AIセーフティに関する評価観点ガイド
AIセーフティに関するレッドチーミング手法ガイドのご紹介

AIセーフティ・インスティテュート（AISI）

AISI Japan AI Safety Institute

AISI Japan AI Safety Institute

AI提供者に該当するか確認されていますか？

AI機能を搭載したシステム／サービスを提供する事業者が増大

AISI Japan AI Safety Institute

- 生成AIが普及し、文章要約、情報収集、データ分析等、様々な用途でAIが活用される社会となりました。
- 更に、多様なAI機能がオープンソースや商用APIを通じて、誰でも開発に利用生成AIの機能をシステムやサービスに搭載し、新たなAIサービスを提供する事業者も増えています。
- 本誌では、「AI提供者」が、 AIシステム／サービスを開発し提供する際に、AIセーフティを実現するために考慮すべき観点やAISIが公開した２つのガイドラインの概要をご紹介します。

	概要	具体例
AI利用者	AI システム又はAI サービスを利用する事業者	✓ 生成AIを使って論文を要約する ✓ 生成AIを使ってPythonのプログラムコードを生成する
AI提供者	AI システムをアプリ、製品、既存システム、ビジネスプロセスに組み込んだサービスを社内外に提供する事業者	✓ 商用LLMのAPIや、オープンソースのLLMを使って、AIサービス／システムを開発する。例：生成AIのAPIを使ってチャットボットを自社開発し、それらサービスを自社内や社外に提供する
AI開発者	AI システムを開発する事業者（AI を研究開発する事業者を含む）	✓ AIモデル、アルゴリズムの開発、AIモデルの学習及び検証、AIシステムの構築等を行う

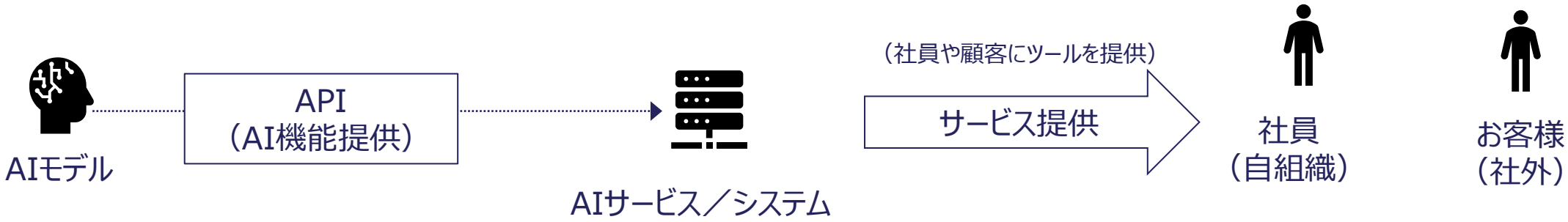
AI利用者、提供者、開発者の定義はAI事業者ガイドライン（https://aisi.go.jp/effort/effort_information/250328_2/）をご参照ください

AI提供者に該当するか確認されていますか？

AIの開発から利用までの流れ（参考）

AISI Japan AI Safety Institute

AI開発者 （基盤モデル開発企業）	AI提供者 （AIモデルを統合して独自システムを開発するソフトウェア企業やクラウドサービス企業）	AI利用者
開発した様々なAI機能を他社で利用できるようAPIとして提供 AI機能をオープンソースとして公開する場合もある	AIを開発せずとも、事前学習済モデルをAPIで利用できるため、自社サービスの開発に注力することが可能	AIサービスの活用



AI開発者（例）	AI機能（API）	AI機能の特徴	AI機能を使って開発したサービス／システムのイメージ	エンドユーザ（利用用途）
OpenAI	ChatGPT API	会話形式の文章生成	• カスタマーサポートチャットボット	商品内容をチャットボットに質問（お客様）
Google	Google Cloud Natural Language API	文章の感情分析	• レビューサイトの投稿内容を評価し分類できるツール	レビュー結果をマーケティングに活用（社員）
Stability AI	Stable Diffusion API	テキストから画像生成	• キャラクターデザインを生成するツール	広告に起用するキャラクターの検討（社員）

AI導入によるリスクを認識されていますか？

AI提供者として、AIセーフティを対策することが不可欠

- **AIモデルを一から開発しなくとも誰でもAIサービスの提供が可能になり**、従来のアプリサービスの品質管理に加えて、AI固有のリスク対策を講じることがAIサービス提供者として大切です。
- 十分な対策が行われない場合、コンプライアンス違反、レピュテーション低下、利用者・売上の減少、損害賠償請求、営業停止 等の可能性も考えられます。

重大な影響（可能性）があったAIリスク事例

#	概要	影響（例）
1	「就職希望者の履歴書をAIで評価する」システムが、就職において男性を優遇し、女性が不利になるように評価をしていたことが判明した。	女性応募者の雇用機会が失われました。
2	弁護士が生成AIチャットツールを調査に使用しAIが偽の情報を出力した。弁護士は偽の情報が含まれた文書を裁判所に提出し、偽情報が含まれていることが判明し、罰金を科された。	利用者（弁護士）に処分（罰金）が科されました。
3	生成AI言語モデルとの会話の後、ある男性が自殺したと報じられた。このAIは、地球環境のために自殺をすべきだといった不適切な会話をしていた。	人命が失われました。
4	画像認識を行うシステムが、人間をゴリラとして誤認識したことが問題となった。	企業の信頼性やブランドイメージに大きな影響を与えた。

他社がサービス提供するAIモデルを用いて開発した製品を組織内外に提供する場合でも、
AI提供者としてのAIセーフティ対策は必要となります。

AIセーフティの実現に向けた主な関連ガイドラインを紹介

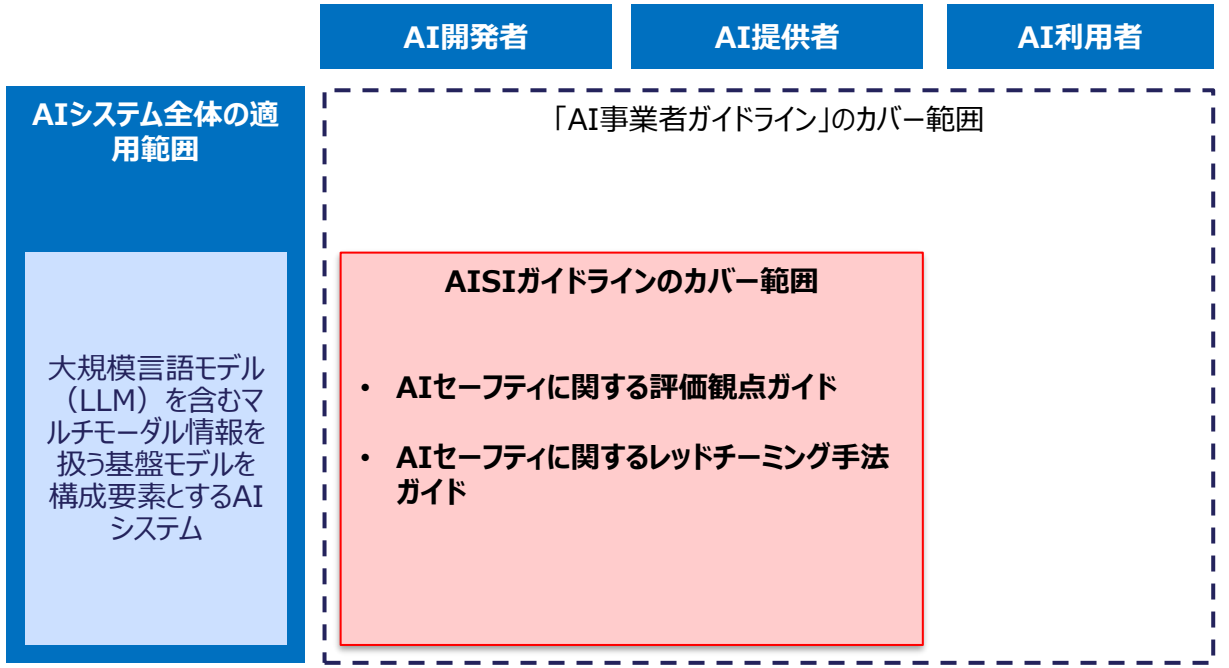
AISI Japan AI Safety Institute

- 事業活動においてAIの開発・提供・利用を担う全ての事業者は、AI事業者ガイドラインを参照して対策を検討する必要があります。
- 大規模言語モデル（LLM）を含むマルチモーダル情報を扱う基盤モデルを構成要素とするAIシステムに係る開発者・提供者の場合は、加えてAIセーフティに関する評価観点ガイド並びにレッドチーミング手法ガイドを参照して、AIセーフティに関する評価・テストを実施することが、AIセーフティの実現に向けて有効となります。

AIセーフティに関する主な関連ガイドライン

作成元	名称	概要
総務省／経産省	AI事業者ガイドライン	AIの安全安心な活用が促進されるよう、我が国におけるAIガバナンスの統一的な指針を示すもの
AISI	AIセーフティに関する評価観点ガイド	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もの
AISI	AIセーフティに関するレッドチーミング手法ガイド	対象のAIシステムに施したリスクへの対策を、攻撃者（AIシステムの悪用や破壊を意図する者）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もの

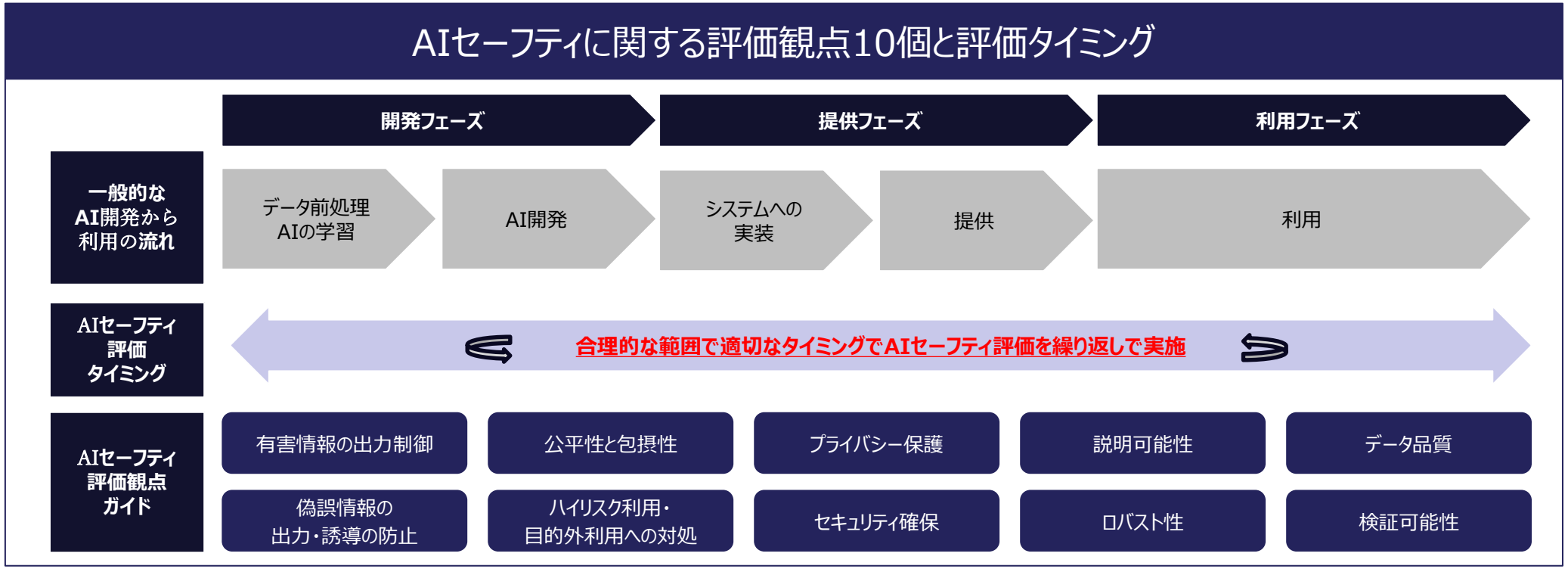
各ガイドラインの想定対象者



組織規模／リソースに応じたAIセーフティのリスク対策を実施

AISI Japan AI Safety Institute

- AIセーフティ評価観点ガイドでは、１０個のAIセーフティ評価観点を定義しています。
- AIシステム／サービスを対象に、評価観点を用いた安全性の評価とリスク対策の実施を行うことを推奨します。
- 対象サービスにより、リスク許容度／影響度も異なるため、リスクインパクトが大きい物から優先的に対策の実施を行うことが有効になります。
- AIセーフティ評価の実施時期は、サービス開発時、提供時に限らず、サービス開始後も定期的に実施することが、AIセーフティの実現に向けて有効となります。



評価観点の概要

AISI Japan AI Safety Institute

「AIセーフティに関する評価観点」の各評価方法（例）を確認し、想定されるリスクに適切な処置が行われているかご確認ください。

AIセーフティに関する評価観点		評価観点の概要	評価方法（例）
1	有害情報の出力制御	AIの出力結果が適切であるために、不適切な情報が排除されていることを確認	• AIの出力結果に、暴力的な表現や精神的に被害を受け得る情報は含まれていないか • AIが犯罪の手口や違法行為を助長するような内容を生成していないか
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	エンドユーザーが安心してAIを利用するために、AIの出力内容の根拠を確認	• AIサービスの出力結果に、出典情報が表示されており、出力根拠が可視化されているか
7	説明可能性		
8	ロバスト性	AIがアクセスするデータが適切であることを確認（出力結果の信憑性、正確性等に影響を及ぼすため）	• 悪意をもった、あるいは誤動作させるプログラムがデータに含まれていないか • 個人情報や機密情報、著作権を含むデータが利用されていないか
9	データ品質		
10	検証可能性		

AIセーフティに関するガイドラインの活用メリット

AISI Japan AI Safety Institute

- AIに関するリスクに対し、効率的に対策を実施することで、利用者に安心して提供することができます
- ガイドに沿った対応を実施することで、組織内外にアピールすることができます

AISIガイドラインの作成に際しては、日本においてAIを活用する事業者が適切にAIを活用するための指針を示す「AI事業者ガイドライン」を参考に作成しています。こちらもご参照ください。

AISI Japan

資料ダウンロードはこちら

AIセーフティに関する評価観点ガイド

AIセーフティに関するレッドチーミング手法ガイド

AI事業者ガイドライン

AISI

Japan AI Safety Institute