人間中心の人工知能―再考

Human-Centered AI -- Various Perspectives

国立研究開発法人産業技術総合研究所　フェロー
辻井　潤一

国立研究開発法人産業技術総合研究所　フェロー、英国マンチェスター大学教授、東大名誉教授、国際計算言語委員会（ICCL）委員長、AAMT／Japio 特許翻訳研究会委員長

j-tsujii@aist.go.jp

1　はじめに

国立研究開発法人産業技術総合研究所が人工知能研究センター（AIRC）を設立し、私がそのセンター長になった当時（2015 年）から、人工知能が幅広い分野の基盤技術となり社会、産業、生活を大きく変革するという議論が盛んであった。2022 年 11 月に ChatGPT が公開され、それに続く生成 AI の進展によって、この変革は、予想を上回る速度で加速している。変革は、社会生活全般に広がり、産業や教育だけでなく、我々の働き方、生き方そのものを変えようとしている。人工知能は、原子力・核技術と匹敵する大きな可能性と危険性を内包した技術との認識が広がっている。

技術と社会のかかわり方は私の専門外であるが、本稿では、専門家からは素朴にすぎるかと思うが、人工知能をどのように社会に受け入れていくかについて、個人的な思いを書くことにする。

2　人工知能と社会

2018 年のカナダ、2019 年のフランスでの G7 サミットで人工知能を取り巻く問題が議論され始めて、それ以降、人工知能は G7 サミットで継続的に議論されている。私が人工知能と社会という問題に興味を持ったのは、フランスでのサミットの準備会議がパリで開催され、日本学術会議の代表として G7 各国のアカデミーからの代表者とともに参加した時からである（図 1）。

人間中心の人工知能 (Human Centered AI　– 以下では HCA と略す) は、カナダでのサミット以来のキャッチフレーズで、パリでの会議でも議論の中核になっていた。技術の視点からではなく、技術によって影響を受ける人間の側から技術のあり方を考えるというのは、至極当然のことで、当時の私は、違和感なく受け止めていた。2 つのサミットでの会議を経て、その後、GPAI(Global Partnership for AI) という専門家集団の集まりが形成され、カナダとフランスに本部的な組織が置かれ、本年度からは、日本にもまとめ役の組織が作られている。

GPAI は、気候変動に関する専門家会議（IPCC）をお手本として、専門家主導のボトムアップな活動を目指している。それなりの活発な活動が行われているが、その中で、HCA は議論の枠組みを与える指導原理（Guiding Principle）の一つとなっている。

ただ、温暖化・気候変動とそれを阻止するための CO2 の削減という明確な方向性を持った IPCC が、求心力のある活動になっているのと比べると、GPAI の活動は多様でいささか分散的になっている。このことは、人工知能という技術の性格とともに、ＨＣＡの解釈に高い多義性があることが一因であろう。

注　2021 年に現れた ChatGPT は、従来の特化型人工知能と異なる汎用人工知能（Artificial General Intelligence）として、人工知能と社会との関連を議論する上でも大きな影響を持っている。ChatGPT の出現以後、ChatGPT やそれを支える大規模言語モデル（Large Language Model －LLM）としての GPT と類似のシステムが多数発表され、それぞれの固有名を持っているが、本稿では、それらを区別せず ChatGPT という用語で総称することとする。

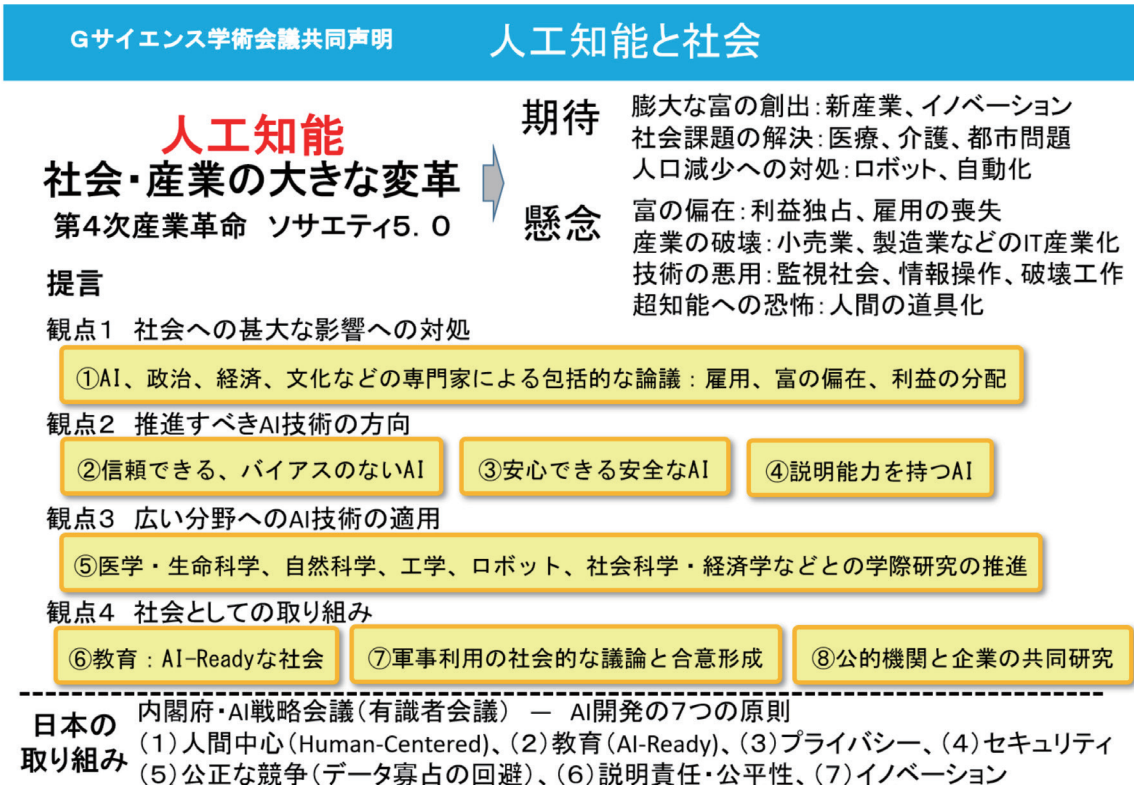

図1　パリ会議のまとめ（当時の安倍首相への報告スライド：2019年）

人工知能は非常に幅広い技術の総称であり、どのような技術と応用を想定するかで、その社会への取り込みの議論も変わり、HCAの解釈も変わる。「人間中心の」の「人間」も総称であり、個々の人間はそれぞれに異なった文化的背景、価値観、信念、利害関係を持っている。人間と人工知能という、多義性の高い2つの関係を総括的に論じようとする議論は、発散的にならざるを得ない。

3 人間中心の人工知能―その多義性

人間と人工知能の多義性だけでなく、「人間中心の」（Human-centered）のcenteredの解釈もそれほど自明ではない。当初の私の漠然とした解釈とは違って、「人間」を上位に置き、「人間」が人工知能を制御・管理するという解釈で議論が進むこともある。キリスト教の影響からなのか、万物の長としての人間が、人工知能を管理しなければならないという意識が、欧米では強いのではないか、と感じることもある。

万物に知を認めて、人間と動物やほかの生物との連続性を普通に受け止め、それぞれに知的能力をもつと考える文化的な背景を持つ者の感性からは、違和感がある。よく言われるように、日本では、無生物までに霊魂を認めるアニミズムがロボットや人工知能にも拡大されたメカノアニミズム的な心の動きがあり、日本社会での人工知能とのかかわり方に影響を持っている。

高齢者の生活の質を向上させる対話ロボットを目指す日本とヨーロッパの共同研究プロジェクト（e-Vita：図2）の対話分析では、ヨーロッパと日本の高齢者の対話行動に差異があることが認められている。日本の高齢者の発話には、対話ロボットをパートナーとみなして、自らの行動を対話ロボットに報告し、激励や称賛、同意を求める発話が多い。また、対話ロボットの誤解にも寛容的で、少しでも自然な発話をすると、ロボットを子供のように見なして、その成長を喜ぶ、といった傾向もある。人工知能を人間が管理するというよりも、人工知能をパートナーとしてみる心の動きが顕著である。

ChatGPTが現れる以前とそれ以後では、HCAの議論も変化している。ChatGPTより以前は、自動運転のように人間から独立して自律的に動作する人工知能をどう社会に取り込むかが議論の中心になっていた。そこでは、

1. 誤動作をした場合の責任の所在をどう考えるのか？
2. 人工知能の動作や判断の根拠をトレースできるのか？

【目的】

対話モデリング技術やビッグデータ解析、エモーショナルコンピューティング等の最新技術を駆使し、高齢者個別の①インタラクティブな生活アドバイスシステムの構築と、本システム開発により得られた技術要件の②国際標準化のための提案作成をもって「アクティブ・ヘルシー・エイジング(AHA：活力ある健康的なエイジング)」の課題とされる認知機能、身体活動、心的健康、社会的交流の改善を図り、高齢者のセルフケア能力の向上を目指す。

【実施内容】

・高度なIoT，NLP，AIに基づくサブシステム（スマートリビングシステム、対話システム等）開発と、それらを統合させた仮想コーチングシステムの構築／相互運用性に関する一連の標準・規格を欧州・日本双方で開発
・欧州と日本の実生活に、個々のウェルビーング、心身の健康状態のモニタリング、社会的交流の新しいサービスを提供
・エイジングを前向きなプロセスとして認識し、有意義に年齢を重ねるための実用的システムの提案、設計
・参加4ヵ国6箇所の概念実証研究実施から、国や文化的背景の多様性に対応したユーザー受容性を検証
・欧州と日本において、産業界や地方公共団体、ICT企業等と連携から市場にサービスを提供し社会実装を促進

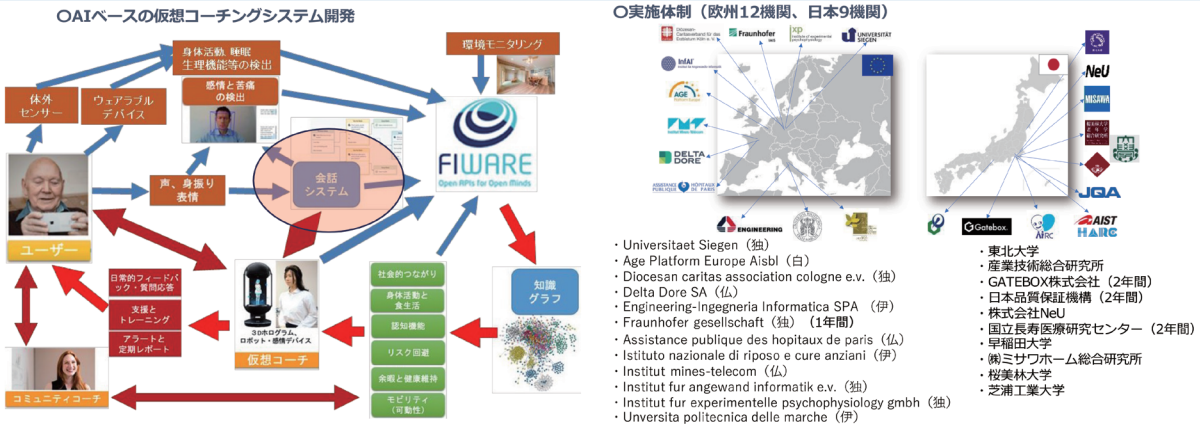

図２　SCOPE 国際標準化型 日 EU 第５次共同研究　e-VITA（EU-Japan Virtual Coach for Smart Aging）

3. 自律的な人工知能を人間がどのように制御するのか？

といった責任の所在、説明可能性 (Explainability) や制御可能性（Controllability）の観点からの議論が盛んであった。これらの論点は、以下でも議論するように、人と人工知能の協働という文脈で、重要となっている。さらに、ChatGPT 的なシステムは、

1. 一つのシステムの影響が社会のひろい範囲とさまざまなユーザに広がったこと、
2. 言語でコミュニケーションする技術がユーザの世界観や認識、価値観に大きな影響を持つ

ことから、ＨＣＡでのユーザとしての「人間」をどうとらえるのかが改めて問われることとなり、議論の範囲が大きく拡大することになった。

4 人工知能と人間―2 つの見方

特定の環境での特定のタスクへ適用されていた従来の特化型人工知能では、個々のシステムによって影響を受ける人間を限定できることが多く、影響を受ける人間の集合を想定することで、HCA の具体的な議論が可能であった。

これに対して、ChatGPT という汎用人工知能（AGI―Artificial General Intelligence）は、様々な状況での様々なタスクに適用され、それによって影響を受ける人間の集合が特定できない。文化的背景、信念、価値観、目的が様々に異なる人間がユーザとなり、様々なタスクの遂行に影響を受ける。

そのために、議論は、人工知能が持つ知能と人間の知能とが受け持つ範囲、そもそも「知能とは何か」を考え、人工知能と人間という異なった知能がどのような関係を持つのかという、抽象度の高い議論となる。

人間と ChatGPT のような人工知能との関係については、少し極端なまとめをすると、２つの見方がある。一つは、人工知能を過大評価するもので、

ChatGPT は世に流通する膨大なテキスト（知識）に基づいて判断しており、そのような膨大な情報処理能力を持たない人間の判断よりも合理的である

とする。この見方は、人工知能の判断を人間のそれよりも上位におくことで、人間を超える超知能の出現というシンギュラリティの主張にもつながる。

これに対して、もう一方の見方は、現在の汎用人工

知能は幻覚（Hallucination）や倫理感や適切な価値観の欠如によるバイアスなど不完全なものであり、「人間」による管理と訓練が不可欠であるとする。この見方は、無意識的ではあるが、「心」を持った「人間」は一様に優れた価値観や真理性を持つ、「人間」を神から理性を与えられた特別な知的存在とする立場につながる。

汎用人工知能が実際に使われるようになって、この２つの極端な見方は徐々に修正されつつあるが、HCA の議論では、無意識的ではあるが、この２つの流れが超知能への恐れとこれを人間が管理しなければならない、という形で現れることもある。

5 人間知能のバイアス

よく知られているように、人間の判断はそれほど優れたものではない。人間の判断には、情報アクセスや処理能力の限界から、膨大な利用可能な情報のごく一部に基づいて判断する限定合理性 (Bounded rationality) がある。膨大な可能的にはアクセスできる情報のごく一部のみを参照する人間の判断には、さまざまな認知バイアスがある（図３）。

実際、自分の考えや意見に合った情報のみを考慮する確証バイアスは、インターネットや SNS での情報流通に深刻な問題を引き起こしている。個人の嗜好に合わせて情報をフィルタリングする人工知能の機能や SNS での同じ意見のグループの形成によって、このバイアスが増幅され、社会の分断を引き起こしている。

人工知能ブームが本格化する前のビッグデータの解析技術は、人間が処理できない膨大なデータを解析し、人間に可視化して提示することで、人間の認知バイアスを矯正するという色彩があった。

定量的な結果の可視化と定量化の理論的根拠が明確であることから、分析者は自らの認知バイアスを矯正し合理的な判断を行う、いわゆるエビデンス・ベース・アプローチ (EBA) が可能になった。人間の持つ様々な認知バイアスを排除するために科学的知見に基づく判断をすべきとする EBA の中で、データと明確な数理理論によるビッグデータ解析が果たした役割は大きい。

ビッグデータ解析や EBA は、人間が処理しきれないデータを分析し、その結果を人間が解釈し納得して自らの判断に反映することを前提とする。重要なことは、データ解析の過程を人間が理解し、その結果に納得して自らの判断に反映できることである。

機械学習や深層学習を基盤とする人工知能では、ビッグデータ解析からさらに一歩進んで判断までも人工知能が行う。ビッグデータ解析の場合でも、背後にある数理理論の理解がないと納得性が低下するが、人工知能は、

1. 人間の介在なしに判断までを行う
2. 深層学習が判断に至る過程に強いブラックボックス性がある

ことによって、人間の納得性が低下し人間との協調がむつかしくなっている。人工知能の説明性 (Explainability) が重要視される所以である。

実データではなく、テキスト集合に基づく ChatGPT

確証バイアス
自分に都合のいい情報だけを集めて、それにより自己の先入観を補強するという傾向。
根本的な帰属の誤り
状況の影響を過小評価し、個人特性を過大評価して人間の行動を説明する傾向。
正常性バイアス
自然災害や火事、事故・事件など何らかの被害が予想される状況下にあっても、自分にとって都合の悪い情報を無視したり、「自分は大丈夫」「今回は大丈夫」などと過小評価したりしてしまう傾向。
生存者バイアス
何らかの選択過程を通過した人・物・事にのみを基準として判断を行い、そうでない人・物・事は見えなくなるため、それを見逃してしまうこと。
アンカリング
ある事象の評価が、ヒントとして与えられた情報に引きずられてしまう傾向。
フレーミング
情報の一部を切り取って、情報発信者の意図にあう情報表現をされると、それにより引きずられてしまう傾向。

図３　人間のもつ様々な認知バイアス

のような汎用人工知能では、この納得の保証がさらに困難になる。世に流通する言語テキストは、そもそも、著者の価値観や見方を反映した解釈の結果であり、いわゆる客観的な真理性に欠けるものである。言語のみを使って判断にいたった推論を外部のデータベースなどを参照することで、その正しさを人間に示していく機能、いわば、言語による汎用人工知能とデータに基づく人工知能の融合によって、説明性を向上することは、必須の研究課題となる。

6 汎用人工知能の受動性

「膨大なテキスト集合に基づく汎用人工知能の判断は、人間の判断よりも合理性がある」という神話は、徐々になくなってきている。いわゆる幻覚現象など、大規模なテキスト集合から言語表現の背後にある意味空間を学習し自然なテキストを生成することと、真理性が保証されたテキストを生成することは、別問題である。

実際、あからさまな誤情報を生成する幻覚現象だけはなく、たとえば、科学論文の要約や参照関係の生成においても、その分野の専門家から見ると誤りとされるものをしばしば生成する[1]。テキスト要約は、ChatGPT の有用な機能であるが、内容を理解し要約をする専門家の要約とは本質的に違う。重要な要約の場合には、生成結果の要約を人間が精査するという協働が必要である[2]。

ChatGPT のような汎用人工知能は、与えられた学習用のテキスト集合の真偽性や倫理的な正しさを吟味することなくそのまま受け入れて動作する受動的な知能である。フェイクニュースにみられるように、そもそも人間が作り出すテキストには真理性が疑わしいものやテキストの真理性の判断が発信者と読み手の立場や価値観に依存するものも多い。

汎用人工知能の運用では、学習用の大規模テキスト集合を用意する段階で不適切なテキストを取り除くクレンジングや生成結果のフィルタリング操作という外付けの処理が必要となっている。

また、現在の汎用人工知能の開発では、学習用のテキスト集合だけでなく、その動作を人間の価値観との整合性をとるための訓練を行う。自然なテキストを生成できる大規模言語モデル（Large Language Model-LLM）の動作を要約や対話など個々のタスクに適応できるようにする訓練である。この訓練過程は、人間の価値観と人工知能の価値観をそろえるという意味で、アライメントと呼ばれる[3]。また、LLM の学習用テキストとは別に正しさが吟味されたテキスト集合を外付けで与え、生成過程に使う RAG(Retrieval Augmented Generation) などの技術研究も盛んである。

学習テキストの慎重な準備、アライメントの必要性、RAG などの技術は、現在の汎用人工知能自体には、与えられたテキストの真理性や倫理面を吟味する能動的な能力はないことを意味している。

7 HCA と社会的コンセンサス

HCA における「人間」の解釈には、政治的・社会的な側面もある。特にヨーロッパでは、人工知能技術を開発し運用する主体としての巨大 IT 企業や国家に対置して、それにより影響を受ける市民の権利や安全を保護する意味合いが強い。人工知能に関する議論に「民主主義的な価値観に基づく」という政治的な色彩がある主張が入りこむが、現実にブラックボックス性のつよい技術が、権威主義的な国家や巨大 IT 企業に独占される危惧はある。

この危惧は、人工知能の中でも汎用人工知能のように人間の価値観や認識に直接影響を与える技術においては、とくに顕著である。巨大 IT 企業も、この危惧には自覚的で、自らを規制する動きがあるが、やはり技術を利益優先の民間企業にすべてゆだねることはできないであろう。

ヨーロッパには、アメリカや中国の技術覇権から自国産業を守る意識からか、民主主義的な価値という言葉で市民の権利保護や安全の保障に重点を置き、技術覇権をもつ主体を規制する態度もみられる。これに対して、アメリカは、自国の企業を考えて、イノベーションや新しい産業の育成の観点から自由を強調し規制的な動きをけん制するなど、同じ民主主義的な価値をうたいながらも、その解釈には違いがみられる。

このように「民主主義的な価値観」というものも、それほど自明なものではない。それぞれの社会の産業の在り方や価値観の差でその解釈が変わる。また、同じ社会の中でも、ある集団で差別とされるものが、別の集団では社会の安全のために必要と考えられることもある。前述のアライメントの過程でも、アライメント作業に従事する人間の訓練、言い換えると人間が持つ多様な価値観

をそろえる必要がある。個々の人間の持つ多様な価値観を、ある組織が望ましいとする価値観を設定し、それに一致させようとすることは、許容できないでは、とおもう。より幅の広い社会的なコンセンサスを構築する機構が必要となろう。

8 おわりに

本稿では、HCA には多義性があり総括的な議論には限界があることを指摘してきた。7 月初めに産総研が主催して、ヨーロッパ、北米、シンガポール、オーストラリア、ナイジェリアなどから ISO 標準化や AI ガバナンスに従事する専門家が多数参加して「Trustworthy AI 実現に向けた AI 標準化」の会議があった。信頼できる AI(Trustworthy AI) も、HCA と同様に多くの解釈が可能で、AI が適用される分野やそれぞれの文化的背景、価値観により、議論が影響を受ける。この会議でも、信頼できる AI のための人工知能の規制や標準化が

1.Global vs Regional：世界全体と地域ごとの個別性
2.Horizontal vs Vertical: 人工知能一般と適用分野ごとの議論の切り分け

の 2 つの軸で議論されていた。1 は地域ごとの文化的背景、価値観、産業構造などの差、2 は個々の適用分野での技術の特殊性から、「信頼できる人工知能」を総括的に議論する困難さが浮き彫りになっていた。会議の最後に、EU の AI Act のような総括的な議論や規制が可能だと思うかという質問には、大多数が否定的であったことが印象的であった。

また、EU の AI Act は、リスクを中核にして EU 市民の安全性を守るという観点からの規制色が強い。個人的には、人工知能の適用がもつリスクを議論するのは当然であるが、人間知能の不完全さがもたらすリスクを人工知能が緩和するというポジティブな面も考えること、言い換えると、人間と人工知能とのあるべき協調を考えるべきではないかとの印象をもった。

人工知能、人間のそれぞれが不完全な知能であることを考えると、2 つの知能がお互いを補完する形で協働し、系としてよりよい知能活動をすること、そのためには、安全性の観点だけでなく、人工知能の動きを人間が納得するための説明可能性、あるいは、人間の意図で人工知能の動きを制御できる制御可能性といった別の論点も重要になるであろう。

人工知能と人間知能の補完的な協働とは逆に、人工知能の故意による悪用、あるいは、誤用から、人工知能が人間知能の弱点である各種の認知バイアスをさらに拡大する可能性も否定できない。言語による情報の生成を行う CahtGPT 的な人工知能は、人間の認識そのものに影響を与えることから、その可能性はさらに大きくなろう。

人工知能と人間知能の負の連鎖を防ぐための社会的な制度とそのための技術とはどのようなものかの議論が必要になっている。

参考文献

[1] Walters, W.H., Wilder, E.I. : Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep 13, 14045 (2023). https://doi.org/10.1038/s41598-023-41032-5

[2] Hannigan,T. R., McCarthy I. P., Spicer A. : ,Beware of botshit: How to manage the epistemic risks of generative chatbots, Business Horizons,2024,

[3] Ouyang, L., Wu, J., Jiang, X., Almeida, Di., et.al. : Training language models to follow instructions with human feedback,arXiv:2203.02155, 2022