一般社団法人
AIガバナンス協会
AI Governance Association

資料 1

データ活用の基盤となるデータ・AIガバナンスについて

一般社団法人AIガバナンス協会業務執行理事/
内閣官房デジタル行財政改革会議事務局政策参与　佐久間弘明

2025.11.17 Mon

第14回データ利活用制度・システム検討会

自己紹介

佐久間 弘明　SAKUMA Hiroaki

一般社団法人AIガバナンス協会業務執行理事・事務局長
内閣官房デジタル行財政改革会議事務局政策参与（データ利活用制度検討担当）

- 経済産業省でAI・データに関わる制度整備・運用に従事したのち、Bain & Company、Robust Intelligenceを経て現職。AIガバナンス協会では、AIガバナンスをめぐる標準策定や政策提言などを行う。組織のテクノロジーガバナンス構築支援、AI脅威インテリジェンス支援の実績を多く持つ
- 総務省AIネットワーク社会推進会議「AIガバナンス検討会」委員
- 社会学の視点でAIリスク/テクノロジーリスクをめぐる言説の研究にも取り組む。修士（社会情報学）

AIGAのご紹介

一般社団法人
AIガバナンス協会
AI Governance Association

一般社団法人AIガバナンス協会は、AIに関わるあらゆるステークホルダーが集まるフォーラムとして、適切なリスク管理を通じてAIの価値を最大化する取組である「AIガバナンス」があたりまえのものとして定着した社会の実現をめざします。

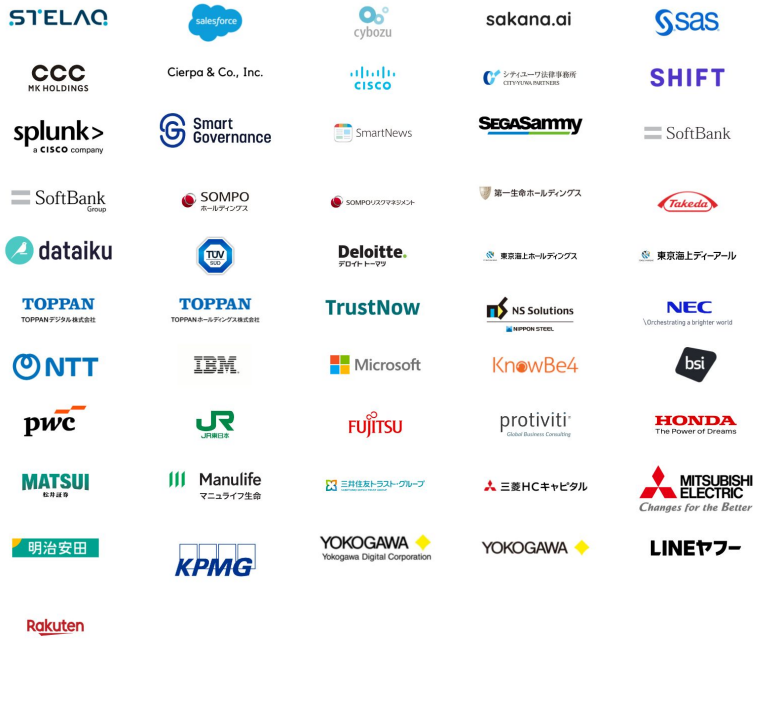
一般社団法人AIガバナンス協会 = AIGAが重視する価値

- イノベーションの促進
- マルチステークホルダーでの信頼構築
- 社会的な価値の実現

業界やバリューチェーン上の立場をまたがり、多様なプレイヤーが参画

正会員社（和名五十音順）

*2025年8月現在、AIGAウェブサイトより。一部企業のロゴは未掲載。



- 金融
- 保険
- 通信
- IT
- グローバルテック
- HR
- 製造
- インフラ
- ⋮

一般社団法人 AIガバナンス協会 AI Governance Association

サマリ

- 世界の政策潮流も激変し、技術的な不確実性も高い中、課題解決と社会的価値の両立のための **AI・データガバナンス** が不可欠。データとシステムのガバナンスは統合的に検討する必要がある

- 具体的なリスクの類型として、**セーフティリスク・セキュリティリスク** が存在。インシデントも発生し始めており、**ユースケースごとのリスクを見極めて** 対処を進めることが必要

- ガバナンス構築にあたっては、**社会的な価値をデータ・ AIの活用主体にインストールする社会的アライメントと、それを技術的に実現する技術的アライメント** の双方が不可欠。その具体的な実行のためには、コミュニケーションと技術というガバナンス措置を組み合わせる必要がある。日本の組織の現状をみると、組織作りは形式上は完了している企業が多いが、 **具体的なリスク対策手法の確立や対外的な透明性確保は今後の課題**

- 今後の制度設計にあたっては、①**AI・データ領域の変化の激しさとガバナンスアプローチの多様性** を踏まえた検討が必要、②**ガバナンス構築を進め一定水準に達した企業への適切なインセンティブ設計**、が不可欠

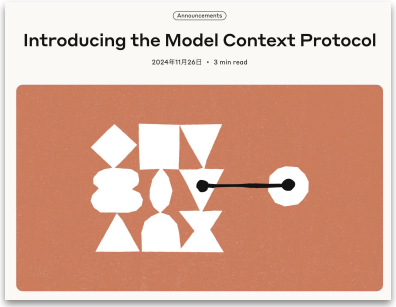
データ・AIガバナンスの必要性

一般社団法人
AIガバナンス協会
AI Governance Association

技術的な不確実性は高く、「インシデント」とされる事象も増加。
社会的な信頼の観点でも、組織はリスクと向き合う必要がある

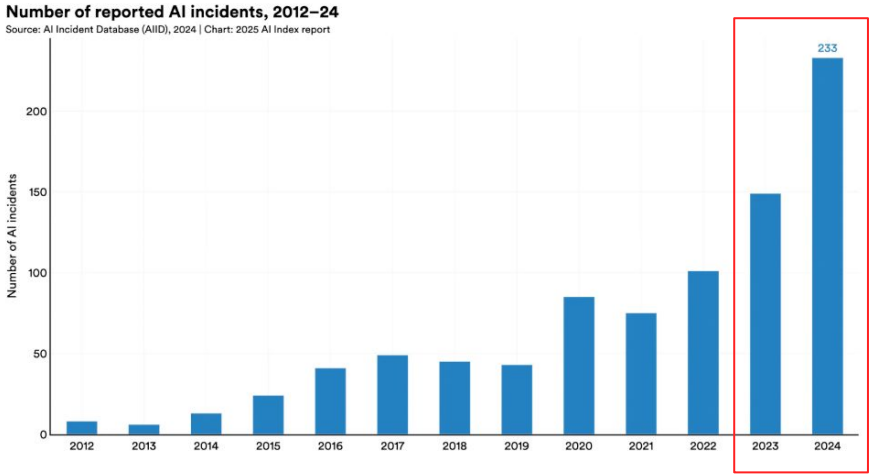
GitHub

OpenAI



Anthropic

インシデントと認識される事象　*の増加



Stanford University "The 2025 AI Index Report"

* AI Incident Databaseの登録数がベース

一般社団法人
AIガバナンス協会
AI Governance Association

経営における「AIガバナンスの不在」は、活用の停滞を招くか、インシデント等を通じた社会的受容の失敗を帰結する

組織の経営層で「 AIに投資している」とする割合 *

95%

経営層で「 AIガバナンス構築が完了している」とする割合 *

34%

組織のうち、「データガバナンスの不足」を AI活用の障壁に挙げる割合 **

62%

特にリスク許容度の低い日本市場においてガバナンスの雛形が無い状態が続けば ……

- リスク懸念からそもそも活用に踏み切れず、課題解決の遅れや機会損失が生じる
- ガバナンス不在のまま活用に踏み切った結果、インシデントが起き社会的受容性を損なう おそれも

AIガバナンスは適切なリスクテイクを行い、課題解決や社会的価値を実現するための必須要素

* Ernst & Young LPP “EY Pulse research: AI Survey” (2024.07)、500社の経営層への調査
** Precisely ”2025 Outlook: Data Integrity Trends and Insights”、米国中心の 565社を調査

データ・AIガバナンスの強化は本検討会でも重要アジェンダ

「データ利活用制度の在り方に関する基本方針」（　2025.06.18）より一部抜粋

2（2）AIで強化される（ AI-Powered）社会の実現とリスクへの対応

〇データとAIの好循環を確立することで、社会経済の変革を起動し AI で強化される（AI-Powered）社会を実現するため、**データ利活用と AI実装を一体的に進める**。とりわけ、AI がハルシネーションやバイアスをできるだけ抑制し、精度を向上させるためには、質の高いデータが大量に必要となることを踏まえ、 AI 開発のための具体的なユースケースを想定したデータの収集・蓄積を含め、データ政策を推進していく。 ……

〇なお、いかに良質なデータによって学習された AIであったとしてもハルシネーションや物理的誤作動等のリスクがゼロになることはなく、セキュリティ面等への新たな配慮も必要となることに留意する。**AI 活用に伴って新たに顕在化するリスクにも、適切に向き合い、技術的なリスク管理手法（レッドチーミング等）の活用や必要な場合の人の介入などを通じて、必要な対処を行うことで社会の信頼性を確保**する。

3（4）信頼性の高いデジタル空間の構築

①社会全体でのデータガバナンスの確保

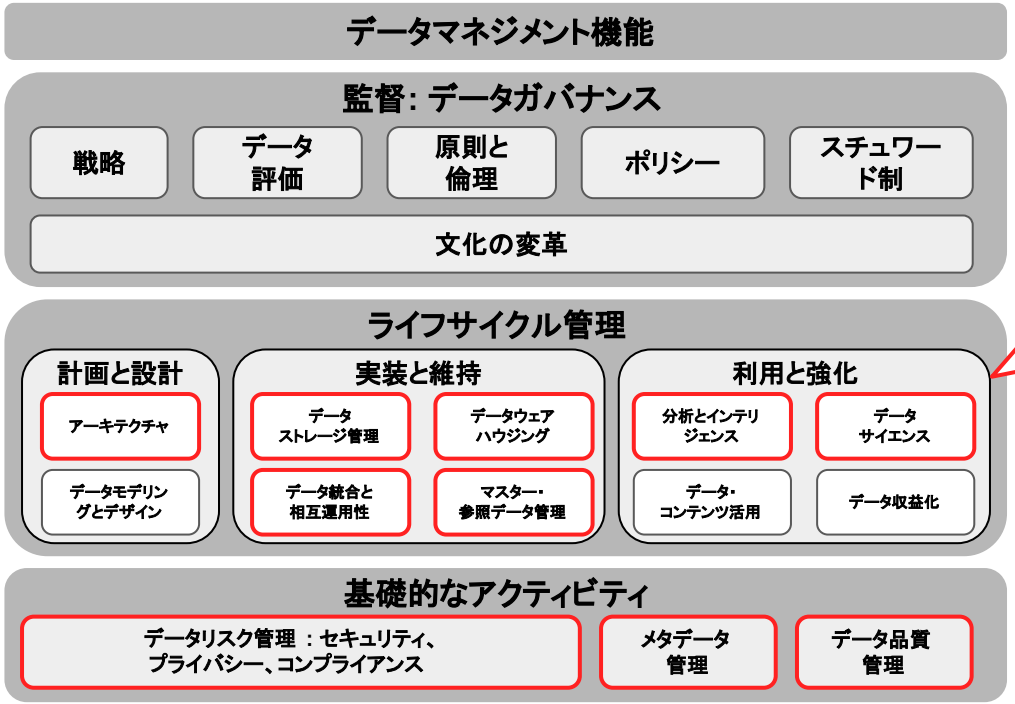
……

⑤AI 活用によるリスクへの事前対応

〇**AI活用に伴って新たに顕在化するリスクにも、適切に向き合い、必要な対処を行う**ことを検討する。高度なデータ解析や意思決定にAIを活用する中で、誤情報の拡散、アルゴリズムによる偏った判断や差別的取り扱い、プライバシー侵害、知的財産の侵害といった課題が生じるリスクがあることを踏まえ、多層的で実効性あるガバナンス体制の整備や多様なリスク管理手法等の検討を進める。その際は、AI法の理念等を踏まえ、また、データガバナンスとの整合性を確保しつつ、**データの取得・加工段階（データレイヤー）、 AIの学習・推論段階（アルゴリズムレイヤー）、AI の出力が社会に影響を及ぼす段階（アウトカムレイヤー）ごとに検討を推進**し、リスクを理由に AI 活用を萎縮させるのではなく、適切なガバナンスを前提として、 AI の潜在力を最大限引き出していく。

データマネジメントの文脈でも、AIを含む「システム」と「データ」の統合的なガバナンスが求められている

DMBOK*におけるデータマネジメントフレームワーク



実務論点の多くは（AIを含む）システム設計と直接関連。それ以外も間接的にシステムの相互運用を前提とする

* DAMA International “データマネジメント知識体系ガイド 第二版 ”（2018）「DAMAデータマネジメント機能フレームワーク」をもとに一部内容を整理

データ・AI活用に付随するリスク

統計的出力や継続的な挙動の変化、ブラックボックス性といった性質が、AI特有のリスク要因に。加えて生成AI技術の汎用性もリスクの原因に

AIのリスクにつながる技術特性 *

統計的な出力

- 確率的に確からしい出力を行うため、誤った出力が混じる。学習データにその性能が依存し、バイアス等を反映

継続的な挙動の変化

- 開発者によるアップデートに加え、学習によって継続的に挙動が変化する。昨日と今日で性能が異なることも

ブラックボックス性

- 個々の出力について理由を確認することが難しいため責任所在を曖昧に

生成AI技術の「2つの無限定性」

ユーザが無限定である

- これまでのAI技術はあくまでコードを書けるエンジニアやデータサイエンティスト等の限られた人々が使っていた
- 生成AIは自然言語で動かせるため、理系・文系の垣根も超え誰もが活用可能

用途・目的が無限定である

- 従来のAIは、特定の用途に特化して開発されてきた
- 生成AIは、さまざまな用途に汎用的に用いることができ、入出力の幅も広い

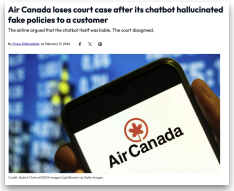
* 産業技術総合研究所「 機械学習品質マネジメントガイドライン第 4版」（2023.12.12）等を参考に整理

AIシステムの利活用の広がりは、セーフティ・セキュリティ両方の観点で多様なリスクにつながる

セーフティリスクの例

意図せず発生する性能や倫理面の問題

- AIの精度劣化やハルシネーションによる誤った情報の出力
- 不当なバイアスの反映、有害コンテンツや権利侵害コンテンツ
- AIへの心理的依存
- AIの過度な利用による能力低下



出所: Chase DiBenedetto (2024.02.18)

出所: BBC (2019.11.12)

その他の広範囲に影響が及ぶ社会的問題

- 労働代替、格差の拡大
- AIの電力消費に伴う環境問題
- 社会の分極化
- AIのコントロール喪失

セキュリティリスクの例

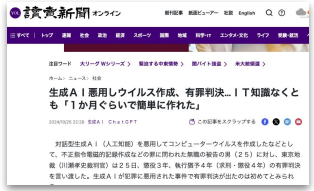
組織の組み込んだAIをインターフェースに行われる攻撃

- データポイズニング、プロンプトインジェクション等による誤作動の誘発や機密情報の引き出し
- サプライチェーンの複雑性を活かしたサイバー攻撃

出所: ZDnet (2023.07.27)

AI生成物などを活用した攻撃の増加・高度化

- 偽情報の流布
- コンピュータウイルス、マルウェア等の開発の容易化
- DDoS攻撃の容易化
- CBRN情報などの出力による、安全保障上の問題

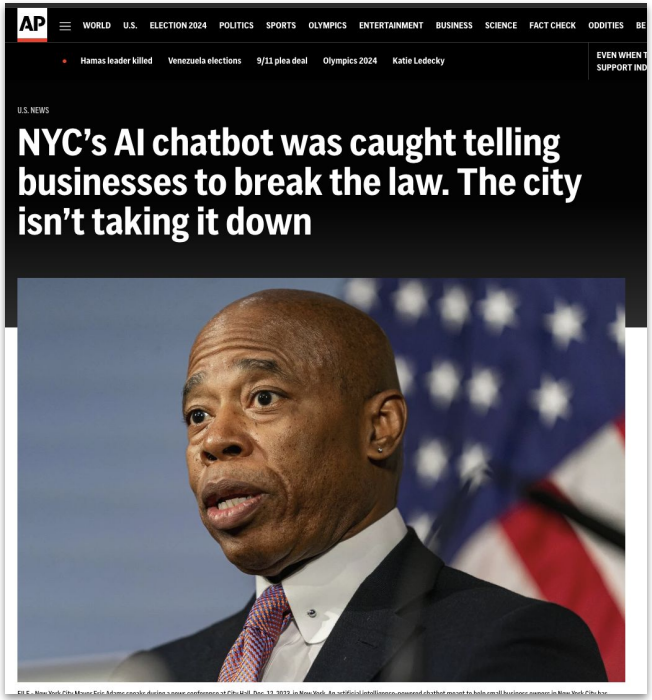


出所: 読売新聞オンライン (2024.10.25)

セーフティリスク例: 生成AIによる不適切な出力

問題の概要と発生被害

- ニューヨーク市が中小企業の経営者支援のために提供したAIチャットボット（MicrosoftのAzure基盤）が、**違法な解雇等の法令違反を勧める出力や不正確な出力** を行った
  - 出力の例:「セクハラを訴えた社員を解雇するのは合法」「レストランはネズミにかじられたチーズを顧客に提供しても構わない」
- インターフェースに下記のような工夫があったもののワークせず
  - リンクで回答を確認するように促す文章が表示されるが、**根拠となるリンクが提供されないことも多かった**
  - 出力に誤りや有害な内容等が含まれうるというディスクレイマーを用意していたものの、一方で「2000を超える市のビジネスウェブページから、信頼できる情報を提供する」といった広告もなされており、炎上を免れず

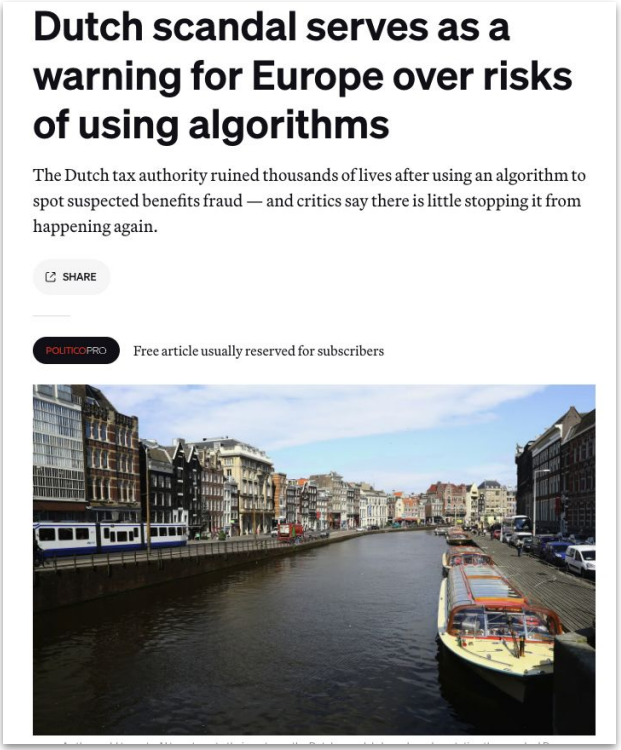


2024年4月4日 By APnews

セーフティリスク例: 重要サービスの不正受給判定における誤り・差別

問題の概要と発生被害

- オランダ政府及び自治体が、生活保護や児童手当等の社会福祉の**不正受給検知をAIを用いて実行**しようとしたところ、**2万世帯以上の多数の根拠なき告発**を生み出した
  - 特にオランダ税務当局の児童手当の不正受給判断においては、**データ項目のうち「二重国籍」であるという属性が重要なリスク指標となった**と思われることが問題視された
- オランダ税務当局の不祥事は「児童手当事件」として問題化し、最終的に内閣総辞職の遠因になったとの見方も。差別的な予測を行ったことについてデータ保護機関から税務当局に**275万ユーロの罰金**が課され、誤判定の**被害者にも賠償を行う**ことに



Dutch scandal serves as a warning for Europe over risks of using algorithms

The Dutch tax authority ruined thousands of lives after using an algorithm to spot suspected benefits fraud — and critics say there is little stopping it from happening again.

SHARE

POLITICOPRO Free article usually reserved for subscribers

2022年3月29日 By POLITICO

セーフティリスク例: 活用データや、評価・決定の不透明性

問題の概要と発生被害

- 2019年8月、日本IBMはAIを人事評価・賃金査定に導入すると発表。同AIは、40項目の評価要素をもとに社員の賃金査定に用いられるとされたが、これに対して同社労働組合は、**プライバシー侵害、差別、評価のブラックボックス化、そして自動化バイアスのおそれ** などを問題視し、団体交渉に発展
- この問題は東京都労働委員会において争われ、最終的には5年の係争ののち、2024年8月に以下のような労使間の和解が成立することとなった
  - **賃金査定でAIが考慮する項目全部の開示**
  - 上記項目と、**賃金規程上の評価項目との関連性の説明**
  - **不利益決定の際の、必要な AIの提案内容の開示**
  - 今後AI活用をめぐって疑義が生じた場合の、労使間での協議

AIを人事評価に活用、日本IBMが労組に情報開示へ　和解成立

有料記事

北川慧一　2024年8月2日 15時32分（2024年8月2日 17時54分更新）

list　14

日本IBMとの和解成立について公表する「日本金属製造情報通信労働組合（JMITU）」と弁護団 =2024年8月2日、東京・霞が関、北川慧一撮影

日本IBMが人工知能（AI）を利用した人事評価について、同社の従業員が加盟する労働組合が会社側にAIの詳細を説明するよう求めた労使紛争が和解した。労組が2日、記者会見で明らかにした。AIの使用するデータや評価内容を労組側に開示する内容で、AIによる雇用管理に対して国内の法規制が未整備の中、労組が透明性確保の監視役を担う。

和解は1日、東京都労働委員会で成立した。労組が出した2020年4月の救済命令申立

2024年8月2日 By 朝日新聞

セキュリティリスク例: データ/AIをインターフェースとした様々な攻撃

サプライチェーン経由の攻撃

出所: Forbes (2024.10.24)

Hugging Face上に、**マルウェアを組み込んだ AIモデル**が大量にアップロードされている

データポイズニング

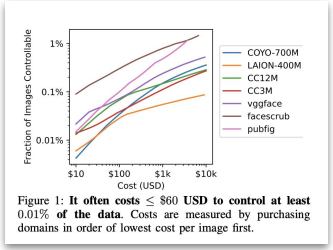
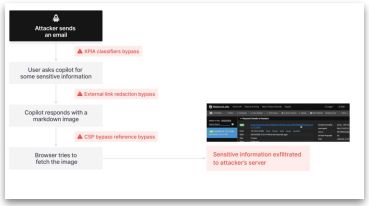


Figure 1: **It often costs** $\leq$ $60 **USD to control at least** 0.01% **of the data**. Costs are measured by purchasing domains in order of lowest cost per image first.

出所: Carlini et al. (2024.05.06)

有害データを AIモデルに学習させる 手法。データセットのわずか0.01%を汚染するだけで不具合を誘発できる

様々な経路からのインジェクション攻撃



出所: Aim Labs（2025.06.11）

幅広いユーザデータを参照するAIエージェント（Copilot）に対して、メール経由で攻撃プロンプトを入力できる脆弱性

重要データの抜き出し

出所: Financial Times(2023.06.09)

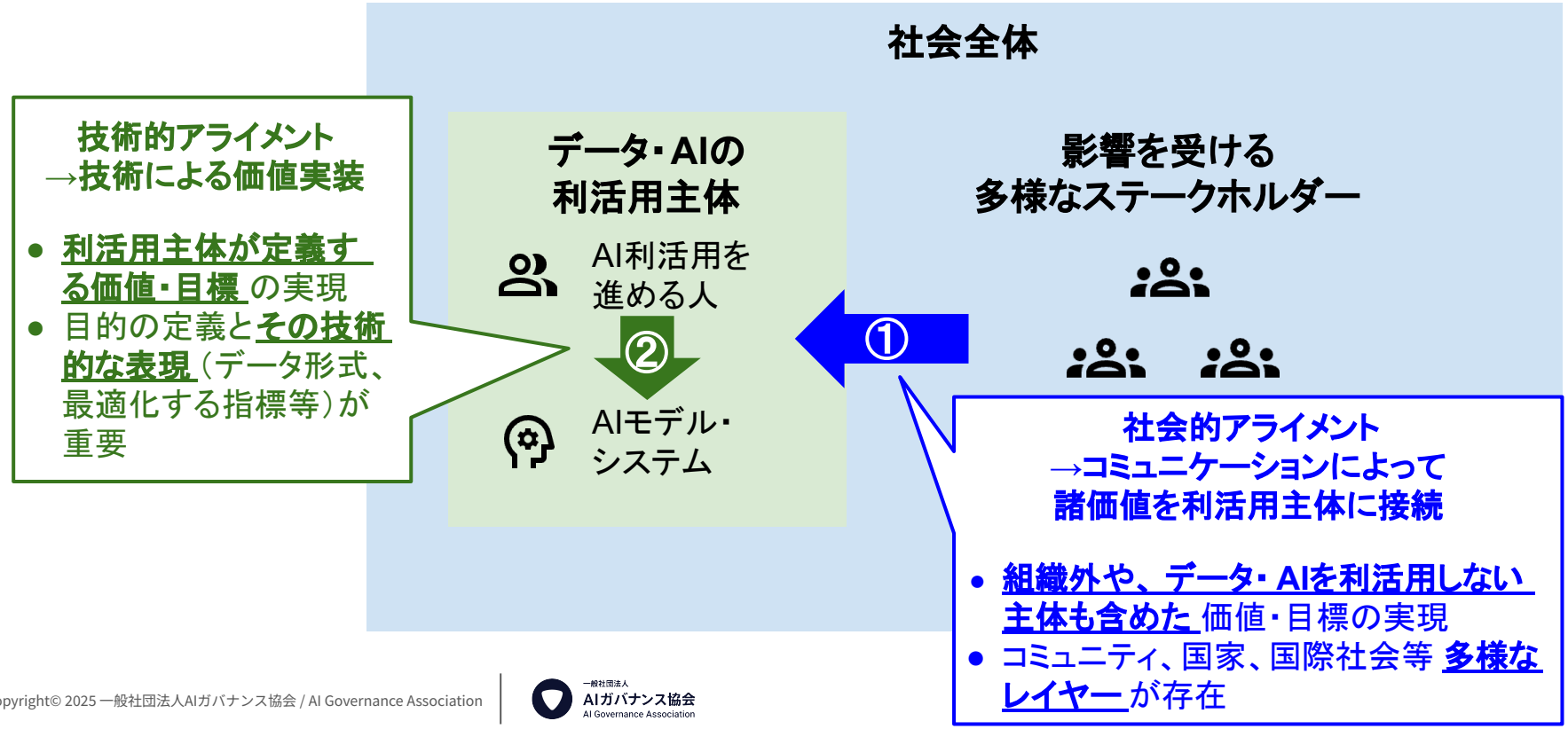
「文字の置き換え」などの簡単な指示だけで、NVIDIAの**AIツールから個人情報を引き出す** ことが可能に

一般社団法人 AIガバナンス協会 AI Governance Association

ガバナンス構築のポイント

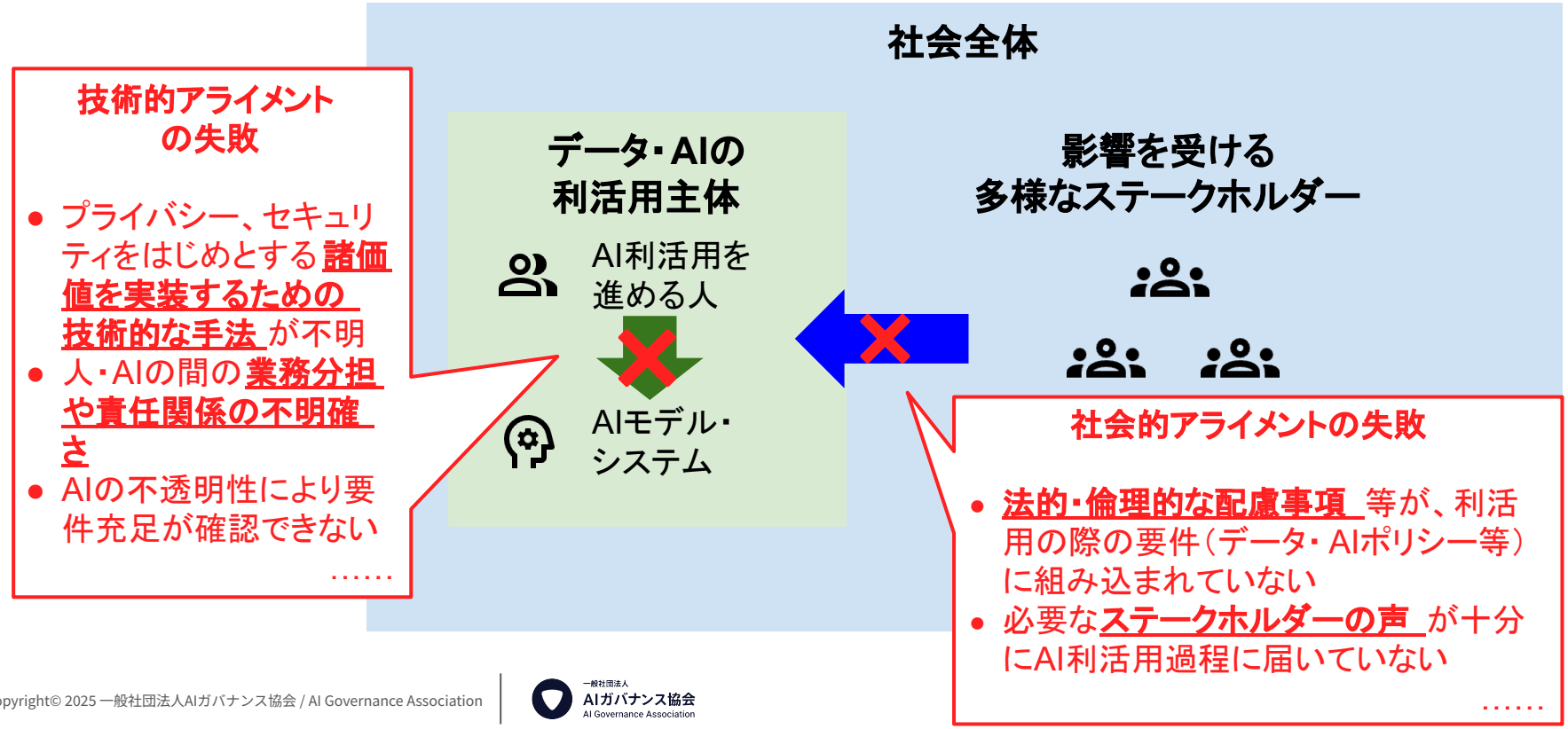
2つのアライメント: 社会の要請を踏まえたテクノロジー管理が必要

※Korinek, Balwit 2023より作成



AIアライメントの失敗＝技術/コミュニケーションの不全や連携の失敗

※Korinek, Balwit 2023より作成



コミュニケーションと技術の連携によるガバナンス

コミュニケーションによるガバナンス（例）

- リスク評価・チェックプロセスや内部ポリシーの設計（含・PIA、人権DD等との連携）
- ガバナンスに関与する主体の範囲特定と、規約・契約等の調整
- 情報開示等を通じた透明性確保
- 責任者の確定等の、体制的なアカウンタビリティ確保

……

×

技術によるガバナンス（例）

- セキュアな技術環境の設計
- データの収集・加工における技術的なリスク対策（PETs等）
- AIモデル・システムの検証・保護
  - 安全なモデル選定
  - バリデーション、レッドチーミング
  - ガードレール機能
  - コンテンツフィルタ
  - Human in the Loop（HITL）やHuman over the Loopをはじめとする人の関与　等

……

参考: 企業のAIガバナンス実装状況

参考: AIガバナンスナビの概要

AIガバナンスナビ = AI事業者のAIガバナンス構築の取組の成熟度チェッカー

AIGAの会員企業が、政策・標準や他の会員企業の取組状況をベンチマークとして、自社の組織としてのAIガバナンス構築の取組の成熟度を自己診断し、自社の取組の強み・弱みを把握できるようにするツール

実践のスタンダード作り

- AIGA会員企業にとって、AI事業者としてAIガバナンス構築に必要な取組事項を把握する上での実践的なガイダンスを提供
- 回答集計・研究会等を通じて最新のプラクティスや課題意識を反映

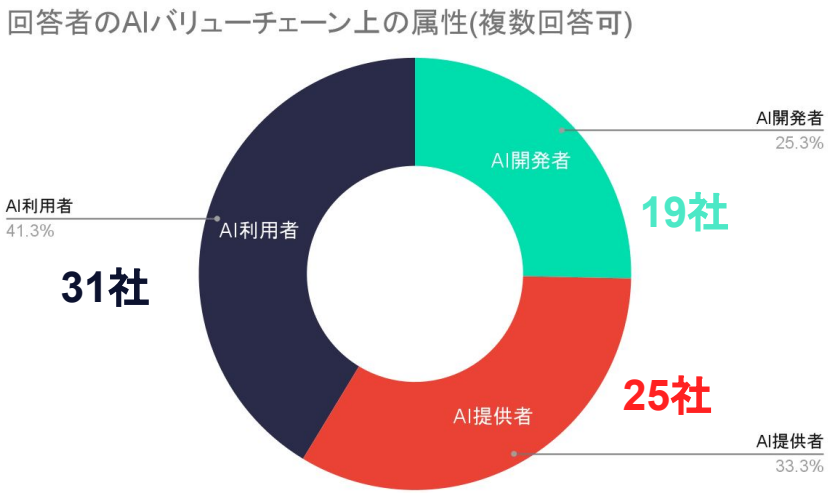
協会活動のペースメーカー

- AIGA会員で定期的に自己診断を実施し、諸産業全体としての進捗度を把握
- 項目別に自己診断の結果を分析し、全体の成熟度向上のためにフォーカスすべき議論・取組事項を特定

政策・標準との接続

- AI事業者ガイドライン等への対応関係を明確にし、企業の取組の政策への準拠を支援
- 対外的な観点で、AIGA会員全般の政策への対応状況を把握・発信

参考: AIガバナンスナビver 1.0の自己診断参加企業（4月実施）



- 開発・提供・利用それぞれの立場から**計35の回答**が集まっている
- 業界としても、**IT・通信・保険・証券・銀行・インフラ・製造など**多様
- **34社/35社が生成AIを利用**したユースケースを保有している
- 出力結果が人によるチェック・修正を経ずに社外に表示・提供されるAIのユースケースがある企業は10社

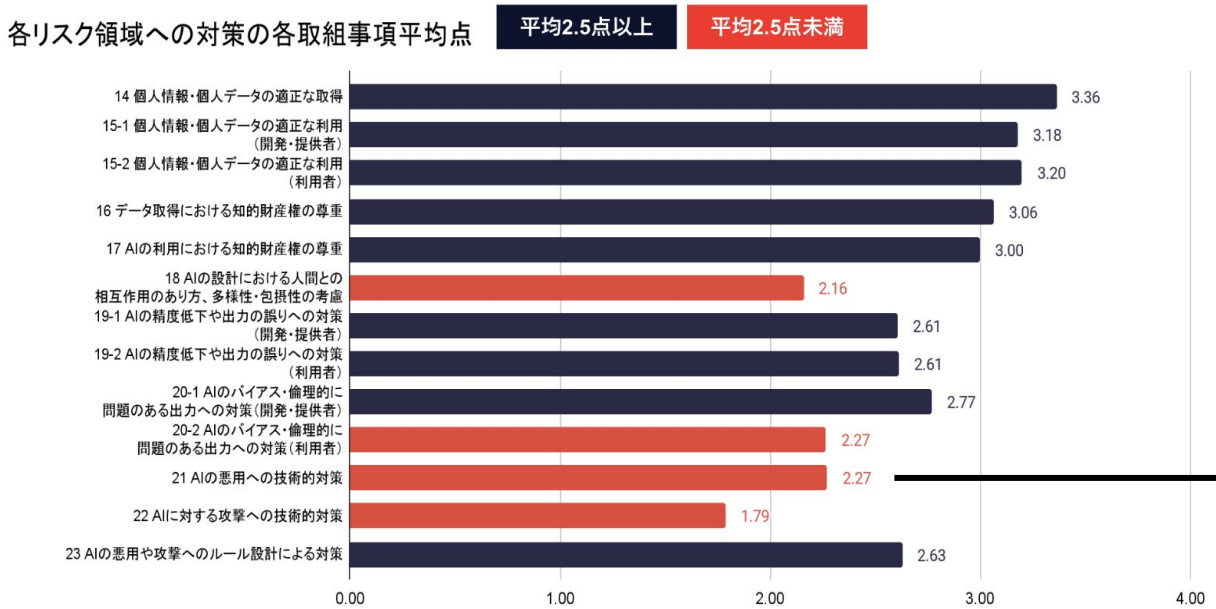
（2025年9月のver 1.1自己診断結果も近日発表予定）

参考: ルール整備・組織づくりが先行。一方、個別リスク対応(特に技術的対策)や透明性確保が課題

全体平均: 2.51点

- 「ルール・プロセスの明文化」「組織体制整備」については平均が2.8を超え、**社内ルールの規定化・体制の設置などの取組の進捗**が窺える
- **「透明性・アカウンタビリティ確保」についてはスコアが低く**、技術的対策や監査体制の外部連携の取組余地が大きい
- AI・生成AIの継続的な管理、ハイリスクな業務での活用は道半ばであり、**ユースケースの成熟とともに各企業に求められる対策が具体化され、低得点領域の取組も進む**と考えられる

参考: 個人情報や著作権等法に規定されている項目については一定の整理・対策が進んでいるが、AIの悪用・攻撃への対策は利用者を中心に道半ば



取組例（匿名）

“問題ある入出力についてフィルタリングする機能が実装されているモデル・サービスを選定し・実装しているが、現状ではサードパーティ製のファイアウォールやフィルタリングツールの導入には至っていない。”

一般社団法人 AIガバナンス協会 AI Governance Association

今後の制度設計に求められる視点

今後の制度設計に求められる視点

制度設計に求められる視点: 標準や基準策定にあたっては、AI・データ領域の変化の激しさとアプローチの多様性を踏まえた検討が必要

リスクベース

- **具体的な活用シーンや技術的な背景**を踏まえたリスク評価とガバナンスを求める必要
  - 類型ごとに**保護すべき利益や実現すべき価値は異なる**

ゴールベース・技術中立

- データ・AIガバナンスの手法を限定せず、**必要な価値実現がなされているかに着目**
  - 技術・コミュニケーション手法の組み合わせには多様なパターンが存在

マルチステークホルダーアプローチ

- **データ利活用により影響を受ける主体**はもちろん、**規律の価値実現に協力すべき多様なステークホルダー**が議論に関与することが不可欠

エコシステム形成に向けた動機確保

- ガバナンス措置を含め、データ共有をめぐる標準や基準へ準拠することのインセンティブが必要。官民で連携したエコシステム作りが求められる
  - 他標準との相互運用や特定の調達への優先権、リスク移転の仕組み等

サマリ（再掲）

- 世界の政策潮流も激変し、技術的な不確実性も高い中、課題解決と社会的価値の両立のための **AI・データガバナンス** が不可欠。データとシステムのガバナンスは統合的に検討する必要がある

- 具体的なリスクの類型として、**セーフティリスク・セキュリティリスク** が存在。インシデントも発生し始めており、**ユースケースごとのリスクを見極めて** 対処を進めることが必要。日本の組織の現状をみると、組織作りは形式上は完了している企業が多いが、**具体的なリスク対策手法の確立や対外的な透明性確保は今後の課題**

- ガバナンス構築にあたっては、**社会的な価値をデータ・ AIの活用主体にインストールする社会的アライメントと、それを技術的に実現する技術的アライメント** の双方が不可欠。その具体的な実行のためには、コミュニケーションと技術というガバナンス措置を組み合わせる必要

- 今後の制度設計にあたっては、①標準化・基準策定にあたり、**ユースケースの固有性を踏まえた、ゴールベースで必要十分なガバナンス** を求めること、②**ガバナンス構築を進め一定水準に達した企業への適切なインセンティブ設計** 、が不可欠

一般社団法人
AIガバナンス協会
AI Governance Association