日本科学技術連盟
スペシャルセッション

「安全、安心で信頼できるAIの実現に向けて
～世界の最新AIセーフティ動向～」

2025-04-22

AIセーフティ・インスティテュート（AISI）技術統括
北村　弘

AISI Japan AI Safety Institute

自己紹介　北村　弘（Hiromu Kitamura）

AISI Japan AI Safety Institute

【所属部門】AIセーフティ・インスティテュート（AISI）技術統括｜
独立行政法人情報処理推進機構（IPA）デジタル基盤センター　デジタルエンジニアリング部エキスパート

【現在、人工知能（Artificial Intelligence）を主軸に活動】

- **AISI国際ネットワーク　Track3（Risk Assessment of Advanced AI Systems）日本代表**
- **東京大学未来ビジョン研究センター客員研究員｜国立研究開発法人産業技術総合研究所　客員研究員**
- **東洋大学総合情報学部 AI 監査研究プロジェクトおよび東洋大学 AI 監査研究会　委員**
- 国立研究開発法人産業技術総合研究所　機械学習品質マネジメントガイドライン詳細検討タスクフォース メンバー
- AI事業者ガイドライン検討会　経産省合同事務局｜総務省オブザーバー
- CDLE（Community of Deep Learning Evangelists）AIリーガルグループ　リーダー｜　AI法研究会メンバー
- 元ISO/IEC JTC1/SC42（人工知能）国内専門委員会　エキスパート
- 2023年度NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）特別講座「AI品質マネジメント講座」講師
- 日科技連　SQiP(Software Quality Profession)ソフトウェア品質保証プロフェッショナルの会　メンバー

【共著、寄稿等】

『Quality World magazine～ Keeping it real How standards in artificial intelligence can prepare quality professionals for the future～』（Chartered Quality Institute　2022年9月）、『AIビジネス大全』（プレジデント社　2022年12月）、『Advancing AI Audits for Enhanced AI Governance』（arXiv　2023年11月）、『デジタルエシックス』（ダイヤモンド社　2024年2月）、『AI事業者ガイドライン パネルディスカッション】（商事法務　NBL1270(2024.7.15)号）、『ソフトウェア品質保証の極意～経験者が語る、組織を強く進化させる勘所～』（オーム社　2024年9月）、『AIリスクアセスメントガイドブック』編著者：（一社）日本品質管理学会　AI 品質アジャイルガバナンス研究会 編（日科技連出版社　2024年9月）等

【受賞】

一般社団法人情報通信ネットワーク産業協会（CIAJ) 功労者表彰受賞　（ISO9001:2015規格解釈WEB教育開発）（2017年12月）

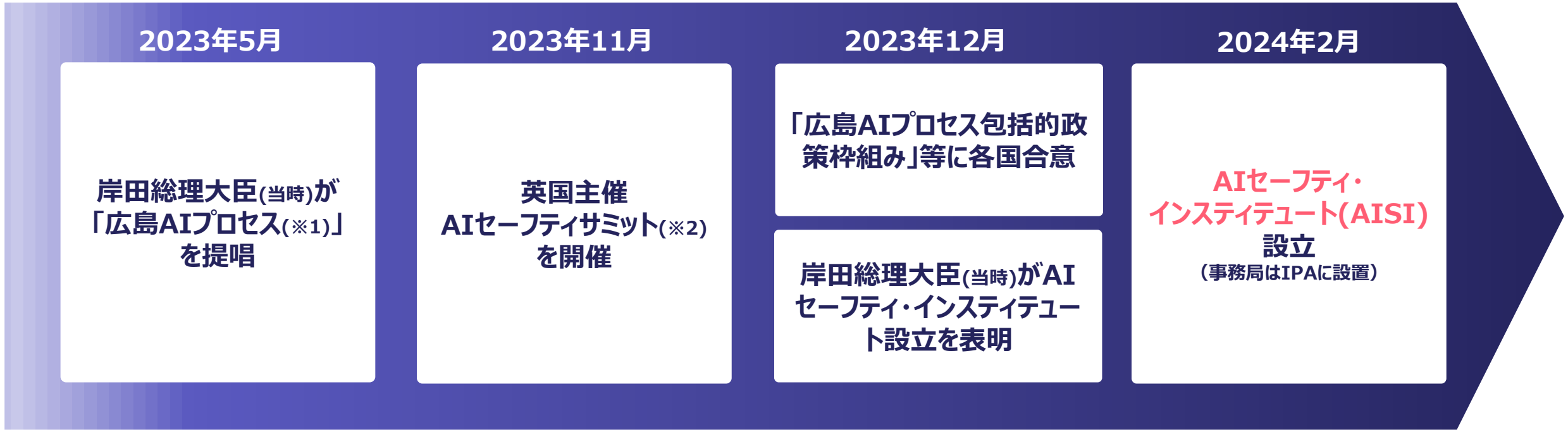
前半：AISIの紹介
後半：世界の最新AIセーフティ動向の紹介

AISIの紹介

日本におけるAISIの設立

AISI Japan AI Safety Institute

広島AIプロセスでの議論や**AIセーフティサミット**を経て
日本でもAIセーフティ・インスティテュート（AISI）を設立（2024年2月）



※1 成果文書｜広島AIプロセス
※2 AI Safety Summit 2023 - GOV.UK

AISIの役割とスコープ

AISI Japan AI Safety Institute

AIの安全安心な活用が促進されるよう
官民の取組を支援することがAISIの役割

役割

- 主に3つの役割を担う。

政府への支援

- AIセーフティに関する調査、評価手法の検討や基準の作成等

日本におけるAIセーフティのハブ

- 産学における関連取組の最新情報の集約
- 関係企業・団体間の連携促進
- 他国のAIセーフティ関係機関との連携

関連の研究機関との連携実施

- AISIは自ら研究開発を行う組織ではない

AIの開発や利用をする者が**AIのリスクを正しく認識**できる仕組みの構築 ＋ ガバナンス確保などの必要となる対策を**ライフサイクル全体で実行**できる仕組みの構築 ⇔ **国内・国際的な関係機関**

▼

イノベーションの促進と**ライフサイクルにわたるリスクの緩和**を**両立**する枠組みを実現

スコープ

- AIによる以下の事象や検討事項の中で、諸外国や国内の動向も見ながら柔軟にスコープを設定し取組を進めていく。

社会への影響	ガバナンス	AIシステム	コンテンツ	データ

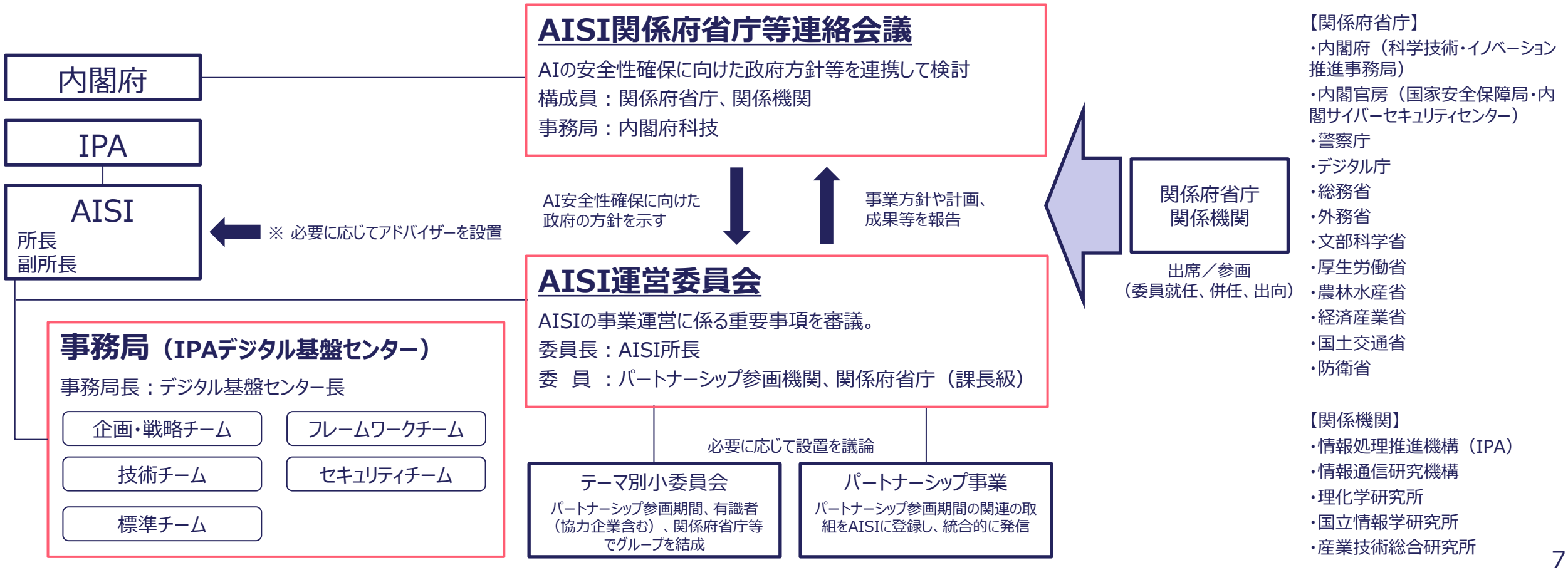
AISIの推進体制

AISI Japan AI Safety Institute

AISIは13府省庁、5関係機関で構成される**政府横断の組織**

＊内閣府を事務局とする「AISI関係府省庁等連絡会議」で政府方針等を検討

＊AISIには「AISI運営委員会」と「事務局」を設置



成果物の概要

AISI Japan AI Safety Institute

AI事業者ガイドラインを軸に、
技術的なレビューから人材育成まで、幅広く取り組む

クロスウォーク
国際的な相互運用性のため

AI事業者ガイドライン

活動マップ
全体像と優先順位付け

評価観点ガイド
評価

レッドチーミング手法ガイド
レッドチーミング

データ品質マネジメントガイドブック
AIに適格なデータを提供するため

多言語/多文化
多国間での問題

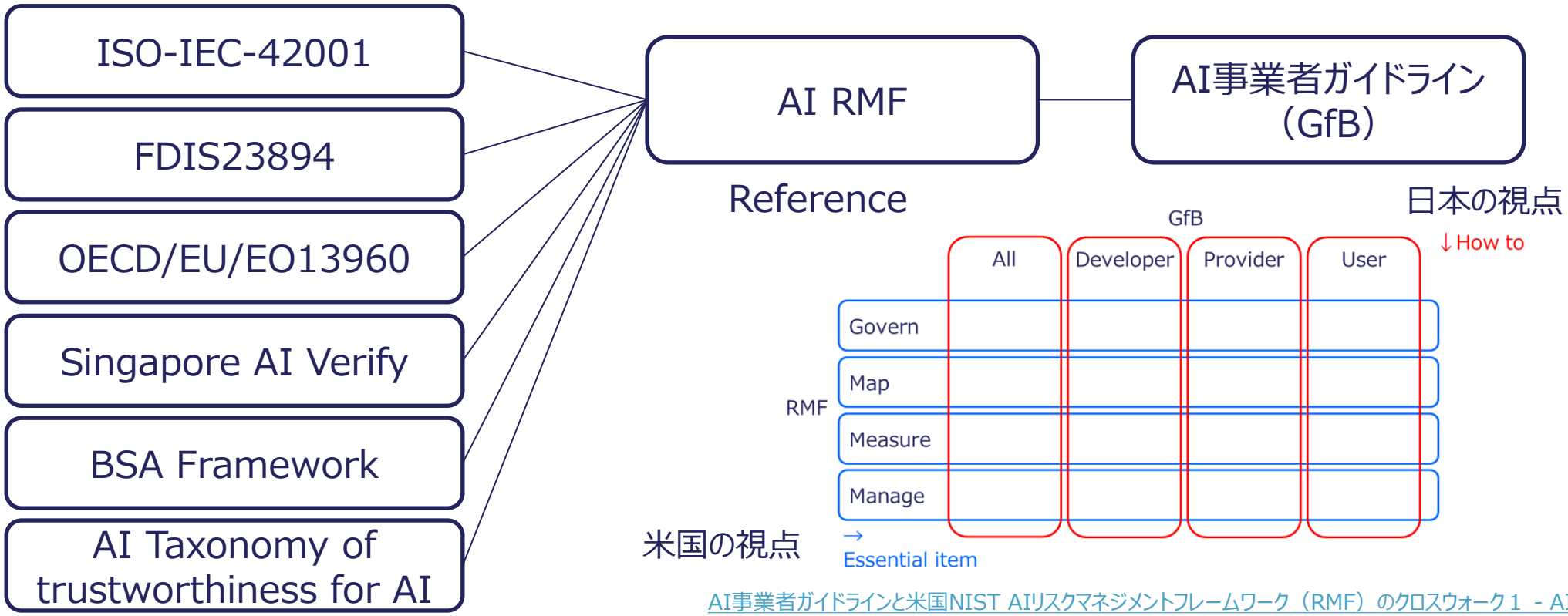
セキュリティレポート
セキュリティに関する知識

デジタルスキル標準
人材育成

日米クロスウォークの概要

米国NISTのAI Risk Management Framework(RMF)と日本のAI事業者ガイドライン(Guidelines for Business; GfB)の相互関係を確認

- 米国のAI RMFをリファレンスに各国ガイドライン等との確認も可能。



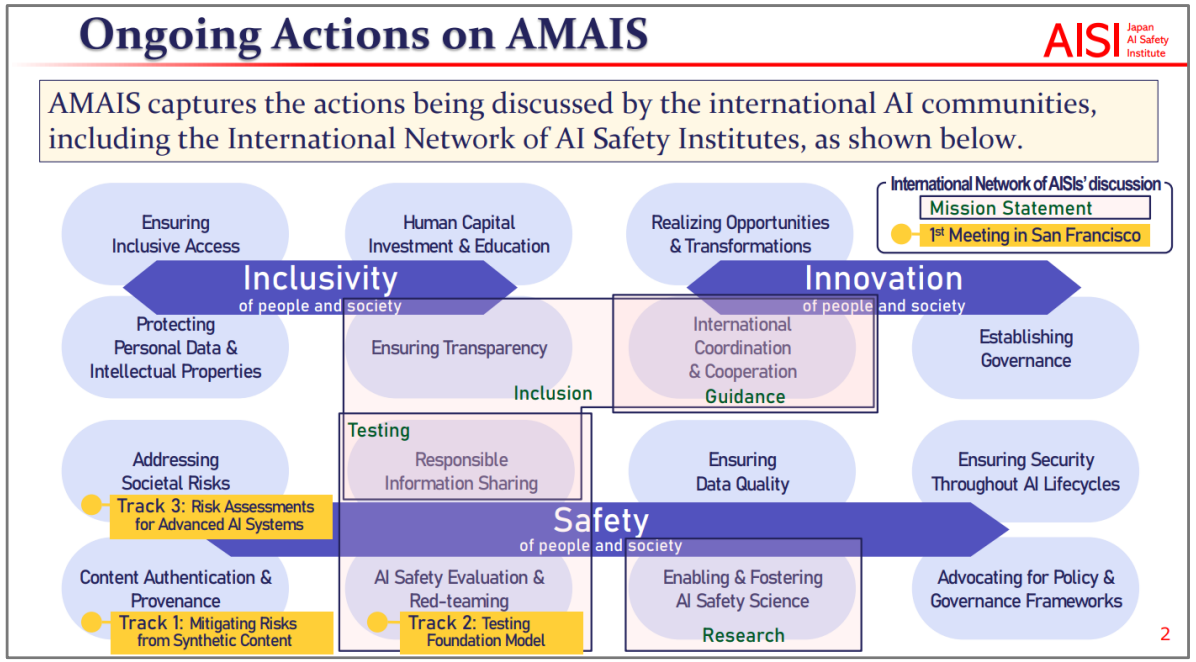
AI事業者ガイドラインと米国NIST AIリスクマネジメントフレームワーク（RMF）のクロスウォーク１ - AISI Japan
AI事業者ガイドラインと米国NIST AIリスクマネジメントフレームワーク（RMF）のクロスウォーク２ - AISI Japan

AIセーフティに関する活動マップ（AMAIS）の概要

AIの安全性に関する活動が急速に変化・進化する中、
見落とされがちな部分や活動間の相関関係を全体像として可視化

- AISIは、**ディスカッションペーパー**として「AMAIS：AIの安全性に関する活動マップ」を公開。

- AISIは、**主要文献のベンチマーク**に基づき、包括的なアクティビティマップと関連用語を開発している。
- この日本主導の取り組みは、AIの安全性に関する国際的な協力体制の基盤をさらに強化し、持続可能で信頼性の高いAI社会の実現に貢献することが期待されている。



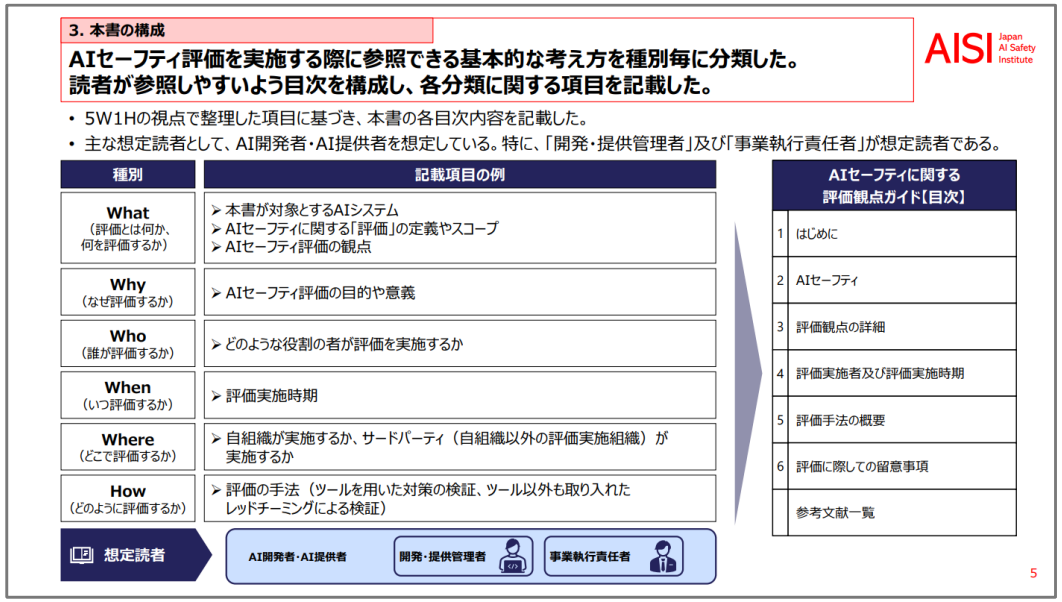
AIセーフティに関する活動マップ（AMAIS）AMAIS: Activity Map on AI Safety - AISI Japan

評価観点ガイドの概要

事業者がAIを開発・提供する際の参考として、
AIシステムの安全性を評価する際の基本的な考え方を示したもの

- 具体的には、以下の事項等が記載されている。
  - **安全性評価で想定するリスクや評価項目**
  - **評価の実施者や実施時期**
  - **評価手法の概要**

- このガイドは、安全・安心で信頼できるAIの実現に向けての第一歩であり、今後のAI開発・提供における安全性の維持・向上に資することを期待している。

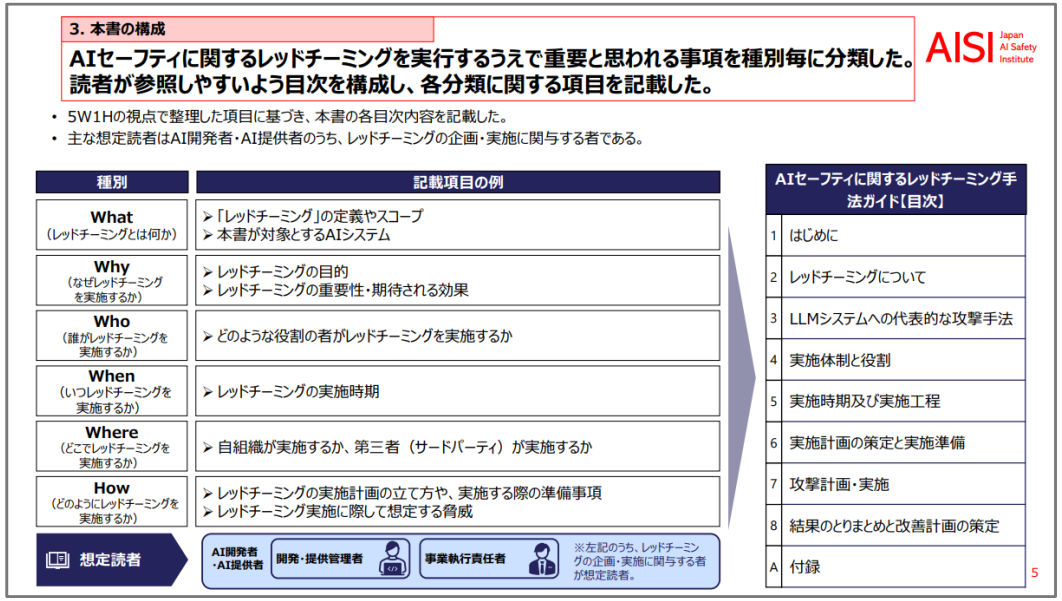


AIセーフティに関する評価観点ガイド - AISI Japan

レッドチーミング手法ガイドの概要

事業者が開発・提供する際の参考として、**AIシステムの安全性を評価する手法**の１つであるレッドチーミング手法について基本的な留意事項を示したもの

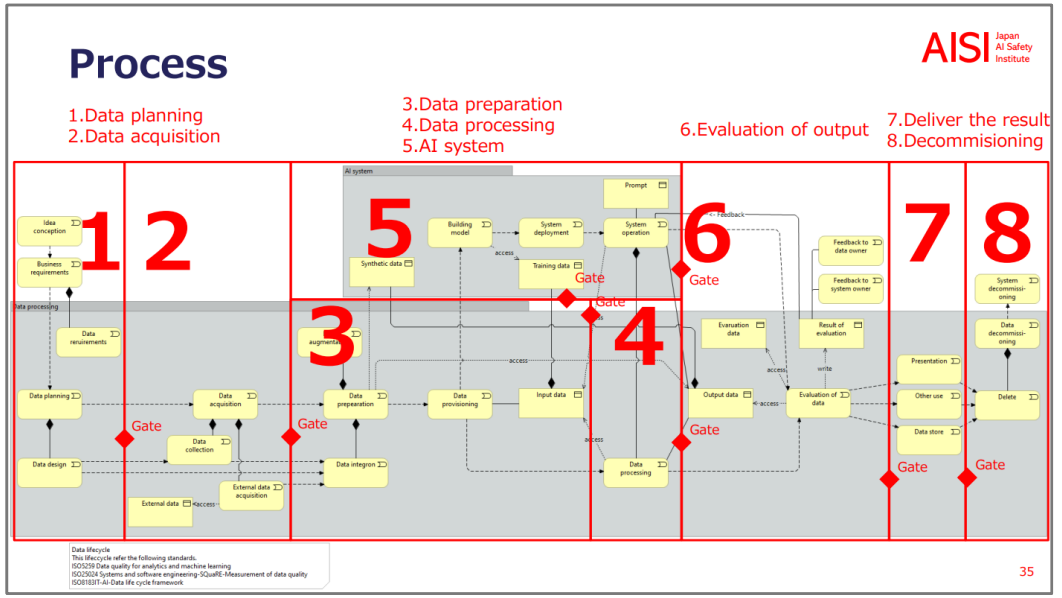
- 具体的には、安全性評価の実施体制、時期、計画、実施方法、改善計画の策定等にあたっての留意点が示されている。
- このガイドは、安全・安心で信頼できるAIの実現に向けての第一歩であり、今後のAI開発・提供における安全性の維持・向上に資することを期待している。



データ品質マネジメントガイドブックの概要

データとAIの価値を最大化するために必要な
データ品質を持続的に確保するため、何をすべきか整理

- データ品質は、AIの卓越性の基礎であり、信頼できるAIの実現に寄与する。AI社会を適切に実現し、データ駆動型社会へと導くため、本ガイドに整理。
- 本ガイドは、英語版が正式版であり、2025年3月に日本語訳サマリが公開。

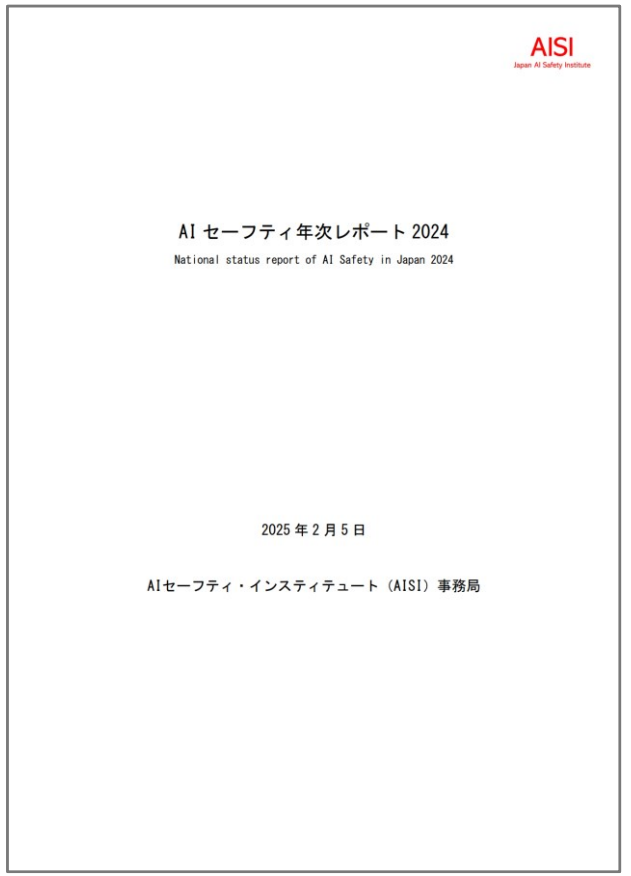


AIセーフティ年次レポート2024の概要

AISIの活動状況を「AIセーフティ年次レポート2024」としてまとめた。

- 「AIセーフティ年次レポート2024」とともに、関連するレポート等についても、年次レポートを補完する参考資料「AIセーフティ　ファクトシート2024」として取りまとめた。

- 本稿においては、AIの急速な進展に対応するための、AISIと国内外の関係機関や企業等との連携など、我々の今後の取り組みやその狙いについても記載している。



AIセーフティ年次レポート2024 National Status Report on AI Safety in Japan 2024 - AISI Japan

今後の取り組み

AISI Japan AI Safety Institute

民間企業との協力、調査研究の拡大

- **民間企業との協力関係**
  - 「**事業実証ワーキンググループ**（ＷＧ）」の設置を検討
    - 第３回AISI運営委員会にて公表
- **調査研究の拡大**
  - 評価観点ガイドやレッドチーミング手法ガイドの改定に向けた調査
- **対象をマルチモーダル基盤モデルに拡大**
  - 事業実証ＷＧの取組に向けた調査
    - **AIセーフティの評価環境の一部を先行して構築**し、WG活動の環境を整備
  - AIセーフティの自動評価に関する調査
    - **AIセーフティ評価を広く一般化**するため、評価の自動化/省力化を検討

AISI Japan AI Safety Institute

世界の最新AIセーフティ動向

2025年3月世界ツアーでの最新情報を共有

AISI Japan AI Safety Institute

1. Measuring AI in the World
2. AI Standards Hub Global Summit

1.主催： **米国DeepMind x サンタフェ研究所**
日時/場所：3月12日-14日@ニューメキシコ州サンタフェ

目的：行動科学、社会科学、コンピュータ科学、そして実践と政策の世界からの専門知識を結集し、関連分野からのAI評価へのアプローチを前進させ、標準化を目指す。また、AIがどのような点で測定のユニークな課題を提示しているのかを明らかにし、このようなシステムのオープンエンド性や複雑さにもかかわらず、どのように評価を扱いやすくすることができるのかを探求する。

https://santafe.edu/events/measuring-ai-in-the-world

2.主催： **英国アラン・チューリング研究所**
日時/場所：3月17日-18日@ロンドン

目的：経済協力開発機構（OECD）、国連人権高等弁務官事務所、パートナーシップ・オン・AIとのパートナーシップにより開催され、国際的なAIエコシステム全体から多様な関係者を集め、AI標準化の現状を検証し、AIガバナンスの枠組みや世界的な規制の台頭との関連で標準の進化する役割を探る。

https://aistandardshub.org/global-summit/

「自社で撮影」

注：世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

【AIリスク(機会と脅威)評価】目的と意味

AISI Japan AI Safety Institute

AIリスク評価の「目的」と「価値」が、AIの未来をデザインする

- AIリスク評価に関する議論は、単なる「どのように評価するか（How）」という技術的な議論だけでなく、「何を評価すべきか（What）」や「なぜ評価するのか（Why）」といった根本的な視点が重要。

- AIリスク評価は、単にパフォーマンスを測定するだけでなく、その評価がどのようにAIの発展を方向づけるかを考慮する必要がある。

- AIリスク評価の目標が、ユーザーの信頼構築であるべきか、社会的影響を測定するべきか、技術進化の指標とするべきかという点が重要な議論ポイント。

注：世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

【AIリスク(機会と脅威)評価】実世界で行う際の課題

AISI Japan AI Safety Institute

AIリスク評価の鍵は、“現実”を見据えた多面的アプローチ

- AIのリスク評価を実際の利用環境で行うには、技術的・社会技術的・法的な課題がある。
- マルチステークホルダーとの連携が必要であり、データの共有や法律の整備が不可欠。
- 「監査（audit）」「ベンチマーク（benchmarking）」「レッドチーミング（red teaming）」などのリスク評価手法を使い分けるべき。
- 中央集権型・分散型のリスク評価の違い（例：ブラウザ拡張機能を使ってAIの動作を記録する vs. 政府主導の監査）を議論。

注：世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

【技術課題】AIのリスク評価基準とその限界

AISI Japan AI Safety Institute

今のAIリスク評価は、エンドユースにも社会にも“十分に”届いていない

- 現在のリスク評価手法（レッドチーミングを含む）は、モデルのパフォーマンスに焦点を当てたものが多く、実際の製品・サービスレベルや社会的な影響を十分に測れていない。

- AIシステムの性能評価と、実際に使用する環境での影響評価に大きなギャップがある。
- リスク評価基準が確立されておらず、実際のユースケースと適合しない場合が多い。

- 既存のリスク評価手法（例：ベンチマークやリーダーボード）が、必ずしもAIの社会的な影響を測るのに適しているわけではない。

注：世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

【技術課題】AIリスク評価における代理指標の問題

AISI Japan AI Safety Institute

AIを測るには“ものさし”がいる――でも、その選び方が未来を左右する

- 経済や都市開発と同様に、AIの影響を測るための代理指標が必要。
- しかし、AIは制御可能なシステムであるため、代理指標を最適化しすぎると「Goodhart’s Law」（指標が目標になると機能しなくなる）が発生する可能性がある。
- 「ジャグド・インテリジェンス（Jagged Intelligence）」の問題（AIが特定の分野で極端に優れ、他の分野で極端に劣る）が代理指標の妥当性を難しくする。

【QA】

Q: AIの影響を代理指標で測ることは可能か？

A: 代理指標の最適化によってバイアスが生じるため、慎重に設計する必要がある。

注: 世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

【技術課題】AIの標準化とAI保証（AI Assurance）

AISI Japan AI Safety Institute

「つながり」と「一貫性」が、AIの信頼をつくる

- AI保証の必要性:
  - AIの社会実装が進むにつれ、セーフティと透明性の確保が求められる。
  - 企業のコンプライアンスを支援する枠組みが必要。
- 現行のAI標準化:
  - ISO/IEC JTC 1/SC 42、IEEE P7000シリーズ、NIST　AIリスクマネジメントフレームワーク他。
  - 既存の標準と新たなAI保証フレームワークの統合の必要性。
- 規制と自主ガイドラインのバランス:
  - 各国政府のAI規制の方向性（例: EU AI Act、米国AI倫理ガイドライン）。
  - 企業による自主規制の促進と標準化の連携。
- 技術的評価:
  - モデルの透明性と説明可能性（XAI）の向上。
  - AIのバイアス評価と公平性の確保。
  - AI監査の自動化ツールとその限界。

注: 世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

【社会技術課題】AIと社会の相互作用のモニタリング

AISI Japan AI Safety Institute

AIの影響、“見える化”してこそ社会に届く

- AIが実社会でどのように影響を与えるかをモニタリング（技術・社会技術的変曲点、ゾーン思考の閾値、リスクパターン）することが重要。
- 公衆衛生のモニタリングモデルを参考にする（例：プライバシーを守りながらAIの影響をモニタリングする）。
- AIの自己評価（self-test kits）を導入し、地方自治体や小規模コミュニティでもモニタリングできる仕組みを整備する。

注：世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

【社会技術課題】AIの社会・経済への影響

AISI Japan AI Safety Institute

生産性だけじゃ見えない、AIが変える社会のかたちも評価を

- AIの経済的影響の評価は、単に生産性向上の指標を用いるだけでは不十分であり、長期的な社会システムへの影響も考慮すべき。
- AIは仕事を単純に自動化するだけでなく、新しい業務プロセスを生み出すことで、労働市場における負担のシフトを引き起こす。
- AIの導入が経済全体の生産性向上にどのように寄与しているのかを評価する枠組みがまだ確立されていない。

注：世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

【社会技術課題】AIと社会制度の関係

AISI Japan AI Safety Institute

AIと社会、かみ合ってこそ前に進む。“社会制度”との適合性のチェックが鍵

- AIが社会制度にどのように適応し、制度がAIをどう受け入れるかを考える必要がある。
- AIが社会制度にどのような影響を与えるかを評価する基準が必要。
- AIを導入する際に、既存の法律や倫理規範と適合するかを評価する必要がある。

注：世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

【社会技術課題】AIリスク評価の科学的アプローチ

AISI Japan AI Safety Institute

AIリスク評価は、理論と実証で裏打ちされた科学の土台から

- AIリスク評価の科学的な方法論を確立するためには、仮説を立て、それを実験的に検証することが重要。
- 現在のAIリスクI評価は、基準が曖昧で信頼性に欠ける場合が多い。
- AIの影響を予測するための理論が不足しているため、社会科学との連携が必要。

注：世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

【社会技術課題】AIリスク評価の科学的方法論

AISI Japan AI Safety Institute

AIリスクを測るなら、“科学という土台”なしに語れない

- リスク評価基準をより科学的に確立する必要がある。
- 仮説を明確にし、検証可能な方法でテストを行うべき。
- 単なるベンチマークではなく、飛行機の風洞実験のように、理論と実験の組み合わせが必要。

【QA】
Q: AI評価に経済学の指標（GDP、失業率など）のような標準的な指標を持ち込めるか？
A: 経済指標のように、AIの影響を測る統一指標が必要。ただし、どの指標を選ぶかが難しい。
Q: AIの評価はどのように進化すべきか？
A: AIを飛行機開発になぞらえ、最初は試行錯誤で進むが、最終的には理論と実験の両面からの評価が必要。

注: 世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

【基盤モデル】性能評価

AISI Japan AI Safety Institute

基盤モデル評価には、“多様・現実・再現”の新たなものさしが必須

- **評価指標の多様性:**
  - 既存の評価指標（BLEUスコア、ROUGEスコア、F1スコア）では限界がある。
  - 人間のフィードバックを活用した評価の重要性。
  - 公正性（Bias）や倫理的影響の評価も必要。
- **ベンチマークの課題:**
  - 既存のベンチマークは欧米のデータに偏っており、グローバルな多様性を反映できていない。
  - より包括的なベンチマークデータセットの開発が求められる。
- **AI評価のための新技術:**
  - 自動評価ツールの開発とその限界。
  - AIシステムのライフサイクルを通じた継続的（超長期）評価の重要性。
  - 実際の使用環境での評価とモデルの適応性の測定。
- **再現性と透明性の確保:**
  - モデルの評価結果が一貫して再現可能（関係性が複雑）であることが重要。
  - 評価方法論を標準化し、異なる研究機関でも同じ結果が得られるようにする。

【BLEUスコア（Bilingual Evaluation Understudy）】
主に機械翻訳の評価に使われる人間の翻訳とAIの翻訳を比べて、どれくらい「同じような単語・フレーズ」を使っているかを見る。

【ROUGEスコア（Recall-Oriented Understudy for Gisting Evaluation）】
主に要約の評価に使われる。どれだけ「参照（正解）の要約」と同じ単語やフレーズが含まれているかを見る。再現率（Recall）が重視されることが多い。

【F1スコア】
正解をどれだけ見逃さずに（再現率）、かつ間違いなく（適合率）見つけられたかのバランスを取るスコア。主に質問応答や情報抽出で使われまる。

出典：ChatGPT

注: 世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

【基盤モデル】セーフティ

AISI Japan AI Safety Institute

基盤モデルのセーフティには、多層的・国際的な連携が不可欠

- **リスク評価の現状:**
  - モデルの誤用リスク（例: フェイクニュース生成、詐欺行為の支援）。
  - AIが意図しないバイアスを持つ可能性とその影響。
  - セーフティを保証するためのテスト手法。
- **モデルセーフティの向上策:**
  - レッドチーミング（Red Teaming）による悪用リスクの特定と対策。
  - フェイルセーフメカニズムの導入。
  - アクセス制御と使用ポリシーの策定。
- **業界標準と国際協力:**
  - 既存のセーフティ基準（ISO、IEEE、NIST他）との整合性。
  - 国際機関や規制当局との協力の必要性。
  - 企業間でのリスク共有とセーフティ対策のベストプラクティスの交換（リスクシナリオやユースケース）。
- **ユーザー教育と透明性:**
  - 一般ユーザーへのAIリスク認識向上。
  - 開発者向けの倫理ガイドラインとトレーニングプログラム。
  - オープンなAI監査プロセスの確立。
- **法規制の役割:**
  - EU AI Actや米国のAIセーフティ基準との整合性。
  - 各国の政策と標準化の調整。
  - 法的枠組みの強化と企業への適用方法。

注: 世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

【基盤モデル】セーフティ評価とその取り組み

AISI Japan AI Safety Institute

AIセーフティの鍵は、動的評価とリスク制御の仕組みづくり

- **セーフティ評価の方法論:**
  - 既存の評価指標と新たな基準の開発。
  - AIモデルの「ブラックボックス化」に伴う評価の難しさ。
  - 動的評価（リアルタイムモニタリング）の必要性。

- **リスク低減策:**
  - モデルの事前トレーニング時における倫理的フィルターの導入。
  - 機密情報や悪意のあるコンテンツの生成を防ぐためのセーフティ制御。
  - フェイルセーフメカニズム（Fail-Safe Mechanisms）の開発。

- **ガバナンスと規制:**
  - AIセーフティ基準の国際整合性の確保（ISO、IEEE、NISTなどの協力）。
  - 企業と政府間でのリスク評価データの共有。
  - 法的枠組みとコンプライアンスの強化。

- **セーフティの透明性:**
  - 開発者向けのガイドラインとエンドユーザー向けの説明責任。
  - 公開ベンチマークと第三者監査の活用。
  - AI倫理委員会の設立とその影響力の拡大。

注: 世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

【基盤モデル】環境影響の測定とマネジメント

AISI Japan AI Safety Institute

セーフティと環境配慮を両立するスマートな設計が求められる

- **環境負荷の現状:**
  - AIモデルのトレーニングに必要な計算資源の大幅な増加。
  - データセンターの冷却システムによる水資源の消費問題。
  - レアメタルの採掘による環境破壊リスク。
- **環境影響の測定:**
  - カーボンフットプリントの定量化手法。
  - AIワークロードのエネルギー効率向上に向けた評価基準。
  - 各クラウドプロバイダーが提供するサステナビリティ指標の活用。
  - 測定単位が課題（理想と現実の乖離）
- **環境負荷の軽減策:**
  - より効率的なハードウェア（TPU、GPU）の開発と活用。
  - 再生可能エネルギーを活用したデータセンターの運用。
  - 分散型コンピューティングを活用したエネルギー効率の向上。
- **規制と業界の取り組み:**
  - 各国の環境政策とAI開発の整合性。
  - 企業による持続可能なAI開発のためのガイドライン策定。
  - ISOやIEEE他による環境負荷評価の標準化の必要性。

注: 世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

【基盤モデル】オープン基盤モデルのガバナンス

AISI Japan AI Safety Institute

オープン基盤は、透明性・リスク対策・ルール整備が不可欠

- **オープンソース基盤モデルの利点:**
  - 研究者や中小企業にとってのアクセスのしやすさ。
  - 透明性の向上と学術研究の促進。
  - グローバルな協力を通じたイノベーションの加速。

- **リスクと課題:**
  - モデルの悪用リスク（例: フェイクニュース生成、ディープフェイクの拡散）。
  - ライセンスマネジメントの複雑さと知的財産権の問題。
  - モデルのバイアスと公平性の確保。

- **ガバナンスのフレームワーク:**
  - オープンソースモデルの適正使用に向けた倫理ガイドラインの策定。
  - 企業と研究機関による共同監査の導入。
  - モデルの透明性を確保するための「オープンモデル審査制度」の構築。

- **規制の役割:**
  - 各国政府によるオープンモデルのリスク評価基準の確立。
  - 既存のAI規制との整合性（例: EU AI Act との関係）。
  - 自主規制と法的規制のバランスの取り方。

注: 世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

【汎用AI（General Purpose AI）】評価と課題

AISI Japan AI Safety Institute

汎用AIの評価は、その性質と位置づけに応じた新たな価値基準が必要

- 汎用AIの評価は非常に難しく、特定のタスクにおける評価指標が必ずしも一般化できるわけではない。
- AIを単なる「ツール」として見るのか、「人間の知能に近づくもの」として扱うのかによって、評価方法が変わる。
- AIが人間よりも優れたパフォーマンスを発揮した場合、どのようにその価値を測るのかが不明確。

注：世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

AIリスク評価の方向性と今後の課題

AISI Japan AI Safety Institute

AI評価には、社会的意義と変化に応じた柔軟な枠組みが求められる

- AI評価を単なる技術的な議論にとどめるのではなく、社会的・経済的な影響まで広げるべき。
- AIが「社会に何をもたらすのか？」という問いに対して、社会学観点から理論的な枠組みを構築することが必要。
- AIの進化が速いため、評価基準も柔軟に適応する必要がある。

注：世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

未来のAIリスク評価基準の再構築

AISI Japan AI Safety Institute

AI評価は「後追い」傾向から脱し、未来を見据えた指標と長期視点が必要

- AIのリスク評価が過去の課題に対して常に「後追い」になりやすい、未来のAIに対する評価基準を先に考えるべきではないかという指摘があった。
- AIシステムが社会にどのように適応し、どのような影響を与えるのかを長期・超長期的に分析する方法を開発する必要がある。
- AIを評価する際に、現行の技術基準だけでなく、進化するテクノロジーに対応した新しい評価指標を導入することが求められる。

注：世界ツアーで収集したマルチステークホルダーの最新の考え方の一例を紹介するもので、AISIの公式な見解ではありません。

AIリスク評価の重要な視点

AISI Japan AI Safety Institute

AIリスク評価の前提としての大切な心得

- AIリスク（機会と脅威）の状況は常にアップデートされる
- AIガバナンスの目的は、リスク（脅威）をゼロにすることではない
- AIリスクは提供する側・開発する側だけでなく、あらゆる組織、個人がAIリスクと対面する
- AIガバナンスはグローバル視点で考える必要がある

最後に！

AISI Japan AI Safety Institute

AIの不確実性に備えるには
基盤となる品質マネジメントが不可欠

AISI

Japan AI Safety Institute