参考資料２-４
情報委員会（第35回）
令和６年１月25日

理化学研究所
革新知能統合研究(AIP)センター

社会における人工知能研究
グループの取り組み

グループディレクター　橋田 浩一

グループの構成

倫理とガバナンス

科学技術
と社会T
佐倉 統

感情

人間性の再定義

科学技術社会論

市民科学

人権

社会におけ
るAI利活用
と法制度T
中川 裕志

安全性と利便性

AI安全性・
信頼性U
荒井 ひろみ

受容性

サービス

構造化文書

説明可能性

プライバシ

分散型
ビッグ
データT
橋田 浩一

パーソナルAI
エージェント

制度

ナッジ

セキュリティ

AIセキュリ
ティ・プラ
イバシーT
佐久間 淳

機械学習

経済

統計

分析と介入

経済経営
情報融合
分析T
星野 崇宏

グループのミッション

AIそのものの研究開発ではなく、

- AIと社会との関係の解明と改善
- AIの開発・導入・運用の社会基盤

研究テーマ

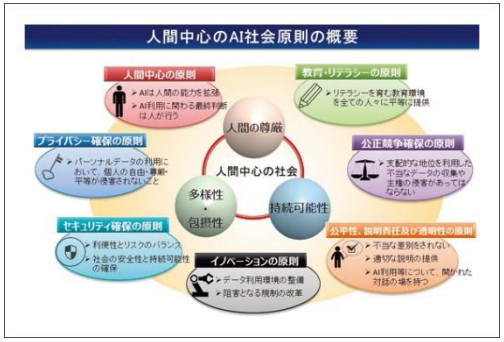
- 倫理とガバナンス
  - 人間・社会とAIとの共進化の可能性と要件
- 安全性と利便性
  - セキュリティとプライバシ
  - 説明可能性と社会的公正
- 分析と介入
  - 社会の分析
  - 実証実験と実運用
    - データの生成・共有・活用

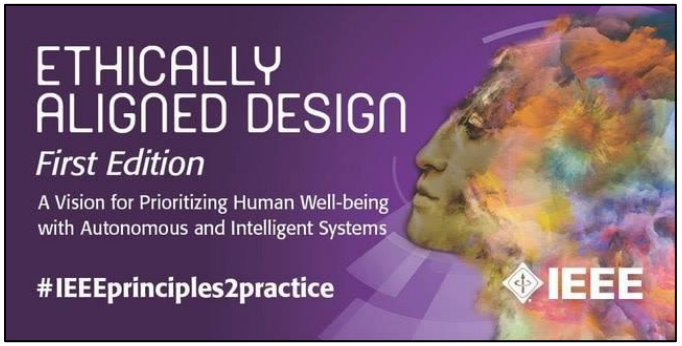
代表的な研究成果

AI倫理指針策定への貢献

社会におけるAI利活用と法制度T

- 人工知能学会倫理指針 (2017)
- 総務省AIネットワーク社会推進委員会
  - AI開発ガイドライン：OECDに提案 (2017)
  - AI利活用ガイドライン (2019)
- 内閣府人間中心のＡＩ社会原則 (2019)
  - AI ready な社会の在り方：G20に提案
- IEEE Ethically Aligned Design (2019)
  - 世界最大規模の電気電子学会の倫理指針



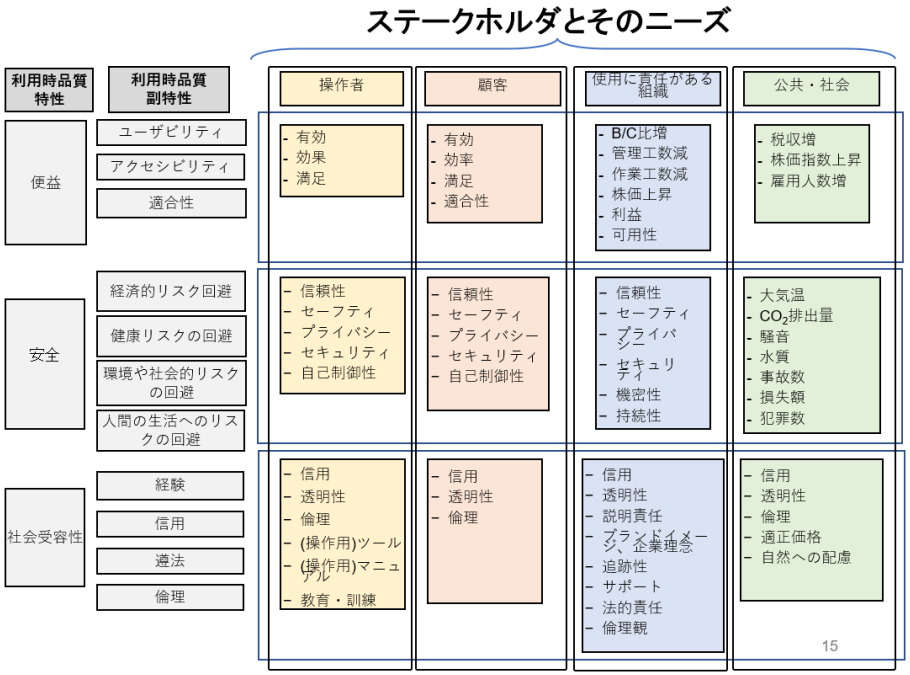


AIの「ユーザビリティ」を良くする

科学技術と社会T

- 技術の社会受容性評価のための利用時品質モデルについて、ステークホルダニーズの詳細化（右図）及び複数システムを対象とした有効性検証とAI製品へ適用し、妥当性を検証
  【成果】国際標準化[1]
- ユーザビリティ向上のための人間中心設計活動の書式(CIF: Common Industry Format)を実開発に適用。課題を明示し、書式へフィードバック。また、顧客視点での未来志向インタラクションとして、RIAS (Robot Intelligence and Autonomous System)を中心に課題を整理
  【成果】CXの観点と合わせて書籍化[2]。CIFはドイツのコンサル会社と競合しているが、書式開発で共同議長を分担。

2021年度「産業標準化事業表彰 経済産業大臣表彰」を受賞。Society5.0における「人間中心の社会」の実現に対して大きく貢献できると期待。



[1] ISO/IEC DIS25019 (2023年発行予定)

[2] 平沢、福住（編著）：顧客経験を指向するインタラクション – 自律システムの社会実装に向けた人間工学国際標準 –. 日本経済評論社(2023年3月発刊予定)

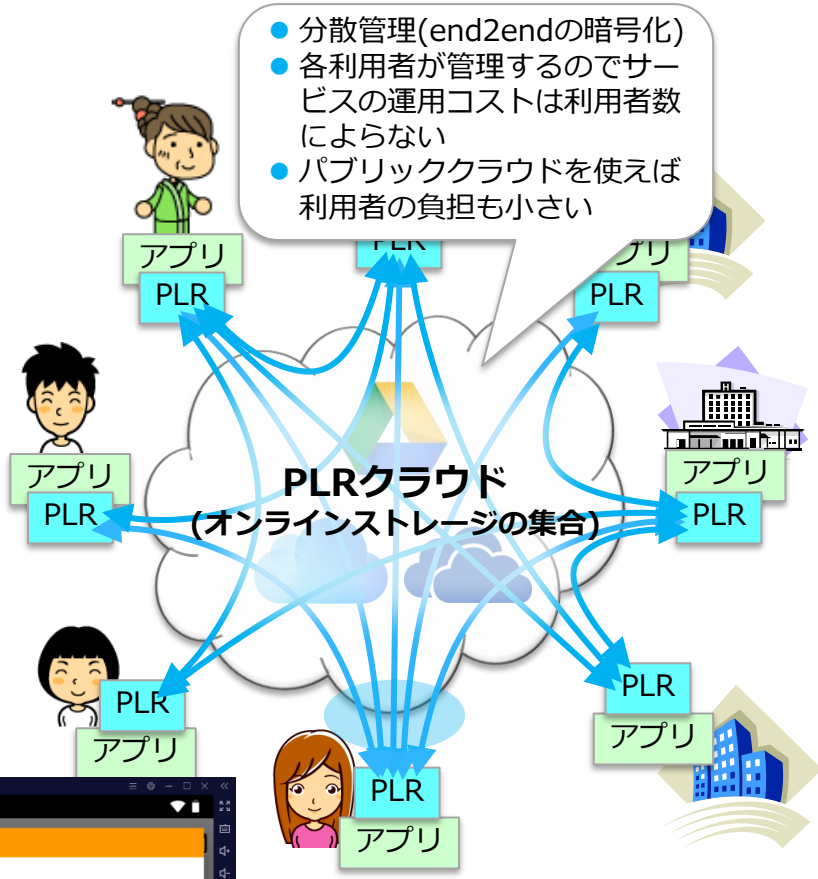
[その他] 福住、西山、梶谷、北村：事例で学ぶ 人を扱う工学研究の倫理. 近代科学社(2023年1月発刊)

パーソナルデータ(PD)の分散管理

分散型ビッグデータT

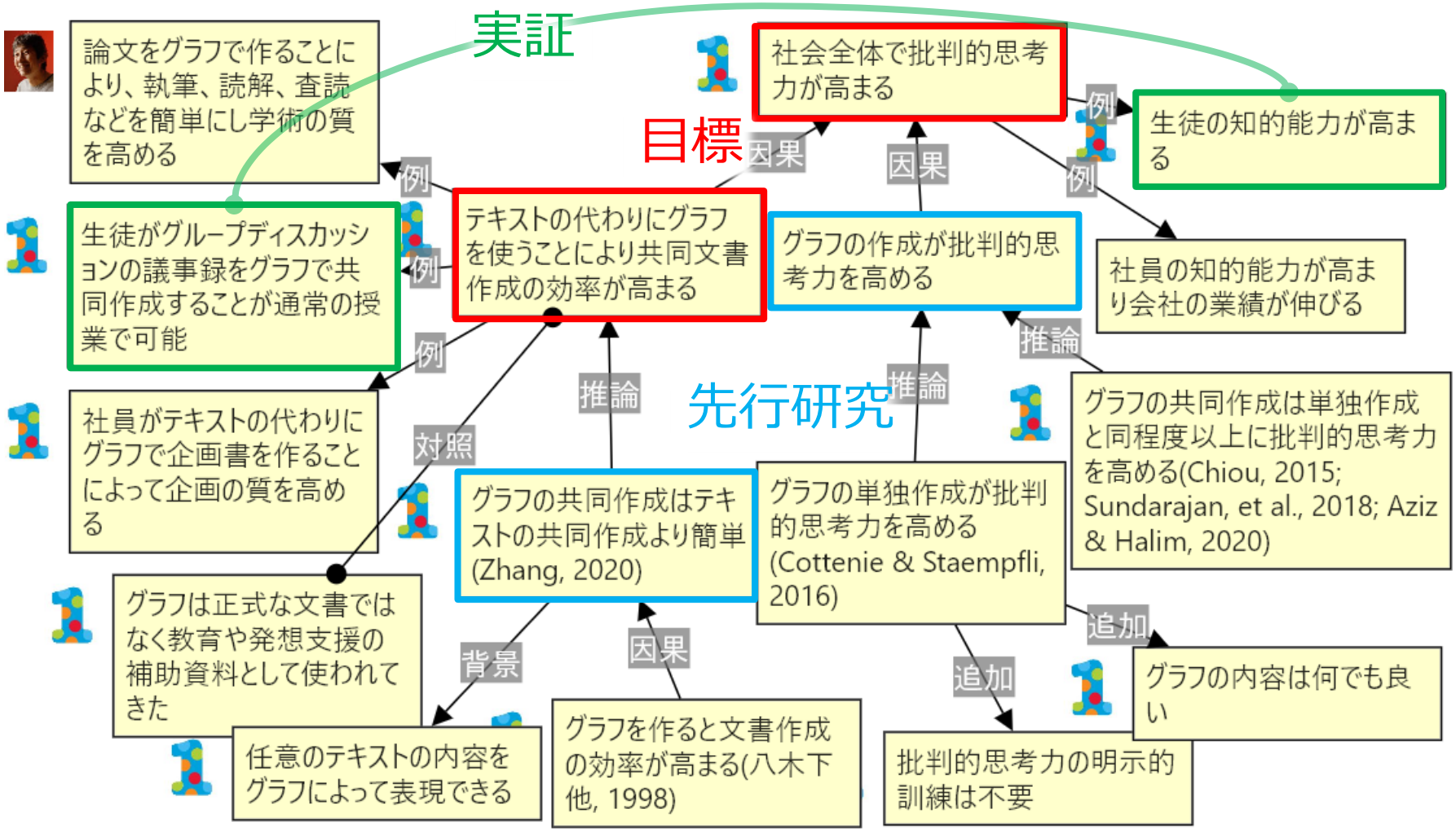
PDの分散管理(本人管理)に基づくサービスの社会実装

- PLR (Personal Life Repository)
  - ソフトウェアライブラリ
  - PDを本人が名寄せしてフル活用
    - 本人のための活用 + 機械学習など
  - **数十億の利用者にアプリの保守コストだけでサービス**
  - **人間の共同作業のほとんど(業務システムやSNS)が可能**
- 電子調査書@埼玉県
  - 2020年秋から実運用中
  - 県立高校の生徒がPLRアプリで作成・管理する課外活動の記録を県が運用する校務支援システムに連携して教員が調査書や推薦状の作成に利用
- 電子母子手帳@市立伊丹病院
  - 2023年実証実験→実運用
  - 産科と小児科のデータを母親に提供
  - 看護師による健康相談
  - 個人アプリで予防接種等を通知
- 乳幼児健診の電子化@熊本県荒尾市
  - 2023年実証実験→実運用
  - 妊娠届出～3歳児健診
    - **全国展開可能**
  - **他の行政手続に拡張予定**



グラフ文書

- テキスト操作の認知負荷を軽減して意味内容に関する思考を促進
  - 文書作成・読解の効率と批判的思考力を高める



グラフ文書の教育効果に関する実験

分散型ビッグデータT

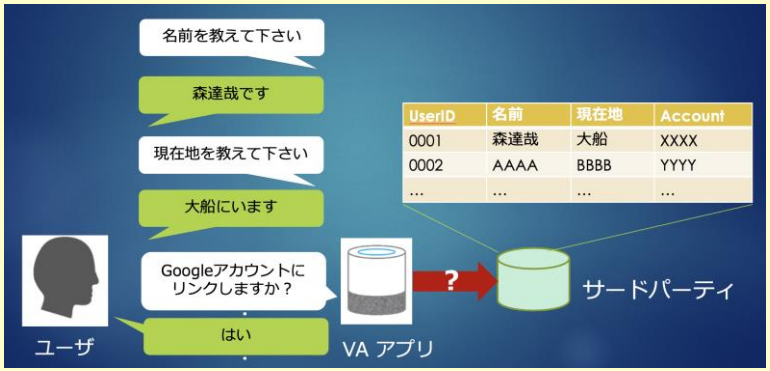
方法	● 2022-10/2023-01、2つの高校の1年生6クラス約100人 ● 批判的思考力試験(CT試験) → グラフ授業5回 → CT試験 ● グラフ授業: 「現代の国語」でグループディスカッションの内容のグラフ文書を各グループの生徒が共著して教員と他の生徒がコメント
結果	● 余分なコストも支障もなくグラフ文書を授業に導入可能 ➢ 教員の負担はグラフ文書の導入で増えない ➢ 生徒は授業が成立する程度にグラフ文書を作れる ● それで生徒の批判的思考力が高まる ➢ グラフ操作(ノードとリンクの作成・編集など)の量とCT試験の成績向上の間の相関係数は0.30、**相関がある確率は99.73%** ➢ グラフ授業を増やせば相関が高まるはず
結論	高校教育へのグラフ文書の導入は現実的に可能で教育効果を高める

- 一般業務でもグラフを使えば文書処理の生産性と勤労者の批判的思考力が向上して組織の業績が高まるはず
  - ◆ 社会全体にわたりテキストをグラフで置換

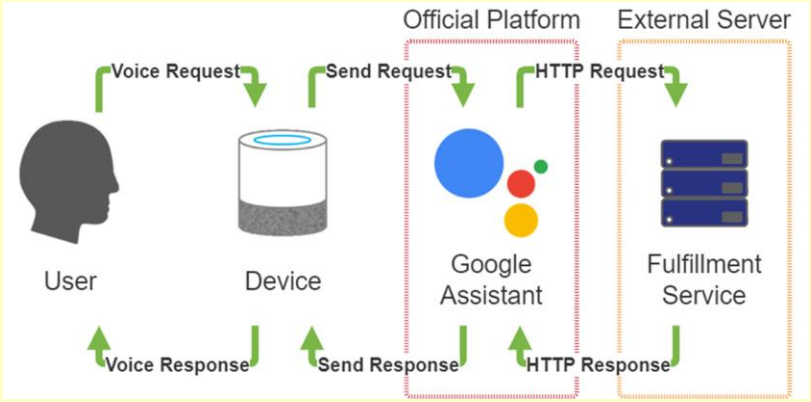
音声アシスタント(VA)アプリによる個人情報収集

RQ: 音声アシスタント(VA)はどの程度個人情報を収集しているか？

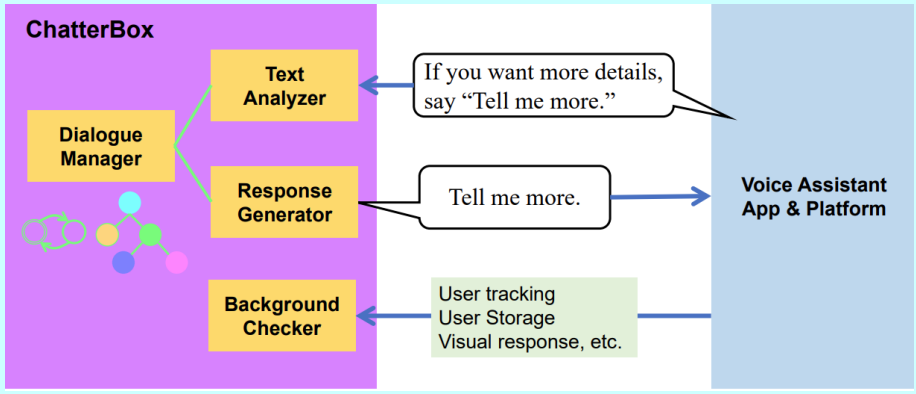
- VAアプリはローカルで実行できない
  - 対話を通じてしか機能が明らかにできない
  - プライバシーリスク不明



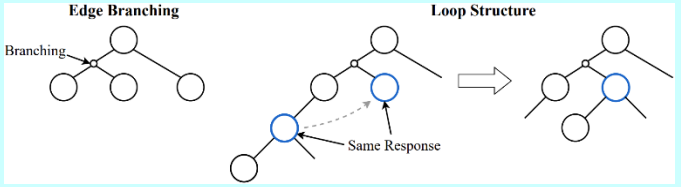
クラウド（見えないところ）で実行される
→ユーザに対する**透明性に欠ける**



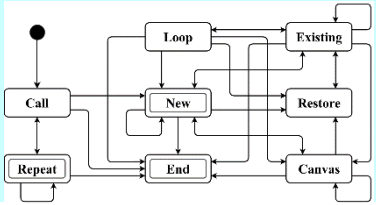
アプローチ: 自然言語処理を利用した対話生成・解析＋
アプリレベル通信解析によるプライバシリスク評価



過去の対話履歴→ツリー構造データ



対話の状態→状態遷移図



Atsuko Natatsuka, Ryo Iijima, Takuya Watanabe, Mitsuaki Akiyama, Tetsuya Sakai, Tatsuya Mori: Understanding the Behavior Transparency of Voice Assistant Applications Using the ChatterBox Framework. RAID 2022: 143-159

音声アシスタント(VA)アプリによる個人情報収集

結果 (VAアプリ en: 9,732、ja: 931)

対話で個人情報を収集

個人情報	アプリ数(en)	アプリ数(ja)
名前	9	2
メールアドレス	6	0
電話番号	7	0
居住住所	3	12
現在位置	0	3
年齢	2	3
性別	0	1
血液型	1	0
誕生日	1	8
勤務地	1	1
通勤経路（駅）	0	2
いずれか1つ以上当てはまる	22 (3.0%)	28 (6.0%)

3～6%は対話を通じて個人情報を収集

ユーザストレージを利用

データ	アプリ数(en)	アプリ数(ja)
ユーザID	133 (18.2%)	42 (8.99%)
最終利用時刻	4 (0.546%)	7 (1.50%)
アプリ利用回数	6 (0.820%)	10 (2.14%)
ユーザの現在位置	1 (0.137%)	0 (0%)
その他	59 (8.06%)	45 (9.64%)
上記のいずれか1つ	160 (21.9%)	79 (16.9%)

秘密裏に情報収集可能（潜在リスク）

プライバシポリシ分析

分類	アプリ数(en)	アプリ数(ja)
妥当	324 (44.3%)	18 (3.85%)
不足あり	367 (50.1%)	431 (92.3%)
掲示なし	41 (5.60%)	18 (3.85%)

- 4～6%はプライバシポリシなし
- あっても76～94%は記述が不十分

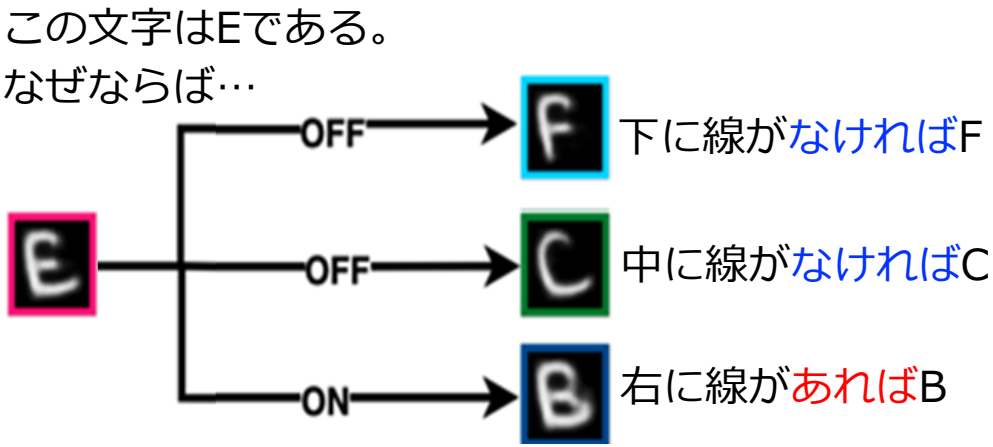
本研究の貢献

- **VAアプリの解析フレームワークを提案**
  - 自然言語処理による対話＋アプリレベル通信解析
  - 日本語＋英語で動作することを実証
  - 自然言語を用いた対話インタフェースを持つシステムのテストに適用可能
  - ※チューニングなしのGPT-3では対話は成立せず
- **VAアプリの実態を解明**
  - スマートフォン等と比べて個人情報の扱いが発展途上
  - 適切なユーザIFの開発が必要

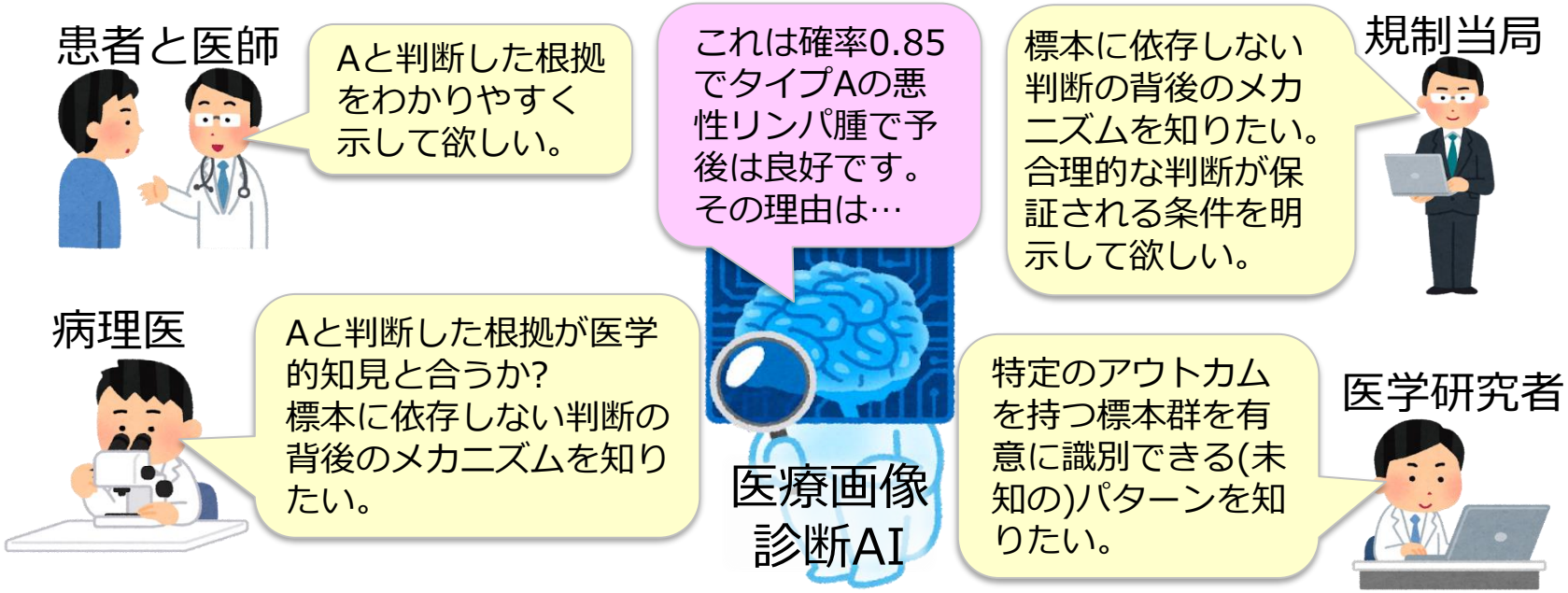
VAEを用いた教師なし因果バイナリ概念の発見

Tran, et al. AAAI2022

- 画像分類の結果をAIが発見した記号的概念で説明



悪性リンパ腫のAI病理診断(データ駆動型生物医科学T・久留米大医学部等と推進中)



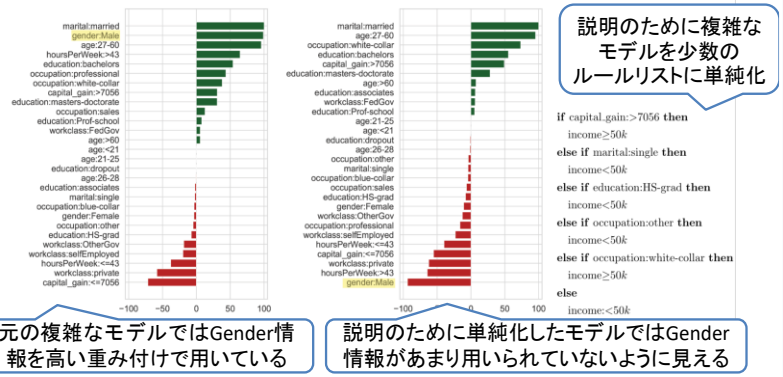
社会におけるAI利用の問題点改善や安全性向上

人工知能安全性・信頼性U

社会におけるAI利用や、AIのためのパーソナルデータ利用の際の潜在的なリスクの指摘、ネット上の有害情報の分析を通じ、問題点の改善や安全性を向上

AIの説明におけるリスク

複雑なAIを単純化して説明する場合に，説明者が不公平なAIを実際より公平なAIであるように説明するリスクが存在することを，実際に説明を生成する方法を提案して指摘．その検出の困難さについて分析(UQAM, 阪大などとの共同研究，ICML2019, NeurIPS2021)



パーソナルデータ利用

パーソナルデータ利用の健全な同意のために，日本語のプライバシーポリシーの記述の適切さを分析．文脈整合性のフレームワークを用い情報の流れの記述が適切か評価(CSS2020優秀論文賞)

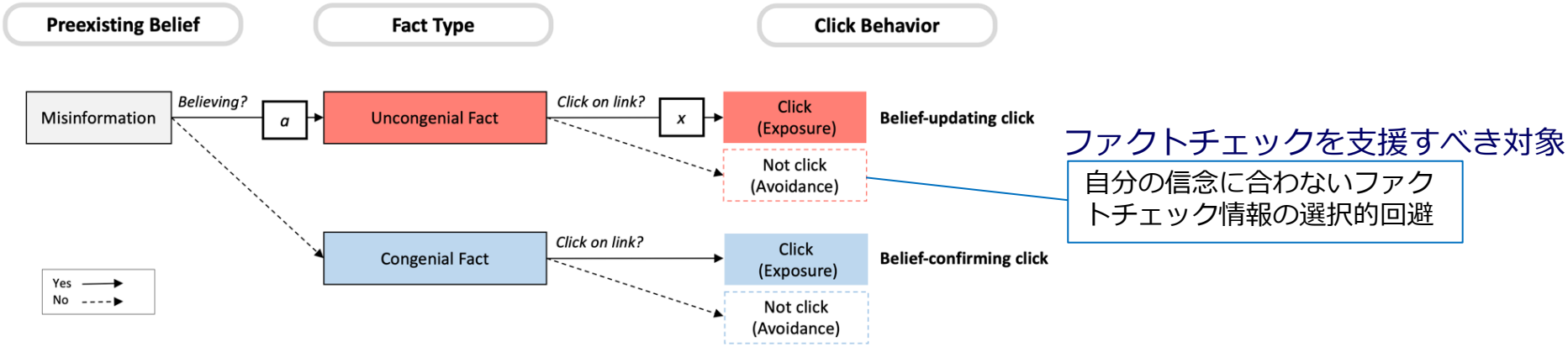
ネット上の有害情報の分析

ソーシャルメディアにおける日本語のヘイトスピーチに関するデータセットの試案を作成．排外主義的な攻撃的発言の分類や事例について考察　(南山大，東大らとの共同研究，NLP2020，岩波「思想」2021年9月号)

ファクトチェック情報のクリック行動の分析

目的：ユーザのファクトチェック行動の支援
ファクトチェック情報の共有において支障となる心理的要因を検証し、より多くの人がファクトチェックの恩恵を受けられる介入方法につなげる

ファクトチェックサイトでのクリック行動



この研究の貢献
(i) 選択的回避を測定するための新しい指標FAEIの提案
(ii) 人を対象としたオンライン実験による選択的回避を予測する心理的特性の調査
(iii) ユーザが偽情報の信憑性を確認することを容易にする将来の設計にユーザの理解を組み込む方法

Tanaka, Inuzuka, Arai, Takahashi, Kukita, & Inui: Who Does Not Benefit from Fact-Checking Website? CHI2023.

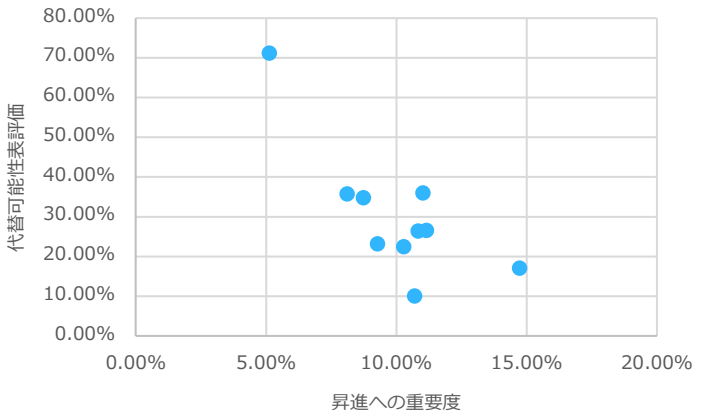
AIの職業への影響

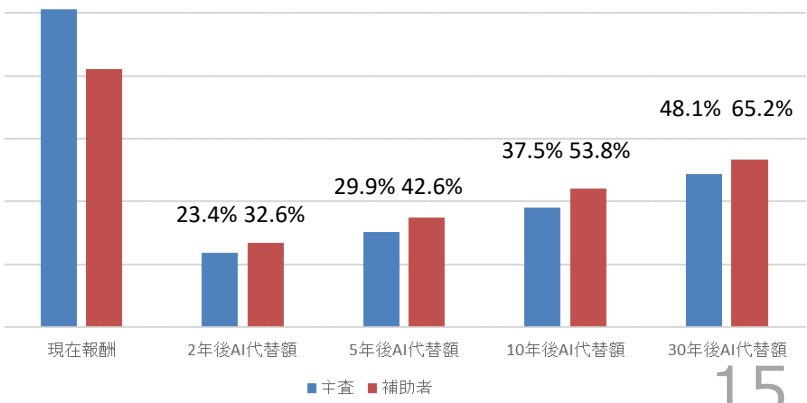
経済経営情報融合分析T

日本公認会計士協会との共同研究

- 背景: AIが人間の労働を代替するかどうか不明
  - Frey & Osborne (2013)への批判
  - 不正確な予測が労働市場をゆがめる
    - 公認会計士試験受験者が激減して人手不足に
- 方法: 公認会計士業務へのAIの影響を調査
  - AI代替可能性の評定：会計主査と補助者の業務を10分類し各分類の代替可能性を評定(デルファイ法)
  - 生産性評価のための調査：会計士協会が計画的に抽出した600人の会計士の給与、労働時間、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年後もほとんどの職務の代替可能性はFrey & Osborneの予想(>90%)より大幅に低い
    - Frey & Osborneは職務内容の詳細に立ち入らず
  - 代替可能性の低い業務ほど人事上の評価が高い
    - クライアントとの調整が最重要
  - 代替可能性が高い業務も補助者の一部の仕事をAIで代替することで生産性が約40%向上する可能性

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%

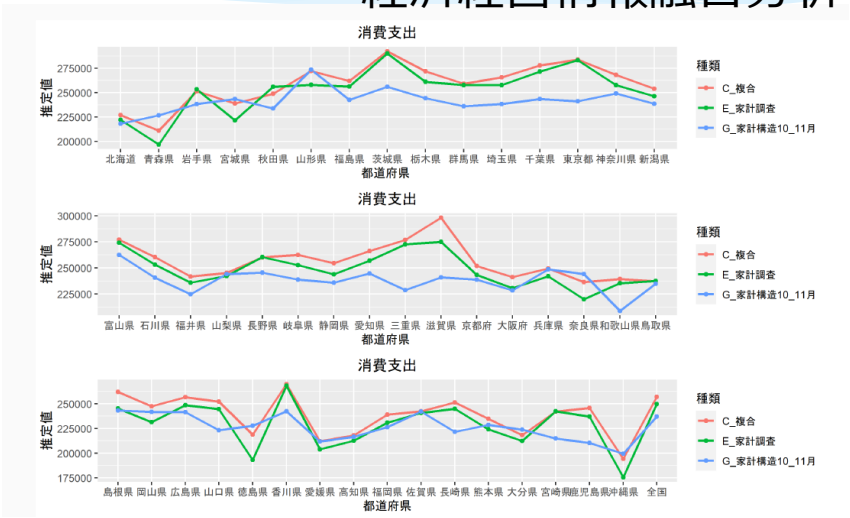




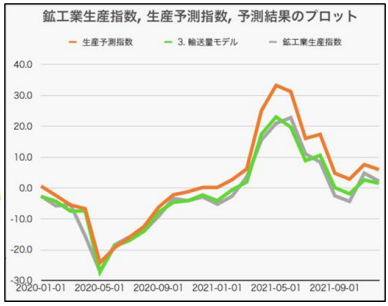
ビッグデータによる行政と経済の支援

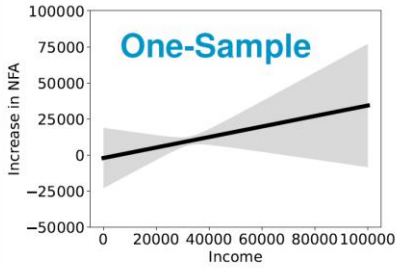
経済経営情報融合分析T

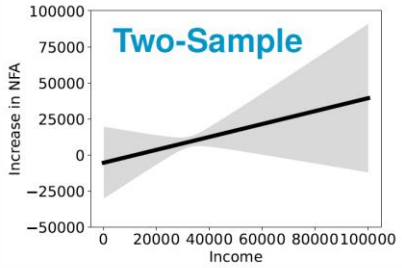
- 政府統計の改善(総務省統計局と共同研究)
  - 家計調査と家計構造調査をデータ融合し、基幹統計である家計構造統計での年次集計法を開発⇒2023年3月に年次集計を公開
  - 家計調査のバイアス補正と家計構造調査の季節性の調整
- 政府EBPMへの助言
  - 星野が2022年1～6月に内閣官房行政改革推進会議アジャイル型政策形成・評価の在り方に関するWG構成員（デジタル大臣・副大臣等臨席）
  - 6月のとりまとめにデータ融合の必要性が入る
  - 経産省・中企庁・内閣府のEBPM委員会の委員なども
- ビッグデータによる経済状況ナウキャスティング
  - 日野自動車・いすゞ自動車の50万台のトラックデータ
  - セブン銀行の2万台のATMデータからGDPや鉱工業生産指数を予測
- データ融合手法の開発
  - 比で表される量に関するDebiased/Double Machine Learningで推定精度が飛躍的に向上
  - 右図two-sampleのデータ融合でも推定精度がほぼ変わらず



2020年Q2まで学習しQ3以降を予測



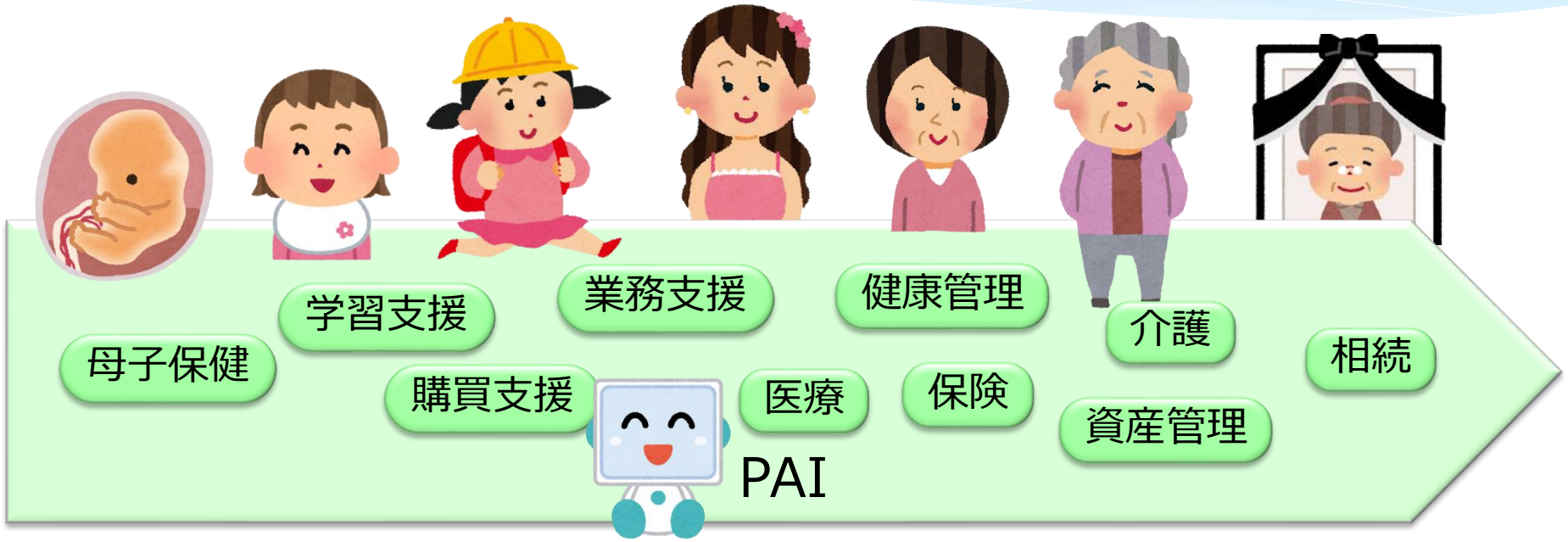




- 傾きは OS では 0.36，TS では 0.45．
- OS では$24000-$68000，TS では$26000-$61000 で有意な正の効果．

展望

社会Gの今後の展望　パーソナルAI (PAI)



- 各個人に専属してパーソナルデータの管理を代行
- 利用者との対話に応じて情報機器やソフトウェアを操作
  - 人間はアプリやWebサイトを使わない → デジタルデバイドの解消
- 多様なサービスの顧客接点を総取りしパーソナルデータをフル活用
  - 既存のデジタルサービスをはるかに凌ぐ潜在的付加価値
- 付加価値を最大化するには利用者のトラストが必須

社会Gの今後の展望

- 人-AIシステムのユーザビリティの国際標準化
- 人–AI共生社会の社会思想基盤の構築
- AIが職に与える社会経済的影響の評価
- AIへの攻撃に対する防御が成功する条件を人間に理解させる技術の開発
- XAIによる説明の方法と効果やリスクの検討
- プライバシーポリシーの可読性の向上

パーソナルAI (PAI)

- 注意経済の終焉へ
  - PAI提供者は利用者のトラストの獲得を競う
  - 偽情報やエコーチェンバーをPAIで抑止
  - PAIで消費者を保護
- 特許の非独占許諾でPAIの普及を促進
- PAIによる代理の技術的・社会的問題の解明
- PAIのUIの研究

PAIを軸にテーマ間の連携を図る