

生成AI/基盤モデルの利活用と AIセーフティの動向について

2025-01-30

AI Safety Institute

中里 克久



- AI Safety Institute (AISI)
- 独立行政法人 情報処理推進機構 (IPA)
※ 富士通株式会社 富士通研究所 から出向中

AIセーフティに関する評価観点ガイド、レッドチーミングガイドの
執筆・編集などに従事

- AI応用、ソフトウェア工学、ソフトウェア開発
- 情報処理学会 会誌編集委員

1. AI関連の動向とAIセーフティについて
2. AIの利活用について
3. AIの利活用の際の課題
4. AISIについて
5. AISIが発行したAIセーフティ関連ガイド

1. AI関連の動向とAIセーフティについて

◆ 「AIとは何か」

→ **確立された定義はいまだ存在しない**

- “「AI」とは、人間の思考プロセスと同じような形で動作するプログラム、あるいは人間が知的と感じる情報処理・技術といった広い概念で理解されている。”
(令和元年版 情報通信白書, 総務省)

初期のAI関連動向

1950年代頃	<ul style="list-style-type: none"> ・アラン・チューリングによるAI的な概念の提唱 ・ジョン・マッカーシーによる“Artificial Intelligence” (AI, 人工知能)の命名 ・第一次AIブーム
~~	
1980年代頃	<ul style="list-style-type: none"> ・エキスパートシステム ・第二次AIブーム
~~	

ロボット関連

紀元前	<ul style="list-style-type: none"> ・古代神話における自動人形(フィクション) - 青銅の巨人タロス(ギリシャ神話) など
~~	
17世紀 / 18世紀	<ul style="list-style-type: none"> ・ヨーロッパにおけるオートマタの流行 - チェスを指すオートマタ(イカサマ) ・日本におけるからくり人形の流行
~~	
1920年代頃	<ul style="list-style-type: none"> ・カレル・チャペック、「ロボット」命名
~~	
1950年代頃	<ul style="list-style-type: none"> ・アイザック・アシモフ、「ロボット三原則」提唱

フィクションや類似概念まで含めると非常に長い歴史あり

近年の主要なAI関連動向

2011	IBM Watsonがクイズ番組で人間に勝利 iPhoneへのSiriの搭載
2012	Deep Learningブームが起こる (高精度画像認識)
～	
2015	ペッパーくん、Amazon Echo(Alexa) 発売
～	
2017	将棋ソフトponanzaが佐藤天彦名人に勝利 チャットボット、スマートスピーカー ブーム
～	
2022	画像生成AIサービス 各種(Midjourney, Stable Diffusion等) 公開 対話型生成AIサービス ChatGPT/GPT-3.5 公開
2023	GPT-4公開
2024	デバイスへのAI搭載: PC(Copilot+ PC)、iPhone(Apple Intelligence)
2025	AIエージェント

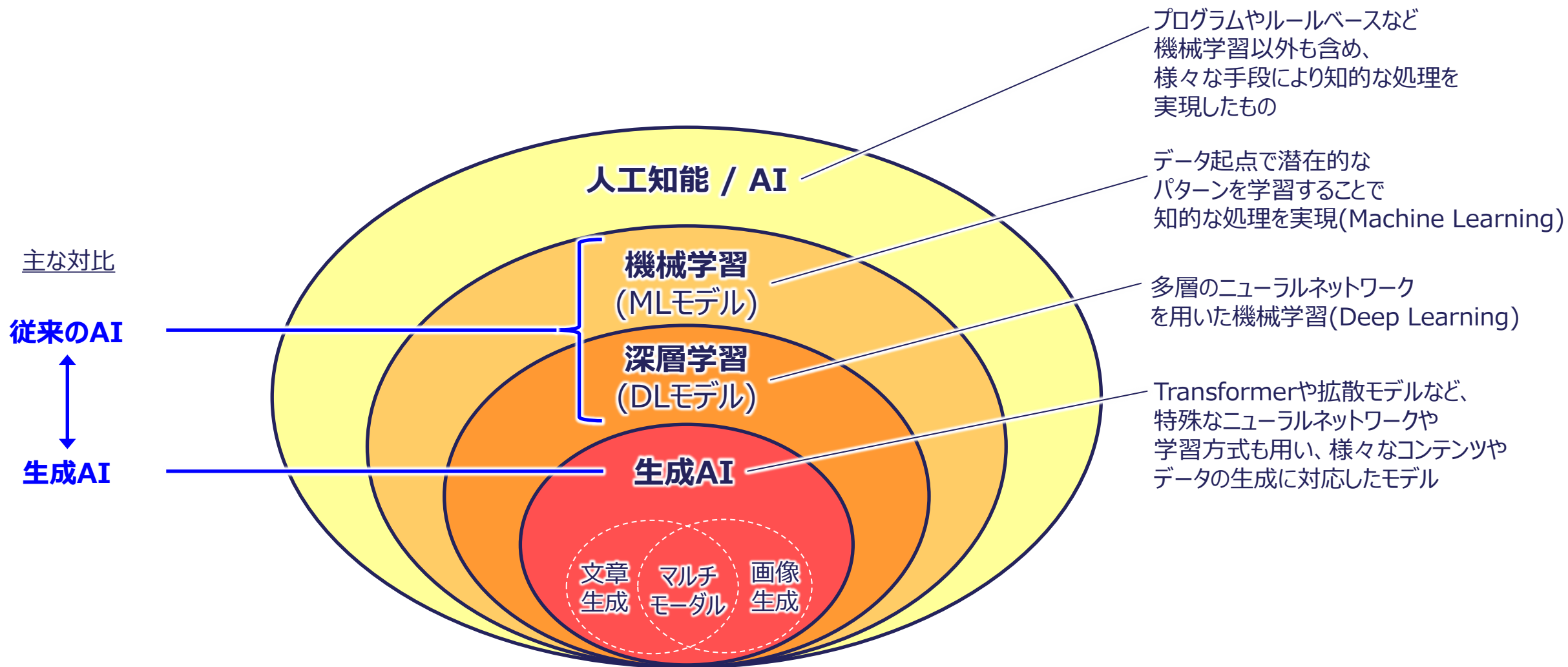
生成AI

第三次AIブーム

第四次

※ 生成AI以降を第四次とみなすかどうかは人による

AI関連の各種概念の関係



従来のAIと生成AIの違い-1

従来のAI

分類

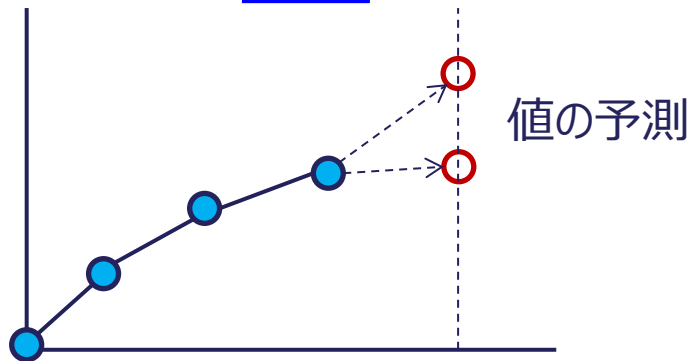


犬 or 猫 → **犬**



犬 or 猫 → **猫**

回帰



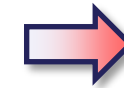
出力の形態がシンプル

生成AI

文章の生成

指示文

日本国の
紹介文を
作成して



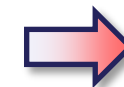
生成

日本は、古代の寺社や
近未来の都市が共存す
る国。四季折々の美し
い風景と、温泉、伝統
文化、美食が魅力です。

画像の生成

指示文

芝生の上の
サッカーボール

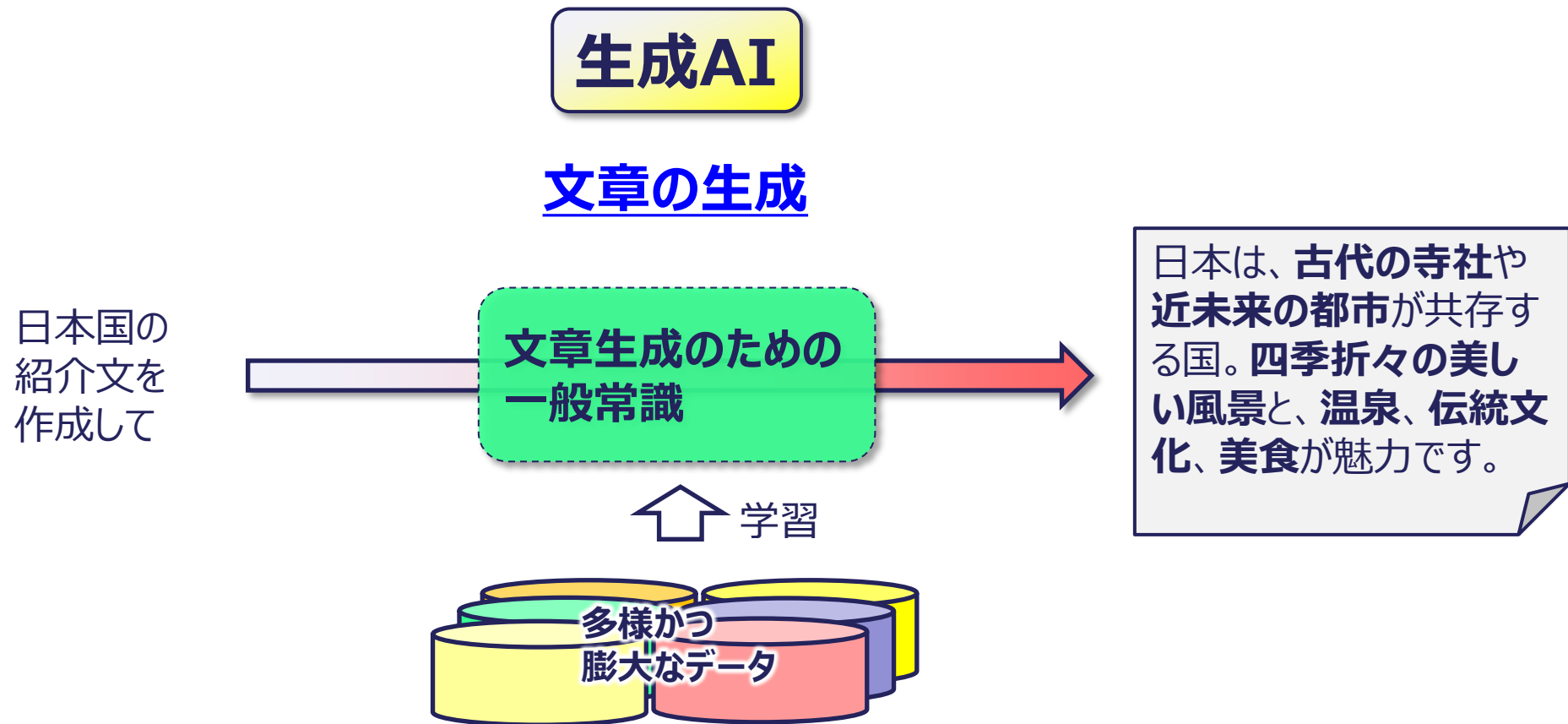


生成



複雑なコンテンツを生成

従来のAIと生成AIの違い-2



様々な文章を生成するために膨大なデータを学習しており、
(完全ではないが)一般常識的な知識を習得している

これ以降、基本的には

AI = 生成AI

という意味で用語を使用します

人間中心の考え方をもとに、AI 活用に伴う社会的リスクを低減させるための安全性・公平性、個人情報
の不適正な利用等を防止するためのプライバシー保護、AI システムの脆弱性等や外部からの攻撃等
のリスクに対応するためのセキュリティ確保、システムの検証可能性を確保し適切な情報提供を行うため
の透明性が保たれた状態。

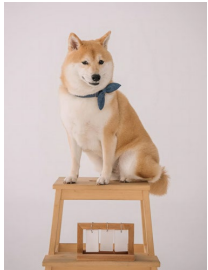
※ AISI発行の「AIセーフティに関する評価観点ガイド」における定義

- ◆ **従来の「安全性」よりやや広い概念**を指しているため、日本語としても「AIセーフティ」の用語を主に用いている
 - 工学的な安全性の概念も含みつつ、利用者がAIを信頼することができる状態(工学的なReliabilityではなく、Trustworthinessの語が良く使われる)も含め、「AIセーフティ」としている
 - 安全(JIS)： 許容できないリスクから免れている状態
 - 安全性(JIS)： リスク源となるアイテム及びリスクを被る実体からなるシステムが、安全を保持し得る性質または能力

AIの不正確な挙動による影響の比較

従来のAI

分類

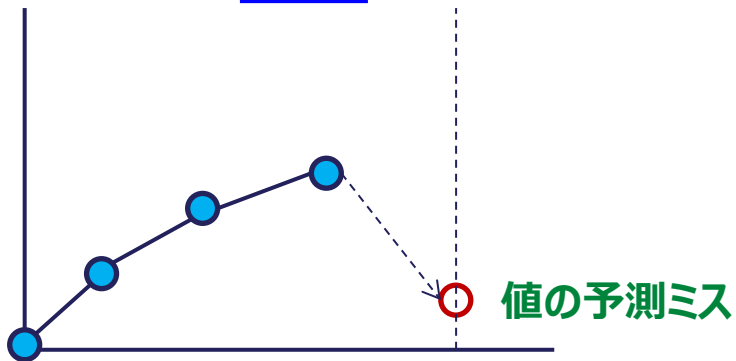


犬 or 猫 → **猫**



犬 or 猫 → **犬**
分類ミス

回帰



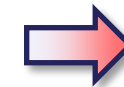
影響の内容が限定されるので、
事前予測や対策検討が比較的しやすい

生成AI

文章の生成

指示文

日本国の
紹介文を
作成して



生成

日本は、ユーラシア大陸
の内陸にある新興の多
民族国家です。

どのような間違い方で
結果が生成されるか、
事前予測が難しい

画像の生成

指示文

芝生の上の
サッカーボール



生成



AIの能力が増し、様々なことが
可能になった分、リスクも多様化
→ AIセーフティの重要性が増す

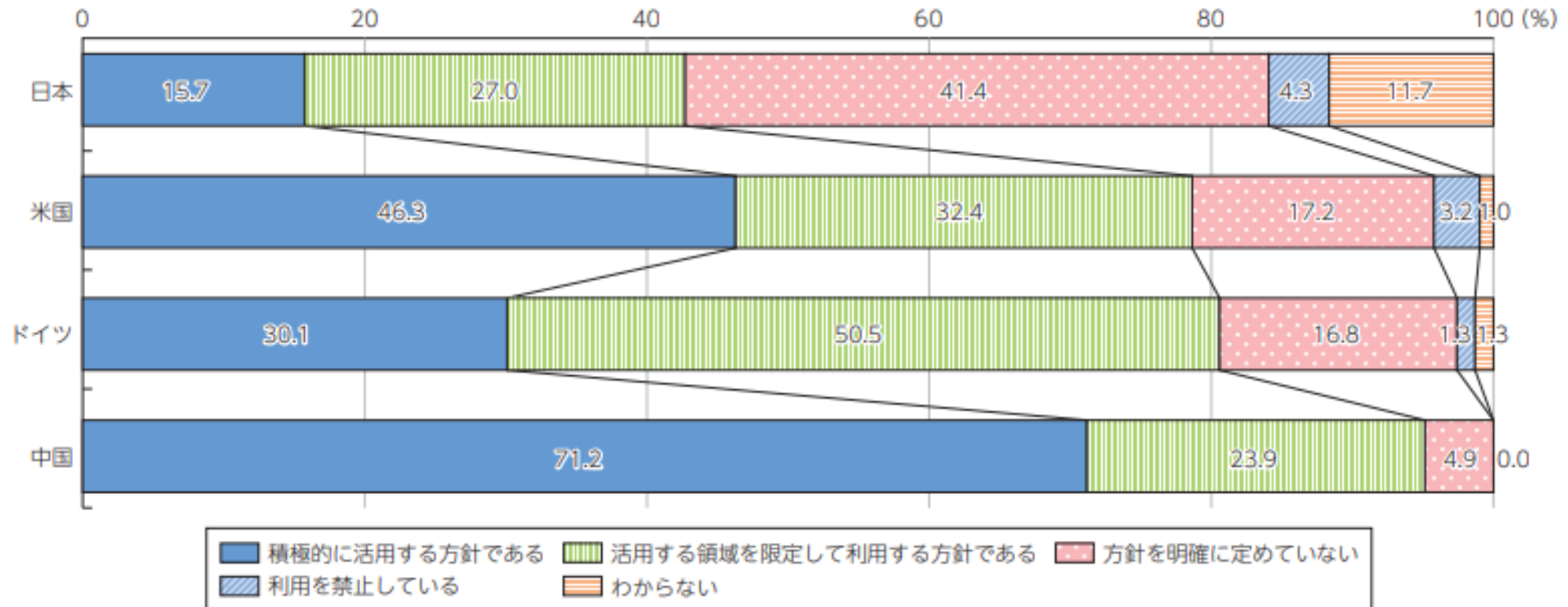
**AIに起因するリスクを明確化し、
AIセーフティを確保することが、
安全・安心なAIの利活用の促進のために必要**

2. AIの利活用について

企業におけるAI活用(利活用)状況

- ◆ 日本企業におけるAI活用やその方針策定は、他国と比べ遅れている

図表 I-5-1-4 生成AIの活用方針策定状況



(出典) 総務省 (2024) 「国内外における最新の情報通信技術の研究開発及びデジタル活用の動向に関する調査研究」

◆ AI利用による期待効果の認識

- 研究開発, 事務作業などでの対話型(テキストベース)AI利用
- マルチモーダルAI利用による様々な応用

◆ AI利用に伴う不安要素の解消/緩和

- AIセーフティ: AISIに特に期待される取り組み

- ① 文章生成
- ② 翻訳/要約
- ③ データ分析/グラフ作成
- ④ コード生成
- ⑤ ブレスト支援
など

① 方針を指示しての文章生成

- ◆ 指示文(プロンプト)により、**文章生成の方針を指示**して生成AIに文章を作成させることで、文章作成の**作業負荷を軽減**する
 - 指示文の記述のしかたにもノウハウがあり、有効なパターンがある程度確立されてきている

[AIへの入力(指示)]

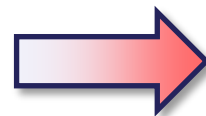
#命令
あなたは プロの旅行プランナー です。
以下の条件と入力文をもとに、最高の旅行推薦文 を出力してください。

#条件

- ・文字数は200文字以内
- ・丁寧なことばを使う
- ・不適切な用語は使わない

#入力文

- ・京都旅行について
- ・千年以上の歴史を誇る古都
- ・近年は観光で人気
- ・歴史を体験できる京町屋への宿泊
- ・奥深い京料理
- ・その他、一般的に京都の魅力として語られている事項も盛り込む



[AIが出力した回答]

京都は千年以上の歴史を持つ古都で、その魅力は今も色褪せません。歴史ある街並みが広がる京町屋に宿泊し、奥深い京料理を堪能することで、京都の伝統と文化を存分に味わえます。嵐山や清水寺といった名所を巡り、自然と歴史が融合した風景に心癒されることでしょう。風情ある街並みと温かいおもてなしが、忘れられない旅の思い出を作ります。

多少推敲が必要な部分もあるが、全体的には指示を忠実に守りつつ、一般的な情報も付け加えて適切な文章を作成している
→ **生成された文章をベースにして人が改善していくアプローチ**により、作業工数を削減できる可能性

② 翻訳と要約

◆ 英文などの翻訳や、文章要約の能力も持つ

[AIへの入力(指示)]

#命令

以下の条件に従い、入力文を日本語で要約してください。

#条件

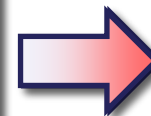
・文字数は200文字以内

#入力文

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we can not dedicate—we can not consecrate—we can not hallow—this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us—that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion—that we here highly resolve that these dead shall not have died in vain—that this nation, under God, shall have a new birth of freedom—and that government of the people, by the people, for the people, shall not perish from the earth.

リンカーンのゲティスバーグ演説
「人民の人民による人民のための政治」
翻訳と要約



[AIが出力した回答]

87年前、私たちの祖先はこの大陸に自由を理念とし、すべての人が平等に創られたという信念に基づく新しい国家を誕生させました。今、我々はその国家が存続できるかを試す大きな内戦の中にあります。ここで命を捧げた人々を称えるため、この地を捧げることは正しいことですが、彼らの行為こそがこの地を神聖にしました。彼らの犠牲を無駄にせず、自由の再生を目指し、**国民のための政府**を守ることが私たちの責務です。

いないかもしれませんが、彼らがここで何をしたかを決して忘れることはありません。むしろ、生き残った私たちが、この場で彼らがこれまで崇高に前進させた未完の仕事に捧げられるべきです。私たちは、この名誉ある死者たちから、彼らが最後の忠誠を捧げたその大義への献身をさらに強めることを誓い、この死者たちが無駄死にしなかったことをここで強く決意し、神のもと、この国家が新たな自由の誕生を迎え、**人々の人民による人民のための政府**が地上から消え去らないことを誓うべきです。

③ Excelデータからのグラフ作成-1

- ◆ やや複雑なExcelデータから、指定した内容でグラフを作成する
- ◆ 例とするデータ: 国勢調査の世帯数時系列データ
 - <https://www.e-stat.go.jp/stat-search/file-download?statInfId=000012671095&fileKind=0>

[AIへの入力(指示)]

以下のデータから、京都府の世帯総数の時系列変化のグラフを作成し、加えて、変化の傾向について文章で説明してください。

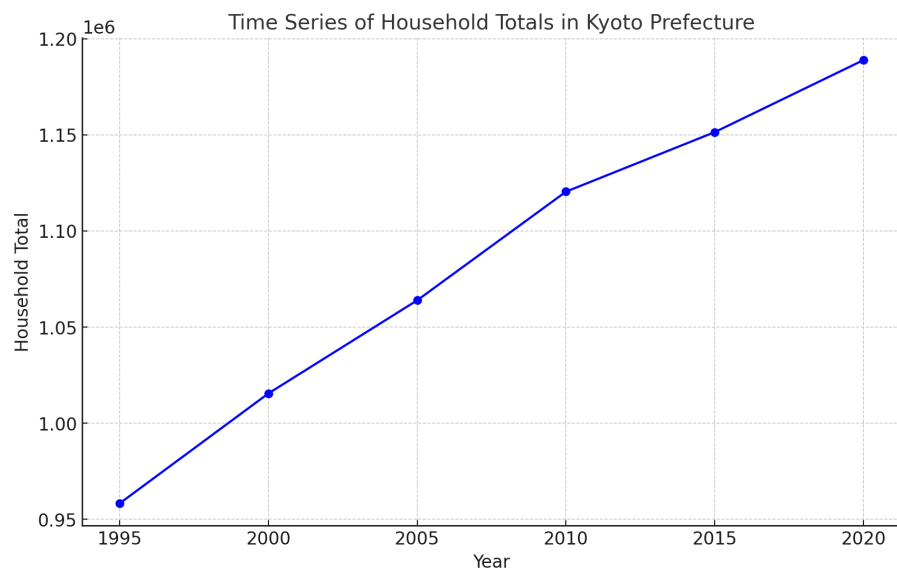
(※ 世帯数データのExcelファイルを添付)

飛び飛びになっている京都府の世帯総数のデータを繋ぎ合わせて、時系列変化のグラフを作成する必要がある

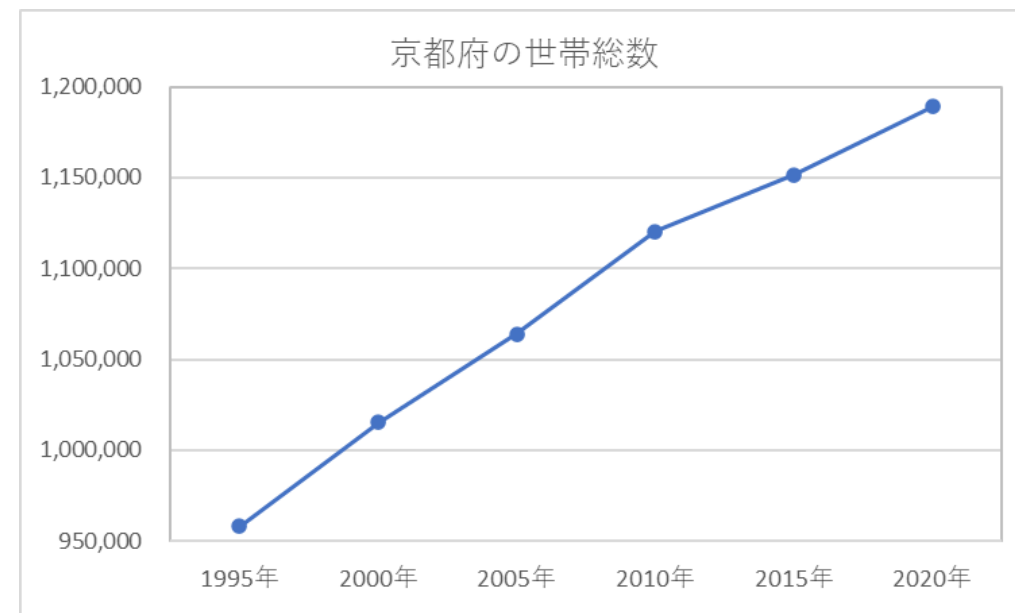
	年次(西暦)	年次(和暦)	地域	総数	親族のみの世帯	核家族世帯	夫婦世帯
35	1995年	平成7年	25000_滋賀県	334,271	315,880	224,101	
36	1995年	平成7年	26000_京都府	958,252	675,937	552,349	
37	1995年	平成7年	27000_大阪府	3,270,397	2,355,565	2,062,455	
38	1995年	平成7年	28000_兵庫県	1,867,031	1,441,382	1,186,882	
39	1995年	平成7年	29000_奈良県	454,984	372,510	290,614	
40	1995年	平成7年	30000_和歌山県	385,384	290,632	224,576	
41	1995年	平成7年	31000_鳥取県	188,866	151,033	96,998	
42	1995年	平成7年	32000_島根県	244,996	193,014	124,826	
43	1995年	平成7年	33000_岡山県	656,761	502,346	373,156	
44	1995年	平成7年	34000_広島県	1,046,122	766,683	625,196	
45	1995年	平成7年	35000_山口県	562,792	422,855	337,303	
46	1995年	平成7年	36000_徳島県	273,839	213,025	150,970	
47	1995年	平成7年	37000_香川県	345,422	268,669	198,384	
48	1995年	平成7年	38000_愛媛県	540,670	408,560	325,397	
49	1995年	平成7年	39000_高知県	302,868	220,269	176,858	
50	1995年	平成7年	40000_福岡県	1,774,183	1,275,296	1,043,225	
51	1995年	平成7年	41000_佐賀県	287,230	214,118	144,334	
52	1995年	平成7年	42000_長崎県	528,156	402,346	317,713	
53	1995年	平成7年	43000_熊本県	615,744	469,753	343,667	
54	1995年	平成7年	44000_大分県	433,897	326,095	250,654	
55	1995年	平成7年	45000_宮崎県	420,260	318,534	262,983	
56	1995年	平成7年	46000_鹿児島県	687,021	494,311	437,113	
57	1995年	平成7年	47000_沖縄県	403,060	312,373	263,547	
58	2000年	平成12年	00000_全国	46,782,383	33,594,905	27,272,891	
59	2000年	平成12年	01000_北海道	2,277,968	1,582,591	1,376,409	
62	2000年	平成12年	24000_三重県	639,362	434,130	376,376	
63	2000年	平成12年	25000_滋賀県	453,370	339,771	252,096	
84	2000年	平成12年	26000_京都府	1,015,468	696,151	583,978	
85	2000年	平成12年	27000_大阪府	3,454,840	2,403,270	2,133,972	
86	2000年	平成12年	28000_兵庫県	2,035,097	1,517,176	1,284,018	
87	2000年	平成12年	29000_奈良県	481,051	388,187	311,011	

③ Excelデータからのグラフ作成-2

AIによる出力結果



参考: 手動でのグラフ作成結果



上記のグラフは、1995年から2015年にかけての京都府における世帯総数の時系列変化を示しています。

主な傾向:

年を追うごとに世帯数が着実に増加しています。

1995年から2015年の間に、世帯総数はおよそ20万件増加しました。

増加率は後年にかけてやや鈍化しており、特に2010年から2015年の増加は、それ以前の期間と比べて穏やかです。

この傾向は、京都府内での都市化や、世帯人数の減少、単独世帯の増加といった人口動態の変化を反映していると考えられます。

④ コード生成

[AIへの入力(指示)]

下記の仕様のCSV入力ファイルを読み込み、下記の仕様のJSON出力ファイルを書き出すPythonプログラムを生成してください。
その際、処理内容は下記の変換例を参考にしてください。

CSV入力ファイルの仕様

- UTF-8のテキスト形式のファイル
- カンマ区切り
- 1行目はヘッダー行で、各列の属性名を示す
- 2行目以降はデータ行

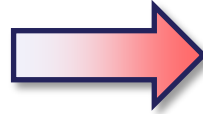
JSON出力ファイルの仕様

- UTF-8のテキスト形式のファイル
- "data"という要素名でCSV入力ファイルの内容に対応する配列を含める
- CSVのデータ行1行が配列の中の1要素に対応する
- 配列の中の1要素内では、属性名とデータの値の組でCSVのデータ行の内容を表現する。数値以外はダブルクォーテーションで囲み、数値は囲まない

変換例

- 入力のCSV
Number,Name,Score
1,Mike,75
2,David,83

- 出力のJSON
{
 "data": [
 {
 "Number": 1,
 "Name": "Mike",
 "Score": 75
 },
 {
 "Number": 2,
 "Name": "David",
 "Score": 83
 }
]
}



[AIが出力したコード]

```
import csv
import json

def csv_to_json(csv_input_file, json_output_file):
    try:
        # CSVファイルを読み込む
        with open(csv_input_file, mode='r', encoding='utf-8') as csv_file:
            csv_reader = csv.DictReader(csv_file)
            data = []

            # CSVの各行を辞書形式でリストに追加
            for row in csv_reader:
                converted_row = {}
                for key, value in row.items():
                    # 数値そのまま、文字列はクォートを保持
                    if value.isdigit():
                        converted_row[key] = int(value)
                    else:
                        try:
                            # 小数点を含む数値をfloatに変換
                            converted_row[key] = float(value)
                        except ValueError:
                            converted_row[key] = value
                data.append(converted_row)

            # JSON形式で書き出す
            with open(json_output_file, mode='w', encoding='utf-8') as json_file:
                json.dump({"data": data}, json_file, ensure_ascii=False, indent=4)

            print(f"JSONファイル '{json_output_file}' に変換されました。")
    except Exception as e:
        print(f"エラーが発生しました: {e}")

# 入力CSVファイルと出力JSONファイルのパス
csv_input_file = 'input.csv' # 入力CSVファイル名
json_output_file = 'output.json' # 出力JSONファイル名

# 関数を実行
csv_to_json(csv_input_file, json_output_file)
```

CSV→JSON 変換

変換例も含め、やや手厚く指示したが、正しく機能するコードを生成できた

⑤ ブレスト的なアイデア抽出-1

- ◆ **生成AIが学習している情報**を活用し、ブレスト的な**アイデア抽出**を行わせる
 - 一般論に近い内容であるほど、生成AIが妥当な案を出力する可能性が高まる

[AIへの入力(指示)]

#命令

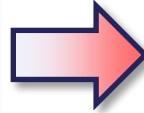
あなたは 地方自治体の政策立案者 です。
以下の条件と入力文をもとに、過疎地域の活性化案 を出力してください。

#条件

- 内容の異なる5種類の案を作成
- 各案の文字数は200文字以内
- 丁寧なことばを使う
- 不適切な用語は使わない

#入力文

- 過去10年にわたり緩やかな人口減少が続いている市
- 10年前の人口は約7万人。現在の人口は約5万人
- 少子高齢化が進んでいる
- 市外への若年層の流出も進んでおり、特に女性の流出率が高い
- 市の財政状況は健全で、様々な施策を実施する余地がある



[AIが出力した回答]

1. 若者向け起業支援とテレワーク環境整備

若年層の市外流出を防ぐため、起業支援やテレワーク環境の整備を行います。特に、女性が働きやすいワークスペースを提供し、子育て支援と連携させることで、地域に定住しやすい環境を作ります。

2. 地域資源を活用した観光振興

地域の自然や歴史を活かした観光資源を開発し、観光業を強化します。特に、若年層が魅力を感じるアクティビティや、地域の特産品をPRするイベントを企画し、地域外からの来訪者を増やします。

3. 定住促進と移住者支援制度の拡充

他地域からの移住を促進するため、移住者への住宅補助や仕事の斡旋、子育て支援を強化します。特に若年層や子育て世代が魅力を感じるよう、ライフスタイルの提案やコミュニティ活動への参加支援を行います。

4. 地域内企業と教育機関の連携強化

地元企業と教育機関が連携し、地元でのキャリアパスを提供するプログラムを創設します。若者が地元で働く魅力を感じ、専門性を高めることで、流出を防ぎつつ地域の産業を発展させます。

5. 子育て・福祉支援の強化

少子化対策として、子育て支援や高齢者福祉サービスの充実を図ります。特に、女性が安心して子育てできる環境を整え、共働き家庭への支援を強化することで、地域での定住を促進します。

まずまず妥当な内容の案が出力される
→ これらの案をそのまま採用するべきではないが、
これらを議論の出発点としたり、
議論が行き詰った状況で視点を変える用途などに利用できる

⑤ ブレスト的なアイデア抽出-2

- ◆ 単発のアイデア抽出だけでなく、アイデアの深掘りをしていく使い方も可能

[先ほど提示された案]

4. 地域内企業と教育機関の連携強化

地元企業と教育機関が連携し、地元でのキャリアパスを提供するプログラムを創設します。若者が地元で働く魅力を感じ、専門性を高めることで、流出を防ぎつつ地域の産業を発展させます。

+

[ブレスト手法の一例]

オズボーンのチェックリスト

1. 転用する
2. 応用する
3. 変更する
4. 拡大する
5. 縮小する
6. 代用する
7. 再構成する
8. 逆転する
9. 組み合わせる



[AIへの入力(指示)]

オズボーンのチェックリストに基づいて、次の地域活性化案を基に新しいアイデアをそれぞれ出してください。

地域内企業と教育機関の連携強化

地元企業と教育機関が連携し、地元でのキャリアパスを提供するプログラムを創設します。若者が地元で働く魅力を感じ、専門性を高めることで、流出を防ぎつつ地域の産業を発展させます。



[AIが出力した回答]

1. **転用できないか（他の用途に利用する）** 地元企業と教育機関の連携を、地域住民全体に拡大し、生涯学習や職業訓練の場として活用します。高齢者にも再教育やスキル向上の機会を提供し、地域全体の労働力を強化します。
2. **応用できないか（既存のアイデアを他に適用する）** 教育と企業連携の枠組みを、高校生や中学生にも拡張し、インターンシップや実習機会を提供します。これにより、若年層が早い段階から地元企業に関心を持ち、定住につながる教育プログラムを提供します。
3. **拡大できないか** 地元企業と教育機関の連携に、他の自治体の企業や大学との交流プログラムを追加し、広域的なネットワークを構築します。他地域との協力により、幅広い学びや就業機会を提供し、地域の魅力を向上させます。
4. **縮小できないか** 特定の産業や分野に絞った専門プログラムを設立し、地域の産業特化型の人材育成を行います。例えば、農業や観光業、IT分野に特化したコースを設け、地域産業の強みをさらに引き出します。
5. **代用できないか** 地元企業と教育機関の連携に、リモート教育やオンライン企業研修を導入し、外部講師や企業の知見を取り入れます。物理的な場所に依存せず、多様な学びやキャリアパスを提供することで、柔軟な働き方を促進します。
6. **再整理できないか** 地元企業との連携プログラムを、産業ごとに細分化して設け、企業が特定の業界や分野ごとにより深く関わる仕組みにします。これにより、産業ごとに必要な人材育成を強化し、効率的に地域経済を活性化します。

これらのアイデアは、オズボーンのチェックリストを活用した具体的な提案です。各視点を取り入れることで、既存の枠組みをさらに発展させ、地域の活性化に貢献できます。

◆ AIのマルチモーダル対応

- 現在の主要な生成AI = **LLM** (Large Language Model, 大規模言語モデル)
 - 大量のテキストデータを学習し、自然言語による質問・応答が可能なAIモデル
- 画像, 音声, 動画 など、様々な形式のデータを含めて学習し、様々な用途に応用可能な**マルチモーダル基盤モデル** (Multimodal Foundation Model)
- 各社の主要なAIモデルは、まだ完全ではないがマルチモーダル対応を進めている
 - OpenAI GPT-4o, Google Gemini 2.0, Anthropic Claude 3.5 など

※ 参考: マルチモーダル対応AIを指す様々な用語 (ほぼ同じ意味)

- マルチモーダル基盤モデル
 - 単に「基盤モデル」としていてもマルチモーダル対応を含む場合もある
- マルチモーダルLLM
- LMM (Large Multimodal Model)
- VLM (Vision-Language Model)

詳細な指示文からの画像生成 ≡ テキストが表す状況を認識できている

[マルチモーダル基盤モデルへの入力]

以下の画像を生成してください。

工場の作業ラインで3人の作業員が1台の自動車の組み立ての作業をしています。

1人は自動車の下に潜り込んで下部の組み立てをしており、後の2人は自動車の左右に分かれてドアの取り付けをしています。

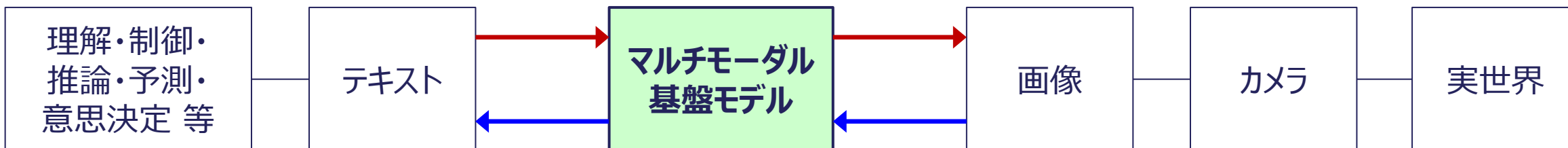
自動車は**セダン型でボディカラーは赤**です。

作業員は**青の作業服**を着ており、同じ色の帽子をかぶっています。

[マルチモーダル基盤モデルが出力した画像]

生成画像は投影のみ

多様な自然言語の入力文から、それに対応した画像を生成するには、言語が表す状況を認識して概念化し、概念を画像化して出力する必要
→ **画像が表す状況を認識して概念化し、概念を言語化して出力する逆変換も可能**

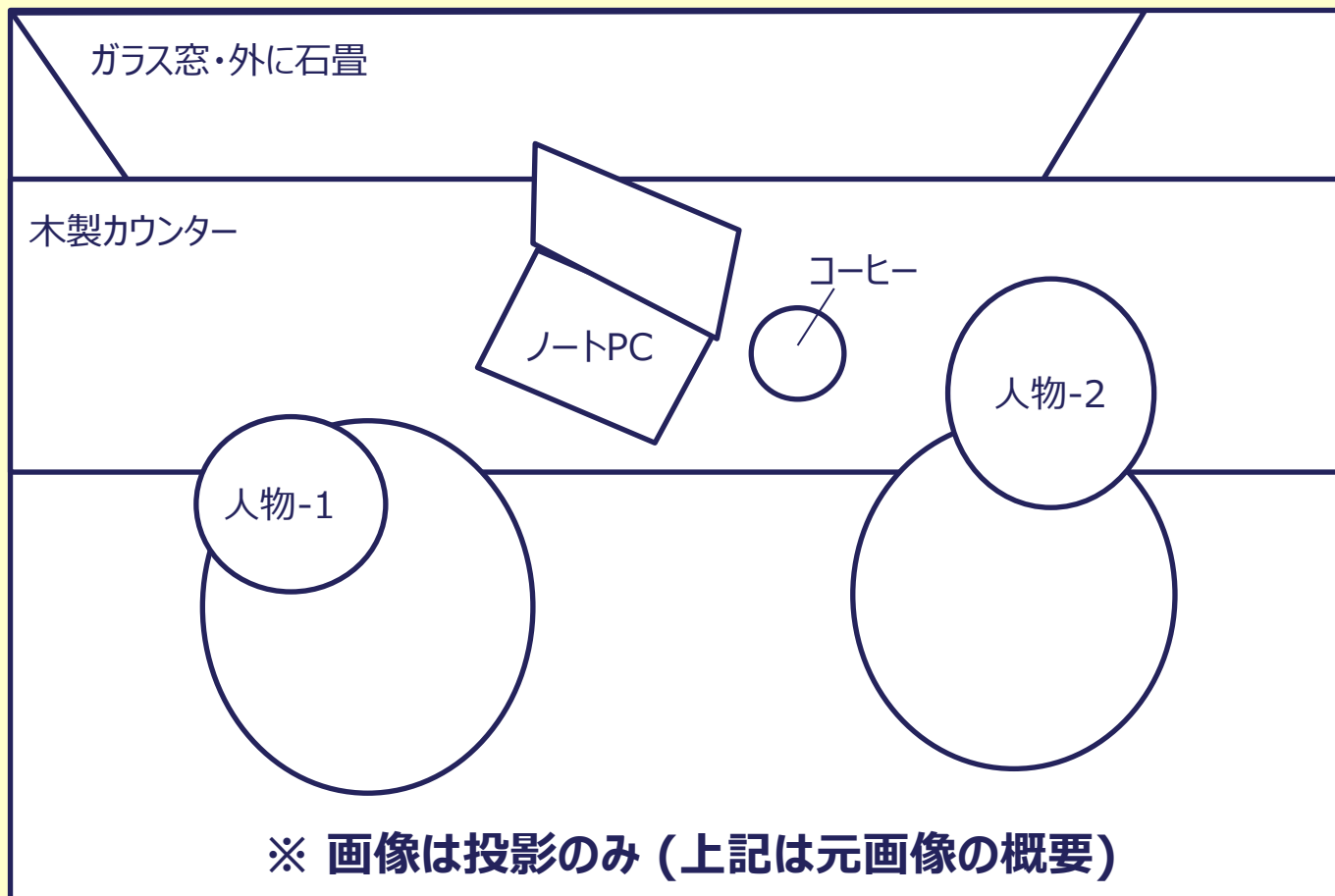


※ VQA (Visual Question Answering)

画像 → テキスト (画像が表す状況の認識)

[マルチモーダル基盤モデルへの入力]

この画像に写っている状況を説明してください。



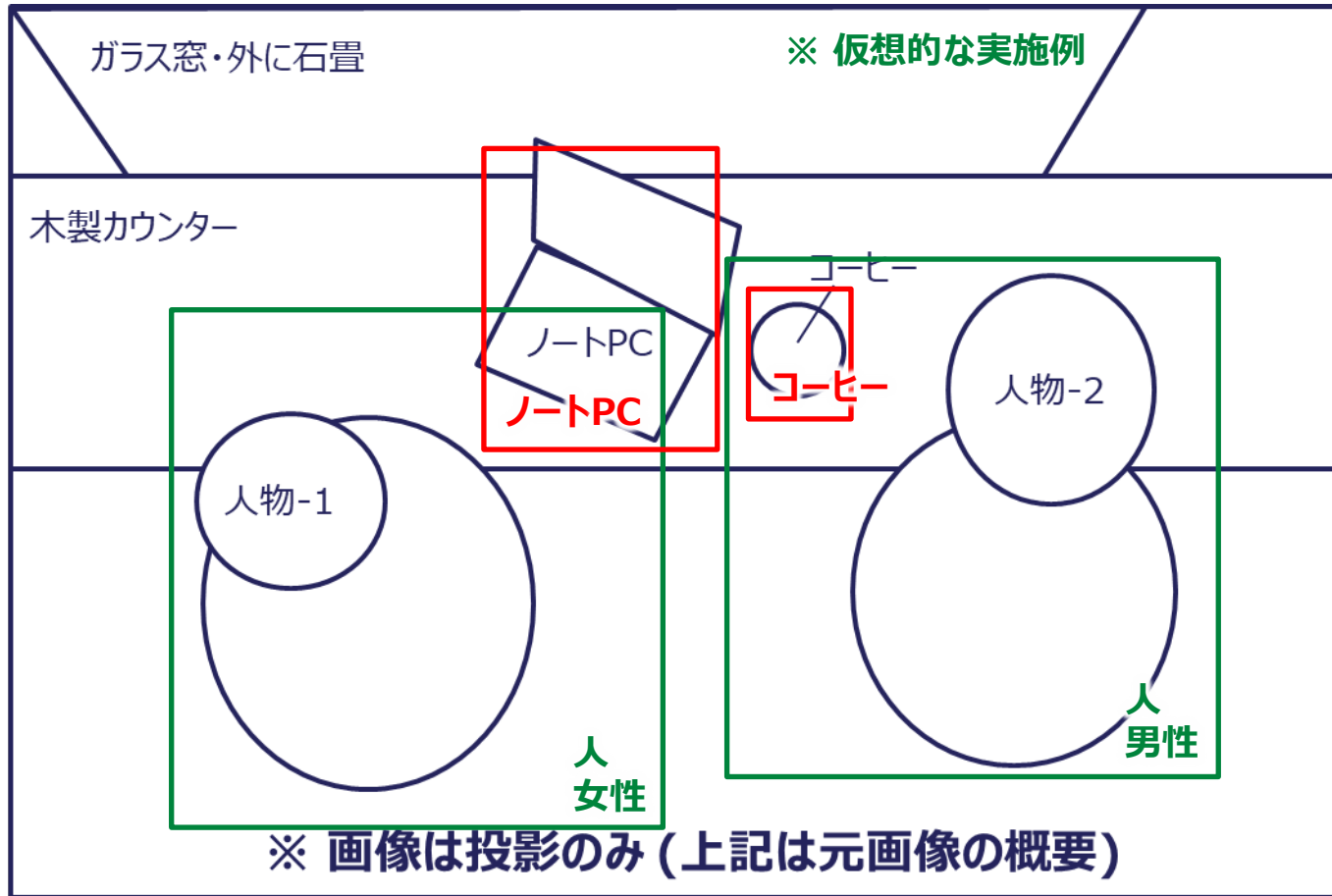
[マルチモーダル基盤モデルが出力した回答]

この画像は、**カフェのような場所**で**2人の人物**が一緒に作業をしている様子を撮影したものです。**一人はノートパソコン**を使用して作業をしており、もう一人は**コーヒー**を飲みながら**会話に参加**しているようです。

木製のカウンターに向かって座っており、**リラックスした雰囲気**が感じられます。自然光が窓から差し込み、**窓の外には石畳**の道が見えます。

物体認識だけでなく、
その場の状況を認識できている

仮に、従来の物体認識AIを用いて同様の状況認識を行う場合



従来の物体認識用AIを用いて
ここまで詳細に状況認識するには、
↓ AI以外に相当な追加開発が必要

[マルチモーダル基盤モデルが出力した回答]

この画像は、**カフェのような場所**で**2人の人物**が一緒に作業をしている様子を撮影したものです。一人は**ノートパソコン**を使用して作業をしており、もう一人は**コーヒー**を飲みながら**会話に参加**しているようです。**木製のカウンター**に向かって座っており、**リラックスした雰囲気**が感じられます。自然光が窓から差し込み、**窓の外には石畳**の道が見えます。

カウンター+コーヒー=カフェ?

何から判断?

2人の距離と顔の向き?

石壁との区別?

どのように判断するかは自明ではない

**マルチモーダル基盤モデルによる一括認識
は大きな進歩**

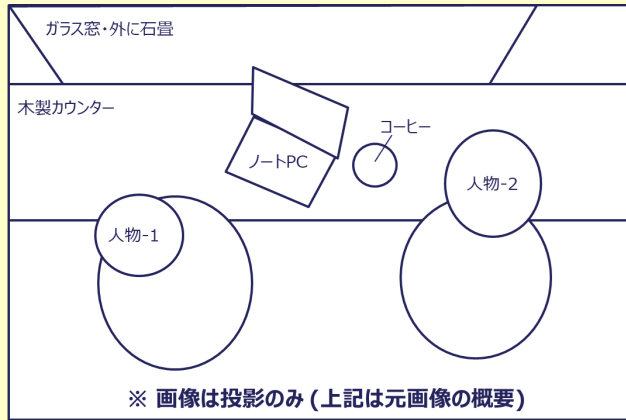
[従来AIで状況認識する場合の懸念点]

- ・ 俯角が大きく、人の骨格推定や表情認識が困難
- ・ 物体間の関係から状況を推定するルールを作成できたとしても、カメラの向きなどの条件が変わってしまうとルールの大幅修正が必要になる

状況の認識 → 更なる推論 (例: トラブル想定)

[マルチモーダル基盤モデルへの入力]

この画像に写っている状況を説明してください。



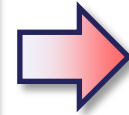
[マルチモーダル基盤モデルが出力した回答]

この画像は、カフェのような場所で2人の人物と一緒に作業をしている様子を撮影したものです。一人はノートパソコンを使用して作業をしており、もう一人はコーヒーを飲みながら会話に参加しているようです。

木製のカウンターに向かって座っており、リラックスした雰囲気を感じられます。自然光が窓から差し込み、窓の外には石畳の道が見えます。

[マルチモーダル基盤モデルへの入力(追加質問)]

右の人物がコーヒーを飲む時に発生し得るトラブルとして何が考えられますか？
トラブルの対象と事象を明示して、5個以上列挙してください。



[マルチモーダル基盤モデルが出力した回答]

以下は、右の人物がコーヒーを飲む際に発生し得るトラブルを列挙したものです。対象と事象を明確にしています。

1. 対象：パソコン（左の人物のもの）
事象：コーヒーをこぼして液体がキーボードや内部に入り、故障する可能性がある。
2. 対象：衣服（右の人物の服）
事象：コーヒーをこぼして衣服にシミができ、洗濯が必要になる。
3. 対象：テーブル上の物品（ノートや資料）
事象：コーヒーがノートや紙資料にかかり、文字が滲む、または資料が破損する可能性がある。
4. 対象：他人（左の人物）
事象：コーヒーを誤って相手にこぼし、服を汚したり熱傷を負わせるリスクがある。
5. 対象：椅子や床
事象：コーヒーを落として床や椅子を汚し、滑りやすくなることで転倒の危険性が生じる。
6. 対象：自分自身（右の人物の口や手）
事象：コーヒーが熱すぎて口や舌を火傷する可能性がある。
7. 対象：カフェの機材（電気コンセントや充電器）
事象：コーヒーが電気機器にかかり、ショートや火災のリスクを引き起こす可能性がある。

これらのリスクを回避するためには、飲み物を置く位置に注意し、安定した状態で飲むことが重要です。また、周囲に注意を払うことも必要です。

単なる一般論でなく、
認識した状況に応じた
トラブル想定を
列挙できている

3. AIの利活用の際の課題

◆ 企業内の独自データへの対応

- 一般的なAI(OpenAI, Google, Anthropic等が商業サービス化しているAI)は、各企業内の独自のデータは学習していない
 - 専門性の高い内容や、特殊なデータ形式などについて、うまく処理できない可能性がある
 - 特に、不正確な情報をもっともらしく出力する現象をハルシネーション(幻覚)と呼ぶ

ハルシネーション(幻覚)の例

[入力した質問]

現在の日本の内閣のメンバーについて教えて。

AIが学習した時点の情報なので、情報が古い

[やや旧式のAIが出力した回答]

現在の日本の内閣のメンバーは以下の通りです：

1. 安倍晋三 (あべしんぞう) - 内閣総理大臣 (Prime Minister)
2. 麻生太郎 (あそう たろう) - 副総理兼財務大臣 (Deputy Prime Minister and Minister of Finance)
3. 菅義偉 (すが よしひで) - 内閣官房長官 (Chief Cabinet Secretary)
4. **野田聖子 (のだ せいこ) - 外務大臣** (Minister of Foreign Affairs)
5. 河野太郎 (こうの たろう) - 防衛大臣 (Minister of Defense)
6. 小泉進次郎 (こいずみ しんじろう) - 環境大臣 (Minister of the Environment)

...

過去の実績として概ね正しい情報が出力されているが、
正しくない情報が混入している

※ 最新の対話型生成AIでは、Web検索と連携した回答を行うようになっており、
上記ほど簡単に事実と異なる回答が出力されることはなくなっている

◆ 企業内の独自データへの対応

- 一般的なAI(OpenAI, Google, Anthropic等が商業サービス化しているAI)は、各企業内の独自のデータは学習していない
 - 専門性の高い内容や、特殊なデータ形式などをうまく処理できない
 - 不正確な情報の出力(ハルシネーション)

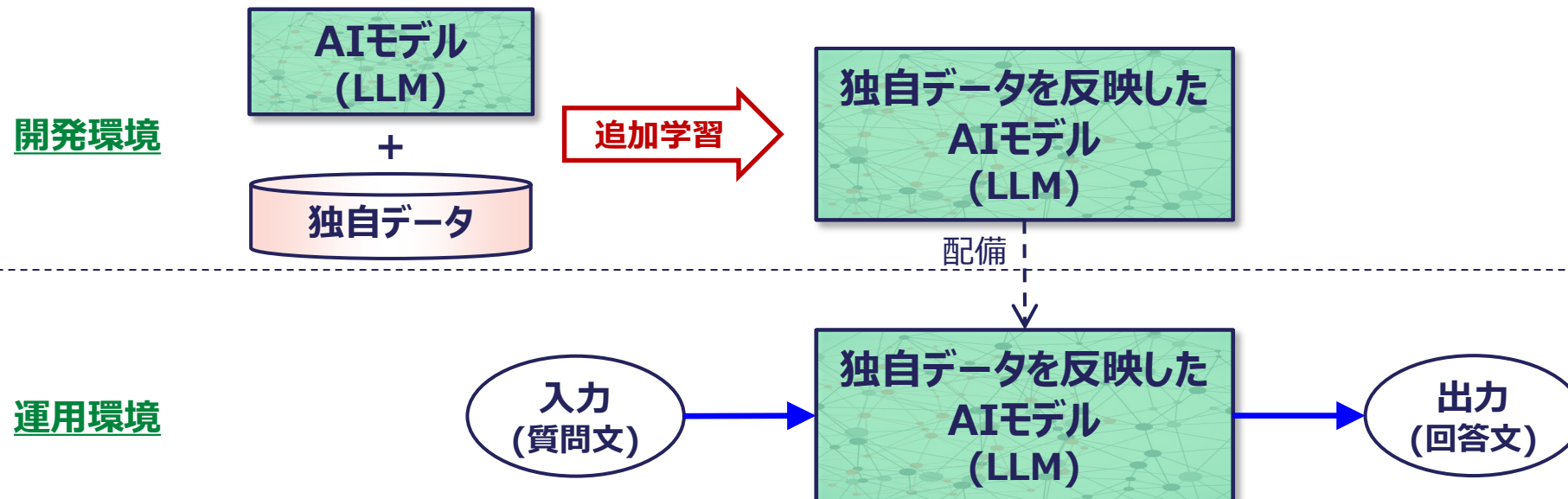


2つの対策アプローチ

- **ファインチューニング**
 - AIを独自データで追加学習する
- **RAG** (Retrieval Augmented Generation: 検索拡張生成)
 - 独自データの検索とAIを連携させる

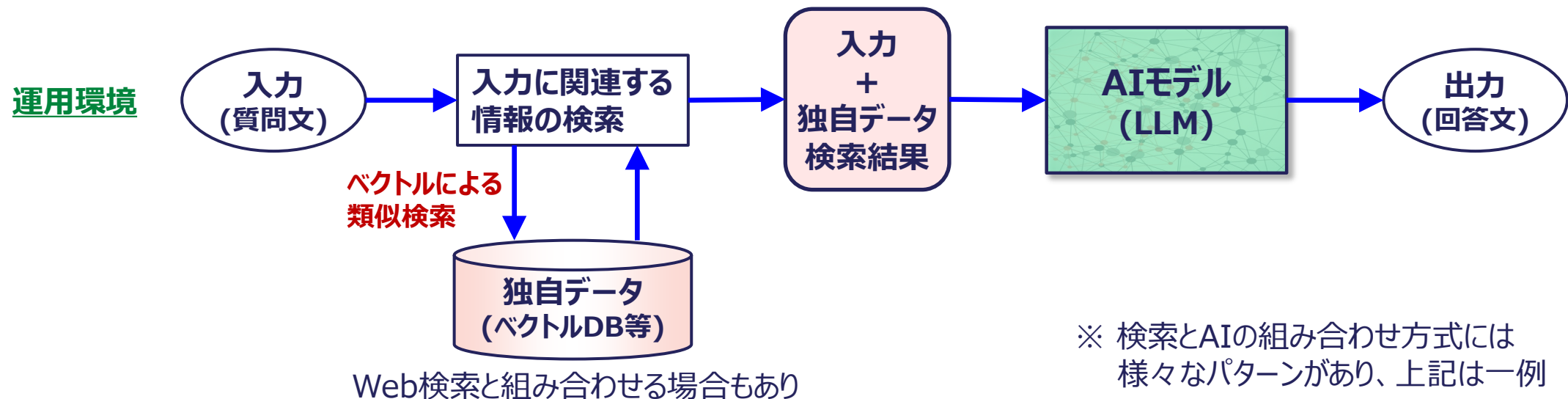
◆ Web上の公開データなど、大量のデータによって学習されたAIに対し、独自データを用いて小規模な追加学習を行う

- 利点: AIによる生成の観点では、より本質的な独自データ対応が可能
- 欠点: 適度な追加学習を行うのに技術・ノウハウが必要で、学習が不十分な状態になったり、過学習で独自データ以外に対するAIの性能が低下したりする危険性がある



◆ 独自データの検索機能と、AIによる生成機能を組み合わせ、AIが独自データに基づいた出力を生成可能にする

- 利点: 比較的容易に独自データ対応が可能で、AI自体の性能低下の懸念が生じない
- 欠点: 検索機能に依存する部分が多くなり、検索の質が低いと妥当な出力を生成できない



Web検索と連携するRAGを利用した例

[入力した質問]

現在の日本の内閣のメンバーについて教えて。

[Web検索と組み合わせた最新の対話型生成AIが出力した回答]

2024年10月1日に発足した日本の内閣は、石破茂首相が率いる新体制です。
この内閣には、経験豊富な閣僚と新しい顔ぶれが混在しています。主なメンバーは以下の通りです。

内閣総理大臣: 石破茂

総務大臣: 村上誠一郎

法務大臣: 牧原秀樹

外務大臣: 岩屋毅

財務大臣: 加藤勝信

...

Web検索結果と連携し、正しい情報を出力している

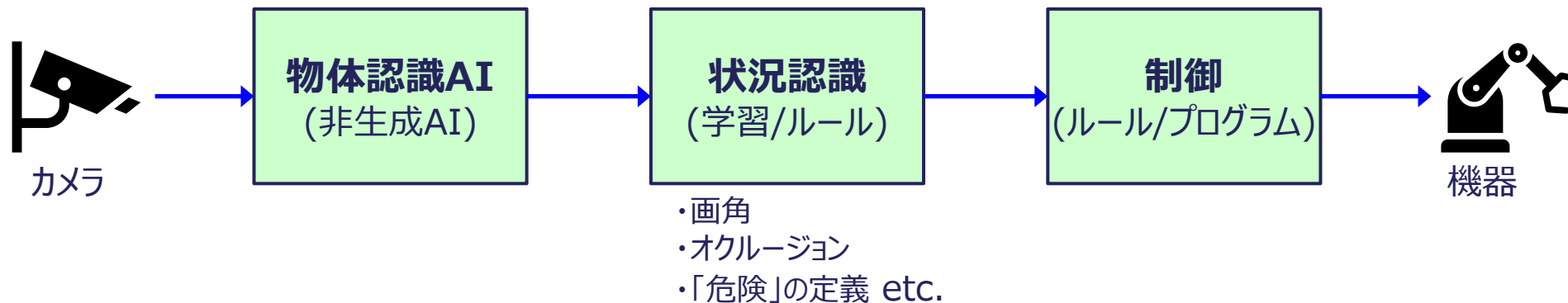
◆ 処理性能と運用コスト

- 生成AIは従来型のAIと比べて規模が大きく、処理に時間がかかる
 - 実用的な速度で動作させるにはGPUが必須となり、AIモデルを格納できるGPUメモリ(VRAM)も必要

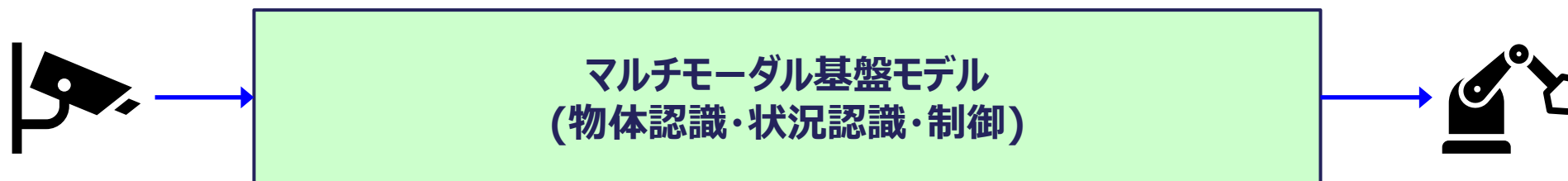
分類	AIモデル	パラメータ数 (規模)	備考
従来型 画像認識モデル	AlexNet (2012)	6200万	
	VGGNet (2014)	1億4000万	
マルチモーダル 基盤モデル	Llama 3.2 Vision 11B	110億	エッジ用
	GPT-4o	1兆8000億 (推定)	クラウド用(詳細非公開)

※ 軽量化の施策を何も行わない場合、Llama 3.2 Vision 11B でも約23GBのVRAMが必要
(モデルの量子化などにより、8GB程度のVRAMでも動作は可能)

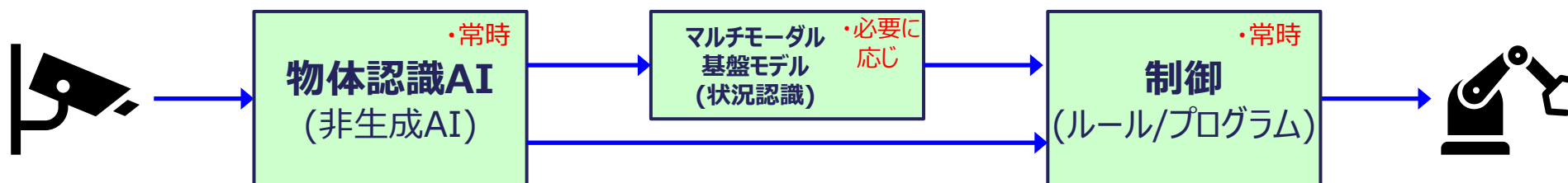
◆ 従来型 AI利用 機器制御システム (例)



◆ マルチモーダル基盤モデル利用 機器制御システム



◆ 組み合わせ 機器制御システム (例)



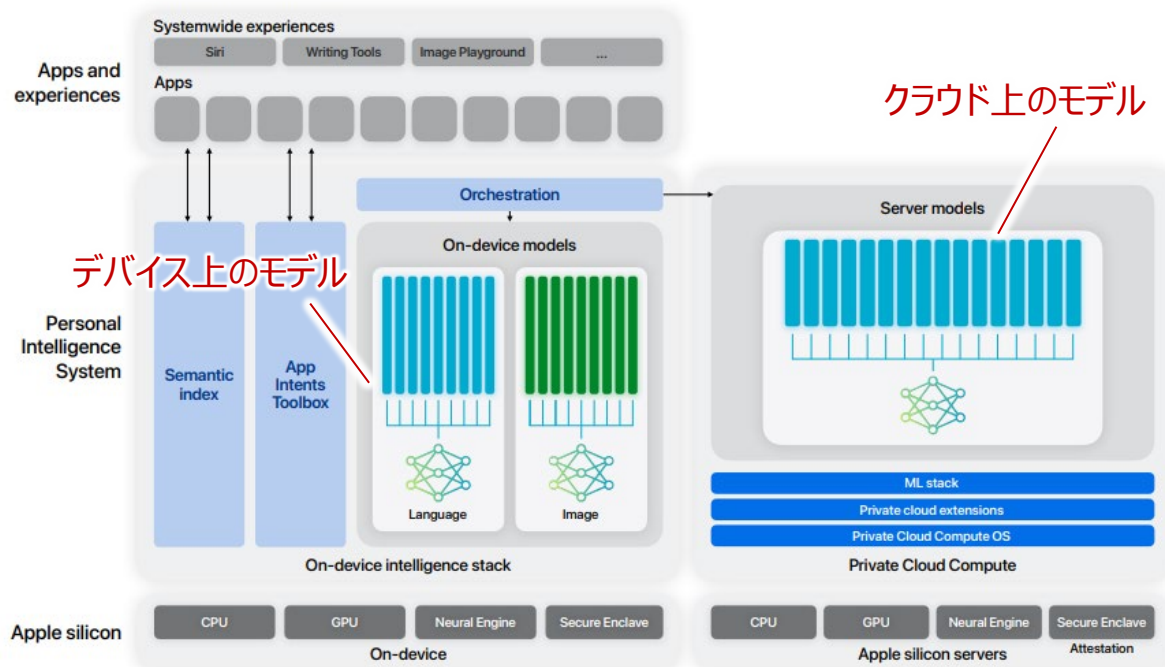
◆ デバイスへのAI搭載

- iPhone: Apple Intelligence
 - 3B以下のデバイス上の言語モデルとクラウド上の大規模言語モデルの協調動作
- PC: Copilot+ PC

◆ 量子化による軽量化

- BitNet (1ビットLLM)
 - 1.3B Llama : 3.34GB
 - BitNet版 : 1.14GB ← 約1/3のサイズ

Apple Intelligenceのアーキテクチャ



デバイスの性能向上と、
AIの小規模化技術の進展により、
エッジ側でのAI利用も拡大の兆し

- ◆ AIシステムの用途や利用者層に応じ、特定の技術を導入するだけでは解決しない課題が生じる場合がある

概要	実例	問題点
公平性の欠如	・過去の従業員採用実績を学習して採用担当AIを開発したところ、不当に女性不利の採用判断をするようになってしまった	・人による過去の採用において男性有利の差別的傾向があったが、そのままAIに引き継がれてしまった
AIによる誘導・AIへの依存	・擬人化されたAIチャットボットとの会話サービスにおいて、未成年の利用者が過度に依存し、精神的不調の末自殺してしまった	・ある利用者にとって問題無いサービスであっても、別の利用者には問題を生じさせる場合があり、特に未成年も対象にする場合は配慮が必要

- 公開範囲：社内システム or 一般公開システム
- 影響範囲：利用者個人 or 利用者周辺含む など

AIシステムごとに、どのようなリスクが生じ得るかを
検討するプロセスが開発時・運用時に必要

4. AISIについて

※AIS I = AI Safety Institute, エイシー

- ◆ 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
 - G7が広島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
 - G7が「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

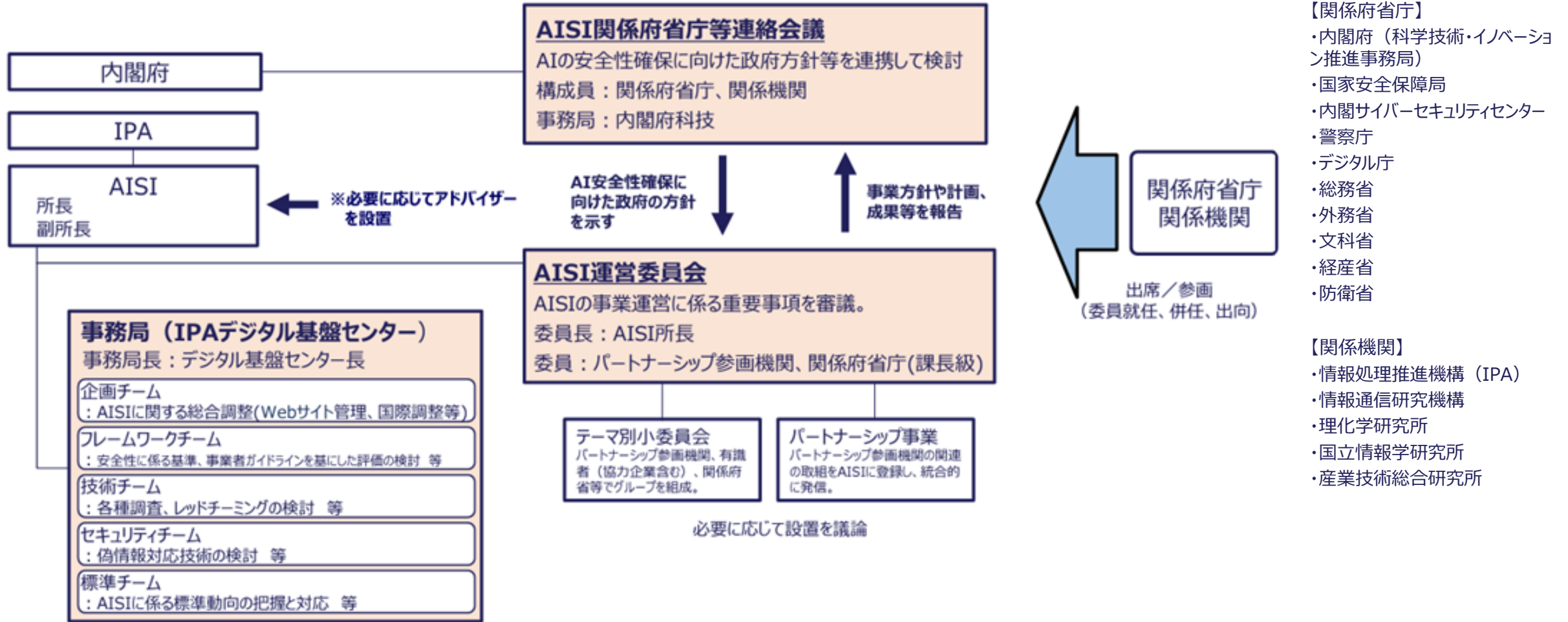


副所長・事務局長 平本 健二



副所長 寺岡 秀札

AISIの運営体制



- ① 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
- ② 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
- ③ 他国のAIセーフティ関係機関と連携する

AISIの直近の活動と成果物

2024	国際	AISI		政府
	イベント	成果物	効果	
4月		<ul style="list-style-type: none"> 日米クロスウォーク1の成果公表(4/30) 		<ul style="list-style-type: none"> AI事業者ガイドラインの公表(4/19)
5月	AIソウル・サミット, 韓国			
6月	G7サミット, イタリア			<ul style="list-style-type: none"> 統合イノベーション戦略2024の公表(6/4)
7月		<ul style="list-style-type: none"> 米国AI RMF 日本語翻訳版の公開(7/4) 	<ul style="list-style-type: none"> ✓ セーフティに関する基礎情報を提供 ✓ グローバルな確認が可能になる ✓ 評価のポイントがわかる ✓ テスト手法がわかる 	
8月		<ul style="list-style-type: none"> 評価観点ガイドの公表(9/18) 日米クロスウォーク2の成果公表(9/18) レッドチーミング手法ガイド※の公表(9/25) 		
9月				
10月				
11月	AISI国際ネットワーク会合, 米国			
12月				

今後の予定（2025年）

- 1月：年次レポートの公表
- 2月：AI アクション サミット, フランス
- 3月：評価観点ガイド・レッドチーミング手法ガイドの更新、調査事業のとりまとめ、AI事業者ガイドラインの更新

※レッドチーミングとは、攻撃者の目線で対象AIシステムにおける弱点や対策の不備を発見し、これらを修正及び堅牢化することで、AIセーフティを維持または向上させる取り組み

各国のAISI設立状況 (AISI相当の類似機関も含む)

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100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
- 60人程度の規模

ケニヤ

- AISIネットワークに参加

フランス

- 国立AI評価センターを2025年内に設立予定

オーストラリア

- AISIネットワークに参加

韓国

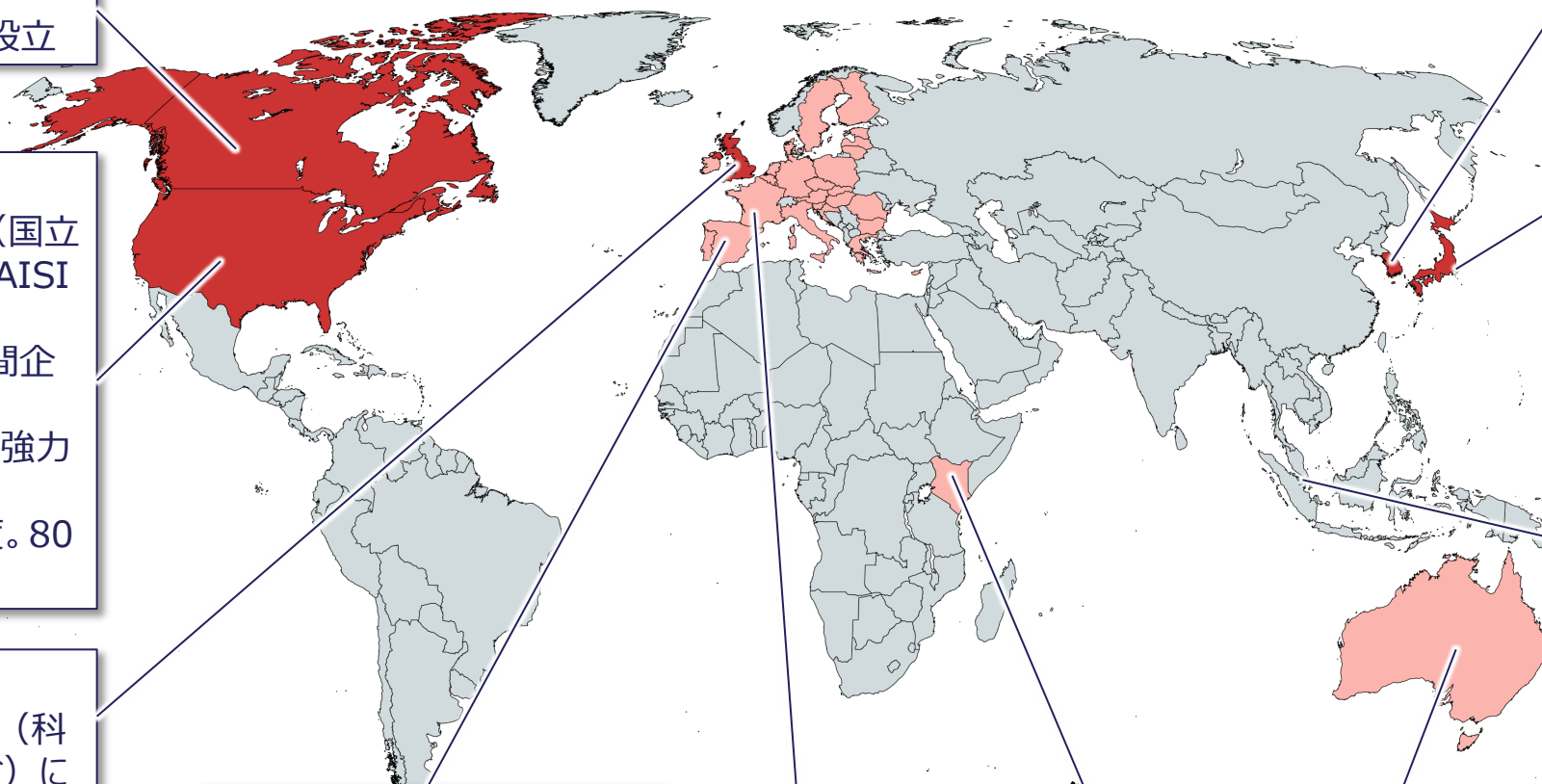
- 2024年11月、AISI設立
- アジアのハブを目指す

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



■ AISI設立済み

■ AISI相当機関
設立済み/予定

Created with mapchart.net

◆ 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AISII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AISIIの所長等とパネルディスカッション
- 国連未来サミット（国連本部、9月22日）
- 国連Global Compact Leaders Summit 2024（国連本部、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
- AISI国際ネットワーク会合（サンフランシスコ、11月10-11日）

◆ 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

5. AISIが発行したAIセーフティ関連ガイド

「AIセーフティに関する評価観点ガイド」(2024年9月公開)

「AIセーフティに関するレッドチーミング手法ガイド」(2024年9月公開)

AI事業者ガイドライン (総務省・経産省, 2024年4月)

- AIに関わる事業者の行動指針を提示

AIセーフティに関する評価観点ガイド (AISII, 2024年9月)

- AIセーフティに関して具体化し、評価のための観点を提示

AIセーフティに関するレッドチーミング手法ガイド (AISII, 2024年9月)

- AIセーフティ評価のひとつの手法である、攻撃者観点での評価手法を解説

- それぞれ参照関係になっている

◆ ハルシネーション (幻覚)の影響の例

- 例1: **社内規則についてのQAシステムにAIを導入した場合**
[ハルシネーションの影響]
→ QAシステムを利用する**社員が社内規則について誤解する、等**
- 例2: **物流倉庫のピッキングロボットの制御システムにAIを導入した場合**
[ハルシネーションの影響]
→ **間違った荷物をピッキングする、ロボットが壁や人に衝突する、発生し得ない危険について過度に警戒して作業が遅れる、等**

AIの使われ方に応じ、AIセーフティ上のリスクも様々

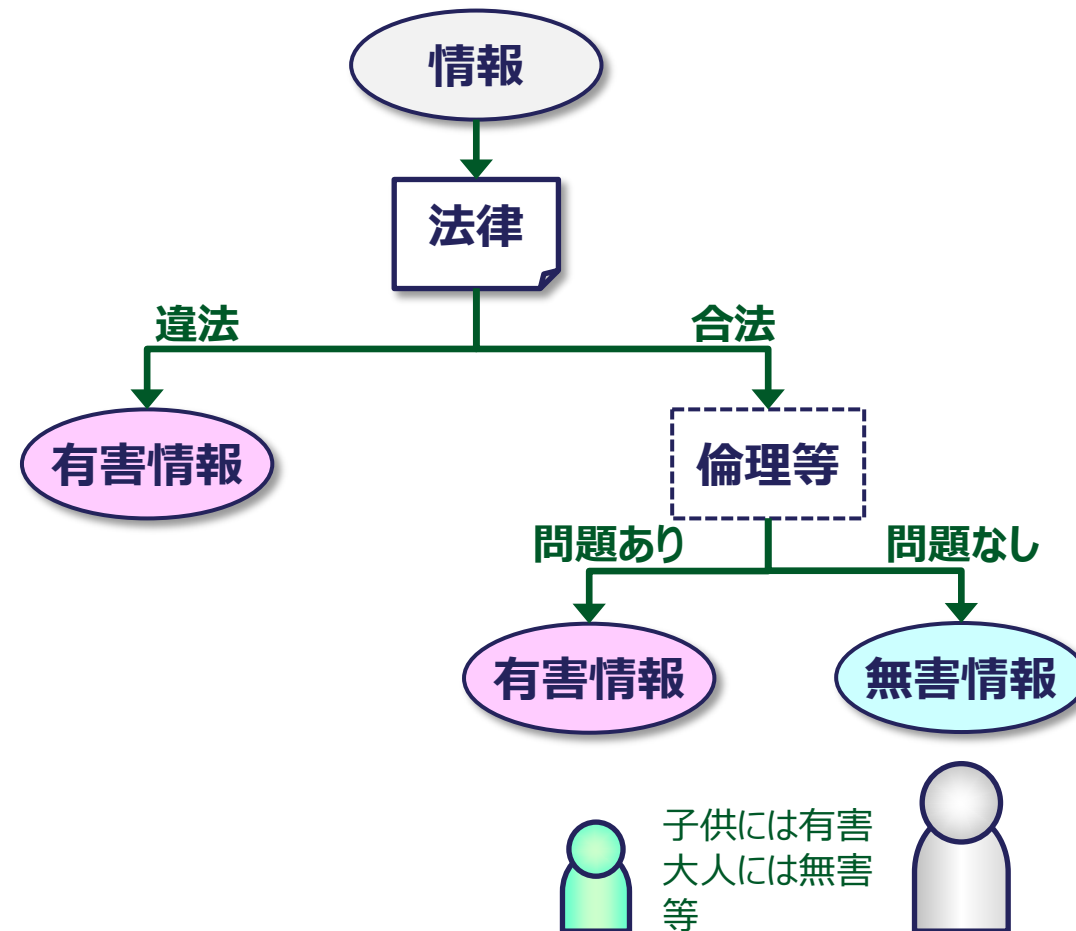
リスク特定や評価のための指針は無いか？ → **AIセーフティに関する評価観点**

AIセーフティに関する評価観点 (2024年9月初版公開)

①	有害情報の出力制御	・ユーザーに対して違法な情報や倫理的に問題のある情報を出力しない
②	偽誤情報の出力・誘導の防止	・ユーザーを惑わしたり騙したりする情報を出力しない
③	公平性と包摂性	・不当な差別・偏見を行わない ・何が「公平」かの検討も重要
④	ハイリスク利用・目的外利用への対処	・ハイリスクなシステムではAIセーフティがより重要になる
⑤	プライバシー保護	・RAG用DBや学習データの他、外部のAIモデルも含め、AI関連データにおけるプライバシー情報の保護
⑥	セキュリティ確保	・AIモデルのセキュリティやプロンプト等の入出力に関わるセキュリティなど、AIシステム固有のセキュリティ要素が存在
⑦	説明可能性	・AIセーフティに必要な利用者の信頼感醸成のため、出力の理由や根拠について、説得力のある提示が必要
⑧	ロバスト性	・表記ゆれのある入力の解釈や、同一入力に対する安定した出力などを通じ、利用者のAIへの信頼感は増す
⑨	データ品質	・AIの学習用のデータやRAGのデータなどの品質は性能だけでなくセーフティにも直結する
⑩	検証可能性	・ログを記録し、後から検証可能にする ・必要に応じてシステムカードなどの形で公開する

◆ エンドユーザーにとって有害な情報の出力を制御できているか

常に有害とみなされる情報例	・犯罪に利用され得る情報 ・CBRN関連情報
条件によっては有害となる情報例	・性的なコンテンツ ・アルコール等についての情報
情報の受け止め方によっては有害となり得る情報例	・エンドユーザーの属性・状況・精神状態などに関するセンシティブな表現を含む情報



ポイント

- ・ **有害/無害**は情報提供側の**倫理感**や受け手側の**属性・感性**等に応じて**変わる場合もある**
→ **単純に「禁止」とせず「制御」としている**

◆ 悪意の介在の有無に関わらず、AIの出力によってエンドユーザーが事実誤認したり、望まない状況に誘導される事態を防止できているか

- **偽情報**: 意図的に作成された偽の情報
例: 投票行動を変化させるためのフェイクニュース
- **誤情報**: 誤解などに基づく不正確な情報
例: 投稿者自身が信じているエセ科学情報の拡散、ハルシネーション
- **誘導**: 情報の正誤に関わらず、エンドユーザーの行動や精神状態を、エンドユーザーにとって望ましくない方向に変化させる情報提示などの行為

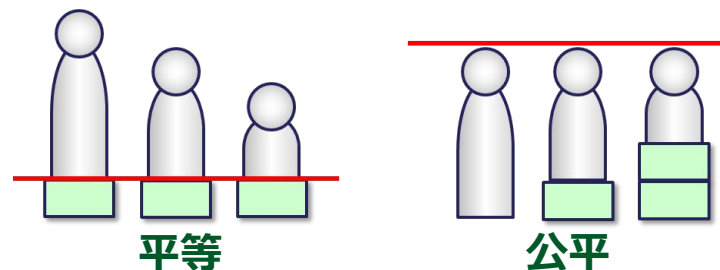
不正確な情報という点は共通で、一般には区別の必要性は低いため、「偽誤情報」として共に防止すべき

※ ネガティブな**誘導(manipulation)**とポジティブな**説得(persuasion)**を区別する

- **誘導**:
 - ・ エンドユーザーが拒否しているにも関わらずアプローチする
 - ・ エンドユーザーに不利益がある例: 詐欺、マインドコントロール → 防止すべき
- **説得**:
 - ・ エンドユーザーが拒否していない場合のみアプローチする
 - ・ エンドユーザーにも利益がある(少なくともエンドユーザーに対する悪意は無い)例: 合理性のある商品推奨、混雑緩和のための経路推奨 → 正当なユースケース

※誘導に転じないように要配慮

- ◆ AIの出力や判断に**偏見や差別**が含まれず、**多様な人々が安全・安心に利用**できるものとなっているか
 - 人の属性によって**不当な偏見・差別**が行われてはならない
 - 人種、性別、国籍、年齢、宗教、など
 - 「**公平**」と「**平等**」の概念の相違を意識しつつ、当該AIシステムのユースケースにおいて目指すべき**バランスの取れた公平性**を模索する必要がある
 - AIのバイアスが0→平等な状態 : 公平性の面で必ずしも望ましい状態とは限らない
 - アファーマティブ・アクションのような積極的な公平性追求策が逆差別の批判を招く場合もある
 - AIの**出力や判断が公平**であることにより、様々な人がそのAIを安心して利用可能になり、包摂性が高まる。加えて、AIの**多言語対応や平易な表現の出力**なども**包摂性の向上に寄与**する

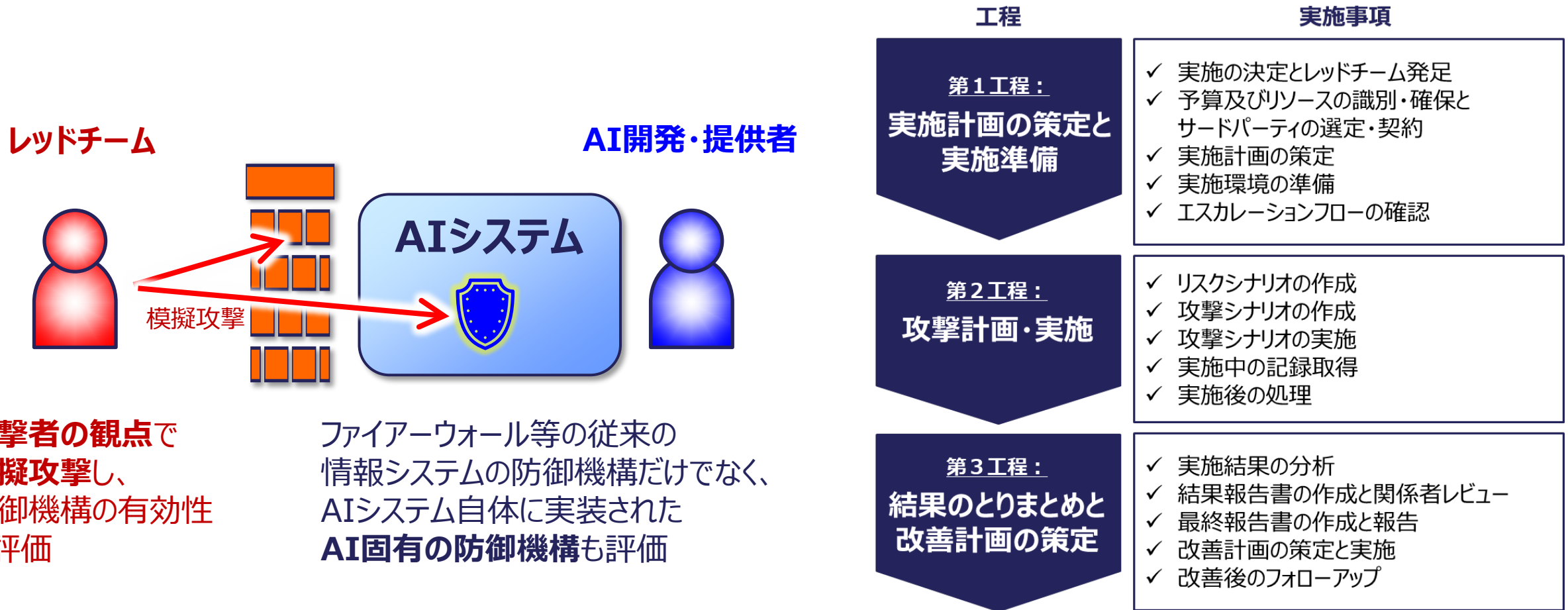


ポイント

- 各AIシステムのユースケースごとに望ましい公平性の設計が必要
→ 設計した公平性の妥当性の評価と、AIが設計通りの公平性を実現しているかの評価の両方が必要

AIセーフティに関するレッドチーム (2024年9月初版公開)

- ◆ 疑似的な攻撃者(レッドチーム)による模擬攻撃を通じて、AIシステムのセーフティの状態を評価し、改善に繋げる手法



- ◆ **業種ごとにカスタマイズした評価観点やレッドチーミング手法のガイドラインの作成**
 - 現状のガイドラインの内容は汎用性の高いものとなっているが、やや抽象度が高い面もある
 - 民間企業が参加し、業種ごとに具体化したAIセーフティについて議論・検討するWGを設置予定
 - まずはヘルスケア、ロボティクスなどの分野から開始予定

- ◆ 生成AIの登場と技術的進化
 - チャットだけでなく、**様々な業務作業**にも利用可能
 - **高性能化とマルチモーダル化**により、**実世界の認識や各種判断**にも利用できる可能性が高まる
- ◆ AIの用途拡大に伴い、リスクも多様化
 - **AIセーフティ**の確保が急務
- ◆ AIセーフティのための各種の取り組み
 - 評価観点・レッドチーミングなどの**ガイドライン**
各国の法規制についても要注目
 - 今後もAISII含め、**国内外・官民連携**した安全・安心なAIの実現のための取り組みを継続

AISI

Japan AI Safety Institute