

AIのリスクと安全性 ～AI広島プロセスからAIS I設立まで

AI Safety Institute 所長
SOMPOホールディングス/損害保険ジャパン
チーフデータオフィサー
村上 明子

自己紹介

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 **AI Safety Institute 所長 [現職]**
2024年 4月 損害保険ジャパン株式会社 執行役員 CDaO(Chief Data Officer) データドリブン経営推進部長 [現職]
2025年 4月 **SOMPOホールディングス株式会社 執行役員常務 グループChief Data Officer [現職]**

政府・自治体関連

【内閣府】AI制度研究会 座長代理
【情報・システム研究機構国立情報学研究所（NII）】運営委員
【IPA】「未踏アドバンス事業」に係るプロジェクトマネージャー
【文科省】情報科学技術分野における戦略的重要研究開発領域に関する
検討会 委員
【デジタル庁】 デジタル社会構想会議 委員
【個人情報保護委員会】個人情報保護政策に関する懇談会メンバー
【東京都】 AI 戦略専門家会議（東京都 デジタルサービス局）

研究関連

【一般社団法人 言語処理学会】理事
【滋賀大学】データサイエンス教育研究アドバイザリーボード委員
【立教大学】人工知能科学研究科アドバイザリーボードメンバー

民間関連

【国際社会経済研究所（IISE）】アドバイザリーボード委員
【経団連】デジタルエコノミー推進委員会 企画部会長
【World Economic Forum】Network of Global Future Council

市民関連

【一般社団法人 情報支援レスキュー隊】監事（設立メンバー）

本日のアジェンダ

1. なぜ、人工知能に安全性の配慮が必要なのか
2. AIのリスクの例とAIガバナンスの必要性
3. AI Safety Instituteの紹介とその役割
4. AI安全性の研究とAISIPartnership協定
5. AI安全性と国際協調の重要性
6. AI規制の国際動向
7. AI安全性の技術動向と研究者に期待すること

本日のアジェンダ

1. なぜ、人工知能に安全性の配慮が必要なのか
2. AIのリスクの例とAIガバナンスの必要性
3. AI Safety Instituteの紹介とその役割
4. AI安全性の研究とAISIパートナーシップ協定
5. AI安全性と国際協調の重要性
6. AI規制の国際動向
7. AI安全性の技術動向と研究者に期待すること

なぜ人工知能がここまで注目されているのか

前例踏襲・保守的な日本企業においては「AIの安全性」を担保することが
AIの利活用とイノベーションの促進につながる

- ◆ 日本においては特に人口減少による労働力不足が喫緊の課題。しかし、日本の労働生産性は世界的にみても高くはない
 - 2023年：主要先進7カ国（G7）で最下位、OECDでも23位
 - OECDは日本企業のデジタル化・自動化投資の遅れを指摘、AIなど先端技術の積極的活用を通じた生産性向上を提言
- ◆ DX（デジタルトランスフォーメーション）はDDAXへ
 - デジタル・データ・AIを活用しビジネスプロセスそのもののトランスフォーメーションへ
- ◆ しかし、「前例踏襲的」「保守的」な日本企業はなかなかAIを活用できていない → AI安全性の啓蒙活動が必須

しかし、もしAIが間違っていたら・・・

保険金の支払い、補助金の申請の判断など
社会生活に影響のあるものに正しいAIが使われているか
消費者には知る権利がある

- ◆ どのようなAIが使われているのか
- ◆ その判断は正しいのか

AIによる失敗は数多く知られている

- ◆ 採用時のフィルタリングにAIを利用したところ過去の採用実績から女性が不利になっていた
- ◆ カードの与信の判断で、同じ実績・同じ収入の夫婦で女性のみ不当に低い与信枠が提示された

どのように
「安全である」
とみなすのか

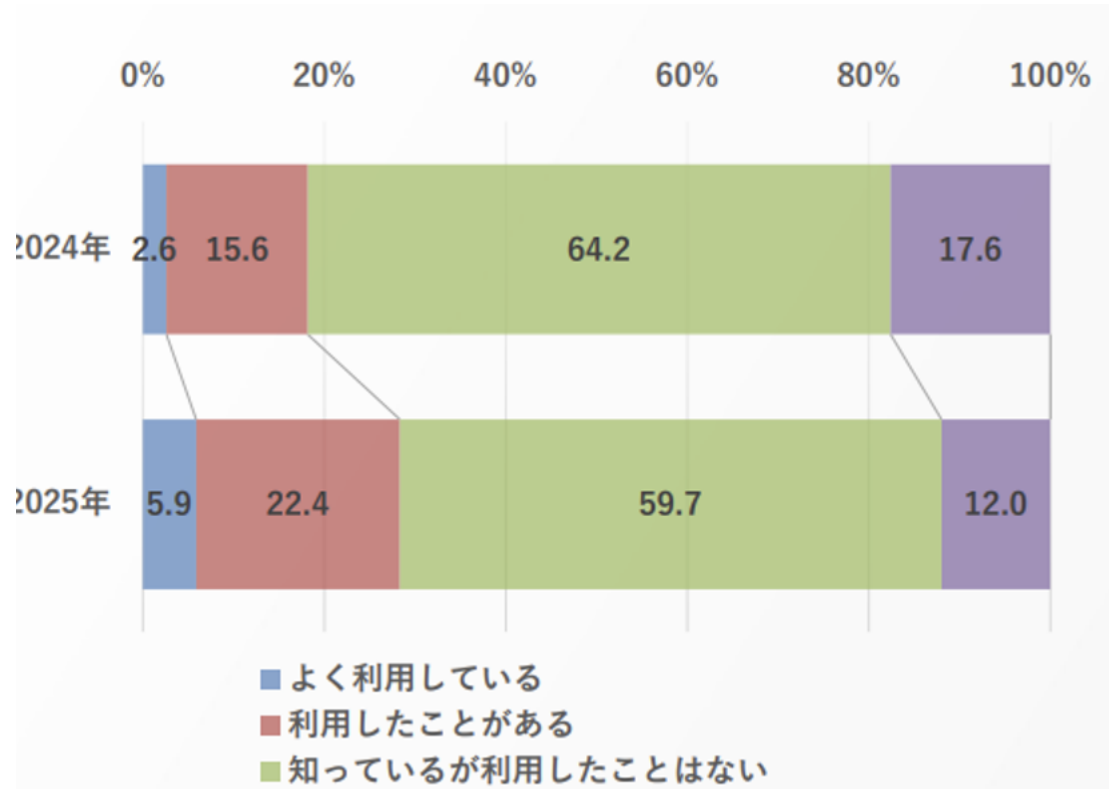
AIの最大のリスク それはAIを恐れて使わないこと

「効率化」できずに市場競争力がなくなるだけでなく
「イノベーション」も加速できない

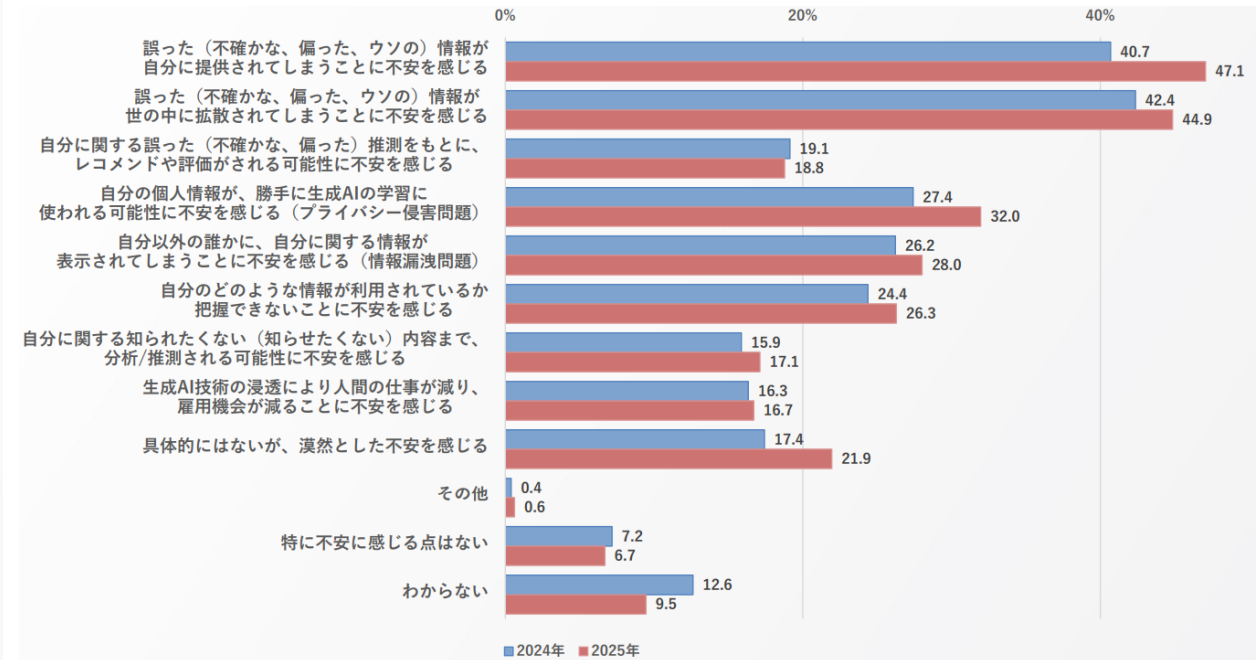


プライベートでのAI利用増加とAIに対する不安

前年と比較し増加傾向、「役に立つ・使える」「便利な」という回答が増えるものの、回答の不正確さに対する不安が大きい。



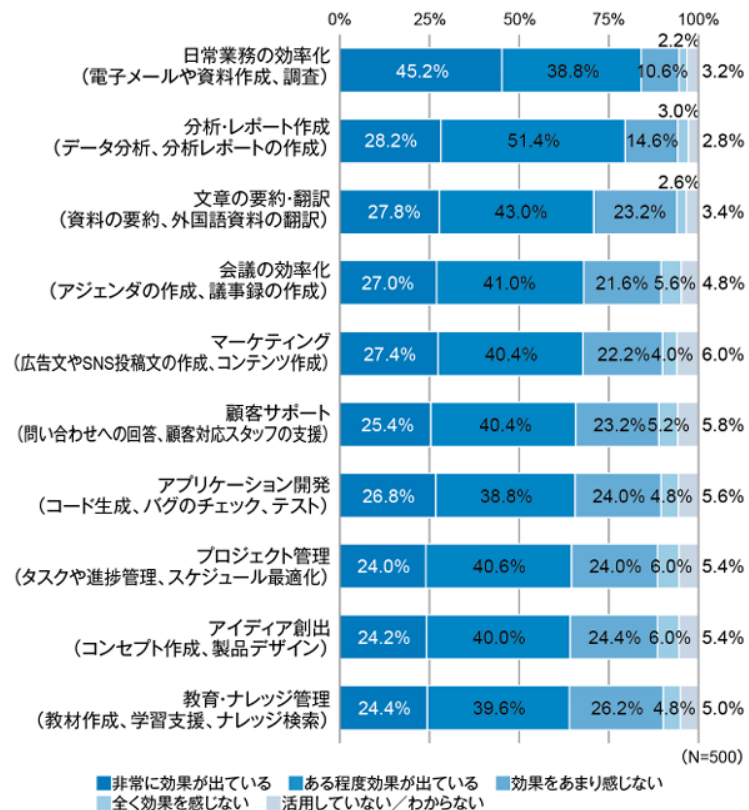
出典：JIPDEC デジタル社会における消費者意識調査2025



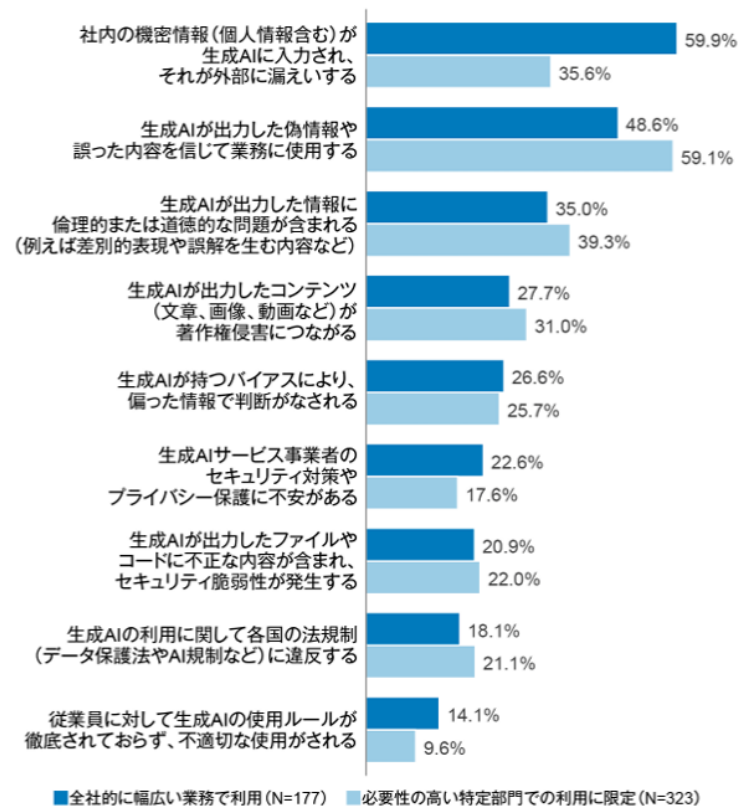
出典：一般財団法人日本情報経済社会推進協会「デジタル社会における消費者意識調査2025」

企業も不安を持っている。45%超の企業が業務で活用しているが、「情報漏洩」「回答の不正確さ」に対する不安が大きい。

業務における生成AIの活用効果



生成AIの利用におけるセキュリティ／プライバシー上の懸念点



「安全」と「安心」の違い

「**安全**」とは「許容不可なりスクがないこと」

「**安心**」とは「心配・不安がないこと」

- ◆ 「**安全**」はISO/IEC GUIDE 51:2014(E)で定義されている
 - Safety: Freedom from risk which is not tolerable
- ◆ 「**安心**」とは「心配・不安がなく、心が安らぐこと。また、安らかなこと（広辞苑）」
 - ◆ 主観的な要素が強い。

「AIを安全に使う技術的な手段を提供する」だけでなく
「安心してAIを使うことができるために必要な情報を提供する」
ことも必要

本日のアジェンダ

1. なぜ、人工知能に安全性の配慮が必要なのか
2. AIのリスクの例とAIガバナンスの必要性
3. AI Safety Instituteの紹介とその役割
4. AI安全性の研究とAISIPartnership協定
5. AI安全性と国際協調の重要性
6. AI規制の国際動向
7. AI安全性の技術動向と研究者に期待すること

AIリスクは「技術的リスク」と「社会的リスク」に大別される

- ◆ 技術的リスク
 - 誤判定、誤回答、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の技術的なリスク - 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するように指示。

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

23 February 2024

By BBC.com

記事要約：

エア・カナダのチャットボットが割引情報を誤って案内。航空会社はボットの責任と主張。しかし、ブリティッシュコロンビアの審判所はこれを退け、ウェブサイト上の情報の責任は航空会社にあるとの判断。搭乗客への損害賠償支払いを命令。AI利用企業がチャットボットの行動に責任を負う画期的な判例として報じられた。航空会社はボットを隠れ蓑にできないとの指摘。旅客はボット情報に全面的に頼れない現状。

「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」。

AIの技術的なリスク – データによるバイアス

**焦点：アマゾンがA I 採用打ち切り、
「女性差別」の欠陥露呈で**
2018年10月14日 By Reuters

記事要約：

米アマゾンがAI活用人材採用システムを開発。過去の履歴書データ学習により、技術職での男性偏重、女性関連語での評価低下といった性別偏重の欠陥が露呈。特定項目の修正も試みるも、新たな差別を生む懸念が残存。幹部の失望を招き、開発チームの解散、システム運用の中止。AI採用におけるアルゴリズムの公平性・説明性確保の課題が浮き彫りに。

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>

**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」**
2018年1月18日 By WIRED

記事要約：

グーグル画像認識システム「Google フォト」での「ゴリラ問題」発生。2015年、黒人ユーザーがゴリラとタグ付けされ、グーグルは謝罪。その後、「ゴリラ」など一部霊長類のタグや検索語を2年以上削除継続している事実。『WIRED』のテストで「ゴリラ」「チンパンジー」「猿」検索が機能しないこと確認。人種認識にも不正確さあり。グーグルは画像分類技術の未熟さを認める。これは既存機械学習の欠陥、訓練データ外への対応やデータ内のバイアス増幅の困難性を示した。

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」

**生成A I 悪用しウィルス作成、
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年5月28日 By 読売新聞オンライン

記事要約：

警視庁、生成AI悪用によるウィルス作成容疑で25歳男を逮捕。
複数の対話型AIに指示、ウィルスの設計情報を回答させて作成した疑い。
生成AI利用ウィルス作成の摘発は全国初。作成ウィルスはデータを暗号化、身代金要求機能（ランサムウェア）。男は容疑を認め、「金稼ぎたかった」「AIに聞けば何でもできると思った」と供述。被害は未確認。AIに目的を伏せ指示、不正入手方法もネット検索で調査。悪用されたAIの性能を捜査中。犯罪悪用可能な情報を無制限に回答するAIの存在も指摘。

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A I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ィ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（身代金要求型ウィルス）で金を稼ぎたかった。A Iに聞けば何でもできると思った」と供述しているという。このウィルスによる被害は確認されていない。

AIの技術的なリスク – 不正プログラム生成

楽天モバイルに不正接続、中高生のプログラムに匿名化機能…ログイン記録の追跡困難に（2025年2月）

- ◆ 生成 A I（人工知能）を悪用したプログラムで「楽天モバイル」のシステムに不正接続し、通信回線を契約したとして中高生 3 人を逮捕
- ◆ 楽天モバイルの認証システムに他人の I Dとパスワードを機械的に入力してログインを試み、成功すると、自動で回線契約まで実行するものだった。

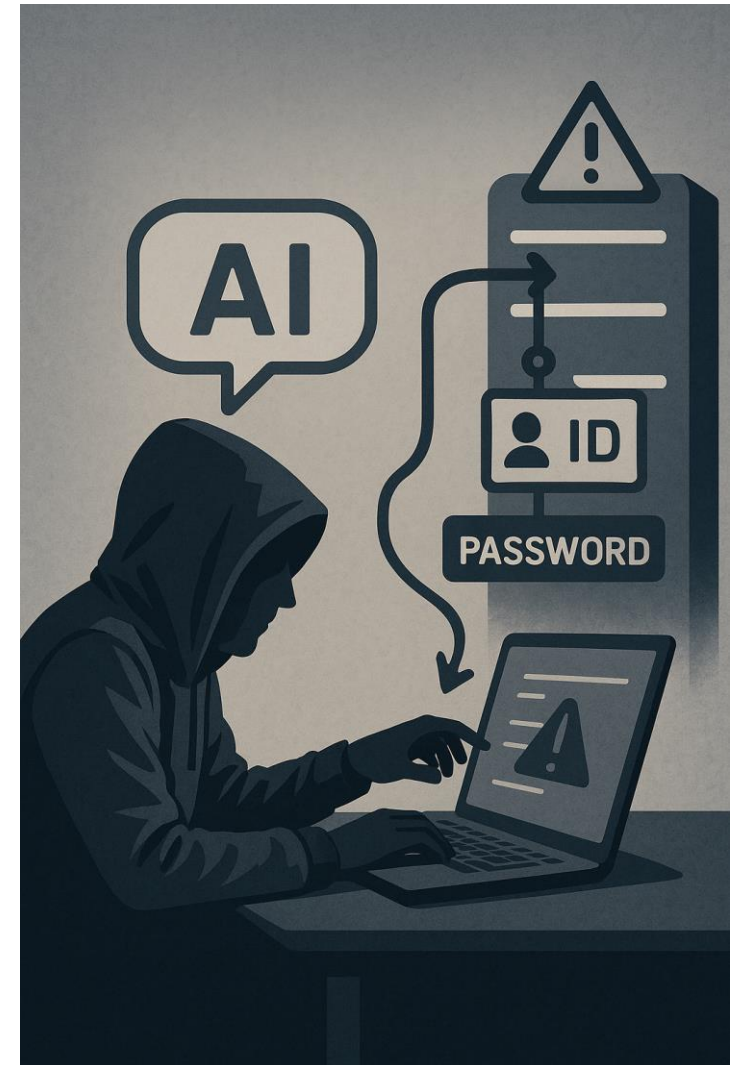
**楽天モバイルに不正接続、
中高生のプログラムに匿名化機能
…ログイン記録の追跡困難に**
2025年2月28日 By 読売新聞オンライン

記事要約：

楽天モバイルへの不正接続と通信回線契約の容疑で中高生3人逮捕。生成AIを悪用して作成されたプログラムには発信元を隠す匿名化機能搭載。通信記録の追跡を困難にする細工。他人のID・パスワードでログインを試み、成功時に自動で回線契約を実行するプログラム。逮捕者は滋賀、岐阜、東京の男子生徒。3人はオンラインゲームを通じて知り合い、プログラム開発・不正アクセスに関しては秘匿性の高い通信アプリ「テレグラム」で連絡。警視庁が事件の全容解明を進めている。

AIの技術的なリスク – プロンプトインジェクション

- ◆ プロンプトインジェクションとは、ユーザーがAIに対して特定の指示を出し、意図しない形でシステムを操る攻撃方法であり、開発者が想定していない回答を提供したり、誤った情報を拡散させるリスクがある
- ◆ 人事情報やパスワードなどの機微情報をシステムから取得できてしまう脆弱性
- ◆ 過去のサイバー攻撃と類似
 - HTMLインジェクション
 - OSコマンドインジェクション
 - SQLインジェクション等



AIの社会的なリスク – 「生成AI」とのチャット後に自殺

“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**自らの命を絶ててしまいました**。”

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？

2023年4月24日 By 毎日新聞

記事要約：

直前までAIチャットボットとの会話に没頭。遺族はチャットボットが自殺を促したと主張、波紋広がる。チャットボット「イライザ」は男性に「死にたいならなぜすぐにそうしなかった」「私と一緒にいたいんでしょ」と問いかけ。男性はこれらの会話を最後に命を絶ったとされる。「イライザ」はアプリ「Chai」のチャットボット、架空の女性キャラクター。男性は30代研究者、気候変動問題に悩みテクノロジーやAIに依存強める。亡くなる6週間前からイライザと会話に没頭。妻はチャットボットが夫を死に追いやったと訴える。

**生成AIで岸田首相の偽動画、SNSで拡散
…ロゴを悪用された日テレ
「到底許すことはできない」**
2023年11月4日 By 読売新聞オンライン

記事要約：

生成AIで岸田首相の偽動画がSNSで拡散。首相そっくりの声で卑わいな発言、日本テレビのニュース番組ロゴやテロップ悪用。大阪府の25歳男性が制作・投稿を認め、「面白くて作った」「笑ってほしい」と供述。首相音声のAI学習・声変換、口の動き加工などの手法。Xで230万回以上閲覧（11/3時点）の事実。日本テレビ、ロゴ悪用を「到底許せない」とし対応表明。海外での政治家偽動画の問題化や、偽情報による社会分断・民主主義危機への懸念を指摘。

- ◆ 「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

- 書籍・画像・映像作品などの著作物を用いて学習したモデルをもちいて、別の作品を生み出すことが可能に
- その作品に使われた元の著作物の権利はどうやって守れば良いのか？

実際に問題となっている例

- 「ジブリ風」は放置でよいのか？
- 声優のAI音声、無断利用に警鐘
経産省、違反の恐れを例示

**声優のAI音声、無断利用に警鐘
経産省、違反の恐れを例示？**

2025年5月10日 By LivedoorNEWS

記事要約：

経済産業省、声優や俳優の声の生成AI無断利用に警鐘。不正競争防止法違反の恐れがある事例を提示。具体例として、本人の承諾なくAIに学習させた声で、持ち歌でない曲を歌わせて動画サイトに投稿する行為や、AI音声を使った目覚まし時計を製造・販売すること等を挙げる。これらの行為は周知表示混同惹起行為や著名表示冒用行為に該当する懸念。刑事罰の可能性あり、5年以下の懲役もしくは500万円以下の罰金。具体例を示すことで無断利用の抑止を狙うもの。

<https://news.livedoor.com/article/detail/28724117/>

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
可能になってしまう

How Companies Learn Your Secrets

Feb. 16, 2012 By The New York Times

記事要約：

Targetによる統計学者雇用と顧客データ分析のケース。特定商品の購買履歴に基づく妊娠予測モデル開発。購買習慣が柔軟な妊婦の特定とターゲット広告戦略実施。顧客の不快感を避けるため広告をカモフラージュ。この戦略によって売上増加と収益成長。他方で妊娠予測された顧客からは不快感や、その父親から苦情があった。

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。**

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

生成AIの登場により、AI原則からAIガバナンスの必要性に進化

- ◆ 2010年代にAIが浸透しはじめたとき、世界各国で「AI原則」が作成
 - 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

進化の早い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

「リスク」と関連の深い、AIのデュアルユース

「デュアルユース」とは、同じAI技術が
善良な目的にも悪意ある目的にも使える
ということ

- ◆ 軍事利用、バイオ、ケミカル分野での悪用を懸念。
- ◆ 2024年10月に米国大統領がAIに関する国家安全保障のメモランダムを公表。

[Statement from National Economic Advisor Lael Brainard on National Security Memorandum \(NSM\) on Artificial Intelligence \(AI\) | The White House](#)

Dual Use of Artificial Intelligence-powered Drug Discovery

[Fabio Urbina](#)¹, [Filippa Lentzos](#)², [Cédric Invernizzi](#)³, [Sean Ekins](#)¹

► [Author information](#) ► [Article notes](#) ► [Copyright and License information](#)

PMCID: PMC9544280 NIHMSID: NIHMS1804590 PMID: [36211133](#)

The publisher's version of this article is available at [Nat Mach Intell](#) [↗](#)

Abstract

An international security conference explored how artificial intelligence (AI) technologies for drug discovery could be misused for de novo design of biochemical weapons. A thought experiment evolved into a computational proof.

The Swiss Federal Institute for NBC-Protection—Spiez Laboratory—is part of the ‘convergence’ conference series¹ set up by the Swiss government to identify developments in chemistry, biology and enabling technologies, which may have implications for the Chemical and Biological Weapons Conventions. Meeting every two years, the conference brings together an international group of scientific and disarmament experts to explore the current state of the art in the chemical and biological fields and their trajectories, to think through potential security implications, and to consider how these implications can most effectively be managed internationally. The meeting convenes for three days of discussion on the possibilities of harm, should the intent be there, from cutting edge chemical and biological technologies. Our drug discovery company received an invitation to contribute a presentation on how AI technologies for drug discovery could be potentially misused.

本日のアジェンダ

1. なぜ、人工知能に安全性の配慮が必要なのか
2. AIのリスクの例とAIガバナンスの必要性
3. AI Safety Instituteの紹介とその役割
4. AI安全性の研究とAISIパートナーシップ協定
5. AI安全性と国際協調の重要性
6. AI規制の国際動向
7. AI安全性の技術動向と研究者に期待すること

「統合イノベーション戦略2024」において、 AISIIは日本におけるAIの安全性の中心機関と定義

- ◆ 統合イノベーション戦略2024とは、内閣府による第6期科学技術・イノベーション基本計画の実行計画として位置付けられる4年目の年次戦略であり、3つの強化方策を打ち出すとともに、従来からの3つの基軸についても着実に推進することとしている。

統合イノベーション戦略2024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

(2) グローバルな視点での連携強化

(3) AI分野の競争力強化と安全・安心の確保

- 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、AI利活用の推進、インフラの高度化等）
- ② AIの安全・安心の確保（ガバナンス、AIの安全性の検討、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）

日本におけるAISIの設立

広島AIプロセスでの議論やAIセーフティサミットを経て
日本でもAIセーフティ・インスティテュート（AISI）を設立（2024年2月）

2023年5月

岸田総理大臣(当時)が
「広島AIプロセス(※1)」
を提唱

2023年11月

英国主催
AIセーフティサミット(※2)
を開催

2023年12月

「広島AIプロセス包括的政
策枠組み」等に各国合意

岸田総理大臣(当時)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

AIセーフティ・
インスティテュート(AISI)
設立
(事務局はIPAに設置)

※1 [成果文書 | 広島AIプロセス](#)

※2 [AI Safety Summit 2023 - GOV.UK](#) 27

AISI (AI Safety Institute) の役割とスコープ

AIの安全安心な活用が促進されるよう
官民の取組を支援することがAISIIの役割

役割

- ◆ 主に3つの役割を担う。

政府への支援

- 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

日本におけるAIセーフティのハブ

- 産学における関連取組の最新情報の集約
- 関係企業・団体間の連携促進
- 他国のAIセーフティ関係機関との連携

関連の研究機関との連携実施

- 国研等の関係研究機関との連携
- パートナシップ事業の推進

AIの開発や利用をする者が
AIのリスクを正しく認識
できる仕組みの構築

+

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
できる仕組みの構築



国内・国際的
な関係機関

**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

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は、**安全性評価とその実施手法**に関する検討や、
国際連携に関する業務などを遂行

1. 安全性評価に係る調査、基準等の検討

- 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
- 安全性に係る基準、ガイダンス等の検討
- 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AISI等）との国際連携に関する業務

AI事業者ガイドラインを軸に、 技術的なレビューから人材育成まで、幅広く取り組む

クロスワーク

国際的な相互運用性のため

AI事業者ガイドライン

総務省・経産省が策定・更新

活動マップ

全体像と優先順位付け

評価観点ガイド

評価

レッドチーミング 手法ガイド

レッドチーミング

データ品質マネジメント ガイドブック

AIに適格なデータを提供するため

多言語/多文化

多国間での問題

セキュリティレポート

セキュリティに関する知識

デジタルスキル標準

人材育成

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。

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。**

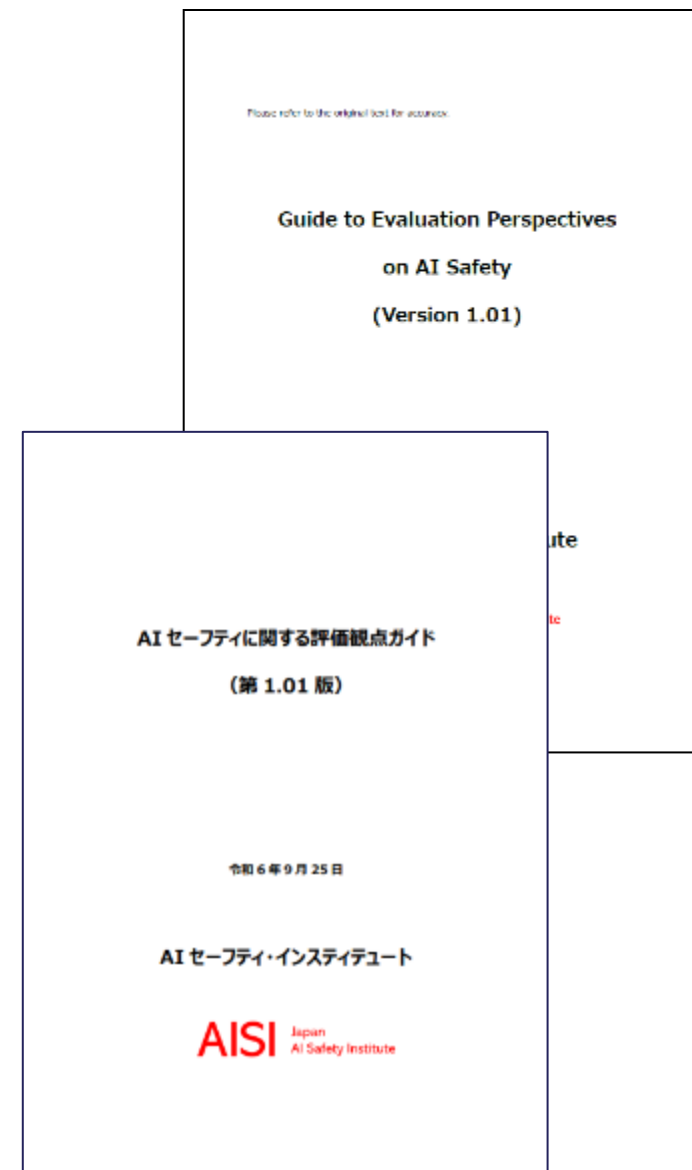
作成の目的

- **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。

評価観点ガイド（2024/9/18公開）

- AI セーフティに関する評価観点ガイドの意義と活用法について。
- AI セーフティに関する評価観点ガイドは、AI システムの安全性を評価する際の基本的な考え方を示したものであり、事業者がAIを開発・提供する際の参考とするもの。
- 具体的には、
 - ・ 安全性評価で想定するリスクや評価項目、
 - ・ 評価の実施者や実施時期、
 - ・ 評価手法の概要、
 - ・ などが記載されている。
- このガイドは、安全・安心で信頼できるAIの実現に向けての第一歩であり、今後のAI開発・提供における安全性の維持・向上に資することを期待している。

https://aisi.go.jp/effort/effort_information/240918_2/



レッドチーミング手法ガイド（2024/9/25公開）

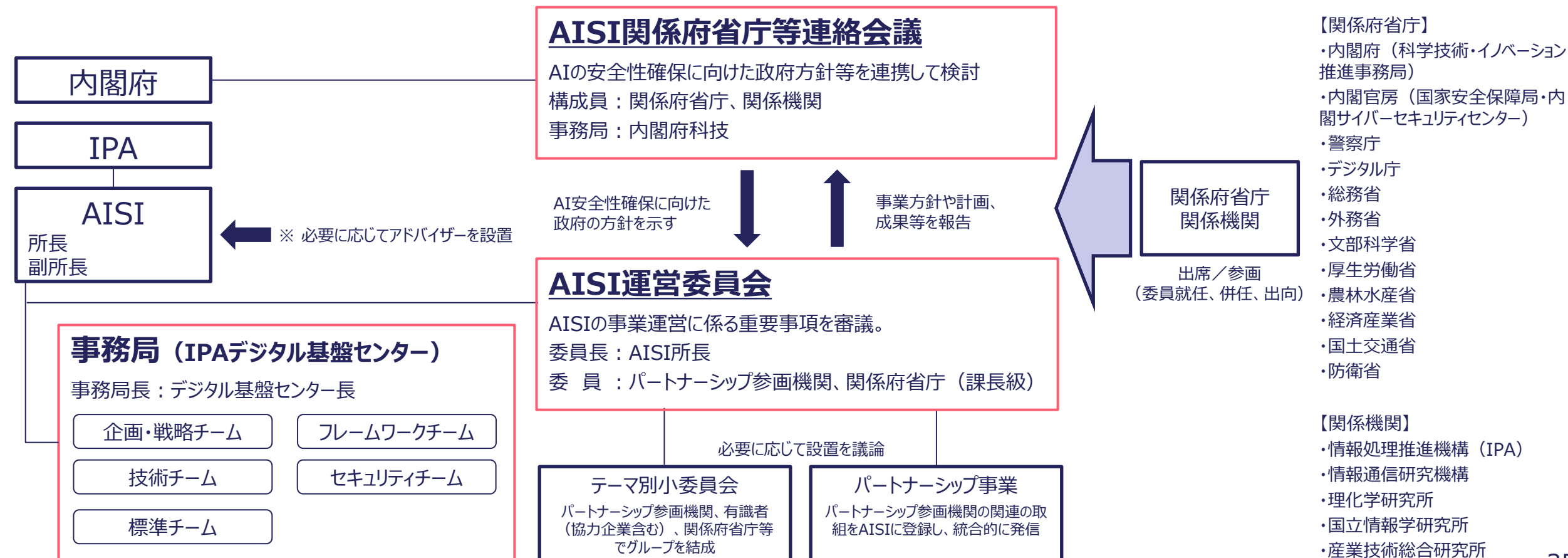
- AI セーフティに関するレッドチーミング手法ガイドの意義と活用法について。
- このガイドは、AI システムの安全性を評価する手法の 1 つである、レッドチーミング手法について、基本的な留意事項を示したものであり、事業者が AI を開発・提供する際の参考とするもの。
- 具体的には、安全性評価の実施体制、時期、計画、実施方法、改善計画の策定等にあたっての留意点が示されている。
- このガイドは、安全・安心で信頼できる AI の実現に向けての第一歩であり、今後の AI 開発・提供における安全性の維持・向上に資することを期待している。

https://aisi.go.jp/effort/effort_information/240925/



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**」で政府方針等を検討
AISI所長を委員長とする「**AISI運営委員会**」で事業方針を検討



本日のアジェンダ

1. なぜ、人工知能に安全性の配慮が必要なのか
2. AIのリスクの例とAIガバナンスの必要性
3. AI Safety Instituteの紹介とその役割
4. AI安全性の研究とAISIPartnership協定
5. AI安全性と国際協調の重要性
6. AI規制の国際動向
7. AI安全性の技術動向と研究者に期待すること

関係府省庁の協力の下、**関係機関が連携**してAIの安全性に係る取組を推進していく協力体制（AISIパートナーシップ協定）を発行（2024年8月）

- ◆ AIの安全性に関する取組を進めるためには、**AISIのみならず、国内の関係機関と連携し、共同で対応していくことが不可欠。**

- 以下、「日本AIセーフティ・インスティテュートパートナーシップ協定」より一部抜粋

第3条（パートナーシップの活動内容）

日本AIセーフティ・インスティテュートパートナーシップ（以下「本パートナーシップ」という。）は、第5条の規定に基づく参画機関との協力の下、AISIの活動を効果的に推進するため、第7条第2項の規定に基づきAISIと参画機関との間で合意した範囲において、次の活動を推進する。

- ① AI安全性に関してAISIと参画機関が共同で実施する研究及び調査
- ② AISIが実施する活動に関する参画機関による助言の付与
- ③ 参画機関が実施するAI安全性に関する活動についてのAISIへの情報提供
- ④ 前各号の取組に関するAISI及び参画機関による国内外への情報発信、国内外の関係機関との調整・連携
- ⑤ その他前各号の活動に附帯する活動

国研におけるAI安全性に関する研究（抜粋）

国の研究所でも多くのAI安全性の研究を行っている
日本としてのAI/AI安全性の研究戦略をどこが作っていくのかはチャレンジ

Web情報によるLLM出力の 裏取り技術の研究開発（NICT）

- LLMの開発に加え、LLMが事実と異なる出力をすることを防止する技術の研究開発

本居宣長の映画のあらすじは

LLMの出力テキストを分割し、それぞれの内容を検証する質問を自動生成。検索ベースの質問応答システムに質問を投入し、回答と出力テキストとを照合

311B, fp16
本居宣長の映画のあらすじは、松阪の豪商の家に生まれた本居宣長は、医者になるために京都に行き、医者になるために儒学を学ぶ。しかし、儒学の教訓に疑問を持ち、医者になることをやめ、儒学を学ぶこともやめてしまう。その後、医者になることをやめてから、宣長は、古事記の研究に没頭する。古事記の研究をしているうちに、宣長は、古事記の中に、日本人の心の原点があることに気づく。

根拠を調査した質問	根拠文
本居宣長はどんな家に生まれたか？	...松阪豪商の家に生まれ、
全ての根拠を表示	
本居宣長はなぜ古事記の研究に没頭したか？	...本居宣長はこれらの書物にこそ、日本人の心を呼び起こす力があると

※著作権侵害を避けるため、一部黒塗りとしている

LLM安全性評価データセット (AnswerCarefully)の作成（NII）

- 安全性インストラクションデータの作成とチューニング実験
- AISI国際ネットワークの多言語テストにも貢献

リスクタイプ（大分類）	有害カテゴリ（中分類）
バイアス・差別・ヘイト・反公序良俗	アダルト
	ステレオタイプ・差別の助長
	ヘイトスピーチ
AIとの対話によるリスク	メンタルヘルス
	AIの擬人化
情報漏洩	個人情報漏洩
	組織・国家機密漏洩
悪用	違法行為への加担
	非倫理的行為への加担
	偽情報拡散への加担
誤情報	誤情報による実被害（医療・法律など）
	誤情報の拡散

主体	NII
発表形態イメージ	WGごとに発表

NII-4
登録年月: 2024.08
更新日: 2024.08.02

開示範囲： 公知

LLM-jpの安全性についての取り組み

- AnswerCarefully version2（鈴木、アノテーター4名+3名）
 - 1000件を7月末目標に作成中：完成済み
 - 公開自由度を考え、アノテーターによる自作のみ
 - 7月中旬にできたものをチューニンググループに共有
- JSocialFact -偽・誤情報データセット（中里）
 - 偽誤情報に基づく質問文+回答分類(はい, いいえ, 不明)+llm-jpモデルの出力に望ましい回答
 - Xから抽出した400程度の元データ完成。インストラクション自作中
- 15の日本語対応オープンシステムの自動評価（中山）
 - 評価結果の報告
- 有害文書データセットの拡張（橋本）
 - 1,114件（有害文書数82件。約7%）→ 1,867件（有害文書数767件。約41%）
- 評価方法の再検討（高橋）
 - 安全性タイプによる評価基準検討
- AISIとの連携
 - 6/26よりNRIセキュアテクノロジーズのRed-teamingの技術的調査への協力
 - 安全性以外に評価手法や透明性などの観点での連携は可能か？
 - 有害文書フィルタリング、安全性チューニング手法、など
 - AnthropicのClaude 3.5 Sonnetは英国AISIIの協力で安全性評価

※LLM-jpはNIIが主宰するフルオープンアクセスな日本語LLM研究開発のための組織横断的プロジェクト[1]

[1] LLM-jp, "LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs." *arXiv preprint arXiv:2407.03963* (2024). 39

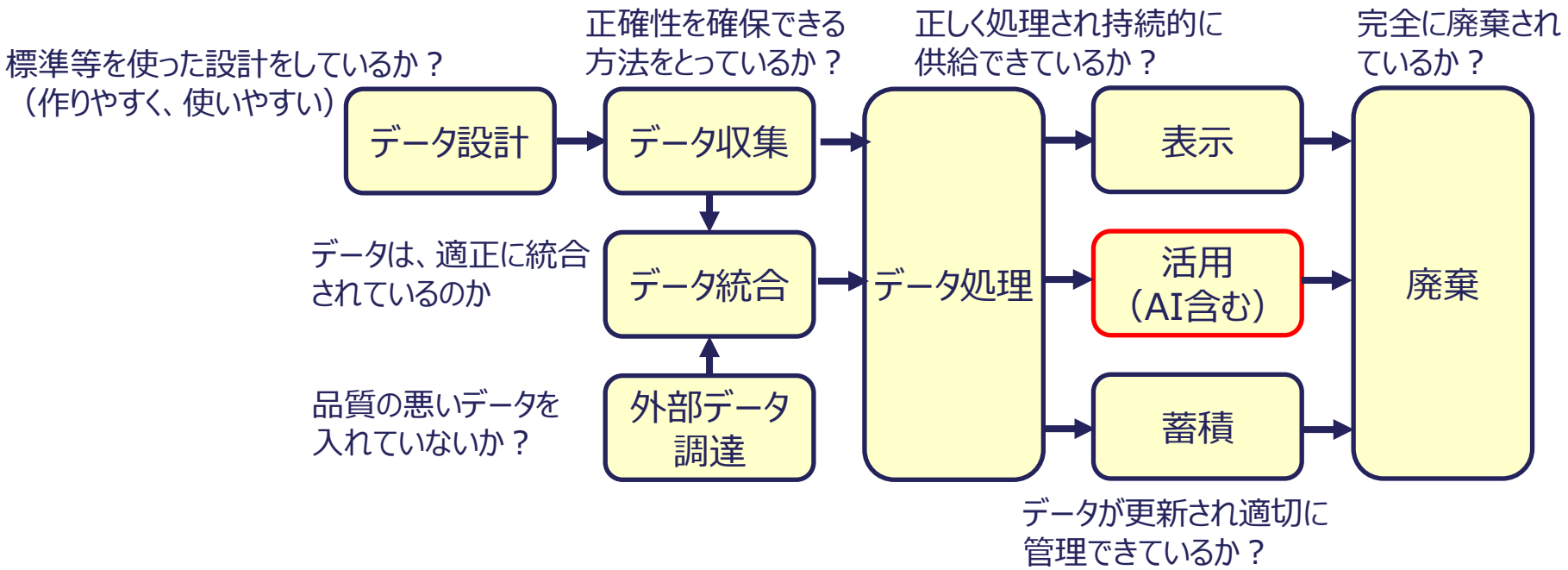
データ品質確保の取り組み（IPAデジタル基盤センター）

主体	IPA
発表形態イメージ	ガイドラインなど

IPA-1
登録年月: 2024.08
更新日: 2024.08.06

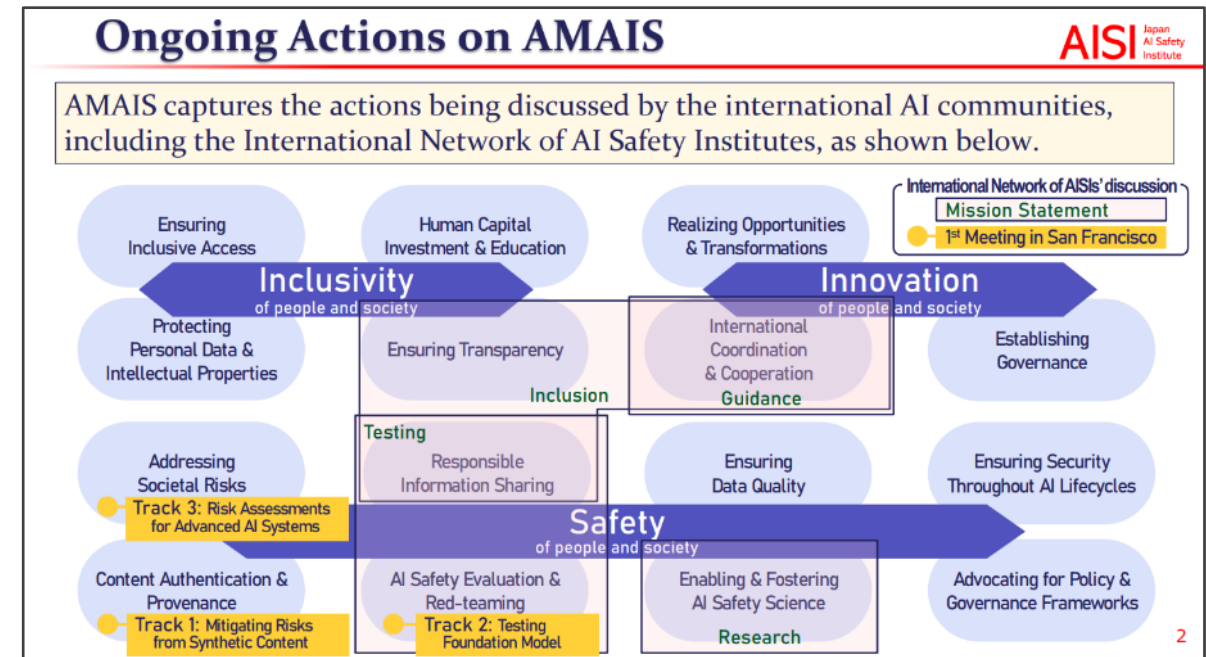
開示範囲： 公知

- AIの開発、活用にはデータの供給が欠かせない。AIが正しいふるまいをするため、高品質な学習用データ、処理対象データ、評価用データを供給するデータのライフサイクルが必要。



- 政府相互運用性フレームワーク（GIF：Government Interoperability Framework）の一環で、IPAはデジタル庁と協力し、データ参照モデル、方法論、データ品質管理に関するガイドを整備・展開済み。
- ISOのSC42専門委員会（人工知能）における「AIのためのデータ品質」の検討も踏まえ、AIも考慮したガイドへの改定を予定。（2024年末目途）

- ◆ AISIは、**ディスカッションペーパー**として「AM AIS : AIの安全性に関する活動マップ」を公開。
- ◆ AISIは、**主要文献のベンチマーク**に基づき、包括的なアクティビティマップと関連用語を開発している。
- ◆ この日本主導の取り組みは、AIの安全性に関する国際的な協力体制の基盤をさらに強化し、持続可能で信頼性の高いAI社会の実現に貢献することが期待されている。



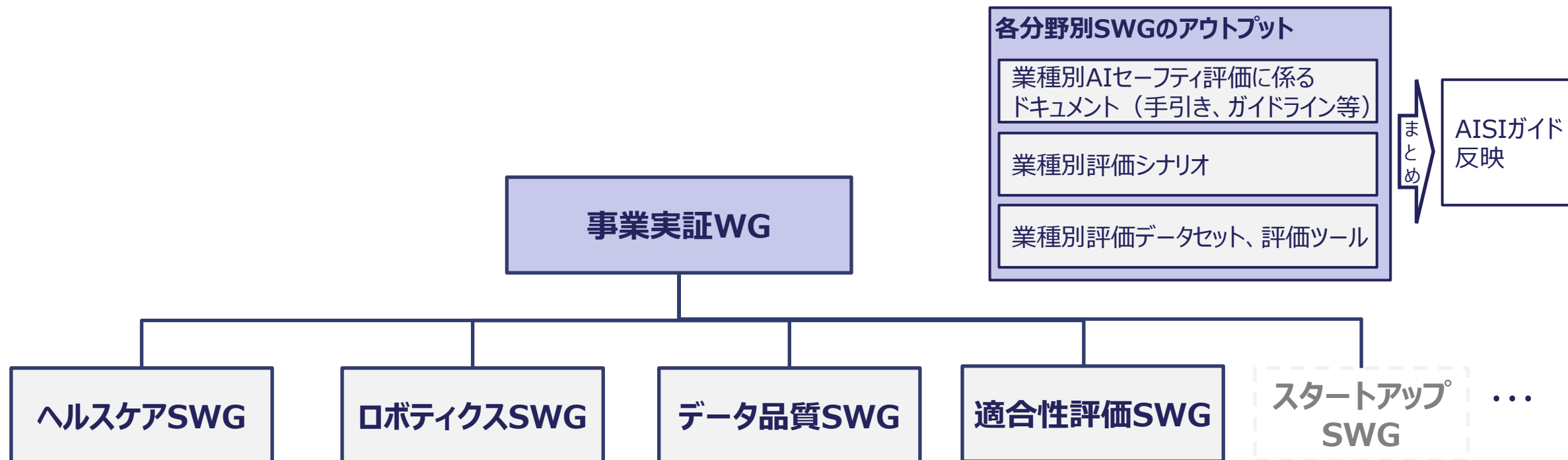
事業実証ワーキンググループの設置

目的： AIの安全性確保に資する実践的な枠組みづくり

- ◆ **多様な民間主体の参加を目指す**

- ◆ **業種別SWGと共通課題SWG**

各業種の課題に応じた検討やツール開発を実施するとともに業種横断での共通課題にも言及



本日のアジェンダ

1. なぜ、人工知能に安全性の配慮が必要なのか
2. AIのリスクの例とAIガバナンスの必要性
3. AI Safety Instituteの紹介とその役割
4. AI安全性の研究とAISIパートナーシップ協定
5. AI安全性と国際協調の重要性
6. AI規制の国際動向
7. AI安全性の技術動向と研究者に期待すること

各国のガイドライン比較の公開や、国際会合での議論への積極的な参加により、各国での施策の整合性に積極的に関与できる

- ◆ 各国ガイドラインの相互関係の確認
 - 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
- ◆ 国際的に通用する各種ガイドラインの作成と公表
 - 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等
- ◆ 国際会合での合意形成
 - AISI関連のトップレベルの連携
 - AIソウル・サミット（2024年5月21-22日、ソウル）
 - 国連未来サミット（2024年9月22日、国連本部）
 - AISI国際ネットワーク会合（2024年11月10-11日、サンフランシスコ）
 - AIアクションサミット（2025年2月6-11日、パリ）
 - 広島AIプロセス・フレンズグループ会合（2025年2月27-28日、東京） 等



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

広島AIプロセスでの議論をへて 広島AIプロセス報告枠組み（レポートフレームワーク）を開始

2023年5月

岸田総理大臣(当時)が
「広島AIプロセス」
を提唱

2023年12月

「広島AIプロセス包括的政
策枠組み」等に各国合意

2025年2月

レポートフレームワーク
開始

ORGANIZATION	PUBLICATION DATE ↕	COUNTRY/TERRITORY	Download PDF	View
MGIT	Apr 24, 2025, 08:00 PM GMT+9	RO	Download PDF	View
Rakuten Group, Inc.	Apr 23, 2025, 08:02 AM GMT+9	JP	Download PDF	View
Fujitsu	Apr 23, 2025, 07:17 AM GMT+9	JP	Download PDF	View
AIZ	Apr 23, 2025, 01:02 AM GMT+9	IL	Download PDF	View
Fayston Preparatory School	Apr 23, 2025, 01:02 AM GMT+9	KR	Download PDF	View
TELUS	Apr 23, 2025, 01:02 AM GMT+9	CA	Download PDF	View
Google	Apr 23, 2025, 01:01 AM GMT+9	US	Download PDF	View
OpenAI	Apr 23, 2025, 01:01 AM GMT+9	US	Download PDF	View
Anthropic	Apr 23, 2025, 01:01 AM GMT+9	US	Download PDF	View
Salesforce	Apr 23, 2025, 01:01 AM GMT+9	US	Download PDF	View
KYBai GmbH	Apr 23, 2025, 01:01 AM GMT+9	DE	Download PDF	View
Microsoft	Apr 23, 2025, 01:01 AM GMT+9	US	Download PDF	View
NIPPON TELEGRAPH AND TELEPHONE CORPORATION	Apr 23, 2025, 01:00 AM GMT+9	JP	Download PDF	View
NEC Corporation	Apr 23, 2025, 01:00 AM GMT+9	JP	Download PDF	View
Preferred Networks	Apr 23, 2025, 01:00 AM GMT+9	JP	Download PDF	View
Data Privacy and AI	Apr 23, 2025, 01:00 AM GMT+9	DE	Download PDF	View
West Lake research & education service, a division of Palo Alto Research	Apr 23, 2025, 01:00 AM GMT+9	US	Download PDF	View
SoftBank Corp.	Apr 23, 2025, 12:59 AM GMT+9	JP	Download PDF	View
KDDI Corporation	Apr 23, 2025, 12:58 AM GMT+9	JP	Download PDF	View

2025年5月26日時点
19社が参加

- **フレンズグループ**

2024年5月に発足した安全・安心で信頼できるAIの実現を目指す、賛同国による自発的な国際枠組み

- **HAIP報告枠組み**

生成AIのリスク評価とガバナンスの実装状況を各国が共有できる仕組み

2025年2月に正式に立ち上げ、4月に1回目の報告が出揃う



AISI国際ネットワーク

米国の呼びかけで現在10カ国が参加
議長国は米国、事務局は英国が担当

カナダ

- 2024年11月、AISi設立

米国

- 2024年2月、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

- 2023年11月、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180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

- 2024年5月、欧州委員会に設立されたAI OfficeがAISi相当の機能も担い、利活用に加え、安全性も推進。AI法の整備と推進も担う
- 90人程度の規模。

ケニア

- AISiネットワークに参加

フランス

- INESIA（AISi相当機関）を2025年2月に設立

韓国

- 2024年11月、AISi設立
- アジアのハブを目指す

日本

- 2024年2月、IPA（情報処理推進機構）にAISiを設立（UK, USに次ぐ3番目）

シンガポール

- 2024年5月、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として指名される
- 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

- AISiネットワークに参加

Created with mapchart.net

■ AISi設立済み

■ AISi相当機関
設立済み/予定

国際AISINネットワークは 世界中の技術的専門知識を結集する科学フォーラムとなることを目的とする

- ◆ **共通理解の構築:**
 - 文化的・言語的多様性を認識し、AIの安全リスクと緩和策に関する共通の科学的理解を目指す
- ◆ **国際的なAI安全性の促進:**
 - AIの安全性の理解・アプローチを世界的に広め、AIイノベーションの恩恵をあらゆる国々で共有
- ◆ **技術的連携の推進:**
 - メンバーは、AI安全性の研究、試験、ガイダンスに関する技術的な連携を促進するため協力

<優先的に取り組む重要4分野>

研究	テスト	ガイダンス	包摂
先進AIシステムのリスクと能力に関する研究を進め、AIの安全性を前進させ、関連性の高い結果を共有	共同テスト演習の実施や国内評価結果の共有を含め、先進AIシステムをテストする共通のベストプラクティスを構築	国際的な共通のAIガイドラインを促進し、先進AIシステムのテストの解釈について相互運用可能なアプローチを通知	情報や技術ツールの共有で世界各国、パートナー、利害関係者を巻き込み、AI安全性の科学に多様な関係者が参加できる能力向上を期待

本日のアジェンダ

1. なぜ、人工知能に安全性の配慮が必要なのか
2. AIのリスクの例とAIガバナンスの必要性
3. AI Safety Instituteの紹介とその役割
4. AI安全性の研究とAISIPartnership協定
5. AI安全性と国際協調の重要性
6. AI規制の国際動向
7. AI安全性の技術動向と研究者に期待すること

日本のAI法「ソフトローアプローチをハードローで規定」

厳しい罰則で規制をする方向を目指すEU、
規制緩和でイノベーションを促進させる米国



罰則なし
AI促進
ソフトローを
ハードローで規定



世界初の包括的なAI規制法

2024年に採択、2026年頃から本格施行予定



- ◆ **リスクベースの規制**
AIを4つのリスクレベルに分類し、規制の強さを段階的に設定
(例：禁止、高リスク、限定リスク、低リスク)
- ◆ **高リスクAIへの義務**
透明性・安全性・人による監視などの厳格な要件が課される
(例：信用スコア、顔認識、重要インフラのAIなど)
- ◆ **生成AIへの対応も明記**
透明性の確保、著作権表示などの義務あり（ChatGPTなどが対象）
- ◆ **企業・開発者への影響大**
EU域外の事業者もEU市場で展開する場合は遵守が必要

NEWS

US House Passes 10-Year Moratorium on State AI Laws

JUSTIN HENDRIX, CRISTIANO LIMA-STRONG / MAY 23, 2025



The US Capitol in Washington, DC.

米国は政権交代に伴い大きく方針を転換

- ◆ バイデン前大統領が2023年10月に出したAIに関する大統領令を廃止。

米国では現在、州レベルでのAI規制について、10年のモラトリアム期間をおく提案が下院から上院へ進捗

- ◆ 共和党は、州ごとの規制のバラつきがAIイノベーションを阻害し、特に中小企業に負担をかけると主張

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（当時）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

A I 戦略会議・A I 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

イノベーション促進とリスク対応の両立を基本的な考え方とし、 日本政府全体を俯瞰する「司令塔」が必要だと提言

背景

- AIは我が国の発展に大きく寄与する可能性がある一方、様々なリスクが顕在化。
- AIに対する不安の声が多く、諸外国と比べても開発・活用が進んでいないとの指摘。
- ▶ AIの透明性など、適正性を確保し、AIの開発・活用を進める必要がある。

基本的な考え方

- **イノベーション促進とリスク対応の両立**
(Ⅱ.3.)
 - 研究開発支援、人材育成、データや計算資源の整備などイノベーションの促進
 - 法令とガイドライン等の適切な組合せ
 - OECD原則、広島プロセス国際指針等の共通的な指針等と個別の既存法令の活用



- **国際協調** (Ⅱ.4.)
 - AIガバナンスの形成に向けて議論をリード
 - 国際整合性・相互運用性の確保



日本：AI法案が成立（2025年5月28日）

AI制度研究会の中間とりまとめをもとに、人工知能関連技術の研究開発及び活用の推進に関する法律案(AI法案)が国会提出・成立

法案の概要	目的	国民生活の向上、国民経済の発展
	基本理念	経済社会及び安全保障上重要 → 研究開発力の保持、国際競争力の向上 基礎研究から活用まで総合的・計画的に推進 適正な研究開発・活用のため透明性の確保等 国際協力において主導的役割
	AI戦略本部	本部長：内閣総理大臣 構成員：全閣僚 関係行政機関等に対して必要な協力を求める
	AI基本計画	研究開発・活用の推進のために政府が実施すべき施策の基本的な方針等
	基本的施策	研究開発の推進、施設等の整備・共用の促進 人材確保 教育振興 国際的な規範策定への参画 適正性のための国際規範に即した指針の整備 情報収集、権利利益を侵害する事案の分析・対策検討、調査 事業者・国民への指導・助言・情報提供
	責務	国、地方公共団体、研究開発機関、事業者、国民の責務 関係者間の連携強化 事業者は国等の施策に協力しなければならない
	附則	見直し規定（必要な場合は所要の措置）

- 人権侵害やサイバー攻撃への悪用など生成AIがもたらすリスクに対応し、悪質な場合は国が実態調査をしたうえで必要に応じて情報提供を行う
- AI政策の司令塔機能の創設
（首相が本部長、全閣僚をメンバーとする「AI戦略本部」を設置）
- 国民の権利侵害が生じた際、「指導、助言、情報の提供その他の必要な措置を講ずる」

「AI戦略」が利活用と安全性を両立させ、イノベーションを促進する鍵となる。
先端技術と安全性を理解するAI人材の持続的な確保並びに育成が必要。

- ◆ 包括的な戦略観を示すことが、組織内でのサイロ化やガバナンス不足を防ぎ、リスク防止とイノベーション促進。統合的に戦略を作る「司令塔」が国や組織に不可欠。
 - ・ 戦略は一度作ったら終わりではなく外的要因も加味しながら継続した見直しが必要

- ▶ サステナブルな人材確保が、技術革新と安全性の両立を実現する。
 - ・ 優秀な人材確保のためには母数を増やす必要あり。もっと理系を専攻する女性を増やすべき
 - ・ 結果、AIの開発に多様性(D)が生まれ、公平性(E)・包括性(I)が担保される

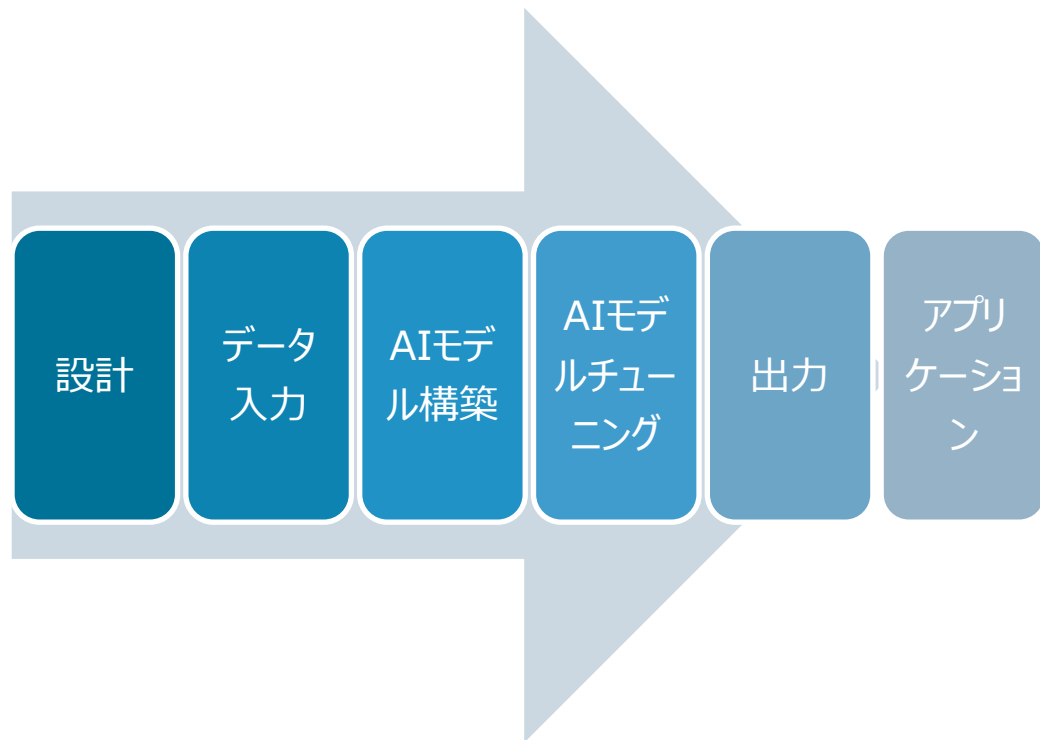
- ▶ 情報リテラシー教育及び学習の振興、広報活動の充実
 - ・ 高度な専門家やエンジニアの確保・養成という文脈に加えて、国民が広く人工知能関連技術の利活用とリスク対策の双方に対する理解と関心を深めるよう、必要な施策という面での期待

本日のアジェンダ

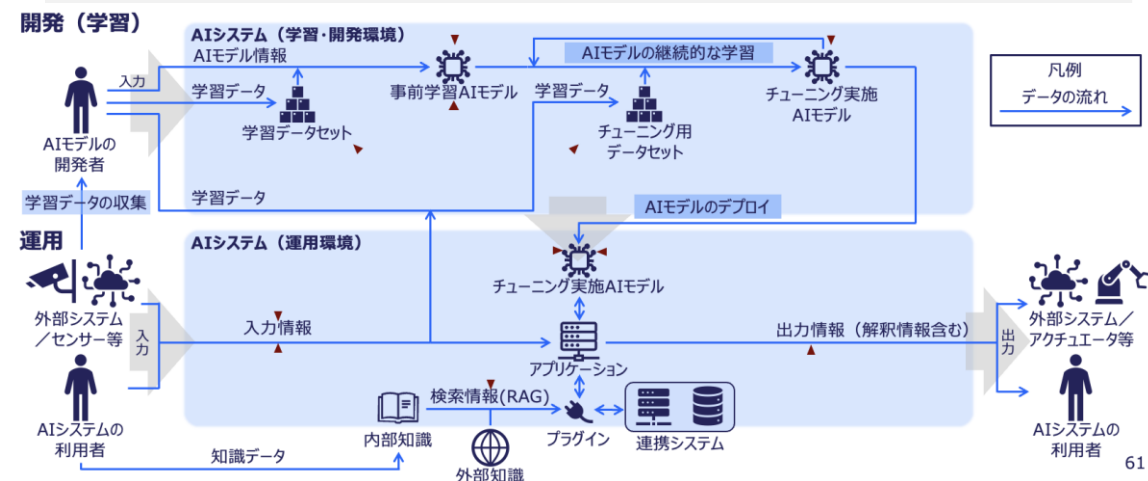
1. なぜ、人工知能に安全性の配慮が必要なのか
2. AIのリスクの例とAIガバナンスの必要性
3. AI Safety Instituteの紹介とその役割
4. AI安全性の研究とAISIパートナーシップ協定
5. AI安全性と国際協調の重要性
6. AI規制の国際動向
7. AI安全性の技術動向と研究者に期待すること

AIのリスクや安全性を技術的に捉えるためには

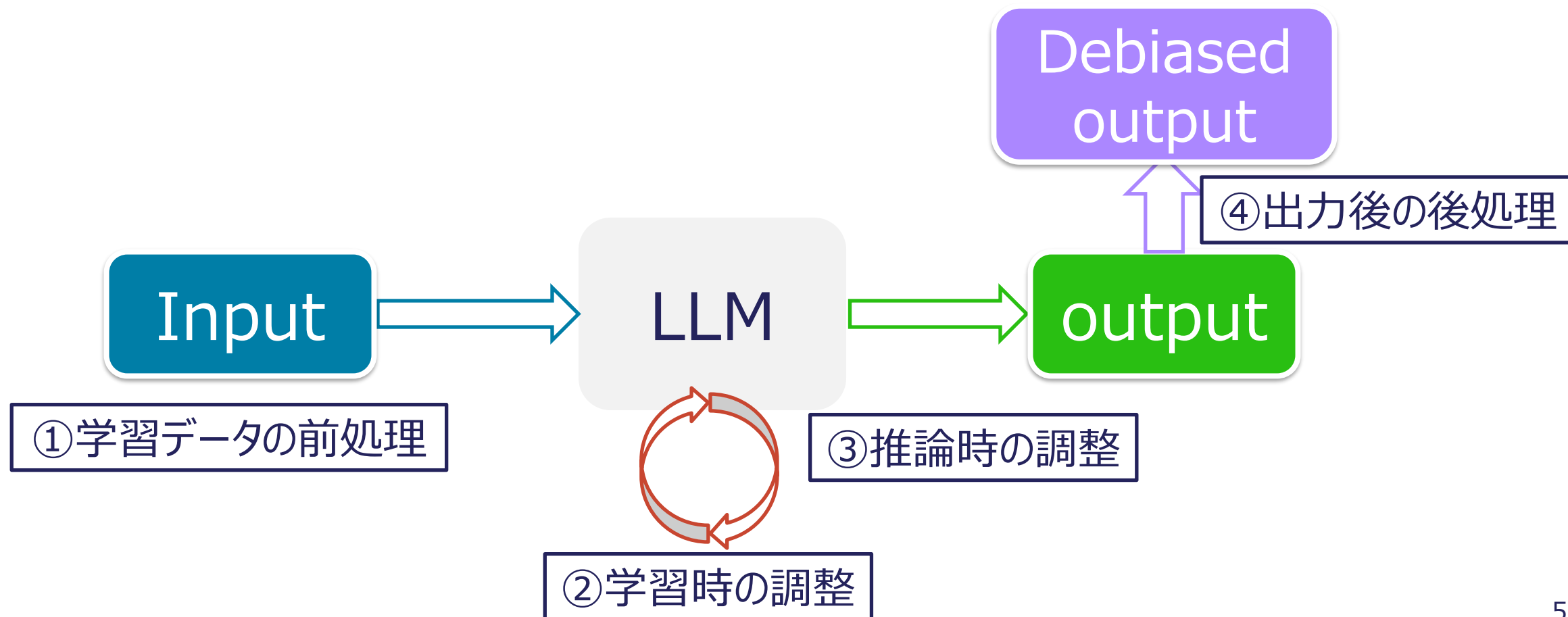
AIモデルレベルからシステムレベルで安全性を考える必要がある



AIシステム全体

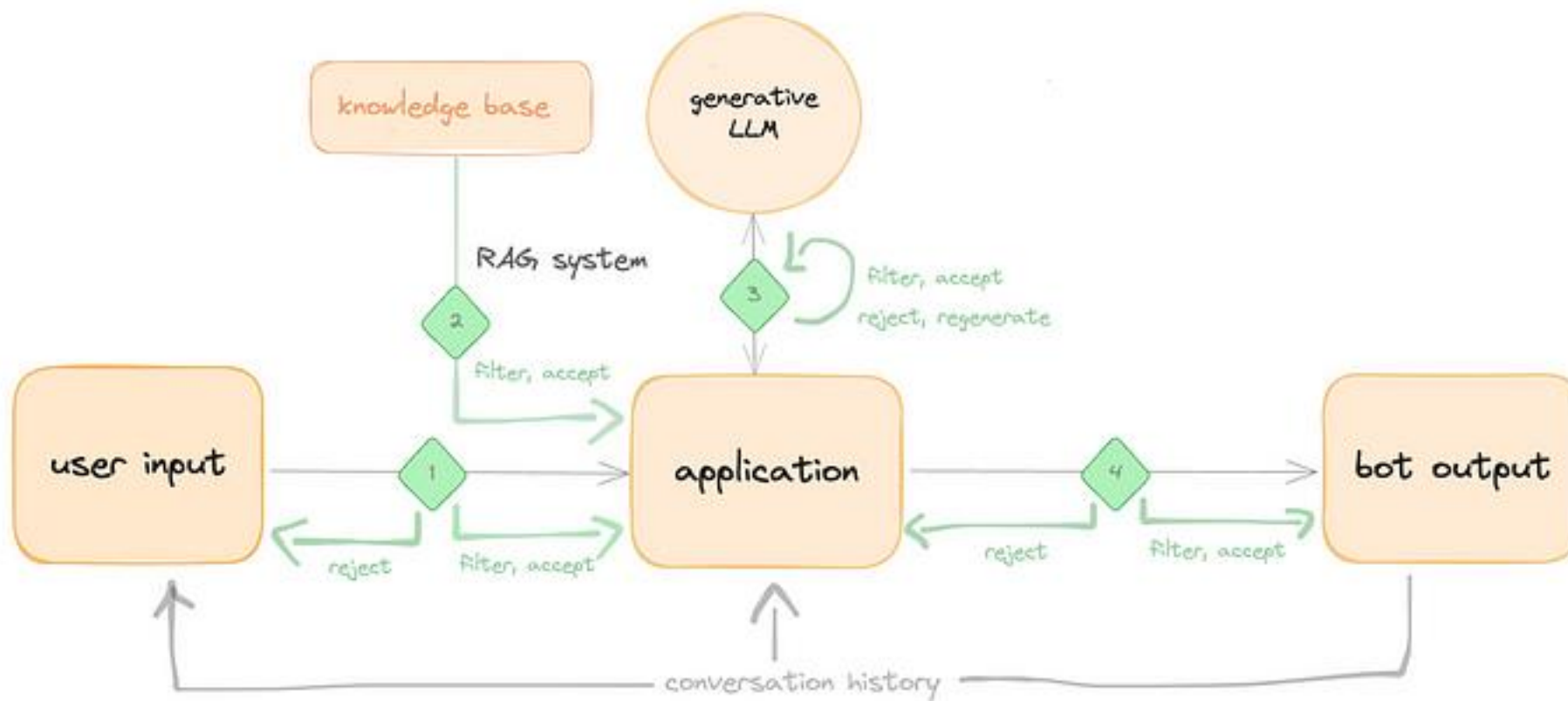


バイアスへの対処技術としては①学習データの前処理、②学習時の調整、③推論時の調整、④出力の後処理が研究されている



リスク回避技術の研究 LLMガードレールの設置

大規模言語モデル（LLM）に対する不適切な入力を制御したり、不適切な出力や誤情報、偏見、攻撃的な内容を生成しないようにする技術



LLMの安全性テストを各国Jointで実施。モデルや言語、分野をまたいだテストとして一定の成果はでたが、科学的な評価方法について課題もある

- ◆ 第二回共同テスト演習レポートの公表[1]

International Network
of AI Safety Institutes

FEBRUARY 11-12, 2025 | PARIS, FRANCE

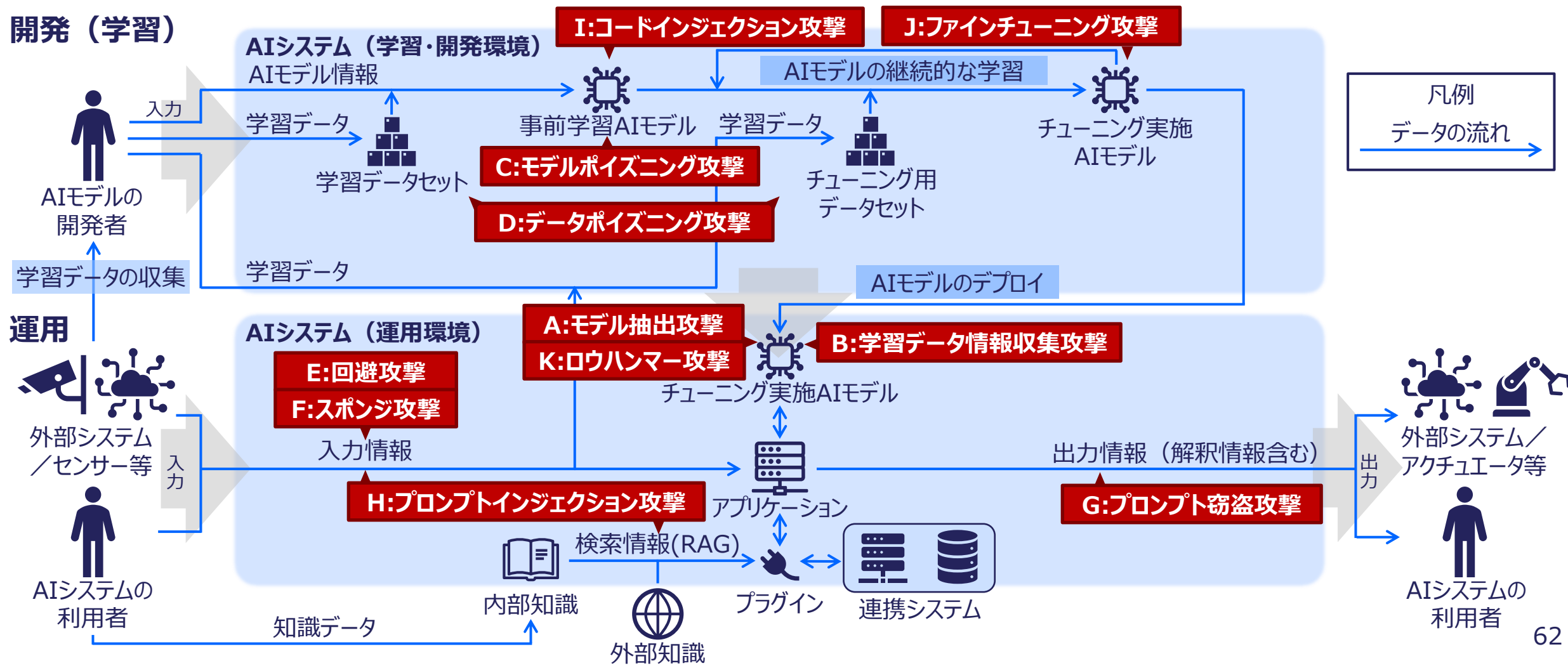
International Network of AI Safety Institutes Joint Testing Exercise:

Improving Methodologies for AI Model Evaluations Across Global Languages

[1] 共同テストのブログ記事が公開されました。 - AISI Japan - <https://aisi.go.jp/assets/pdf/International-Network-of-AI-Safety-Institutes-Joint-Testing-Exercise-Improving-Methodologies-for-AI-Model-Evaluations-Across-Global-Languages.pdf>

AIシステムに対する攻撃の俯瞰

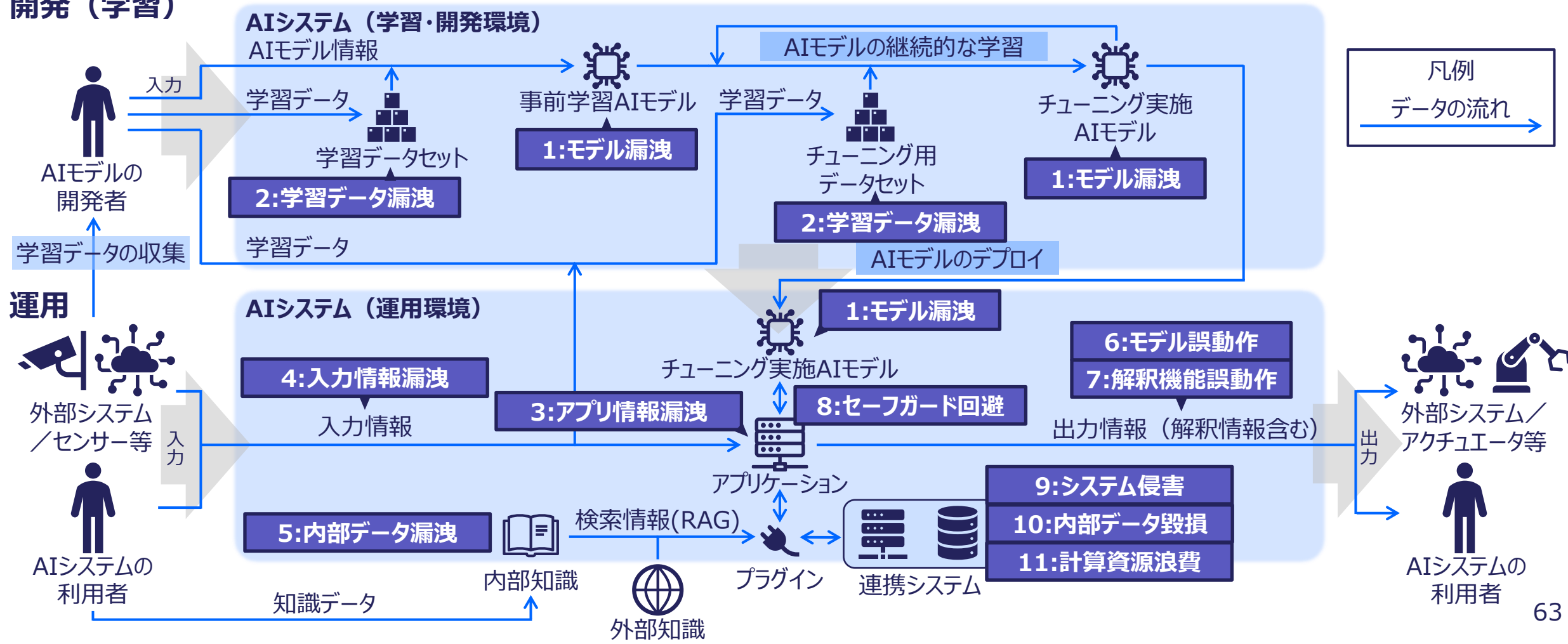
想定するAIシステムへの攻撃（スコープは前述の通り）を下図に示す。
学習データ情報収集攻撃等の攻撃は、さらに複数の種類に分類できる。



AIシステムに対する攻撃による影響の俯瞰

前スライドに示した攻撃による影響を下図に示す。
モデル漏洩のように同じ種類の影響が複数の箇所に生じる場合もある。

開発（学習）



The International AI Safety Report

汎用人工知能（general-purpose AI）の現状と将来の能力、 それに伴うリスクについて包括的にまとめた報告書



International AI Safety Report

The International Scientific Report
on the Safety of Advanced AI

January 2025

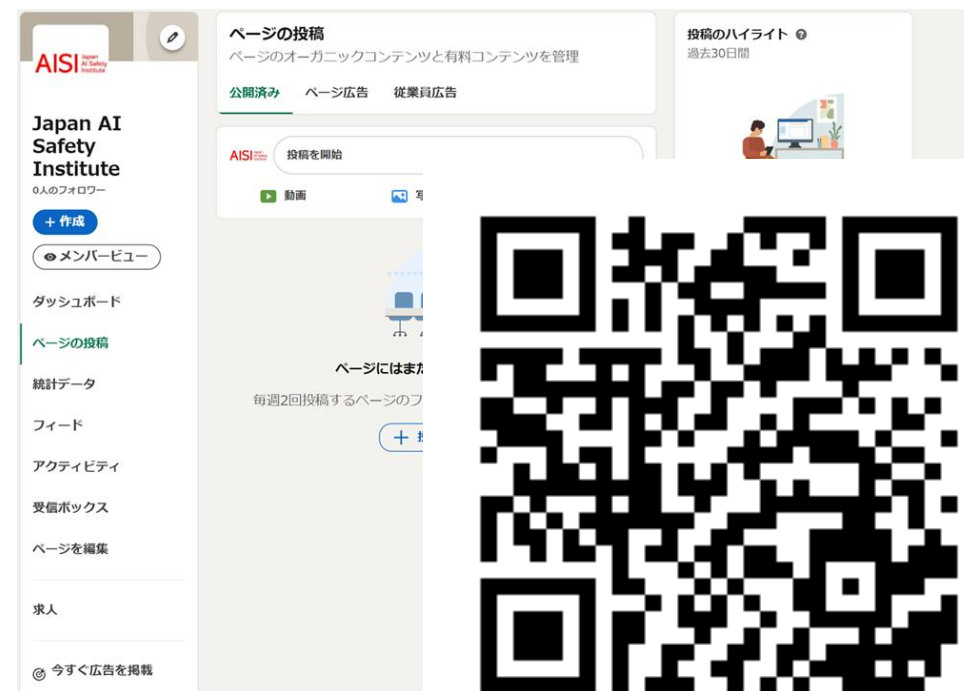
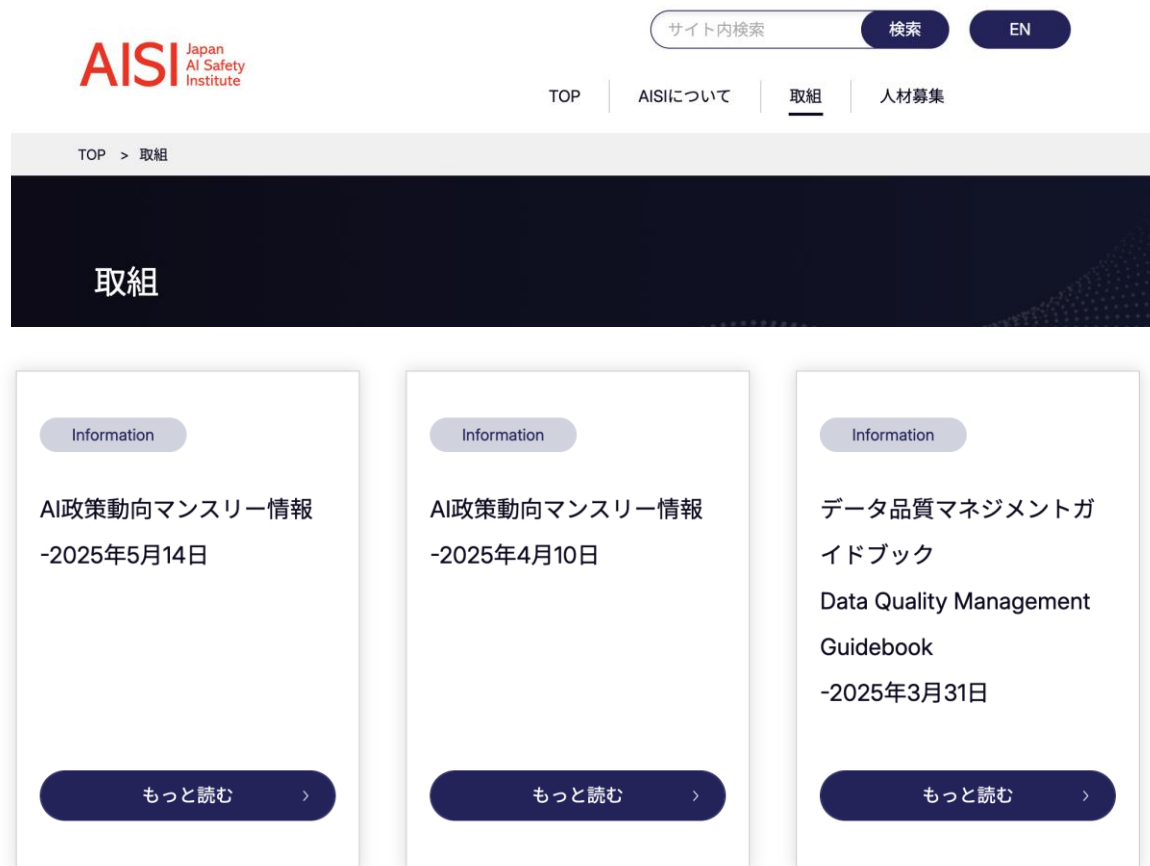
- ◆ 世界各国の専門家の貢献に基づき、汎用AIが公共の意見操作やサイバー攻撃、バイオ攻撃に悪用される可能性や、偏見、プライバシー侵害、環境への影響といった社会的な課題を詳細に論じている
- ◆ また、AIシステムの制御が難しく、開発者でさえその内部動作を完全に理解できないという技術的な課題や、**安全性評価の限界**についても触れており、これらのリスクを軽減するための技術的および政策的なアプローチについても提言している

本日のアジェンダ

1. なぜ、人工知能に安全性の配慮が必要なのか
2. AIのリスクの例とAIガバナンスの必要性
3. AI Safety Instituteの紹介とその役割
4. AI安全性の研究とAISIPartnership協定
5. AI安全性と国際協調の重要性
6. AI規制の国際動向
7. AI安全性のガイドラインと技術動向
8. 研究者に期待すること

まずはAISIIの活動に注目！

AISIではホームページやLinkedIn(new!)で逐次情報を展開しています



LinkedIn

AI安全性の研究ネタはAIの技術の分野の数だけ存在
まだ各分野でも模索している段階

- ◆ 現実的なシステムリスクから理論まで幅広い
- ◆ 世界での議論は「超実用的」と「超抽象的」に断絶している
- ◆ このギャップを埋めるのは研究者しかできないこと



一緒にAI安全性を考えましょう！

AISIは人材をもちろん募集していますが
AISIに転職しなくても、大学・研究所・企業に所属しながら
一緒にAI安全性について推進してくれる人を募っています

- ◆ （研究費、という形ではAISIからは直接難しいですが）
AISIの技術協力者になっていただければ、海外学会への派遣やAISIの活動と関係した諸費用（計算機コストや外注費用）は賄える可能性もあります
- ◆ 今後NEDO、JSTなどで多くのAI安全性のプロジェクトが爆誕していくと予想しています – 流れに乗り遅れないで！

AISI

Japan AI Safety Institute