4th Grand Canvas: AI品質の未来を共に描く

# 生成AIを活用したシステムのリスクマネジメントについて

2025年3月4日

NEC セキュアシステムプラットフォーム研究所
ISO/IEC JTC 1/SC 42（AI） 国際エキスパート
林谷　昌洋

# アジェンダ

1. AIに対する世界の動き
2. AIのリスクとは
3. 生成AIを活用したシステムのリスクマネジメント
4. NECの取り組み

# AIに対する世界の動き

# 広がるAIの社会実装

# AIの誤った利活用により発生可能性のある社会への不利益

## AIの誤った利用は、差別やプライバシー侵害などの人権課題を引き起こし、生活者や消費者に多大な影響を及ぼす可能性がある

### 生活者や消費者が被る不利益の例

**人種・性差別**
- 人材採用で女性が不利になる
- 犯罪リスクで黒人の犯罪確率が高く評価
- 顔認証の精度問題が差別を助長

**プライバシー・自由の侵害**
- 個人の行動や趣味嗜好を同意なく把握
- 国家による監視の乱用につながる

**生命・身体の侵害**
- 自動運転車が事故を起こした場合の責任の所在が不明
- 使用者の命令によらず自律的に稼働し、人間に危害を加える恐れがある

### 事業者への影響の例

**法令違反に対する疑い**

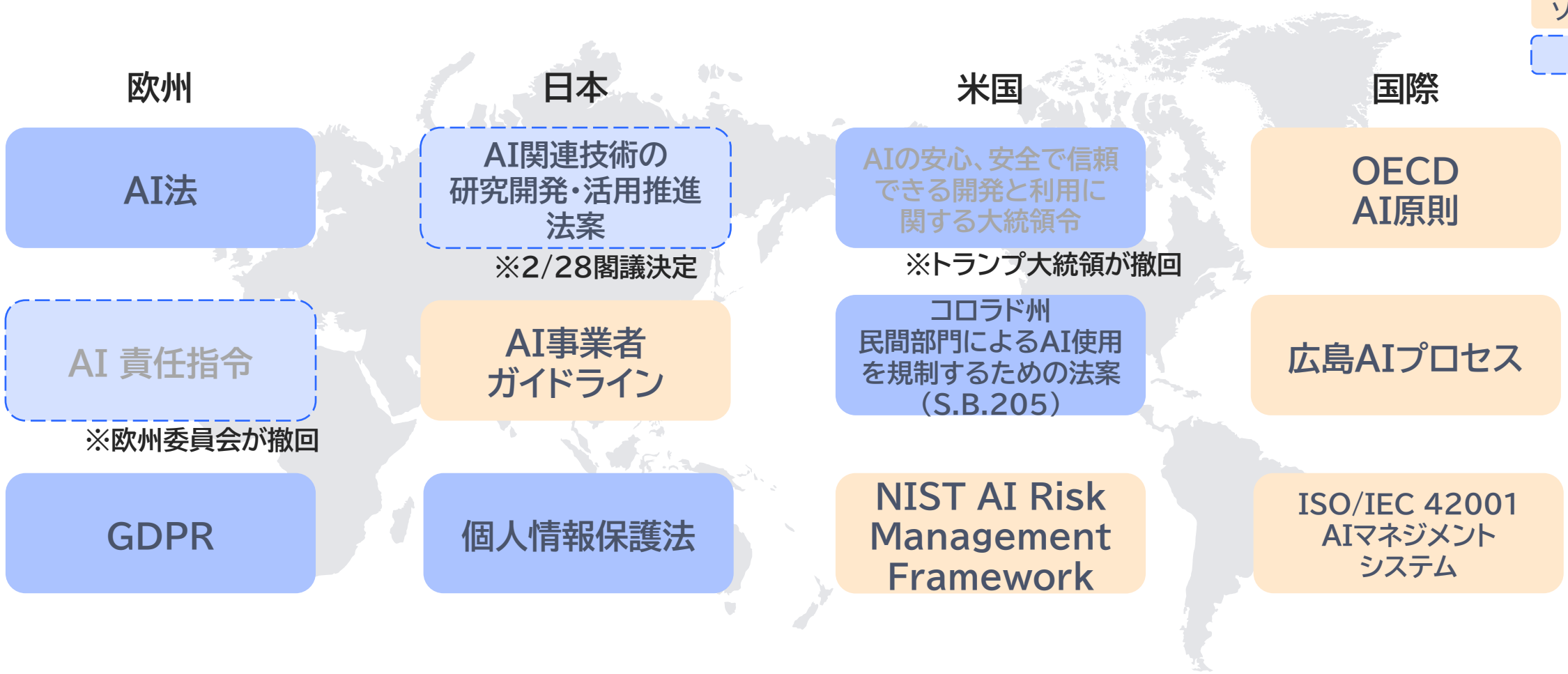
**社会的評判・信頼が低下**

**技術に対する不安の増大**

# 事業を取り巻く環境

## AIがもたらす価値の最大化とリスクの極小化を目指し、様々なレギュレーションが存在
## 既存の法令への対応のみならず、新たなレギュレーションへの対応が必要

ハードロー
ソフトロー
検討中



# ほかの世界の動き

## 英、米、日でAISI（AIセーフティ・インスティテュート）が誕生
## 韓国、シンガポールにも動き

### 英国　AISI

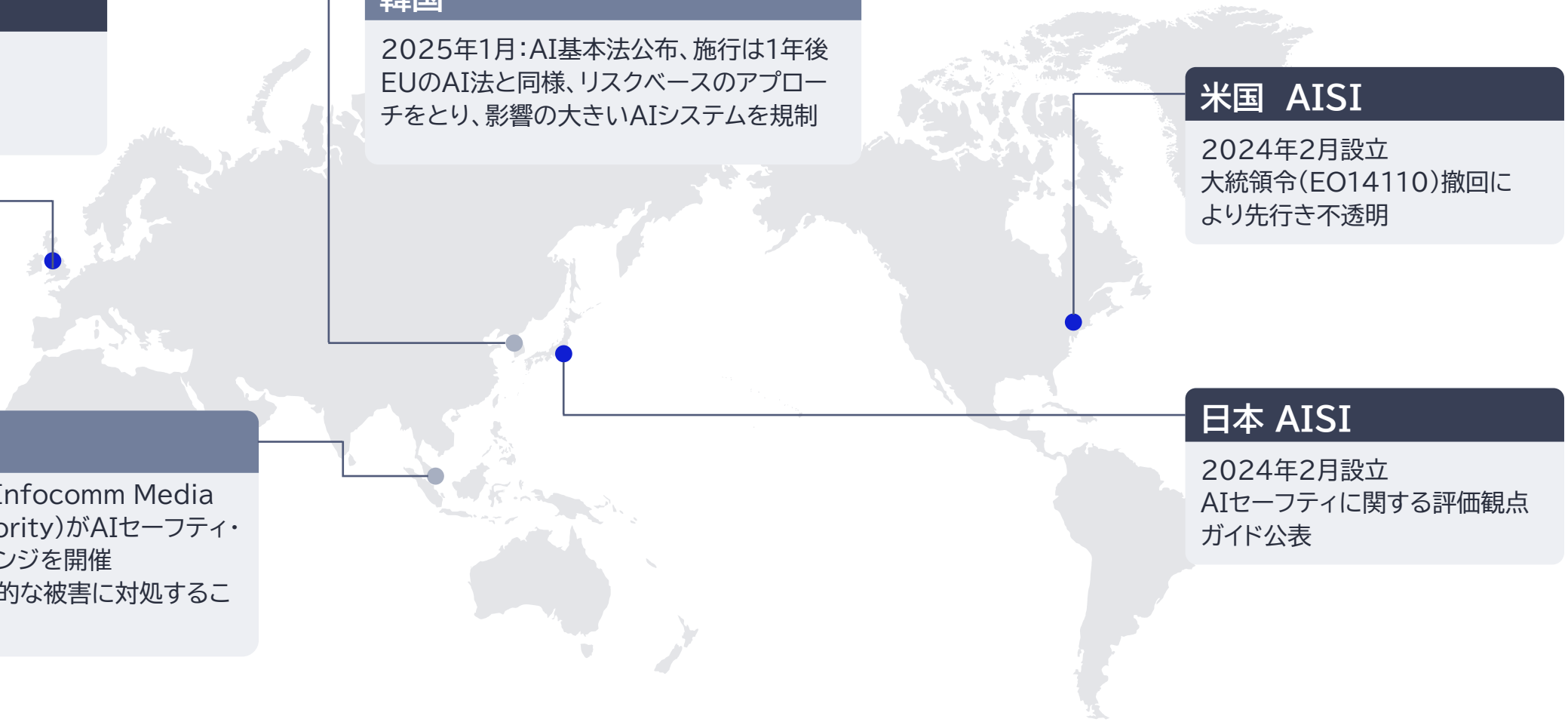
2023年11月設立
現在：AIセキュリティ・
インスティテュート

### 韓国

2025年1月：AI基本法公布、施行は1年後
EUのAI法と同様、リスクベースのアプローチをとり、影響の大きいAIシステムを規制

### 米国　AISI

2024年2月設立
大統領令（EO14110）撤回により先行き不透明



### シンガポール

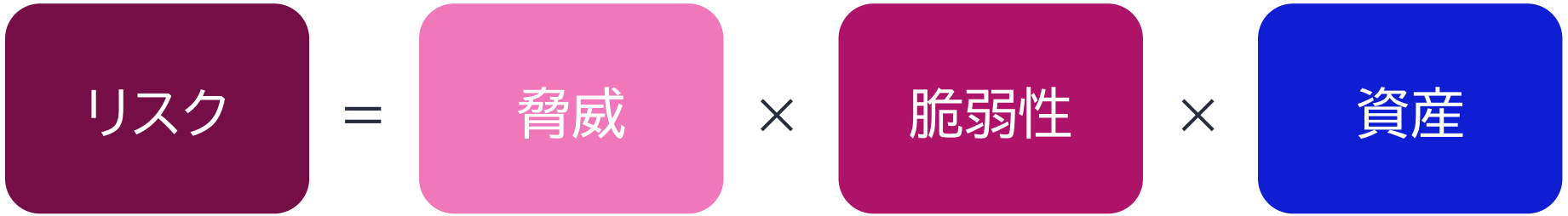
2024年11月：IMDA（Infocomm Media Development Authority）がAIセーフティ・レッドチーミング・チャレンジを開催
欧米中心ではなく、地域的な被害に対処することが目的

### 日本 AISI

2024年2月設立
AIセーフティに関する評価観点ガイド公表

# AIのリスクとは

# リスクとは




- 脅威
  - 該当の脆弱性に対して攻撃者視点から攻撃可能か
- 脆弱性
  - 該当の脆弱性がシステムやデータに与える潜在的な影響の大きさ
- 資産
  - 該当の脆弱性を保有するシステムの重要度

脆弱性対応における リスク評価手法のまとめ:
https://www.ipa.go.jp/jinzai/ics/core_human_resource/final_project/2024/f55m8k0000003v30-att/f55m8k0000003v94.pdf

# AIのリスク

AIによる便益は広がる一方で利用の拡大および新技術の台頭に伴い、それらが生み出すリスクも増大

| リスク | AI事業者ガイドラインにおける指針 |
|---|---|
| バイアスのある結果および差別的な結果の出力 | 人間中心、公平性 |
| フィルターバブルおよびエコーチェンバー現象 | 人間中心 |
| 不適切な個人情報の取り扱い | プライバシー保護、人間中心 |
| 生命・身体・財産の侵害 | 安全性、公平性 |
| データ汚染攻撃 | セキュリティ確保 |
| ブラックボックス化、判断に関する説明の要求 | 透明性、アカウンタビリティ |
| エネルギー使用量および環境の負荷 | 人間中心 |

# 生成AIのリスク

**生成AIの普及に伴い、偽情報・誤情報の生成・発信等のリスクの多様化・増大が進むほか、知的財産権の尊重を求める声も高まっている**

| リスク | AI事業者ガイドラインにおける指針 |
|---|---|
| 機密情報の流出 | セキュリティ確保、教育・リテラシー |
| 悪用 | 安全性、教育・リテラシー |
| ハルシネーション | 安全性、教育・リテラシー |
| 偽情報・誤情報を鵜呑みにすること | 人間中心、教育・リテラシー |
| 著作権との関係 | 安全性 |
| 資格等との関係 | 安全性 |
| バイアスの再生成 | 安全性 |

# 生成AIを活用したシステムのリスクマネジメント

# 生成AIを活用したシステムのリスクへの対応

## プロジェクトマネジメントの知識を体系化しているPMBOK（Project Management Body Of Knowledge）のプロジェクトリスクマネジメントを参考

| リスク対応 | 概要 |
|---|---|
| 回避 | リスクを完全になくして排除 |
| 転嫁 | リスクを第三者に移管 |
| 軽減 | リスクを受容可能なレベルまで低減 |
| 受容 | 事前対処はせず、リスクが顕在化した時点で対処 |
| エスカレーション | 他の組織に報告 |

# 生成AI活用システムに対するリスク対応の例

| リスク対応 | 提供元・サービス | 生成AIのリスク |
|---|---|---|
| 回避 | Microsoft・Azure OpenAI Service | 機密情報の流出 |
| 転嫁 | あいおいニッセイ同和損保・生成AI専用保険 | 機密情報の流出<br>ハルシネーション<br>著作権との関係 |
| | Google・訴訟リスク対応 | 著作権との関係 |
| | Adobe・Adobe FireFly | 著作権との関係 |
| 軽減 | Citadel AI・Len for LLM | 悪用<br>バイアスの再生成 |
| | Amazon・Amazon CodeWhisperer | 悪用 |
| | NEC・ハルシネーション、誤情報対策 | ハルシネーション<br>偽情報・誤情報を鵜呑みにすること |
| 受容 | ー | ー |
| エスカレーション | NEC・AIガバナンス | ー |

# リスク回避の例

## Microsoft: Azure OpenAI Service

AzureサービスとしてMicrosoftによって運営。Microsoftは、MicrosoftのAzure環境でOpenAIモデルをホストしており、AzureサービスはOpenAIが運営するサービス（例：ChatGPT、またはOpenAI API）とは一切やり取りを行わないことにより、リスク回避を実現。

顧客のプロンプト（入力）および補完（出力）、顧客の埋め込みおよびトレーニングデータは、

・他の顧客には利用できない。
・OpenAIには利用できない。
・OpenAIモデルの改善には使用されない。
・Azure OpenAI Serviceの基礎モデルのトレーニング、再トレーニング、または改善には使用されない。
・顧客の許可または指示なしに、Microsoftまたはサードパーティの製品またはサービスの改善には使用されない。
・顧客が微調整したAzure OpenAIモデルは、お客様のみが利用できる。

Data, privacy, and security for Azure OpenAI Service - Azure AI services | Microsoft Learn:
https://learn.microsoft.com/ja-jp/legal/cognitive-services/openai/data-privacy

# リスク転嫁の例1

## あいおいニッセイ同和損保: 生成AI専用保険

Airchaic(生成AIを使用したサービスを開発・提供)社のサービスを
利用する企業を補償することで、リスク転嫁を実現

| リスク | 補償例 |
|---|---|
| 機密情報の流出 | 生成AI使用に起因して、自社の機密情報が外部に漏洩し、そのことが新聞やテレビ等で報道された場合 |
| ハルシネーション<br>(人格権侵害、名誉棄損、その他不適切表現) | 生成AI使用に伴い、口頭、文書、図画その他これらに類する表示行為による名誉棄損またはプライバシー侵害、その他不適切な表現が新聞やテレビ等で報道された場合 |
| 著作権との関係 | 生成AIを使用し生成した製造物が著作権をはじめとする知的財産権を侵害したとして、権利者から訴訟を起こされた場合(国内限定) |

【国内初】生成 AI のリスクを補償する「生成 AI 専用保険」の提供開始:
https://www.aioinissaydowa.co.jp/corporate/about/news/pdf/2024/news_2024022701277.pdf

# リスク転嫁の例2

## Google：知財関係の訴訟リスクからの保護により、リスク転嫁

- トレーニングデータに関する補償
  - AIのトレーニングデータに著作権を侵害するデータが含まれているといった訴えを起こされたケースに対応する補償
- 出力結果に対する補償
  - Googleの生成AIを利用して出力された結果に対して第三者の知的財産権を侵害しているという申し立てがあった際の補償
    - ユーザーが他者の権利を侵害することを目的としてAIに生成させたデータについては、補償対象外

Shared fate: Protecting customers with generative AI indemnification:
https://cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification?hl=en

# リスク転嫁の例3

## Adobe: Adobe Fireflyによる出力の補償によりリスク転嫁

知財関係の訴訟リスクからの保護のため

- 顧客が使用したFirefly出力について、個人または法人が第三者の著作権、商標権、パブリシティ権、またはプライバシー権の直接的な侵害だとして申立て、訴訟、または法的手続を提起した場合に、Adobeは当該侵害申し立てに対する防御を行う
- Firefly 出力についてのAdobeの損害賠償責任の最大総額
  - Firefly アウトプット 1 点あたり、または、侵害申し立て 1 件あたり10,000 米ドルを上限

Adobe 生成 AI 追加条件:
https://wwwimages2.adobe.com/content/dam/cc/jp/legal/servicetou/Adobe-Generative-AI-Additional-Terms_ja_JP_20240305.pdf

# リスク軽減の例1

## Citadel AI: Lens for LLMs

生成AIのリスクを自動で可視化し、継続的にモニタリングする仕組みを提供することで、安全安心な生成AIの普及と、企業ユーザーによる利活用を促進

- 自動レッドチーム機能の追加
  - 安心・安全な生成AIの活用環境の実現
- カスタムメトリクスの導入
  - ジェイルブレイクの検知など、各アプリケーションの用途に即した評価と制御
- 人手評価機能の拡充
  - ヒューマン・イン・ザ・ループによるダブルチェック

大規模言語モデル向け品質改善ツール「 Lens for LLMs」の商用サービス開始:
https://citadel-ai.com/ja/news/2024/09/30/lens-for-llms-launch/

# リスク軽減の例2

## Amazon: CodeWhisperer

AI により生成されたコードを提案・自動挿入できるツールで、コードをスキャンしてセキュリティの脆弱性を検知

- コードをスキャンしてセキュリティ上の問題点を検知し、その修正候補の提示
  - 例えば Open Worldwide Application Security Project（OWASP）で公開されている脆弱性を含んだコードや、暗号ライブラリのベストプラクティスが満たされていないコードを検知し、迅速に修正を検討
- 脆弱性の検知および修正するコード提案は Java、Python、JavaScriptで利用可能。

AI がコーディングを支援 !:
https://citadel-ai.com/ja/news/2024/09/30/lens-for-llms-launch/https://aws.amazon.com/jp/builders-flash/202402/try-codewhisperer/

# NECの取り組み

# 紹介する取り組み

- リスク軽減
  - ハルシネーション対策機能
  - 偽情報・誤情報分析レポート
- リスクエスカレーション
  - AIガバナンス

2024年9月18日 プレスリリース

# リスク軽減：ハルシネーション対策機能

## 生成AI/LLMの出力に含まれるハルシネーションを検知し、確認作業を効率化

### LLMのみ

要約文

国内において、企業のIT活用と行政のデジタル化などDX投資の需要が増えており、NECはグループ会社のアビームコンサルティングと連携して対応力を強化しています。マイクロソフト社などの大手企業とのアライアンスも競争力を高めています。NECの技術を用いた自社実験やDigital ID、サイバーセキュリティサービスの共通基盤を通じて、NX提案を推進しています。

要約が正しいか確認するためには、
原文と要約文を**全て読むことが必要**

➔ ユーザーは資料すべての確認が必要

### LLM＋NEC技術を用いたハルシネーション対策

国内では、企業によるITを活用した経営変革や、行政のデジタル化など、いわゆるDX投資の需要が旺盛です。[…]においては、お客さまの課題分析、発見から[…]End to Endで対応できるパートナーとしての対応が重要です。そのために、国内でも有数のコンサルタントリソースを有するグループ会社のアビームコンサルティング(株)と連携しながら、さらなるアプローチ力の強化を図っています。またこれまで、マイクロソフト社、AWS社、オラクル社、SAP社やService Now社などとのグローバルベースでのアライアンスを締結し、競争力を強化してきました。(中略)2023年6月にはAvaloq社が米国の資産運用会社であるBlackRock社と戦略的パートナーシップを締結しました。ウェルスマネージャーやプライベートバンク向けに両社の強みを掛け合わせ、統合ソリューションを提供します。こうして市場競争力を強化し、幅広い顧客へとアプローチすることで、事業を拡大できると確信しています。

原文

原文との対応関係

国内において、企業のIT活用と行政のデジタル化などDX投資の需要が増えており、NECはグループ会社のアビームコンサルティングと連携して対応力を強化しています。マイクロソフト社などの大手企業とのアライアンスも競争力を高めています。NECの技術を用いた自社実験やDigital ID、サイバーセキュリティサービスの共通基盤を通じて、NX提案を推進しています。

要約文

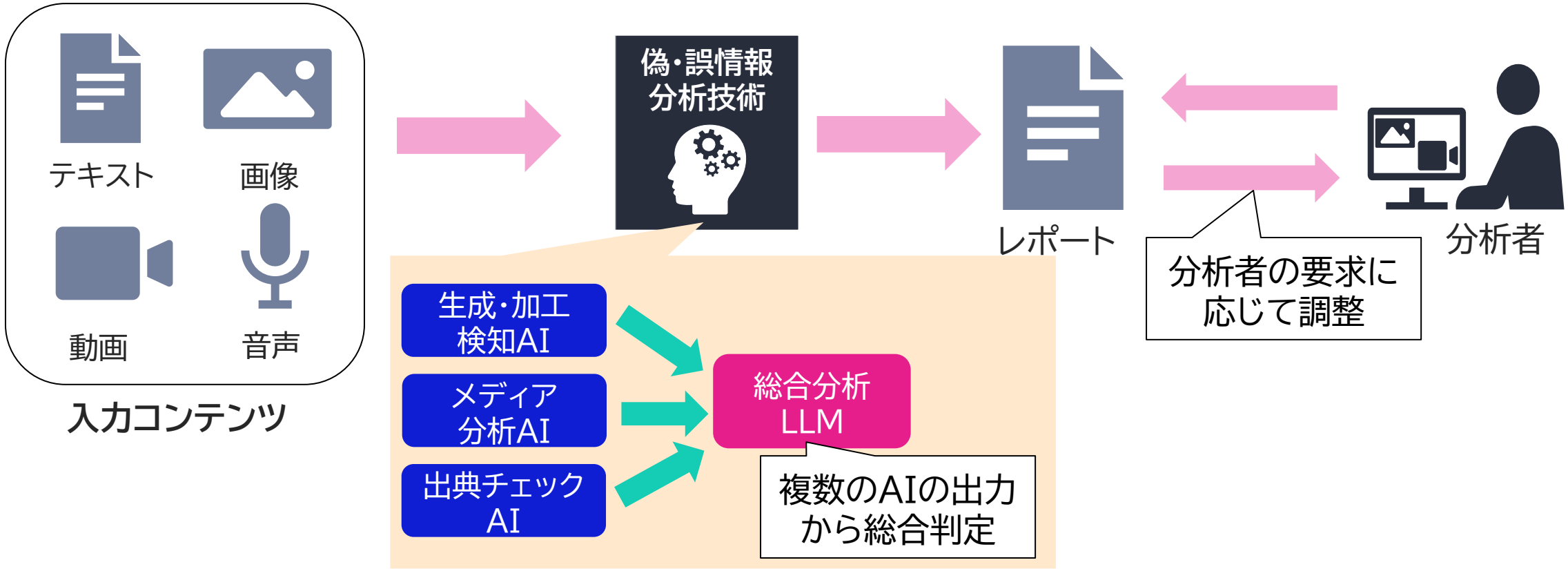
警告：ハルシネーションの可能性
（原文にない単語）

単語・文単位で、関連する箇所を絞り込まれるから
読む範囲が限定され、簡単に確認できる

➔ 確認ポイントが限定されるため、ユーザーの確認作業が効率化

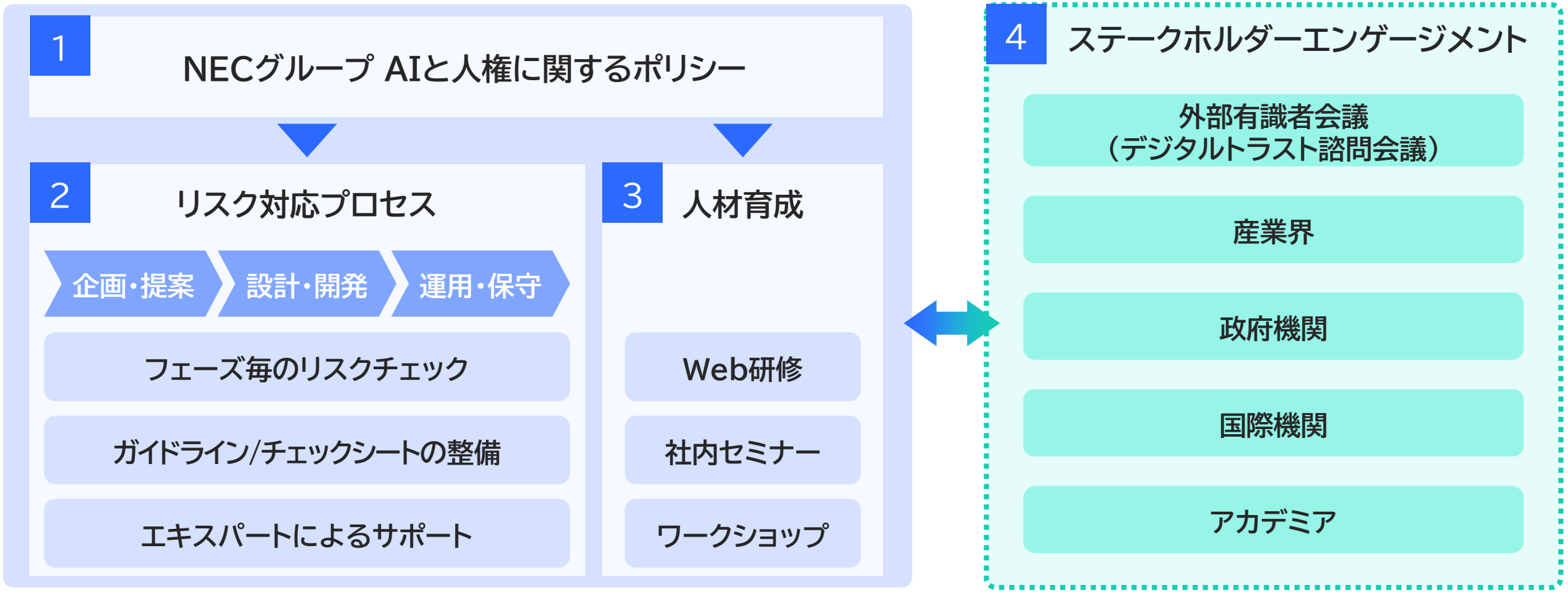
# リスク軽減：偽情報・誤情報分析レポート

ファクトチェック機関のレポートに近い形式で出力し、ファクトチェック業務を効率化



# リスクエスカレーション：AIガバナンス

## AIの利活用に関する事業活動が人権を尊重したものとなるよう、リスク対応、人材育成、ステークホルダーエンゲージメントを推進



# 1.NECグループ AIと人権に関するポリシー

AIの利活用において大切にすべき価値観を事業活動を推進するための指針として2019年4月に「NECグループ AIと人権に関するポリシー」を策定

1. 

公平性
Fairness

2. 

プライバシー
Privacy

3. 

透明性
Transparency

4. 

説明する責任
Responsibility to Explain

5. 

適正利用
Proper Utilization

6. 

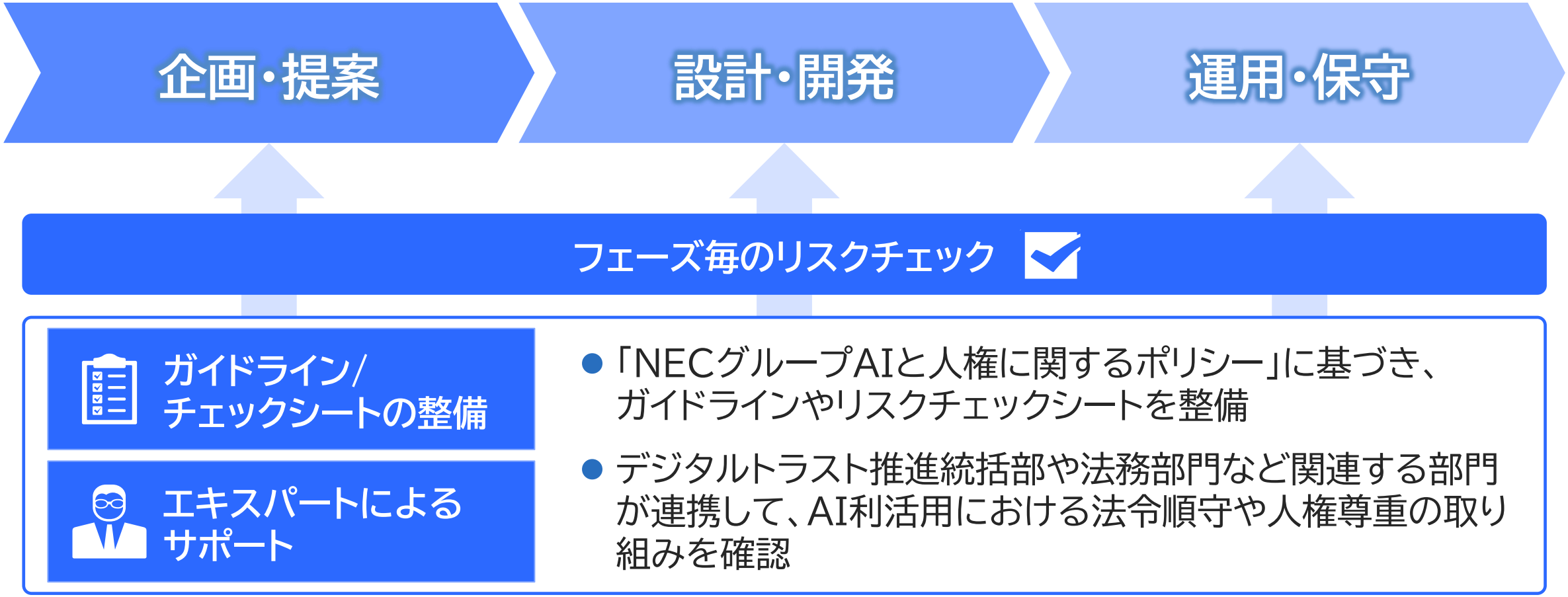
AIの発展と人材育成
AI and Talent Development

7. 

マルチステークホルダーとの対話
Dialogue with Multiple Stakeholders

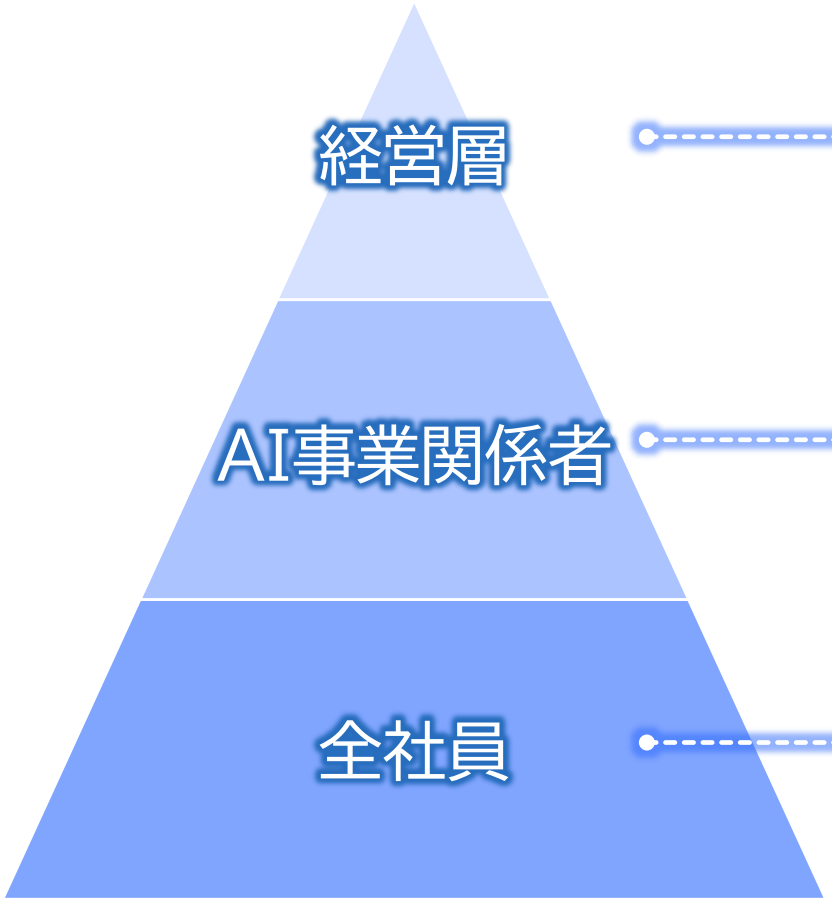
# 2.リスク対応プロセス

適正な用途での利用に向け、リスクチェックと対策を企画フェーズから実施



# 3.人材育成

全社ポリシーに基づき、事業活動において人権を尊重した適切な行動がとれるよう、
NECグループ12社 5.4万人へのeラーニング、有識者による規制動向解説、炎上事例解説などのセミナーを実施




| | 到達目標 | 実施内容 |
|---|---|---|
| 経営層 | 人権に配慮した事業判断 | 新技術に関連する規制動向をインプット |
| AI事業関係者 | お客様に対して法規制の最新動向を踏まえた提案・サポート | 規制動向のインプットや具体的なリスク対策事例の解説 |
| 全社員 | AI・生体認証技術が引き起こす人権課題を理解した業務遂行 | 人権課題の理解、全社ポリシー、案件相談ルール等の周知 |

# 4.ステークホルダーエンゲージメント

## 法規制や受容性など社会動向の変化に応じた活動を行うため、外部有識者会議を実施

### デジタルトラスト諮問会議（外部有識者会議）

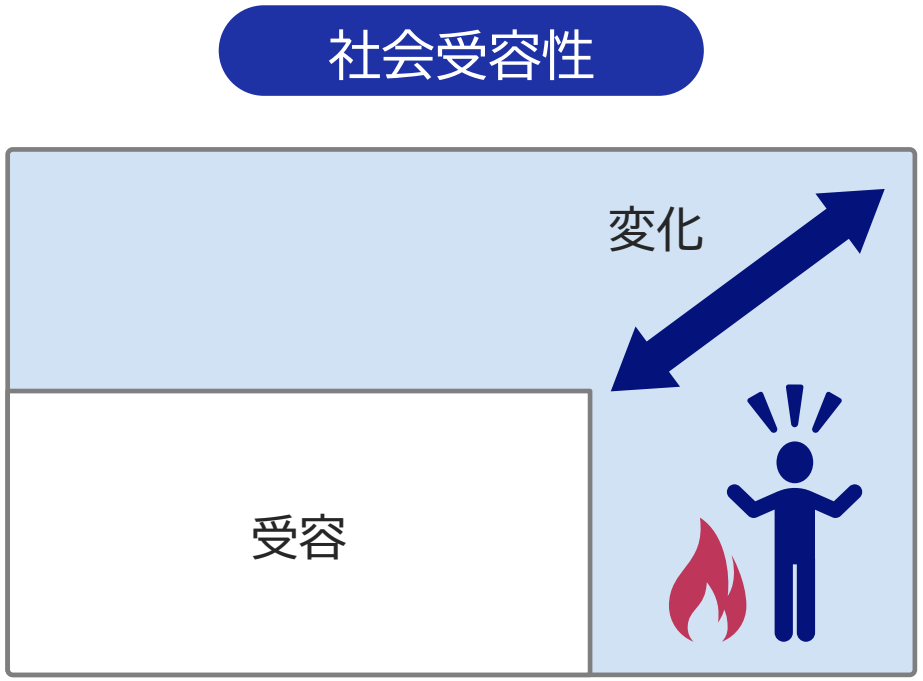
| | |
|---|---|
| 目的 | 法制度や人権・プライバシー、倫理に関し専門的な知見を有する外部有識者から<br>多様な意見を取り込み、AIの利活用において生じる新たな課題への対応を強化を図る |
| メンバー | 法律家、法学者、サステナビリティや人権などの分野のNPO関係者、消費者団体代表など、<br>法制度や人権・プライバシー、倫理に関する専門的な知見を有する次の4名の有識者で構成<br>● 議長： 板倉 陽一郎　（ひかり総合法律事務所）<br>● 議員： 永井 朝子　（BSR東京事務所 マネジング・ディレクター）<br>古谷 由紀子　（サステナビリティ消費者会議 代表）<br>山本 龍彦　（慶應義塾大学 法科大学院教授） |

# 終わりに

# 生成AIを活用したシステムに向けて大切にすべき観点

## 生成AIのような新たな技術の利活用において、機能の向上や法規制への準拠のみならず、法規制が未整備の「グレーゾーン」や「社会受容性」への配慮も重要

- LLMをはじめとする生成AIは技術の進展が早く、それに応じてリスクも変化
- システムを担当するプロジェクトマネージャだけでは対応が難しいケースもあり
- 新しく現れるリスクに対して管理できるようにガバナンス体制の構築が今後より重要に

\Orchestrating a brighter world