総務省 学術雑誌『情報通信政策研究』 第３巻第２号
Journal of Information and Communications Policy Vol.3 No.2

寄稿論文

AI 倫理指針の動向とパーソナル AI エージェント

中川　裕志[1]（理化学研究所）

要　旨

シンギュラリティによって人間と同じような知的能力を持つ AI が出現し、人間への脅威になりかねないという言説が流布した。これによって、AI にも倫理を守らせようという機運が高まったという状況もあってか、2016 年ころから AI 倫理指針の作成と公開が盛んになった。本論文では、2017 年から 2019 年にかけて国内外で公開された多数の AI 倫理指針のうち、影響力の大きな主要な指針に関して、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、幸福などの倫理的テーマを各 AI 倫理指針がどのように扱ってきたかをまとめた。種々の AI 倫理指針の公開の時間順序と合わせてみれば、AI 倫理の内容の変遷を探ることができ、同時に AI 技術、AI 応用システムの開発を行うにあたって留意すべき点が明らかになる。また、これらの指針が誰を対象に起草されているか、すなわち名宛人を考察することによって、AI 倫理指針を作成した組織の意図が見えてくる。

次に、AI 倫理指針のうち IEEE EAD ver2、1e で提案された個人データの収集、管理、保護をおこなう代理ソフトウェア、すなわちパーソナル AI エージェントの概念設計について述べる。これは、データ主体本人の個人データとその利用条件の記述されたデータベースであるので、これをデータ主体の死後に残されたディジタル遺産の管理に適用する場合の検討課題について述べた。

キーワード：AI 倫理、アカウンタビリティ、AI エージェント、パーソナルデータ、人間中心

１．AI 倫理指針の背景

2005 年刊行のカーツワイルの『ポスト･ヒューマン誕生』[2]が示した人間の知的能力を凌駕する汎用人工知能（以下 AGI と略記）、そして 2014 年刊行のボストロムの『Super intelligence』[3]は、人工知能（以下、AI と略記）が発展して AGI、さらに強力な超知能にたどり着いたとき、人類にとってとてつもない脅威になるという言説を、AI 関係者をはじめとする多くの人々の間に流布させた[4]。そのような背景から AI は野放図に開発すると脅威になるという感覚が生まれ、AI を人間の制御下に置くための方策として AI 倫理の議論が

[1] 理化学研究所・革新知能統合研究センターPI
[2] R. カーツワイル：『ポスト･ヒューマン誕生』，NHK 出版，2005
[3] N. Bostrom: Super intelligence, Oxford University Press.2014
[4] もっともボストロムは人類を支配するような Super intelligence が実現する確率は非常に低いと言っている。ただし、確率が低くても起きてしまったら人類には手の施しようがないので、今から注意を怠ってはいけないと説いている。

盛んになったと考えられる。

例えば、初期の AI 倫理指針として知られる FLI の Asilomar Principles[5]の項目 10 には「自動的な AI システムは、目標と行動が倫理的に人間の価値観と一致するようデザインする」と要求し、さらに項目 19 には「一致する意見がない以上、未来の AI の可能性に上限があると決めてかかるべきではない」と警告し、項目 22「あまりに急速な進歩や増殖を行なうような自己改善、または自己複製するようにデザインされた AI は、厳格な安全、管理対策の対象にならなければならない」、項目 23「超知能は、広く認知されている倫理的な理想や、人類全ての利益のためにのみ開発されるべきである」と多数の項目において AGI や超知能への警告および人間の制御下におくことを要求している。

また、IEEE Ethically Aligned Design version2 [6]（以下では IEEE EAD ver2 と略記）では、Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence(ASI)、という章(pp.73-82)で AGI や超知能について考察し、safety by design という考え方を提唱している。

ただし、その後 AGI の可能性の研究が進み、カーツワイルのポストヒューマンのアイデアの分析が進むにつれて、AGI の可能性はまだまだ先のことであり[7]、まして超知能の実現性は強く疑われはじめた。このような背景から、AI 倫理の当初の動機の一つであった AGI や超知能による AI 脅威論は退潮し、代わって現在ないし近い将来において重要な AI 倫理の課題が語られるようになった。

本論文の以下の部分では、AI 倫理における課題の最近の動向、およびその結果として重視されているパーソナルデータの管理の問題について説明する。

２．AI 倫理指針の焦点の移動経緯

２．１．主要な AI 倫理指針

2017 年以降、国内、国外において AI の開発や利活用に関する多くの文書（以下では AI 倫理指針と総称する）が公開された。すべてを列挙することはとてもできないが、参照されることが多いものをおおよその公表時期の古いものから順に挙げると以下の表 1 のようになる。これらの AI 倫理指針は 2019 年末現在、Web で公開されているため容易に内容を確認することができる。脚注に URL を記した（2019 年末時点）。

表１．AI 倫理指針

名称	略称	作成した組織	公開時期
Asilomar AI Principles	Asilomar Principle	Future Life Institute	2017

[5] Future Life Institute: ASILOMAR AI PRINCIPLES, https://futureoflife. org/ai-principles/?cn-reloaded=1

[6] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: Ethically Aligned Design version2: A Vision for Prioritizing Human Well-being with Autonomous and with Autonomous and Intelligent Systems. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

[7] カーツワイルは 2045 年にシンギュラリティが起こり AGI が実現すると予言している。

人工知能学会　倫理指針[8]	人工知能学会・倫理指針	人工知能学会・倫理委員会	2017
国際的な議論のためのAI開発ガイドライン[9]	総務省 AI 開発ガイドライン	総務省・AI ネットワーク社会推進会議	2017
Ethically Aligned Design version2: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems[10]	IEEE EAD ver2	The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	2017
Ethically Aligned Design (first edition): A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems[11]	IEEE EAD 1e	The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	2019
人間中心の AI 社会原則[12]	人間中心 AI 社会原則	AI 戦略実行会議、内閣府	2019
Ethics Guidelines for Trustworthy AI[13]	Trustworthy AI	The European Commission's High-Level Expert Group on Artificial Intelligence	2019
Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449[14]	OECD Recommendation	OECD	2019
AI 利活用ガイドライン[15]	総務省 AI 利活用ガイドライン	総務省・AI ネットワーク社会推進会議	2019

表 1 に列挙した AI 倫理指針において重視されている分野ないし項目を一覧表にして表 2 にまとめた。第 2 行には表 1 の AI 倫理指針の略称を公開された時期の古いものから順に左か

[8] http://ai-elsi.org/wp-content/uploads/2017/02/%E4%BA%BA%E5%B7%A5%E7%9F%A5%E8%83%BD%E5%AD%A6%E4%BC%9A%E5%80%AB%E7%90%86%E6%8C%87%E9%87%9D.pdf
[9] https://www.soumu.go.jp/main_content/000490299.pdf
[10] https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf
[11] https://ethicsinaction.ieee.org/
[12] https://www8.cao.go.jp/cstp/aigensoku.pdf
英語版は次の URL:https://www8.cao.go.jp/cstp/stmain/aisocialprinciples.pdf
[13] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
[14] https://www.soumu.go.jp/main_content/000642218.pdf
[15] https://www.soumu.go.jp/main_content/000624438.pdf

ら右に並べた。表 1 の最左列の 3 行目から下には AI 倫理指針で言及される項目を列挙した。表中、各 AI 倫理指針で重視されている項目には○をつけている。例えば、Asilomar Principle は、AI 制御を重視していることが表されている。

表２．AI 倫理指針で重視されている項目

公開順序	古い				→				新しい
AI 倫理指針略称	Asilomar Principle	人工知能学会・倫理指針	総務省 AI 開発ガイドライン	IEEE EAD ver2	IEEE EAD 1e	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
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				○	○		△[16]		
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	○	○	○	○	○	○	○	○	○

[16] Guidelines for Trustworthy AI で述べられている human oversight と human in the loop を組み合わせると必然的にパーソナル AI エージェントになるが、直接の言及はない。

AI 技術に進展が早いため、ハードローで制約してしまうと、技術的発展を阻害するという考え方が支配的である[17]。よって AI 倫理指針は非拘束的ガイドライン[18]である。このような状況なので、AI 倫理指針の実効性の持たせ方としては標準として制定することが有力である。特に国際標準は各国の開発者への影響が大きい。代表的には ISO 標準、IEEE P7000 シリーズ（IEEE EAD ver2, 1e を思想的基盤としている）である。

本章の以下の各節で項目ごとに AI 倫理指針での重視された経緯などを説明する。

２．２．AI 制御

1 章で述べたように、AGI や超知能は人類にとって脅威だから、AI を人間の制御下におけるように AI を設計、開発しなければいけないという言説に則り Asilomar Principle や IEEE EAD ver2 ではこの項目を重視していた。しかし、すでに述べたように AGI はまだ遠い将来の話であることが理解されはじめ、むしろ現在ないし近い将来の AI で留意すべき項目が多数存在することが AI 倫理に係わる議論の中で明らかになった。よって、IEEE EAD ver2 以降では取り上げられていない。

興味深いのは表 2 で AI 制御に△をつけた人工知能学会・倫理指針である。この指針の最後の第 9 項目に

「人工知能が社会の構成員またはそれに準じるものとなるためには、上に定めた人工知能学会員と同等に倫理指針を遵守できなければならない。」

と記されている。第 8 項目までは、人工知能学会の会員に対して、会員が AI 開発において遵守すべき指針を示している。それが第 9 項目ではあたかも AI が社会の構成員となることを想定した場合、すなわち AGI の実現した状況を想定し、AGI にも第 8 項目までを守れと書いている。人工知能学会・倫理指針は Asilomar Principle と同時期に書かれているので、同様に AGI を意識した項目が入ったと推測される。

２．３．人権

人権はほとんどすべての AI 倫理指針において直接的にせよ間接的にせよ触れられている。ただし、その扱いは簡素な言及から、厚みのある記述まで様々である。多角的に人権に焦点を当てているのは IEEE EAD ver2、IEEE EAD 1e であり、この両者が大部の指針であるにしても human rights という表現の出現回数は前者で 90 回、後者で 140 回に及ぶ。また、human rights を種々の視点から取り上げており、IEEE EAD ver2 では principle、personal data、misuse、well-being、autonomous weapon などにおいて人権との関連を記述している。IEEE EAD 1e では人権は焦点を当てる 8 つの一般原則のうちの一つになっている。具体的には、AI は人権を尊重するように設計せよと宣言している。さらに、IEEE EAD 1e では 10 章全てで人権に言及している。

ちなみに、IEEE EAD 1e の 8 つの一般原則は以下の通りである。

Human Rights、Well-being、Data Agency、Effectiveness、Transparency、Accountability、

---

17 政府の委員会（総務省の AI ネットワーク推進会議、内閣府の人間中心の AI 社会原則検討会）などの議論による。

18 日本では政府諸官庁が公開する法律のガイドラインには指令的な意味合いが強い。

Awareness of Misuse、Competence である。

２．４．公平性、非差別

Fairnessあるいは公平性については表面的には多くの倫理指針で触れている。公平性を守れという言い方が多くみられるが、公平性の定義にまで踏み込んだものは少ない。例えば、Trustworthy AIは公平性と非差別(non-discrimination)を平行して記述している。ただし、公平性や非差別に直接的な寄与をするバイアスを意識した記述はTrustworthy AIでは結果のバイアスという言い方に終始しており、バイアスの原因であるアルゴリズムやデータへの直接の言及がない。OECD　Recommendationと人工知能学会・倫理指針も公平性に関しては同様の方向で書かれている。IEEE EAD 1eのLawの章の266ページには、脚注に「公平性ないしバイアスの統一的な定義を与えることは適切ではない」[19]と書かれている。

現在のAIの出力のデータ依存性の高さを鑑みれば、データバイアスはあって当然と見なせる。人間中心AIでは、データおよびアルゴリズムの双方のバイアスに言及し、これらのバイアスが公平性や非差別に影響しうる可能性に留意せよと記している。2019年に公開された総務省AI利活用ガイドラインでは、AIに利用者らが守るべき公平性の原則として以下のように明記している。

「AIサービスプロバイダ、ビジネス利用者及びデータ提供者は、AIシステム又はAIサービスの判断にバイアスが含まれる可能性があることに留意し、また、AIシステム又はAIサービスの判断によって個人が不当に差別されないよう配慮する。」

個人データの種々の属性、例えば性別、人種、年齢などのうち、公平に扱うべき属性が決まれば、AIによる判断が、それらの属性に対して公平であるように制約をかける方法はすでによく研究[20]されている。ただし、公平性はその確保のためにアファーマティブアクションのような方法を用いると、逆差別も起こしやすい[21]。IEEE EAD 1eではこういった困難さを自覚して、抑制した記述にしたとも考えられる。一方で、公開時期がこれより古いとはいえ、IEEE EAD ver2ではアルゴリズム、機械学習の教師データ、さらに文化によるバイアスにも言及しており、この問題への意識の高さがうかがえる。

２．５．透明性、アカウンタビリティ、トラスト

初期の倫理指針の Asilomar Principle や人工知能学会・倫理指針では透明性[22]、アカウン

---

19 原文は以下の通り：Fairness (as well as bias) can be defined in more than one way. For purposes of this chapter, a commitment is not made to any one definition—and indeed, it may not be either desirable or feasible to arrive at a single definition that would be applied in all circumstances.

20 直接的に公平に扱う属性だけでなく、当然ながら、その属性に間接的に作用する属性も含めて公平化する。詳細は以下の文献が参考になる。神嶌 敏弘、『公平配慮型データマイニング技術の進展』、第 31 回人工知能学会大会、1E1-OS-24a-1、2017.

21 拙著『裏側から視る AI－脅威・歴史・倫理―』近代科学社(2019 年 9 月刊行)の 5 章を参照していただけると幸いである。

22 transparency

タビリティ[23]、トラスト[24]という概念は扱われていなかった。これらの概念は、AGI の脅威とは無関係に現在の AI を社会で実利用するにあたって避けて通れないことが認識されたといえる。総務省 AI 開発ガイドラインでは透明性が重要項目として列挙され、その後の IEEE EAD ver2 以降の倫理指針では重要な項目として丁寧に記述され続けている。

まず、これらの諸概念の関係を説明する。

1)AI システムの動作やその結果に対する説明を生成できる AI の説明可能性[25]、そしてその説明の理解可能性、

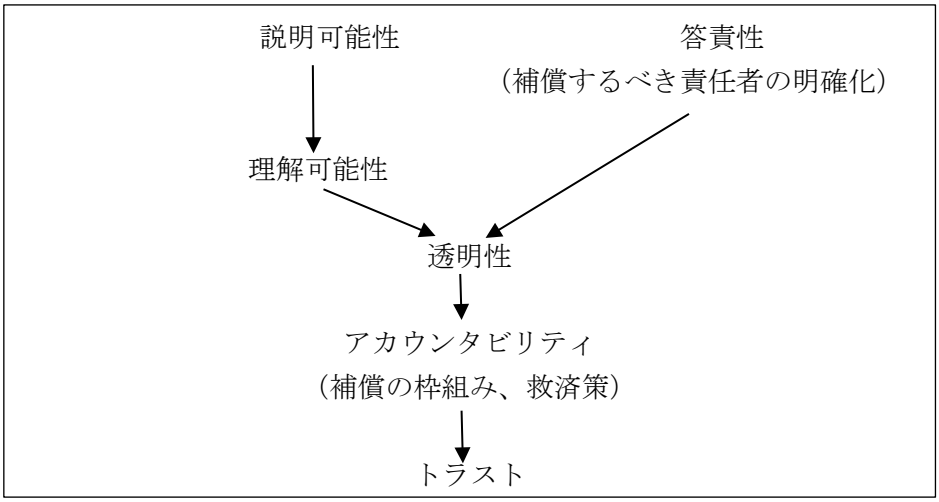
2)AI の起こした事故に対する責任者の明確化（法律分野では「答責性」という単語が使われることがあるが、答責性はアカウンタビリティと同義で使う場合もある。用語の専門的な定義は避け、ここでは「誰が責任をとるか」という部分を答責性の意味として使うことにする。）

上記 1)、2)という主要な二つの項目が総合されて透明性という概念に至る。透明性にはさらになぜその事故を引き起こしたかという原因究明を辿れる追跡可能性[26]と当事者ないし関係者間のコミュニケーションという要素が含まれる。さらに透明性に関するこれらの要素が与えられたときに、実際の補償の枠組みや救済策[27]、被害の報告義務、今後の被害を最小化する方策などが総合されたアカンウタビリティという概念になる。ただし、この説明からもお分かりいただけるようにアカウンタビリティは複雑な構造を持っているため、よほどの専門家でなければ理解し使いこなすことはできない。まして一般の AI 利用者にとって理解するにはハードルが高いだろう。そこで、アカウンタビリティの内容はさておき、「とりあえず信用して使ってみよう」という態度すなわちトラストに落ち着くことになる。この全体像を図 1 に示す。なお、Trustworthy AI では、トラストできる AI という意味で trustworthy AI という概念を前面に打ち出している[28]。

---

[23] accountability
[24] trust
[25] explainability
[26] traceability
[27] redress という単語が使われることが多い。
[28] トラスト(trust)の意味は「A が B をトラストする」という関係を表す。一方、トラストという関係を十分に期待できる、あるいは信頼できるという概念は trustworthy という単語を用いる。日本では、トラスト、trusted という表現を trustworthy の意味で使うことが多いが、これは英語的には誤用に近いだろう。

図 1. 諸概念の関係



この図で注意をしておくべきことは、答責性における責任者の範囲である。事故が起きたときに主として経済的補償をする責任者は、AI システムのオペレータ、設計者、運営者すなわち事業主体、さらにはその事業への出資者であろう。オペレータが明白なミスをした場合は責任を問うことは容易だが、AI システムの複雑性、ブラックボックス化、自ら学習する機能によって、動作の予見性が低い。したがって、オペレータが正しく運用しても、結果がまずいことがありえる。すると、責任は AI システムの設計者なのか、あるいは AI システムで使うツール開発者なのかという技術的に解明困難な状況に陥る。当然、保険もかかっているから、保険会社はこれらを勘案して保険ポリシーの契約を作ることになる。したがって、この状況では契約作成も難問となる。

ここで期待されるのが、AI システムの動作を明らかにしてくれる説明であり、この説明能力を持つ AI が説明可能 AI[29]である。ただし作られた説明は少なくとも AI システムのオペレータや運営者には理解可能でなければ意味はない。理想的には、末端利用者にも理解可能であってほしい。しかし、技術的に見ると、AI システムが生成した説明として、内部変数の値の変化などを表示されても開発者でもなければ理解はできない。一般人の利用者を含む多くの人々に理解可能な説明を生成することは困難な課題である。AI の挙動を AI が使っているアルゴリズムに沿って説明するような内部動作由来の説明は研究が続けられてきた。しかし、捗々しい成果があがっていない。そこで、最近は AI の動作を外部から見て理解できる簡単なシミュレータ、例えば、決定リストや、各ノードが条件で、yes、no の各々の場合に移動する先のノードが与えられる決定木のような直観的理解が可能な形式[30]で表現し、入力から出力結果に至るルートを表示するような方法が研究されている。ちなみに Trustworthy AI では説明する相手は human beings と書かれており陽に特定されていないが、このような現状からみてやむをえない記述であると思われる。

事故時に対応できる説明可能性、責任をとる対象者を明記する答責性、およびすでに述べたように救済策、報告義務、被害最小化方策などが揃うとアカウンタビリティという概念が

[29] eXplainable AI　略称 XAI
[30] このように AI システムを近似するように作られた決定木を Born Again Trees と呼ぶ。

成立する。このような流れは IEEE EAD ver2、1e の法律の章で詳しく述べられている。アカウンタビリティは事故の説明を行う責任と補償や救済を総合した概念であるが、日本語でしばしば和訳として使われる「説明責任」は説明をすれば責任を果たしたというイメージで使われており、誤訳ではないかと思われる[31]。

ところで、アカウンタビリティはすでに起きてしまった結果についての説明や対応策を表すのに対して、将来、AI システムが安全に動作することを保証する概念が responsibility である。説明可能性を AI の能力として作りこむことや救済策は responsibility の確立にも必要である。ただし、これはむしろ後に述べる AI の安全性との関係が強い。

アカウンタビリティは、背後にある人的、組織的問題、技術的問題の両者が絡み合うため、一般の AI システム利用者にとって理解が困難である。そこで先に述べたように、一般利用者にも通用するトラスト[32]という概念が浮上する。表 1 の AI 倫理指針では Asilomar Principle、人工知能学会・倫理指針ではトラストは扱われず、IEEE EAD ver2、人間中心 AI 社会原則ではトラストという表現は散見されるが、一般的用語としてのトラストを直接定義せずに技術、AI、社会などにおけるトラストと同じ意味で、かなり抽象的に扱っている。これらに比べて、IEEE EAD 1e ではより積極的にトラストを位置づける作業を行っている。すなわち、透明性からトラストを導く条件を以下のように列挙している。

- AI システムの動作と結果に関する説明可能性があり、かつバイアスがなく公平であること。
- AI システムの動作と結果が予め定められた規範的基準を満たすこと。
- AI システムが効率的に作業できること
- AI システムの出した結果に再現性があること
- AI システムの設計、開発、調達、供用、動作、効率化の証明において、良くない結果が出力された場合もそれらを説明可能であり、保証の範囲内では懲罰措置と是正措置が取れること

たしかにこれらの条件を満たしていると AI システムの運営業者が保証していると言えば、一般利用者もトラストするであろう。ただし、私見では、一般の末端利用者は、むしろこのような保証をする AI システムの運営業者自体が信用できるかどうかに関心があると思われる。具体的に言えば、この AI システム、あるいは同じ業者の提供する AI システムが、過去に事故や不具合をおこしていないこと、あるいは AI システムを提供している業者の過去の行動に上記諸条件を守るという点で落ち度がなかったこと、あるいは補償がきちんと行われたことなどによって、その業者の提供する AI システムがトラストできるということになる。

Trustworthy AI では、trustworthy を lawful(遵法的)、ethical(倫理的)、robust(頑強性がある)と定義している。最後の robust は通常の意味でシステムが robust であること、すなわち同じ入力に対して常に同じ出力を出し、壊れないということと社会的な安定性を意味

---

[31] 総務省 AI 開発ガイドライン、AI 利活用ガイドラインにおいては説明責任という表現を避けアカウンタビリティを用いており、人間中心の AI 社会原則においても「説明責任（アカウンタビリティ）」と注意深く併記している。

[32] trust は A と B の 2 者の間に成立する関係である。A や B が他の者から trust されうる存在であることを trustworthy と言う。

している。ethical の部分は、倫理の内容が明確に記載されていないため、その意図を理解しにくい。古典的倫理としては、功利主義的倫理、規範的倫理、義務論的倫理、徳倫理、などがあるが、どのような倫理に沿うべきか書かれていない。しかし、これらはどうやら標語またはモットーらしく、Trustworthy AI の満たすべき具体的条件は以下のように列挙されている。

- 人間の主体性および人間による監視
- 技術的ロバストさと安全性
- プライバシー保護とデータのガバナンス[33]ができている
- 透明性
- 多様性、非差別性、公平性
- 環境および社会的な幸福への寄与
- アカウンタビリティ

IEEE EAD 1e の諸条件が技術的であるのに比べて、これら Trustworthy AI の条件はより抽象的ではあるが、その意図は明確である。もちろん、これらが抽象的であることは十分に意識しているらしく、これらの諸条件の改善、追加、明確化に継続的努力を要すると記している。また、Trustworthy AI を基礎におく OECD Recommendation においても、同様の trustworthy の条件を明言している。

以上まとめると、手段としての説明可能性、透明性、アカウンタビリティに留意しつつも、より広範囲の人々に理解できるトラストないし trustworthy に強く焦点が当たり始めたといえよう。

２．６．悪用、誤用

悪用[34]、誤用[35]については Asilomar Principle では AGI そのものを技術自体の悪用とみなしているような書き方に読める。人工知能学会・倫理指針では AI 研究者が悪用の発見や告発者の不利益になることをしてはいけないとやや踏み込んでいる。

IEEE EAD ver2、1eでは悪用、誤用をいかに防ぐかという観点から“Awareness of misuse”という標語を打ち出し、AI開発、利活用の種々の局面で誤用、悪用に関する具体的な提案を行っている。法律の章では、AIの仕組みで防ぐだけではなく、悪用、誤用による被害を受けた者の保護、発見した者からの内部通報制度を法制度化すること、通報者の制度的に保護すること、組織内で悪用・誤用が起きないように教育すること、被害者に対する保険による補償の活用などを提言している。

Trustworthy AI では誤用、悪用に関しては malicious misuse として 1 回触れているが、その内容や対策については書かれていない。OECD Recommendation では安全性やセキュリティの一部として一言言及されているだけである。人間中心 AI 社会原則では悪用についての警告は 1 回されているが、誤用についての言及はない。総務省 AI 利活用ガイドライン

[33] この意図は理解が難しいが、おそらく人間が統御できることと思われる。
[34] abuse
[35] misuse

では誤用、悪用とも言及はない。

以上まとめると、悪用、誤用については IEEE EAD ver2、1e が質量ともに圧倒的に優れた分析、提言を行っている。一方、他の倫理指針では言及はあるものの具体性のある記述に欠ける。AI 技術、あるいはその技術的基礎となる機械学習に関して、なにかまずいことが起こると、機械学習や AI の所為にするという行為も見られる[36]現在、この課題に対する IEEE EAD ver2、1e の取り組みは高く評価されるべきであろう。

２．７．プライバシー

すべての倫理指針で継続的に取り上げられているテーマとして、プライバシーの保護がある。よく耳にするように、プライバシー保護が技術発展を阻止していると仮定してみよう。そうだとすれば、産業界からも多くの人々が参加している IEEE EAD ver2、1e、あるいは産業界からの圧力もあった[37]と言われる Trustworthy AI で、プライバシー保護を強く主張することには、ならなかったかもしれない。しかし、これらの指針でもプライバシー保護が取り上げられているということは、裏を返せば、プライバシー保護は今や IT 産業で利益を生む分野になってきているからともいえるだろう。このような考え方を最初に広く述べたのは当時カナダのオンタリオ州のプライバシー・コミッショナーだったアン・カブキアンである。カブキアンが提案したプライバシー・バイ・デザイン[38]の 7 原則の中の第 4 原則「プライバシー保護はゼロサムではなくポジティムサムである」がこれに相当する。言い換えると、プライバシー保護は消費者にとって望ましいことであるだけではなく、サービスや製品供給を行う事業者にとって、プライバシー保護が消費者に与える安心感から消費者からの好評につながり、さらにプライバシー保護の法制度が整備されるにつれて法的リスクも回避できるという観点で有益なものになるということである。最近では、GDPR[39]、日本の個人情報保護法の改正、アメリカにおける California Consumer Privacy Act (CCPA)[40]などに見られるようなプライバシー保護の世界的潮流が強いことが窺われる。

IEEE EAD ver2 では、プライバシー保護という原則だけでなく、法律的観点から分析して、透明性、アカウンタビリティとにおける位置づけとプライバシー保護の実装についての問題を列挙している。さらに IEEE EAD 1e では Personal Data and Individual Agency というタイトルの章を立て、個人データの create、curate[41]、control の側面から人間自身の個人データ管理、および人間の代理で個人データを管理する AI エージェント[42]の役割につ

---

36 実名は避けるが、特定の国の応募者を採用しないという発言を公開し非難された人物が、後になってそれは AI が出した結果だと謝罪することが 2019 年末にあった。

37 Wired:こうして大手テック企業は、「AI の規制」に影響力を行使する、Regulation 2019.08.19 MON 17:00.参照先：https://wired.jp/2019/08/19/how-tech-companies-shaping-rules-governing-ai/

38 Cavoukian A. (2011). Privacy by Design The 7 Foundational Principles.
参照先: https://www.iab.org/wp-content/IAB-uploads/2011/03/fred_carter.pdf

39 EU の General Data Protection Regulation

40 https://www.caprivacy.org/

41 ここで curate は、自分の個人データがどのように使われたかを知る機能も含む。

42 IEEE EAD 1e では、algorithmic agent という名称を使っている。

いて提言をしている。また、間接的ではあるが、エージェントは Trustworthy AI や OECD RECOMMENDATION でも触れられている。この論点の展開を次節で試みる。

２．８．AI エージェント

IEEE EAD 1e の Personal Data and Individual Agency の章で導入されている A person's A/IS agent は A person すなわちデータ主体の意思に沿う条件に従いつつ、事前的にデータ主体の個人データの処理を行うソフトウェアツールである。事前的という表現の意味するところは、個人データを外部の事業者などに無条件に渡してサービスを受けるのではなく、事業者の提示する個人データの利用方法とサービスがデータ主体自らが決めた条件に合致するかどうかを事前にテストし、条件に合致すれば事業者に個人データを渡してサービスを受けることである。このようなツールは AI 技術によって実現するものであるので、ここでは AI エージェントと呼ぶ。AI 技術としては、事業者の提示する個人データの利用法と自ら決めた条件の一致の可否判断を行う推論処理、さらなる他の類似サービスとの比較などを行うための情報収集、条件の妥協など、人間が通常行う知的処理が想定される。

このような AI エージェントの概念設計を具体的レベルまで記載している倫理指針は IEEE EAD 1e をおいて他になく、AI の利用方法の極めて具体的かつ先進的な提案となっている。なお、これに先立つ IEEE EAD ver2 では、Personal Data and Individual Access Control の章で、個人データのプライバシー保護を行うためのアクセス制御とプライバシー保護法制度の関係などを議論し、IEEE EAD 1e の AI エージェントの提案の下準備をしている[43]。

２．９．安全性ないしセキュリティ確保

安全性については Asilomar Principle では直接言及がないが、他の倫理指針ではニュアンスの差はあれ、意見が表明されている。人工知能学会・倫理指針では表面的に安全性を重視しているが、同じ項目に制御可能性、つまり AGI 対応策も書かれているので、AGI を危険視しているように読める。

総務省 AI 開発ガイドラインと AI 利活用ガイドラインの双方では、より現実的に安全性を「利用者及び第三者の生命・身体・財産に危害を及ぼすことがないよう配慮する」としているが、これは技術的ツール一般に成り立つ言明であろう。加えて、セキュリティ確保にも言及しているが、これはネットワーク経由での AI への攻撃を意識したものであり、現実に喫緊の課題である。人間中心 AI 社会原則では、サイバーセキュリティ確保を重視している。

IEEE EAD ver2、1e では、安全性(safety)は AGI が人類にとって安全であるべきという視点で問題を捉えている。また、サイバーセキュリティについても一定の関心を示している。Trustworthy AI と OECD Recommendation ではあくまで AI をツールとしてみた

[43] IEEE EAD ver2、1e のこれらの章は、committee に参加した筆者の経験からすると Co-Chair の John Havens の貢献が大きい。なお、John はもう一人の Co-Chair と共同で内容を練ったと committee meeting で発言していた。

場合の安全性、つまりツールが壊れて人間に危害が及ばないようにするという観点で議論され、またサイバーセキュリティについても若干の言及がある。

２．１０．SDGs

SDGs は Asilomar Principle や人工知能学会・倫理指針では触れられていない。同時期から作成されていた IEEE EAD ver2 では、国連の Sustainable Development Goals（SDGs）について economics の章で若干触れている程度だった。しかし、IEEE EAD 1e では A/IS for Sustainable Development の章で AI が貢献できる項目として中進国、発展途上国への均等な機会提供、雇用の問題（いわゆる AI が職を奪う話）について AI の観点からきちんと書き込んでおり、一読の価値がある。2018 年時点で重視されるスキル、2020 年に落ち目のスキル、2020 年にトレンドになっているスキルの比較[44]は興味深い。

Trustworthy AI と OECD Recommendation では国連の SDGs を参照しつつ、環境、文化と幸福を SDGs のターゲットにし、そのための方向性を人間中心の AI という建付けにしている。人間中心 AI 社会原則でも SDGs を意識して取り込んでいるが、IEEE EAD のような具体策へのブレークダウンはされていない。

２．１１．教育

教育に触れ始めたのはIEEE EAD ver2からである。倫理研究の方向性の一つとして工学

44

2018 年時点の重視スキル	2020 年の落ち目スキル	2020 年のトレンドスキル
1. Analytical thinking and innovation	1. Manual dexterity, endurance, and precision	1. Analytical thinking and innovation
2. Complex problem-solving	2. Memory, verbal, auditory, and spatial abilities	2. Active learning and learning strategies
3. Critical thinking and analysis	3. Management of financial and material resources	3. Creativity, originality, and initiative
4. Active learning and learning strategies	4. Technology installation and maintenance	4. Technology design and programming
5. Creativity, originality, and initiative	5. Reading, writing, math, and active listening	5. Critical thinking and analysis
6. Attention to detail, trustworthiness	6. Management of personnel	6. Complex problem-solving
7. Emotional Intelligence	7. Quality control and safety awareness	7. Leadership and social influence
8. Reasoning, problem solving, and ideation	8. Coordination and time-management	8. Emotional intelligence
9. Leadership and social influence	9. Visual, auditory, and speech abilities	9. Reasoning, problem-solving, and ideation
10. Coordination and time management	10. Technology use, monitoring, and control	10. Systems analysis and evaluation

系の学生に倫理、正義、人権、およびAIに係わるビジネスの実践における倫理を教えるべきだとしている。また、プライバシー保護においても消費者教育が重用だとしている。

IEEE EAD 1e では持続可能な発展の観点から Methods to Guide Ethical Research and Design の章において、種々の課題を提示している。数理系の教育だけでは不足であるため、将来の技術、ビジネス、政策を担う学生、および一般人のすべてを対象にした分野横断的、学際的な教育の必要性を訴えている。さらに AI から派生するリスクとその管理方法も教育すべきとしている。すでに SDGs の節でも書いたように教育の必要性は SDGs の実現の有力な手段であり、IEEE EAD 1e では、このことが強く意識されている。

Trustworthy AI では教育の機会均等と倫理教育の必要性を主張している。それを引き継ぐ OECD Recommendation では教育については触れていない。

人間中心 AI 社会原則では、教育・リテラシーの原則として、文系ないし一般人への AI 技術や STEM 教育だけでなく、AI 技術開発者には法制度、経済、社会、文化への理解が必要であるとしている。すべての技術開発者がこういった社会制度等に関する知識をマスターするのは無理にしても、開発の方向を誤らないためは、サービスや製品の開発チームに 1 名はこのような素養を持った人を配置しておくべきであろう。また、初等教育、情報弱者への手当についても丁寧に触れている。

２．１２．独占禁止、協調、政策

人間中心 AI 社会原則では、公正競争確保の原則において、特定の企業や国による AI 技術やデータ資源の独占への警鐘を強く鳴らしている[45]。また、イノベーションの原則において、人材・研究の両面から、国際化・多様化と産学官民連携を推進するべきとしている。これと似た主旨で、国際協調ないし開発組織間の協調は Asilomar Principle と Trustworthy AI、 OECD RECOMMENDATION で陽に言及されている。また、Trustworthy AI と OECD Recommendation では AI が活躍できる環境を推進する政策を奨励している。

IEEE EAD ver2、1eでは独占については触れていない。これは政治的ないし企業経営の観点からは難しい問題なので、他の倫理指針ではあえて触れなかったのであろうか。

２．１３．軍事利用

Asilomar Principle では Lethal Autonomous Weapon System(LAWS)すなわち自律型致死兵器システムを単純に禁止せよと主張している。

IEEE EAD ver2 では、AI 兵器の定義の再構築から始めている。大雑把にいえば、引き金を引く操作を AI の判断で行う兵器と定義される。しかし、積極的な攻撃なのか、攻撃された場合の防衛なのか、など複雑な戦場の状況では明確な定義が困難であろう。IEEE が工学、技術系の学会であるから、このような議論になるのは当然である。また、直接的に AI 兵器禁止を声高に記載しないのは、IEEE には多くの兵器製造に関連するメーカも入っているからではないかと思われる。なお最新の IEEE EAD 1e ではこの問題に全く触れていない。他の指針でもほとんど触れられていない。ひとつの理由は AI の軍事利用は好ましくないとい

[45] 人間中心 AI の場合、日本の経済的位置、地政学的立ち位置を反映しているのかもしれない。

う主張は当然であるが、これを実現するための CCW[46]のような国際政治の場は各国の利害対立があまりに生々しく、倫理云々という場ではないことがある。仮にこのような議論をしている先進国が AI 兵器を禁止しても、テロリストなどが AI 兵器を使うようになれば対応措置が必要になる。AI の軍事利用の倫理的側面については拙著[47]でまとめている。

Trustworthy AI では AI 兵器も Article 3 of the Treaty of the European Union に述べられた平和主義に従うべきとしている。また、日本の指針は軍事的な発信を控えるこれまでの傾向を踏襲して、AI 兵器については一切触れていない。

２．１４．法律的位置づけ

AI システムが実用に供されて生み出された結果を法律的にどのような扱うべきかに関しては、すでに公平性、非差別、アカウンタビリティ、プライバシー、ないし政策などに関して扱ってきた。ここでは、以上の諸点の底流にある AI と法律の関係すなわち AI を法律的にどのように位置付けるかについての議論、端的にいえば AI に人格権などの固有の権利を与えるか否かについての議論を紹介する。

多くの AI 倫理指針では AI の人格権は議論すらしていない。AI はあくまで人間が使うツールであるとしている。Trustworthy AI、人間中心 AI 社会原則でも AI はツールという立場は明確にしている。Asilomar Principle では人間と同等の知的能力を想定しつつも、その法律的位置づけについては何も言及していない。人工知能学会・倫理指針はもっと微妙であり、第 9 項目で人間と同定度の知的能力を持つ AI を想定し、それが法令遵守することを第 2 項目で述べているので、AI に人格権を認めたと解釈する可能性が残ってしまっている。

AI が人間が使うツールであるなら、AI システムが人間にとって被害を与えた場合は、AI システムないし AI ツールそのものではなく、AI システムの運用者、管理者などの関係者が責任をとるべきであり、論点は誰が責任を取るべきかに集約される。透明性、アカウンタビリティの議論は責任をだれが取るかに焦点が当たっていた。

ところが、IEEE EAD ver2 の法律の章の最初の節、および 1e の法律の章の最期の節では、AI の人格論的な法律的位置づけの議論が展開されている。以下でこれらをまとめてみる。

IEEE EAD ver2では、いきなりAI[48]に人間のような法的位置づけを与えるのは時期早尚であり、まずコンパニオンアニマル、動物、被雇用者、法人のような部分的権利を考えることを考えるべきとしている。また、AIにその判断を委譲してはいけない事象、例えば、戦争開始の可否、死刑の判決、を明確化すべきとしている。判断を委譲する場合は、当然ながらAI開発者、運用者はAIの判断が法令遵守していることを十分に確認しなければいけない。

IEEE EAD 1e では、さらに踏み込み、AI に自律性をもった存在としての権利、例えば人格権を与えないとすると、事故が起きた時、AI ツールの開発、AI システムの開発、AI システムの運用の長い連鎖を辿って責任者を特定するという困難な作業が必要であることを注

46 特定通常兵器使用禁止制限条約
47 中川裕志：『裏側から視る AI－脅威・歴史・倫理―』.近代科学社.2019 年 9 月
48 IEEE EAD ではこの論文でいう AI を AI/S(Autonomous and Intelligent Systems)と書いている。

意喚起している。この辿る作業の結果は、個々の事故においてケース依存性が高いのみならず、関係者、国、文化、そして法律に依存しており、難解な代物である。一方で、AI に人格権とまではいかなくても、何らかの責任を問える権利、たとえば法人格のような権利を与えるほうが、法体系の見通しが良いこと、そして被害者の泣き寝入りを避けるために役立つ可能性についても述べている。最後にやや皮肉っぽく、現代の法律で前提とされている人間の自律性のほうが疑われるような事態[49]になったら、AI に人格権を与えるというのは突飛な話ではなくなるかもしれないと述べている。なお、筆者が検討している自動運転車の事故の責任においては、以下のような見方もある。すなわち、自動運転車によって事故率が大きく低下し、鉄道事故や航空機事故のような低いレベルになったら、事故の法的責任を誰かに問うのではなく、事故原因の究明を優先して、将来の事故原因をなくする方向で社会が動くべきであるという見方である。自動運転車のように AI が自律的な行動をするような時期は予想より早く訪れる可能性もあるため、AI の法律的位置づけは案外、近々に解決を迫られるかもしれない。

２．１５．幸福

幸福（well-being）を AI の目的とするのは自明すぎることである。技術的に興味深いのは、IEEE EAD ver2 で GDP 以外の幸福度指標の必要性を述べ、種々の指標の比較をしていることである。IEEE EAD 1e では、幸福度を測る対象を、経済指標だけでなく、肉体的健康および精神的健康、環境、教育条件、コミュニティ、仕事（ワークライフバランス）などに拡大している。さらに、考案された幸福度指標を人々のデータからアルゴリズム的に計算する方法を示唆している。

２．１６．名宛人

以上紹介してきた倫理指針において期待される読み手、すなわち名宛人としては、AIツール、AIシステムの開発者、AIシステムを使ったサービスを行う事業者、サービスの受け手である消費者、政策担当者などが考えられる。総務省AI利活用原則には、実質的に名宛人の種別に関する記述があり、開発者、利用者、データ提供者、AIサービスプロバイダ、最終利用者（自らのビジネスとして利用する者、行政機関、消費者的利用者）に分類している。ただし、ここには政策立案者が抜けているので追加しなければならない。このような分類を行うことにより、名宛人に直接影響する倫理指針の項目が明らかになる。筆者の理解範囲での名宛人の一覧を表３に示す。

表３．倫理指針の名宛人

名称（略称）	名宛人
Asilomar　Principle	開発者、政策立案者

[49] 例えば、ユヴァル・ノア・ハラリは人間とは、遺伝子で決まる生命活動のアルゴリズムにほぼしたがって、環境から得たデータによって行動する存在と定義している。ユヴァル・ノア・ハラリ（柴田裕之訳)、『ホモ・デウス:テクノロジーとサピエンスの未来』河出書房新社、2018

人工知能学会・倫理指針	開発者、AI 自体
総務省 AI 開発ガイドライン	開発者
IEEE EAD ver2	主に開発者、ただし政策立案者も含む
IEEE EAD 1e	主に開発者、ただし政策立案者も含む
人間中心 AI 社会原則	開発者、利用者、政策立案者
Trustworthy AI	開発者、政策立案者
OECD Recommendation	政策立案者
総務省 AI 利活用ガイドライン	開発者、政策立案者、消費者（最終利用者）

名宛人としてAIの開発者はほぼ全ての倫理指針で共通する。次に多い名宛人は政策立案者である。AI開発者は、本質的に先進技術の追求あるいは売れる商品の開発を狙うため、してよいこと、すべきことを示す倫理指針を必ずしも歓迎するわけではなく、倫理指針にマッチしない開発に進むこともありえる。したがって、倫理指針に実効性を持たせたければ、より強制力のある国の政策に反映させるべきである。これが、政策立案者が多くの指針で名宛人とされる所以である。人間中心AI社会原則やTrustworthy AIにはAI研究開発の推進するAI開発ないし投資政策を促す文言が明示されている。また、AIの利用者を意識している倫理指針も多い。

別の側面の例としては、Asilomar Principleで明記されているAIの軍事応用や自律AI兵器の抑制と禁止を訴える名宛人がある。この訴えの第一義的な名宛人は政策立案者であるものの、政策は世論を反映するものだから、世論形成を行う一般人も名宛人になる。

特色があるのは人工知能学会・倫理指針である。人工知能学会・倫理指針は明快に名宛人を人工知能学会会員としている。ただし、会員宛ての前半の 8 項目の後の第 9 項目に「人工知能が社会の構成員またはそれに準じるものとなるためには、上に定めた人工知能学会員と同等に倫理指針を遵守できなければならない。」と明記され、素直に読めば AI 自体にも名宛人として人間並みの倫理観を要求している。思うに、人間なみの倫理観を持つ AI は、人間と同レベルの知的能力を持ついわゆる汎用 AI であり、これは超知能の一歩手前の AI である。したがって、人工知能学会・倫理指針の第 9 項目は特異な構造を持っているといえよう。

３．パーソナル AI エージェント

３．１．概念設計

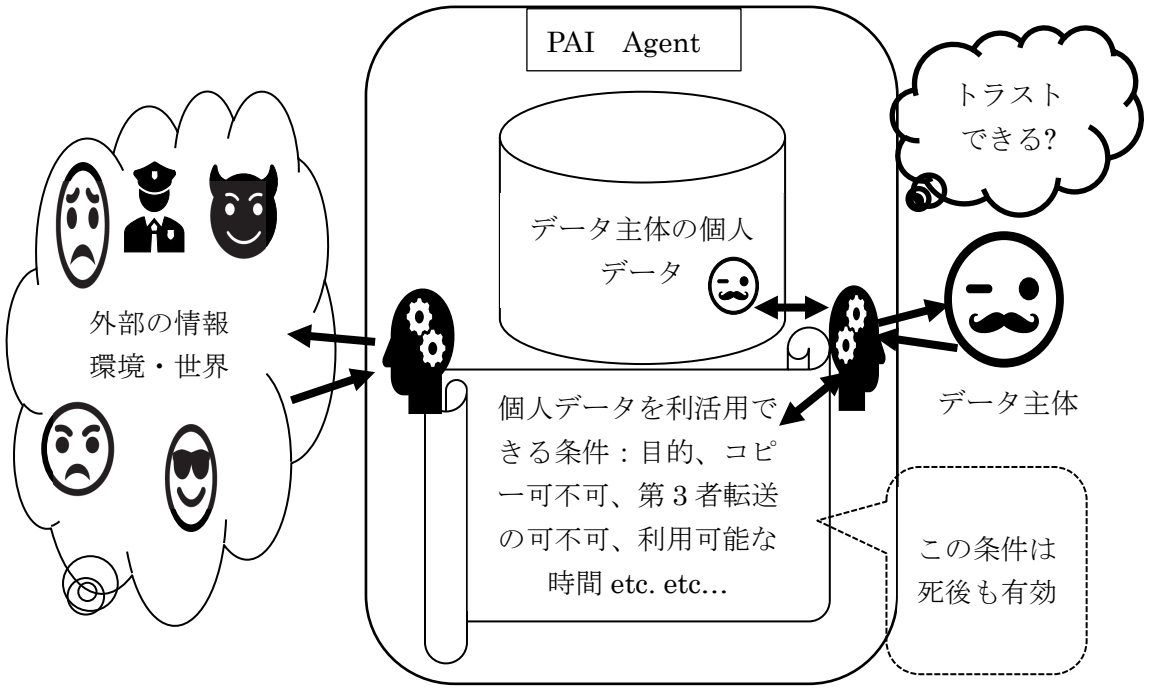
この節では、IEEE EAD 1e の Personal Data and Individual Agency の章で導入されている A person's A/IS agent すなわち個人を代理する Data Agent を AI 技術によって実現するパーソナル AI エージェント（以下では、PAI Agent と略記する）について説明する[50]。

情報が溢れかえり、複雑化する一方の情報社会において、生身の人間が対峙できる時代は終わり、外界と個人を仲介してくれる AI による個人の代理、すなわち PAI Agent が各個人にとって不可欠な存在になることが予想される。すでに IT プラットフォームないし IT 経

[50] ここでの PAI Agent は IEEE EAD 1e に書かれていることそのものではなく、そこに書かれた指針を実装する場合の私案、および補足である。

由の種々のサービスプロバイダが個人の生活に入り込んでいる。また個人情報を預けて運用を任せる情報銀行なども存在しており、これらを一般人が利活用、運用するには生身では困難なことが多く、PAI Agent の支援が必要になるのではないかと思われる。PAI Agent の概念を図２に示す。

図２．PAI Agent の概念



データ主体である個人の代理をする PAI Agent（図の中央）のデータ内容は、図の右側のデータ主体の個人データと、個人データを外部の事業者など（図の左側）が使う場合の利活用条件群[51]からなる。個人データも利活用条件も PAI Agent に最初から与えられているわけではない[52]。最初からデータ主体が自身の個人データの利活用条件を記述することはデータ主体にとって負担が大きいし、また種々の利活用ケースを数え上げることは現実的ではない。よって、外部事業者との利活用に関するやり取りにおいて、既存の利活用条件に当てはまらない場合は、PAI Agent がその利活用の可否をデータ主体に伺いを立て、データ主体の可否判断の結果を使って利活用条件を徐々に更新し拡充していくことが現実的であろう。このような処理はデータベースの解釈、更新を含む知的処理が必要であるため、PAI Agent のデータ主体向けインタフェースと外部事業者向けインタフェースには AI 技術が使われる。

[51] 英語では、terms and condition と表現する。
[52] 最初から利用条件をデータ主体が記述するシステムとして、PDE(Personal Data Ecosystem)が提案されていたが、あまり普及していないようである。筆者がみたところでは、利用条件の記述が難しく、IT 技術の専門家でもなければとても書けそうにないと思われた。

図２では PAI Agent は独立したソフトウェアのように描かれている。たしかにこのような独立したソフトとして個人にスマホや個人対応するクラウドサーバ上で実現することもありえるだろう。一方で、情報銀行や SNS 業者の個人個人に対応したサービスを行う AI インタフェースとして実現されることもあるだろう。PAI Agent の具体的な実装は今後の課題である。

３．２．死後のディジタル遺産データの扱い

次に考えられるのは、PAI Agent はデータ主体の個人データを扱うとするなら、どのような期間においてデータ主体の代理をするかという問題である。人間は生まれる前、すなわち胎児のときから両親の氏名や DNA という個人情報を持っている。これを他者が知ろうと思えば知られてしまう可能性がある。成長して学校に通い、社会人として仕事をし、最後に退職して死にいたるまで、常に自分の外側にある膨大かつ複雑な情報の世界に係わり続けなければならない。死後も SNS メッセージやメールのような大量のディジタル遺産として個人データが残されるが、当然、死後は自分自身の個人データを管理できない。あるいは認知症によって生物学的な死の前に自分の個人データが管理できなくなるかもしれない。つまり、データ主体自身が自らの意思で個人データを管理できる期間とその前後にある自らの意思で管理できない長い期間にわたって、データ主体を PAI Agent が代理をするようになれば、PAI Agent はどのような機能を持つべきかという問題がある[53]。このような状況を図３に示した。以下でこの図に沿って検討課題を説明する。

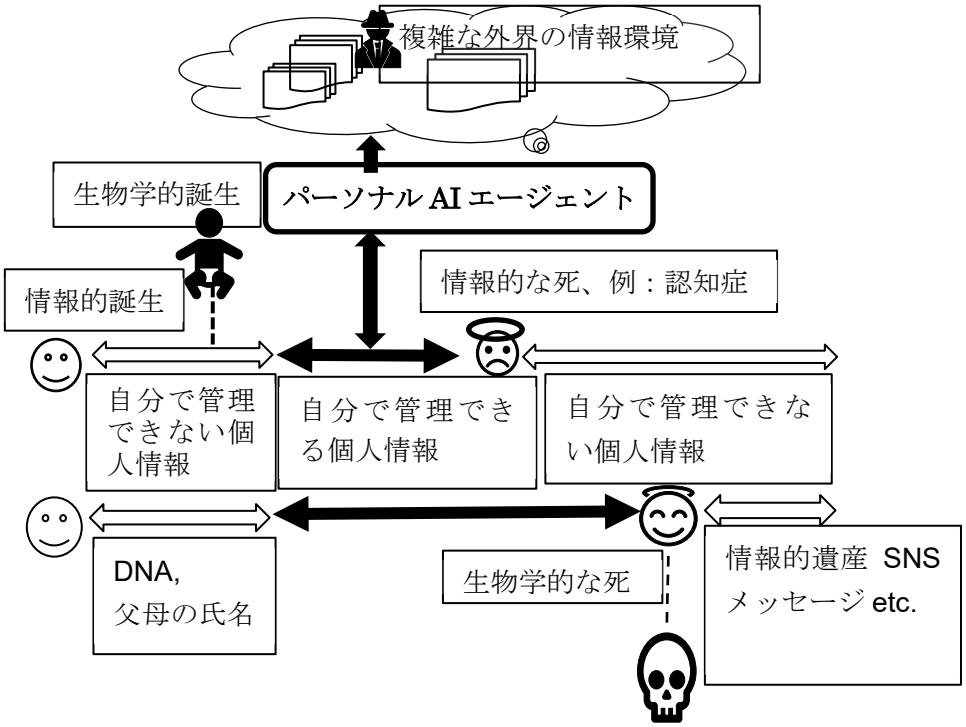
妊娠と同時に発生する胎児に関する種々の個人データを格納するメディアとしては、長く使われてきたのは母子手帳である。これを電子化し、内容も拡大した電子母子手帳は個人データ格納メディアとして有力である。母子手帳と同様に電子母子手帳も母親あるいは父親、場合によっては親権者が管理することになる。ただし、身長、体重や予防接種記録のように誰にでもわかりやすい情報に比べて、遺伝子など一般人にはわかりにくいデータの記載も増えるだろう。すると、個人データがたくさん詰まった電子母子手帳の管理は一般人の手に余ることが予想される。そこで電子母子手帳の情報管理を支援する手段として PAI Agent が期待される。電子母子手帳に書かれた種々の個人データとその利用条件は図２に示される形でデータベース化される。ただし、図２の右側にデータ主体は本人ではなく母親などである。

子供はやがて成長し、自分の個人データを自分で意識するようになると、PAI Agent のデータ主体としての地位を母親などから譲り受ける。だが、この委譲に関しては、母親などと子供の意見をすり合わせるという教育的ないし心理的な問題があり、実用化においてはリアルなケースを想定した分析、調査が行ったうえで、場合に応じた委譲方法を準備するという課題がある。

---

53 科学研究費基盤(B)『情報ネットワーク社会における「死」の再定義』代表者：折田明子。筆者も分担者として共同研究を行っている。

図３．誕生前と死後の個人データの扱い



さて、高齢になって個人データを自分でまったく管理できなくなる状況を考えてみる。生物学的に死ねば当然、それ以後は管理できない。病気などで体力を失って死亡する場合は、病状によっては自身の個人データを死のかなり前から管理できくなり、生物学的な死よりも情報的な死のほうが早く訪れる。認知症の場合は、情報的な死が生物学的死よりもかなり早い。事故になどによる不慮の死では、生物学的な死と情報的な死がほぼ同時に起きる。

いずれの場合でも、PAI Agent は仕えていたデータ主体を失う。ただし、それまでのデータ主体が存命中の個人データの処理で蓄積された個人データの利活用条件は存在する。よって、これを使えば、生前と同じような個人データの処理ができる。問題はディジタル遺産の処置である。ディジタル遺産とは、生前の SNS で発信した情報、送受信した電子メール、さらに個人 PC ないし Dropbox や Google Drive などのクラウドストレージの残された個人データを意味する。また、残されたファイルに書かれていた銀行アカウントやパスワード、さらには暗号資産など、金銭的価値のあるディジタル遺産もある[54]。

まず、金銭的なディジタル遺産は、通常の金融資産と同様に相続人が決まれば相続されるものである。遺言があれば、相続の手続きは従来と同じような方法で行えるように思える。しかし、不慮の事故死など遺言がない場合は、厄介なことが起きる。すなわち、金融資産に関する情報を知らなければ相続できないが、金融機関自身が保持している情報の場合、金融機関が相続人候補には開示しないという問題がある。そのため故人の金融資産が塩漬けに

[54] Edina Harbinja (2017) Post-mortem privacy 2.0: theory, law, and technology,International Review of Law, Computers & Technology, 31:1, 26-42,DOI:10.1080/13600869. 2017.1275116.
https://doi.org/10.1080/13600869.2017.1275116

なって利用できなくなる問題がありえる。さらにディジタル銀行口座についての情報が含まれたディジタル遺産を開示しないという Yahoo!のポリシーも報告されている[55]。

では、相続人の候補者は金融資産情報が書いてある可能性のあるディジタル遺産を見てよいのだろうか？ディジタル遺産には個人のプライバシーに係わるデータが記載されている可能性が高いが、死者の個人データには個人情報保護法や GDPR は適用されないので、法律的には一見許容されそうである。ところが、ディジタル遺産の中に存命中の第三者のプライバシーに係わる個人データが書かれていると、これを相続人が見るのは、個人データの第三者移転あるいは目的外使用になり、法律を遵守していないことになる。つまり、ディジタル遺産の内容は、仮に相続人が見ることを許容するように故人が遺言していたとしても、相続人が見ることはできない[56]。

この状況への解決策の一つが PAI Agent に蓄積されている個人データおよびその利用方法のデータベースを死後に利用する方法である。蓄積されている情報へのアクセスを以下の場合について考えてみる。

(1)PAI Agent 内部に蓄積された個人データの処理（財産的情報）：この場合、個人データ中に記載されている金融資産など財産的価値のある情報の記述を PAI Agent 自身が分析して抽出し、提示する機能が必要である。

(2)PAI Agent 内部に蓄積された個人データの処理（第三者個人情報）：PAI Agent に蓄積された存命中の第三者のプライバシー情報は、自然人が見てしまうとその時点で個人情報の第三者提供になり法的問題となる。したがって、PAI Agent 自身が、第三者に関するプライバシー情報を認識して、消去ないしはアクセスできない処理を施す必要がある。これは、外部情報、例えば当該第三者が存命中かどうかの調査、などの文脈依存的な自然言語の意味解析技術が必要になり、解決可能性はあるにしても大きな技術課題である。

(3) PAI Agent の内部に個人データとして蓄積されているもの以外に SNS やメールサービス事業者などの外部の組織に蓄積されているデータへのアカウントへのアクセス情報がある。これはいわゆるディジタル遺産であり、外部組織との間での死後データの扱いを決めてあればその決め事の通りに処理できる。しかし、決めていない場合は、PAI Agent の利用条件の記載、外部事業者との法的関係によることならざるを得ない。ここに法律的な資格をもつ第三者の介入する必要性があると考えられる。

データ主体の個人の生物学的死亡、ないしは情報的死亡に際しての PAI Agent への権限の委譲方法は課題である。データ主体が情報的死亡以前に自分の意思で権限委譲の方法を個人データ利用条件に記載しておくことが望ましい。財産的価値のあるディジタル遺産の相続では、データ主体が個人データの利用条件にディジタル遺産の処理方法を記載しておくことができる。ただし、法律的な資格をもつ第三者がこの条件の正統性、真正性を保証しなければならないだろう。この問題は、PAI Agent の技術と、その法的位置づけの双方に関連している。

---

[55] Natasha Chu, Protecting Privacy after Death, 13 Nw. J. Tech. & Intell. Prop 255 (2015).https://scholarlycommons.law.northwestern.edu/njtip/vol13/iss2/8

[56] 他者の個人データが書かれていなければ見ることはできるが、そのようなデータが書かれているかどうかは、読んでみなければ分からないという矛盾に陥る。

また、死後のディジタル遺産の扱いついて利用条件に記載がない場合の処理はさらに問題であり、法的措置の介入方法を定める必要があるが、これは将来課題である。

４．おわりに

本論文では、第２章で 2017 年から 2019 年にかけて国内外で公開された多数の AI 倫理指針のうち、影響力の大きな主要な指針に関して、以下の各項目に焦点が当たったかどうか、および各指針でどのような扱いをされたかをまとめた。

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

これによって、AI 倫理の内容の変遷を探ることができ、同時に AI 技術、AI 応用システムの開発を行うにあたって助けになる。また、これらの指針が誰を対象に起草されているか、すなわち名宛人を考察することによって、AI 倫理指針を作成した組織の意図が見えてきた。

第３章では、IEEE EAD ver2、1e で提案された個人データの管理をおこなう代理ソフト、すなわちパーソナル AI エージェントの概念設計について述べた。次に、これを死後のディジタル遺産管理に適用する場合の検討課題について述べた。引き続きの概念設計の詳細化、機能の実用化手法、法律的観点からの位置づけと、パーソナル AI エージェントの利用が適法であるように法律的な建付けの改善の提言が今後の課題である。

AI 倫理指針は依然として新規公開が続いているが、その内容はかなりコンセンサスが得られつつある。ただし、将来公開される AI 倫理指針は注意深く読み解く必要があろう。また、企業等では AI 倫理指針を参考にして、企業独自に AI 製品開発倫理基準などを策定しつつあり、AI 倫理は実用化の段階に入ったと考えられる。

参考文献

R.カーツワイル：『ポスト・ヒューマン誕生』，NHK 出版. 2005

N.Bostrom: Super intelligence, Oxford University Press. 2014

Future Life Institute: ASILOMAR AI PRINCIPLES. 2017.
(https://futureoflife.org/ai-principles/?cn-reloaded=1)

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: Ethically Aligned Design version2: A Vision for Prioritizing Human Well-being with Autonomous and with Autonomous and Intelligent Systems.
(https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf)

人工知能学会・倫理委員会.倫理指針. 2017.
(http://ai-elsi.org/wp-content/uploads/2017/02/%E4%BA%BA%E5%B7%A5%E7%9F%A5%E8%83%BD%E5%AD%A6%E4%BC%9A%E5%80%AB%E7%90%86%E6%8C%87%E9%87%9D.pdf)

総務省・AI ネットワーク社会推進会議.国際的な議論のためのＡＩ開発ガイドライン. 2017.
(https://www.soumu.go.jp/main_content/000490299.pdf)

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design version2:A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. 2017.
(https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf)

AI 戦略実行会議、内閣府.人間中心の AI 社会原則. 2017.
(https://www8.cao.go.jp/cstp/aigensoku.pdf)

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design (first edition):A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. 2019. (https://ethicsinaction.ieee.org/)

The European Commission's High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. 2019. (https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai)

OECD. Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. 2019.(https://www.soumu.go.jp/main_content/000642218.pdf)

総務省・AI ネットワーク社会推進会議.AI 利活用ガイドライン. 2019.
(https://www.soumu.go.jp/main_content/000624438.pdf)

神嶌 敏弘.『公平配慮型データマイニング技術の進展』.第 31 回人工知能学会大会,1E1-OS-24a-1. 2017.

中川裕志.『裏側から視る AI ―脅威・歴史・倫理―』.近代科学社. 2019.

Wired:こうして大手テック企業は、「AI の規制」に影響力を行使する、Regulation 2019.08.19 MON 17:00. (https://wired.jp/2019/08/19/how-tech-companies-shaping-rules-governing-ai/)

Cavoukian A. Privacy by Design The 7 Foundational Principles. 2011.
(https://www.iab.org/wp-content/IAB-uploads/2011/03/fred_carter.pdf)

California Consumer Privacy Act (CCPA). 2019.(https://www.caprivacy.org/)

ユヴァル・ノア・ハラリ（柴田裕之訳）.『ホモ・デウス:テクノロジーとサピエンスの未来』. 河出書房新社. 2018.

E. Harbinja. Post-mortem privacy 2.0:theory, law, and technology, International Review of Law, Computers & Technology,31:1, 26-42, 2017. DOI:10.1080/13600869.2017.1275116. (https://doi.org/10.1080/13600869.2017.1275116)

Natasha Chu. Protecting Privacy after Death. 13 Nw. J. Tech. & Intell. Prop 255. 2015.
(https://scholarlycommons.law.northwestern.edu/njtip/vol13/iss2/8)