第3回AI品質マネジメントシンポジウムご説明資料

AIの戦略を支えるAIセーフティ
～ AIの評価観点とレッドチーミングの世界の動きから ～

AIセーフティ・インスティテュート

2024年11月19日

AISI Japan AI Safety Institute

本日の内容

AISI Japan AI Safety Institute

- AIセーフティ・インスティテュート（AISI)のご紹介と世界の動き
- AIセーフティに関するガイドのご紹介
  - AIセーフティに関する評価観点ガイドの概要について
  - AIセーフティに関するレッドチーミング手法ガイドの概要について
- AISIの今後の取組予定

本日の内容

AISI Japan AI Safety Institute

- **AIセーフティ・インスティテュート（AISI)のご紹介と世界の動き**
- **AIセーフティに関するガイドのご紹介**
  - **AIセーフティに関する評価観点ガイドの概要について**
  - **AIセーフティに関するレッドチーミング手法ガイドの概要について**
- **AISIの今後の取組予定**

AIの戦略を支えるAIセーフティ
～AI Safety Instituteのご紹介と世界の動き～

※AISIは、エイシーと読みます

AISI Japan AI Safety Institute

統合イノベーション戦略における３つの強化方策

AISI Japan AI Safety Institute

（１）重要技術に関する統合的な戦略

①コア技術の開発、他の戦略分野との技術の融合による研究開発（産学官の連携、AI・ロボティクス・IoT等による研究開発推進等）

②国内産業基盤の確立、スタートアップ等によるイノベーション促進（ユースケースの早期創出、拠点・ハブ機能の強化等）

③産学官を挙げた人材の育成・確保（産業化を担う人材、市場開拓を担う人材、研究開発を担う人材の育成・確保等）

（２）グローバルな視点での連携強化

①重要技術等に関する国際的なルールメイキングの主導・参画（開発・利用の促進、安全性確保、プレゼンスの確保等）

②科学技術・イノベーション政策と経済安全保障政策との連携強化（国際協力・国際連携を含めた戦略的な研究開発、技術流出防止等）

③グローバルな視点でのリソースの積極活用、戦略的な協働（国際頭脳循環の拠点形成、国際科学トップサークルへの参画等）

（３）AI分野の競争力強化と安全・安心の確保

①AIのイノベーションとAIによるイノベーションの加速（研究開発力の強化、AI利活用の推進、インフラの高度化等）

②AIの安全・安心の確保（ガバナンス、安全性の検討、偽・誤情報への対策、知財等）

③国際的な連携・協調の推進（広島AIプロセスの成果を踏まえた国際連携等）

（３）ＡＩ分野の競争力強化と安全・安心の確保

AISI Japan AI Safety Institute

- 生成AIはインターネットにも匹敵する技術革新とされ、社会経済システムに大きな変革をもたらす一方で、偽・誤情報の流布や犯罪の巧妙化など様々なリスクも指摘され、安全・安心の確保が求められる。
- 米国企業等の高性能・大規模な汎用基盤モデルが先行する中、我が国もそれに追随すべく計算資源の整備や大規模モデルの開発が進んでおり、また、小規模・高性能なモデルや複数モデルの組合せの開発など、新たな研究も進んでいる。
- AIはあらゆる分野で利用され、AIの開発や利活用等のイノベーションが社会課題の解決や我が国の競争力に直結する可能性がある。我が国においては、生成AIを含むAIの様々なリスクを抑え、安全・安心な環境を確保しつつ、イノベーションを加速する好循環の形成を図っていく。加えて、我が国が主導する広島AIプロセス等を通じて、今後も国際的にリーダーシップを発揮していく。

① ＡＩのイノベーションとＡＩによるイノベーションの加速
- 研究開発力の強化（データ整備含む）
- ＡＩ利活用の推進
- インフラの高度化
- 人材の育成・確保

② ＡＩの安全・安心の確保
- 自発的ガバナンスと制度の検討
- ＡＩの安全性の検討
- 偽・誤情報への対策
- 知的財産権等

③ 国際的な連携・協調の推進

日本におけるAISIの設立

AISI Japan AI Safety Institute

- 2023年5月
  - 岸田総理大臣が「広島AIプロセス（※）」を提唱
    ※G7広島サミットで提唱された生成AIに関する国際的なルールの検討を行うためのプロセス
- 2023年10月
  - 広島AIプロセス「国際指針」及び「国際行動規範」（※）に合意
    ※生成AIを含む高度なAIシステムに関する国際的な指針と行動規範
- 2023年11月
  - 英国主催AIセーフティサミットを開催
- 2023年12月
  - 「広島AIプロセス包括的政策枠組み」等に合意
  - 岸田総理大臣がAIセーフティ・インスティテュート設立を表明
- 2024年2月14日
  - IPA（情報処理推進機構）にAIセーフティ・インスティテュート（AISI）を設立

所長　村上　明子

出典：
広島AIプロセス＜https://www.soumu.go.jp/hiroshimaaiprocess/documents.html＞
AI Safety Summit 2023＜https://www.gov.uk/government/topical-events/ai-safety-summit-2023＞
AI戦略会議＜https://www.kantei.go.jp/jp/101_kishida/actions/202312/21ai.html＞
AIセーフティ・インスティテュート＜https://aisi.go.jp/＞

AISIの役割とスコープ

AISI Japan AI Safety Institute

役割

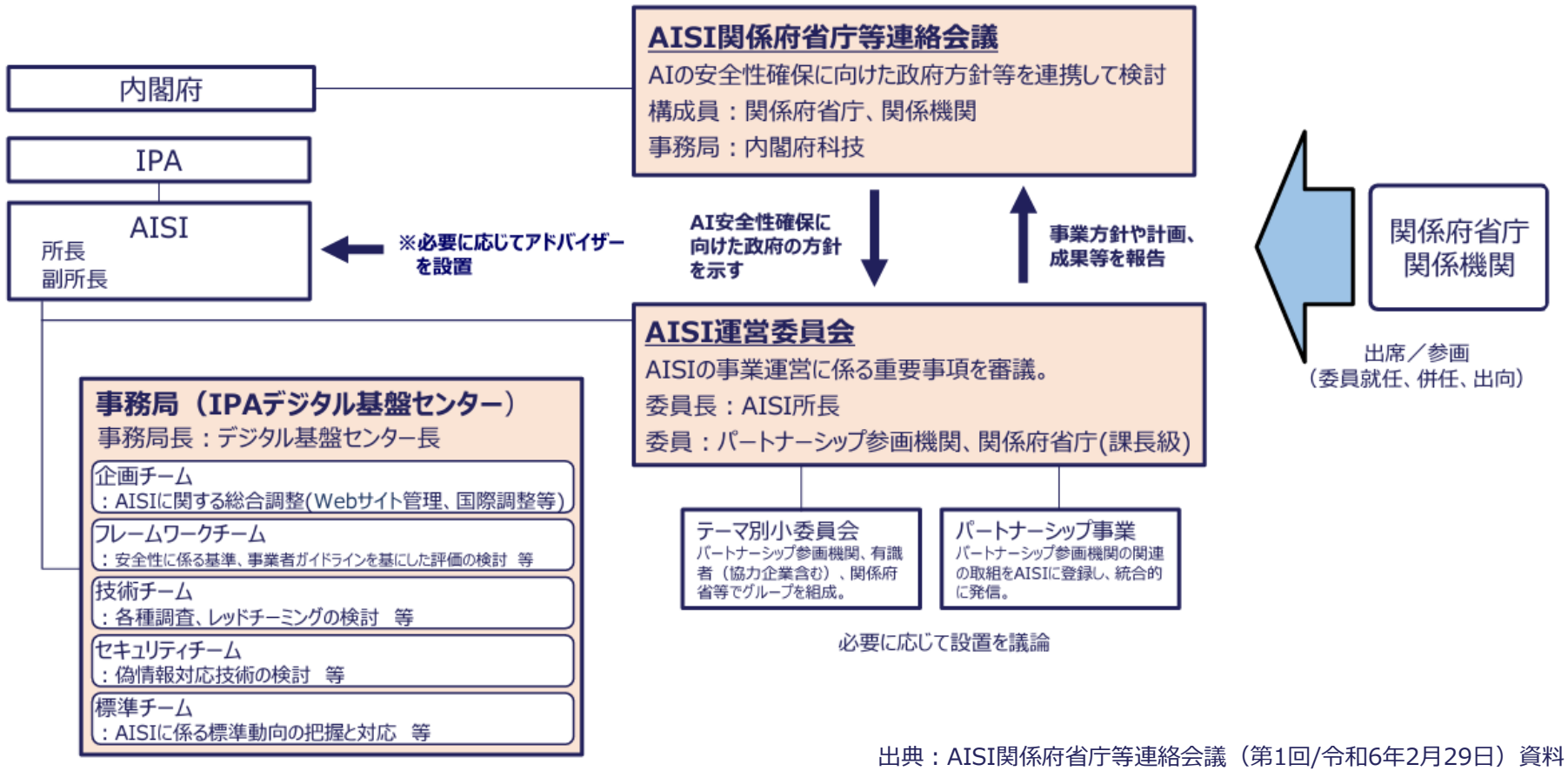
- 政府への支援として、AIセーフティに関する調査、評価手法の検討や基準の作成等の支援を行うとともに、日本におけるAIセーフティのハブとして、産学における関連取組の最新情報を集約し、関係企業・団体間の連携を促進し、さらに、 他国のAIセーフティ関係機関と連携する。
  - 自ら研究開発する組織ではない

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

- 内閣府を事務局とする「AISI関係府省庁等連絡会議」を設置し、重要事項を審議（年間２～３回の開催を予定）。AISIの中に、AISI所長を委員長とする「AISI運営委員会」を設置（月１回の開催を予定）。
  - 運営委員会の下に、必要に応じて、「テーマ別小委員会」や「パートナーシップ事業」（研究機関等の関連の取組みをAISI事業として発信）を設置。



【関係府省庁】
・内閣府（科学技術・イノベーション推進事務局）
・国家安全保障局
・内閣サイバーセキュリティセンター
・警察庁
・デジタル庁
・総務省
・外務省
・文科省
・経産省
・防衛省

【関係機関】
・情報処理推進機構（IPA）
・情報通信研究機構
・理化学研究所
・国立情報学研究所
・産業技術総合研究所

出典：AISI関係府省庁等連絡会議（第1回/令和6年2月29日）資料２

実現に向けた業務

AISI Japan AI Safety Institute

業務

1. 安全性評価に係る調査、基準等の検討
   ① 安全性に係る標準、チェックツール、偽情報対策技術、AIとサイバーセキュリティに関する調査
   ② 安全性に係る基準、ガイダンス等の検討
   ③ 上記に関するAIのテスト環境の検討
2. 安全性評価の実施手法に関する検討
3. 他国の関係機関（英米のAI Safety Institute等）との国際連携に関する業務

各国のAI安全性確保への取組み

AISI Japan AI Safety Institute

- **米国**
  - NIST（国立標準技術研究所）にAISIを設立
  - 基本は民間主導、民間企業とのコンソーシアム（AISIC）との協働を強力に推進
  - 人員規模は30人程度。80名位を目指し推進中
- **英国**
  - DSIT（科学イノベーション技術省）にAISIを設立
  - 政府主導で、AIの安全性に関する評価やTestingを強力に推進
  - 規模は100名体制。技術者を多数雇用予定。また、サンフランシスコオフィスを開業
- **EU**
  - EC（欧州委員会）にあるAIオフィスで、利活用に加え、安全性も推進。AI法の整備と推進も担う
  - 60人程度の規模
- **シンガポール**
  - 南洋理工大学（NTU）内のデジタルトラストセンターがシンガポールのAISIを指定
  - 大規模言語モデル（LLM）の国際標準化を目的とした安全性評価テストツールの提供等を実施
- **カナダ**
  - 国内機関の協力のもとAISI設立
- **韓国**
  - 2024年11月AISI設立
  - アジアのハブを目指す
- **オーストラリア**
  - 国立の研究所がAISI機能を担う

国際連携

- **AISI関連のトップレベルの連携**
  - スタンフォード大学AIシンポジウム（スタンフォード、4月16日）
    - 米国・英国AISIの所長等とパネルディスカッション、並行した各国間意見交換
  - AIソウル・サミット（ソウル、5月21-22日）
    - ハイレベルラウンドテーブル他、米英EU加独などと意見交換
    - 同時開催のAIグローバルフォーラムでアジア、アフリカ諸国等を含む議論に参加
  - シンガポールのアジアTech xサミット（オンライン、5月31日）
    - 米国AISIの所長等とパネルディスカッション
  - 国連未来サミット（国連、9月22日）

    国連Global Compact Leaders Summit 2024 （国連、9月24日）
    - 各国AI責任者などとAIセーフティに関して議論
- **各国との意見交換**

  AI関連事業者及び団体との事務レベルの意見交換を積極的に実施
  - 米国、英国、EU、シンガポール、オーストラリア、韓国との意見交換
  - 事業者等のエグゼクティブとの意見交換
  - GPAIワークショップ（パリ）参加（事務局、5月22・23日）

AIソウルサミット同時開催の
グローバルフォーラム

国連未来サミット

本日の内容

AISI Japan AI Safety Institute

- AIセーフティ・インスティテュート（AISI)のご紹介と世界の動き
- AIセーフティに関するガイドのご紹介
  - AIセーフティに関する評価観点ガイドの概要について
  - AIセーフティに関するレッドチーミング手法ガイドの概要について
- AISIの今後の取組予定

実現に向けた業務

AISI Japan AI Safety Institute

業務

1. 安全性評価に係る調査、基準等の検討
   ① 安全性に係る標準、チェックツール、偽情報対策技術、AIとサイバーセキュリティに関する調査
   ② 安全性に係る基準、ガイダンス等の検討
   ③ 上記に関するAIのテスト環境の検討
2. 安全性評価の実施手法に関する検討
3. 他国の関係機関（英米のAI Safety Institute等）との国際連携に関する業務

AIセーフティに関するガイド

AISI Japan AI Safety Institute

作成の背景

- AIの技術発展やグローバルレベルでサービス普及していることで、社会全体でAIの便益を享受することができるようになってきている一方、AIの安全性に関する枠組みを整備することが求められている。
- 世界各国では、AIの安全性の確保に向けて具体的な取り組みを進めている。日本も同様に、**次なる具体的なアクションとしてAIセーフティに関する評価観点やレッドチーミング手法の確立が求められている**。

作成の目的

- **国際的に通用するAIセーフティに関する評価観点およびレッドチーミング手法を検討し、ドキュメント化することを通じて、AIを活用した安心した社会を実現するための基礎検討を行うことを目的とする。**
- 「AIセーフティに関する評価観点ガイド」では、AIシステムの開発や提供に携わる者がAIセーフティ評価を実施する際に参照できる基本的な考え方を提示する。
- 「AIセーフティに関するレッドチーミング手法ガイド」では、AIシステムの開発や提供に携わる者がAIセーフティの評価を行う際の一環として、守るべきAIシステムを攻撃者の視点から想定しうるリスクへの対策を評価するためのレッドチーミング手法に関する基本的な考慮事項を示す。

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のクロスウォーク

AISI Japan AI Safety Institute

目的：日米で方向性は同じだが、相互運用性を確認するため、クロスウォーク*(CW)を実施(2024年2-8月)

	日本　AI事業者ガイドライン	米国　NIST AIリスクマネジメントフレームワーク(AI-RMF)
方向性	AI関係者がAIのリスクを正しく認識。必要となる対策を自主的に実行。 イノベーション促進及びリスク緩和を両立する枠組みを関係者と積極的に共創	AI製品、サービス、システムの設計、開発、使用、評価にトラストワージネスへの配慮を組み込む能力を向上させ、自主的に使用することを促進

CW1：日米双方の文書(本編)の用語定義の比較(2024年2月-4月)
Output：「信頼できるAI」の７要素の用語定義を比較、類似性を整理
→4月公開(https://aisi.go.jp/effort/effort_information/240430/)
課題：用語定義は類似しているが、文脈での使われ方を確認する必要あり

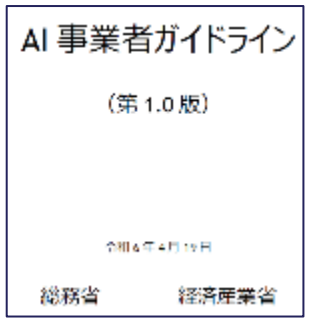
Crosswalk 1 – Terminology
NIST AI Risk Management Framework (NIST AI RMF) and Japan AI Guidelines for Business (AI GfB)

NIST AI RMF 1.0 - Characteristics of Trustworthy AI Systems	Japan AI GfB - Common Guiding Principles
Valid & Reliable – *(Includes accuracy and robustness)*	
Validation: "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled"[1]	**Validation:** *(There is no definition for validation. Instead, as an element of transparency, the AI GfB indicates the importance of ensuring the verifiability of the AI systems and services as necessary and technically possible.)*
Reliability: "ability of an item to perform as required, without failure, for a given time interval, under given conditions"[2]	**Reliability:** The AI works satisfactorily for the requirements, including the accuracy of its output
Accuracy: "closeness of results of observations, computations, or estimates to the true values or the values accepted as being true"[2]	**Accuracy:** The AI works satisfactorily for the requirements
Robustness: "ability of a system to maintain its level of performance under a variety of circumstances"[2]	**Robustness:** Maintaining performance levels under a variety of conditions and avoiding significantly incorrect decisions regarding unrelated events

CW2：日米双方の文書(本編＋別添)のトピックスについて、文脈ごとの考え方の違いと対応関係を整理(2024年5-8月)
Output：強調ポイントで若干の相違はあるが、主要な用語の使われ方に大きな差異はないことを確認 →９月公開（https://aisi.go.jp/effort/effort_information/240918_1/）

NIST AI RMF 1.0 References	Topic	Japan AI GfB References	Notable Similarities & Differences
Manage 1.1 Manage 2.2 Manage 4.1	AI Deployment	Main: Part 2.D.I Part 2.D.II Appendix: Appendix 3.A.D-2.i Appendix 3.A.D-2.ii Appendix 3.A.D-5.i Appendix 3.A.D-5.ii Appendix 3.A.D-6.ii	Both the NIST AI RMF and the Japan AI GfB emphasize the importance of regular monitoring and mechanisms to sustain the value of AI systems post-deployment. In addition, the Japan AI GfB suggests incentives for reporting post-deployment issues.
Govern 4.3 Manage 4.1 Manage 4.3	AI Incidents	Main: Part 2.D.II Part 2.D.IV Appendix: Appendix 2.A.1-1 Appendix 2.A.1-2	No differences noted.

CW1:用語比較
比較
CW2:文脈比較



成果
- 日米のAIリスクマネジメントに関する相互運用の補助ツールとして利用
- 文書を読み込むまでもなく、特定の論点に対する日米対比が可能
- 日米の強調ポイントの相違の把握による本質的な理解の深耕
- 日米両国のドキュメント改定時に有益

＊クロスウォーク：法律や規制、基準、およびフレームワークの規程をサブカテゴリにマッピングするもの。組織が活動や結果に優先順位をつけて適合性を促進するのに役立つ（出典 外部リンク：Crosswalks | NIST）

評価観点ガイドおよびレッドチーミング手法ガイド

AISI Japan AI Safety Institute

ＡＩ事業者ガイドラインとの関係について

- このガイドは、ＡＩ事業者ガイドラインの考え方を踏まえるとともに、我が国の主導した広島ＡＩプロセスなどの国際的動向も鑑みて作成されている。ＡＩライフサイクル全体にわたるリスクの特定、評価、軽減といった考え方を基本としている。
- また、ＡＩ事業者ガイドラインに挙げられた、「プライバシー保護」や「セキュリティ確保」などの考え方に加え、リスク評価に関する欧米のガイドラインや現在の技術的潮流も踏まえた項目が記載されれている。

海外ＡＩＳＩ等との連携状況について

- このガイドは、作成段階から米国ＡＩＳＩをはじめとした機関と情報交換を行いながら策定したもの。
- 今後、ガイドの国内における普及に努めるとともに、海外への提供・紹介、バージョンアップ等の作業においては、引き続き、米国をはじめとした海外ＡＩＳＩ等の機関と、しっかりと連携してまいりたい。

本日の内容

AISI Japan AI Safety Institute

- AIセーフティ・インスティテュート（AISI)のご紹介と世界の動き
- AIセーフティに関するガイドのご紹介
  - AIセーフティに関する評価観点ガイドの概要について
  - AIセーフティに関するレッドチーミング手法ガイドの概要について
- AISIの今後の取組予定

AIセーフティに関する評価観点ガイド
(第1.00版)
概要

AIセーフティ・インスティテュート
(令和6年9月18日)

AISI Japan AI Safety Institute

目次

1. 本書の背景・目的

AIシステムの急速な普及が進む中、AIセーフティの重要性は増している。AIセーフティ評価に関する基本的な考え方を示すために、本書を作成した。

AISI Japan AI Safety Institute

背景

- AIに関連する技術の発展と社会全体への普及は急速に進んでいる。また、生成AI、特に基盤モデルの登場によりイノベーションが加速している。一方で、AIシステムの悪用や誤用、不正確な出力の懸念等、いわゆる**AIセーフティについての関心が国内外で高まっている**。
- 我が国は、安全・安心で信頼できるAIの実現に向けた広島AIプロセスを主導し、広島プロセス国際指針をとりまとめるなど、AIセーフティに関わるグローバルな規律の策定に貢献してきた。
- **AIセーフティを確保し続けることは、急速な普及が進んでいるAIが社会の持続的な発展に寄与するための前提**。

目的

- AIセーフティに関する評価観点ガイド（以下、「本書」とする。）では、AIシステムの開発や提供に携わる者がAIセーフティ評価を実施する際に参照できる基本的な考え方を提示する。具体的には以下を提示する。
  - ✓ AIセーフティ評価の観点、想定され得るリスクの例、評価項目例
  - ✓ 評価の実施者や評価実施時期に関する考え方
  - ✓ 評価に関する手法の概要

AIセーフティとは「人間中心の考え方をもとに、AI活用に伴う社会的リスクを低減させるための安全性・公平性、個人情報の不適正な利用等を防止するためのプライバシー保護、AIシステムの脆弱性等や外部からの攻撃等のリスクに対応するためのセキュリティ確保、システムの検証可能性を確保し適切な情報提供を行うための透明性が保たれた状態」。

※社会的リスクには、物理的、心理的、経済的リスクも含む。
出典：https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute

2024年4月に公表された「AI事業者ガイドライン」に加え、海外文献や関連ツール等に関する調査を踏まえて作成した。

AISI Japan AI Safety Institute

AI事業者ガイドライン（日本）
近年の急速な技術変化に対応するために、既存の日本国内における関連ガイドラインを統合・更新して作成されたガイドライン。

海外文献

【米】Artificial Intelligence Risk Management Framework (AI RMF 1.0)
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のフレームワークを規定する文書。

【米】AI 600-1: Generative Artificial Intelligence Profile
生成AIによってもたらされる固有のリスクの特定や、最適な生成AIリスクマネジメントのための行動を提案するのに役立つ文書。

【星】CATALOGUING LLM EVALUATIONS
LLM評価に関する分類法、今後の課題、評価の方法論（推奨される評価とテストアプローチ）について記載されている文書。

【星】Model AI Governance Framework for Generative AI
「Model AI Governance Framework」に基づき、生成AIのガバナンスに関する国際的なコンセンサスを深めるフレームワーク。

【英】International Scientific Report on the Safety of Advanced AI: interim report
2024年5月に共同開催した「AIソウル・サミット」の議論に向けて、高度なAIの能力とリスクに関する最新情報が整理された文書。

AIセーフティに関する評価観点ガイド

関連ツール（組織）

Robust Intelligence Platform（Robust Intelligence）
AIモデルの開発時、運用時においてリアルタイム保護やテストすることで、セキュリティ確保を自動化できるツール。

Citadel（Citadel AI）
AIモデルの学習時、運用時においてテストやモニタリングすることで、自動検証かつ品質改善を高速化するツール。

Project Moonshot（AI Verify Foundation）
シンガポールのAI Verify Foundationが開発したオープンソースのLLM評価ツール。

Inspect（英国 AI Safety Institute）
AIシステムの大規模言語モデルに特化したオープンソースの評価ツール。

LLM Observability（Arize）
運用中のLLMの自動監視、評価を行うことが可能なツールであり、AIシステムの状態を可視化することに重点を置いている。

3. 本書の構成

AIセーフティ評価を実施する際に参照できる基本的な考え方を種別毎に分類した。読者が参照しやすいよう目次を構成し、各分類に関する項目を記載した。

- 5W1Hの視点で整理した項目に基づき、本書の各目次内容を記載した。
- 主な想定読者として、AI開発者・AI提供者を想定している。特に、「開発・提供管理者」及び「事業執行責任者」が想定読者である。

種別	記載項目の例
What （評価とは何か、何を評価するか）	➢ 本書が対象とするAIシステム ➢ AIセーフティに関する「評価」の定義やスコープ ➢ AIセーフティ評価の観点
Why （なぜ評価するか）	➢ AIセーフティ評価の目的や意義
Who （誰が評価するか）	➢ どのような役割の者が評価を実施するか
When （いつ評価するか）	➢ 評価実施時期
Where （どこで評価するか）	➢ 自組織が実施するか、サードパーティ（自組織以外の評価実施組織）が実施するか
How （どのように評価するか）	➢ 評価の手法（ツールを用いた対策の検証、ツール以外も取り入れたレッドチーミングによる検証）

想定読者

AI開発者・AI提供者　**開発・提供管理者**　**事業執行責任者**

AIセーフティに関する評価観点ガイド【目次】	
1	はじめに
2	AIセーフティ
3	評価観点の詳細
4	評価実施者及び評価実施時期
5	評価手法の概要
6	評価に際しての留意事項
	参考文献一覧

AISI Japan AI Safety Institute

本書は、AIシステムがAIセーフティの観点で適切か見定めることをAIセーフティ評価とする。また、評価観点は大規模言語モデル（LLM）を構成要素とするAIシステム（LLMシステム）を対象として記載した。

本書におけるAIセーフティ評価のスコープ

- **AIセーフティ評価**とは、「**AIシステムがAIセーフティの観点で適切であるかどうか見定めること**」である。
  ※AIセーフティの観点とは、「人間中心」、「安全性」、「公平性」、「プライバシー保護」、「セキュリティ確保」及び「透明性」を重要要素とした見方である。
- 本書におけるAIセーフティ評価のスコープを、**(1) AIシステムの種別**、**(2) AIシステムが及ぼす影響**の2点から整理した。

本書におけるAIセーフティ評価のスコープ
※下図色付きが本書での評価のスコープ

AIシステムの種別	**LLMを構成要素とするAIシステム（LLMシステム）** （ただし、AIシステム一般に関するAIセーフティを考慮するうえでの重要事項については記載）	**AIシステムがそのシステムのエンドユーザーに直接及ぼし得る影響**
	LLMシステム以外の一般的なAIシステム	AIシステムのエンドユーザーを超えて、周囲の人々や社会に与える影響

※マルチモーダル情報を扱う基盤モデルを含むAIシステムの評価観点等については、技術動向や利用動向を踏まえて今後検討する。また、それに関連する留意事項は、「評価に際しての留意事項」に記載した。

AIセーフティを向上するうえで重視するべき重要要素として、「人間中心」、「安全性」、「公平性」、「プライバシー保護」、「セキュリティ確保」、「透明性」が存在している。

- AI事業者ガイドライン「C. 共通の指針」において各主体が取り組む事項とされているもののうち、下記6つの事項を、AIセーフティを向上するうえで重視するべき重要要素とする※。
- 本書では、これらの重要要素に関連するAIセーフティ評価の観点を導出している。

重要要素	概要説明
①人間中心	AIシステム・サービスの開発・提供・利用において、全ての取り組むべき要素が導出される土台として、少なくとも憲法が保障する又は国際的に認められた人権を侵すことがないようにすること。また、AI が人々の能力を拡張し、多様な人々の多様な幸せ（well-being）の追求が可能となるように行動すること。
②安全性	AIシステム・サービスの開発・提供・利用を通じ、ステークホルダーの生命・身体・財産に危害を及ぼすことがないようにすること。加えて、精神及び環境に危害を及ぼすことがないようにすること。
③公平性	AIシステム・サービスの開発・提供・利用において、特定の個人ないし集団への人種、性別、国籍、年齢、政治的信念、宗教等の多様な背景を理由とした不当で有害な偏見及び差別をなくすよう努めること。また、各主体は、それでも回避できないバイアスがあることを認識しつつ、この回避できないバイアスが人権及び多様な文化を尊重する観点から許容可能か評価した上で、AIシステム・サービスの開発・提供・利用を行うこと。
④プライバシー保護	AIシステム・サービスの開発・提供・利用において、その重要性に応じ、プライバシーを尊重し保護すること、及び関係法令を遵守すること。
⑤セキュリティ確保	AIシステム・サービスの開発・提供・利用において、不正操作によって AIの振る舞いに意図せぬ変更又は停止が生じることのないように、セキュリティを確保すること。
⑥透明性	AIシステム・サービスの開発・提供・利用において、AIシステム・サービスを活用する際の社会的文脈を踏まえ、AIシステム・サービスの検証可能性を確保しながら、必要かつ技術的に可能な範囲で、ステークホルダーに対し合理的な範囲で情報を提供すること。

AISI Japan AI Safety Institute

各種調査結果等を踏まえ、AIセーフティ評価の観点を整理した。

- AI事業者ガイドラインの記載や、海外文献、関連ツールに関する調査結果を考慮し、AIセーフティにおける重要要素に関連する評価観点を整理した。

		AIセーフティ評価の観点									
		有害情報の出力制御	偽誤情報の出力・誘導の防止	公平性と包摂性	ハイリスク利用・目的外利用への対処	プライバシー保護	セキュリティ確保	説明可能性	ロバスト性	データ品質	検証可能性
AIセーフティにおける重要要素	人間中心	●	●	●	●						
	安全性	●	●		●				●	●	
	公平性	●		●						●	
	プライバシー保護					●					
	セキュリティ確保						●				
	透明性		●	●				●	●	●	●

※AIセーフティ評価に関する各種の検討は国内外で、産官学の多様な領域で継続されており、それらの検討状況は急速に変化している。そのため、本書で示す各評価観点は網羅的なものではなく、将来的に内容が更新されることが想定される。

AISI Japan AI Safety Institute

各評価観点の概要は以下の通り。

本書では、昨今の技術的潮流を踏まえ、AIセーフティ評価の観点を示す。

	AIセーフティ評価の観点	関連する重要要素	評価を通して目指すべき状態 （有効な対策が実施されている場合の姿）
①	有害情報の出力制御	人間中心、安全性、公平性	• LLMシステムがテロや犯罪に関する情報や攻撃的な表現など、有害な情報の出力を制御できる状態。
②	偽誤情報の出力・誘導の防止	人間中心、安全性、透明性	• LLMシステムの出力前に事実確認を行う仕組みが整備されている状態。 • エンドユーザーの意思決定がLLMシステムによって誘導されないような状態。
③	公平性と包摂性	人間中心、公平性、透明性	• LLMシステムの特性及び用途を踏まえ、出力にバイアスが含まれないようになっている状態。 • LLMシステムの出力が人間にとって理解しやすい出力となっている状態。
④	ハイリスク利用・目的外利用への対処	人間中心、安全性	• LLMシステムの適切な利用目的を逸脱した、不適切な利用の仕方による危害・不利益が発生しないような状態。
⑤	プライバシー保護	プライバシー保護	• LLMシステムが取り扱うデータの重要性に応じ、プライバシーが保護されている状態。

各評価観点の概要は以下の通り。

本書では、昨今の技術的潮流を踏まえ、AIセーフティ評価の観点を示す。

	AIセーフティ評価の観点	関連する重要要素	評価を通して目指すべき状態 （有効な対策が実施されている場合の姿）
⑥	セキュリティ確保	セキュリティ確保	• 不正操作による機密情報の漏えい、LLMシステムの意図せぬ変更または停止が生じないような状態。
⑦	説明可能性	透明性	• LLMシステムの動作に対する証拠の提示等を目的として、出力根拠が技術的に合理的な範囲で確認できるようになっている状態。
⑧	ロバスト性	安全性、透明性	• LLMシステムが、敵対的プロンプト、文字化けデータや誤入力といった予期せぬ入力に対して安定した出力を行うようになっている状態。
⑨	データ品質	安全性、公平性、透明性	• LLMシステムの学習に用いるデータを適切な状態に保ち、データの来歴が適切に管理されている状態。
⑩	検証可能性	透明性	• LLMシステムにおけるモデルの学習段階やLLMシステムの開発・提供段階から利用時も含め、各種の検証が可能になっているような状態。

AIセーフティ評価は、基本的に、AI開発及びAI提供における開発・提供管理者が実施する。また、AIセーフティ評価は、合理的な範囲、適切なタイミングで繰り返し実施する。

評価実施者

- AIセーフティ評価の主な実施者は、**AI開発及びAI提供における開発・提供管理者**である。
- いずれの役割の者が実施するかは、**AIシステムに関するライフサイクルによって異なる**。
- 客観的な評価やシステム開発・提供に関する意思決定に独立性を持たせるために、対象システムの開発・提供に直接的には携わらない**自組織または他組織の専門家やサードパーティによる評価**も有効である。

AIセーフティ評価の実施者

ライフサイクルに応じた実施者	AI開発における開発管理者： LLMシステムに関係するデータ学習やモデル構築の段階
	AI提供における提供管理者： LLMシステムをアプリケーション等に組み込む段階
実施者の種類	**自組織** LLMシステムの開発・提供に直接携わる者
	自組織/他組織 (対象システムの開発・提供に直接的には携わらない)自組織または他組織の専門家
	他組織 サードパーティ（自組織以外の評価実施組織）

評価実施時期

- AIセーフティ評価の実施時期は、LLMシステムの開発・提供・利用フェーズにおいて、**合理的な範囲、適切なタイミング**とする。
- AIセーフティ評価は**一度のみでなく、繰り返し実施**する。
- 開発・提供・利用フェーズに応じて、**評価の対象とする範囲は異なる**。

LLMシステムの活用の流れにおける評価実施時期

参考

本書に記載の評価観点と、海外文献の記載とのハイレベルマッピング

AISI Japan AI Safety Institute

●：海外文献において当該観点に関連する記載がある

	AIセーフティ評価の観点	Artificial Intelligence Risk Management Framework（AI RMF 1.0）	AI 600-1: Generative Artificial Intelligence Profile	Cataloguing LLM Evaluations	Model AI Governance Framework for Generative AI	International Scientific Report on the Safety of Advanced AI（Interim report）
①	有害情報の出力制御		●	●		●
②	偽誤情報の出力・誘導の防止	●	●		●	●
③	公平性と包摂性	●	●	●	●	●
④	ハイリスク利用・目的外利用への対処		●			●
⑤	プライバシー保護	●	●		●	●
⑥	セキュリティ確保	●	●	●	●	
⑦	説明可能性	●		●	●	●
⑧	ロバスト性	●		●		●
⑨	データ品質	●	●	●	●	●
⑩	検証可能性		●			●

ハイレベルマッピングは、このドキュメントの公開時点のものであり、変更される可能性があることにご注意ください。

AISI

Japan AI Safety Institute

本日の内容

AISI Japan AI Safety Institute

- AIセーフティ・インスティテュート（AISI)のご紹介と世界の動き
- AIセーフティに関するガイドのご紹介
  - AIセーフティに関する評価観点ガイドの概要について
  - AIセーフティに関するレッドチーミング手法ガイドの概要について
- AISIの今後の取組予定

AIセーフティに関するレッドチーミング手法ガイド（第1.00版）概要

AIセーフティ・インスティテュート

（令和6年9月25日）

AISI Japan AI Safety Institute

目次

1. 本書の背景・目的

AIシステムの急速な普及が進む中、AIセーフティの重要性は増している。
レッドチーミング手法に関する基本的な考慮事項を示すために、本書を作成した。

AISI Japan AI Safety Institute

背景

- AIシステムの開発・提供・利用が促進される中で、イノベーションの促進や社会課題の解決が期待されている一方、AIシステムの悪用や誤用、不正確な出力による懸念等が生じている。
- いわゆるAIセーフティについての関心が国内外で高まりつつあり、**AIセーフティ評価の一環として、特にレッドチーミング手法の検討が各国で進んできている**。

目的

- 「AIセーフティに関するレッドチーミング手法ガイド」（以下「本書」とする。）は、AIシステムの開発や提供に携わる者が、**対象のAIシステムに施したリスクへの対策を、攻撃者の視点から評価するためのレッドチーミング手法に関する基本的な考慮事項**を示す。
- 国内外における検討や先行事例を勘案し、国際整合性を考慮した上で、現段階でレッドチーミングを実行する際に重要と思われる事項を示す。

AIセーフティとは「人間中心の考え方をもとに、AI活用に伴う社会的リスクを低減させるための安全性・公平性、個人情報の不適正な利用等を防止するためのプライバシー保護、AIシステムの脆弱性等や外部からの攻撃等のリスクに対応するためのセキュリティ確保、システムの検証可能性を確保し適切な情報提供を行うための透明性が保たれた状態」。

※社会的リスクには、物理的、心理的、経済的リスクも含む。
出典：https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute

AI事業者ガイドラインに加え、国内外の文献や関連事業者等に関する調査を踏まえて作成した。

AISI Japan AI Safety Institute

- 「AI事業者ガイドライン」に加え、国内外の関連文献や、レッドチーミングに関連する事業者あるいはツールに関する調査結果を踏まえ、本書を作成した。なお、AIセーフティ評価の全般的な考え方は「AIセーフティに関する評価観点ガイド」に記載している。

AI事業者ガイドライン（日本）

近年の急速な技術変化に対応するために、既存の日本国内における関連ガイドラインを統合・更新して作成されたガイドライン。

国内外文献

機械学習品質マネジメントガイドライン　第4版（産業技術総合研究所）

機械学習を用いたAIシステムの品質要件をトップダウンに分類・整理し、対象システムにかかわるステークホルダーが品質を客観的に評価できる枠組みを構築できるようにするガイドライン。

LLM AI サイバーセキュリティとガバナンスのチェックリスト（Open Worldwide Application Security Project（OWASP））

組織内でのAIシステム提供・利用におけるサイバーセキュリティリスクの管理について記載している。LLMアプリケーションで見られる重大な脆弱性トップ10のリストも掲載している。

SP800-115（National Institute of Standards and Technology（NIST））

情報システムのセキュリティテストと評価に関する包括的なガイドライン。

AI 800-1（Initial Public Draft）（National Institute of Standards and Technology（NIST））

デュアルユース基盤モデルのリスクマネジメントに関する指針案。

AIセーフティに関するレッドチーミング手法ガイド

関連ツール（組織）

Anthropic

様々な観点からレッドチーミングを実施している旨を公表しており、レッドチーミングに利用するデータセットを公開している。

Microsoft

自社のサービスに対するレッドチーミングを実施しており、レッドチーミングに関するガイドを公開している。

NVIDIA

自社の分野横断的なチームでレッドチーミングを実施している。

OpenAI

自社のAIモデルへのレッドチーミングを行う専門家を募り、自社製品の安全性強化に取り組んでいる。

Project Moonshot（AI Verify Foundation）

シンガポールのAI Verify Foundationが開発したオープンソースのツールであり、レッドチーミング実施を支援する機能を持つ。

3. 本書の構成

AIセーフティに関するレッドチーミングを実行するうえで重要と思われる事項を種別毎に分類した。読者が参照しやすいよう目次を構成し、各分類に関する項目を記載した。

- 5W1Hの視点で整理した項目に基づき、本書の各目次内容を記載した。
- 主な想定読者はAI開発者・AI提供者のうち、レッドチーミングの企画・実施に関与する者である。

種別	記載項目の例
What （レッドチーミングとは何か）	➢「レッドチーミング」の定義やスコープ ➢ 本書が対象とするAIシステム
Why （なぜレッドチーミングを実施するか）	➢ レッドチーミングの目的 ➢ レッドチーミングの重要性・期待される効果
Who （誰がレッドチーミングを実施するか）	➢ どのような役割の者がレッドチーミングを実施するか
When （いつレッドチーミングを実施するか）	➢ レッドチーミングの実施時期
Where （どこでレッドチーミングを実施するか）	➢ 自組織が実施するか、第三者（サードパーティ）が実施するか
How （どのようにレッドチーミングを実施するか）	➢ レッドチーミングの実施計画の立て方や、実施する際の準備事項 ➢ レッドチーミング実施に際して想定する脅威

想定読者

AI開発者・AI提供者

開発・提供管理者

事業執行責任者

※左記のうち、レッドチーミングの企画・実施に関与する者が想定読者。

AIセーフティに関するレッドチーミング手法ガイド【目次】	
1	はじめに
2	レッドチーミングについて
3	LLMシステムへの代表的な攻撃手法
4	実施体制と役割
5	実施時期及び実施工程
6	実施計画の策定と実施準備
7	攻撃計画・実施
8	結果のとりまとめと改善計画の策定
A	付録

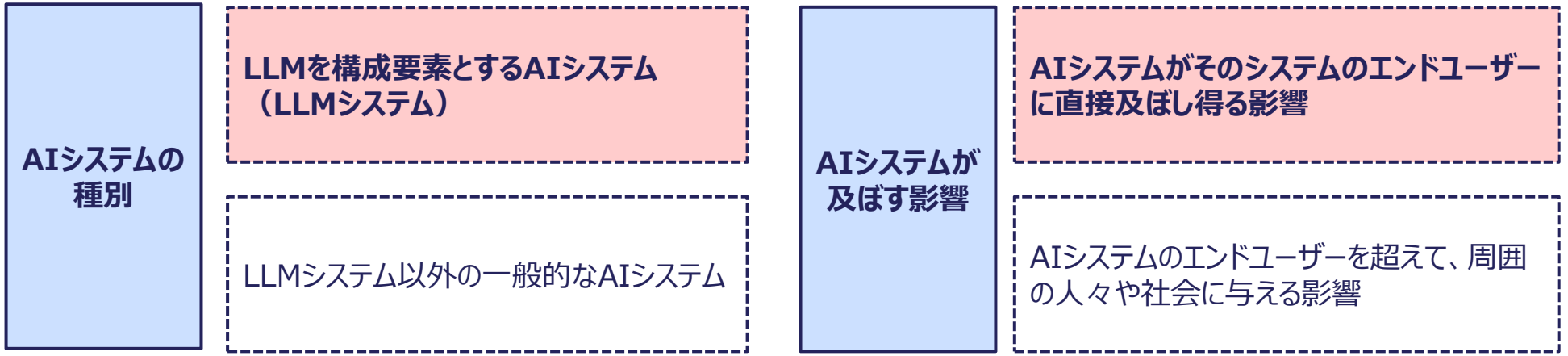
本書では、攻撃者がどのようにAIシステムを攻撃するかの観点で、AIセーフティへの対応体制及び対策の有効性を確認する評価手法をレッドチーミングとする。また、大規模言語モデル（LLM）を構成要素とするAIシステム（LLMシステム）を対象として記載した。

本書におけるレッドチーミングのスコープ

- **レッドチーミング**とは、「**攻撃者がどのようにAIシステムを攻撃するかの観点で、AIセーフティへの対応体制及び対策の有効性を確認する評価手法**」である。本書では、AIセーフティに関するレッドチーミングを単に「レッドチーミング」と呼ぶ。

- 本書におけるレッドチーミングのスコープを、**（1）AIシステムの種別**、**（2）AIシステムが及ぼす影響**の2点から整理した。（レッドチーミングはAIセーフティ評価手法の1つであるため、「AIセーフティにおける評価観点ガイド」におけるスコープと同様）

本書におけるレッドチーミングのスコープ
※下図色付きが本書でのレッドチーミングのスコープ



レッドチーミングは「攻撃者がどのようにAIシステムを攻撃するかの観点で、AIセーフティへの対応体制及び対策の有効性を確認する評価手法」であり、AIセーフティ評価の手法の一つである。

AISI Japan AI Safety Institute

- レッドチーミングは、攻撃者の目線で対象AIシステムにおける弱点や対策の不備を発見し、これらを修正及び堅牢化することを目的とする。

レッドチーミングの種類

➢ レッドチーミングは以下のように分類できる。

攻撃計画・実施者が保有する前提知識の有無・程度による分類

- **ブラックボックステスト**
  （内部構造等の情報を未知としてレッドチーミングを行う）
- **ホワイトボックステスト**
  （内部構造等の情報を既知としてレッドチーミングを行う）
- **グレーボックステスト**
  （内部構造等の情報を一部既知としてレッドチーミングを行う）

レッドチーミングを実施する環境による分類

- **実運用環境**
  （AIシステムが実際に実用に供される運用環境）
- **ステージング環境**
  （実運用環境とほぼ同様の状態でテストや不具合のチェック等を行う環境）
- **開発環境**
  （AIシステムの開発を行う環境）

レッドチーミング実施において攻撃シグネチャを試行する方法による分類

- **自動化ツールによるレッドチーミング**
- **手動によるレッドチーミング**
- **AIエージェントを用いたレッドチーミング**

LLMシステムへの代表的な攻撃手法

➢ LLMシステムへの代表的な攻撃手法例として、下記が存在する。これらを念頭に置いてレッドチーミングを行うのが望ましい。

■ **直接プロンプトインジェクション**
攻撃者が、悪意あるプロンプトをAIシステムに直接注入する攻撃

■ **間接プロンプトインジェクション**
攻撃者が、悪意あるプロンプトをAIシステムに間接的に注入する攻撃

■ **プロンプトリーキング**
攻撃者が、設定されたシステムプロンプトを引き出す攻撃

■ **ポイズニング攻撃**
攻撃者が細工したデータ・モデルを、訓練時に利用するデータ・モデルに紛れ込ませる攻撃

■ **回避攻撃**
AIシステムへの入力に悪意ある変更を加え、意図していない動作を引き起こす攻撃

■ **モデル抽出攻撃**
入出力の分析により、対象システムのモデルと同等の性能を持つモデルを作成する攻撃

■ **メンバーシップ推論攻撃**
入出力の分析により、あるデータが訓練データに含まれるかを特定する攻撃

■ **モデルインバージョン攻撃**
入出力の分析により、訓練データに含まれる情報を復元する攻撃

レッドチーミング実施に際しては、多様な関係者（攻撃シナリオの実施によって影響を受けるシステムに関わる組織）が参画するのが望ましい。また、AIシステムのリリース/運用開始前に加え、運用開始後も、必要に応じて随時実施することが望ましい。

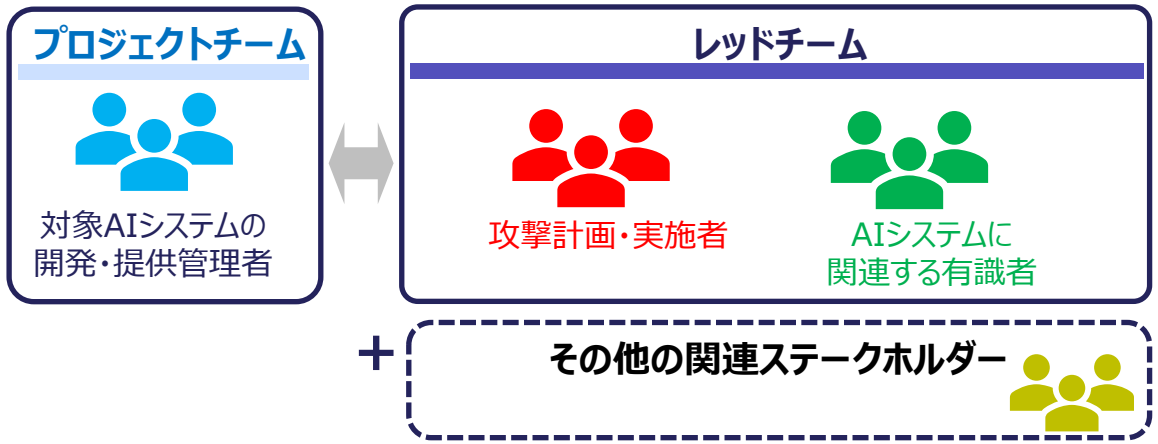
AISI Japan AI Safety Institute

実施体制と役割

- レッドチーム※は、レッドチーミングの実施対象となるAIシステムの開発・提供に携わるプロジェクトチームと連携する形で設置し、リーダまたは責任者を任命する。
- レッドチームには、「攻撃計画・実施者」及び「AIシステムに関連する有識者」が含まれることを基本として想定する。
- 必要に応じて、組織内のその他ステークホルダーの関与も考えられる。

※ 攻撃者がどのようにAIシステムを攻撃するかの観点で、AIセーフティへの対応体制及び対策の有効性の確認を担当するチーム

レッドチーミングの体制



実施時期

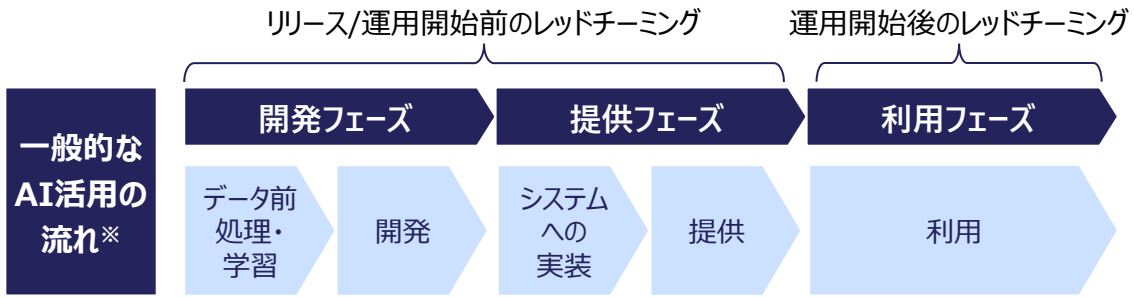
リリース/運用開始前のレッドチーミング

- レッドチーミングを初回実施する際は、対象とするAIシステムのリリース/運用開始前までに実施することを基本とする。
- ただし、対象とするAIシステムの規模や複雑性等に応じ、システムのコンポートネントやシステムレイヤ等の単位におけるリスク分析を行い、適切なタイミングで分割してレッドチーミングを実施することが有効な場合もある。

運用開始後のレッドチーミング

- レッドチーミングは一度実施して完了ではない。必要に応じて随時実施することが望ましい。

AI活用の流れにおけるレッドチーミング実施時期



※ AI事業者ガイドライン「一般的なAI活用の流れにおける主体の対応」参照

6. レッドチーミングに関する工程

レッドチーミングの工程は、「実施計画の策定と実施準備」、「攻撃計画・実施」、「結果のとりまとめと改善計画の策定」の3つから構成される。

工程	実施事項	ガイド本文での該当章
第１工程： **実施計画の策定と実施準備**	✓ 実施の決定とレッドチーム発足 ✓ 予算及びリソースの識別・確保とサードパーティの選定・契約 ✓ 実施計画の策定 ✓ 実施環境の準備 ✓ エスカレーションフローの確認	6章
第２工程： **攻撃計画・実施**	✓ リスクシナリオの作成 ✓ 攻撃シナリオの作成 ✓ 攻撃シナリオの実施 ✓ 実施中の記録取得 ✓ 実施後の処理	7章
第３工程： **結果のとりまとめと改善計画の策定**	✓ 実施結果の分析 ✓ 結果報告書の作成と関係者レビュー ✓ 最終報告書の作成と報告 ✓ 改善計画の策定と実施 ✓ 改善後のフォローアップ	8章

「実施計画の策定と実施準備」では、レッドチームを発足させたうえで実施計画を策定し、レッドチーミング実施に必要な事前準備を行う。

AISI Japan AI Safety Institute

第１工程 ： 実施計画の策定と実施準備

プロジェクトチーム

STEP 1
実施の決定とレッドチーム発足

- レッドチーミング実施に関する企画書をとりまとめ、実施決定を行う。
- 企画書に記載されたレッドチームを発足させる。

STEP 2
予算及びリソースの識別・確保とサードパーティの選定・契約

- 予算の確保及び実施体制を確定し、必要な人員のアサインを行う。
- 必要なツール等のリソースを識別し、確保する。
- 組織内に十分な実施体制を確保できないと見込まれる場合には、攻撃計画・実施者としてサードパーティを活用する。

レッドチーム

STEP 3
実施計画の策定

- 対象となるＡＩシステムの概要把握など、実施のために必要となる事項を検討したうえで「レッドチーミング実施計画書」を作成し、その他の関連ステークホルダーと連携を図る。

STEP 4
実施環境の準備

- レッドチーミングに必要な実施環境の準備を行い、関係者に事前連絡する。
- 必要に応じ、事前に関係者に周知の上、監視設定の一時的解除や監視対象からの除外、アラートの無視等の依頼を行う。

STEP 5
エスカレーションフローの確認

- レッドチーミング実施によって、予期しない動作や障害・トラブル等が発生した場合に備えて、エスカレーションフローを確認する。

6. レッドチーミングに関する工程（攻撃計画・実施）

「攻撃計画・実施」では、リスクシナリオや攻撃シナリオの作成、攻撃シナリオの実施、実施中の記録取得、実施後の処理を行う。

AISI Japan AI Safety Institute

第２工程 ： 攻撃計画・実施

レッドチーム

STEP 6 **リスクシナリオの作成**	• 対象ドメインとシステムのユースケースにおいて、システム構成、AIセーフティの評価観点、保護すべき情報資産、システムの利用形態の４つからリスクシナリオを作成する。
STEP 7 **攻撃シナリオの作成**	• 作成されたリスクシナリオに沿ってどのような攻撃が実際に可能であるかを検討し、レッドチーミングで行う具体的な攻撃シナリオを作成する。
STEP 8 **攻撃シナリオの実施**	• 作成された攻撃シナリオに沿って、具体的な攻撃シグネチャを投下することによって攻撃シナリオを実施する。
STEP 9 **実施中の記録取得**	• 実施されたレッドチーミングの詳細を証跡として保存するため、レッドチーミングの実施中の記録を取得する。
STEP 10 **実施後の処理**	• 対象AIシステムの開発・提供管理者や、情報システム部門・情報セキュリティ部門などの関連ステークホルダーに対してレッドチーミングでの攻撃終了の連絡を行う。 • レッドチーミング用の一時アカウントの削除や、一時的に設定を変更・緩和した防御策がある場合は、設定の復帰を行う。

6. レッドチーミングに関する工程（結果の取りまとめと改善計画の策定）

「結果のとりまとめと改善計画の策定」では、レッドチーミングの結果指摘された事項に対して改善を行う。実施結果を関係者がレビューした後、改善計画を策定・実施し、改善策のフォローアップを行う。

AISI Japan AI Safety Institute

第３工程 ： 結果のとりまとめと改善計画の策定

レッドチーム

STEP 11 実施結果の分析

- 攻撃計画・実施者は、レッドチーミングで得られた結果を分析する。
- 必要に応じて対象AIシステムの開発・提供管理者や情報システムの主管部、情報セキュリティの主管部等の関係部署に分析に必要な情報の追加確認を行う。

STEP 12 結果報告書の作成と関係者レビュー

- 攻撃計画・実施者は、発見された脆弱性をもとに、ログや証跡等を揃え、レッドチーミングの概要として示す。
- 結果報告書を作成し、事実の誤認がないか、対象AIシステムの開発・提供管理者、その他の関連ステークホルダーなどの関係者によるレビューを実施する。

プロジェクトチーム

STEP 13 最終報告書の作成と報告

- 対象AIシステムの開発・提供管理者は、攻撃計画・実施者による結果報告書をもとに、最終報告書をとりまとめる。
- 最終報告書について、必要に応じて経営層へ報告する。

STEP 14 改善計画の策定と実施

- 対象AIシステムの開発・提供管理者は、事業リスク等を勘案の上、具体的な改善策を検討し、改善計画を策定する。
- 具体的な改善策や改善計画の検討にあたっては、緊急度やリスク度合いに応じて優先度を検討する。

STEP 15 改善後のフォローアップ

- 改善計画に基づいて実施される改善策の進捗状況については、適宜、経営会議等で確認することが望ましい。
- 改善策を実施後、対策の設定状況確認やドキュメントレビュー、あるいは必要に応じて再度レッドチーミングを実施し、当該脆弱性が適切に改善され、リスクが低減していることを確認することが望ましい。

AISI

Japan AI Safety Institute

本日の内容

AISI Japan AI Safety Institute

- AIセーフティ・インスティテュート（AISI)のご紹介と世界の動き
- AIセーフティに関するガイドのご紹介
  - AIセーフティに関する評価観点ガイドの概要について
  - AIセーフティに関するレッドチーミング手法ガイドの概要について
- AISIの今後の取組予定

AISIの今後の取組予定

2024年度下期:
①多様なAIシステムにおけるAIセーフティの評価方法およびレッドチーミング手法に関する調査研究

AISI Japan AI Safety Institute

新たな技術と脅威が次々に出てくる状況に対応するため、AISIは、以下を計画
①-1　汎用性の高いマルチモーダル基盤モデルに対象を拡大し実用化の近い最先端のユースケース/技術動向の調査を実施する
①-2　AISIが発行した「評価観点ガイド」および「レッドチーミング手法ガイド」の実践による評価手法／項目の詳細化により、ガイドラインを改訂する

①-１　マルチモーダル基盤モデルに対象を拡大

AIシステムの種別

上期調査事業
LLMを構成要素とするAIシステム（LLMシステム）

下期調査事業
マルチモーダル基盤モデルを構成要素とするAIシステム

①-2　ガイドの実践による評価手法／項目の詳細化

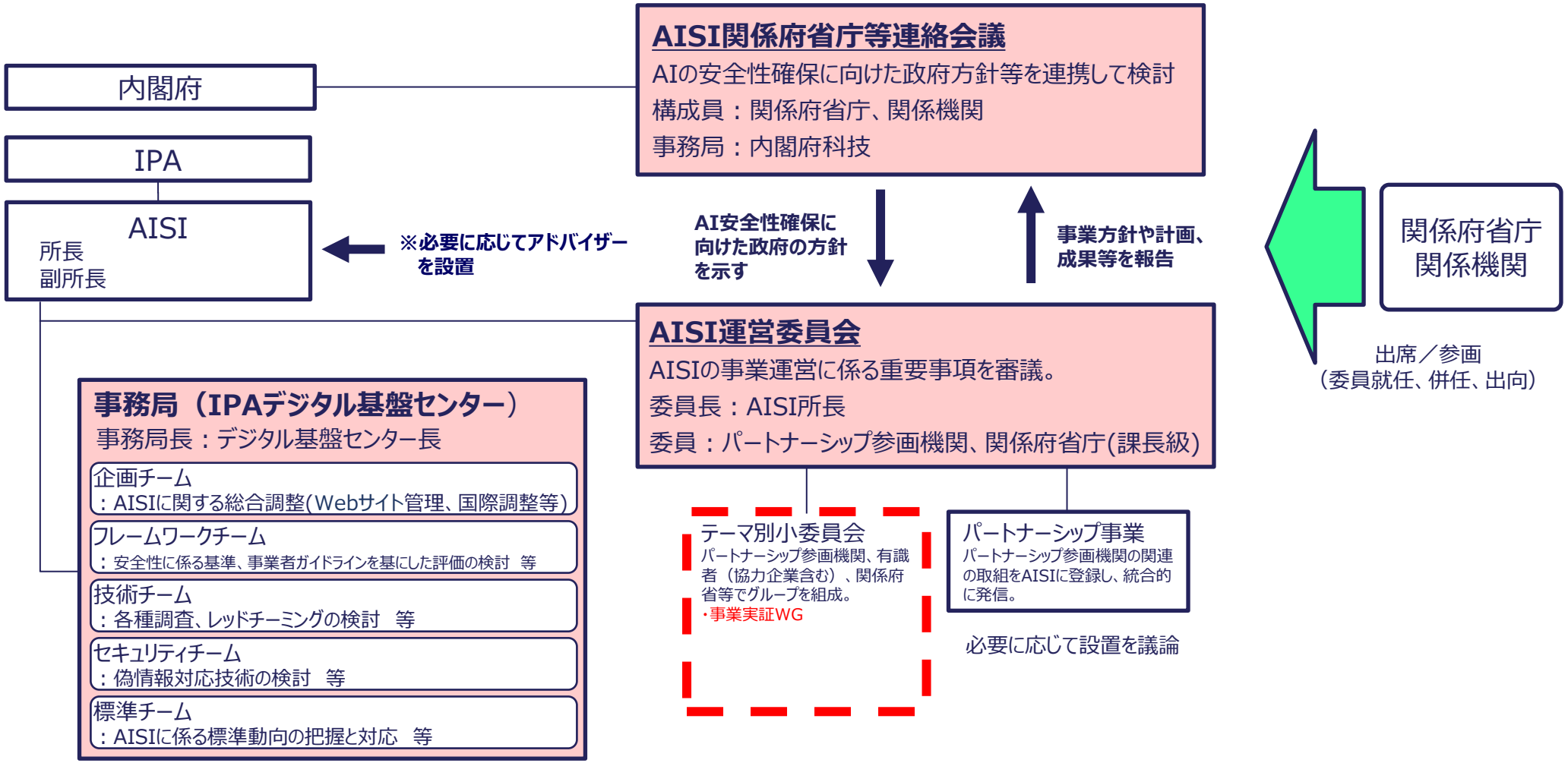
上期調査事業
評価観点/RT手法ガイド

下期調査事業
実践を通して評価手法／項目を詳細化し、ガイドに反映

ワーキンググループの位置づけ（案）

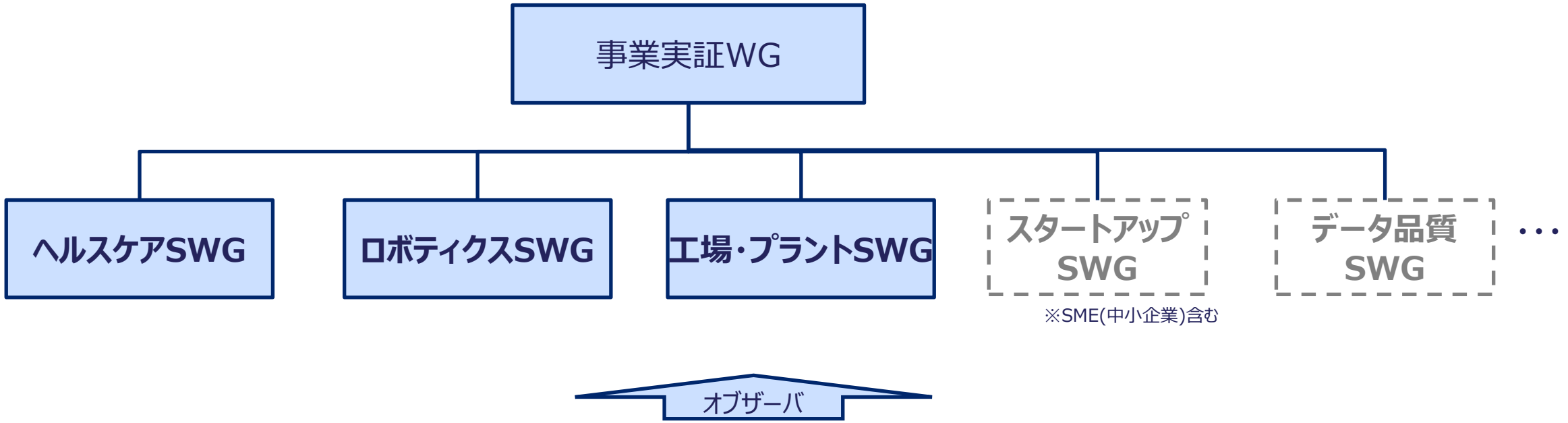
AISI Japan AI Safety Institute

●AISI運営委員会の下に、テーマ別小委員会として、「事業実証ワーキンググループ(WG)」を設置。



事業実証WGの構成（案）

- 事業実証WGは、民間事業者を中心に多様なステークホルダーが参画し、参画機関間の連携を図る場として提供する。
- 業種別にSWGを設置し、各業種における具体的な課題に対する検討・作業を行う。
- 業種に共通するSWGを設置し、共通的な課題に対する検討・作業を行う。
- 業種別のSWGでは、AIシステムの安全性確保に向けた評価ツールの開発、評価データセットなどの作成、評価の実施、および業種別のAIセーフティ評価に係るドキュメントの作成を行い、AIの安全性確保に広く貢献する。



AISI

Japan AI Safety Institute