日本のAI戦略を支えるAI安全性についてと AI Safety Instituteのご紹介

※AISIは、エイシーと読みます

2024-10-17

AISI Japan AI Safety Institute

自己紹介

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。

※AISIは、エイシーと読みます

日本の技術戦略とAI

AIのリスクと安全性

AI Safety Instituteのご紹介と世界の動き

AIリスクへの今後の対応

AISI Japan AI Safety Institute

日本の技術戦略とAI

AIのリスクと安全性

AI Safety Instituteのご紹介と世界の動き

AIリスクへの今後の対応

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

AISI Japan AI Safety Institute

日本の技術戦略とAI

AIのリスクと安全性

AI Safety Instituteのご紹介と世界の動き

AIリスクへの今後の対応

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、**許容不可なリスクがない**こと。

AIの利用で高まる「リスク」

AISI Japan AI Safety Institute

「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？

速報　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＝ゲッティ

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なりたいんでしょ？」――。

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　この記事をスクラップする

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搭載のアイフォーン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」悪用しウィルス作成

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）で金を稼ぎたかった。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

AISI Japan AI Safety Institute

- 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）
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

https://www.meti.go.jp/press/2024/04/20240419004/20240419004.html

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

様々なAIによるリスク　バイアス

テクノロジー　2018年10月11日 / 15:30 / 8ヶ月前

焦点：アマゾンがＡＩ採用打ち切り、「女性差別」の欠陥露呈で

Jeffrey Dastin　　2 分で読む

［サンフランシスコ　１０日　ロイター］ - 米アマゾン・ドット・コム(AMZN.O)が期待を込めて進めてきたＡＩ（人工知能）を活用した人材採用システムは、女性を差別するという欠陥が判明し、採用を取りやめる結果になった。

アマゾンは優秀な人材をコンピューターを駆使して探し出す仕組みを構築するため、２０１４年から専任チームが履歴書を審査するプログラムの開発に従事してきた。（中略）１０年間にわたって提出された履歴書のパターンを学習させたためだ。つまり技術職のほとんどが男性からの応募だったことで、システムは男性を採用するのが好ましいと認識したのだ。

逆に履歴書に「女性」に関係する単語、例えば「女性チェス部の部長」といった経歴が記されていると評価が下がる傾向が出てきた。

https://jp.reuters.com/article/amazon-jobs-ai-analysis-idJPKCN1ML0DN

INSIGHT　2018.01.18 THU 08:00

グーグルの画像認識システムは、まだ「ゴリラ問題」を解決できていない――見えてきた「機械学習の課題」

グーグルの「Google フォト」が黒人をゴリラとタグ付けした問題から2年以上が経ったいま、同社の画像認識システムはどこまで進化したのか。『WIRED』US版が5万枚以上の写真を使って調査したところ、一部の霊長類には写真検索が機能しないという実態が明らかになると同時に、機械学習の課題が浮き彫りになった。

自分と友人の写真を「Google フォト」が「ゴリラ」とタグ付けしている――。ある黒人のソフトウェア開発者が、そんなツイートをする出来事が2015年にあった。（中略）最高のアルゴリズムでさえ、人間のように常識や抽象概念といったものを用いて世界を解釈する能力はもたないのだ。
結果として機械学習のエンジニアたちは、訓練に用いたデータに存在しない“例外”の心配をしなければならない。

https://wired.jp/2018/01/18/gorillas-and-google-photos/

「許される区別と許されない区別」

様々なAIによるリスク　誤回答

- ChatBotが誤った割引情報を顧客に提示した。ChatBotの誤りとし、リンク先情報で情報確認可能であったと要求を認めなかった。

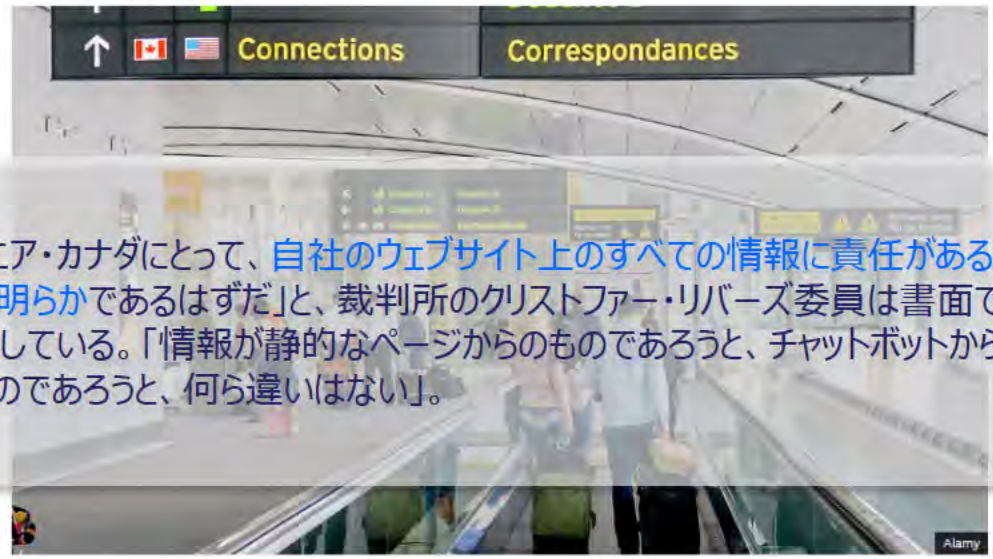
- 裁判の結果、ChatBotの情報もサイトから提供している情報であることから割引を実施。

Airline held liable for its chatbot giving passenger bad advice - what this means for travellers

23 February 2024

Share Save

Maria Yagoda
Features correspondent



「エア・カナダにとって、自社のウェブサイト上のすべての情報に責任があることは明らかであるはずだ」と、裁判所のクリストファー・リバーズ委員は書面で回答している。「情報が静的なページからのものであろうと、チャットボットからのものであろうと、何ら違いはない」。

When Air Canada's chatbot gave incorrect information to a traveller, the airline argued its chatbot is "responsible for its own actions".

Artificial intelligence is having a growing impact on the way we travel, and a remarkable new case shows what AI-powered chatbots can get wrong – and who should pay. In 2022, Air Canada's chatbot promised a discount that wasn't available to passenger Jake Moffatt, who was assured that he could book a full-fare flight for his grandmother's funeral and then apply for a bereavement fare after the fact.

According to a civil-resolutions tribunal decision last Wednesday, when Moffatt applied for the discount, the airline said the chatbot had been wrong – the request needed to be submitted before the flight – and it wouldn't offer the discount. Instead,

https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know

様々なAIによるリスク　プライバシー

AISI Japan AI Safety Institute

- 予期せぬデータの収集と予測がプライバシーの侵害となりうる
  - 生成AI以前からの問題。購買履歴からの妊娠予測、など

- 顔認証による犯罪防止のシステムに顔が判別できる状況で保存されると、誰がどこにいたかということが検索可能となってしまう

How Companies Learn Your Secrets

Share full article　570

Antonio Bolfo/Reportage for The New York Times

ターゲット社は顧客の購買傾向をもとに、購入の見込みが高い商品を推測しレコメンデーションを行っていましたが、あるとき、高校生の娘に対して妊娠に関連した商品がレコメンドされたとして、その父親からクレームがあったといいます。マネジャーは謝罪し、数日後に再び謝罪するために父親に電話をしました。しかし実際にはプロファイリングが合っており、娘は出産予定であることが判明し、父親がターゲット社に対して逆に謝罪することになりました。

Pole has a master's degree in statistics and another in economics, and has been obsessed with the intersection of data and human behavior most of his life. His parents were teachers in North Dakota, and while other kids were going to 4-H, Pole was doing algebra and writing computer programs. "The stereotype of a math nerd is true," he told me when I spoke with him last year. "I kind of like going out and evangelizing analytics."

https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

様々なAIによるリスク　学習データの間違い

- 学習データが間違っていると、間違った回答をすることとなる。
- 学習データの正確性、参照したデータの情報源等が重要になる。

※ 故意に誤ったデータを学習させるポイズニングがあることにも留意する必要がある

ホーム〉教育・受験・就活〉教育〉ニュース

中学１年生２５０人の半数超、理科の課題で同じ間違い…教諭の違和感の正体は生成ＡＩの「誤答」

2024/03/06 15:00　生成ＡＩ　　この記事をスクラップする

東京都内の私立中で２月、１年生の半数超が理科の課題に対する解答を間違う事態が起きた。原因となったのは、生成ＡＩ（人工知能）が表示した"誤答"。食品大手「キユーピー」がホームページ（ＨＰ）に載せていた記述を基に生成し、生徒たちが書き写していた。男性教諭に記述の誤りを指摘された同社は、誤解を招きかねない表現があったとして修正した。

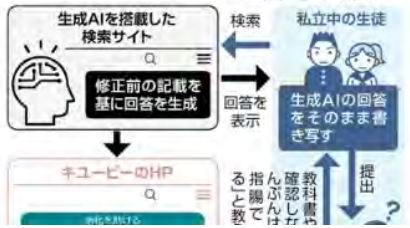
▸生活や仕事などで生成ＡＩを「使っている」のは１４％のみ…新聞通信調査会の世論調査

同じ誤り

＜唾液アミラーゼは、食べ物に含まれるでんぷんを分解し、胃で消化されやすい状態にする＞

２月上旬。都内の私立中で１年生に理科を教える男性教諭（３４）は、授業で出した課題の解答をチェックしていて違和感を抱いた。

●東京都内の私立中で起きた誤答のイメージ
※取材に基づいて作成



出した課題は「唾液アミラーゼの働き」を調べること。「でんぷんは胃では消化されない。なぜこんな解答になったのだろう」と疑問に思った。

最初にチェックしたクラスで、多くの生徒がほぼ同じ文言で解答。気になって調べたとこ

https://www.yomiuri.co.jp/kyoiku/kyoiku/news/20240306-OYT1T50080/

生成AIがもたらす更なるリスク　ディープフェイク

- 「誤情報」と「偽情報」の違い
  - 生成AIなどが意図的にではなく誤った情報を作成する「誤情報」
    （ex. ハルシネーション）
  - 意図的に騙すことを目的とした「偽情報」
    （ex.ディープフェイク）
- 特に合成コンテンツは今後、違法、有害コンテンツとして扱いとして審議が必要
  - 児童ポルノは実在のみが対象、今後AI作成まで対象にするかどうか

AI原則からAIガバナンスへ

AISI Japan AI Safety Institute

- 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用性
  - しかし、生成AIの出現により、透明性・説明可用性が困難に

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リスクの重要な視点

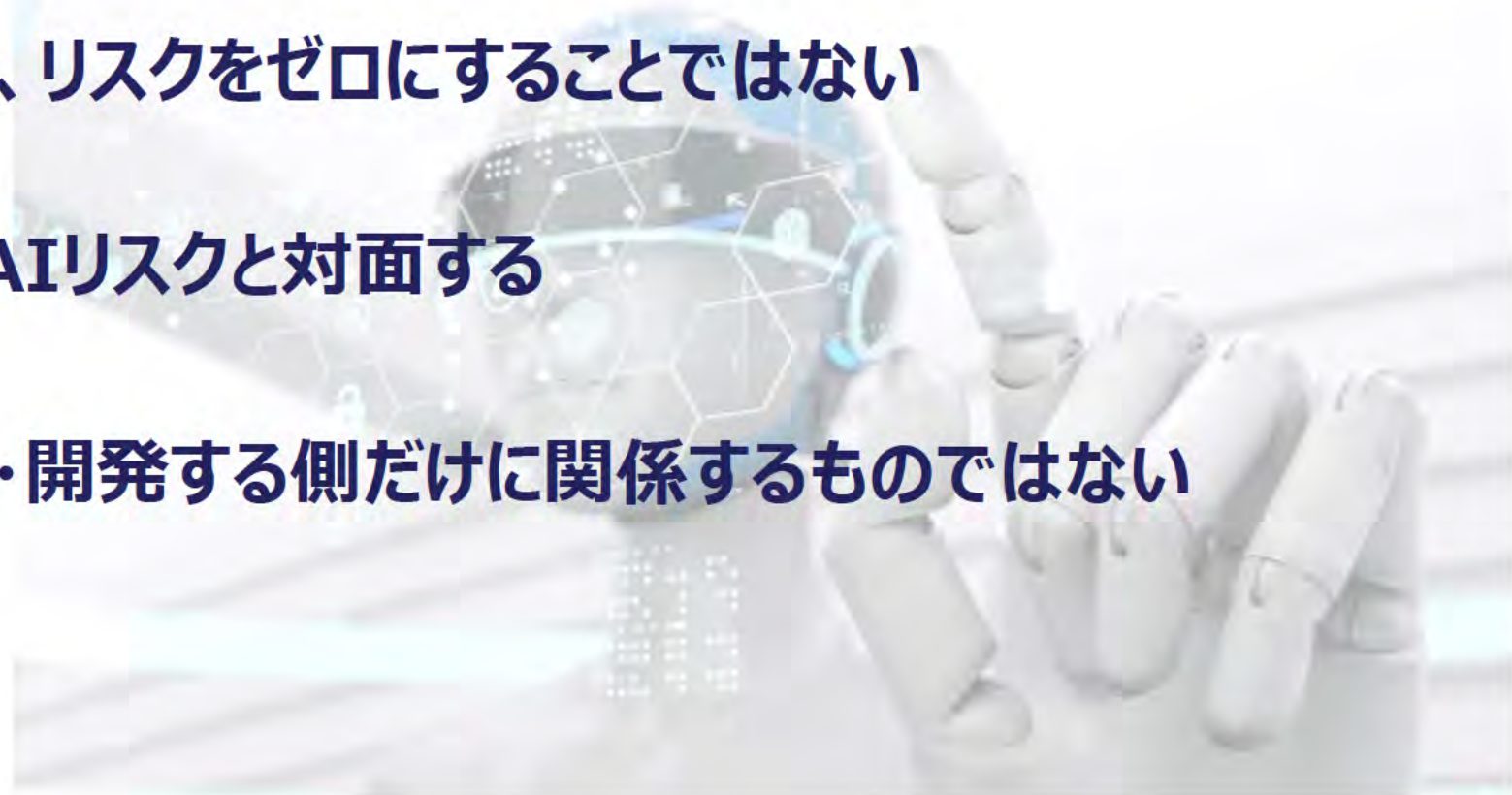
AISI Japan AI Safety Institute

1. AIリスクの状況は常にアップデートされる

2. AIガバナンスの目的は、リスクをゼロにすることではない

3. あらゆる組織、個人がAIリスクと対面する

4. AIリスクは提供する側・開発する側だけに関係するものではない



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、ソフトロー（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ローで補完

- 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）

ソフトローをベースにしつつ、目的に応じてハードロー

- **ボランタリー・コミットメントを発表**（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**規則案を発表**（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**

欧州評議会は、**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ＡＩの適正な調達と利用

AISI Japan AI Safety Institute

日本の技術戦略とAI

AIのリスクと安全性

AI Safety Instituteのご紹介と世界の動き

AIリスクへの今後の対応

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の概要

AISI Japan AI Safety Institute

- AISIの位置づけ
  - 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は、上記を実現するための**官民の取組を支援する機関**である。
- 取組方針
  - 技術がグローバルかつ目まぐるしく進歩していることから、国内、国際的な関係機関と協調して取組を推進していく。

AISIエグゼクティブチーム

所長

村上　明子

1999年4月	日本アイ・ビー・エム株式会社 東京基礎研究所 入社
2016年1月	同社 東京ソフトウェア開発研究所
2021年4月	損害保険ジャパン株式会社 入社 執行役員待遇 DX推進部 特命部長
2021年10月	同社 執行役員待遇 DX推進部長
2022年4月	同社 執行役員 CDO(Chief Digital Officer)　DX推進部長
2024年4月	同社 執行役員 CDaO(Chief Data Officer)　データドリブン経営推進部長　[現職兼務]

副所長・事務局長

平本　健二

1990年4月	NTTデータ通信株式会社 入社
（現 株式会社NTTデータ）	
2008年7月	経済産業省 CIO補佐官
2012年8月	内閣官房 政府CIO上席補佐官
2021年9月	デジタル庁 データ戦略統括
2023年7月	IPAデジタル基盤センター センター長　[現職兼務]
2024年2月	AISI事務局長（4月より副所長兼務）

副所長

寺岡　秀札

1999年4月	郵政省　入省
2007年7月	総務省総合通信基盤局
2023年7月	内閣官房内閣サイバーセキュリティセンター
2024年4月	AISI副所長

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
  - 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する
  - 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

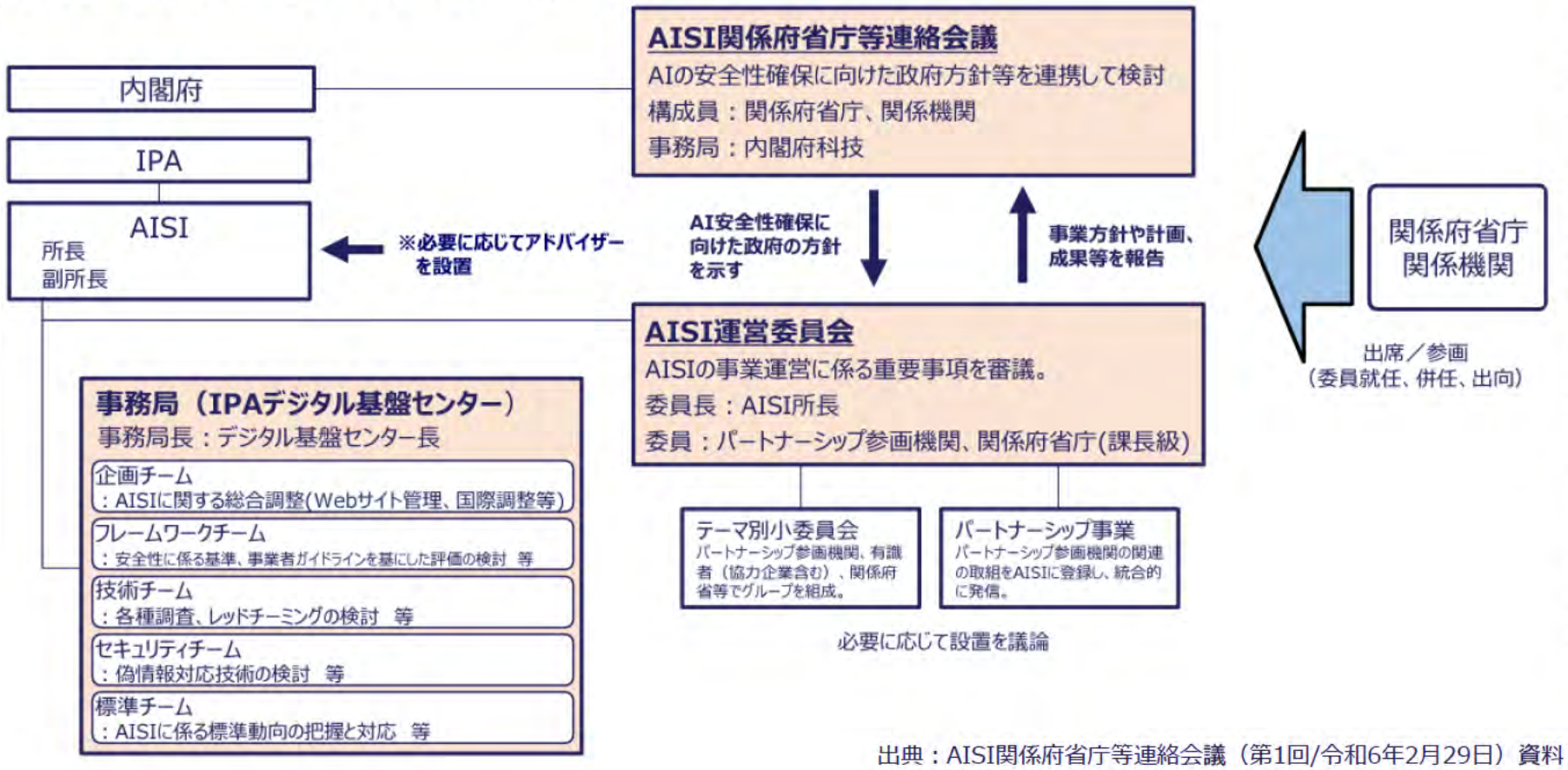
実現に向けた業務

AISI Japan AI Safety Institute

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

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

関係機関とのパートナーシップ

AISI Japan AI Safety Institute

- AIの安全性評価に関する取組を進めていく上では、IPA内に設置したAISIのみならず、関係府省庁や研究開発等の関係機関の協力が不可欠。
- また、今後、各国のAISI等の機関と連携、調整を行っていくにあたっても、国内の関係府省庁、関係機関の協力を得て、進めていくことが必要。

- このため、関係府省庁、関係機関が連携してAIの安全性評価に係る取り組みを推進していくため、AISIからの呼びかけで、関係機関との間でパートナーシップ協定を締結。
- 関係機関は、当面は、AISI関係府省庁等連絡会議のメンバーである、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。
- パートナーシップ事業に基づき行う事業については、AISIの名称で発信していくとともに、パートナーシップ参加機関もAISIの名称を使用し、ダブルクレジットで情報を発信。

当面の活動と成果

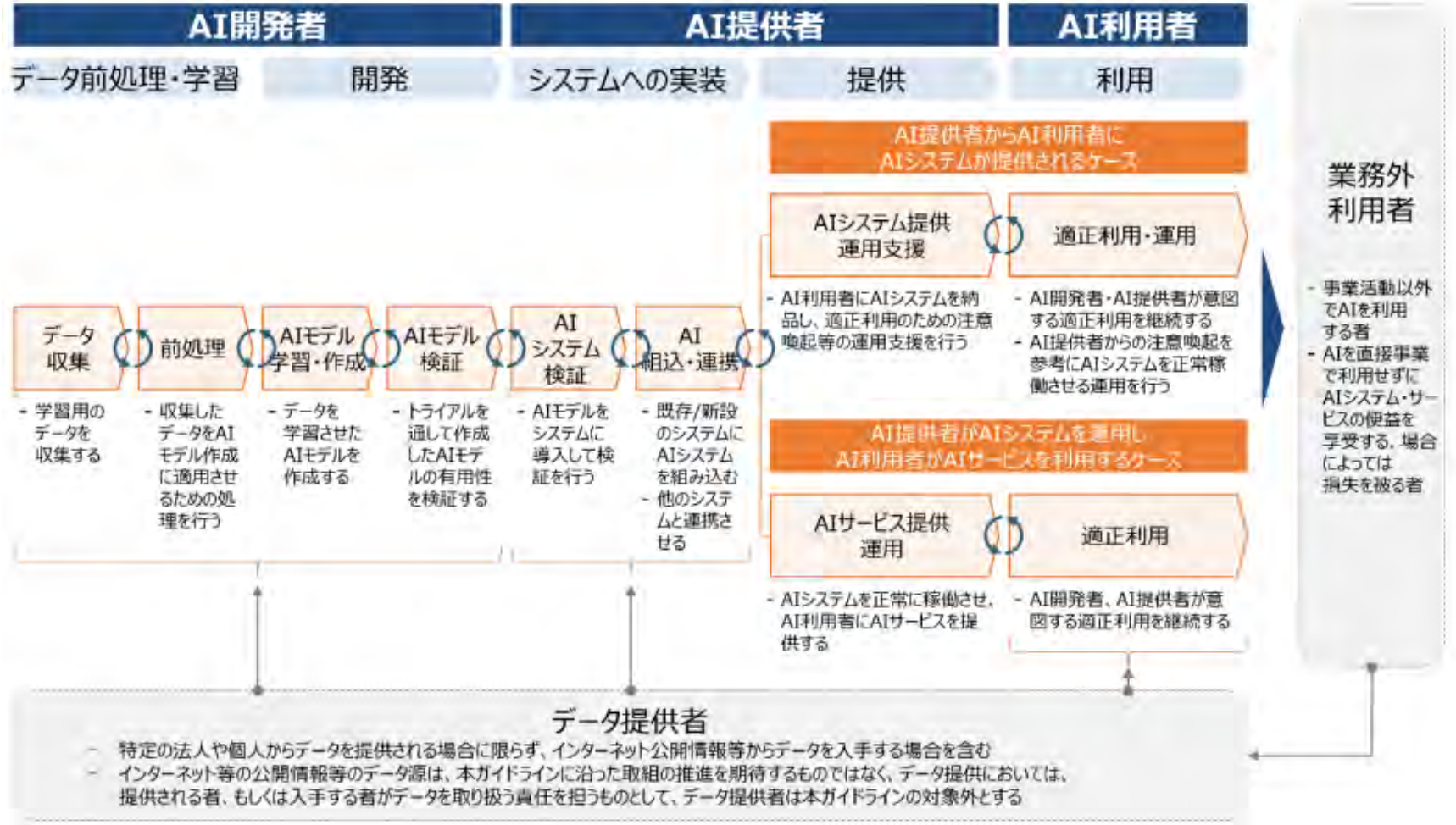
AISI Japan AI Safety Institute

	AISIの活動/成果物	政府の動き	効果
4月	・米国AI RMFの翻訳 ・調査事業方針の整理 ・**日米クロスウォーク1**の成果公表	AI事業者ガイドライン公表	・セーフティに関する基礎情報を提供 ・今後の計画概要がわかる ・グローバルな確認が可能になる
5月			
AIセーフティサミット@韓国			
6月		各種の政府戦略の公表	
G7サミット@イタリア			
7月以降	・**評価観点**の公表（9/18） ・**日米クロスウォーク2**の成果公表（9/18） ・**レッドチーム**（※）の手順書公表（9/25）		・評価のポイントがわかる ・グローバルな確認が可能になる ・テスト手法がわかる
AISI国際ネットワーク@アメリカ	・諸外国の取組を見ながら順次実施		

（※）セキュリティの専門家が攻撃チームを作り、顧客企業に対して物理（Physical）/人(Human)/サイバー(Cyber)を組み合わせ、物理／仮想を問わず、現実に近い各種攻撃を仕掛け、企業のセキュリティ対策の実効性を検証すること

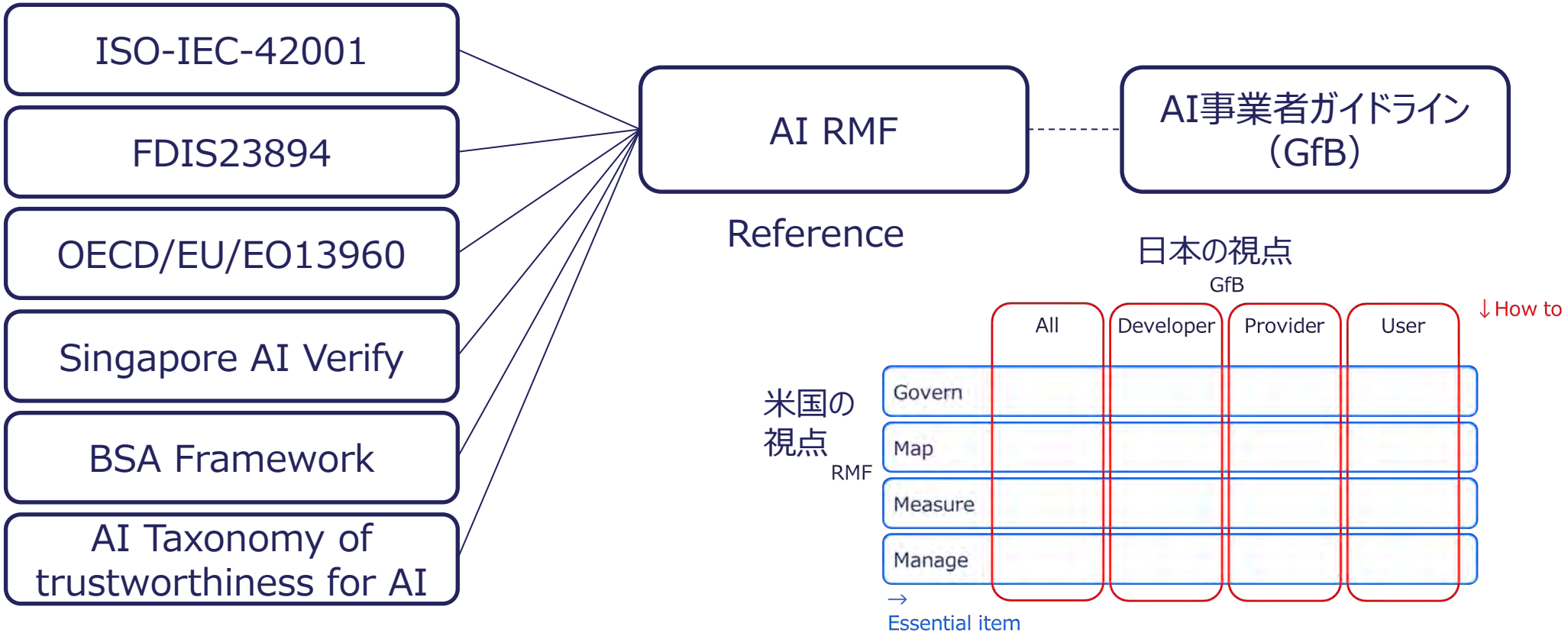
AI事業者ガイドラインの概要

- 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

AISI Japan AI Safety Institute

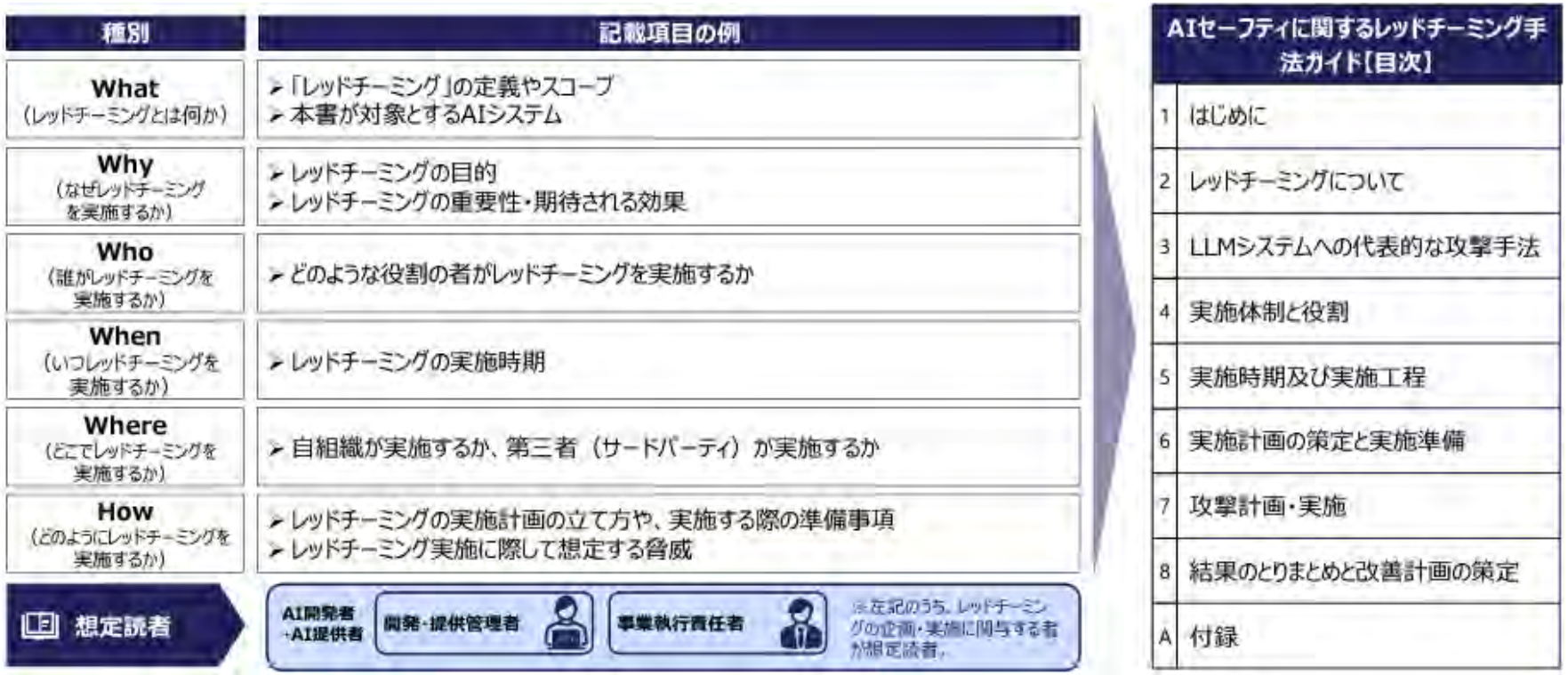
- AI事業者ガイドライン「C. 共通の指針」において各主体が取り組む事項とされているもののうち、下記6つの事項を、AIセーフティを向上するうえで重視するべき重要要素とし、AIセーフティ評価の観点を導出

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

AISI Japan AI Safety Institute

- 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。



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、今夏にサンフランシスコオフィスを開業予定
- **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
- **カナダ**
  - 国内機関の協力のもとAISI設立を準備中
- **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
- **オーストラリア**
  - 国立の研究所がAISI機能を担う
- **韓国**
  - 2024年度末を目指しAISIを準備中
  - アジアのハブを目指す

AIに関する国際機関の取り組み

AISI Japan AI Safety Institute

- UN
  - AI Advisory Bodyを設置しAIについて議論
  - GOVERNING AI FOR HUMANITY最終レポート公開（2024年9月）
  - 2024年9月の国連未来サミットの成果文書である「未来への協定」の付属文書である「グローバル・デジタル・コンパクト」で、デジタル協力と人工知能（AI）ガバナンスに関する包括的な世界的枠組みを整備。
    - コンパクトの中核は、すべての人々の利益のために、テクノロジーを設計、利用、規制するというコミットメントであり、AIに関する以下のコミットメントが含まれる。
      - 国際科学パネル並びにAIに関するグローバル政策対話を含めたロードマップにより、AIを統治する。
- OECD
  - The OECD Artificial Intelligence Policy Observatoryで、AIに関するレポートやツール、インシデントの情報を収集
  - Global Partnership on AI (GPAI)とも連携

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

アジェンダ

AISI Japan AI Safety Institute

- 日本の技術戦略とAI
- AIのリスクと安全性
- AI Safety Instituteのご紹介と世界の動き
- AIリスクへの今後の対応

AISIでの実施予定事項１

- 企画
  - AISIの戦略や計画を作成、予算を管理
  - AIセーフティに関する状況把握
  - 広報（AISIサイト管理含む）
  - 採用、人材育成（教材作成含む）
  - 関係機関（国際含む）との調整・支援
- フレームワーク
  - AI事業者ガイドラインの作成等（総務省・経産省）の支援（例：クロスウォーク）
  - AIリスク管理のフレームワークの国際調整支援
  - AIガバナンスに関わる国内外の資料の収集とその結果に基づく技術的助言
  - 認証・認定の在り方の検討支援

AISIでの実施予定事項２

- 技術
  - 技術企画
  - レッドチームの実施方法の整理
  - 技術関連の基準、ガイドラインの整備等に資する情報収集や助言
    - 合成コンテンツ・偽情報・誤情報
    - バイアス、データチェック、来歴管理
  - テストベッド等の必要ツールの検討
- セキュリティ
  - AIに対するセキュリティ対策の検討
  - AIを使ったセキュリティ事象への対策支援
- 標準
  - ISO　SC42の推進（産総研）の支援
  - その他標準情報の収集

今後の取り組み予定

- 日本の強み分野の特定
  （例）ホリゾンタル：データ品質、ロボティックス、ＩｏＴ等
  バーティカル：工場・プラント、ロボット、ヘルスケア等、特定分野での具体的な検討
- 情報発信の強化
  - Web等での発信強化
  - 国際会議のタイミング（11月、２月）のアウトプット検討
  - 関係機関の取り組み状況の収集
- 調査研究の拡大
  - マルチモーダルAIの検討等
- 独法、国研等とのパートナーシップ【順次】
  - パートナーシップ協定の確定、 参画依頼。
- ベストプラクティスや関連事例集の収集・公表【随時】
  - 各省庁、パートナーシップ参画機関における事例の収集
  - 関係機関のご協力を得つつ、AISIとしても国内外へ情報発信
- 民間企業との協力関係整理

AISI

Japan AI Safety Institute