」といった AI のトラストワージ特性に取り組む取り組みの一環として、組織は、NIST サイバーセキュリティ・フレームワーク、NIST プライバシー・フレームワーク、NIST リスクマネジメント・フレームワーク、セキュアソフトウェア開発フレームワークなど（ただしこれらに限定されない）、セキュリティとプライバシーのリスクを低減するための幅広いガイダンスを組織に提供する、利用可能な標準やガイダンスの使用を検討することができます。これらのフレームワークには、AI RMF と共通する特徴がいくつかあります。ほとんどのリスクマネジメントアプローチと同様に、これらのフレームワークは規定的というよりむしろ成果ベースであり、多くの場合、機能、カテゴリー、サブカテゴリーのコアセットを中心に構成されています。AI リスクマネジメントは他の多くの種類のリスクに対処する必要があるため、これらのフレームワークの間には、対処する領域によって大きな違いがあるが、上記のようなフレームワークは、AI RMF の **MAP**、**MEASURE** および **MANAGE** の機能におけるセキュリティおよびプライバシーの検討に役立つ可能性があります。

同時に、この AI RMF が公表される前に利用可能だったガイダンスは、多くの AI システムのリスクを包括的に扱っていません。例えば、既存の枠組みやガイダンスでは、以下のことができません：

- AI システムにおける有害なバイアスの問題を適切に管理する。
- 生成型 AI に関連する困難なリスクに立ち向かう。
- 回避、モデル抽出、メンバシップ推論、可用性、またはその他の機械学習攻撃に関連するセキュリティ上の懸念に包括的に対処する。
- AI システムの複雑な攻撃対象領域、または AI システムによって可能となるその他のセキュリティ侵害を考慮する。
- AI システムが組織のセキュリティ管理外の意思決定のために学習されたり、ある領域で学習された後に別の領域用に「微調整」されたりするような、サードパーティの AI 技術、転移学習、適応外使用に関連するリスクを考慮する。

AI も従来のソフトウェア技術やシステムも、急速な技術革新にさらされています。技術の進歩をモニターし、その進歩を使用するようにデプロイし、トラストワージネスと責任の両方を備えた AI の未来に向けて取り組む必要があります。

付録 C：
AI のリスクマネジメントと人間と AI の相互作用

業務環境で使用する AI システムを設計、開発、またはデプロイする組織は、人間と AI の相互作用の現在の限界を理解することにより、AI のリスクマネジメントを強化することができます。AI RMF は、AI システムを使用、相互作用、マネジメントする際の様々な人間の役割と責任を明確に定義し、区別する機会を提供します。

AI システムが依拠するデータ駆動型アプローチの多くは、個人や社会の観察および意思決定の慣行を測定可能な量に変換または表現しようとするものです。複雑な人間現象を数学的モデルで表現することは、必要なコンテクストを取り除くという代償を払うことになりかねません。このようなコンテクストの喪失は、ひいては AI のリスクマネジメントの取り組みの鍵となる個人や社会へのインパクトを理解することを困難にする可能性があります。

さらなる検討と研究に値する課題には、以下のようなものがあります：

1. **意思決定と AI システムのオーバーサイトにおける人間の役割と責任を明確に定義し、区別する必要がある。**人間と AI の構成は、完全な自律型から完全な手動型まで様々である。AI システムは自律的に意思決定を行うことも、人間の専門家に意思決定を委ねることも、人間の意思決定者が追加的な意見として利用することもできる。AI システムの中には、ビデオ圧縮の改善に使われるモデルのように、ヒューマンオーバーサイトを必要としないものもある。また、ヒューマンオーバーサイトを特に必要とするシステムもある。

2. **AI システムの設計、開発、デプロイ、評価、使用における意思決定は、システムおよび人間の認知バイアスを反映する。**AI アクターは、個人と集団の両方の認知バイアスをプロセスに持ち込む。バイアスは、エンドユーザの意思決定タスクに起因することもあれば、設計やモデリングタスクにおける人間の仮定、期待、意思決定を通じて、AI のライフサイクル全体にわたって持ち込まれることもある。これらのバイアスは、必ずしも常に有害であるとは限らないが、AI システムの不透明性や、その結果として生じる透明性の欠如によって悪化する可能性がある。組織レベルのシステム的バイアスは、AI ライフサイクルを通じて、チームの構成や意思決定プロセスを誰がコントロールするかに影響を与える可能性がある。これらのバイアスは、エンドユーザ、意思決定者、政策立案者による下流の意思決定にも影響を及ぼし、ネガティブインパクトをもたらす可能性がある。

3. **人間と AI の相互作用の結果は様々である。**特定の条件下では、例えば知覚に基づく判断タスクでは、人間と AI の相互作用の AI 部分が人間のバイアスを増幅させ、AI または人間単独よりも偏った判断につながる可能性がある。しかし、人間と AI のチームを編成する際に、このようなばらつきを注意深く考慮すれば、相互補完をもたらし、全体的なパフォーマンスを向上させることができる。

4. **AI システムの情報を人間に提示するのは複雑である。**人間は、AI システムの出力や説明をさまざまな方法で知覚し、そこから意味を導き出すが、これは個人の嗜好、特徴、スキルの違いを反映している。

GOVERN 機能は、Human-AI チームの構成員と AI システムの性能をオーバーサイトする人間の役割と責任を明確にし、定義する機会を組織に提供します。**GOVERN** 機能はまた、組織が意思決定プロセスをより明確にし、系統的バイアスに対抗するためのメカニズムを構築します。

MAP 機能は、AI システムの性能とトラストワージネスの概念に関するオペレータと実務者の習熟のためのプロセスを定義して文書化し、関連する技術標準と認証を定義する機会を提案します。**MAP** 機能のカテゴリーとサブカテゴリーを導入することは、組織が、コンテクストの分析、手続きとシステムの限界の特定、現実世界における AI ベースのシステムのインパクトの調査と検討、AI のライフサイクル全体を通しての意思決定プロセスの評価のための内部能力を向上させるのに役立つ可能性があります。

GOVERN と **MAP** の機能では、学際性、人口統計学的に多様なチームの重要性、インパクトを受ける可能性のある個人やコミュニティからのフィードバックの使用について説明しています。AI RMF で言及されているヒューマンファクターのタスクや活動を行う AI アクターは、設計や開発の実践において、ユーザの意図や、より広範な AI コミュニティの代表者、そして社会的価値観に軸足を置くことで、技術チームを支援することができます。これらのアクターはさらに、システム設計にコンテクスト特有の規範や価値観を取り入れ、AI システムと連携してエンドユーザエクスペリエンスを評価するのに役立ちます。

人間と AI の構成のための AI リスクマネジメントアプローチは、継続的な研究と評価によって強化されます。例えば、人間がどの程度 AI システムの出力に挑戦する権限を与えられ、インセンティブを与えられているかについては、さらなる研究が必要です。配備されたシステムにおいて、人間が AI システムの出力を覆す頻度や根拠に関するデータを収集・分析することは有益であると思われます。

付録 D：
AI RMF の属性

NIST は、フレームワークの作業が最初に始まったときに、AI RMF のいくつかの主要な属性について説明しました。これらの属性はそのままで、AI RMF の開発の指針として使用された。参考としてここに示されています。

AI RMF は以下のことに努めます：

1. リスクベースであること、リソース効率であること、イノベーションを促進すること、自発的である。
2. コンセンサス主導であり、オープンで透明性の高いプロセスを通じて策定され、定期的に更新される。すべての利害関係者が AI RMF の策定に貢献する機会を持つ。
3. 上位管理職、政府関係者、非政府組織のリーダー、AI の専門家でない人々など、幅広い読者が理解できる明確で平易な言葉を使用する。AI RMF は、組織全体、組織間、顧客、そして一般市民に対して、AI リスクを伝えることを可能にするものでなければならない。
4. AI リスクをマネジメントするための共通言語と理解を提供する。AI RMF は、AI リスクの分類法、用語、定義、測定基準、特徴付けを提供する。
5. 使いやすく、リスクマネジメントの他の側面と調和する。フレームワークの使用は、組織の広範なリスクマネジメント戦略およびプロセスの一部として、直感的で容易に適応可能であるべきである。また、AI リスクをマネジメントするための他のアプローチと整合的であるべきである。
6. 幅広い視点、セクター、技術領域にとって有用であること。AI RMF は、どのような AI 技術にも、また、状況特有のユースケースにも普遍的に適用できるものでなければならない。
7. 成果に焦点を当て、非規定的であること。フレームワークは、画一的な要件を規定するのではなく、アウトカムとアプローチのカタログを提供すべきである。
8. AI リスクをマネジメントするための既存の基準、ガイドライン、ベストプラクティス、方法論、ツールを使用し、その認知度を高める。
9. 法律や規制にとらわれない。フレームワークは、適用される国内および国際的な法律や規制体制の下で活動する組織の能力を支援するものでなければならない。
10. リビングドキュメントであること。AI RMF は、技術、理解、AI のトラストワージネスおよび AI の利用に対するアプローチが変化するにつれて、また、一般的な AI リスクマネジメントおよび特にこのフレームワークの実施から利害関係者が学ぶにつれて、容易に更新せざるを得ない。

本書は https://doi.org/10.6028/NIST.AI.100-1 から無料で入手できます。