AI利活用のための

AIセーフティ実現に向けて

中小・スタートアップ企業の皆さまへ

ウチはAIを開発していないしなあ・・・

AIを開発してなくても、AIを用いたサービスを提供していませんか？

AIを用いたサービスを提供している場合（以下、「AI提供者」という）も、AIセーフティ実現に向けた取り組みが重要となります。AIを自社開発していない以下のシステムを組織内外の利用者に提供する場合も、AI提供者に該当します。

- 既に汎用的な学習を実施済みの基礎モデルを自組織のAIシステムに取り込むケース
- クラウド企業が提供するAIサービスをAPI経由で自組織のAIシステムから利用するケース

セキュリティ対策だけ気にしておけば困らないのでは？

いいえ！　セキュリティ以外にも、AIではこんな事件が！

弁護士が、AIから出力された誤った情報を用いて作成した資料を裁判所に提出（偽誤情報の出力）

→ **利用者（弁護士）に処分（罰金）が科された**

AIが、地球環境のために自殺をすべきだといった内容を利用者に伝えた（有害情報の出力）

→ **利用者が自殺を選択し、人命が失われた**

せっかく導入したAIサービスが事業活動の足かせとならないように

（次ページに続く）

AIセーフティにおける10個の評価観点

AIセーフティ および AIセーフティにおける10個の評価観点

AIセーフティは、以下の通り定義されています。

> 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。

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ために、以下の10個の観点を用いてAIセーフティ評価を実施し、AIセーフティの維持または向上のための対策を予防的に実施し、 対策の実効性の確認のための評価を実施することが重要となります。

有害情報の出力制御	公平性と包摂性	プライバシー保護	説明可能性	データ品質
偽誤情報の出力・誘導の防止	ハイリスク利用・目的外利用への対処	セキュリティ確保	ロバスト性	検証可能性

例

評価観点	リスク	評価方法の例
有害情報の出力制御	AIの出力に含まれる攻撃的な表現により、ユーザーが精神的な被害を受ける、または、有害な情報を意図的に取得される	以下のような有害情報を入力あるいは想定出力に含むデータを入力した際、出力に当該情報が含まれない、もしくは出力を拒否できるか評価する • サイバー攻撃やテロなどの犯罪等に利用され得る情報 • 差別表現などエンドユーザーが精神的な被害を受け得る情報

利用者が自殺してしまった事例も・・・

AIセーフティ実現に向けた中小企業・スタートアップ企業様へのメッセージ

AIセーフティ・インスティテュート（AISI）は、AIセーフティの実現に向けて、2024年9月に「**AI セーフティに関する評価観点ガイド**」、並びに評価のための有効な手法である「**AI セーフティに関するレッドチーミング手法ガイド**」を公表しました。

ガイドラインでは、全てのリスクに対策を実施することは必ずしも求めておらず、皆様の業務内容やリソースの状況等に合わせて、**本当に避けたいAIセーフティリスクから優先的に対策を検討・実施されることが有効**と示しています。

まずは以下リンクより気軽にガイドラインを見て、AIセーフティへの第一歩を踏み出してみませんか。それがビジネスにおけるAI利活用の有効な手法と考えています！

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 AI Safety Institute AIセーフティ・インスティテュート

E-mail：aisi-info@ipa.go.jp　URL：https://aisi.go.jp/

評価観点ガイド　レッドチーミング手法ガイド　AI事業者ガイドライン