

生成 AI 時代のリスクと安心・安全について

2025年2月5日

損害保険ジャパン株式会社

執行役員CDaO データドリブン経営推進部長

AI セーフティ・インスティテュート 所長

村上 明子



村上 明子 (むらかみ あきこ)

AI Safety Institute 所長
損害保険ジャパン株式会社 執行役員CDaO

1999年日本アイ・ビー・エム（株）入社、同社東京基礎研究所において研究に従事。

2021年に損害保険ジャパン株式会社に転職、損害保険のデジタル・データの利活用の推進をしている。

2022年4月より同社執行役員CDO（チーフデジタルオフィサー）としてDXを牽引、2024年よりCDaO（チーフデータオフィサー）となり、データ戦略を担う。

2024年2月、AI Safety Instituteの設立とともに、初代所長となる。
損害保険ジャパンとは兼任となる。

※AISIIは、エイシーと読みます

デジタルの役割

人の役割

人をアシスト×人が出来ない領域を補完

人にしかできない価値創造に注力

引受判断



保険金支払



営業



引受情報の自動収集

人が出来ない領域を補完

引受判断

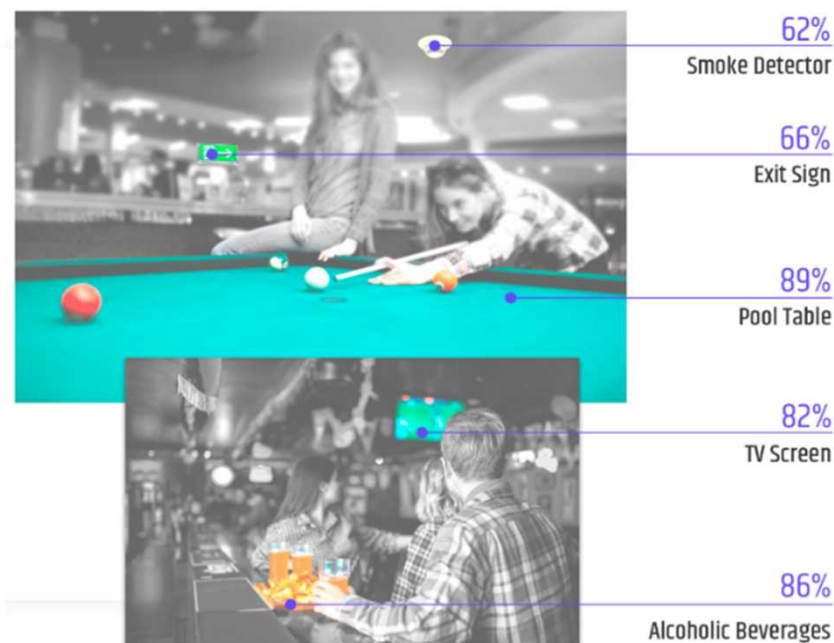


航空写真データやオンライン上のデータからAIにより保険見積りに必要なデータを情報を自動取得 → **人間が収集出来ない情報の量・種類**

航空写真データから建物情報の自動取得
(イスラエルGeoX社との協業)



オンラインデータから店内設備に関する情報を取得(イスラエルPlanck社との協業)



統合イノベーション戦略における3つの強化方策

(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

- ◆ 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

AI Safety Instituteについて

- ◆ 安全、安心で信頼できるAIの実現に向けて、AIの安全性に関する評価手法や基準の検討・推進を行うための国の機関
 - 今後、官民が協力して、AIの安全安心な活用が促進されるよう、AIの開発や利用をする全ての関係者がAIのリスクを正しく認識し、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ようにしていく必要がある。
 - また、これらの取組を通じ、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していく必要がある。

AISIとは上記を実現するための**官民の取組を支援する機関**

そもそも安全性（Safety）とは

- ◆ ISO/IEC GUIDE 51:2014(E)
 - Safety: Freedom from risk which is not tolerable
 - 安全とは、許容不可なりスクがないこと。



AIの利用で高まる「リスク」

「生成AI」とチャット後に自殺

“温暖化のことが心配で居ても立ってもいられなくなったベルギー人男性が、AI企業 *Chai Research* のAIチャットボット *Eliza* と6週間対話を続けているうちに地球の未来を託せるのはAIしかないと思い詰めるようになり、「**自分が犠牲になるから地球を救ってほしい**」と *Eliza* に言い残して**自らの命を絶ってしまいました。**”

特集 生成AI

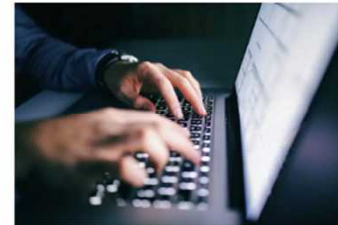
+ この特集をフォロー

デジタルを問う 欧州からの報告

AIとチャット後に死亡 「イライザ」は男性を追いやったのか？

深堀り 国際 | 速報 | 欧州

毎日新聞 | 2023/4/24 17:00(最終更新 11/30 12:43) 有料記事 3829文字



写真はイメージ=Getty

ある男性の自殺が3月下旬、ベルギーのメディアで報じられた。男性は直前まで人工知能(AI)を用いたチャットボット(自動会話システム)との会話にのめり込んでいた。遺族はチャットボットが男性に自殺を促したと主張し、波紋を広げている。【ブリュッセル岩佐淳士】

「イライザと会話しなければ…」

「死にたいのなら、なぜすぐにそうしなかったの?」。イライザが問いかけると、男性は答えた。「たぶんまだ、準備ができていなかったんだ」。しばらくしてイライザはこう切り出した。「でも、あなたはやっぱり私と一緒にいたいんでしょう?」――。

ベルギー紙「ラ・リーブル」によると、男性はこうした会話を最後に、自ら命を絶った。相手の「イライザ」は、米国のスタートアップ(新興企業)が運営するアプリ「Chai(チャイ)」のチャットボット。デジタル空間に作り出された架空の女性キャラクターだった。

<https://mainichi.jp/articles/20230423/k00/00m/030/156000c>

AIの利用で高まる「リスク」

ホーム > ニュース > 社会

生成AI 悪用しウイルス作成、警視庁が25歳の男を容疑で逮捕...設計情報を回答させたか

2024/05/28 07:35 生成AI

この記事をストックする

インターネット上で公開されている対話型生成AI（人工知能）を悪用してコンピューターウイルスを作成したとして、警視庁は27日、川崎市、無職の男（25）を不正指令電磁的記録作成容疑で逮捕した。複数の対話型生成AIに指示を出してウイルスの設計情報を回答させ、組み合わせて作成したという。生成AIを使ったウイルス作成の摘発は全国初とみられる。

▶生成AI搭載のiPhone 16、アップル幹部「日本は非常に重要な市場」...機能説明や下取り強化の方針



警視庁

捜査関係者によると、男は昨年3月、自宅のパソコンやスマートフォンを使い、対話型生成AIを通じて入手した不正プログラムの設計情報を組み合わせてウイルスを作成した疑い。

作成されたウイルスは攻撃対象のデータを暗号化したり、暗号資産を要求したりする機能があった。

<https://www.yomiuri.co.jp/news/national/20240528-OYT1T50015/>

「生成AI」悪用しウイルス作成

捜査関係者によると、男は昨年3月、自宅のパソコンやスマートフォンを使い、対話型生成AIを通じて入手した不正プログラムの設計情報を組み合わせてウイルスを作成した疑い。（中略）

男は調べに、容疑を認め「ランサムウェア（身代金要求型ウイルス）で金を稼いだかった。AIに聞けば何でもできると思った」と供述しているという。このウイルスによる被害は確認されていない。

AI事業者ガイドラインが想定するリスク

- ◆ ガイドラインの目的は、**安全安心な活用**。
 - 「本ガイドラインは、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。」
- ◆ 安全性確保のため、以下の共通の指針を示す。

共通の指針

- 1) 人間中心（社会的文脈、倫理、偽情報）
- 2) 安全性（AIシステムの信頼性・堅牢性、知的財産、制御可能性、目的を逸脱した利用、学習データの品質）
- 3) 公平性（バイアス）
- 4) プライバシ保護（プライバシー）
- 5) セキュリティ確保（不正操作、機密性・完全性・安全性、データの改ざん）
- 6) 透明性（ログ、AI情報の提供）
- 7) アカウンタビリティ（トレーサビリティ、ガバナンス、関係者情報の開示）
- 8) 教育・リテラシー
- 9) 公正競争確保
- 10) イノベーション

AIのリスクとは

◆ 生成AIなどの台頭により顕著となったAIのリスク

- AIリスクとは以下の2つに大別
 - 技術的リスク（誤判定、データやロジックによるバイアス、ハルシネーション、安全性、セキュリティ等）
 - 社会的リスク（プライバシー侵害、政治活動への悪用、不正目的、権力集中、財産権の侵害、環境負荷等）
- 企業のAI活用には、さらに、法的なリスク、レピュテーションのリスクも存在

AI原則からAIガバナンスへ

- ◆ 2010年代にAIが浸透しはじめたときに世界各国で作られたAI原則
 - 安全性・セキュリティ・プライバシー・公平性・透明性あるいは説明可能性
 - しかし、生成AIの出現により、透明性・説明可能性が困難に

AIの安全性の担保のためにはAIガバナンスが必要

AIガバナンスとはAIのリスクを受容可能な最小限に抑えつつ、AIがもたらす価値を最大化することを目的とする

AI規制のアプローチ

生成AIなどの新規技術に対するアプローチは特殊

- ◆ 権利ベースアプローチとリスクベースアプローチ
- ◆ 対応は、「ハードロー」か「ソフトロー」か
 - ハードローであっても多くの国は軍事関連は対象外
- ◆ ハードローで対応しようとするのはEU、カナダ。分野ごとではない水平アプローチ。
- ◆ 一方アメリカや日本はドメインごとのガイドラインを出す
セクターごとのアプローチ

欧米の制度の動向

AI制度研究会
第1回資料より抜粋

AISI Japan
AI Safety
Institute

欧米ともに、ソフトウェア（標準、ガイドライン等）とハードロー（法令）の組合せを模索

広範なハードローをソフトウェアで補完



- EU理事会は**AI法案**を採択（2024年5月）
- 欧州委員会は通信ネットワーク・コンテンツ・技術総局内に**AIオフィス**を設置（2024年5月）
- **デジタルサービス法（DSA）**を採択（2022年10月）

ソフトウェアをベースにしつつ、目的に応じてハードロー



- **ボランタリー・コミットメント**を公表（2023年7月） **大統領令**を発出（2023年10月）
- 海外顧客にIaaSを提供する際、身元を確認し政府へ報告を義務付ける**規則案**を公表（2024年1月）
- 選挙の保護やディープフェイクへの対応などを念頭に**複数の州で法規制**

欧州評議会、**AI、人権、民主主義、法の支配に関する枠組み条約**を採択（2024年5月）



国連では、「安全、安心で信頼できるAI」に関する**国連総会決議**を採択（2024年3月） **UNESCO、ITU**等においても議論。



※出典：各国の公開情報等を基に内閣府が作成した資料より抜粋

日本でのAI制度に関する動き

- ◆ AI戦略会議の下部組織として「AI制度研究会」を発足
- ◆ 岸田総理の指摘する4つの原則
 - リスク対応とイノベーション促進の両立
 - 技術・ビジネスの変化の速さに対応できる柔軟な制度の設計
 - 国際的な相互運用性、国際的な指針への準拠
 - 政府によるA I の適正な調達と利用

AI戦略会議・AI制度研究会合同会議

更新日：令和6年8月2日 | 総理の一日

📄 ポスト 📌 シェアする 📞 LINEで送る



会議のまとめを行う岸田総理1

日本におけるAISIの設立

- ◆ 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

出典：

広島AIプロセス <<https://www.soumu.go.jp/hiroshimaaiprocess/documents.html>>

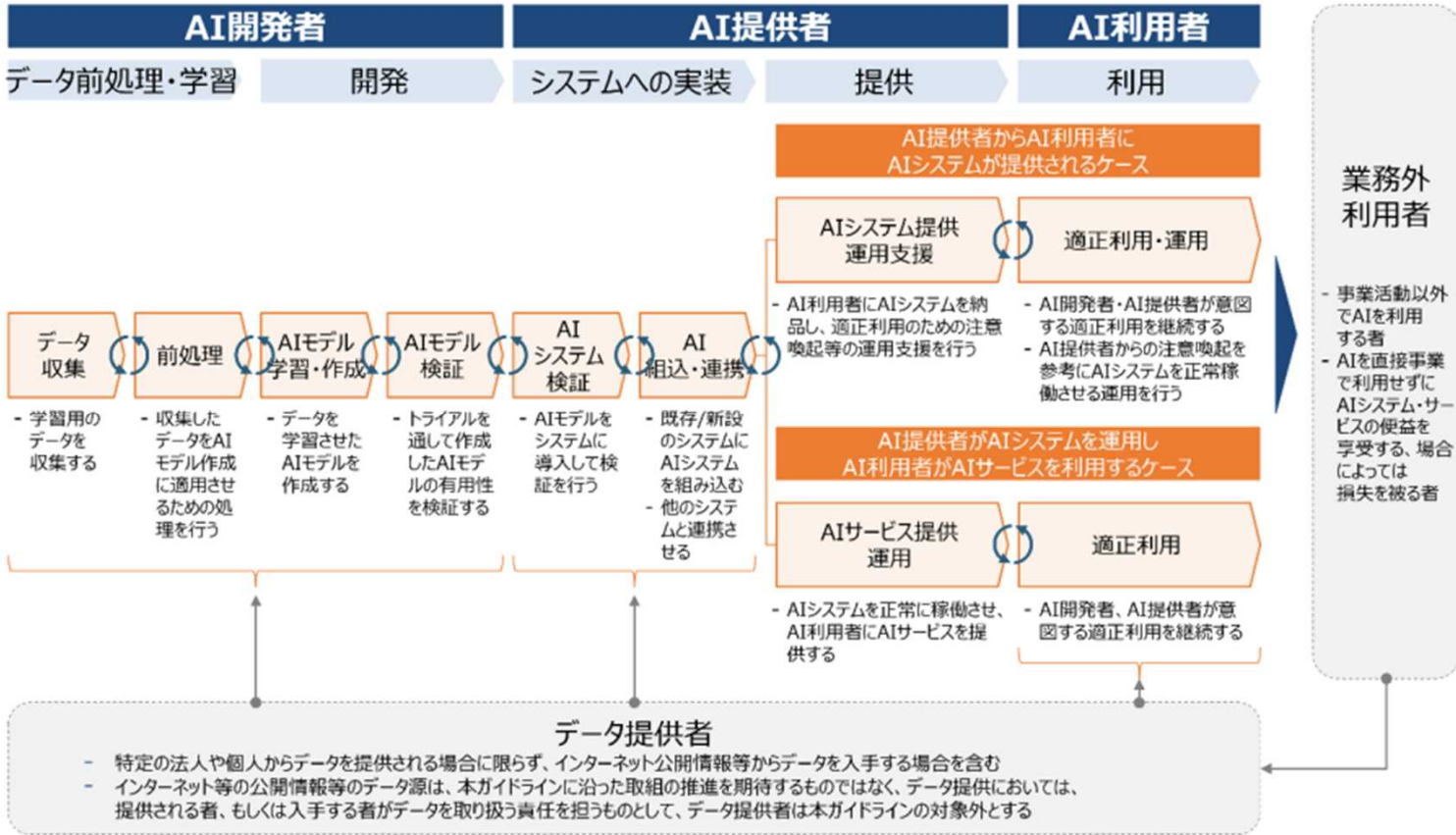
AI Safety Summit 2023 <<https://www.gov.uk/government/topical-events/ai-safety-summit-2023>>

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/>>

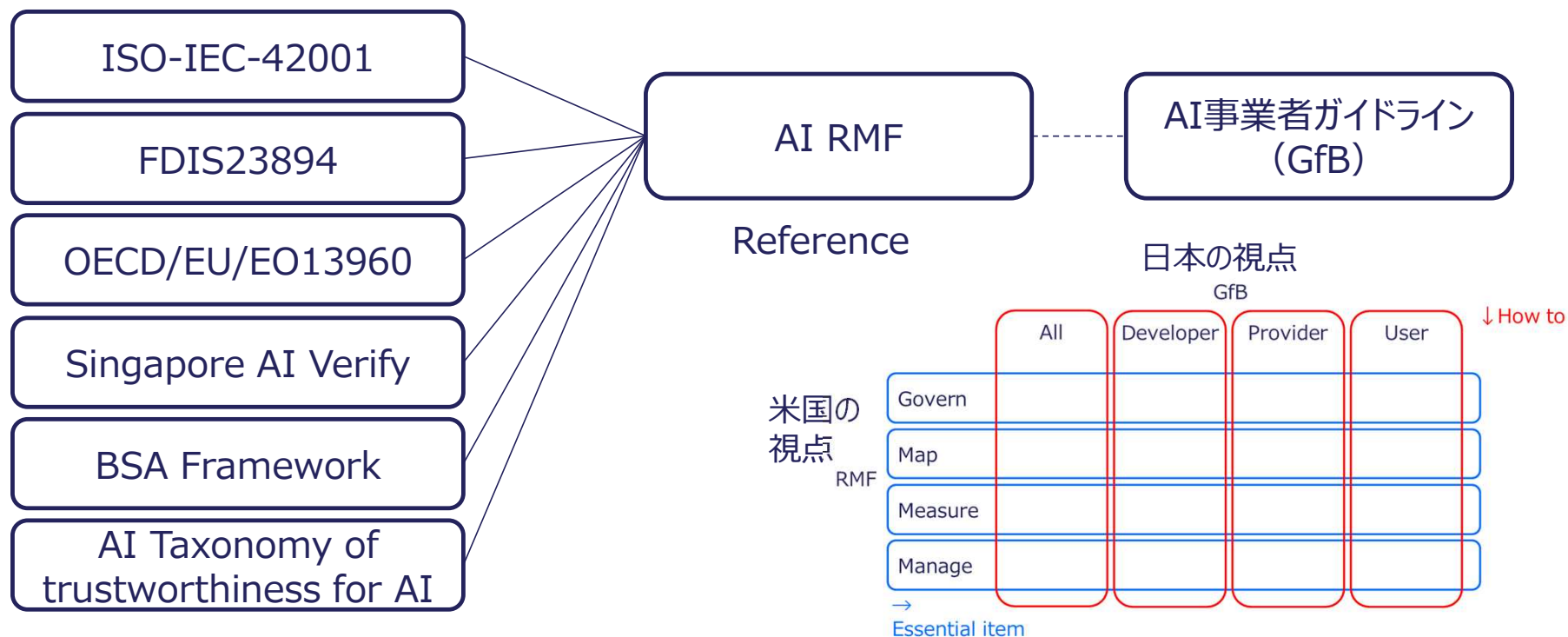
AI事業者ガイドラインの概要

◆ AI活用の流れの中で、各ステークホルダが対応すべきことを明確化









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セーフティに関する評価観点ガイドの公開

- ◆ 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べき重要要素とし、AIセーフティ評価の観点を導出

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。

AIセーフティに関するレッドチーミング手法ガイド

- ◆ 本ガイドは、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。

種別	記載項目の例
What (レッドチーミングとは何か)	<ul style="list-style-type: none"> ➢ 「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)	<ul style="list-style-type: none"> ➢ 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)	<ul style="list-style-type: none"> ➢ 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)	<ul style="list-style-type: none"> ➢ 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)	<ul style="list-style-type: none"> ➢ 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)	<ul style="list-style-type: none"> ➢ 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

そこにある「リスク」を認識し、備えて前へ進む

