2019/12/19
総務省　情報通信法学研究会・ＡＩ分科会

AI倫理とエージェント

中川裕志
(理化学研究所・革新知能統合研究センター)

目次

- 国内外の組織が提案している人工知能の倫理
- 弱そうに見えるAIの脅威
- パーソナルAIエージェント
- 付録

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 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

- 総務省　AIネットワーク社会推進委員会(2018)
  - AI開発ガイドライン　OECDに提案

- ICDPPC（40th International Conference of Data Protection and Privacy Commission 2018)

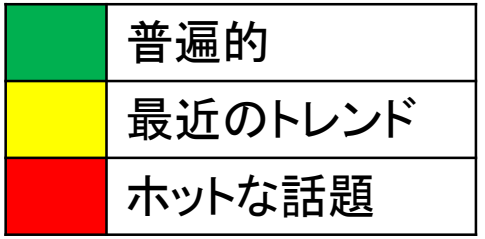
国内外の組織が提案している人工知能の倫理

- 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

- 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)



古い → 新しい

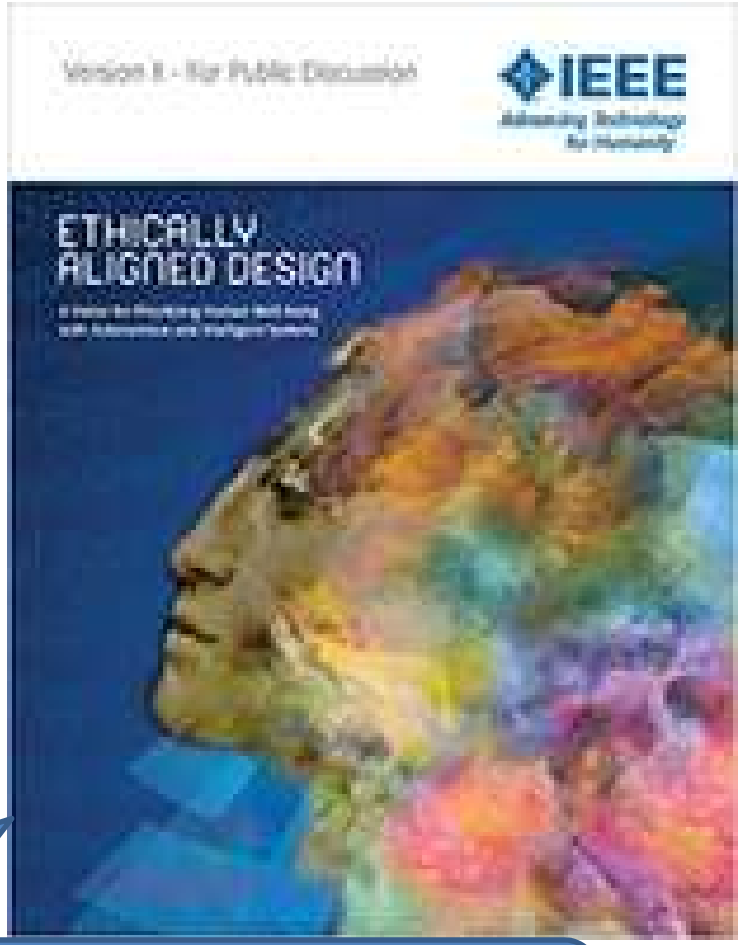
	FLI Asilomar 23	人工知能学会倫理委員会	ICDPPC	IEEE EAD v1,2	IEEE EAD e1	人間中心AI	EC Trustworthy AI	OECD AI Recomm.
人権								
公平性、差別								
法令遵守								
透明性								
アカウンタビリティ								
トラスト								
自律的AI・自由意志								

ICDPPC: International Conference of Data Protection Commisioners

	FLI Asiloma 23	人工知能学会倫理委員会	ICDP PC	IEEE EAD v1,2	IEEE EAD e1	人間中心AI	EC Trust worthy AI	OECD AI Recomm.
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								
軍事利用								
幸福								

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 Intelligen
12. Policy
13. Mixed Reality
14. Well-being



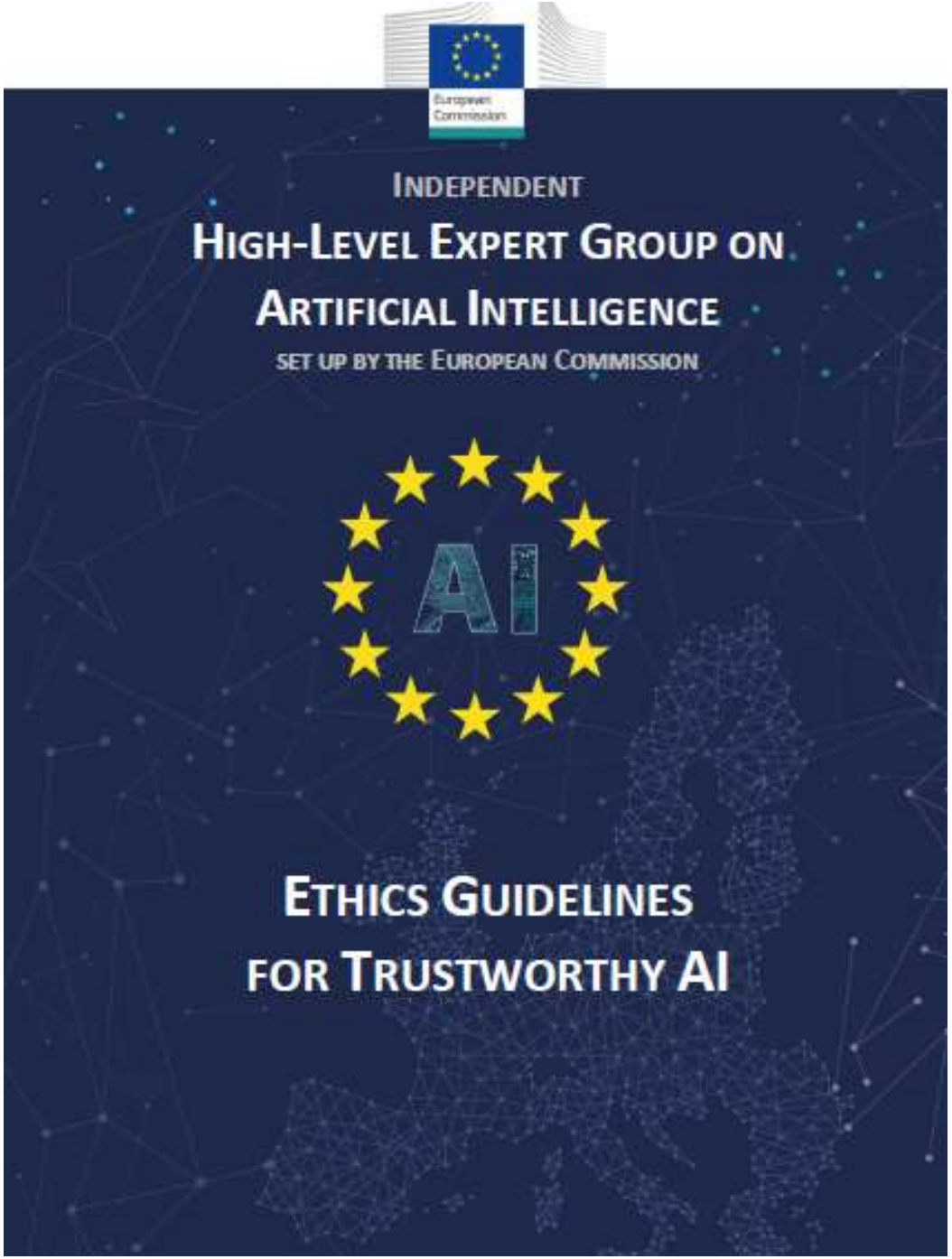
The final version was published

IEEE　EAD　（Final）　on April 2019

- 1.人権
  - 国際的に認められた人権を尊重し、促進し、保護
- 2.幸福
  - 主要な成功基準として人間の幸福の増加
- 3.データ・エージェント
  - 個人データを自分自身が管理できるする能力
- 4.有効性
  - ＡＩの開発者および運営者は、AIの目的に対する有効性および適合性の証拠を提供する

IEEE　EAD　（Final）　on April 2019

- 5.透明性
  - AIの決定の根拠の提供
- 6. アカウンタビリティ
  - 行われたすべての決定に対して明確な根拠を提供（＝説明＋責任者）
- 7.　誤用の認識
  - 潜在的な誤用およびリスクから保護する設計
- 8.能力
  - 運用者は安全かつ効果的な運用に必要な知識とスキルを遵守しなければなりません。


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

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。

EC HLEG Ethical Guidelines for Trustworthy AI

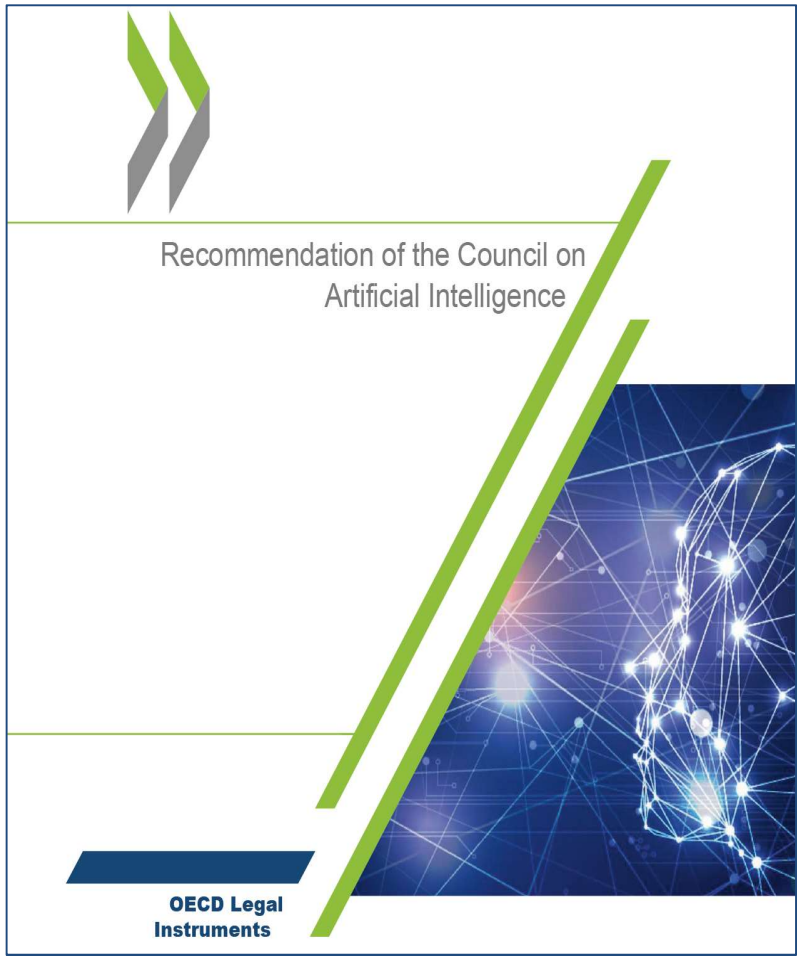
- 合法的、倫理的、ロバスト

- 必要条件

1. 人による見通しと査察
2. 技術的なロバスト性と安全性
3. プライバシーとデータガバナンス
4. 透明性
5. 多様性、非差別と公平性
6. 社会的および環境的幸福
7. アカウンタビリティ

Recommendation of the Council on OECD Legal Instruments Artificial Intelligence

- Approved on May 22, 2019 by OECD Councuil.
- No forcing power but strong guideline.
  - Ex. OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data （1980） was a baseline of privacy protection laws of many countries.



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

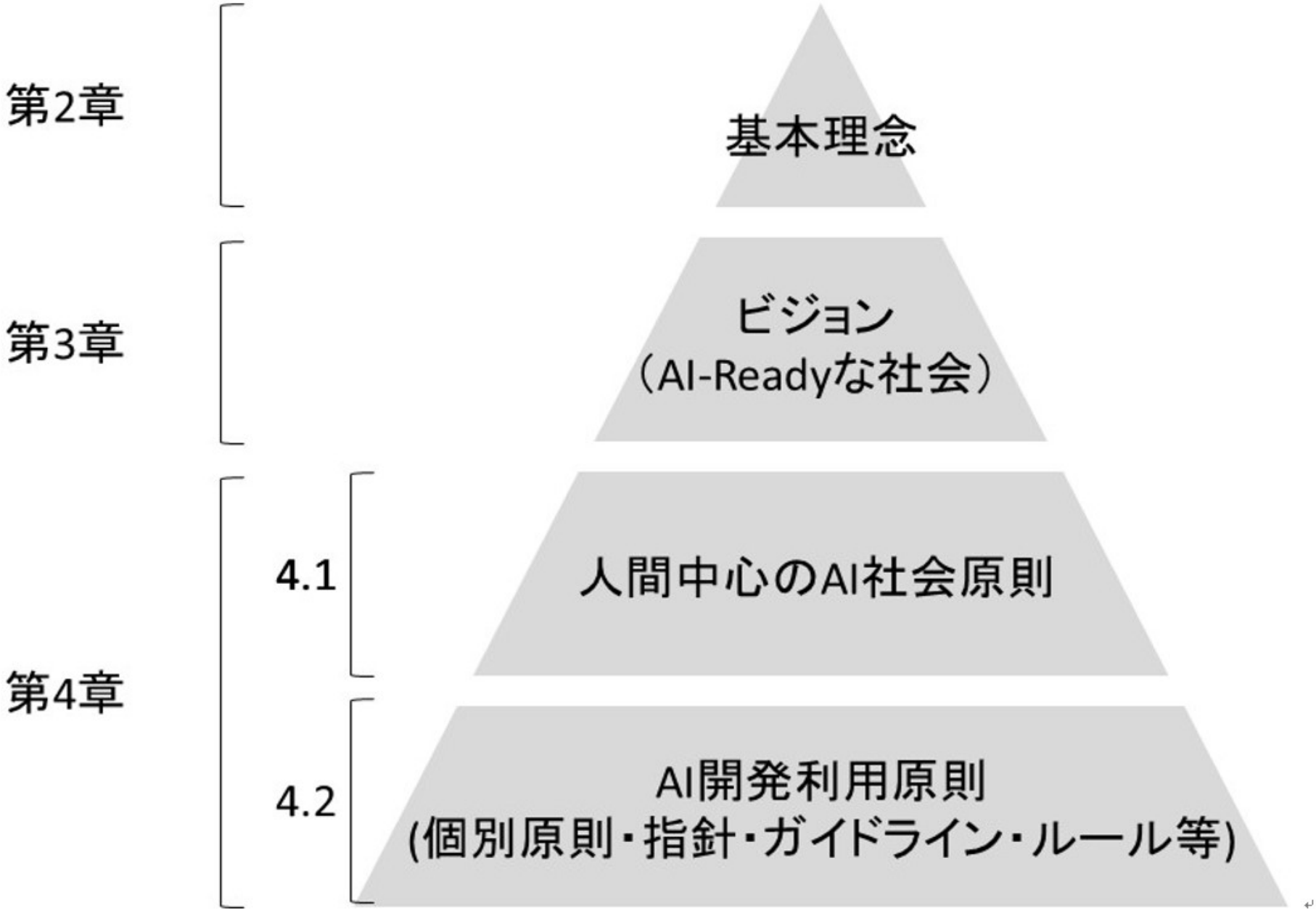
- 技術的:
- インクルーシブな成長、持続可能な開発、そして幸福
- 人間中心の価値観と公正さ
- 透明性と説明可能性
- ロバスト性、セキュリティ、そして安全性
- アカウンタビリティ

- 非技術的:
- AIの研究開発に投資する
- AIのためのデジタルエコシステムの育成
- AIを可能にする政策環境の形成
- 人的能力の構築と労働市場の変革への準備
- 信頼できるAIのための国際協力

人間中心のAI社会原則

人間中心のAI社会原則会議

内閣府


第2章
第3章
第4章
4.1
4.2
基本理念
ビジョン
（AI-Readyな社会）
人間中心のAI社会原則
AI開発利用原則
(個別原則・指針・ガイドライン・ルール等)

- Dignity、
- Diversity　＆　Inclusion、
- Sustainability

- Society 5.0 → 「AI-Readyな社会」
  - 「人」、「社会システム」、「産業構造」、
  - 「イノベーションシステム(イノベーションを支援する環境)」、
  - 「ガバナンス」

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

寄り道：弱そうに見えるAIの脅威

- 現在はまだまだ弱そうに見えるAIも多数が共謀して悪さをするかもしれません。
  - 弱いAIたちが共通の言語を持てば、共謀が可能
  - しかも、人間は気づかないかもしれない
  - 例：フラッシュ・クラッシュ
  - 　　クウォンツ・クウェーク

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。

ブラックボックス化＋ネット接続の行き着く先は？

- つまり、1個の人工知能では小さな影響しかないようでも、
- ネット経由でデータが交換されると、多数の人工知能が制御不能な動きをしてしまうことが実例として存在するわけです。

- 現在の人工知能の倫理はえてして、単独の人工知能が倫理的、道徳的に行動するかという視点で語られています。

- しかし、ネットワークで接続された多数の人工知能たちがかってに動き出すと、
- 制御不能かつ収拾不能になりそうです。

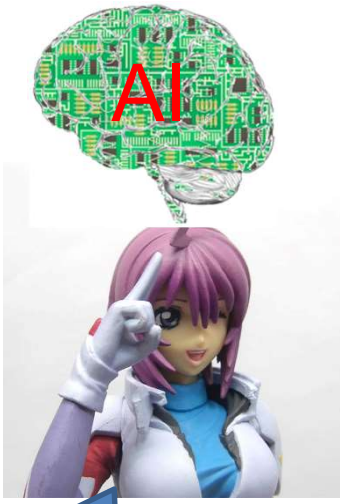
Outside observer AI

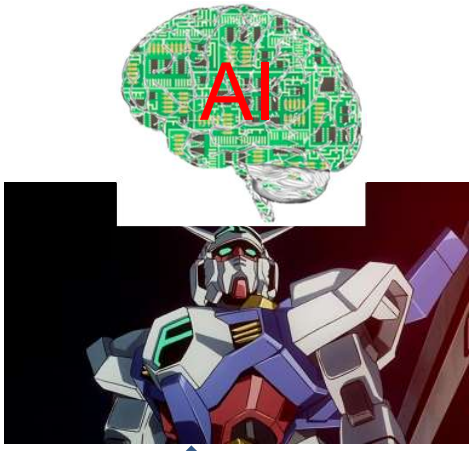
- 許容損害値超過を早期発見するチェック機構が必要

- 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

AIの行動をAIが観察してテスト

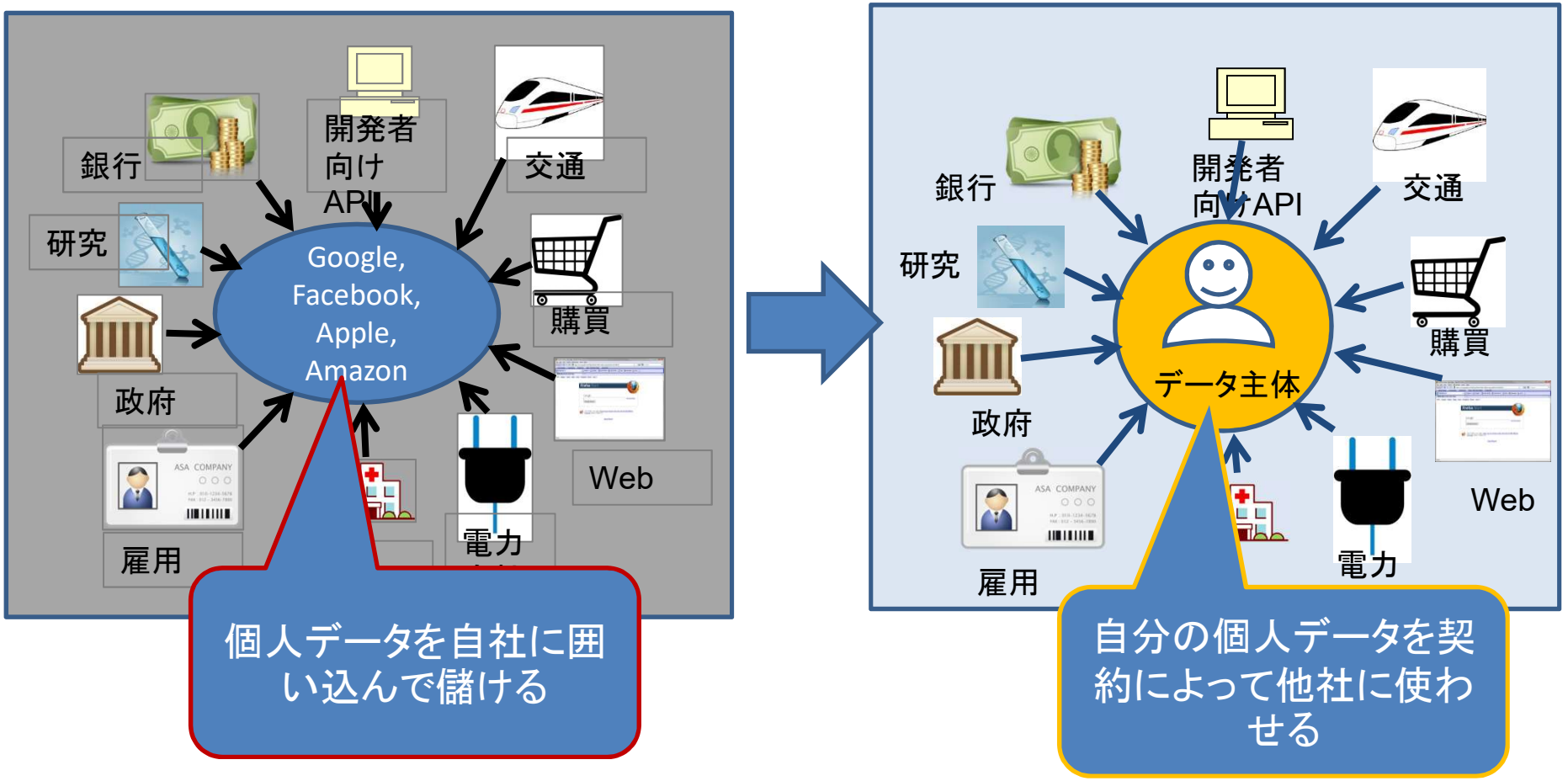




このAIが正しく機能しているかどうかを人間がtrust（信用）できる必要がある

分散型かつチェック機構が独立したAI群からなるシステム

本筋に戻ります：
個人データ管理はデータ主体の個人へ



背景：IT企業と個人データ

- 米国のIT企業GAFA: Google Amazon Facebook Apple がパーソナルデータをどんどん収集して囲い込み、利益を上げている現状

- 収奪されるEU、収奪されるデータ主体の個人
- GDPRで反撃しているが、それだけではEUの産業は育たない
- EUの個人データのプライバシー（＝人権）の危機。だが、産業は興さないと低落するのみ

- 個人データはデータ発生源であるデータ主体の個人が管理
- その枠組みの標榜と、ビジネス育成がテーマ

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

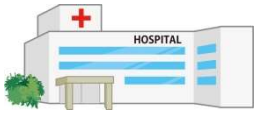







個人データ
個人データ
パーソナル・データ・ストレージ
個人データ
個人データ

仲介IT業者（AIを活用）



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 による個人の Identity 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
- サービスごとの契約による利用許可

- 上記技術の実装と、AIによる使い易い柔軟なインタフェースが必須
  - マネタイズ可能なビジネスモデルを確立できるかがカギ

個人を信頼できるID認証

- 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ができるような個人認証

- 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

- 認証に必要な個人データは最小限にしたい
  - SSI：Self Sovereign Identity

- MyData (Global)
  - 2016年よりヘルシンキで毎年9月に開催

- MyData Japan
  - https://mydatajapan.org/

- 情報銀行
  - IT連が「情報信託機能の認定に係る指針ver1.0」を平成30年6月に公開
  - 参照
  - 興味を示している企業はある（銀行など）
  - ◆Data Portabilityに触れていない。

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。

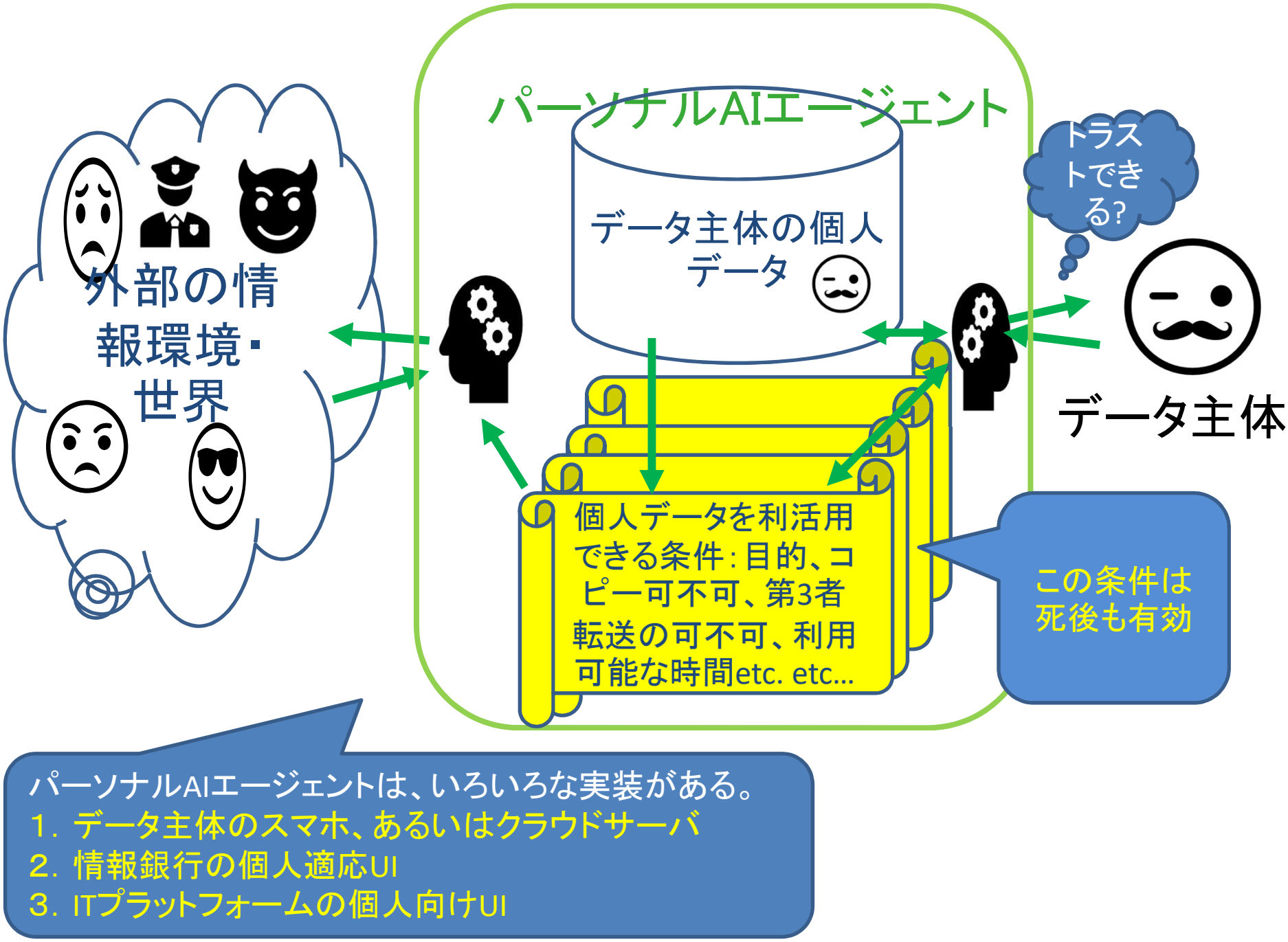
* IEEE EAD ： Personal Data Agent

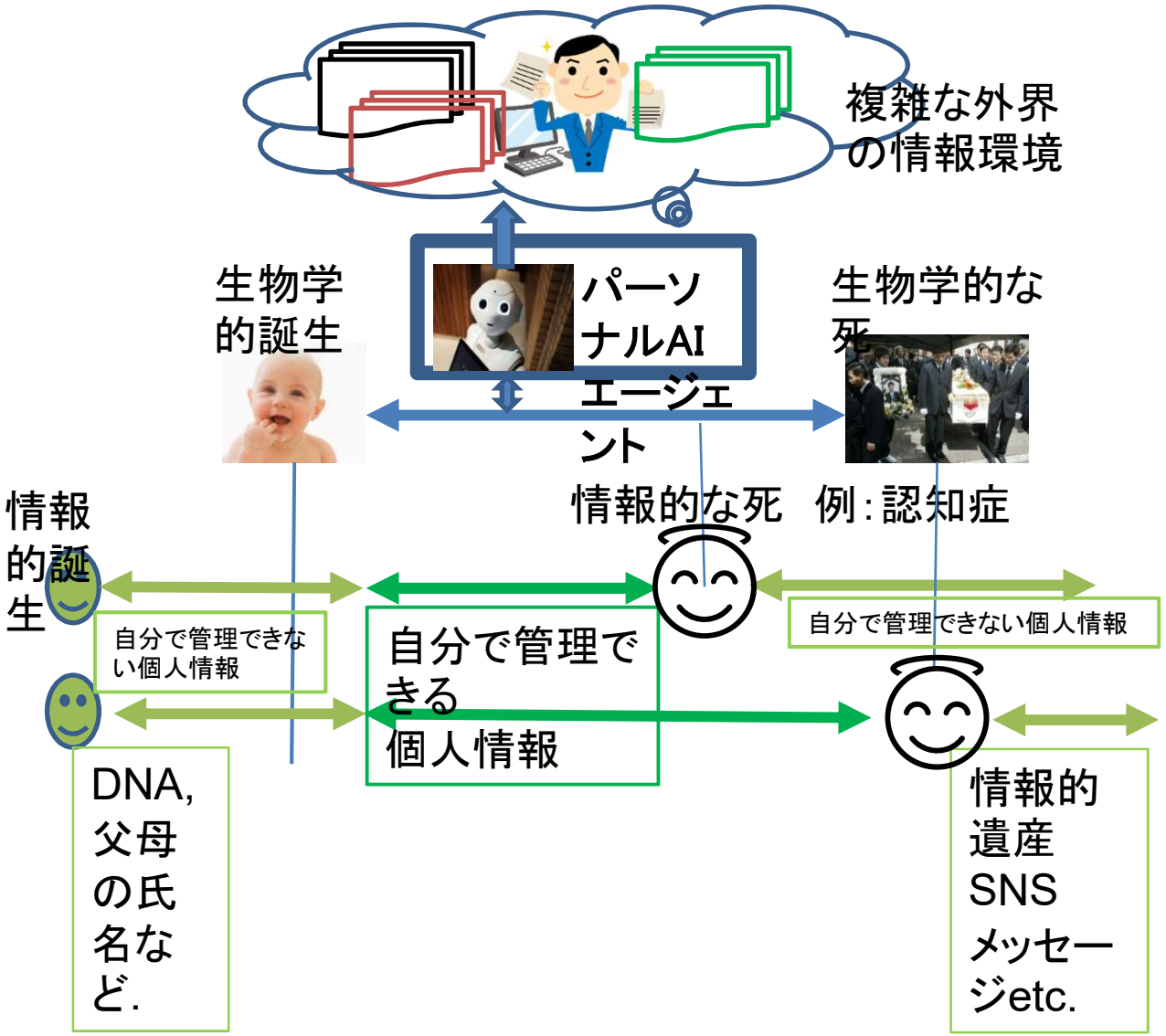
パーソナルAIエージェント*
- 誕生から死まで継続的にサポート-

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
データ主体の個人データ
外部の情報環境・世界
トラストできる?
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
パーソナルAIエージェントは、いろいろな実装がある。
1. データ主体のスマホ、あるいはクラウドサーバ
2. 情報銀行の個人適応UI
3. ITプラットフォームの個人向けUI


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
情報的な死
例：認知症
情報的誕生
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.

今後の検討課題：パーソナルAIエージェントの拡張

- 個人のパーソナルAIエージェントから

➢ グループのパーソナルAIエージェントへ拡張
  – 家族、仕事仲間、趣味仲間、などなど

➢ 自宅の家のパーソナルAIエージェントへ拡張
  – 自宅の在-不在
    – 宅急便の配達対応に使えるかも
  – 電気、ガス、水道のIoTメータ、IoT家電

➢ 所有する自動車のパーソナルAIエージェントへ拡張
  ➢ いろいろな使い道がありそう

ご清聴ありがとうございました

付録：AIの著作権と知的財産権

- AIの助けを借りて作成された芸術等の様々な著作権
- AIの支援で創り出された知的財産権

- 著作権/知的財産の所有者は誰か？
- 2人以上の当事者が彼らの権利を主張するとき、AIはどのような役割をすべきか？

- AIの説明可能性とアカウンタビリティは重要な役割を果たす

AIの創作活動　:創作活動にコミットしたのは誰？

＊知財の場合は過去に類似したものがないことを確認となる

AIによって、くまなく創作物候補を生成

AIによって評価の高い創作物に類似＊し、ヒットしそうな創作物を選別

AIを利用して作った創作作品

このループで考える必要あり

鑑賞者たちの思想、感情:著作権法2条

過去の評価の高い創作物の選択　(by AI)

過去の評価の高い創作物

鑑賞者たちの反応＝鑑賞者たちの思想、感情（著作権法2条）

末端の鑑賞者

AIの創作活動の位置づけ
知的財産作成にコミットしたのは誰？

＊知財の場合は過去に類似したものがないことを確認となる

AIによって、くまなく知的財産候補補を生成

AIによって評価の高い知財に類似＊しているが、明らかに異なる点を持つ創作物を選別

過去の評価の高い知財選択　(by AI)

このループで考える必要あり

AIを利用して作った知財

知財審査官が審査
（実はここもAI)

知財権侵害権侵害の訴え

過去の評価の高い知財など

末端の利用者：儲け

紛争

◆AIを使う芸術家、あるいはAI自身は自律的か？

- 著作権侵害の場合、誰がそれに責任を負うべきか？

- 自律AIそれ自体
- 自律AIの開発者