2020/2/12
NICTサーバーセキュリティシンポジウム2020

AIの利活用を巡る課題

中川裕志
(理化学研究所　革新知能統合研究センター)

生体情報のプライバシー

■　ある国ではごみについたDNAから顔を復元し、ポイ捨てした人をポスターにするキャンペーンがあるらしい
WRED.jpより

◆　一方で、EUでは顔認証システムを公共の場所で使うことを禁じる
CPDP2019より

フラッシュクラッシュ
ブラックボックス化の金融への悪影響

- 人工知能技術のブラックボックス化が社会にリアルな損害を与えています

- 金融取引（株の売買など）は、既にネットワークを介してエージェントベースで秒以下の売り買いされる世界です。
  - エージェントに人工知能が使われています。
- 人間（トレーダー）が介入して判断するより早く事態は進行します。
- 世界中の金融センターも似たような状況なので、なにかのトリガがかかると連鎖反応が瞬時におこり、とんでもないことになります。

ウォールストリート・暴走するアルゴリズム
Wired.jp 2016年9月　より　1

- だが最悪の場合、それら(AIトレーダー）は不可解でコントロール不能のフィードバックのループとなる。

- これらのアルゴリズムは、ひとつひとつは容易にコントロールできるものなのだが、ひとたび互いに作用し合うようになると、予測不能な振る舞いを–売買を誘導するためのシステムを破壊しかねないようなコンピューターの対話を引き起こしかねないのだ。

ウォールストリート・暴走するアルゴリズム　Wired.jp 2016年9月　より　3

- アルゴリズムを用いた取引は個人投資家にとっては利益となった。以前よりもはるかに速く、安く、容易に売買できるのだ。
- だがシステムの観点からすると、市場は迷走してコントロールを失う恐れがある。

- たとえ個々のアルゴリズムは理にかなったものであっても、集まると別の論理–人工知能の論理に従うようになる。人工知能は人工の「人間の」知能ではない。

- それは、ニューロンやシナプスではなくシリコンの尺度で動く、われわれとはまったく異質なものだ。その速度を遅くすることはできるかもしれない。だが決してそれを封じ込めたり、コントロールしたり、理解したりはできない。

ウォールストリート・暴走するアルゴリズム
Wired.jp 2016年9月　より　２

- 2010年５月６日、ダウ・ジョーンズ工業株平均はのちに「フラッシュクラッシュ（瞬間暴落）」と呼ばれるようになる説明のつかない一連の下落を見た。一時は５分間で573ポイントも下げたのだ。

- ノースカロライナ州の公益事業体であるプログレス・エナジー社は、自社の株価が5カ月足らずで90%も下がるのをなすすべもなく見守るよりほかなかった。９月下旬にはアップル社の株価が、数分後には回復したものの30秒で４%近く下落した。

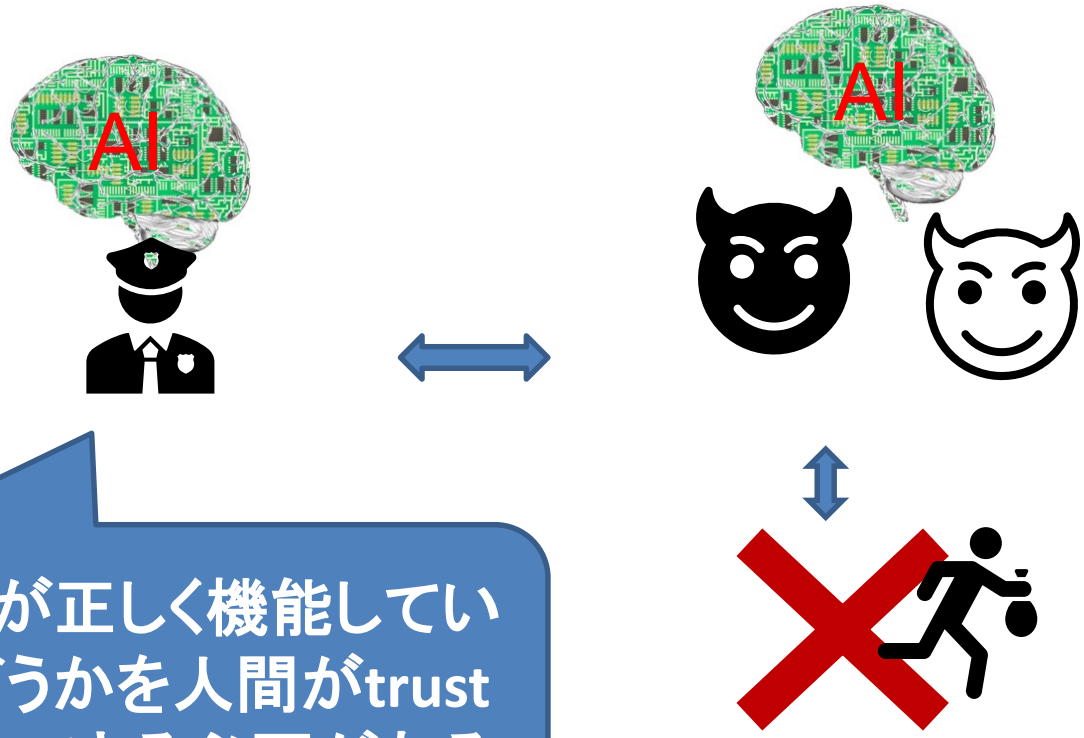
AI内部に説明能力を組み込む XAI(eXplainable AI) の研究が盛んですが

- フラッシュクラッシュのような多数のAIが相互作用する場合にはとても追いついていない。

AIトレーダーの損害検出能力

- 適格な行為者の高速トレードの誤動作（フラッシュクラッシュ）の早期発見、検証が人間にできそうもない
- 許容損害値超過を早期発見するチェック機構が必要
- ネットワーク接続された多数のAIトレーダーの行動チェック機構はAIトレーダーを外部観察、あるいはネットワークの挙動を外部観察するAIとして作る
  - チェック自体を学習機能を持つ人工知能に任せるような局面が考えられますが、果たして可能か？

AIの行動をAIが観察してテスト

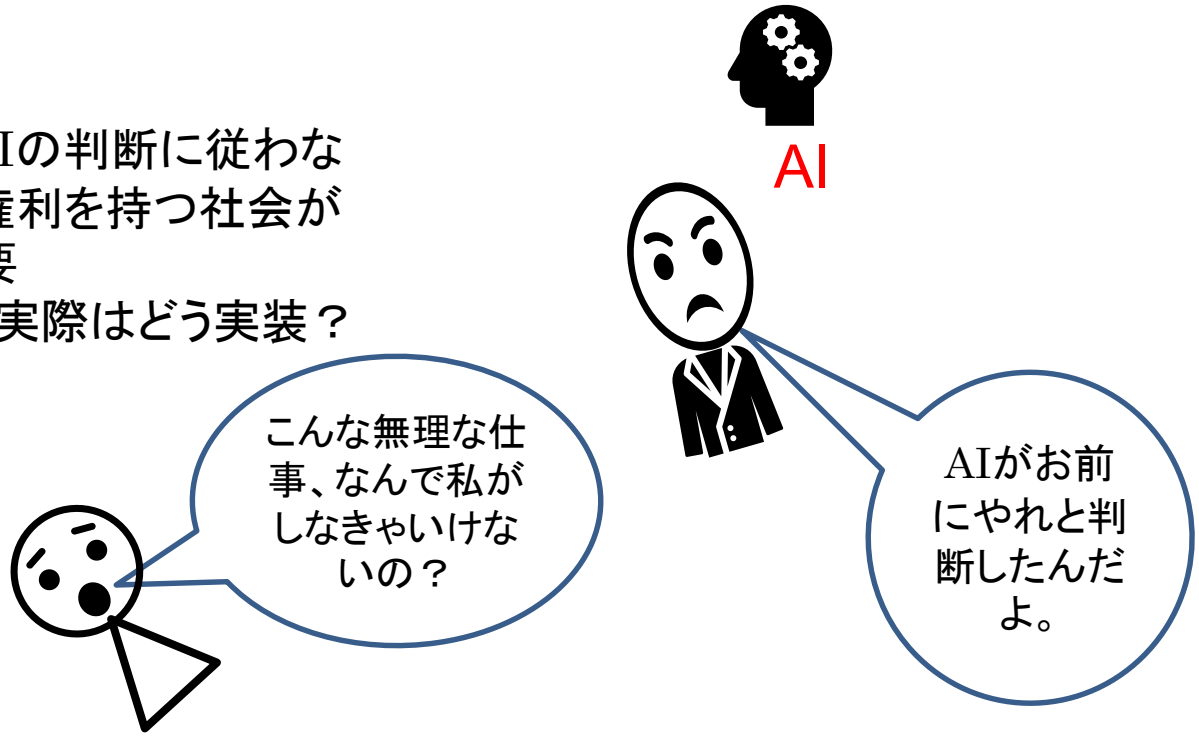


- 早く止めすぎると、収益の機会を奪ったという文句がくる
- 止めるのが遅すぎるとクラッシュして大きな被害を出す

➢ 収益ロスとクラッシュの被害の和を最小化する止め方　→ 最適化問題

AIの悪用

- AIの判断に従わない権利を持つ社会が必要
- 実際はどう実装？



- 悪用された人工知能に一般人が文句を言うことができなくなってくると、実質的に言論の自由も人権もない状況になりかねません

➢ 人工知能に対して文句を言える社会制度を考える必要があります。

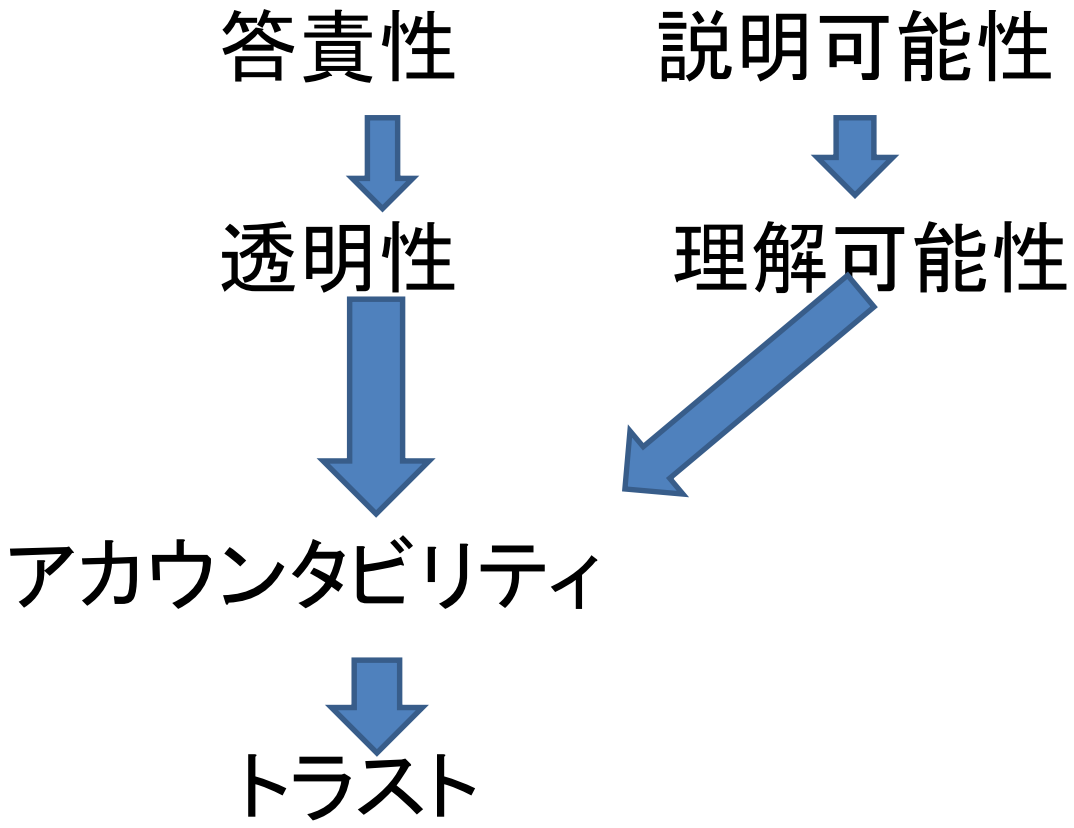
  ➢ GDPR 22条： 計算機（人工知能）のプロファイリングから出てきた決定に服さなくてよい権利
  ➢ 具体的には以下のようにします。
    ➢ プロファイリングに使った入力データの開示
    ➢ 出力された決定に対する説明責任を人間が果たす。
    ➢ ただし、この権利の社会実装、技術による実現はかなり困難
    ➢ 守秘義務や企業秘密の壁もあります。

AIのブラックボックス化

- 人間の仕事を自律的AIで置き換えるにせよ、人間の能力を拡張するにせよ
  - 開発者に責任はあるはず
  - だが、既に関係者たちが把握しきれない状態に突入しているのかもしれません
  - 関係者：　Multi-stakeholder
    - 人工知能開発者
    - 人工知能へ学習に使う素材データを提供した者
    - 人工知能製品を宣伝、販売した者
    - 人工知能製品を利用する消費者

➢したがって、事故時の責任の所在を法制度として明確化しておく必要がある時期になってきています

信用できるAIへ向けての取り組み：
透明性、説明可能性、トラスト



トラストの補足

- 利用者がサービス提供側をトラストするという局面ばかり考えてきたが

- サービス提供側が利用者をトラストするという問題もある
  - 利用者認証（多くはネットワーク越し）
  - Bad userではないことを推定
  - Self Sovereign Identity　→ 対応するサービスに必要なことだけ identify できればよい

AI倫理指針

国内外の組織が提案している人工知能の倫理（古い順）

- FLI: Asilomar AI Principles (23原則) (2017)
- IEEE Ethically Aligned Design, version 2(2017/12)
  - AIおよび開発者が持つべき倫理
- Partnership on AI (2016~
- 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(2019/3/29)
  - AI ready な社会の在り方 G20に提案

国内外の組織が提案している人工知能の倫理

- IEEE Ethically Aligned Design, first edition (2019/3)
  - 倫理的なAIの設計指針
- EU: High Level Expert Group: Ethics Guidelines for Trustworthy AI (2019/4/8)
  - 倫理的なAIの設計指針
- Recommendation of the Council on OECD Legal Instruments Artificial Intelligence
  - OECD 閣僚理事会承認 (2019/5/22)
- Beijing AI Principle (2019/5/25)
- Guidance for Regulation of Artificial Intelligence Applications:
  - USA Whitehouse. MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES (Draft 2019/4/24)

IEEE Ethically Aligned Design version 2

1. Executive Summary
2. General Principles
3. Embedding Values Into Autonomous Intelligent Systems
4. Methodologies to Guide Ethical Research and Design
5. Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI)
6. Personal Data and Individual Access Control
7. Reframing Autonomous Weapons Systems
8. Economics/Humanitarian Issues
9. Law
10. Affective Computing
11. Classical Ethics in Artificial Intelligence
12. Policy
13. Mixed Reality
14. Well-being

The final version was published

IEEE EAD （Final） on April 2019

- 1. Human Rights
  - A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
- 2. Well-being
  - A/IS creators shall adopt increased human well-being as a primary success criterion for development.
- **3. Data Agency**
  - **A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.**
- 4. Effectiveness
  - A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.

IEEE EAD （Final） on April 2019

- 5. Transparency
  - The basis of a particular A/IS decision should always be discoverable.
- 6. Accountability
  - A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.
- **7. Awareness of Misuse**
  - **A/IS creators shall guard against all potential misuses and risks of A/IS in operation.**
- 8. Competence
  - A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

人間中心AIの社会原則
Social Principles of Human-Centric AI
2109/3/29

Council for Social Principles of Human-centric AI

内閣府, Japan

AI社会原則

1. 人間中心の原則、
2. 教育・リテラシーの原則、
3. プライバシー確保の原則
4. セキュリティ確保の原則、
5. 公正競争確保の原則
6. 公平性、説明責任及び透明性の原則、
7. イノベーションの原則

EC High Level Expert Group

Ethical Guideline for Trustworthy AI

- 2019/4/8

Trustworthy AI

- Lawful, Ethical, Robust

- 具体的要件
  1. Human agency and oversight
  2. Technical robustness and safety
  3. Privacy and data governance
  4. Transparency
  5. Diversity non-discrimination and fairness
  6. Societal and environmental well-being
  7. accountability

Recommendation of the Council on　OECD Legal　Instruments　Artificial Intelligence

- 2019年5月22日のOECD　閣僚理事会Iで採択

- 強制力はないが、各国での立法の指針になる可能性大

  – 例：　OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data　（1980）は各国のプライバシー保護法の基礎になった

OECD, Recommendation of the Council on A I, OECD/LEGAL/0449, 2019/5/23

- 技術的目標:
- inclusive growth, sustainable development and well-being
- human-centred values and fairness
- transparency and explainability
- robustness, security and safety
- accountability
-
- 政策的目標:
- investing in AI research and development
- fostering a digital ecosystem for AI
- shaping an enabling policy environment for AI
- building human capacity and preparing for labour market transformation
- international co-operation for trustworthy AI

Guidance for Regulation of Artificial Intelligence Applications: USA Whitehouse.

- AI倫理というよりはむしろ、AIシステム開発のガイダンスで、AI産業の育成が目標
  - 名宛人は産業界と読める
  - 例えば「 AIアプリケーションの技術仕様を規定しようとする厳格な設計ベースの規制は、AIが進化する予想されるペースを考えると、ほとんどの場合、非実用的で非効率的」

- 指針は以下の10項目からなります。

1. **Public Trust** in AI
2. Public Participation
3. Scientific Integrity and Information Quality
4. **Risk Assessment and Management**
5. **Benefits and Costs**
6. Flexibility
7. Fairness and Non-Discrimination
8. Disclosure and Transparency
9. Safety and Security
10. Interagency Coordination

◆いかに規制しないかが根底

- ただし、無制限な開発への歯止めとして他の倫理指針と違うのは
- risk assessment、risk management
- リスク評価をサボると public trust　を失うぞ、という言い方
  - 下記の引用を参照

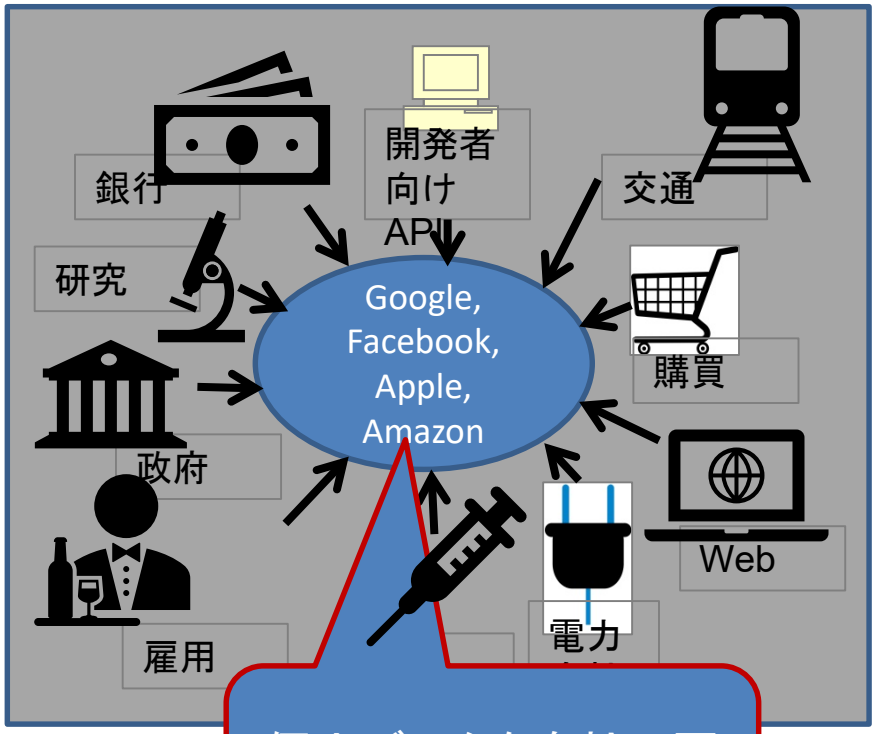
- A risk-based approach should be used to determine **which risks are acceptable and which risks present the possibility of unacceptable harm, or harm that has expected costs greater than expected benefits.**
- Agencies should be transparent about their evaluations of risk and re-evaluate their assumptions

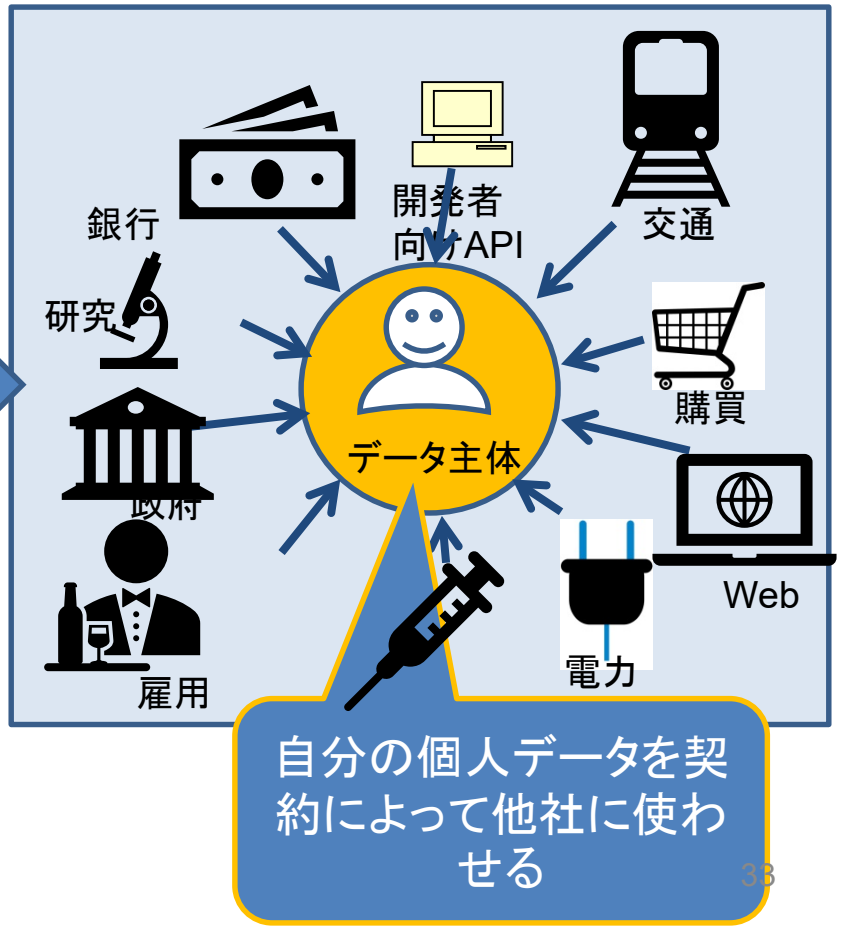
	AI制御	人権	公平性　非差別	透明性	アカウンタビリティ	トラスト	悪用、誤用	プライバシー	AIエージェント	安全性	SDGs	教育	独占禁止・協調、政策	軍事利用	法律的位置づけ	幸福
Asilomar Principles	○	○						○						○		○
人工知能学会・倫理指針	△	○					○	○		○					○	○
総務省AI開発ガイドライン	○	○	○	○	○			○		○						○
Partnership on AI		○	○	○	○			○	○	○	○	○	○			○
IEEE EAD ver2	○	○	○	○	○	○	○	○	○	○		○		○	○	○
IEEE　EAD 1e		○	○	○	○	○	○	○	○	○	○	○			○	○

	AI制御	人権	公平性　非差別	透明性	アカウンタビリティ	トラスト	悪用、誤用	プライバシー	AIエージェント	安全性	SDGs	教育	独占禁止・協調、政策	軍事利用	法律的位置づけ	幸福
人間中心AI社会原則		○	○	○	○	○	○	○		○	○	○	○			○
Trustworthy AI		○	○	○	○	○	○	○	○	○	○	○	○	○	△	○
OECD Recommendation		○	○	○	○	○	△	○		○	○		○			○
総務省AI利活用ガイドライン			○	○	○	○	○	○		○						○
Beijing Principle	○	○		○	○	○		○		○		○	○		△	○
Whitehouse Guidance	×		○	○	○	○	○	○		○			○		△	○

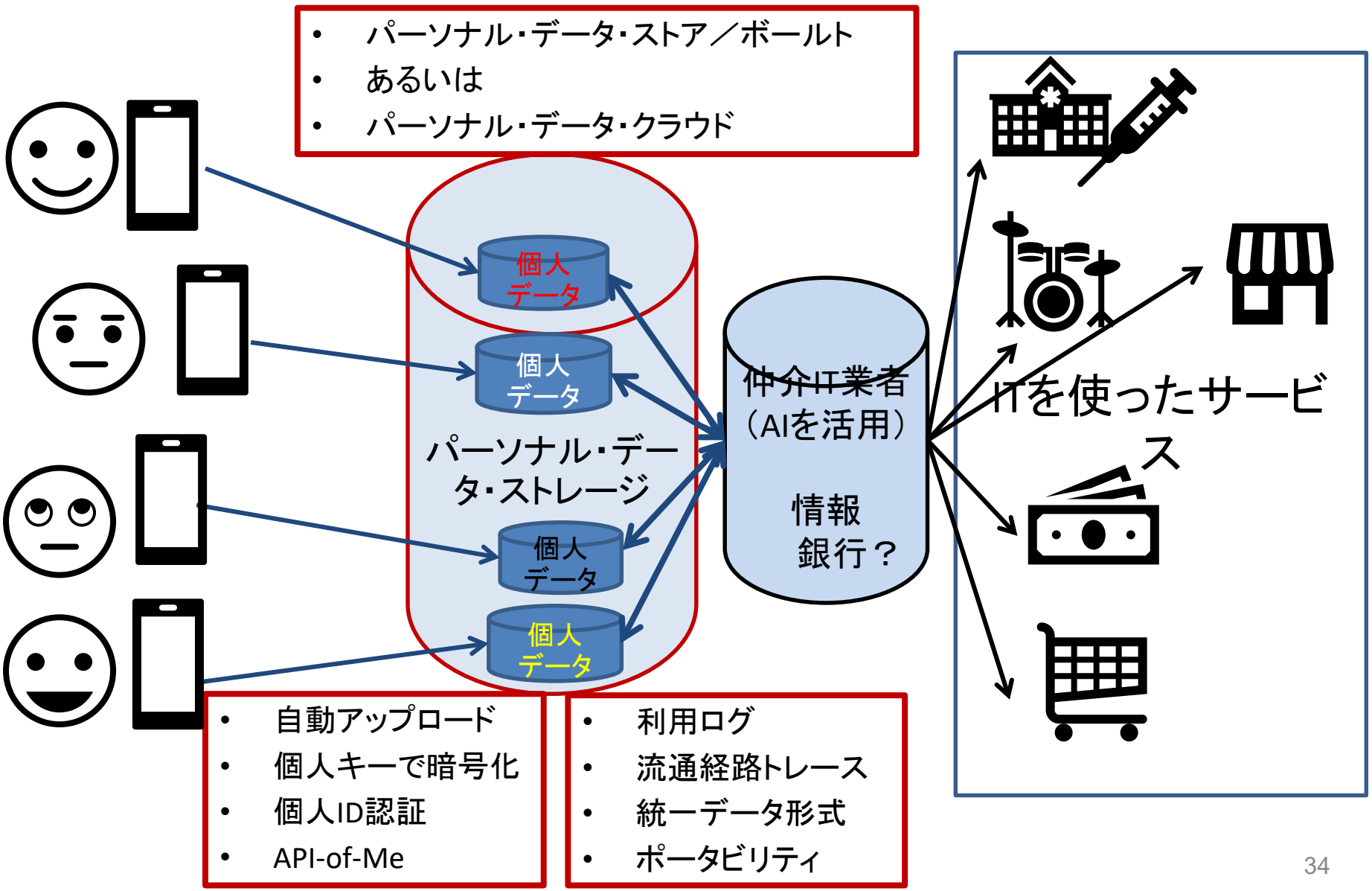
法的位置づけ：○＝AI人格権、△＝AIの現行法への適法性

個人データ管理は
データ主体の個人へ：MyData





パーソナル・データ・ストレージ（PDS)



主要な技術的ポイント

- パーソナルクラウド
- インターネットにおける Identity 認証
- 個人データのポータビリティ
- Block Chain による個人データの真正性認証
- プライバシー保護(暗号化,複数当事者による計算 : MPC , etc.)
- 公平性、透明性の確保手段

データポータビリティ

- GDPR20条
- データ主体は、個人データを機械可読性のある形式で受け取る権利があり、
- 当該データを、個人データが提供された管理者の妨害なしに、他の管理者に移行する権利がある。
  - ただし、20条3項 「職責によって収集した個人データ（例えば医療データ、医療カルテ）にはデータポータビリティは適用しない」

- Googleの個人データAPI

- 日本
  - 銀行API、個人医療履歴、

残された問題

- トラストするのは
  - 組織か　（利用者からみたら）
    - 情報銀行、ITサービス企業
  - 人か　（サービス業者からみたら）
    - 利用者の個人をトラストできるか
    - Self Sovereign Identity
  - データか
    - 流通する個人データか、サービスの結果か？

個人データ個人管理の問題点

- 個人データがどう使われているかにsensitiveな人は多いのか？
  - 痛い目をみるまで分からない
  - ポイントの餌に釣られる？　目先の利益を優先する人々が大多数
  - だからこそ、きちんと規制すべきという意見もあるが……
- 個人データを自分で管理するスキルがない人が大部分

➢ 近代的個人の消失につながるのか？

パーソナルAIエージェントとガバナンス

- **背景**：既存のガバナンスの枠組みがBrexit、トランプ現象、中国の台頭などで揺らいでいる現状への危機感

➢ **新しい方向性**

(1)ディジタル・レーニズム

(2)GAFAのような国境を超えるITプラットフォームによる情報支配

(3)既存の民主主義を基礎にするガバナンスの拡充

(3)は望ましいが、多くの人間は近代的法制度、政治制度が前提にした完全な自我と自由意志に基づいて行動する主体には程遠い

パーソナルAIエージェントとガバナンス

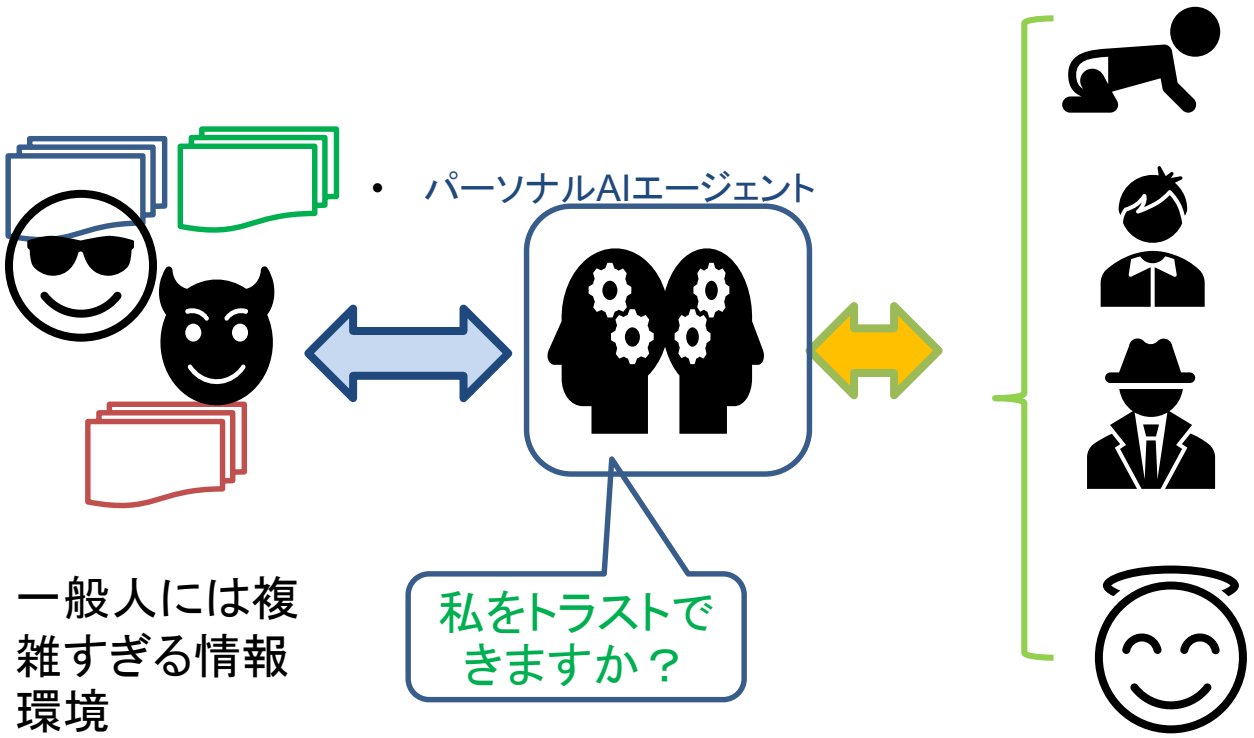
”(3)既存の民主主義を基礎にするガバナンスの拡充“

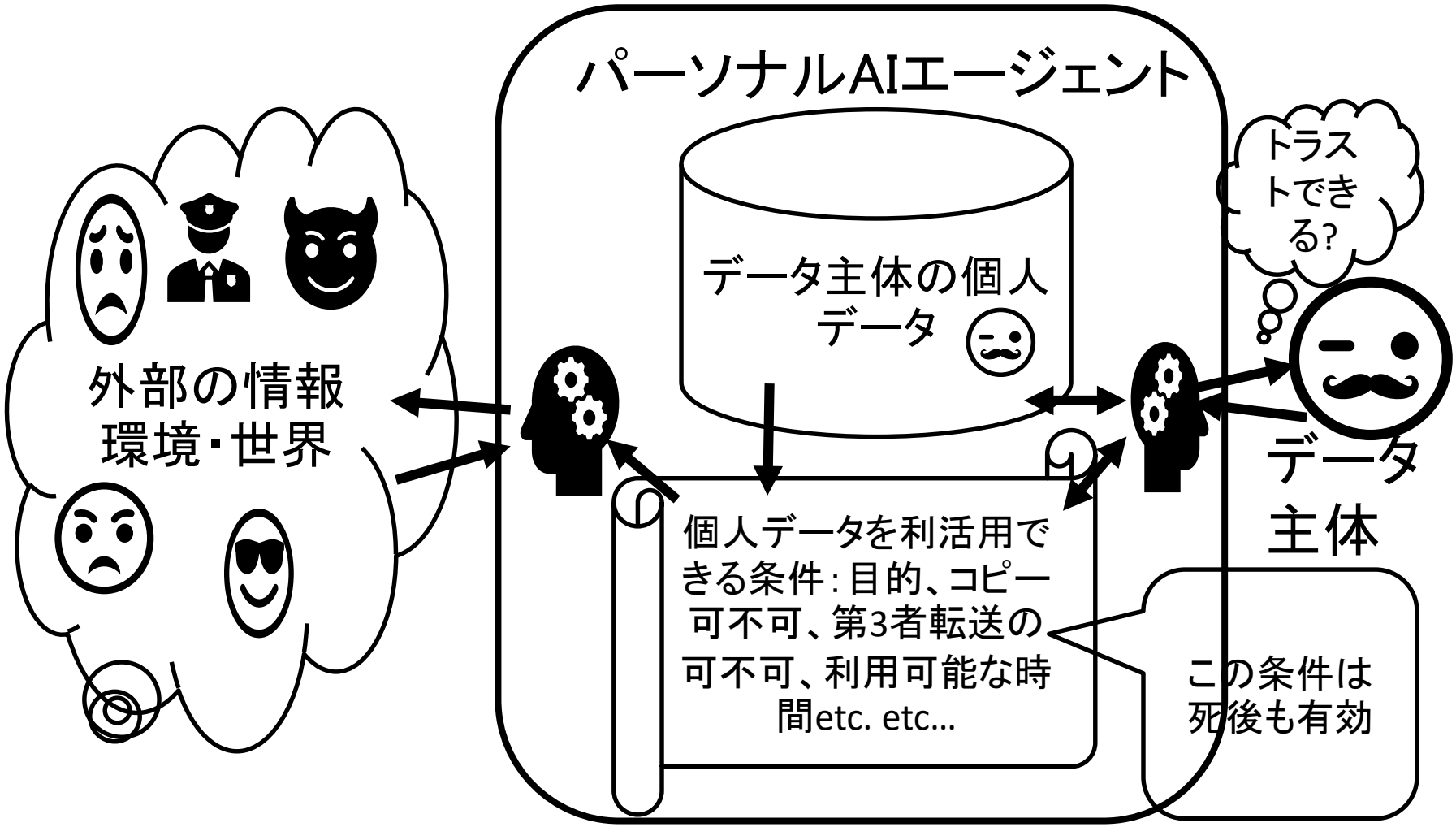
- **提案**：生身の人間には対処しきれない複雑な情報環境をパーソナルAIエージェントが支援し、人間の情報能力を増強することで対処する

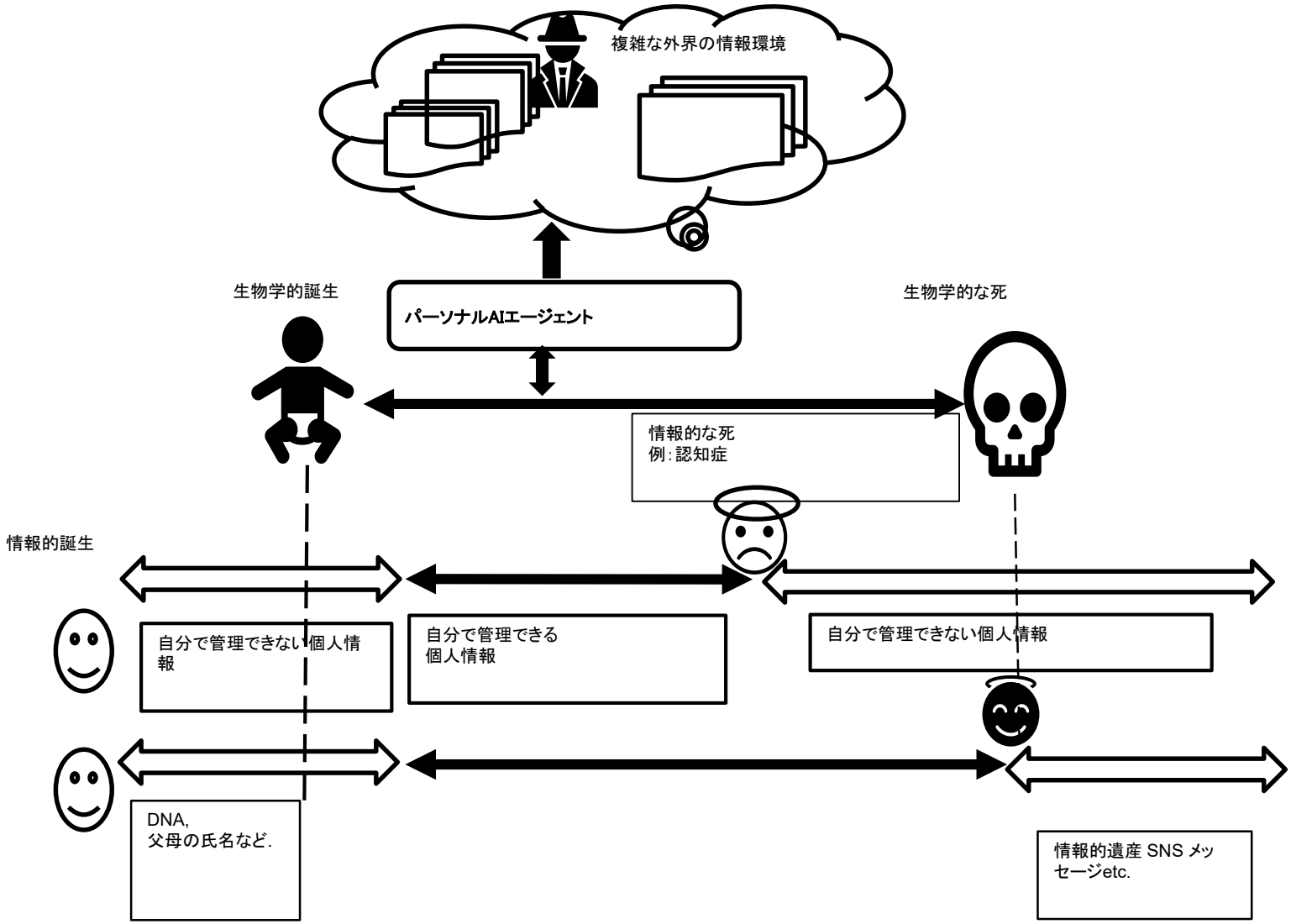
- こうして(3)に近づこうとする枠組みが民主主義国家に住む人々とっては最も受け入れやすくかつWell beingに資する。

* IEEE EAD ： Personal Data Agent

パーソナルAIエージェント*
- 誕生から死まで継続的にサポート-




パーソナルAIエージェント
外部の情報
環境・世界
データ主体の個人
データ
トラスト
トでき
る?
データ
主体
個人データを利活用で
きる条件：目的、コピー
可不可、第3者転送の
可不可、利用可能な時
間etc. etc...
この条件は
死後も有効



個人データは自分で管理できない期間のほうが長い