2020/5/21
サイバー犯罪に関する白浜シンポジウム

人間中心のAIの倫理

中川裕志
(理化学研究所　革新知能統合研究センター
情報ネットワーク法学会　理事長
東京大学　名誉教授)

概要

- AIは依然として人間が扱うツールであるにもかかわらず、その高機能化とブラックボックス化は進行している。
- また、AIが直接、間接の要因となって人間社会に対して被害を及ぼしている事案も頻発している。
- この状況において、AIが行ってはいけないこと、あるいはAIが行うべきことを指針としてまとめたAI倫理指針が2017年から国内外で公開されてきている。
- ここでは、まず公開されている主要なAI倫理指針を概観し、各々の指針の狙いや時間的変化を説明する。
- 次に具体的事案に対してAI倫理ないしAI技術の観点からの展望と対策などについて説明する。

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
- IEEE Ethically Aligned Design, first edition (2019/3)
  - 倫理的なAIの設計指針

国内外の組織が提案している人工知能の倫理

- EU: High Level Expert Group: Ethics Guidelines for Trustworthy AI (2019/4/8)
  - 倫理的なAIの設計指針

- Guidance for Regulation of Artificial Intelligence Applications:
  - USA Whitehouse. MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES (Draft 2019/4/24)

- Recommendation of the Council on OECD Legal Instruments Artificial Intelligence
  - OECD 閣僚理事会承認 (2019/5/22)

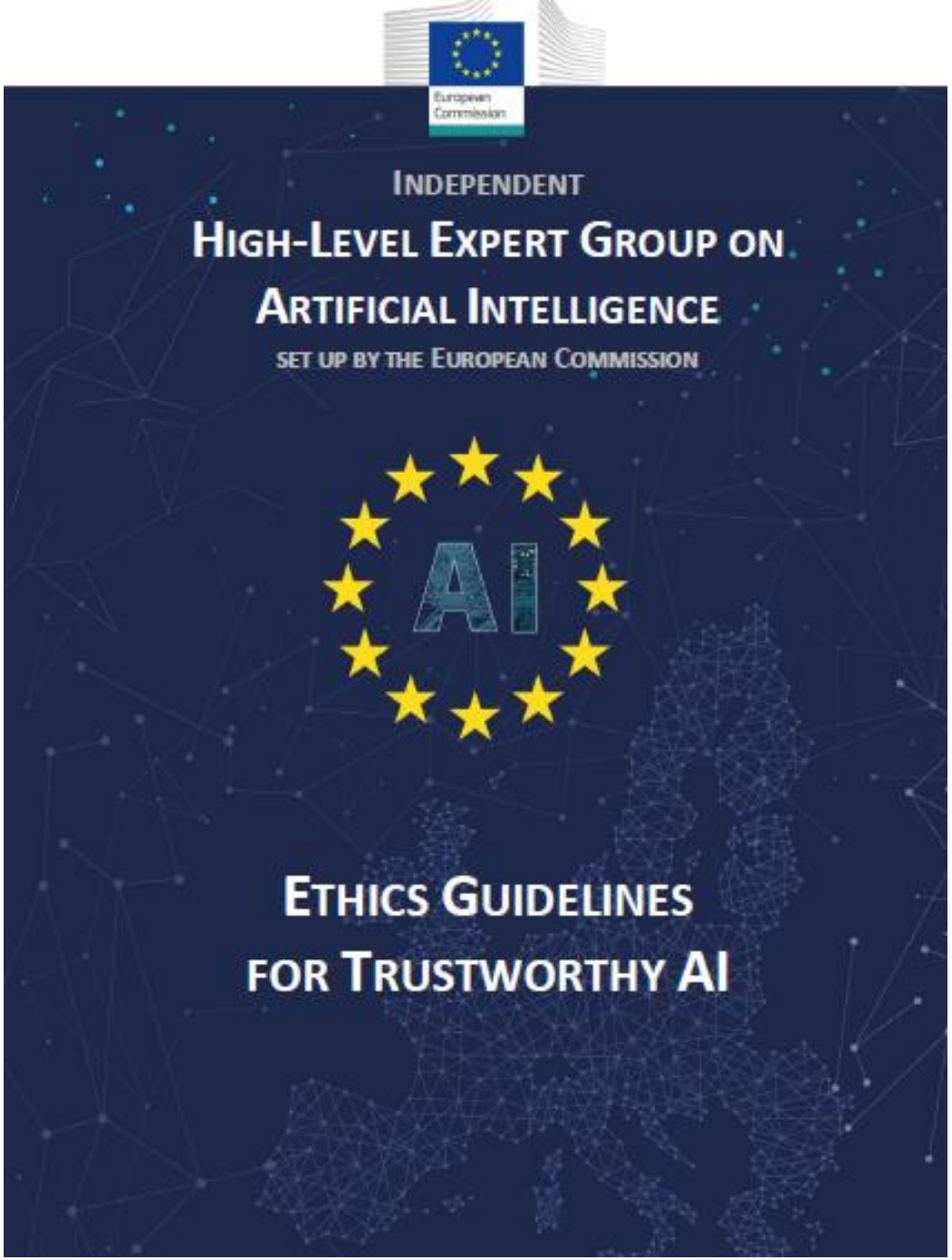
- Beijing AI Principle (2019/5/25)

- EUROPEAN COMMISSION：White Paper on Artificial Intelligence A European approach to excellence and trust, Brussels, 19.2., 2020.

IEEE Ethically Aligned Design version 2

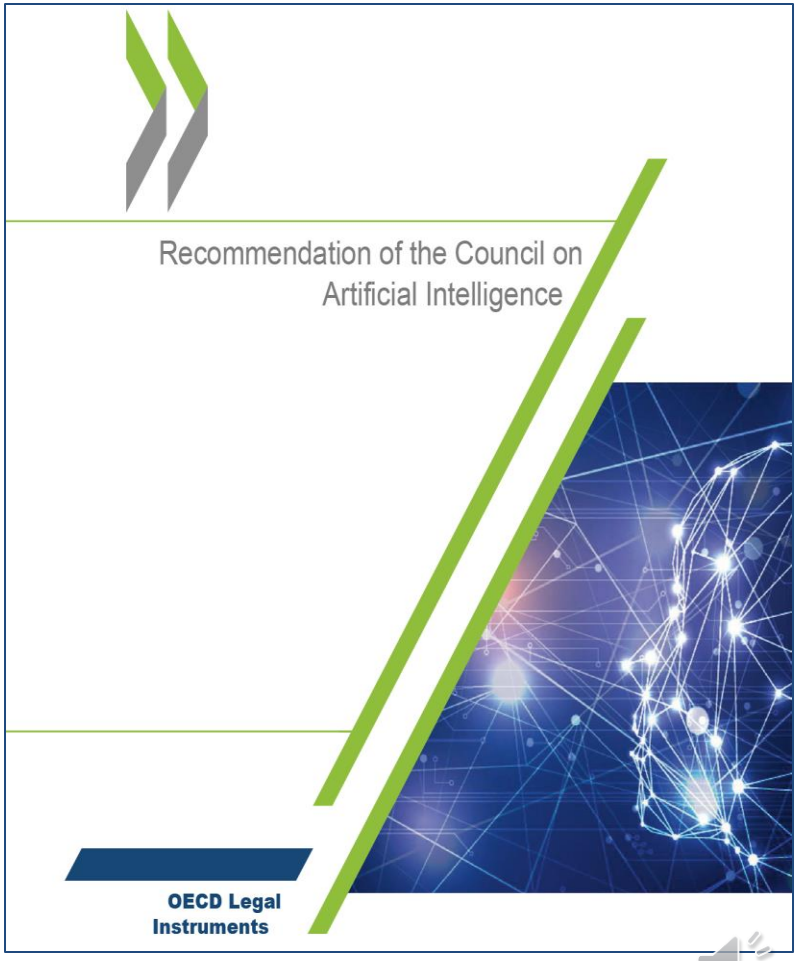
1. Executive Summary
2. General Principles
3. Embedding Values Into Autonomous Intelligent Systems
4. Methodologies to Guide Ethical Research and Design
5. Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI)
6. Personal Data and Individual Access Control
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


European Commission
INDEPENDENT
HIGH-LEVEL EXPERT GROUP ON
ARTIFICIAL INTELLIGENCE
SET UP BY THE EUROPEAN COMMISSION
AI
ETHICS GUIDELINES
FOR TRUSTWORTHY AI

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



	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		○	○	○	○	○	○	○	○	○	○	○			○	○
人間中心AI社会原則		○	○	○	○	○	○	○		○	○	○	○			○

	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
Trustworthy AI		○	○	○	○	○	○	○	○	○	○	○	○	○	△	○
OECD Recommendation		○	○	○	○	○	△	○		○	○		○			○
総務省AI利活用ガイドライン			○	○	○	○	○	○		○						○
Beijing Principle	○	○		○	○	○		○		○		○	○		△	○
Whitehouse Guidance	×		○	○	○	○	○	○		○			○		△	○
EU Whitepaper		○	○	○	○	○		○		○		○	▲		◎	

法的位置づけ：○＝AI人格権、△＝AIの現行法への適法性

▲　EU域内優遇的

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.**
  - **許容不可能なリスク、期待される有益さより大きなコストを払うリスクをきちんと評価すべき**
- Agencies should be transparent about their evaluations of risk and re-evaluate their assumptions

EUROPEAN
COMMISSION

Brussels, 19.2.2020
COM(2020) 65 final

WHITE PAPER

On Artificial Intelligence - A European approach to excellence and trust

- promotes the respect of

基本的人権、人間の尊厳、多様性、許容性、非差別、プライバシーと個人データ保護

  - このような価値観を全世界に押し出すべき

- **既存のEUないしEU域内国家の法律を効果的に適用**

- **AI応用において発生する可能性のあるリスクを洗い出せる手段をフル活用**
  - including using the experience of the EU Cybersecurity Agency (ENISA) for assessing the AI threat landscape

  - 必要とあれば、既存の法制度を改善する：あくまで法律でAI技術を制御しようという立場

– **サプライチェーン（EU域外も含む）全般に対して適用する方向**

- ソフトの更新、AI製品が実利用で学習する場合のリスクもターゲット

- AI システムはそのライフサイクルを通じて監視すべき

- AIシステムの結果は、事前のレビューと妥当性確認を人間が確認する以前は無効

  – ただし、結果が有効になった後も、人間は常に介入できなければならない

- **AI システムの稼働中の監視、そして必要に応じて介入、停止を人間ができること**

- こういった制約を設計段階で組み込むこと

- 不味い点が見つかったら、その訂正や再学習はEU域内で行うこと
  - in case an AI system does not meet the requirements for example relating to the data used to train it,

- 法的規制範囲外の動作は**EU域内で通用するベンチマーク**でテストしなければならない

USA とEU の比較： EU

- AIサービスは事前に徹底的なリスク予測を行うべき
- リスク発生の予測はAI技術に複雑さや発展の早さから技術的に抑えることは困難であることも意識

- AIシステム製造の**サプライチェーンの各段階**で倫理指針ないしは法制度に基づくリスク管理や公平性，非差別性を徹底することを求めている

USA とEU の比較： EU

- AIシステムが実利用において再学習によって動きが変化する場合，開発者側はその都度リスク管理，公平性などを確認することを求めている.

- つまり，AIシステムの仕様を倫理指針，法制度で管理しようとする指針を打ち出している. 開発者にとっては非常な重荷になると予想される.

USA とEU の比較： EU

- 以上のEUの規則主導の指針はUSAの市場評価主導の方針と全く異なる

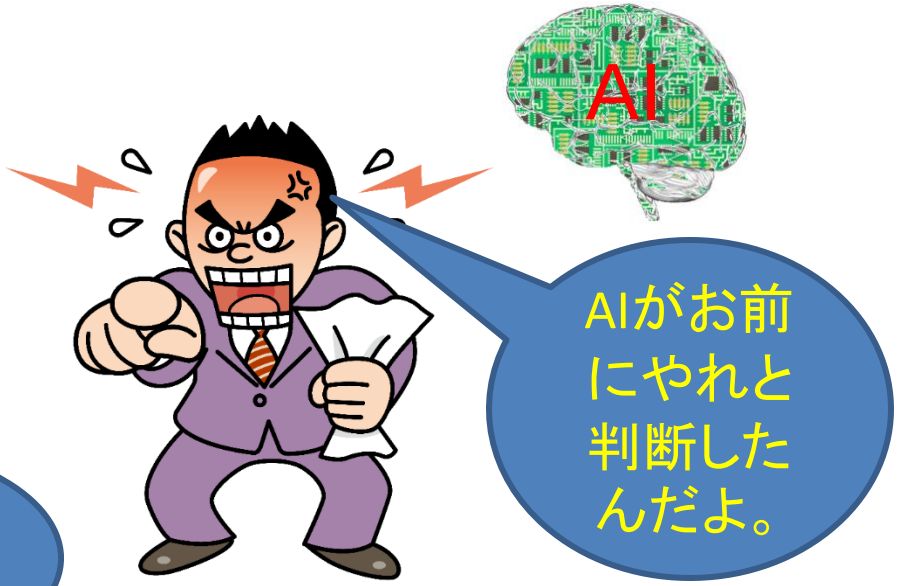
- EUはさらにEU域内で使われるAIシステムはEU域外の開発者が作ったものでも，EU域内においてテストすることを義務付け
  - 保護主義的な傾向が感じられる.
  - EUでは大企業中心の高リスク分野の他に中小規模開発者の保護や支援を強く打ち出していることにも特徴がある.

少し話題を変えます 

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？

こんな無理な仕事、なんで私がしなきゃいけないの？



- 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。

IEEE EAD ver.2　の指針

- 社内の悪用への対策として
  - 内部告発を可能にする匿名化告発（技術　Tor　など）
  - 内部告発者が不利にならないように社内規則ではなく、より上位の国の法制度で内部告発者を守る
  - 告発者を経済的に支援する保険制度

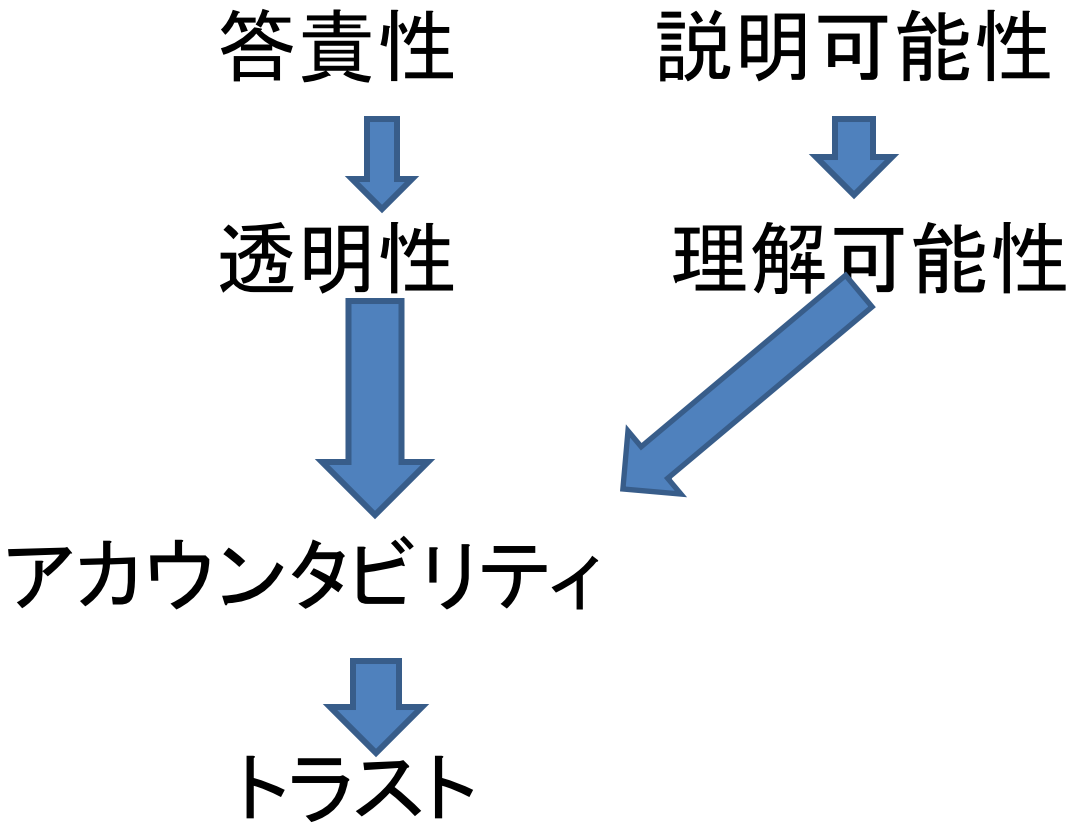
- 告発者が裁判などで戦えるために
  - AIのブラックボックス化を排除→ XAI
  - AIシステムの透明性
    ＋アカウンタビリティ

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



トラストはどうやって確保するか？

- 過去の学問的成果の集積を信用してもらう
  - 数学、物理学、医学、…
- 専門家をトラストしてもらうライセンス制度
  - 専門家のスキルのトラスト：医師国家試験、司法試験など
  - ライセンスする側（国など）へのトラスト
- それでも事故は起きる
  - 補償制度の確立：保険など
- これら全部を統合したシステム体系がトラストの基本

Trust(信頼)

- （1）能力、誠実さ、および予測可能性を扱う一連の特定の信念
- （2）危険な状況下で、ある当事者が他の当事者に依存する意思

  （証明したわけではないが…）

- 「信頼」はAIシステムのライフサイクルに関与するすべての人々とプロセスに帰することができます。

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
- 認証されたIDで金融、政府、通信などのサービスを受けることを保証するAI

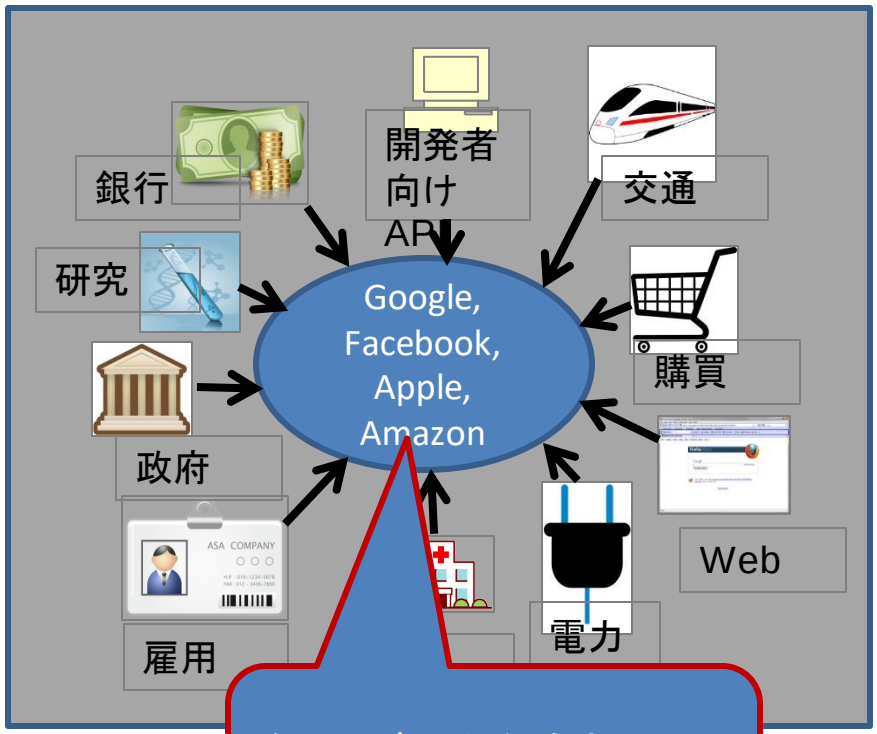
個人は信頼できるID認証にアクセスできること

- 認証はインターネットにおける個人の存在の証拠
  - 対面認証でないので、技術的な問題が多い。
  - 特に生体認証の危険性
  - 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
    - 必要最小限の個人データによる認証：比例原則

トラストする対象

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

個人データは個人が管理したほうが質が良く、データ自体をトラストできるのではないか？



パーソナル・データ・ストレージ（PDS)

- パーソナル・データ・ストア／ボールト
- あるいは
- パーソナル・データ・クラウド

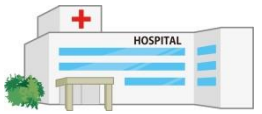







個人データ
個人データ
パーソナル・データ・ストレージ
個人データ
個人データ

仲介IT業者（AIを活用）
情報銀行？



ITを使ったサービス


- 自動アップロード
- 個人キーで暗号化
- 個人ID認証
- API-of-Me

- 利用ログ
- 流通経路トレース
- 統一データ形式
- ポータビリティ

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、

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。

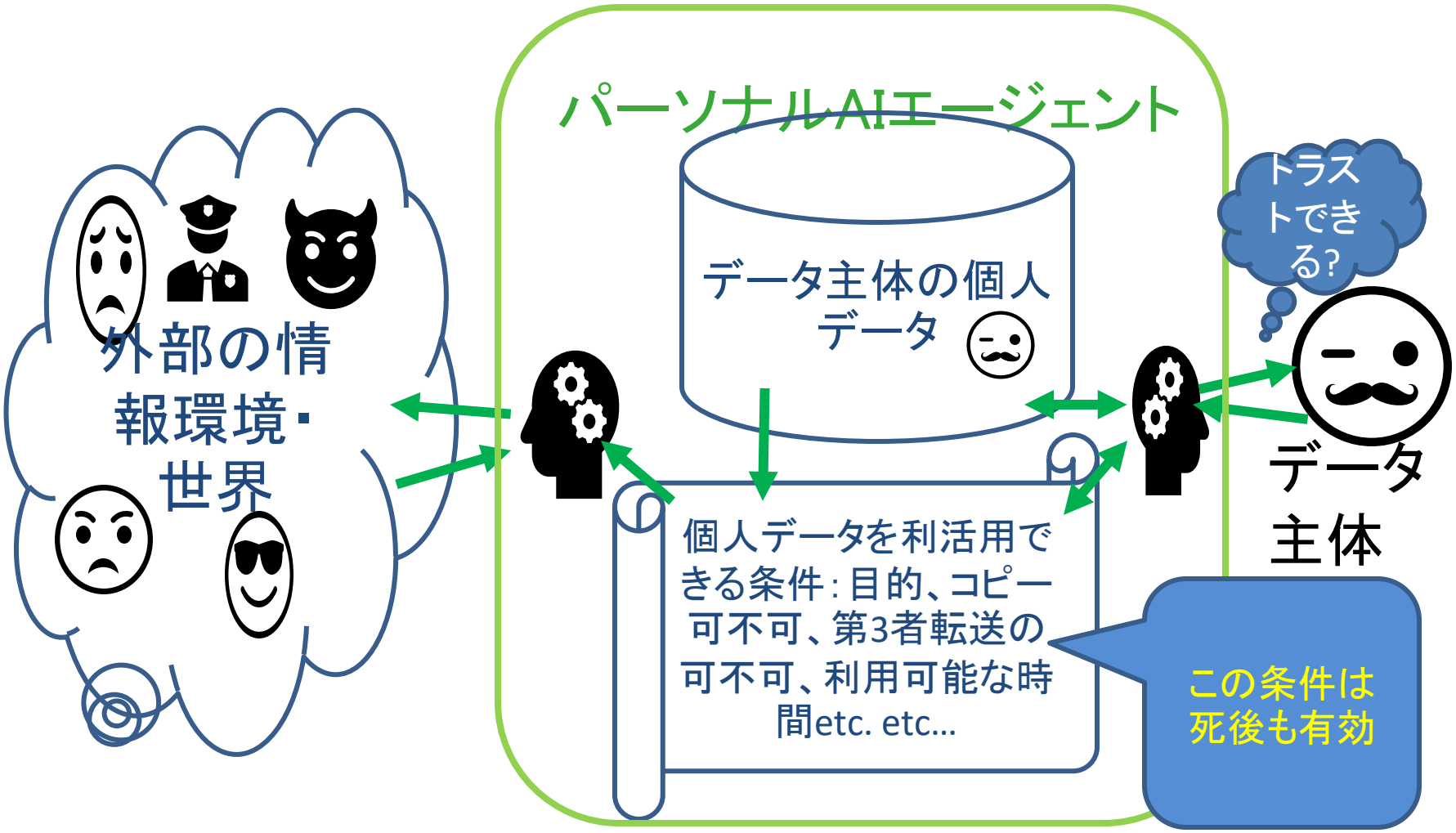
* IEEE EAD ： Personal Data Agent

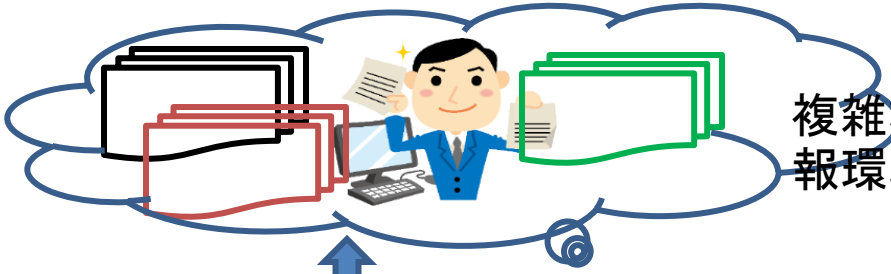
パーソナルAIエージェント*
- 誕生から死まで継続的にサポート-

一般人には複雑すぎる情報環境

- パーソナルAIエージェント

私をトラストできますか？


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
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
トラストできる?
データ主体



複雑な外界の情報環境

生物学的誕生

パーソナルAIエージェント

生物学的な死

情報的誕生

情報的な死
例：認知症

自分で管理できない個人情報

自分で管理できる
個人情報

自分で管理できない個人情報

DNA,
父母の氏名など.

情報的遺産 SNS メッセージ etc.

死後の個人データの諸問題

- 著作人格権
  - 一身専属性
  - パブリシティ権
  - 著作隣接権（儲けるネタ）

- 倫理的感覚
  - 死者を冒涜 （AI美空ひばり問題）
  - 死者の知的能力を備えたアバター
    - 不変
    - 可変　→ 生存者とやり取りすることによって変化、成長
      - ゲーム能力など

ご清聴ありがとうございました