Tokyo DXセミナー

都民に信頼されるAI：
「設計段階からの安全性」が創る未来

2025年8月25日

SOMPOホールディングス株式会社
執行役員常務 グループチーフデータオフィサー
損害保険ジャパン株式会社
執行役員チーフデータオフィサー、データドリブン経営推進部長
AIセーフティーインスティテュート
所長
村上 明子

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

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】「未踏アドバンスト事業」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ト委員
【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）

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

本日のアジェンダ

1. SOMPOが目指すAIとの共生
2. AI安全性と考慮しなければならないAIのリスク
3. 日本政府のAI安全性の活動とAI Safety Instituteの紹介
4. AIのリスクや安全性を技術的に捉えるために
5. 東京都のAIに必要な「Safety by Design」

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

本日のアジェンダ

1. SOMPOが目指すAIとの共生
2. AI安全性と考慮しなければならないAIのリスク
3. 日本政府のAI安全性の活動とAI Safety Instituteの紹介
4. AIのリスクや安全性を技術的に捉えるために
5. 東京都のAIに必要な「Safety by Design」

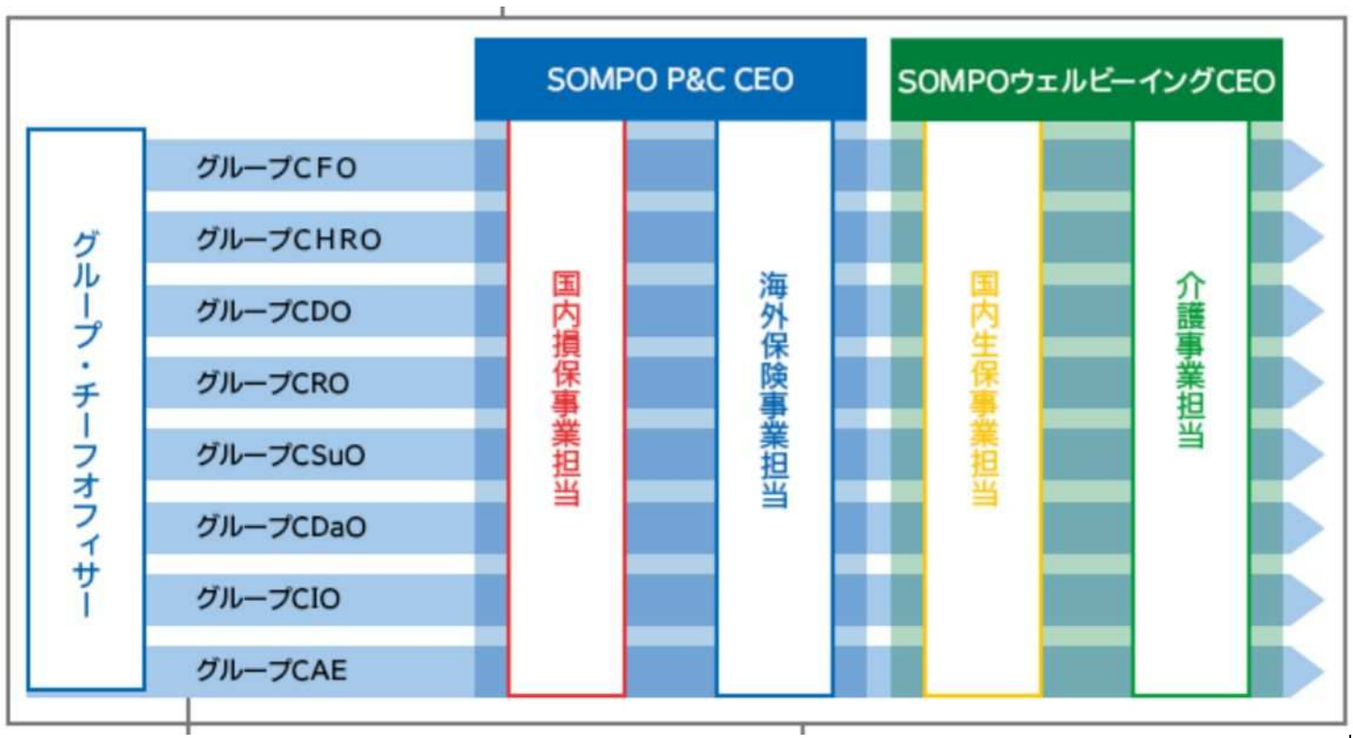
© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

SOMPOホールディングスと損保ジャパンについて

SOMPO

ＳＯＭＰＯのパーパス
“安心・安全・健康” であふれる未来へ

SOMPO



© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

SOMPOが目指すAIとの共生

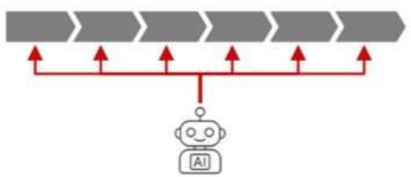
人とAIの共生でビジネスモデルの再構築と信頼醸成を実現

目指す方向性

人とAIが共生する世界へ！

～AIを前提にビジネスモデルを根本的に再構築。AIを活用した人間が顧客との信頼醸成において替え難い能力を発揮～

AIの存在をビジネスに不可欠な存在と捉え、あらゆる工程、顧客接点においてAIを活用する。大量の価値あるデータをAIが学習することによる付加価値が付いたサービスをお客さまに提供する。人間はAIにできない「顧客との信頼醸成」「新たなビジネスモデルの創造」などを担う

AI 1.0	AI 2.0	AI 3.0
個別業務代替	オペレーション全体の変革	ビジネスモデル変革
		
AIによって個別業務の生産性向上 人間が行なっている業務にAIを導入することで生産性向上、ないしは代行。組織として「議事録作成にかかる時間を○時間削減」など成果を示せるようになる	**複数部門の業務をAIがコントロールし、生産性を最大化** 部門を横断する、全オペレーションにおいてAIを利用。保サ業務のように他部門との連携が必要な工程も、AIが適切に連携し、部門間業務の効率を最大化する	**AIを前提とした人とAIが共生するビジネスモデルへと変革する** 今までの常識にとらわれず、既存のビジネスモデル自体を見直し新たなビジネスモデルへの変革を行う（商品の最適化、拠点統廃合、顧客対応のあり方の変革などを含む）

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

人とデジタル技術のハイブリッドによる業務の効率化と高度化

デジタルの役割	人をアシスト×人が出来ない領域を補完
人の役割	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

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

本日のアジェンダ

1. SOMPOが目指すAIとの共生
2. AI安全性と考慮しなければならないAIのリスク
3. 日本政府のAI安全性の活動とAI Safety Instituteの紹介
4. AIのリスクや安全性を技術的に捉えるために
5. 東京都のAIに必要な「Safety by Design」

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

なぜいま、人工知能がここまで注目されているのか

AIの活用は選択肢や差別化ではなく、労働力不足の現在では必須

- 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
- バブル崩壊後、前例踏襲・保守的となった日本においては、AIの導入や活用に遅れがあるのが事実

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

自治体におけるAIの利用例（東京都AI戦略会議資料より）

SOMPO

2-② 自治体の利活用事例１

凡例 実証 導入済

自治体

住民・事業者主体

AIを活用したオンライン就労支援サービス

行政機関	京都市
分野	労働
AIの機能	チャットボット/マッチング

■実施内容（2021年度）
就職氷河期世代の方を対象に、SNS上で「仕事紹介」や「キャリア相談」などが利用できるオンライン就労支援サービスを提供。SNSを活用することで、24時間いつでもカウンセリングの予約が可能。AIチャットボット機能によりニーズに応じた各種サービスの情報を取得可能。自身のスキルや価値観等に関する質問に回答することで、相性のよい企業を提案する「AIマッチング」も搭

■効果（見込み）
窓口で就職相談をすることに抵抗感を持つ方や、現職が忙しく日中に来所することが難しい方でも、少ない負担で就労に関する相談が可能。

多言語翻訳AIチャットボットを活用した外国人への情報発信強化

行政機関	蔵越町等
分野	観光
AIの機能	チャットボット

■実施内容（2020年度）
外国人観光客の割合が多いため、観光案内も外国人に分かりやすいものである必要があり、外国人観光客をターゲットとしてAIチャットボットを運用している。地域で多言語対応のAIチャットボット機能を導入することにより、利用者は、各町の境界に囚われずにサービスを利用することができる。

■効果
観光案内窓口を訪れる観光客が減少し、窓口業務負荷が軽減。
閲覧される情報の傾向を分析したマーケティング戦略の改定、地域の飲食やアクティビティ事業者の情報の積極的な発信により、消費拡大に寄与。

職員主体

AIカメラによる混雑の見える化

行政機関	秩父市
分野	観光
AIの機能	画像認識

■実施内容（2024年度）
道路上や神社近くの駐車場にAIカメラを設置し、渋滞や駐車場の空き状況を撮影しAIで解析。「駐車場まで渋滞」「空きあり」といった解析結果をポータルサイトで発信するとともに、付近の飲食店や観光スポット、土産物店などの情報も合わせて提供する取り組みを開始。

■効果（見込み）
観光客の分散を図り、**混雑を緩和。観光客の満足度の向上。**

AIによる避難指示の自動音声変換/架電・会話のテキスト変換

行政機関	陸前高田市
分野	消防・防災
AIの機能	チャットボット/作業自動化

■実施内容（2023年度）
市災害対策本部の職員がパソコンを操作して、避難を呼びかける指示文を入力すると、AIが自動的に音声に変換し、通報対象の登録者に自動で架電。電話を受けた住民が受話器に向かって自身の状況について説明すると、AIが再び音声をテキストに変換し、市災害対策本部で対象者の状況を一覧で確認可能。

■効果
架電に係る職員の負担軽減。
SNSや電子メールの利用に不慣れな方々への情報伝達が可能。

※（総務省）地域社会DXナビを基に作成 | 地域社会DXナビ

※記載内容は、出典の資料に記載されている当時の状況

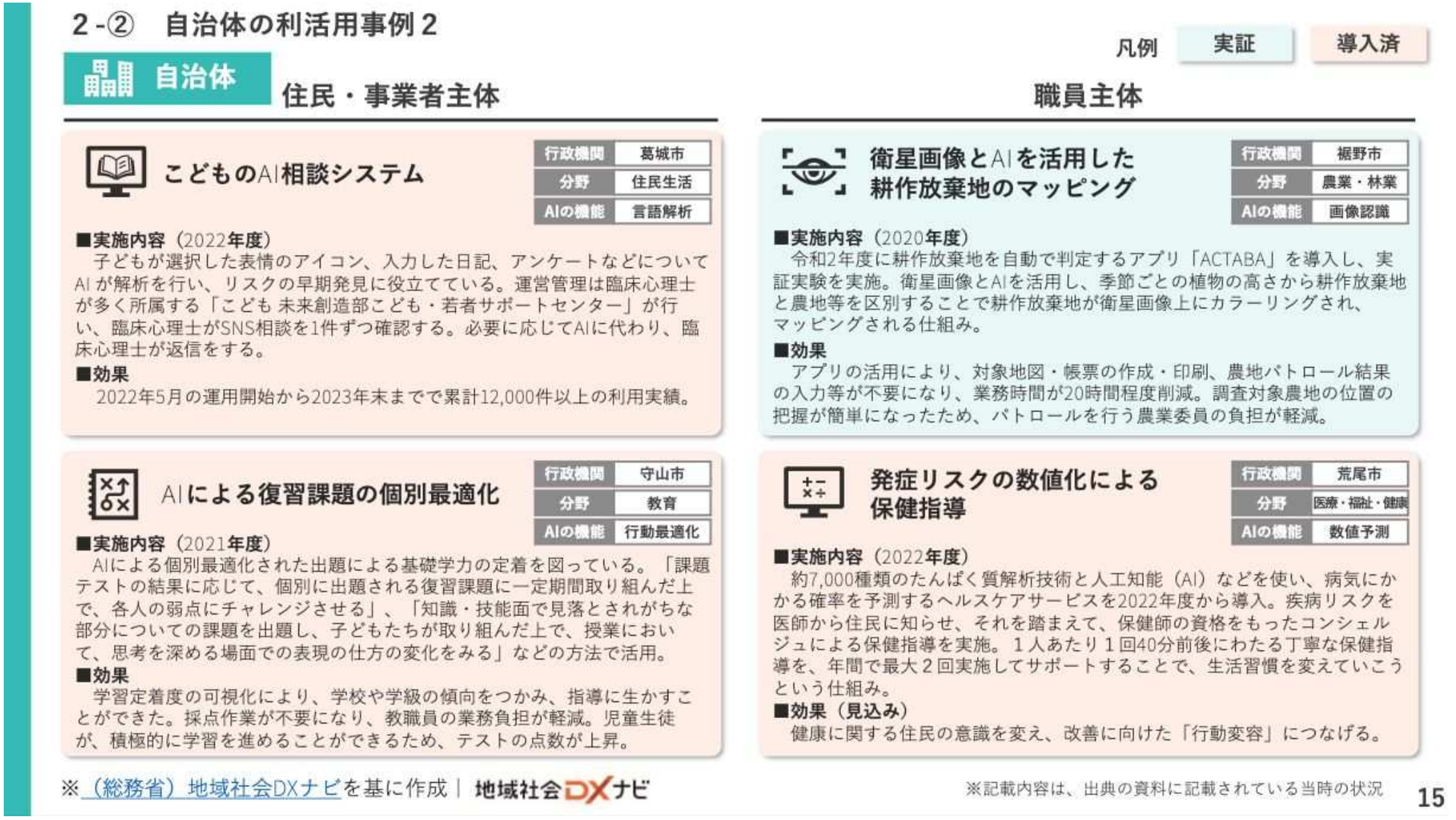
14

出典：第3回東京都AI戦略会議　表示資料より抜粋
https://www.digitalservice.metro.tokyo.lg.jp/business/ai-strategy-council/council03

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

自治体におけるAIの利用例（東京都AI戦略会議資料より）

SOMPO



出典：第3回東京都AI戦略会議　表示資料より抜粋
https://www.digitalservice.metro.tokyo.lg.jp/business/ai-strategy-council/council03

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

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による判断の失敗は数多く知られている（具体例参照）

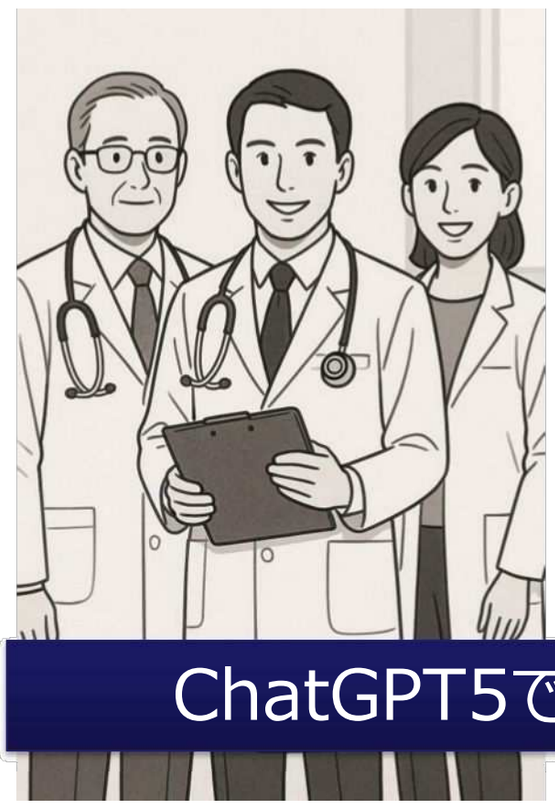
- 「安全」であると「安心」できなければ、AIの導入は進まない

どのように「安全である」とみなすのか

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

AIによるバイアス

画像生成ができた初期の頃、医者、弁護士、CEOを書いてと指示し
生成された画像はほとんどが男性の白人の画像だった



ChatGPT5では多様性が確保されるようになっている

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

生成AIなどの台頭により顕著となったAIのリスク

AISI Japan AI Safety Institute

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
  - データやロジックによるバイアスのほか、誤判定、誤回答、ハルシネーション、安全性、セキュリティ等多く存在
- 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の技術的なリスクー データによるバイアス

AISI Japan AI Safety Institute

焦点：アマゾンがＡＩ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/

AIの技術的なリスク - 誤回答

AISI Japan AI Safety Institute

- 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」。

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

AIの技術的なリスク – プロンプトインジェクション

AISI Japan AI Safety Institute

- 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を悪用しウィルス作成

AISI Japan AI Safety Institute

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悪用によるウイルス作成容疑で25歳男を逮捕。
複数の対話型AIに指示、ウイルスの設計情報を回答させて作成した疑い。
生成AI利用ウイルス作成の摘発は全国初。作成ウイ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。

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

AIの技術的なリスク – 不正プログラム生成

AISI Japan AI Safety Institute

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ＡＩ（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ＩＤ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。

出典：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：読売新聞

AIの社会的なリスク –「生成AI」とのチャット後に自殺

AISI Japan AI Safety Institute

“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。**”

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」「私と一緒になりたいんでしょ」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。

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

AIの社会的なリスク – ディープフェイクによる悪用

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。

https://www.yomiuri.co.jp/national/20231103-OYT1T50260/
https://www.yomiuri.co.jp/national/20231103-OYT1T50260/2/

- 「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の社会的なリスク – 著作権侵害

- 書籍・画像・映像作品などの諸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

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

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。

AIによるレピュテーションリスク – AIによるプライバシー侵害

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

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。

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

「リスク」と関連の深い、AIのデュアルユース

AISI Japan AI Safety Institute

「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 series[1] 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.

出典： Dual Use of Artificial Intelligence-powered Drug Discovery - PMC

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原則からAIガバナンスへ

AISI Japan AI Safety Institute

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

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

本日のアジェンダ

1. SOMPOが目指すAIとの共生
2. AI安全性と考慮しなければならないAIのリスク
3. 日本政府のAI安全性の活動とAI Safety Instituteの紹介
4. AIのリスクや安全性を技術的に捉えるために
5. 東京都のAIに必要な「Safety by Design」

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

「安全」と「安心」の違い

AISI Japan AI Safety Institute

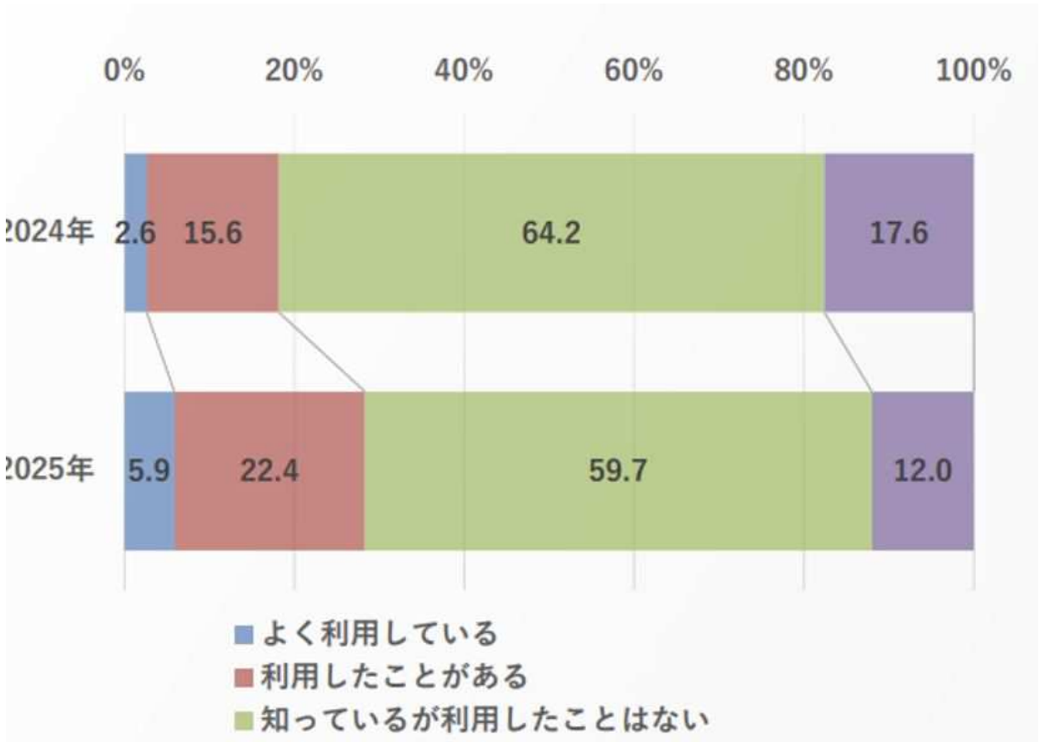
「**安全**」とは「許容不可なリスクがないこと」
「**安心**」とは「心配・不安がないこと」

- 「**安全**」はISO/IEC GUIDE 51:2014(E)で定義されている
  - Safety:Freedam from risk which is not tolerable
- 「**安心**」とは「心配・不安がなくて、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。

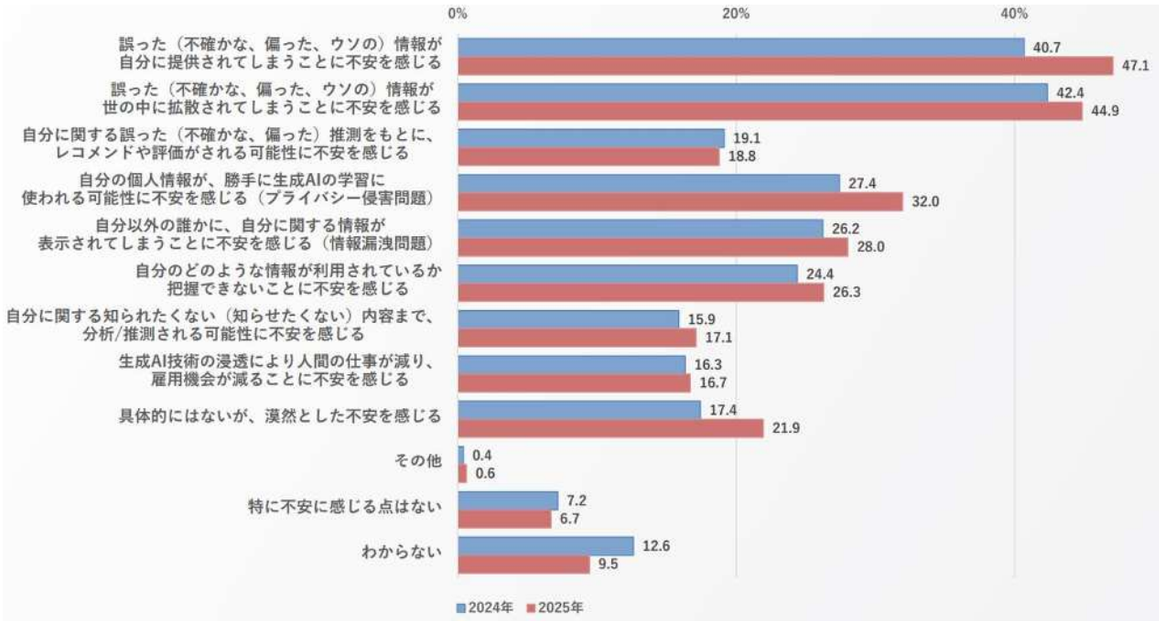
「AIを安全に使う技術的な手段を提供する」だけではなく
「安心してAIを使うことができるために必要な情報を提供する」
ことも必要

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



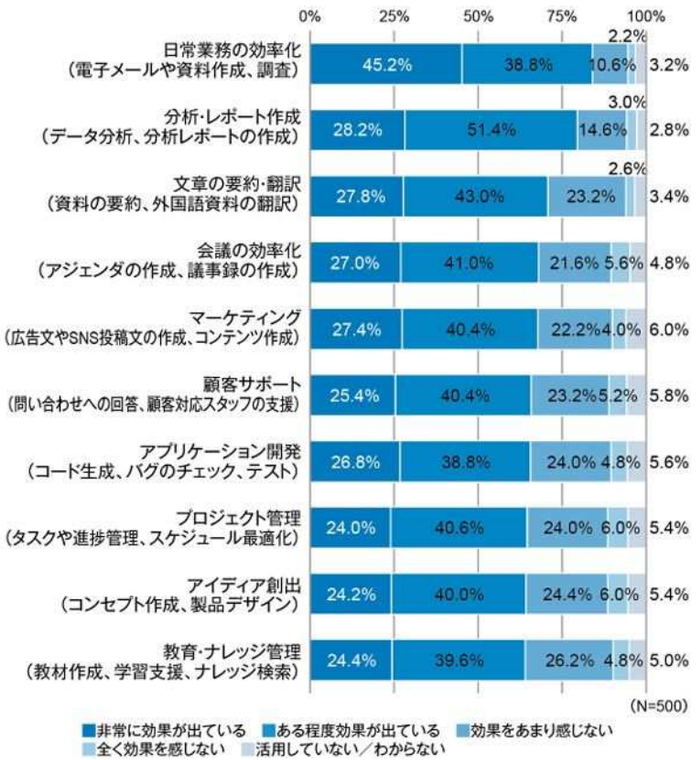
出典：一般財団法人日本情報経済社会推進協会「デジタル社会における消費者意識調査2025」

企業の利用状況と不安

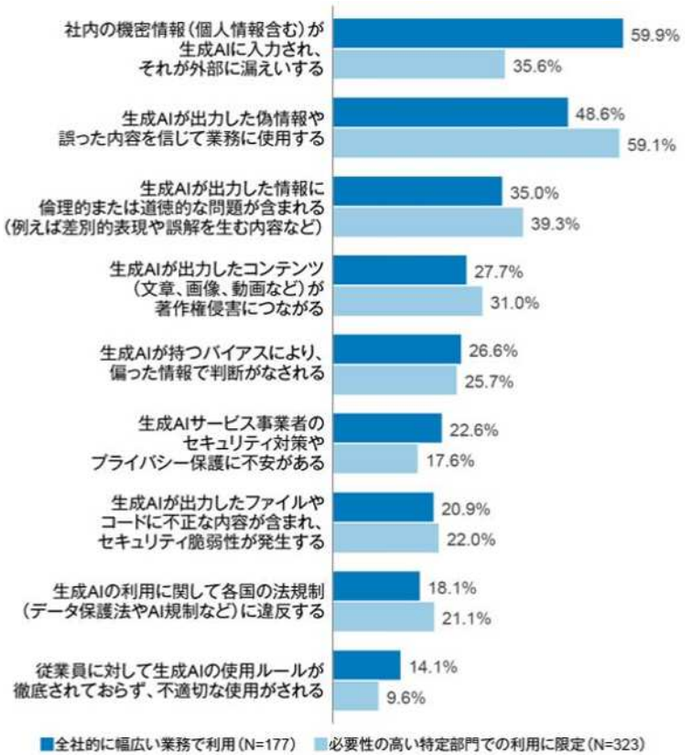
AISI Japan AI Safety Institute

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



「企業IT利活用動向調査2025」一般財団法人日本情報経済社会推進協会、ITR　https://www.jipdec.or.jp/news/pressrelease/20250314.html

統合イノベーション戦略

AISI Japan AI Safety Institute

統合イノベーション戦略2024

AISIを**日本におけるAIの安全性の中心機関**と定義

３つの強化方策の1つの柱がAI

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

統合イノベーション戦略2025

（１）先端科学技術の戦略的な推進

①重要分野の戦略的な推進

- AIイノベーション促進とリスク対応の両立
- AIの研究開発の推進等
- AI関連施設等の整備及び共用の促進
- AI活用の推進
- AIの適正性の確保
- AI関連人材の確保と教育振興等
- AIに関する調査研究等
- AI分野の国際的協調の推進

AISI (AI Safety Institute) の役割とスコープ

AISI Japan AI Safety Institute

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AISI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
- パートナーシップ事業の推進

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

AIセーフティ実現に向けた業務

AISI Japan AI Safety Institute

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

成果物の概要

AISI Japan AI Safety Institute

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

クロスウォーク
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

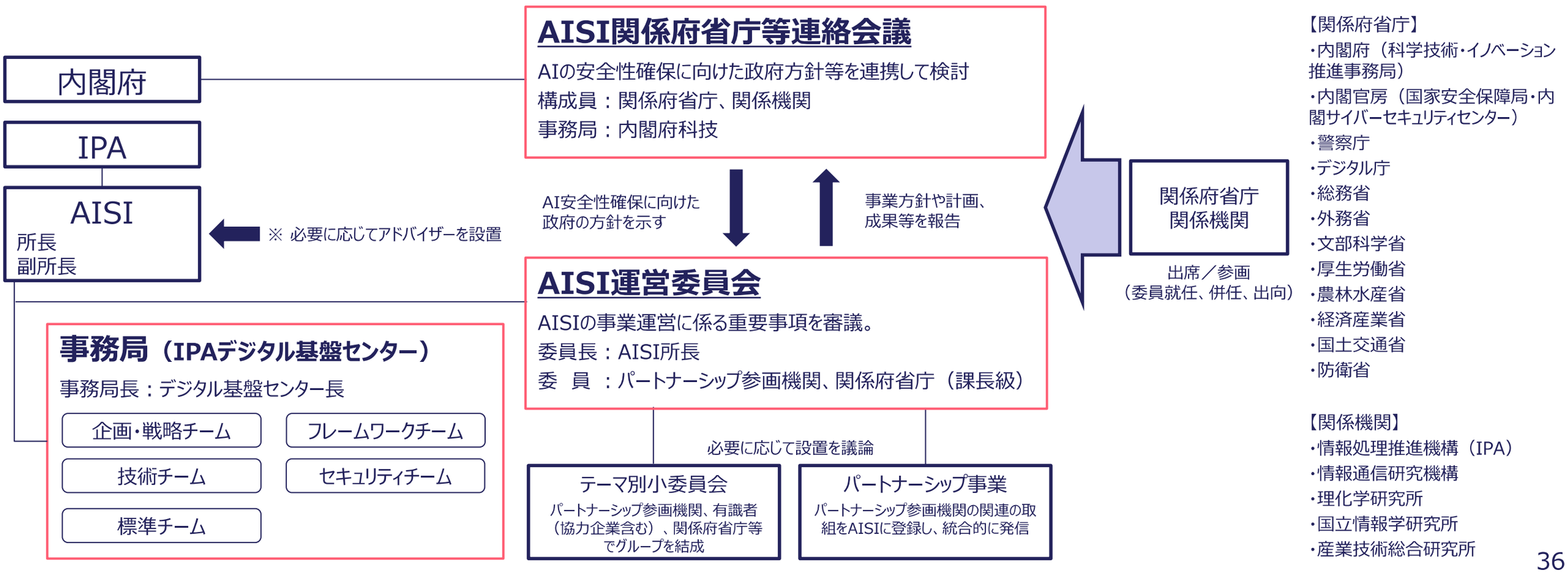
セキュリティレポート
セキュリティに関する知識

デジタルスキル標準
人材育成

AISIの推進体制

AISI Japan AI Safety Institute

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



【関係府省庁】
・内閣府（科学技術・イノベーション推進事務局）
・内閣官房（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
・文部科学省
・厚生労働省
・農林水産省
・経済産業省
・国土交通省
・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

本日のアジェンダ

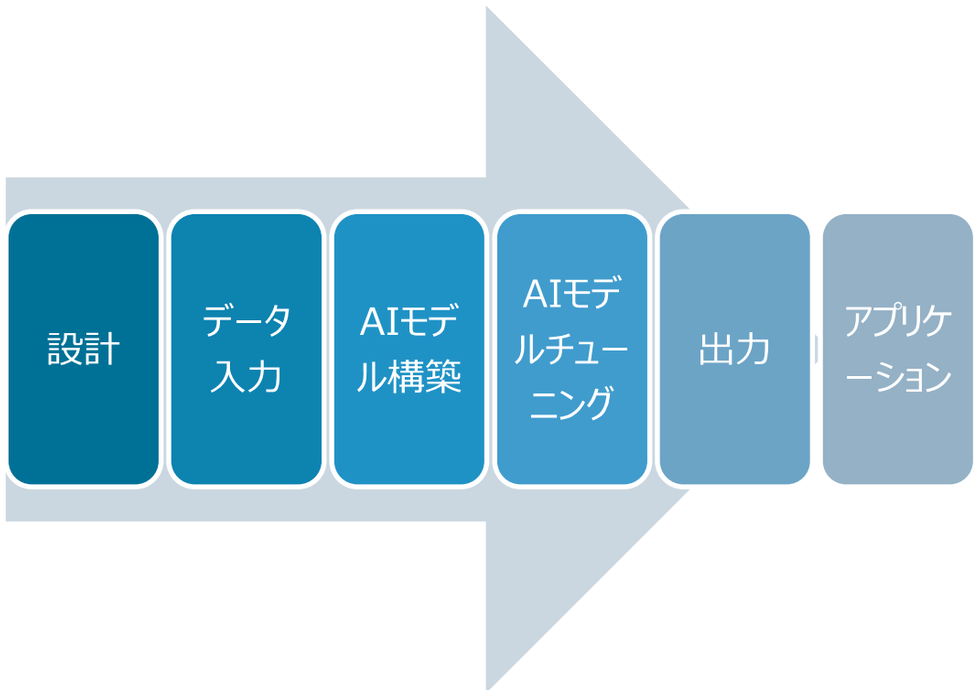
1. SOMPOが目指すAIとの共生
2. AI安全性と考慮しなければならないAIのリスク
3. 日本政府のAI安全性の活動とAI Safety Instituteの紹介
4. AIのリスクや安全性を技術的に捉えるために
5. 東京都のAIに必要な「Safety by Design」

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

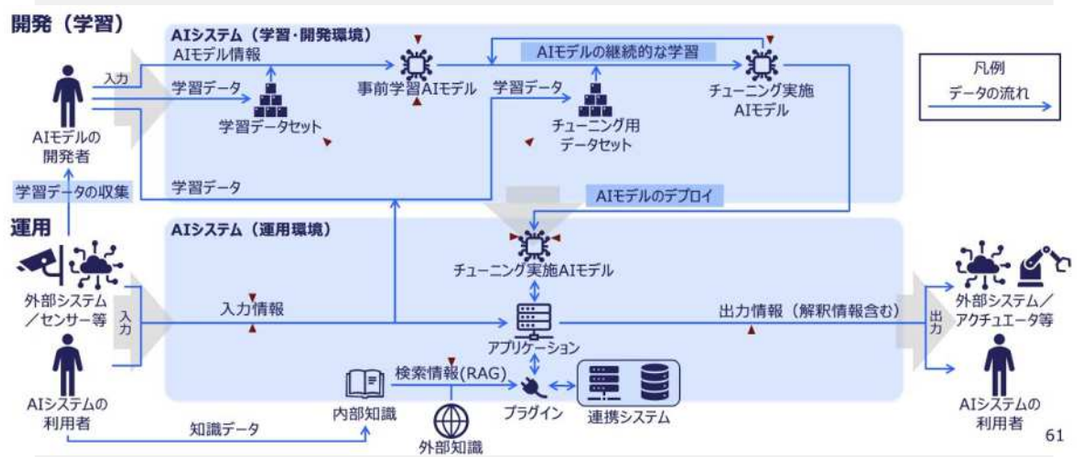
AIのリスクや安全性を技術的に捉えるためには

AISI Japan AI Safety Institute

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



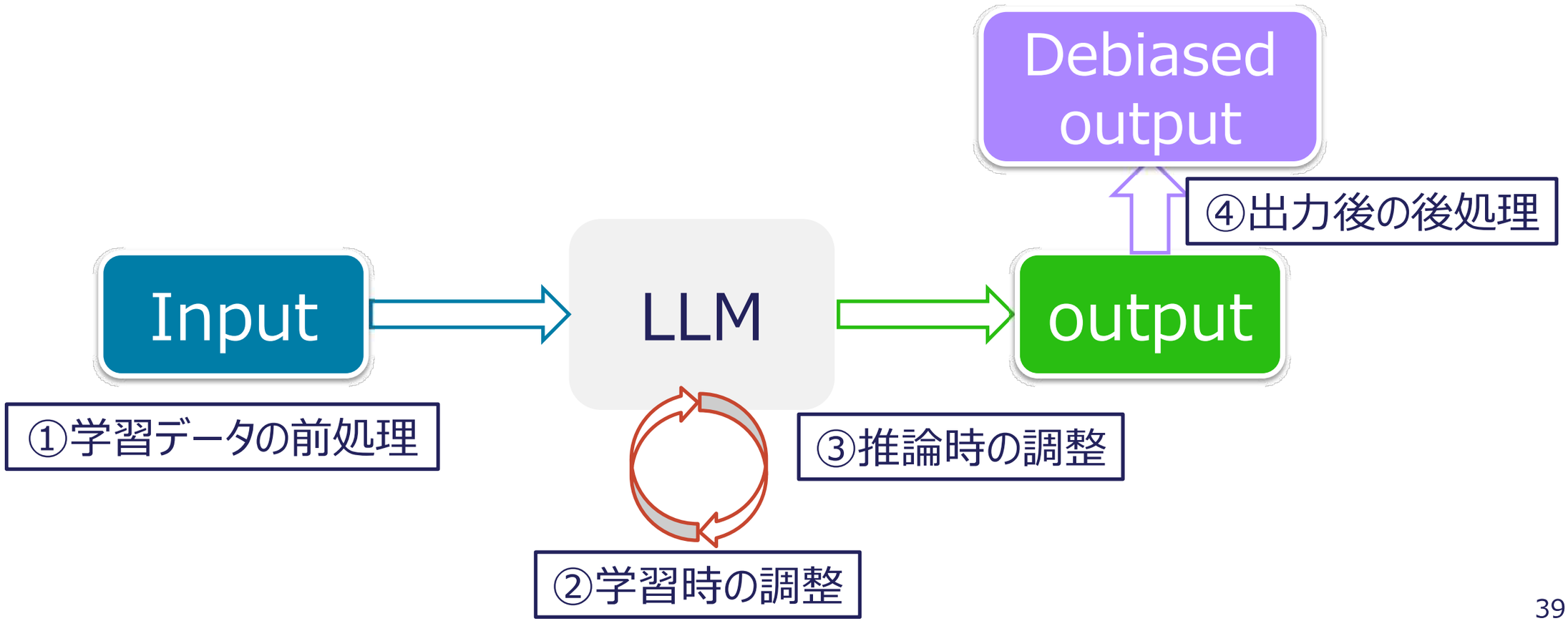
AIシステム全体



リスク回避技術の研究　バイアスへの対応

AISI Japan AI Safety Institute

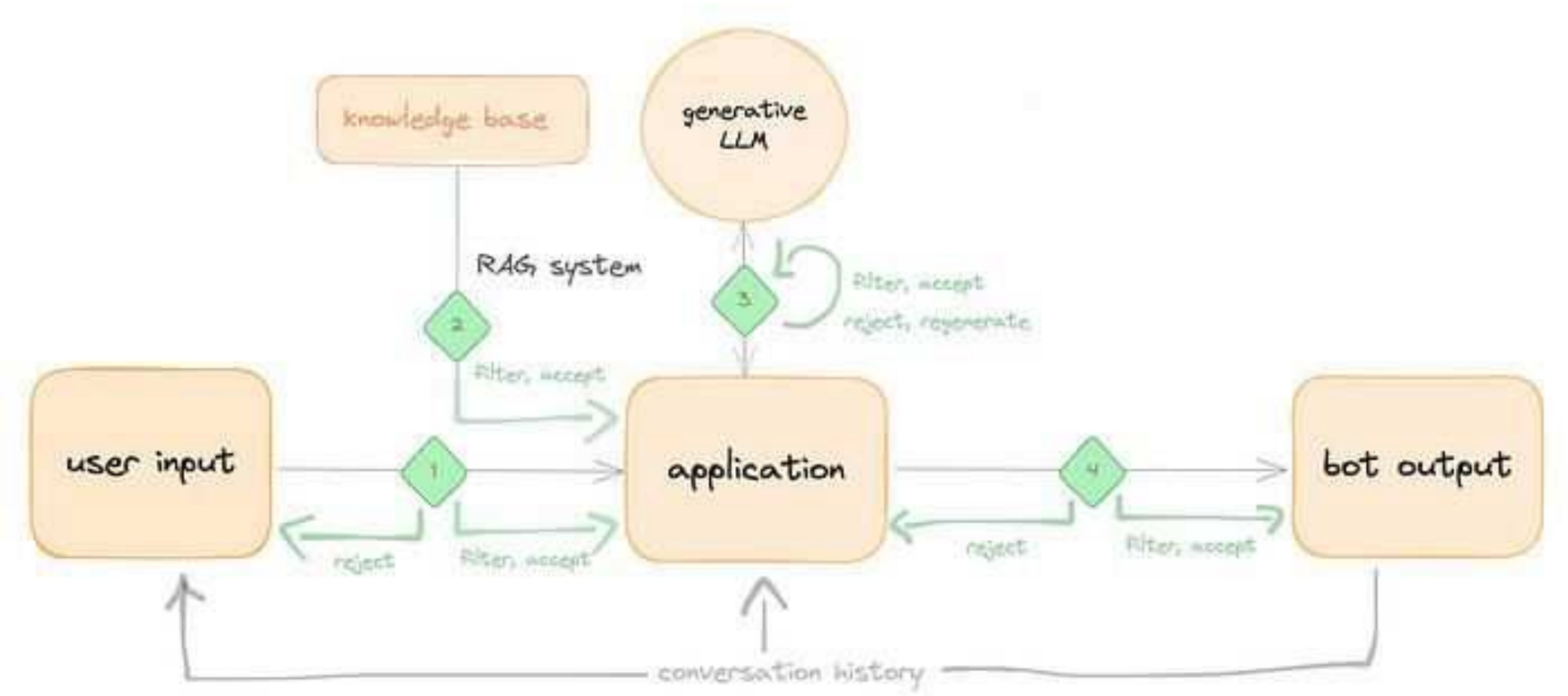
バイアスへの対処技術としては①学習データの前処理、②学習時の調整、
③推論時の調整、④出力の後処理が研究されている



リスク回避技術の研究　LLMガードレールの設置

AISI Japan AI Safety Institute

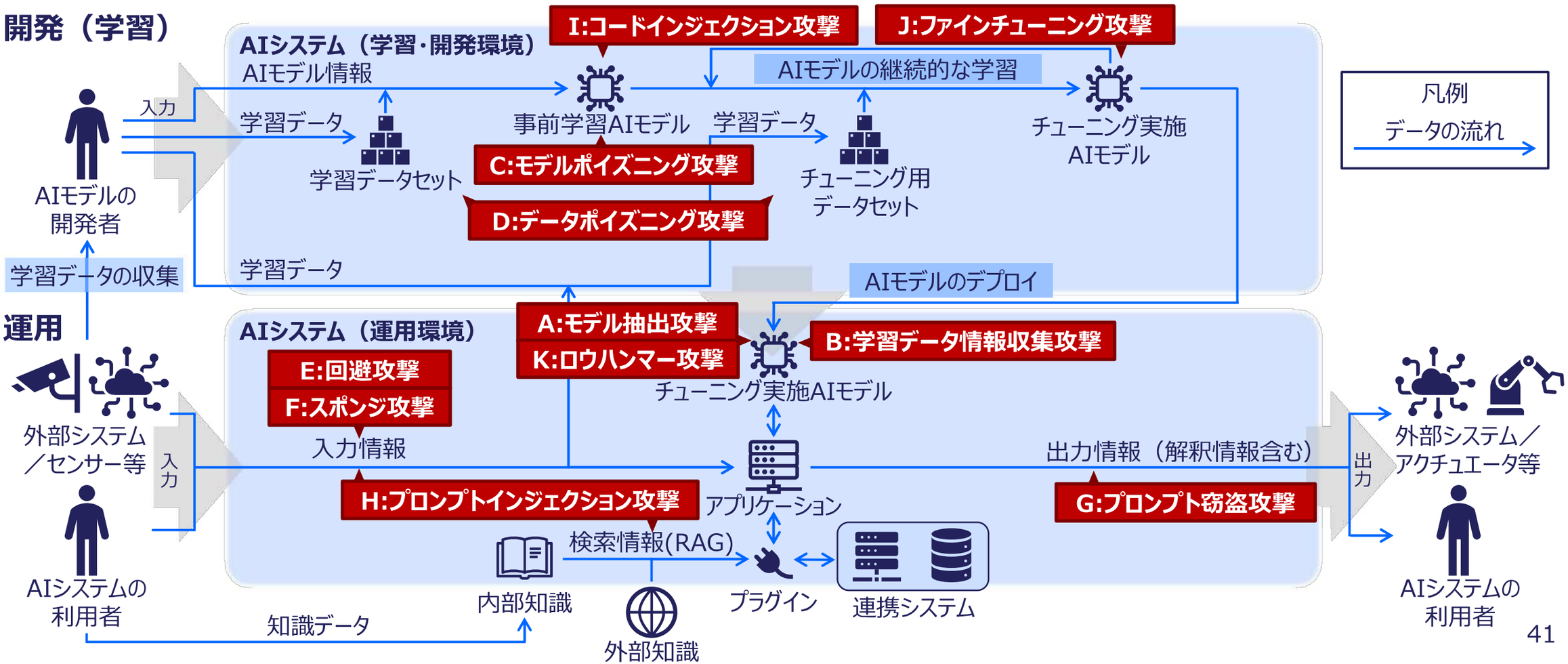
大規模言語モデル（LLM）に対する不適切な入力を制御したり、不適切な出力や誤情報、偏見、攻撃的な内容を生成しないようにする技術



出典：The landscape of LLM guardrails: intervention levels and techniques

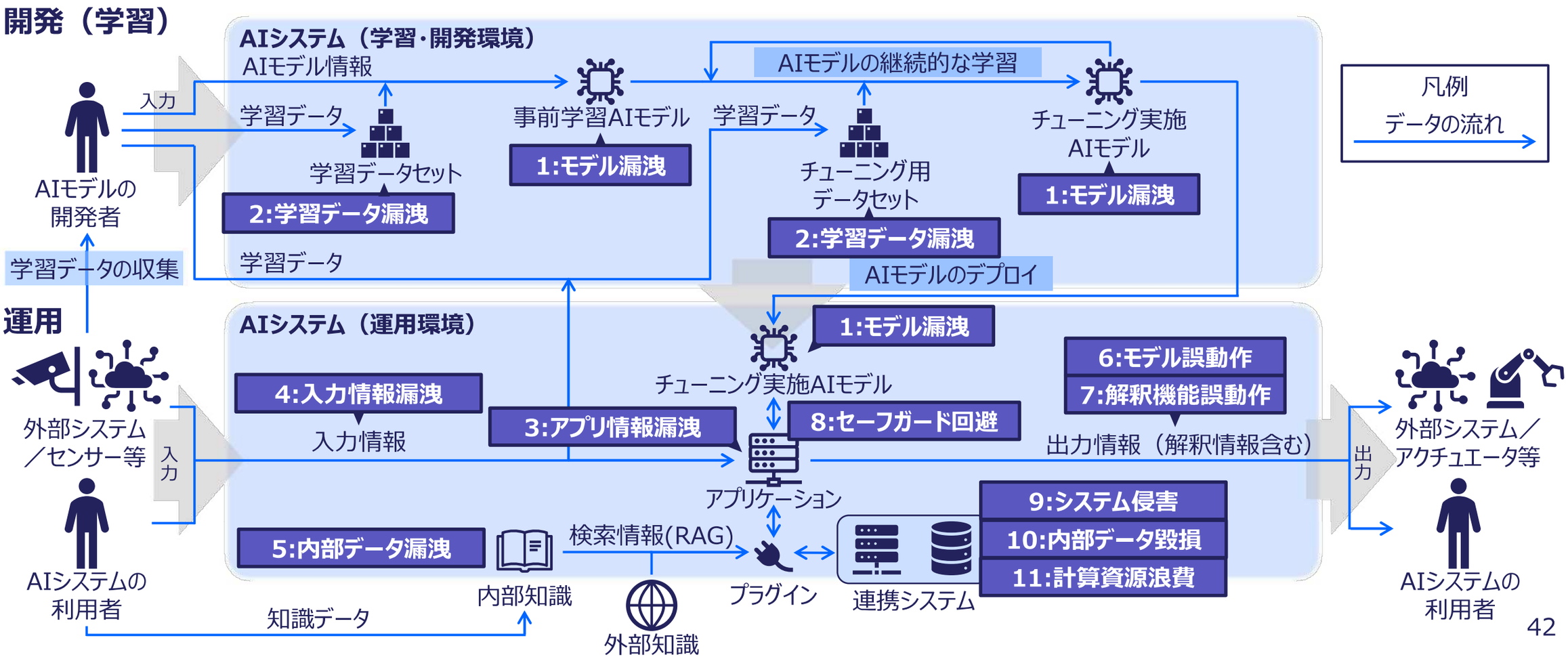
AIシステムに対する攻撃の俯瞰

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。



本日のアジェンダ

1. SOMPOが目指すAIとの共生
2. AI安全性と考慮しなければならないAIのリスク
3. 日本政府のAI安全性の活動とAI Safety Instituteの紹介
4. AIのリスクや安全性を技術的に捉えるために
5. 東京都のAIに必要な「Safety by Design」

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

東京都のAI利活用の考え方

3-④　AI利活用に当たっての考え方

業務の効率化、都民サービスの質向上に寄与する業務について、次の３段階のレベルで利活用を進める

青　比較的リスクが低く、積極的に利活用を進めていく

黄　リスクに十分配慮した上で利活用を進めていく

赤　今後の技術動向や法制度の整備状況等を注視

出典：第3回東京都AI戦略会議　表示資料より抜粋
https://www.digitalservice.metro.tokyo.lg.jp/business/ai-strategy-council/council03

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

東京都のAI利活用の考え方

「都民サービス」「都民サービス関連業務」「職員内部業務」
誰が対象のAIなのかで安全性の考え方は異なる

都民サービス		利活用の考え方
段階1	情報提供	青 比較的リスクが低く、積極的に利活用を進めていく
段階2	行動支援	
段階3	パーソナライズド支援	黄 リスクに十分配慮した上で利活用を進めていく
段階4	一部自動化	
段階5	完全自動化	赤 今後の技術動向や法制度の整備状況等を注視

都民サービス関連業務		利活用の考え方
段階1	定型業務補助	青 比較的リスクが低く、積極的に利活用を進めていく
段階2	データ分析	
段階3	予測・判断支援	黄 リスクに十分配慮した上で利活用を進めていく
段階4	一部自動化	
段階5	完全自動化	赤 今後の技術動向や法制度の整備状況等を注視

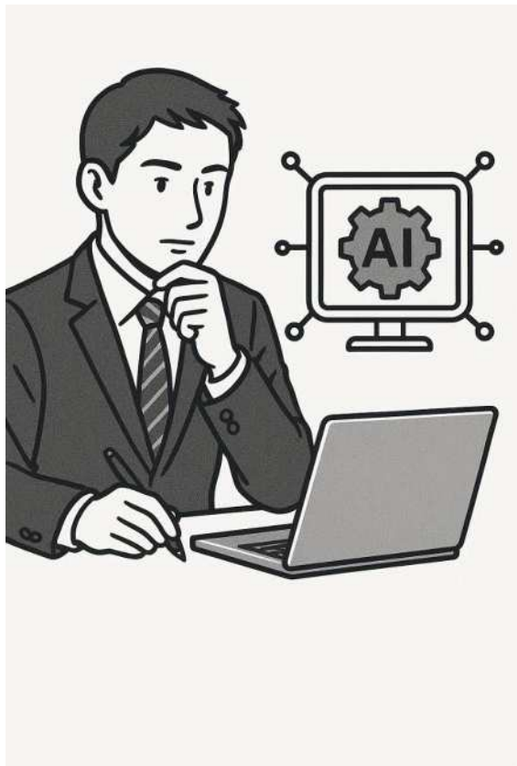
職員内部業務		利活用の考え方
段階1	情報検索・定型業務補助	青 比較的リスクが低く、積極的に利活用を進めていく
段階2	文書作成支援	
段階3	企画案の提案・支援	
段階4	専門知識支援	黄 リスクに十分配慮した上で利活用を進めていく
段階5	完全自動化	赤 今後の技術動向や法制度の整備状況等を注視

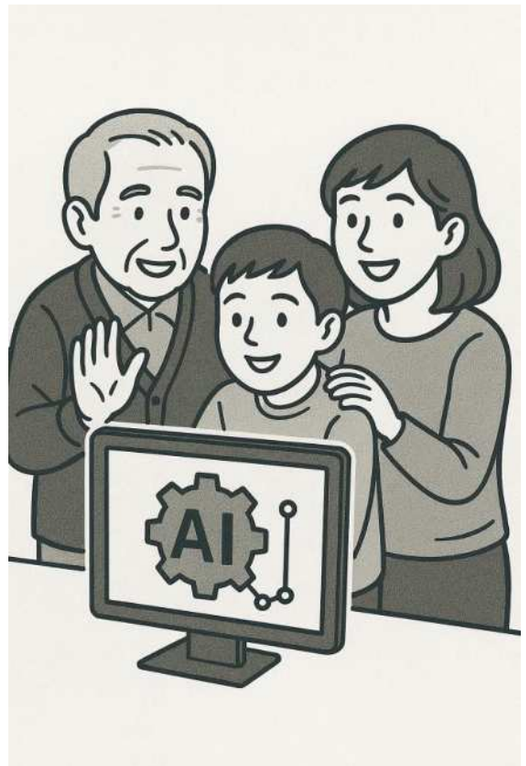
出典：第3回東京都AI戦略会議　表示資料より抜粋
https://www.digitalservice.metro.tokyo.lg.jp/business/ai-strategy-council/council03

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

Safety by Designという考え方

システムの企画や設計の初期段階から安全性を考慮することが重要





© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

Safety by Designに必須なAIガバナンス

SOMPO

4-② ガバナンスの方向性

・AIの利活用促進のため、東京都AI戦略等を随時更新しながら全体統括を行うとともに、効率的なツール活用、各部署のサポートという視点からも各局での活用を後押し

全体統括（随時更新）

全体方針 全件把握等

東京都AI戦略で示した統一方針・ルールに基づくAI活用やツール整備を進めるため、**AIを利用する全施策を政策・財務・技術部門が把握**し、**政策部門が取組の進捗管理、財務部門は予算面での統制、技術部門が必要な技術的支援**を行う

AI導入・活用ガイドライン

各局の課題解決のために**AIを導入するプロセス**や、**AI活用の留意点への具体的な対応方法**などを示した**ガイドラインを年度末までに策定**

AI技術の進展・国の動向・都でのAI活用事例等を踏まえ、戦略・ガイドライン等を随時更新して施策に反映

効率的なツール活用

ツールの共通化

全庁共通する業務など**共通化が可能なものは、デジタルサービス局がツールを一括調達**（AIチャットボット、音声議事録、Copilotなどの共通ツールの使い方や好事例等を共有し、更なる活用を推進

AIアプリの共通化

AIアプリについても、共通化を進めていくため、**GovTech東京が構築した、生成AIプラットフォームを積極的に活用**していく

各部署のサポート

職員のAIリテラシー向上研修

AI利活用の目的の他、安全性やセキュリティ、ハルシネーションなど**AIの業務活用の留意点やリスク対策**などを内容とする**AIリテラシー向上研修**を、職員向けに実施

相談窓口

AIの利活用やリスク対策に関する相談窓口
海外都市や他自治体の先進事例、AIの技術動向や民間AI関連サービス、国等の規制・ルールなどを調査・把握した上で、助言を行う

© 2025 Sompo Japan Insurance Inc. All Rights Reserved.

出典：第3回東京都AI戦略会議　表示資料より抜粋 https://www.digitalservice.metro.tokyo.lg.jp/business/ai-strategy-council/council03

AIの最大のリスク　それはAIを恐れて使わないこと

AISI Japan AI Safety Institute

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


SOMPO