2020年度

AIPシンポジウム 成果報告会

RIKEN Center for Advanced Intelligence Project

社会における人工知能研究グループの活動紹介

橋田 浩一

社会における人工知能研究グループ

グループディレクター

グループの構成

科学技術
と社会T
佐倉 統

AI倫理・
社会T
鈴木 晶子

感情

人間性の再定義

科学技術社会論

倫理とガバナンス

市民科学

人権

制度

AI安全性・
信頼性U
荒井 ひろみ

受容性

サービス

パーソナルAI
エージェント

構造化文書

説明可能性

社会におけ
るAI利活用
と法制度T
中川 裕志

分散型
ビッグ
データT
橋田 浩一

プライバシ

安全性と利便性

ナッジ

セキュリティ

経済

AIセキュリ
ティ・プラ
イバシーT
佐久間 淳

機械学習

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

グループの構成

科学技術
と社会T
佐倉 統

AI倫理・
社会T
鈴木 晶子

倫理とガバナンス

AI安全性・
信頼性U
荒井 ひろみ

社会におけ
るAI利活用
と法制度T
中川 裕志

分散型
ビッグ
データT
橋田 浩一

安全性と利便性

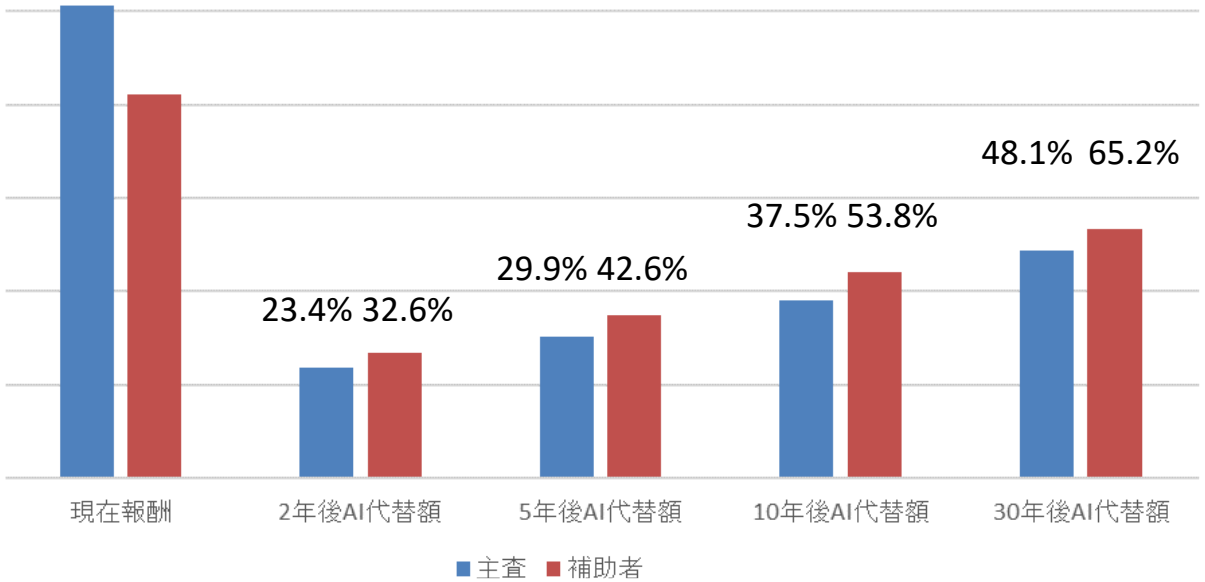
AIセキュリ
ティ・プラ
イバシーT
佐久間 淳

分析と介入

経済経営
情報融合
分析T
星野 崇宏

AIによる労働市場の変化

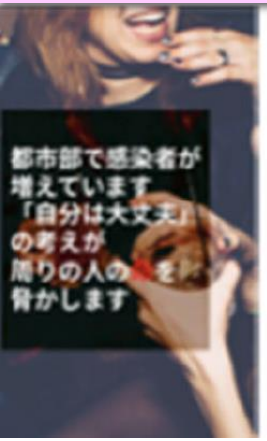
- 10～20年内に労働人口の47%が機械に代替されるリスクが70%以上(Frey & Osborne, 2013)というのは本当?
- 会計士報酬、昇給時の職務のウェイト、職務のAI代替可能性の調査
  - 日本公認会計士協会との共同研究
- 各業務の重要度・生産性を分析
- AI代替による労働時間配分と生産性向上の予測
  - 労働集約的要素がAIで代替され、付加価値の高い業務にシフト



ナッジ広告の感染リスク抑制効果

- 2020年7月に60万人を対象とする外出抑制のためのナッジ(行動のちょっとした後押し)広告の実験
- 経済損失訴求ナッジで週末の外出が52分/日減少
  - 経済的インセンティブ(>3千円)よりはるかに低コスト

ナッジ



非ナッジ

行動履歴データの蓄積

3,000超のアプリから**2,600万人**の精緻な行動履歴(外食や小売店の利用)を取得

位置情報計測

アルゴリズム最適化

事後に消費行動を調査し経済活動への影響を低減する通知を開発

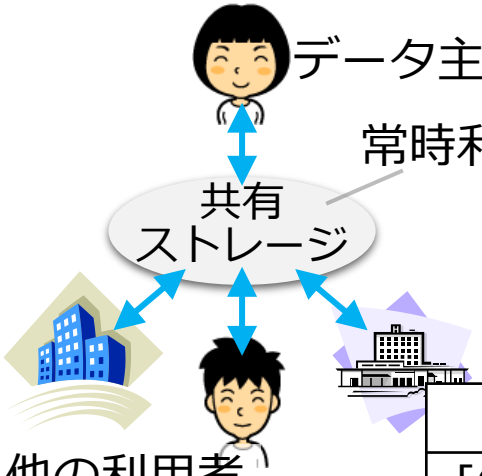
広告配信

本人主導のパーソナルデータ(PD)活用

PDを本人に集約 → 価値が最大化

- PDが**名寄せ**されて価値が高まる
  - 複数事業者に共通のID等は不要
- 1次利用: 本人へのサービスでの活用 ～ **AIの運用**
  - **各個人がPDをフル活用**できる
  - 機微性(と価値)が高いPDは他者に開示せずに活用
- 2次利用: 多数の人々のPDの統計分析 ～ **AIの開発**
  - **名寄せされたPDを本人同意だけで容易に収集**できる
  - 十分多くの人々がPDを提供してくれる
    - ✓ドナー登録やCOVID-19のアンケートに国民の10%がオプトイン
- PDの管理を個人に分散すれば**大規模な漏洩等がない**

データ共有技術の分類 [橋田2021]

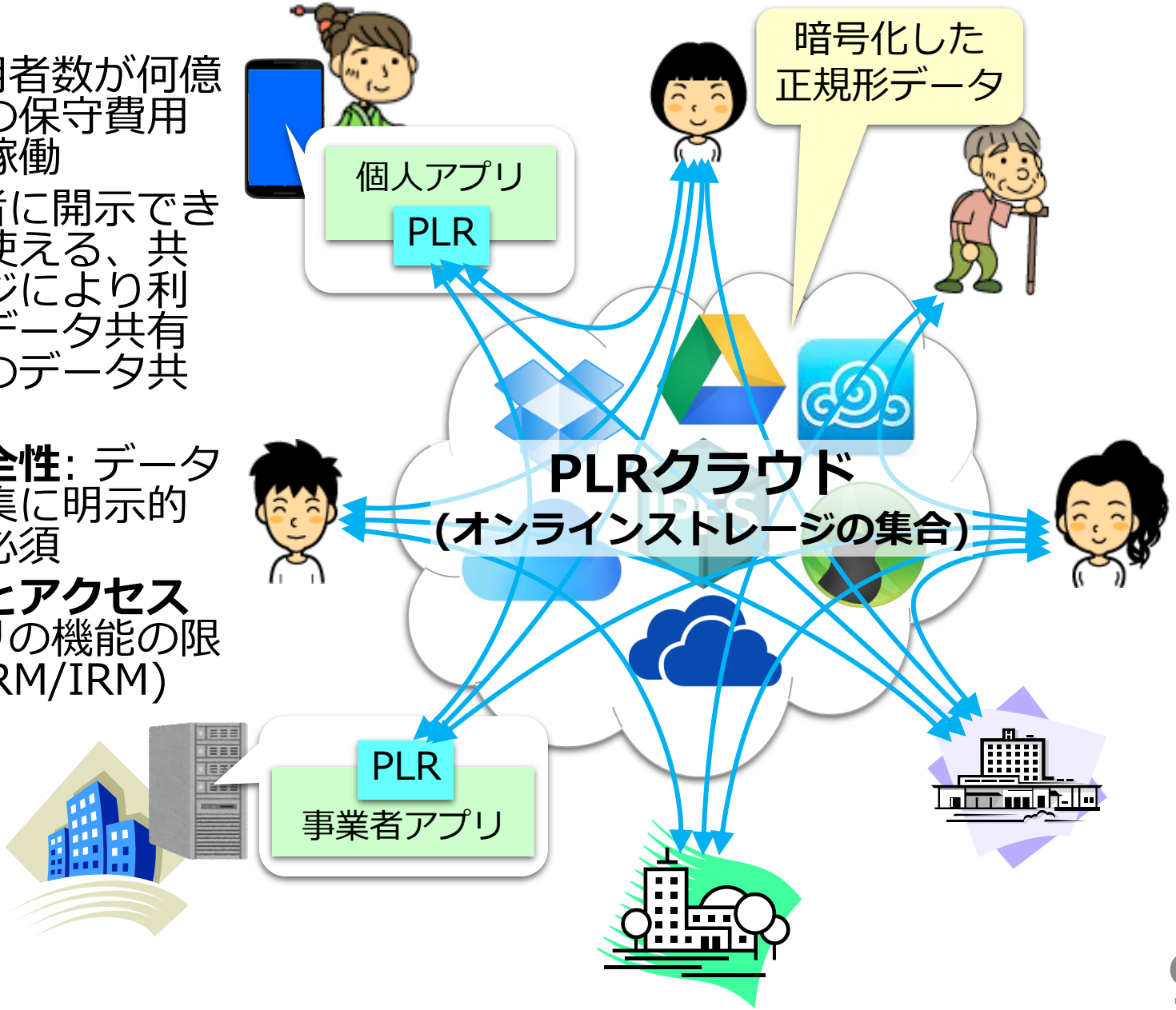


- 複数の利用者を持つ情報システム
- **AIの運用と開発**に必須
- どの技術も下記[O]~[S+]のいずれかに属する
- [S+]が最良で、実用化された実装はPLRのみ

	他者管理(集中型)		**自己管理(分散型)**	
	[O] X-Road、MedRec、…	[O+] ほとんどの情報システム	[S] digi.me、CitizenMe、…	[S+] PLR
共有関係	他者管理	他者管理	自己管理	自己管理
共有ストレージ	なし	他者管理	なし	自己管理
経済性	×集中管理のコストが高い		✓	
可用性	×他者に開示しない情報を活用できない		×個人端末による共有のみ	✓
機密性・完全性	×集中管理機能の誤用・悪用による大量データの漏洩等のリスク		✓	✓end-to-endの暗号化で過失による漏洩等も防げる
追跡可能性・アクセス制御	×	✓共有ストレージ内	×	✓共有ストレージ外DRM/IRM

[S+]共有関係と共有ストレージの自己管理(PLR)

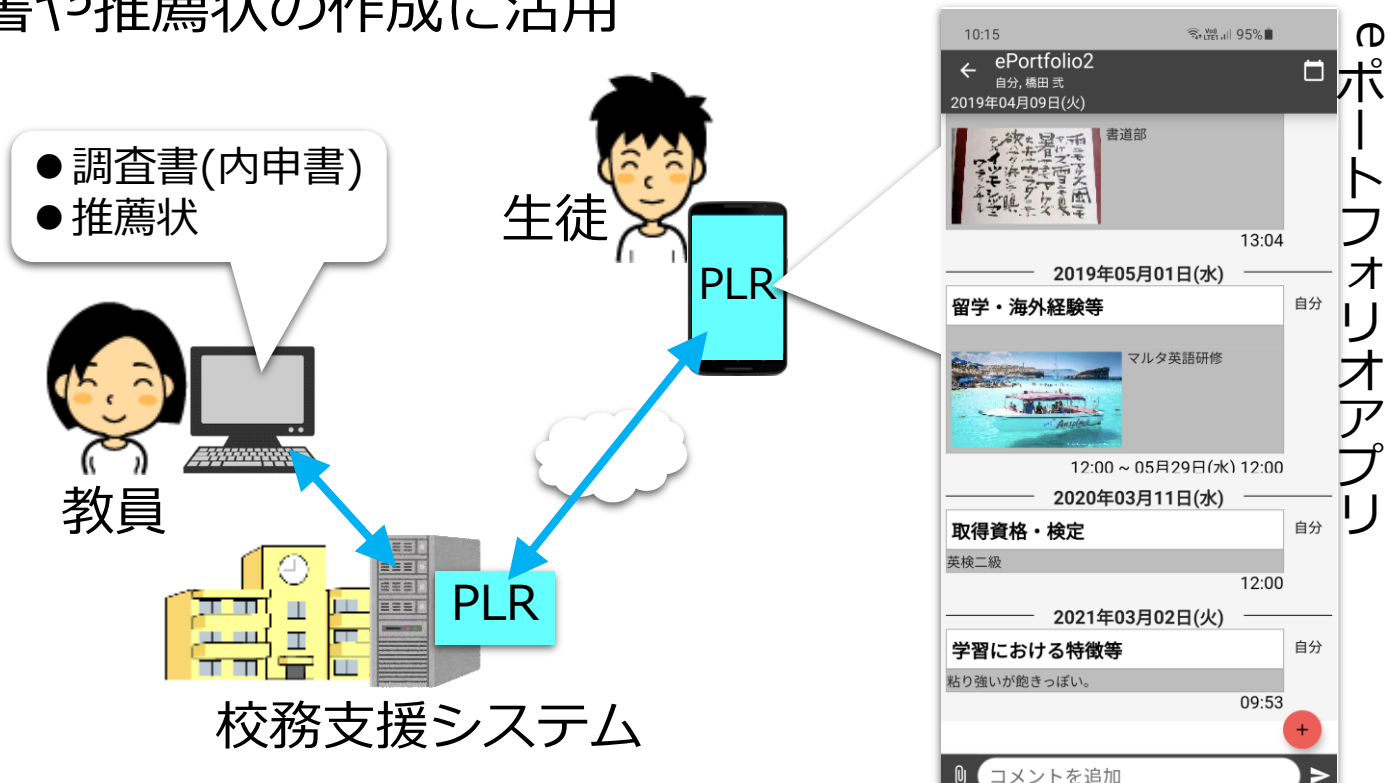
- **経済性**: 利用者数が何億でもアプリの保守費用だけで安定稼働
- **可用性**: 他者に開示できない情報も使える、共有ストレージにより利用者同士のデータ共有や常時大量のデータ共有が可能
- **機密性・完全性**: データの閲覧・編集に明示的本人同意が必須
- **追跡可能性とアクセス制御**: アプリの機能の限定と強制(DRM/IRM)



分散eポートフォリオ

[Hasida2020; 橋田2021]

- 課外活動を記録するeポートフォリオPLRアプリ
  - データポータビリティとセキュリティを確保
  - 電子調査書や生涯スタディログの基盤
- 埼玉県教育局が2020年度から実運用
  - 生徒がeポートフォリオ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



グループの構成

科学技術
と社会T
佐倉 統

AI倫理・
社会T
鈴木 晶子

倫理とガバナンス

AI安全性・
信頼性U
荒井 ひろみ

社会におけ
るAI利活用
と法制度T
中川 裕志

分散型
ビッグ
データT
橋田 浩一

安全性と利便性

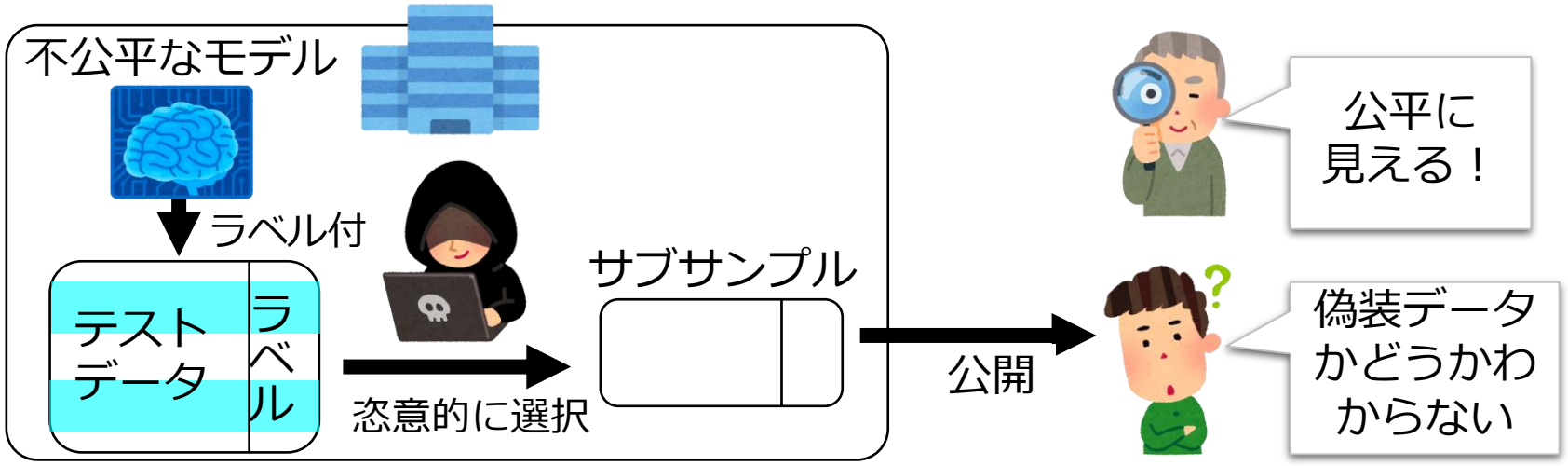
AIセキュリ
ティ・プラ
イバシーT
佐久間 淳

分析と介入

経済経営
情報融合
分析T
星野 崇宏

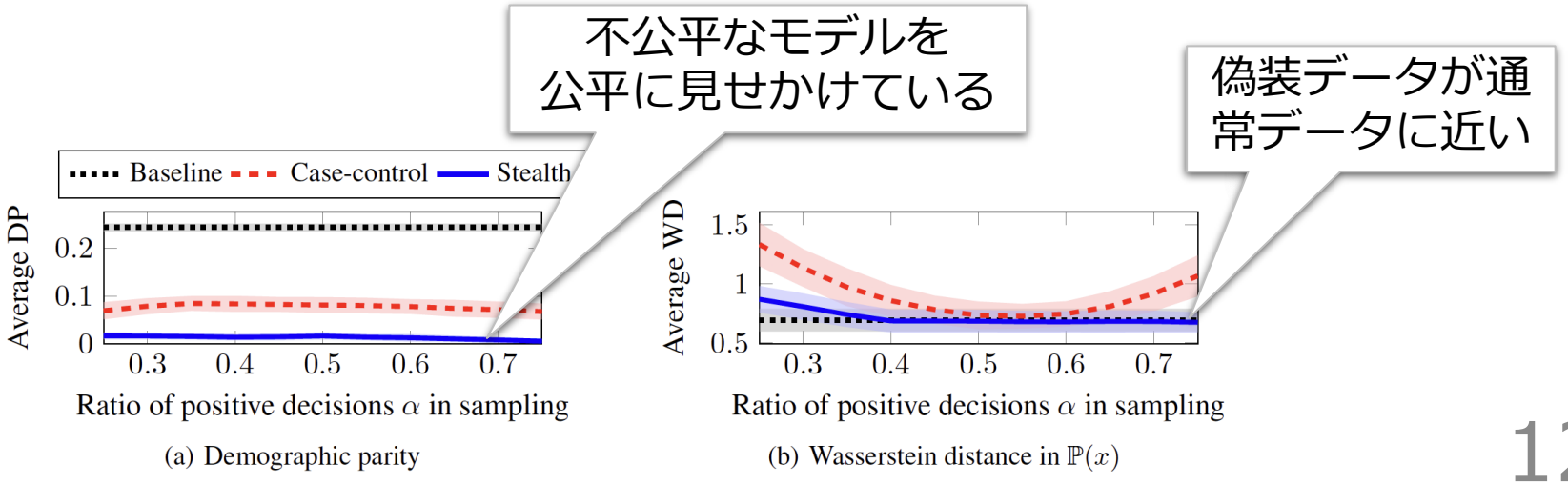
AIは公平性を偽装できる

[Fukuchi+2020]



不公平なモデルを公平に見せるサブサンプルを生成するアルゴリズムを開発

→ **公平性偽装対策の必要性を指摘**

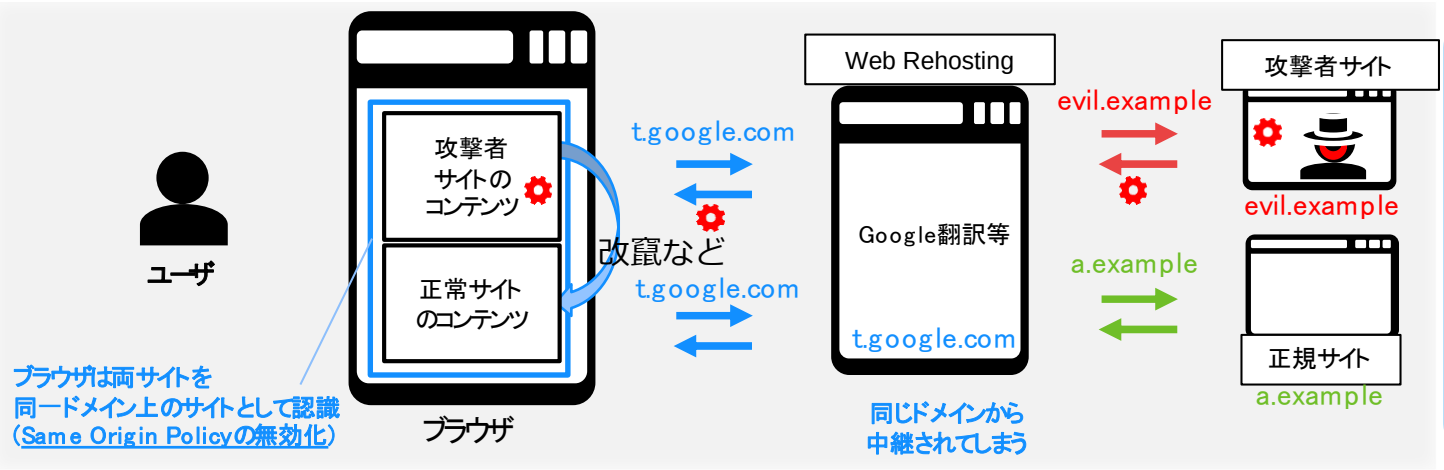


Web Rehosting※のセキュリティリスク

[Watanabe+2020] Distinguished Paper Award

※他サイトのコンテンツを再ホストして中継するサービス

- 通信内容を半永久的に傍受・改竄されるリスク



どういった攻撃につながるか?

1. 悪性スクリプト感染による中間者攻撃
2. パスワードマネージャーの共有によるID/PW 漏洩
3. Cookieの共有によるセッションハイジャック
4. GPSやカメラへのアクセスパーミッション流用
5. ストレージ共有による閲覧履歴漏洩

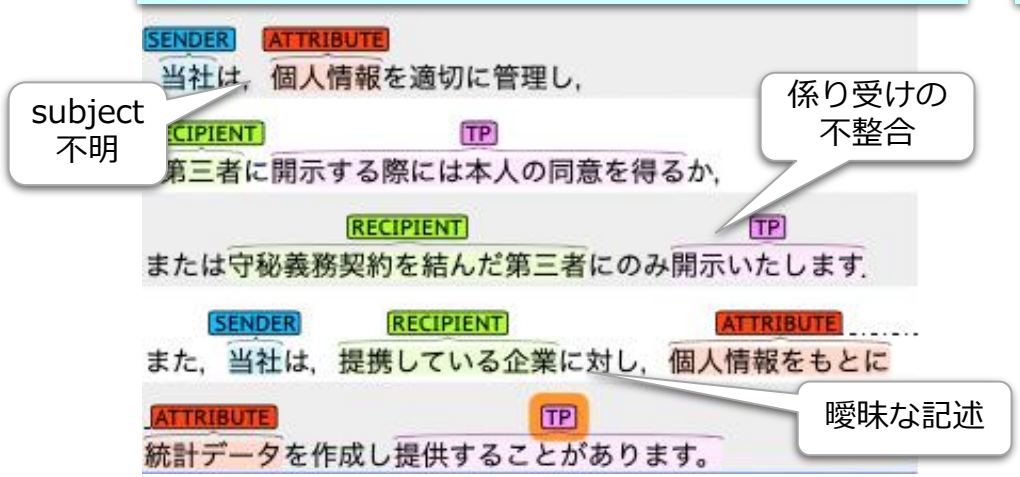
- 多様なリスクを発見して網羅的・体系的に分類
- 調査した21サービスのうち18が脆弱
  - Google翻訳、Googleキャッシュ、Internet Archiveなどは1日に数千万～数億件のアクセス
- 対策
  - 悪性スクリプト取得時に現れる文字列をブラウザが検知してブロック
  - 再ホストするページごとにドメインを分ける
  - ➡Googleが採用、他サービスでも検討中

プライバシーポリシーの評価

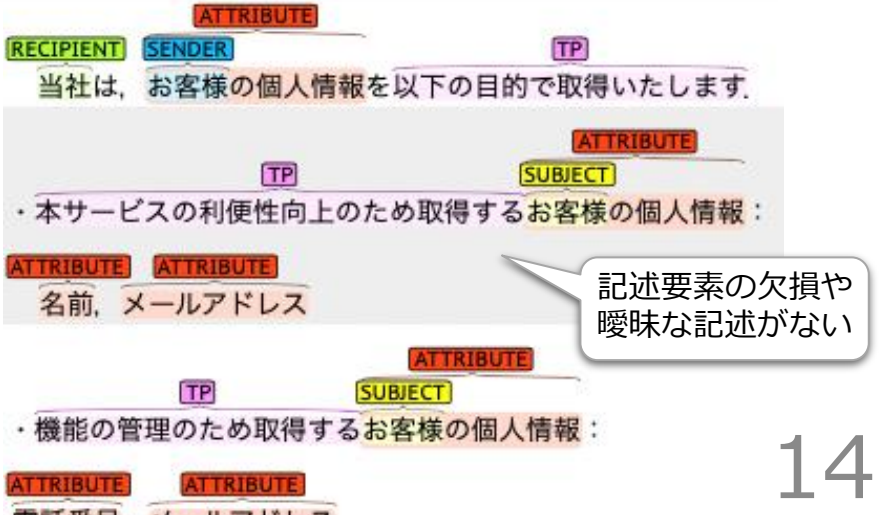
[荒井+2020] 優秀論文賞

- プライバシーポリシーの問題
  - 曖昧な表現
  - 不正確な記述
  - 複雑で難解な長文のため読まれない
- 情報の流れのアノーテションで曖昧性や情報不足を明示
  - subject (情報の主体)、sender (情報の送り手)、…

情報の流れが未整理のプライバシーポリシー



情報の流れが整理されたプライバシーポリシー



グループの構成

科学技術
と社会T
佐倉 統

AI倫理・
社会T
鈴木 晶子

倫理とガバナンス

AI安全性・
信頼性U
荒井 ひろみ

社会におけ
るAI利活用
と法制度T
中川 裕志

分散型
ビッグ
データT
橋田 浩一

安全性と利便性

AIセキュリ
ティ・プラ
イバシーT
佐久間 淳

分析と介入

経済経営
情報融合
分析T
星野 崇宏

技術の社会的形成と社会受容性

- 社会・文化的側面
  - 日本のテクノアニミズム
    [佐倉2020; 猪口2020; 前田 2020]
    - ✓ペットロボット、ロボット供養、…
    - ✓仲間としてのAIと共生する社会の可能性
  - AIによるファッションデザイン
    [藤嶋 準備中]
    - ✓AIが「人間的な作業」をするというイメージが「人／機械」の二分法を崩し、「手仕事＝高級」・「機械生産＝大衆的」というステレオタイプを変える?
- 人間的側面
  - 利用者視点でのAIの倫理ガイドライン項目をモデルケースから抽出
    - ✓ISO/TR 9241-810:2020 Ergonomics of human-system interaction — Part 810: Robotic, intelligent and autonomous systems に採用

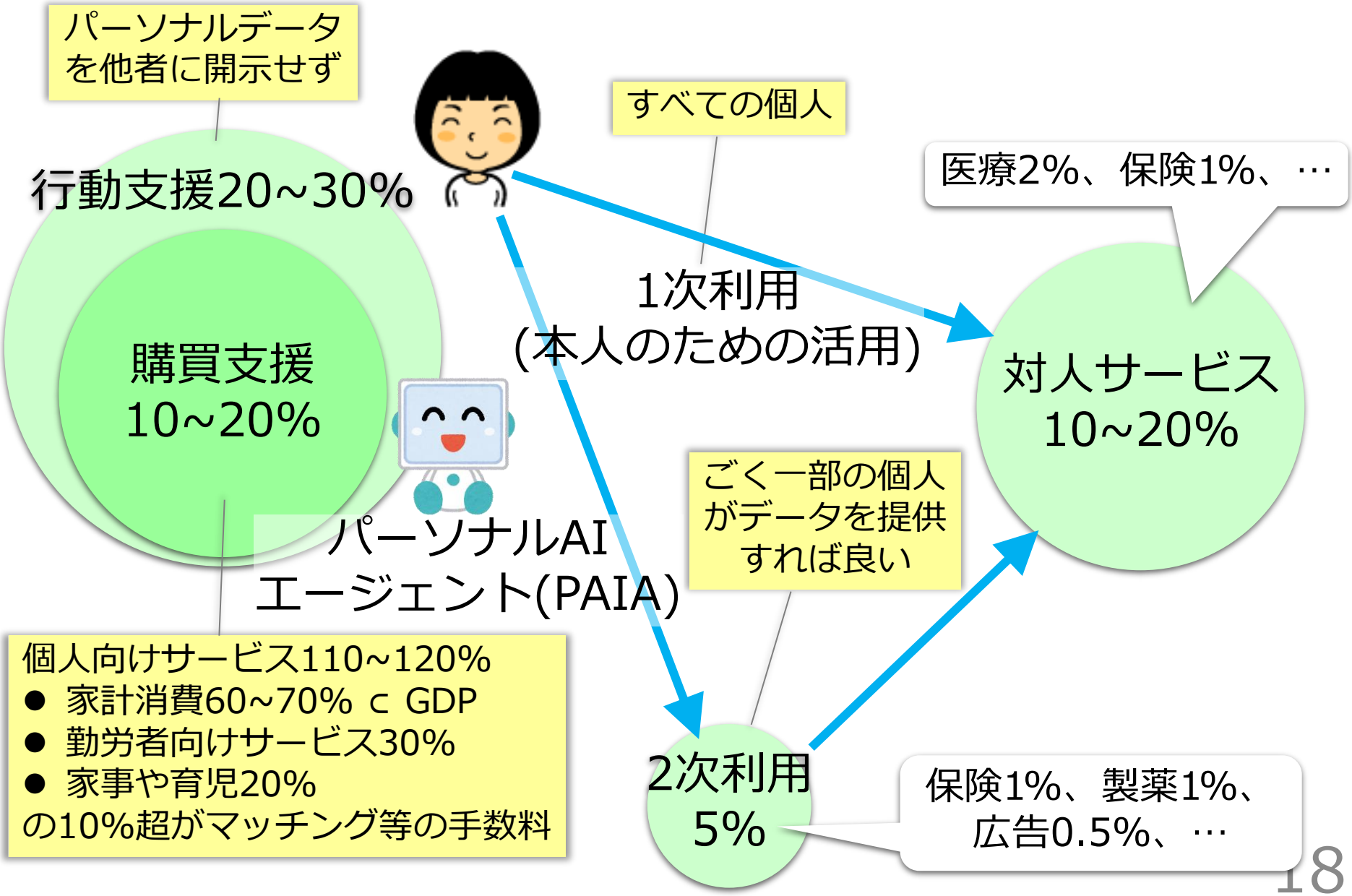
- ユースケースの多様性考慮
- AIに対する理解の浸透
- AIと人の信頼度
- サービスにおけるAIの役割
- AIとコミュニケーション
- AI予測の受容性
- 自己決定権

人間と技術を共進化させる総合倫理

- レジリエンス: 状況の変化に対する適応力
  [Ueno+2020; Kobayashi+2020; Miyagi+2020]
  - 東洋の現象学的視点: 遠目の目、不動心、…
  - 脳機能画像解析…デフォルトモードネットワークの機能結合
    - ✓ レジリエンスが高いほど注意課題で安静を保ちやすい
    - ✓ 中等度以下のデジタルマルチタスク状態も注意課題で安静を保ちやすい
- タクト(takt): 状況に適応して秩序を保つ制御能力
  [Suzuki2020; Berberich+2020; 鈴木2020]
  - 触覚知性、リズム、身体知、暗黙知等にわたる知
  - タクト～レジリエンス?
  - AIと共存する社会における人間に必要な能力・実践的技倆
- 感情(～行動制御機能)の役割の変容
  - 感情労働、ネットでの感情表現、…
- AIの根本問題との関連

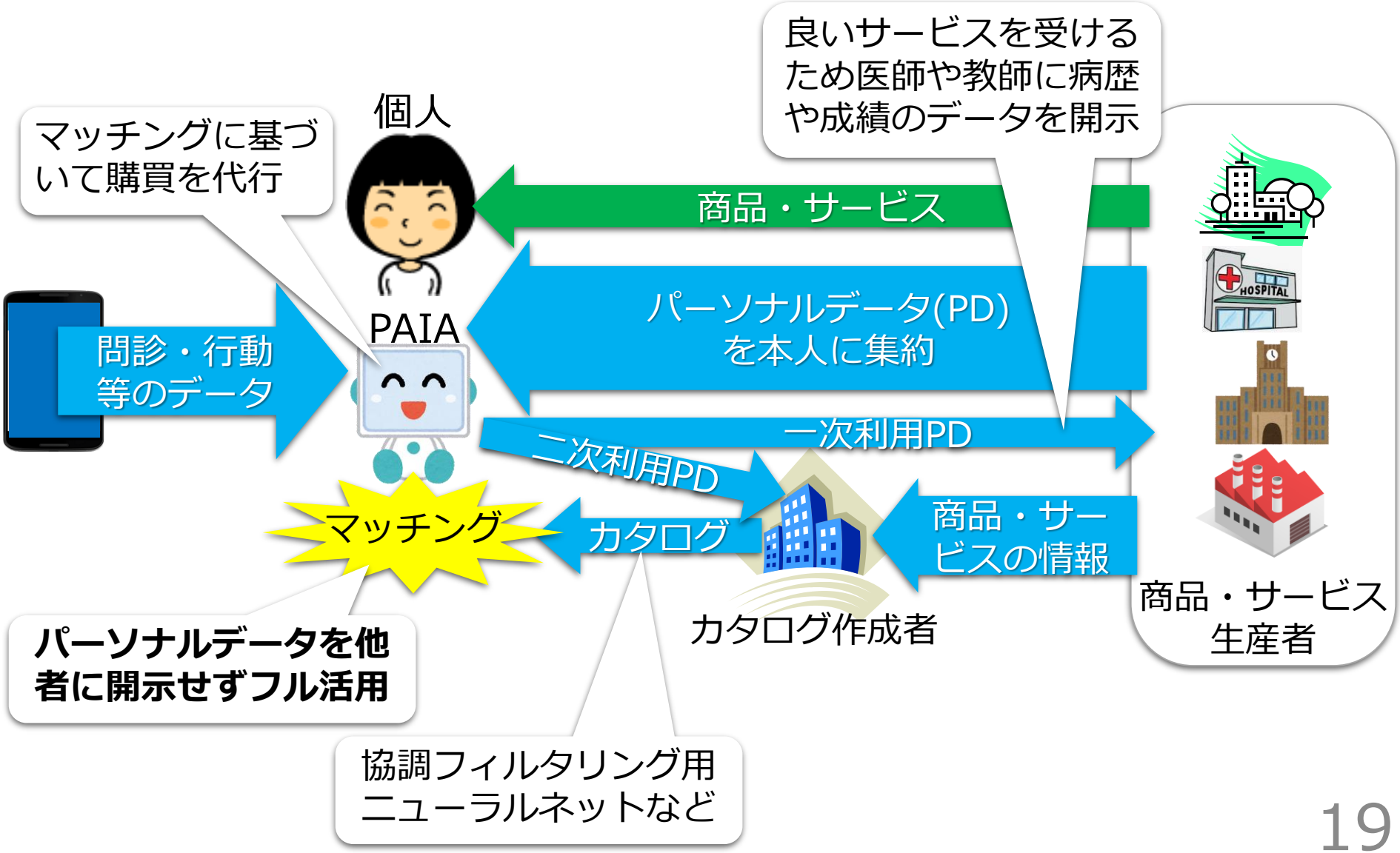
木を見て森を見ず

パーソナルデータの価値(対GDP%)



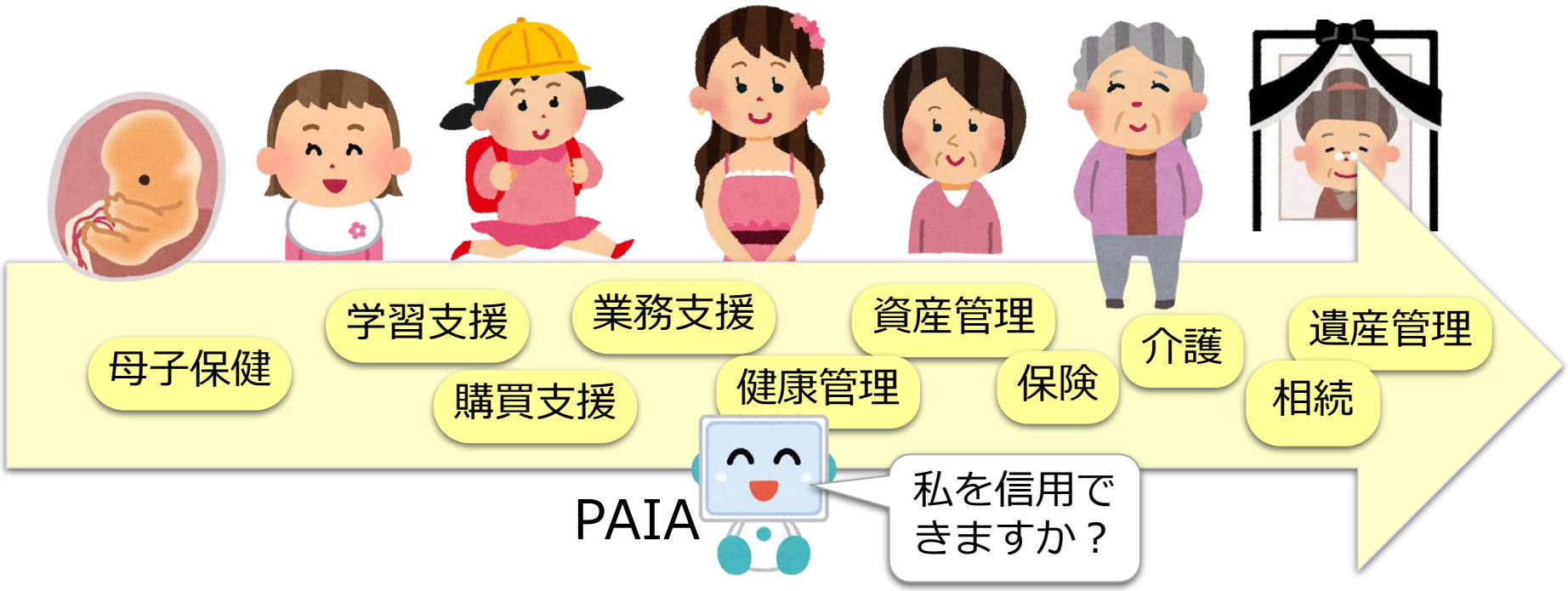
購買支援

市場規模 > GDPの10%



誕生前から死後まで個人に寄り沿うPAIA

[吹金原+2020; 橋田2020; 中川2021, Nakagawa2021]



- 認知限界: 人間は複雑な情報環境に十分対処できない
- PAIAが個人の生活行動全般に介入し行動変容を促して人生を全体最適化
  - パーソナルデータをフル活用して個人を支援・代理・支配?
- 期待と脅威
  - 経済の安定/崩壊、民主主義の強化/弱体化、…

まとめ

- 成果
  - 対外発表: 国内外の招待講演など多数
  - 社会活動
    - ✓ 政府の委員会の委員など多数 → 立法など
    - ✓ ワークショップ等の開催
  - 国際標準化
    - ✓ 構造化文書 … 分散型ビッグデータT
    - ✓ AI倫理 … 科学技術と社会T
  - 社会実装
    - ✓ 政府統計 … 経済経営情報融合分析T
    - ✓ 公教育 … 分散型ビッグデータT
    - ✓ Web … AIセキュリティ・プライバシT
    - ✓ 個人情報保護制度 … 社会におけるAI利活用と法制度T
- 課題
  - AIと社会との関係の解明と改善
    - ✓ 人間・社会とAIの共進化の可能性とリスクの分析
  - AIの開発・導入・運用の社会基盤
    - ✓ データの生成・共有・活用を促進し最適化する実証と社会実装
    - ✓ 社会的公正を高め民主主義を強化する技術とガバナンス
  - 人文社会的研究と数理的研究との融合
    - ✓ パーソナルAIエージェントなど