参考資料1-3
科学技術・学術審議会
情報委員会（第47回）
令和8年5月12日

理研AIPセンター
社会における人工知能研究グループ研究成果

グループディレクター　橋田 浩一

RIKEN　AIP 革新知能統合研究センター
Center for Advanced Intelligence Project

概要

- 目標
  - 技術・倫理・経済・制度・文化を横断してAIと人間社会の持続可能な共生の基盤を構築
- 成果
  - セキュリティ: 攻撃耐性と機密性を両立する基盤技術
    - モデルマージ防御
    - 秘密計算・差分プライバシー・敵対的攻撃対策
  - 公平性・説明可能性
    - Demographic Parity制約下の最適回帰
    - fairwashingの分析
    - 医療AIの説明可能分類
  - 情報空間とデータ健全性: 情報生態系のリスク評価・低減
    - 偽誤情報回避行動
    - 生成AIによるデータ汚染
    - ダークパターンの社会影響
  - ビッグデータによる行政の高度化
    - バイアス補正・因果推論・政府統計融合
    - 政府統計の改善
    - AIと労働市場の実証分析
  - 人間とAIの共生エコシステム設計
    - 利用時品質モデルと人間中心設計
    - テクノアニミズム再考
    - AIの法的人格・責任モデル
    - AIの総合的ガバナンスの設計…EU AI法の整合標準の作成
    - パーソナルAI…総合的ガバナンスの確実な普及
    - グラフ文書…認知的オフローディング防止

人工知能のセキュリティ・プライバシー（佐久間）

- セキュリティ・プライバシー分野で多くの「世界初」
- 世界初のモデルマージに対するプロアクティブ防御 (ICCV'25)
- 世界初のTransformerに対するマルチパーティー秘密計算 (EurpS&P'23)
- 世界初の自然物体/自然音を模した敵対的サンプル (AAAI'20)
- 世界初の物理的敵対的サンプル (IJCAI'19)
- 世界初の局所差分プライバシー保証下における経験リスク最小化 (DS'17)
- 世界初の実用規模データを用いた完全準同型暗号による秘密計算 (NDSS'17)

経済経営情報融合分析（星野）

①データ融合と因果効果に関する方法論開発

中間欠測のある繰り返し継続時間モデルでのマクロ情報を用いた識別(Igari&Hoshino,2018,CSDA)

バイアス標本での二重にロバストな推定(Shimizu&Hoshino,2019,Stats)

ガウス過程正準モデルによる複数データ融合(Mitsuhiro&Hoshino,JJSD)

入れ子型構造の離散選択モデル(Miyazaki,Hoshino & Böckenholt,2021,JBES)

介入ラベルとアウトカムが紐づかない状況での因果効果推定(Shinoda & Hoshino,2022,AAAI)

内生性バイアスを除去した広告効果推定法の開発と実データへの応用(Emori… Hoshino,2024,KDD)

ノンコンプライアンス下での因果推論(Ota, Hoshino and Otsu, 2025,DP)

②政府統計の改善

総務省統計センターと理研AIPの連携協定を締結(2018-)

総務省消費統計関連の方法論開発と公表系列化(総務省HP:松永・・・星野,2021)

特に「CTIミクロ(内閣の月例経済報告に採用)のモニター融合」「全国家計構造調査の年集計の開始」

⇒関連して政府EBPMにおけるビッグデータ活用についての各種委員会参加(内閣官房・内閣府・総務省・経済産業省等)

③経済状況のナウキャスティングとメカニズム理解

オルタナティブデータを用いたREIT価格のナウキャスティング (Nakakita et al., 2025, IIAI-AAI)

ファイナンシャル時系列でのLLMを利用したレジームスイッチ(Morita et al.,2025,IIAI-AAI)

④AIと労働市場の研究

AIによって労働のどの部分が特に代替されるか、生産性向上の程度の理解

日本公認会計士協会との共同研究(2018-2022)

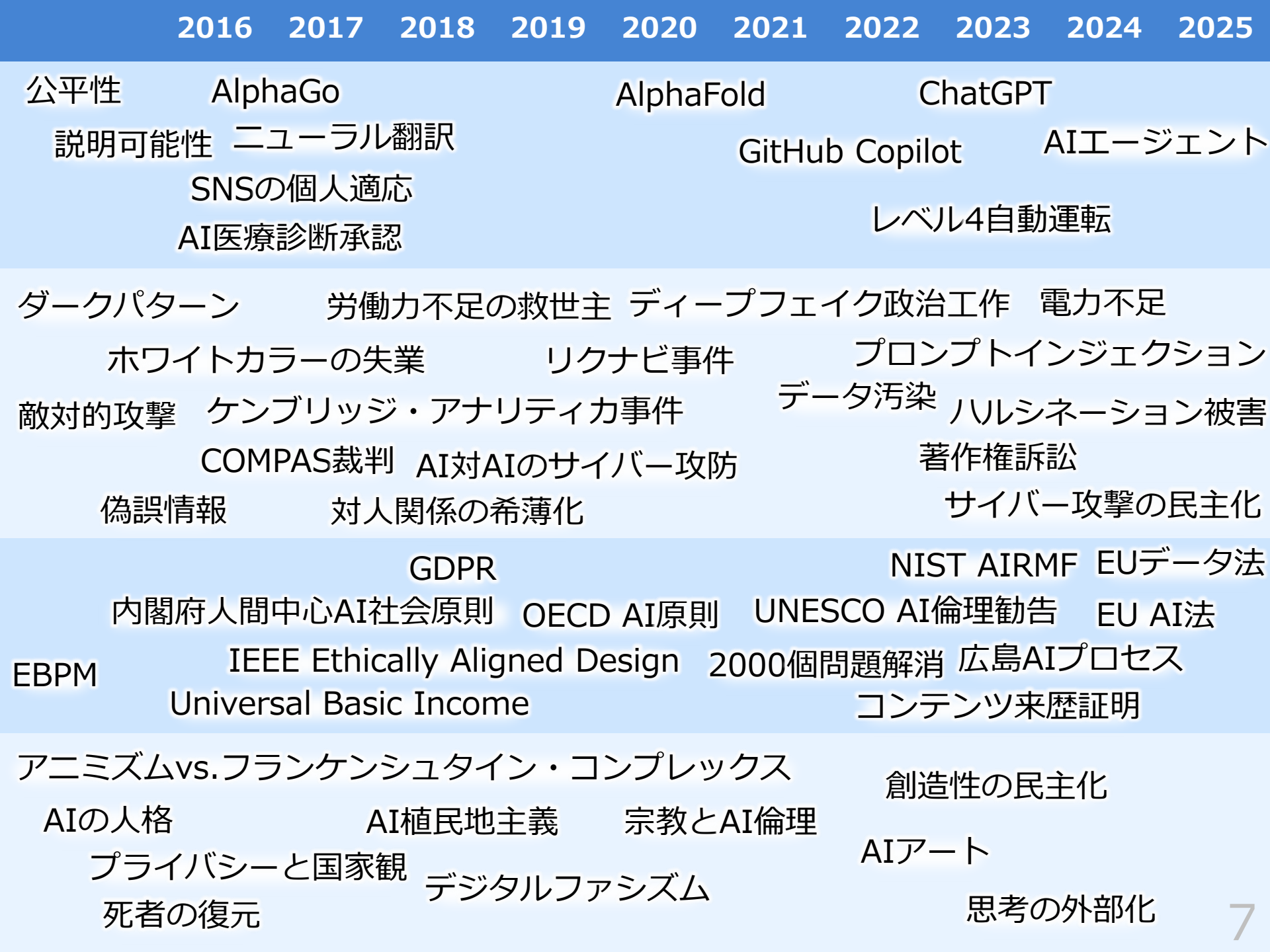
日本税理士会連合会との共同研究(2023-)

分散型ビッグデータ処理（橋田）

- パーソナルデータの分散管理
  - パーソナルデータを本人に集約してフル活用
    - パーソナルデータを本人が私的目的に利用することは無制限
  - PLR (分散パーソナルデータストア)を教育用に実運用
- パーソナルAI
  - 各個人に専属のAIエージェント
  - 10年で確実に普及 → AIの総合的ガバナンス
- AIの総合的ガバナンスの標準化
  - ログデータを利用者に集約することで利用者とサードパーティがAIのリスク管理に関与
  - 欧州AI法の整合標準にその旨を記述
- グラフ文書: 文書としての知識グラフ
  - テキスト文書より作成効率が高く他の方法より批判思考力を高める効果が大きい
  - AIを使っても批判思考力を高める効果が保たれる
  - 知識グラフはAI活用の基盤…人間にもAIにも理解できるデータ

社会におけるAI利活用と法制度（中川）

- プライバシー保護のための匿名化技術タスク　PWSCUPの開催と参加（優勝１回）
- 内閣府：人間中心社会AI原則　（分担執筆）
- IEEE　Ethically Aligned Design (一部執筆)
- AIエージェントの法的人格付与の調査分析（情報ネットワークローレビューなど）
- AIエージェント（Agentic AIとも言う） を個人および企業において社会利用する場合の問題点，課題の調査提案および個人代理に使う方法の提案


2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
公平性
AlphaGo
AlphaFold
ChatGPT
説明可能性
ニューラル翻訳
GitHub Copilot
AIエージェント
SNSの個人適応
AI医療診断承認
レベル4自動運転
ダークパターン
労働力不足の救世主
ディープフェイク政治工作
電力不足
ホワイトカラーの失業
リクナビ事件
プロンプトインジェクション
敵対的攻撃
ケンブリッジ・アナリティカ事件
データ汚染
ハルシネーション被害
COMPAS裁判
AI対AIのサイバー攻防
著作権訴訟
偽誤情報
対人関係の希薄化
サイバー攻撃の民主化
GDPR
NIST AIRMF
EUデータ法
内閣府人間中心AI社会原則
OECD AI原則
UNESCO AI倫理勧告
EU AI法
EBPM
IEEE Ethically Aligned Design
2000個問題解消
広島AIプロセス
Universal Basic Income
コンテンツ来歴証明
アニミズムvs.フランケンシュタイン・コンプレックス
創造性の民主化
AIの人格
AI植民地主義
宗教とAI倫理
AIアート
プライバシーと国家観
デジタルファシズム
死者の復元
思考の外部化

以下，主要な研究成果の詳細説明です

人工知能セキュリティ・プライバシーチーム(佐久間)

- セキュリティ・プライバシー分野で多くの「世界初」
- 世界初のモデルマージに対するプロアクティブ防御 (ICCV'25)
- 世界初のTransformerに対するマルチパーティー秘密計算 (EurpS&P'23)
- 世界初の自然物体/自然音を模した敵対的サンプル (AAAI'20)
- 世界初の物理的敵対的サンプル (IJCAI'19)
- 世界初の局所差分プライバシー保証下における経験リスク最小化 (DS'17)
- 世界初の実用規模データを用いた完全準同型暗号による秘密計算 (NDSS'17)

AI医療診断承認

説明可能な悪性リンパ腫病理画像の病型分類 (MICCAI'25)

要求

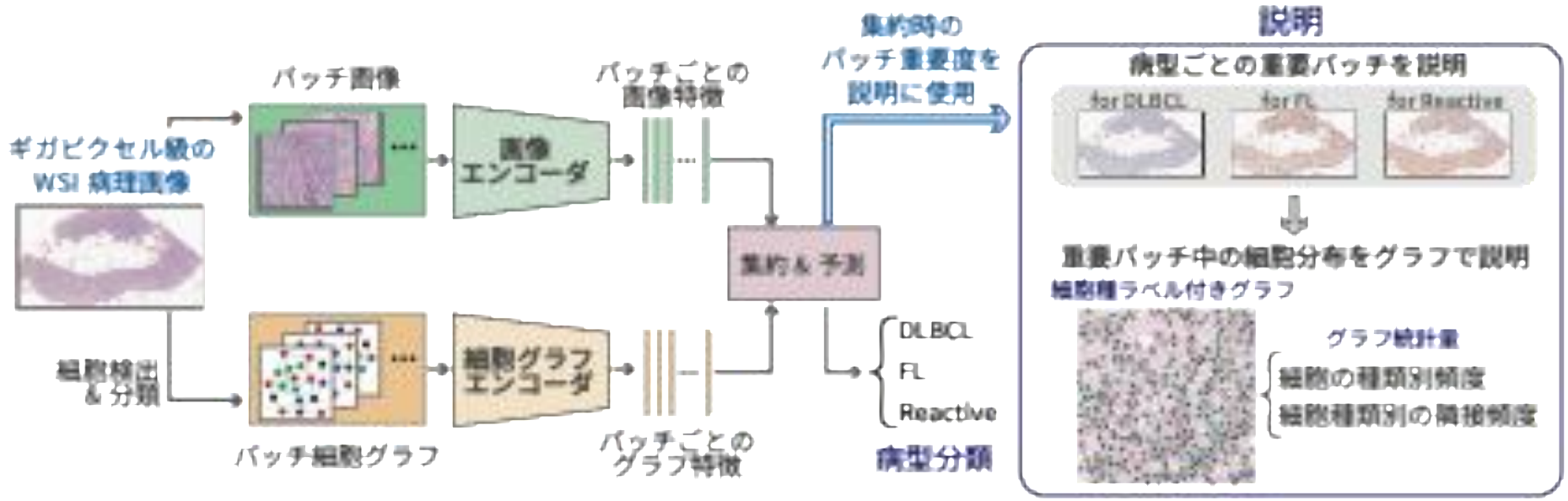
1. WSI（超高解像度）の高精度分類
2. WSI中の病型特有の病変部の説明
3. 病変部中の細胞空間分布の説明

→ 病理医の診断根拠と対応

アプローチ

1. → 画像と細胞グラフのマルチモーダル活用
2. → パッチの病型ごとの重要度で病変部を説明
3. → 細胞グラフのグラフ統計量で空間分布を説明

説明例: 予測病型Aのこの症例は細胞X頻度が50個

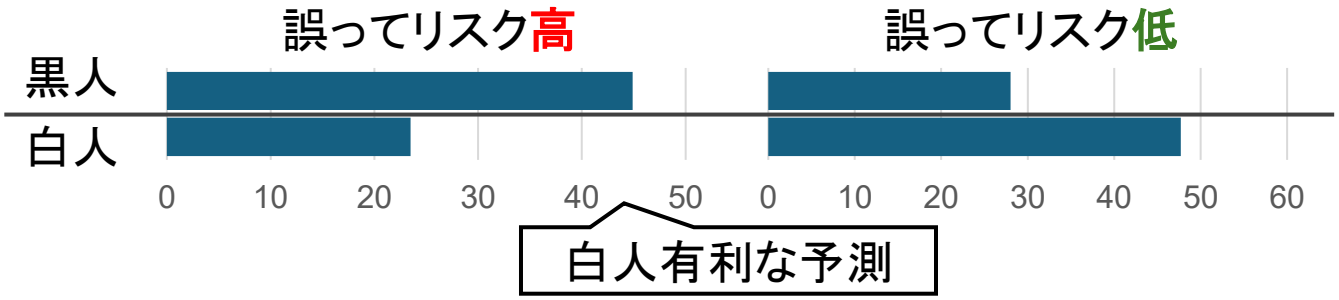


Daiki Nishiyama, Hiroaki Miyoshi, Noriaki Hashimoto, Koichi Ohshima, Hidekata Hontani, Ichiro Takeuchi and Jun Sakuma. Explainable Classifier for Malignant Lymphoma Subtyping via Cell Graph and Image Fusion. *MICCAI'25*

COMPAS裁判

差別しないAIモデル構築
(NeurIPS'23, ICML'25)

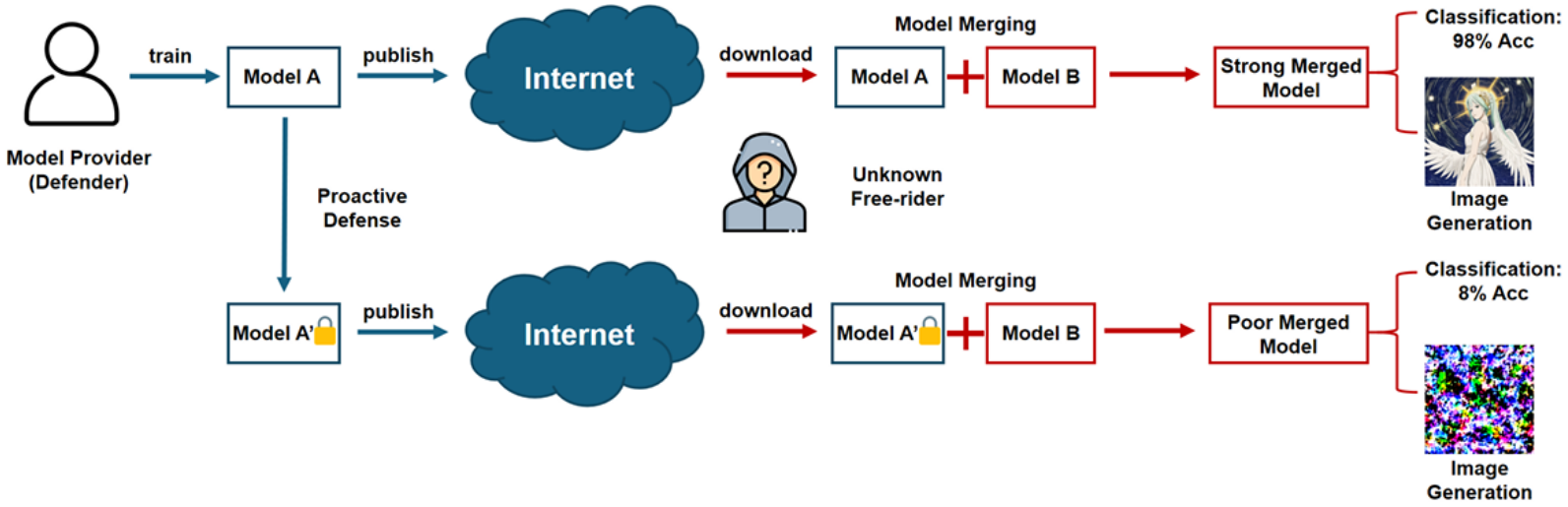
- COMPASによる累犯リスク予測
  - アメリカの裁判で過去に量刑に使用
  - 10段階のスコアを算出



- 世界のAI法・ガイドラインが**公平性を要求**
  - 広島AIプロセス，EU AI法，IEEE EAD，OECD AI原則，UNESCO AI倫理勧告
- **成果**: 公平かつ最も予測性能の良いアルゴリズムの解明
  - 線形モデルにおける最適回帰 (K. Fukuchi and J. Sakuma. NeurIPS'23.)
  - 後処理型の最適回帰 (K. Fukuchi. ICML'25.)

Kazuto Fukuchi: Meta Optimality for Demographic Parity Constrained Regression via Post-Processing. ICML 2025
Kazuto Fukuchi, Jun Sakuma: Demographic Parity Constrained Minimax Optimal Regression under Linear Model. NeurIPS 2023

モデルマージに対するプロアクティブ防御 (ICCV 2025)



Research Question

- モデルマージはそれぞれ異なるタスクに関する能力を持つ複数のopen weightモデルをトレーニングなしで一つのマルチタスクモデルに統合可能
- モデルマージを特定の者のみに許可し、フリーライドを防ぐことは可能か？

The Solution: PaRaMS

- モデルマージに対するプロアクティブ防御を実現初めての手法を提案
  - 秘密鍵を持つ者はモデルマージが可能
  - 秘密鍵を持たない物はモデルマージをするとモデル性能が大きく劣化
- 画像生成モデル、大規模言語モデルに適用可能

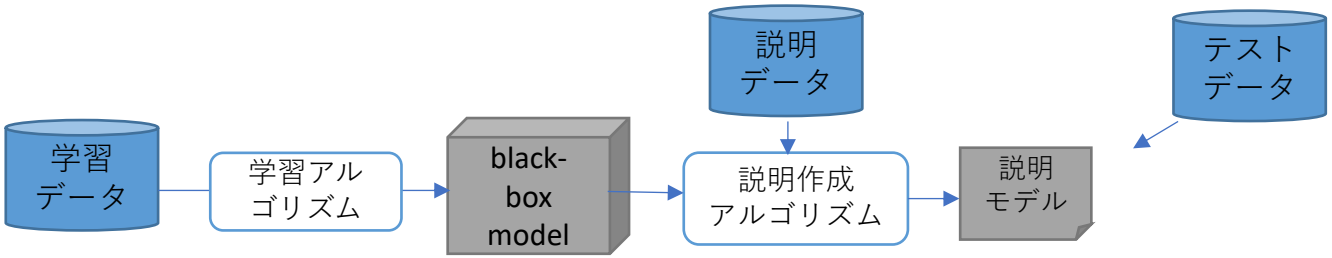
W Junhao, Y Zhe, J Sakuma, Disrupting model merging: A parameter-level defense without sacrificing accuracy, ICCV'25

人工知能安全性・信頼性ユニット(荒井)

- 説明可能AI
  - 機械学習モデルの説明における倫理的なリスクの検証
  - 説明可能AIの目的やニーズについての議論の整理
- AIの公平性・多様性
  - 公平な判断の機械教示
  - クラウドソーシングにおける多様な判断
- プライバシー保護に関する諸問題
  - 秘密計算や匿名化などのプライバシー保護技術
  - プライバシーポリシーの分析
- 情報空間の安全性
  - ヘイトスピーチ検出に向けた日本語データセットの試案
  - ダークパターンの悪影響についての調査
  - ファクトチェック情報に対するクリック行動の分析
  - 生成データによるコンタミネーションの悪影響

機械学習におけるfairwashing

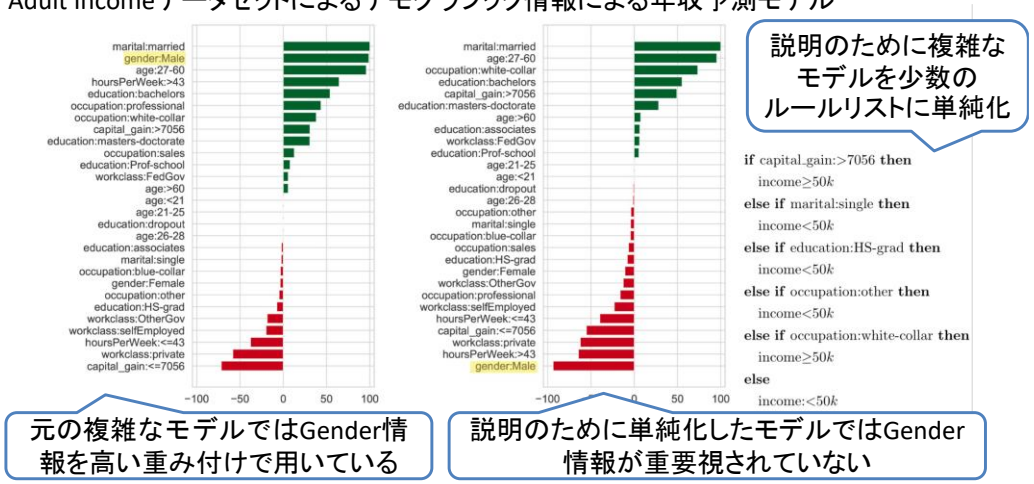
- 機械学習モデルの説明におけるfairwashing
  - 学習モデルがある倫理的な値（fairnessなど）を実際は満たしていないが，満たせていると見せかけること



Fairwashing [Aïvodji+2019]
Unfairなモデルをfairだと見せかけるモデルを生成する手法の指摘

Fairwashingの特徴分析
[Aïvodji+2021]
一般化可能性、検出可能性の検証

Adult Incomeデータセットによるデモグラフック情報による年収予測モデル



AÏVODJI et al. Fairwashing: the risk of rationalization. (ICML2019)
AÏVODJI et al. Characterizing the risk of fairwashing. (Neurips2021)

公平な判断の機械教示

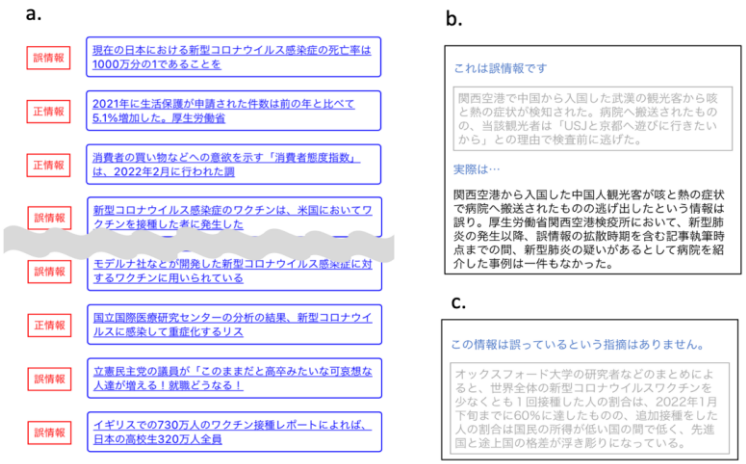
AIから人間へのフィードバックによって公平な判断を促す

- ワーカーは社会的バイアスを含んだラベル付けを行う場合がある
- より公平な判断をクラウドワーカーにさせるための機械教示
- 人間のモデルと公平な機械学習モデルを比較して差異を視覚化
  - 公平配慮機械学習＋XAI＋機械教示による人間への教示

Yang *et al.* Fair Machine Teaches Fair Decisions to Biased People. (CHI2024)

ファクトチェック情報に対するクリック行動の分析

誤情報に対するクリック行動測定

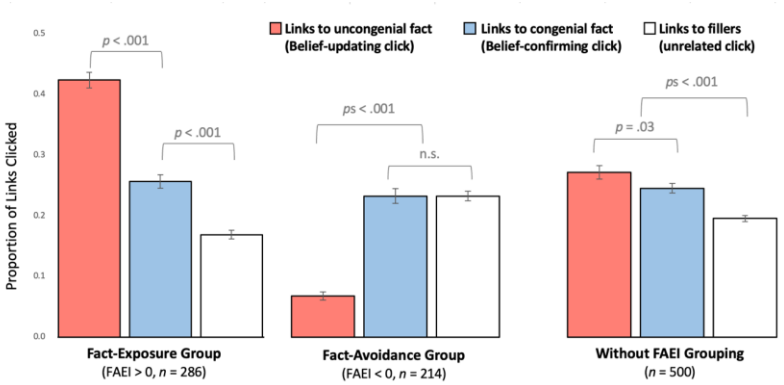


1. いくつかの心理尺度の質問紙調査
2. ユーザーの信じている誤情報を測定
3. 誤情報、誤情報訂正記事、実ニュース記事（filler）のリンクをランダムに提示し（左図）各ユーザーのクリック行動を測定

実験結果より

1. 提案指標FAEIは既存手法に比べ選択的回避行動を区別できた
2. デタラメ受容性尺度は選択的回避行動と関連する

→反射的な閉鎖的思考を緩和し、ユーザーが事実に基づいて誤った信念を振り返ることを促すデザイン介入の必要性



FAEIにより分けられたfact-avoidance groupとfact-exposure groupのクリック行動.

Tanaka et al., Who Does Not Benefit from Fact-checking Websites? A Psychological Characteristic Predicts the Selective Avoidance of Clicking Uncongenial Facts (CHI2023)

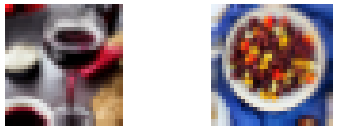
生成データによるコンタミネーションの悪影響

Dataset Creation

To answer this question, we experimentally simulate such a contamination scenario and observe its effects. First, we created ImageNet-like and COCO-like datasets using Stable Diffusion v1.

SD-ImageNet

SD-COCO

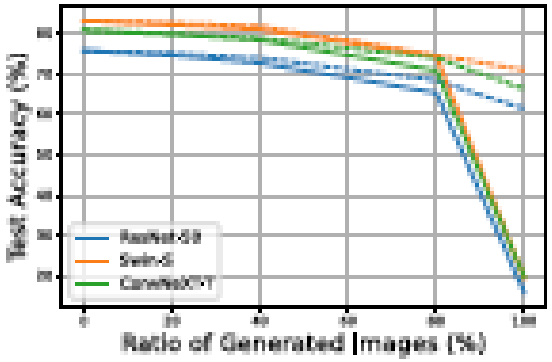


Then, we mixed real images from the original ImageNet (COCO) with generated images from SD-ImageNet (SD-COCO).

Experiments

We observed **performance degeneration** caused by contamination of generated images.

Image Classification



"contaminated" by real images from another dataset (WebVision)

contaminated by generated images

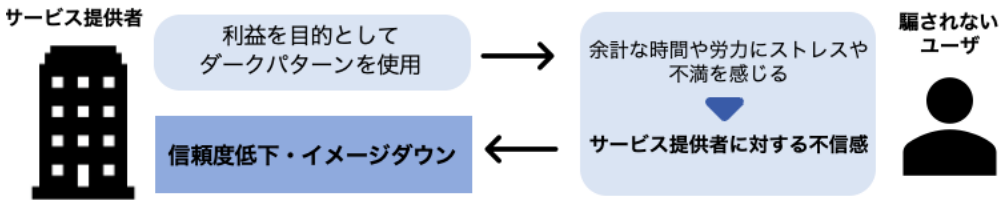
Image Captioning

Test metrics of the BLIP model

	Train Data		BLEU-4↑	SPICE↑	CIDEr↑
contaminated by generated images	Ratio of SD-COCO	0%	0.400	0.340	1.335
		20%	0.391	0.235	1.320
		40%	0.390	0.236	1.319
		80%	0.377	0.233	1.279
		100%	0.287	0.191	1.000
"contaminated" by real images from another dataset	Ratio of Flickr 30k	40%	0.393	0.238	1.326
		100%	0.321	0.215	1.092

Hataya et al., Will large-scale generative models corrupt future datasets? (ICCV2023)

ダークパターンの悪影響



最終的には**ダークパターンの利用が，**
企業の目的に反して利益を損なう結果につながる可能性

調査方法

タスク調査
- ダークパターンを含んだWebサイト上でタスク調査を実施

調査シナリオ
- Webサイト上でデリバリーサービスを利用し，6人分の料理を注文
- 料理の選択から注文完了までの一連の流れをWebサイト上で擬似的に体験

図2：作成したタスク調査用Webサイト

- 参加者からは、騙されたことへの不快感が表明された
- 購買意欲を低下させる可能性があることが示唆された
- ユーザーによっては、ダークパターンに気づいても、回避にかかるコストを見積もりながら、ダークパターンを受け入れることにしているケースもあり、ユーザー体験が悪化している可能性がある。

Kojima et.al., Are Dark Patterns Self-destructive for Service Providers?: Revealing Their Impacts on Usability and User Satisfaction, HCII 2024.

経済経営情報融合分析チーム(星野)

①データ融合と因果効果に関する方法論開発

中間欠測のある繰り返し継続時間モデルでのマクロ情報を用いた識別(Igari&Hoshino,2018,CSDA)

バイアス標本での二重にロバストな推定(Shimizu&Hoshino,2019,Stats)

ガウス過程正準モデルによる複数データ融合(Mitsuhiro&Hoshino,JJSD)

入れ子型構造の離散選択モデル(Miyazaki,Hoshino & Böckenholt,2021,JBES)

介入ラベルとアウトカムが紐づかない状況での因果効果推定(Shinoda & Hoshino,2022,AAAI)

内生性バイアスを除去した広告効果推定法の開発と実データへの応用(Emori… Hoshino,2024,KDD)

ノンコンプライアンス下での因果推論(Ota, Hoshino and Otsu, 2025,DP)

②政府統計の改善

総務省統計センターと理研AIPの連携協定を締結(2018-)

総務省消費統計関連の方法論開発と公表系列化(総務省HP:松永・・・星野,2021)

特に「CTIミクロ(内閣の月例経済報告に採用)のモニター融合」「全国家計構造調査の年集計の開始」

⇒関連して政府EBPMにおけるビッグデータ活用についての各種委員会参加(内閣官房・内閣府・総務省・経済産業省等)

③経済状況のナウキャスティングとメカニズム理解

オルタナティブデータを用いたREIT価格のナウキャスティング (Nakakita et al., 2025, IIAI-AAI)

ファイナンシャル時系列でのLLMを利用したレジームスイッチ(Morita et al.,2025,IIAI-AAI)

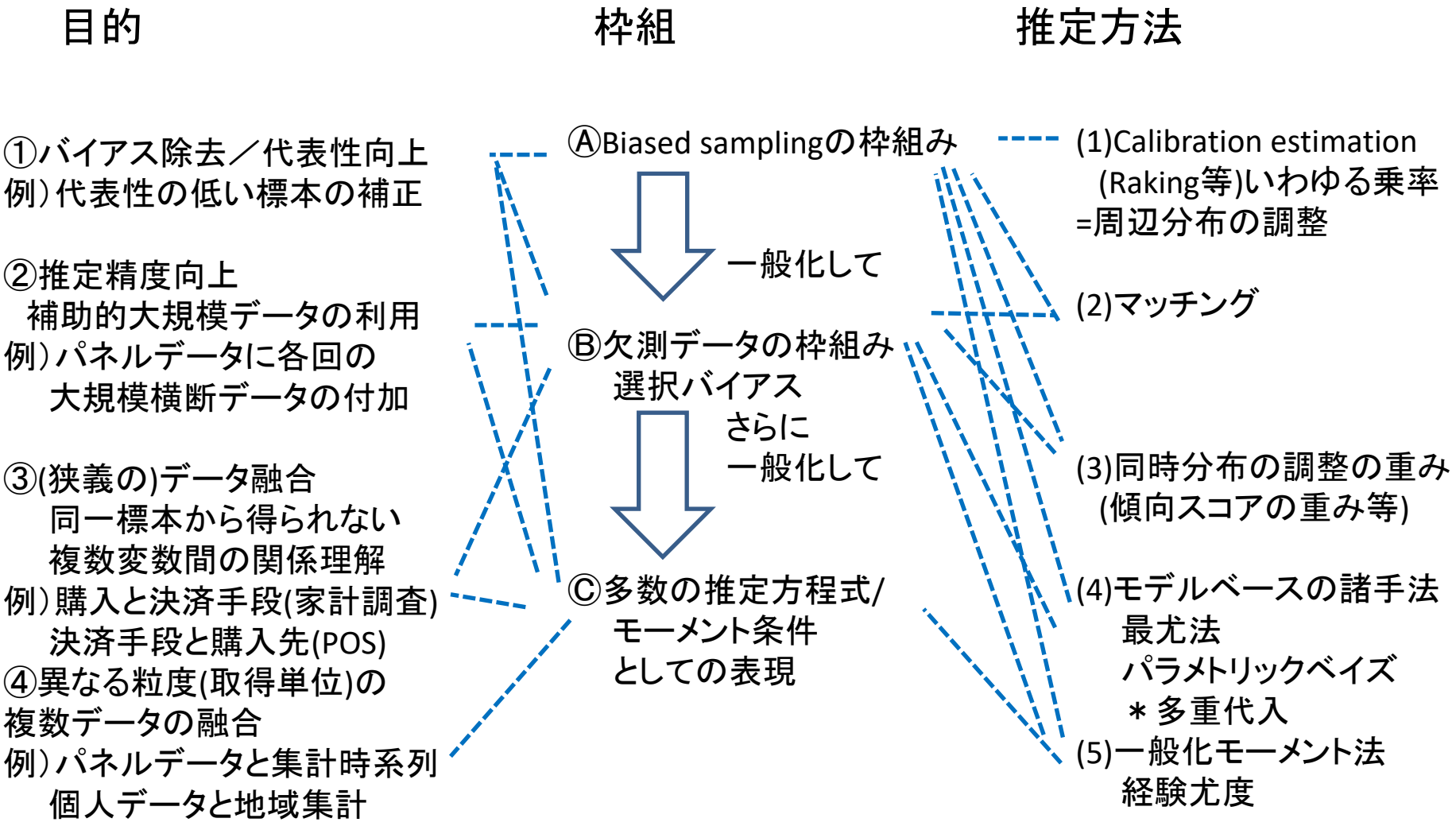
④AIと労働市場の研究

AIによって労働のどの部分が特に代替されるか、生産性向上の程度の理解

日本公認会計士協会との共同研究(2018-2022)

日本税理士会連合会との共同研究(2023-)

複数データを融合する目的/枠組み/手法



・時系列データや個人・家計レベルデータなど様々で一部バイアスのあるデータを統合し政策意思決定や経済活動を理解するための様々な分析法を開発
・データ間で変数やユニットが自動対応されるような機械学習的アプローチの開発

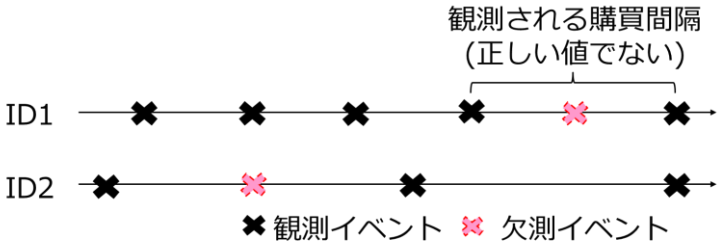
EBPM

政府統計等複数データの融合のための方法論開発

【外部周辺情報が利用可能である場合の母数推定についての理論開発】

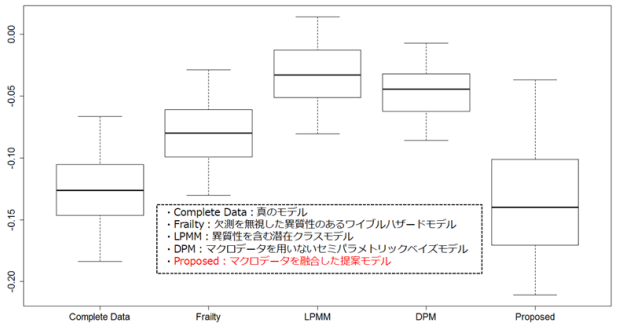
中間欠測の問題とマクロ情報を用いた解決

観測データのみから推定を適切に行うことは難しいことから、マクロデータを融合することで適切に母数を推定 (Hoshino&Igari,2018)



解析例：「価格」が購買間隔に与える影響バイアスを無視すると価格の影響を3分の1程度に過小評価

＊企業の意思決定や政府統計への応用へ



$$q(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}) \exp\left\{-\frac{n}{2} L_n(\boldsymbol{\theta})\right\} p(\boldsymbol{\theta})$$

尤度 = ミクロデータ部分

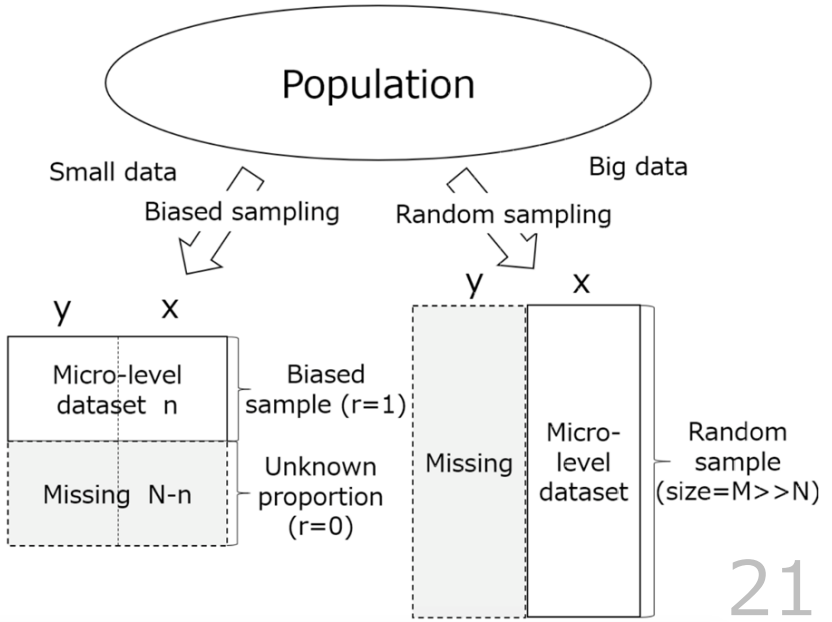
モーメント制約 = マクロデータの融合

事前分布

Biased sampling下の推測

元データ(共変量＋結果変数)が母集団からのランダムサンプルではないが共変量のみ観測されているサイズの大きいランダムサンプルが利用可能

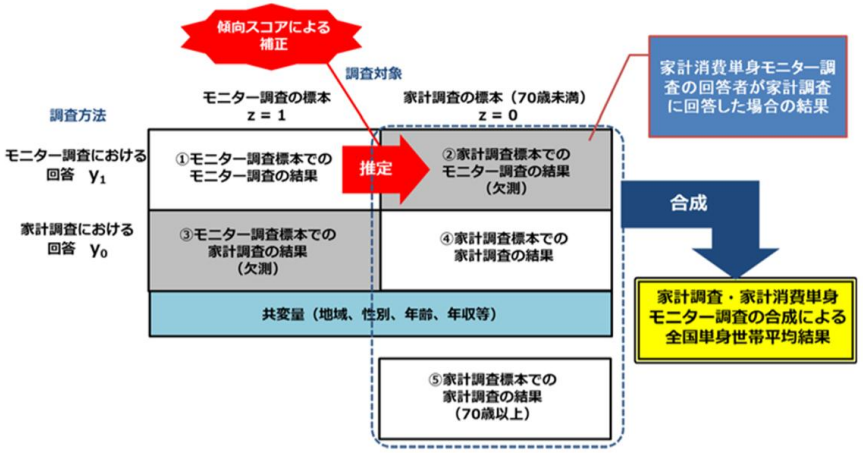
⇒それらを融合してbiasを補正する手法の開発 (Shimizu&Hoshno, 2019)



EBPM

総務省CTIミクロでのデータ融合

家計調査の単身世帯は700人弱
単身モニター調査と融合
2400人程度の単身モニター対象者の
偏りを傾向スコアで調整



消費動向指数（ＣＴＩ）の概要

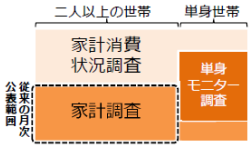
ビッグデータ等を活用し、消費動向をマクロ・ミクロの両面から捉える速報性の高い消費指標の体系：**消費動向指数（CTI：Consumption Trend Index）**を新たに開発し、

- 2018年１月分から**参考指標として公表開始**
- 2021年７月分公表時に、**2020年基準改定を実施**

世帯消費動向指数（ＣＴＩミクロ）
世帯の平均消費支出額（10大費目別、世帯類型別など）の月次動向を示す統計指標

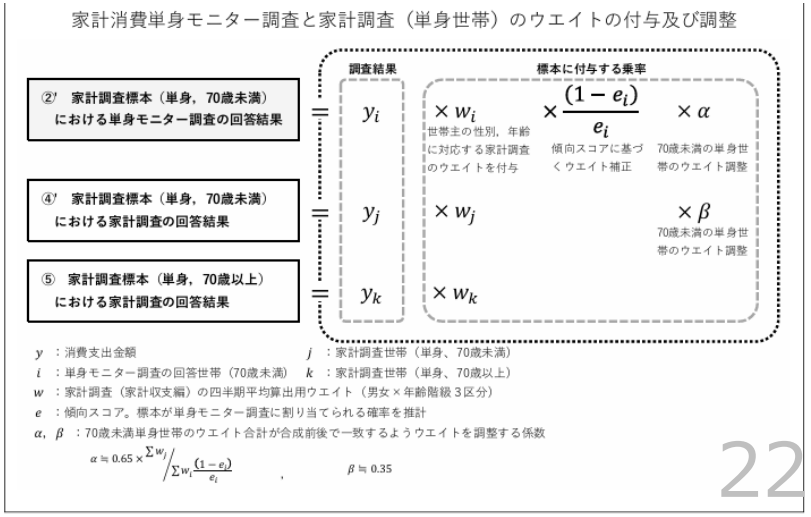
- 家計調査（標本規模：二人以上の世帯 約８千、単身世帯 約7百）の結果を、
  - 家計消費単身モニター調査（標本規模：2千4百）
  - 家計消費状況調査（標本規模：約３万）

  の結果等と統計的手法によって補正・補強し、標本規模を擬似的に拡大、**推計精度を向上**



総消費動向指数（ＣＴＩマクロ）
国内経済における個人消費総額（GDPにおける家計最終消費支出）の月次動向を示す統計指標

- GDP統計（家計最終消費支出）をターゲットとして、最新の動向を推測
- GDP統計の四半期別公表値では観測できない月次の値を時系列回帰モデルによって推計
- 2022年12月に、**ビッグデータ利活用の成果に関する報告書をウエブサイトに掲載**



総務省HPより抜粋
https://www.stat.go.jp/data/cti/pdf/micro_ref_2020.pdf
2020年基準 世帯消費動向指数（CTIミクロ）の推定方法（2024年１月分～）

（中略）

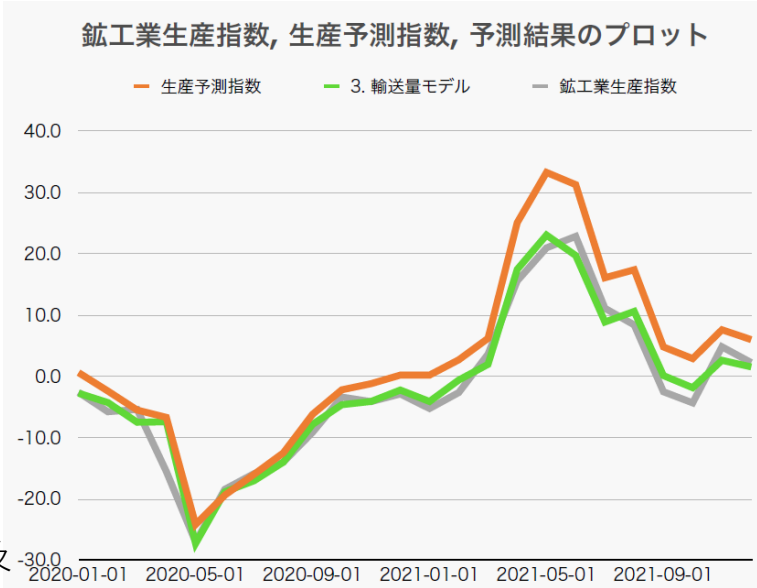
参考文献
[1] 星野崇宏(2005) 欠測群の周辺分布の母数に対する傾向スコアを用いた重み付きM推定量の提案と介入効果研究への応用、行動計量学、 32(2)、pp. 121-132.
[2] 星野崇宏(2009) 、『調査観察データの統計科学：因果推論・選択バイアス・データ融合』、岩波書店.
[3] 消費統計研究会（2017年度 第１回～第３回、2020年度 第１回、2021年度 第１回、2022年度 第１回、2023年度 第１回、第２回）資料
https://www.stat.go.jp/info/kenkyu/skenkyu/index.html

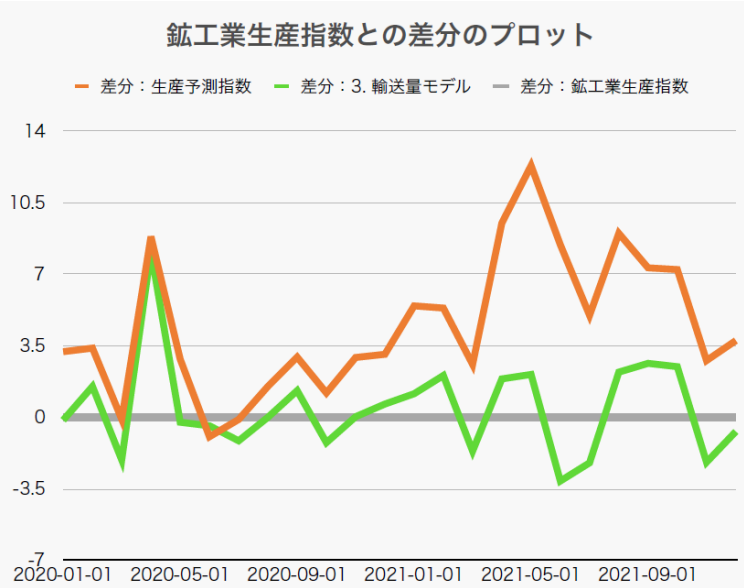
鉱工業生産指数の予測

携帯電話GPS（ブログウォッチャーにて取得）
　携帯電話アプリを通じて提供されるGPS位置情報（UID, 時間, 緯度, 経度, 他） アクティブ台数：約2,500万台
トラック走行履歴GPS（日野自動車にて取得）
　トラック走行時に提供されるGPS位置情報（UID, 時間, 緯度, 経度, 他）　 アクティブトラック数：50万台以上
対象地域：全国　対象期間：2018年1月～2021年12月
【説明力】
相関は0.88→0.96
RMSEは58.1%減少

【現在進行中】
都道府県別
×財別・都市圏別
など公表データの
無い系列の予測
OD(出発ー到着)
データ化
⇒産業連関の
時系列推移

経済インパクトの波及





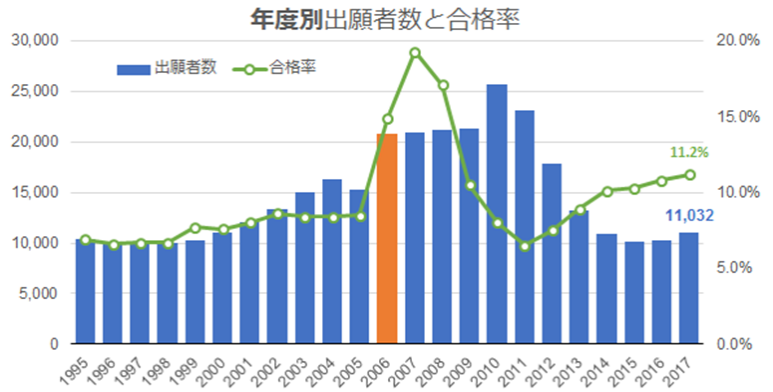
AIと職

【背景】AIが人間の労働を代替するとの議論(Frey&Osborne, 2013等)で労働市場に歪み

公認会計士試験受験者が急減する一方でコンプライアンス整備やコンサルの必要性から人手不足

【日本公認会計士協会との共同研究】

「会計士の業務がどの程度代替されるか」「AIが導入されることで生産性はどう変化するか」を研究



【方法】

（1）AI代替可能性評定：会計主査と補助者での業務を１０分類し各分類の代替可能性を評定(デルファイ法)

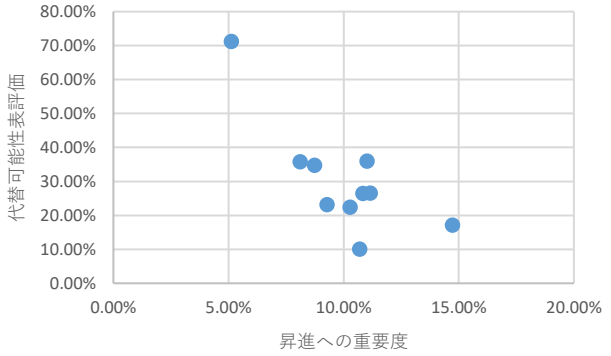
（2）生産性評価の調査：会計士協会が計画的に抽出した600人の会計士の給与情報、労働時間情報、上司による職階昇格条件の調査（コンジョイント分析）

【結果】

◘30年後も職務内容のほとんどの部分で先行研究より代替可能性確率は大幅に低い

◘一部の仕事をAIで代替することで40%ほど生産性向上の可能性

主査 （n=101）		
業務内容	代替可能性（10年後）	昇進への重要度
①クライアントとの調整	10.11%	10.70%
②監査チームのマネジメント	36.00%	11.02%
③監査契約時（新規締結・更新時）のリスク評価	35.78%	8.12%
④企業環境の理解及び監査リスクの評価	26.56%	11.16%
⑤適切な監査手続の立案と必要な修正	26.44%	10.84%
⑥定型的な監査手続の実施	71.22%	5.12%
⑦非定型な監査手続	22.44%	10.29%
⑧監査上の重要事項に係る検討及び判断	17.11%	14.73%
⑨監査調書の査閲と監査意見案の作成	23.22%	9.28%
⑩マネジメントレター案等の作成	34.78%	8.74%



科学技術と社会チーム(佐倉／福住)

AIを人間と社会の側から考える

社会や人間が新しい技術と安心して共存できるように、社会やそれを構成する人間や組織、さらに文化や地域の違いが新しい技術に何を求めていくのか？といった技術の社会的形成と技術の社会受容性について、《I. 人間的側面》と 《II. 社会・文化的側面》から考える。

Ⅰ. 人間的側面

①国際標準化された利用時品質モデルの開発とそのリスク評価方法［図 1］

②人間中心設計活動の書式(CIF: Common Industry Format)についてガイドライン

③ポストコロナ時代のウェルビーイング実現～コロナ前後での働き方変化の調査［図 2］

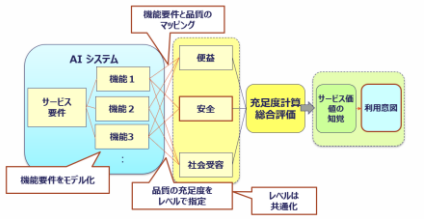


図 1　利用時品質を用いたリスク回避のモデル化

	n(Total)	全く変化ない	ほぼ変化ない	どちらとも…	少し変化した	大きく変化…
Total	177	9.6%	14.1%	26.0%	31.6%	18.6%

図 2　自宅内の環境や理解度の変化－在宅勤務／テレワークに対する家庭内の理解

Ⅱ. 社会・文化的側面

①テクノアニミズム論の実証調査～人とロボットの構図の日米差［図 3］

②テクノアニミズムを日本のアニミズム思想の文脈から見直す［図 4］

図3　人とロボットの並び方の文化差（左）、日本の構図には浮世絵との共通点（第三項共視）が見られる（右）

図4　テクノアニミズム論の元祖・梅棹忠夫（左）とアニミズム思想の中心人物・折口信夫（右）

■ 指標化

	ステークホルダ１（従業員）	**ステークホルダ２（管理者）**	**ステークホルダ３（経営者、組織）**	**ステークホルダ４（自治体）**
便益(B)	仕事のやりやすさ（主観評価）	仕事のフォーメーションの組みやすさ（主観評価）	生産性の向上（導入前後比等）	税収増、雇用増（導入前後比等）
安全(S)	ワークライフバランスの向上（主観評価）	チームメンバーの健康維持、安全確保（主観評価）	セキュリティレベルの向上（事故件数等）	事故減（事故件数等）
社会受容性(A)	コミュニケーション活性化（主観評価）	チームの一体化（主観評価）	導入によるブランドイメージ向上（市場調査）	モデル地域（総務省調査等）

便益向上度(B)＝f(SH(B)1, SH(B)2, SH(B)3, SH(B)4)

安全向上度(S)＝f(SH(S)1, SH (S) 2, SH (S) 3, SH (S) 4)

社会受容性向上度(A) ＝f(SH (A) 1, SH (A) 2, SH (A) 3, SH (A) 4)

主な発表成果

[1] Fukuzumi S, Kasamatsu K: Changing Work Style in Activity Based Working (ABW), AHFE2025

[2] Fukuzumi S: Quality-in-Use for AI, HCII2025

[3] Ogawa R, Sagawa, Y., Shima S, Takemura T, Fukuzumi S: A Case Study on AI System Evaluation from Users' Viewpoints, HCI International 2024.

[4] Fukuzumi, S. and Hirasawa, N.: HCD and software development process, AHFE2024.

[5] Umematsu, T. and Fukuzumi, S.: Applying a Quality-in-use Model to Health Conditions Estimation Systems Using Facial Videos, HCII2025

[6] 前田・翁・佐倉：人とAIの共生を考えるには「文化」が重要である、科学 242, 2024

[7] Sakura O: Robot and Ukiyo-e: implications to cultural varieties in human-robot relationships, AI & Society 37, 1563-73, 2021, doi.org/10.1007/s00146-021-01243-8

[8] Sakura O, Yuki M: Animism and Techno-animism in Japan: Their Roots and Modern Transformations, In: Global Perspectives on Animism and Autonomous Technologies, (Eds. Becker R, Luz Costa AC, Ventimiglia A) Springer, 2025

[9] 佐倉統：テクノロジーによる教育ではなく教育のためのテクノロジーを,国際教育夏季研究大会，京都科学大学, 2025年8月6日（招待講演＋パネル討論）

[10] 佐倉統：文化的アイデンティティとしての科学技術──健全な多元主義をめざして，文藻外語大学日本語学科国際シンポジウム～日本研究の独自性と学際性，文藻外語大学（台湾）, 2025年10月25日

イントロバージョン（約1分）

ダイジェストバージョン（約15分）

フルバージョン（約1時間）

社会におけるAI利活用と法制度チーム（中川）

- プライバシー保護のための匿名化技術タスク　PWSCUPの開催と参加（優勝１回）
- 内閣府：人間中心社会AI原則　（分担執筆）
- IEEE　Ethically Aligned Design (一部執筆)
- AIエージェントの法的人格付与の調査分析（情報ネットワークローレビューなど）
- AIエージェント（Agentic AIとも言う） を個人および企業において社会利用する場合の問題点，課題の調査提案および個人代理に使う方法の提案

AIの法的人格

◆「何かが足りない論」~AIに法的人格を与えるには，
魂，意識，意図，感情，善悪の感覚，自由意思
- Solum：AIがこれらを持つように振る舞うなら，それを否定するだけの具体的根拠がない以上「何かが足りない」論は成立しない

Lawrence B. Solum
1992

◆AIが本人のツールなら，AIの知識は自動的に本人の知識．しかし，AIの複雑さから，この前提はあやしい．だから，その点に関する議論を避けるためにAIに法的立場を一切与えてはならないという臭いものに蓋は良くない

Chopra & White
2011

◆法的実験などの実用的なアプローチ

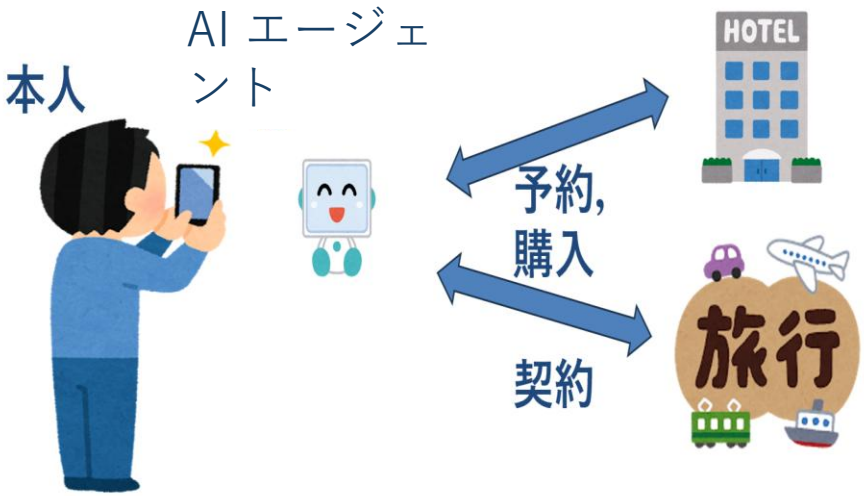
a. AI ロボットに完全な法的人格という仮説は避ける（EU 委員会　2018）
b. 説明責任と賠償責任の可能性に対する新形態の AI 責任を模索すべき．
c. 法的実験で新しい説明責任と賠償責任をオープンな環境でテスト．
既存領域におけるいくつかの法制度のポリシー (たとえば，2003 年以降の日本の特区制度) を拡大し，エビデンスを基礎にして，困難なケースに対して合理的かつ効率的な新しい法的代理形態を探す．

Ugo Pagallo
2018

AIエージェントのトラスト

◆ **AIエージェント ＝ AI＋実社会での行為.**

行為してもらうかどうかはトラストの強さによる



３種類の個人代理するAIエージェント

(1)　単なるツール

・利用者本人が完全にコントロールするツール
・利用者本人は事細かに命令
・AIエージェント といえども，その行為は利用者の責任.
　損失が起きても利用者が被る

(2) 保険付きツール

・ AIエージェントが自律的に行った行為の結果,
本人の意図に沿わない不利益，ないし本人
にとって損失が起きる，あるいはAIエージェントの
行為が相手側に損害を与える

・ AIエージェントの行為に保険をかけ，本人が保険料を払う
保険料と免責額という保険ポリシーは，本人が選べる

✓ただし，本人の意図に沿わない場合，本人の依頼のし方が不十分だったのか，AIエージェントのソフトが，本人の依頼した自然言語の理解能力が不十分だったのか.

これを争うのはなかなか難しい問題

(3) 法人AI

・AIエージェントが資産を持ち，部分的にせよ法人格を持っていたとすると，損失はAIエージェントが手持ちの資産で支払い，本人には損失が及ばないようにできる．本人への支払いは迅速化できる

・資産額が大きければ， AIエージェントの信用度，具体的には許容範囲が上がる．

・法人AI自身が保険に入ることもある．この場合も保険ポリシーを現実の場面で解釈する問題は残る．

・高い許容度を持つトラストで関係つけられた自然人と法人格を持つAIエージェントが共生するエコシステムが構成できると，本人にとっては時間や行動の自由度が増すという大きなメリット．

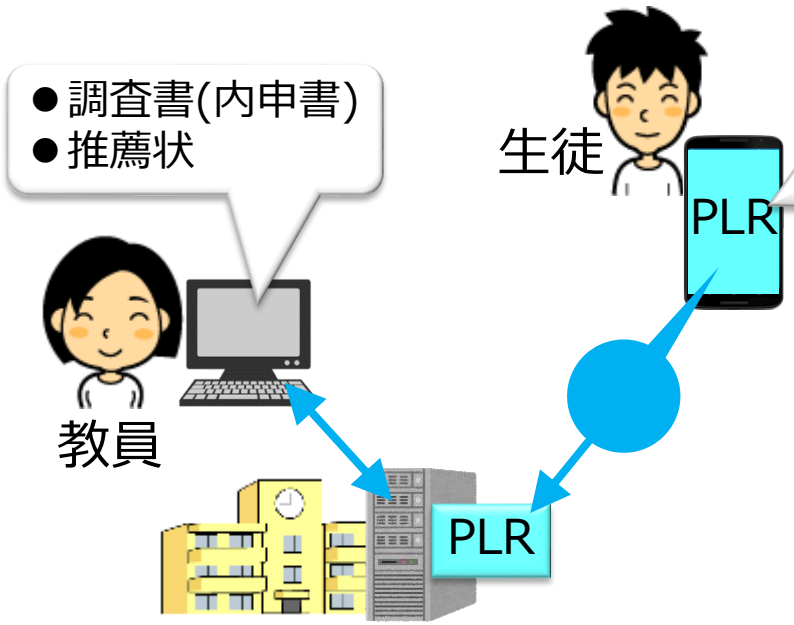
分散型ビッグデータチーム(橋田)

- パーソナルデータの分散管理
  - パーソナルデータを本人に集約してフル活用
    - パーソナルデータを本人が私的目的に利用することは無制限
  - PLR (分散パーソナルデータストア)を教育用に実運用
- パーソナルAI
  - 各個人に専属のAIエージェント
  - 10年で確実に普及 → AIの総合的ガバナンス
- AIの総合的ガバナンスの標準化
  - ログデータを利用者に集約することで利用者とサードパーティがAIのリスク管理に関与
  - 欧州AI法の整合標準にその旨を記述
- グラフ文書: 文書としての知識グラフ
  - テキスト文書より作成効率が高く他の方法より批判思考力を高める効果が大きい
  - AIを使っても批判思考力を高める効果が保たれる
  - 知識グラフはAI活用の基盤…人間にもAIにも理解できるデータ

パーソナルデータの分散管理

GDPR
2000個問題解消

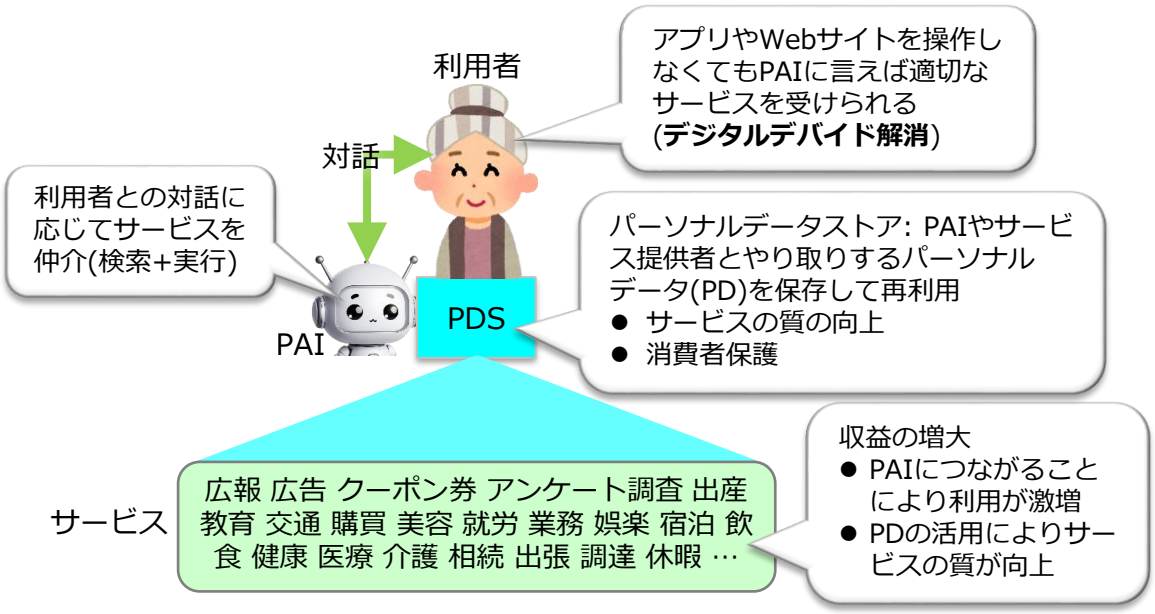
- PLR (Personal Life Repository)
  - パーソナルデータを本人の管理下に集約してフル活用
- 分散eポートフォリオ
  - 埼玉県教育局が2020~2023年度に実運用
  - 生徒がPLRアプリで入力した課外活動のデータを教員が調査書や推薦状の作成に活用

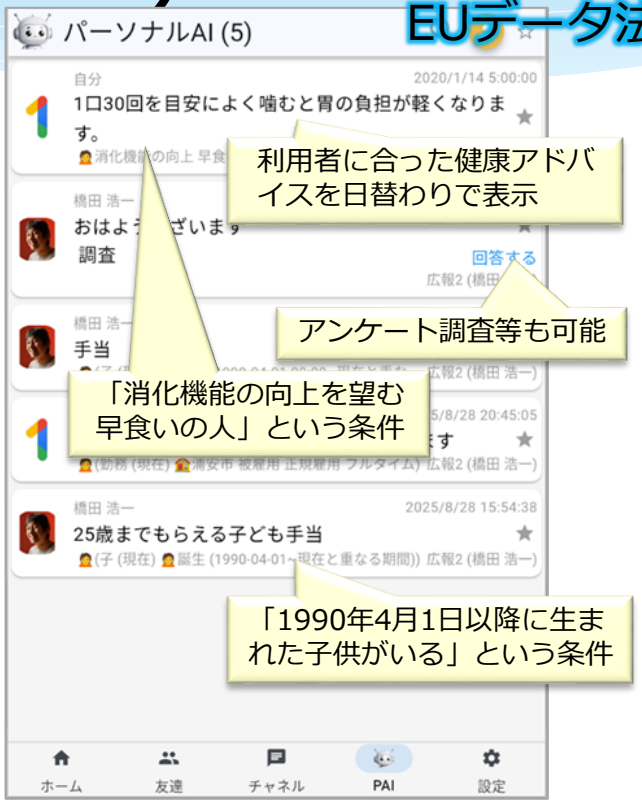


「埼玉県eポートフォリオ」チャネル

パーソナルAI (PAI)

AIエージェント
GDPR
EUデータ法

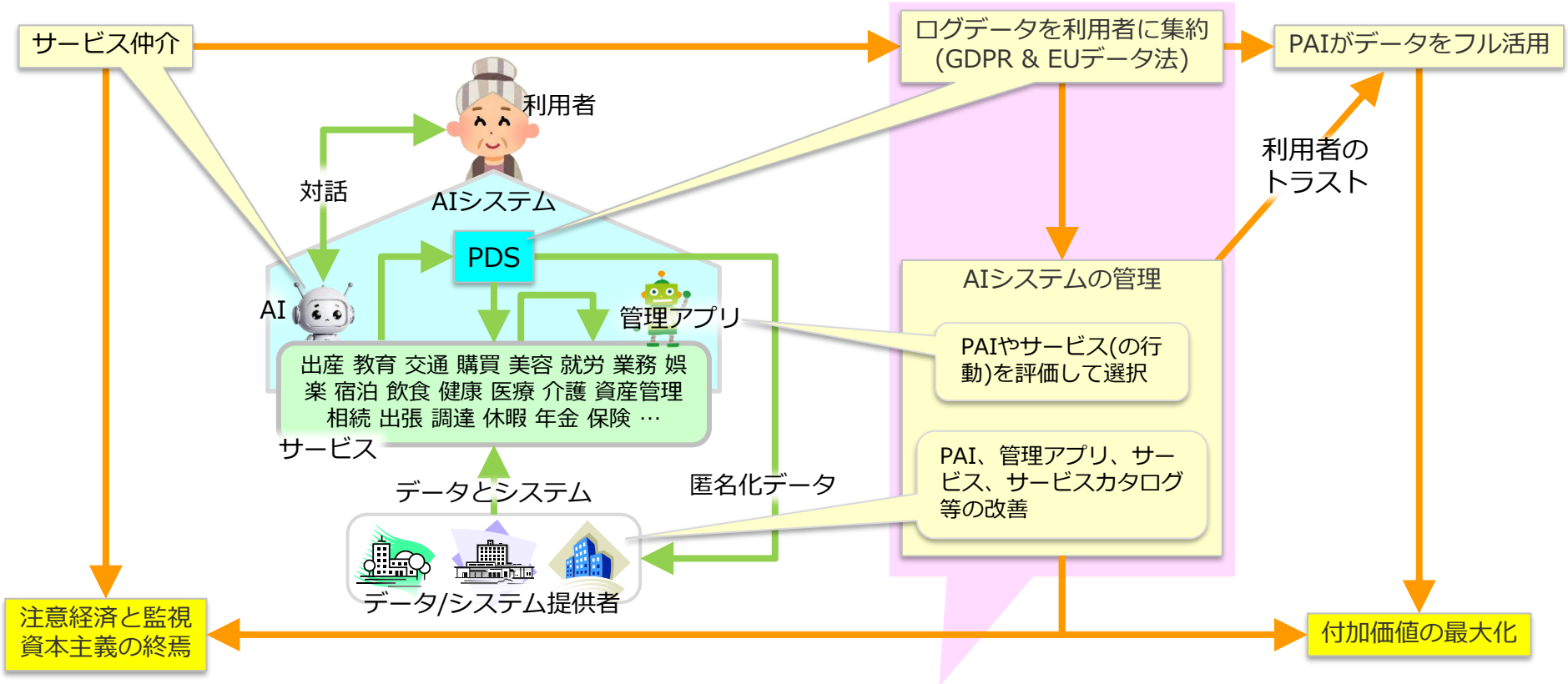




- 各利用者に専属するAIエージェント(AIA)
- 確実に普及
  - AIAはサービス利用を圧倒的に楽にし提供事業者の収益も大きいので普及は不可避
  - GDPRと欧州データ法とブリュッセル効果により世界中でAIAはPAIでなければならない

AIの総合的ガバナンスの標準化

- AIシステムの開発者や提供者だけでなく利用者やサードパーティが関与する総合的ガバナンス
- EU AI法の整合標準となる予定のISO/IEC 24970 “AI system logging”による義務付け



ISO/IEC 24970 “AI system logging”

グラフ文書: 文書としての知識グラフ

- グラフ文書はテキスト文書より作成の効率が高い
- グラフ文書の作成は批判的思考力を高める効果が他の学習法より大きい
  - AIを使ってもその効果が保たれる
  - 教育に限らないさまざまな文書作成においてAIによる認知的オフローディングを防ぐ
- 知識グラフはAI活用の基盤
  - 人間にもAIにも理解できるデータ