F-030

人格を持つAIエージェントの自我と意識ー唯識とI-JEPAの応用機序

Cogito and Consciousness of AI Agents with Personality - a Mechanism to Apply “Yuisiki” (Consciousness-only) and I-JEPA

水野義之[1]

Yoshiyuki Mizuno

1.はじめに

本稿では、近年のAI技術を背景として、今後に予想される自律的なAI技術として「人格を持つAIエージェント」という概念を提案し、またその内容について議論する。

AI技術を使って何かのタスク目的を達成するためには、自律的なAIエージェントに「人格」を付与することが必須である。なぜならタスクは人間のために行われるからである。何が人間のためであるか、それは人間にしか分からない。また人間にも、それは容易には分からない。従って「人格」を持った存在にしない限り、それは想定し判断できないと考えられる。すなわち有用なAIシステムのエージェントに対しては、自律的動作を内発的かつ適切に実現するための「人格」が必要である。

そこで本稿では「人格を持つAIエージェント」の設計において、2つの重要な要素が、新たに必要だと指摘する。第1に自我の存在、第2に自我を駆動する意識と無意識の存在である。ここで第1の自我とは、デカルトの自我 (コギト Cogito)、または仏教哲学の唯識論の末那識(まなしき)をモデルとする。第2の意識と無意識は心理学でも知られるが、ここでは唯識論で知られる阿頼耶識(あらやしき)をモデルとして考える。

これらの自我と意識・無意識を実装できて初めて、システム内からシステムを自律的かつ自己無撞着に統御できると考えられる。またこれらを含む諸要素の統合には、Meta社のY.LeCunによるI-JEPA[12]が有用であると考える。そこでこの論考では、このようなシステムを着想した背景と、その実装可能性について議論する。

2.研究の背景

AI技術の最近の進展は激越である。歴史的には過去200年の間に機械的・電気的に発明されたネットワークには3種類あり、鉄道網、道路網、通信網である。この意味で最近のAI技術の進展の激越さにも既視感はある。しかし従来の網は、そのどこかに人間が居た。これが網の運用の暴走に、一定の歯止めを掛けることが出来た。

ところが近年のAI技術の問題の一つは、自動化が進んだ結果、自律的なAIシステムがいずれ実現されること、つまり人間がどこにも居なくなることである。従って仮に暴走しても歯止めがかからない。これは大きな問題である。

これを人間側や設計者の問題だとして、人間にボールを預けるわけにはいかない。なぜなら今後のAIシステムは自律性を持つことが予測されるからである。例えば山川[1]はこの問題を「ポストシンギュラリティ共生学」という形で社会的に警告を発している。またその対応策をも検討している[2]。これはAIのAlignment問題だと考えることも出来る。しかし例えばIEEEのEAD (Ethically Aligned Design) の枠内に問題は収まらない。なぜなら自律的な判断は状況依存的であり、これを事前の設計では制御できないと考えられるからである。

この問題は、自律的なAIエージェントに「人格」を付与することで解決可能である。なぜなら、いかなるタスクも人間のために行われ、しかも何が人間のためであるか、人間にしか分からないからである。実際これは、人間にも分からないことが多い。しかし少なくとも「人格」を持った存在にしない限り、システムはそれが人間のためかどうか、原理的に判断できないと考えられる。もちろん自律的AIシステムに「人格」を持たせたからといって、暴走の問題が解決できるとは限らない。しかし問題はそれからだ。その後なのだ。言い換えれば、AIへの「人格」付与は十分ではないが必要、ということである。

山川[2]はこの問題を「慈悲深いAI」で解決できると考え、善良収束仮説（Benevolent Convergence Hypothesis）を提案している。本稿ではこのような概念的提案を、別の視点で提案しているということも出来るかもしれない。

本稿ではこのような「人格を持つAIエージェント」の設計において、2つの重要な要素が、新たに必要であると指摘したい。第1に自我の存在である。また第2に自我を駆動する意識と無意識の存在である。

ここで第1の自我とは、哲学者R.デカルトがいうところの自我（コギト、Cogito）を想定して良い。本稿では仏教哲学の唯識論の末那識(まなしき)をモデルとする。これは唯識論では概念上、一般的認識と自我とを同時に扱えるからである。また第2の意識と無意識は、近代以降の心理学でも知られる。本稿では唯識論の中で知られる阿頼耶識(あらやしき)を、そのモデルとして考える。これも同じ理由である。

それらの自我と意識・無意識は、デカルトの心身2元論的に振る舞い、システム内からシステムを自律的に統御することになる。そこで、これらの諸要素の統合には、米国Meta社のLeCunが提案するI-JEPAが有用であると指摘したい。以下、本稿ではこのような形で人格を持たせた自律的AIシステムについて、その着想を得た背景と、その実装可能性について議論を進める。

3.研究内容の背景とモジュール構造の可能性

本研究では、まず人間の行う次の４要素を明確に区別する。第1に、人間が扱う**対象**としてのデータ・情報・知識・知恵・統合（脳内オブジェクトの5形態）、これは[3]で整理し統合した。第2に情報処理としてこれらの対象を扱う人間の能力である**知能**（知性・理性・感性・悟性の4形態）、これは[4-8]で示した。第3に知能を作動・駆動するための**思考**（考えること、抽象化等の動作プロセス）、こ

[1] 京都女子大学

れは[9]で示した。第4にこれら全てを駆動する上で必須の統合的背景としての**意識・無意識**である。

この第4は新規で困難な課題である。実際、著者は文献[9]で、こう記した：「逆に意識の謎や心の謎まで、ここで相手にすると恐らく収拾がつかないと思われる。」

しかし本稿ではこの問題を、次の二つのguiding principle（指導原理）を得た上で、敢えて考察を進める。この二つとは、第1に唯識論、第2にI-JEPAである。

確かに人工知能（AI）は元々、哲学や人間学的な人間理解を抜きに語れないはずである[10]。また文献[11]は、本稿と並行する問題意識で出版されている。例えば[11]において、AIに個性を持たせる可能性が議論されている（同書 p.15, 26等）。ここでの「個性」とは、本論考でいうところの「人格」の別表現であると考えられる。

3.1.第1の指導原理：唯識論

このようなAIの個性あるいは人格を考える上で、本論考では次の二つのguiding principle（指導原理）を参照する。

第1に唯識論である。これは仏教哲学で知られる。唯識論では人間の認識は次の八段階で深まると考える。

まず人間のセンサーに対応する五感である。これは「眼耳鼻舌身意」（げんにびぜっしんに）と言われる6つのうち、最初の5つである（いわゆる五感）。6番目の「意」は、意識が覚醒して感覚センサーがオン（つまり知覚可能）を意味する。

7番目の末那識（まなしき）は、これら6つを駆動する自我である。この7つで自律的なAIマルチエージェントモデルに対応すると考えられる。この末那識（まなしき）はデカルトの自我に対応し、「考える我」を見つめる我である。

8番目は阿頼耶識（あらやしき）である。これは末那識（まなしき）の背景にある意識・無意識である。その存在は明らかだといえる。心理学で明らかにされつつあるとも言える。ここでフロイトやユングを正しいといっているのではない。仏教哲学の唯識論では、この問題が理路整然と言語化、対象化されている。

自律的なAIエージェント設計では、この部分を過去の行動記録（AIのライフログ）として、これをその当該AIエージェントシステムの個性すなわち唯一無二で一回性を有する「人格」（人格形成）であると定義できる。人間の場合、過去の経験や出会った人、考えた事、行動、場所、国際環境、政治環境、地域環境、保護者や家族、地域・国の地理・歴史・文化等が人格を形成する。その対応物である。

3.2.第2の指導原理：I-JEPA

第２はYann LeCunが提唱し推進するI-JEPA（Image-based Joint-Embedding Predictive Architecture）[12]である。この特徴は、画像を大枠のカタマリで捉えて、これをエンコード、学習、デコードする仕組みである。画像の場合、大枠で捉えることは、個別のピクセルではなく画像の特徴を抽象化することに対応する。これは本著者の提唱する、思考とは抽象化であるとの指摘に対応する。このため、I-JEPAは、思考モジュールの設計とその統合的なモジュール構築に有用であり、取り入れることができると考えられる。

この実装は、今後の課題とする。

4.おわりに

本稿では、最近のAI技術の先に予測される、自律的AIエージェントにおいて、「人格」の付与が、社会的に有用、かつ必須であると考えられることを議論し、示唆した。

またAIエージェントに、仮に「人格」を付与するならば、その際に現状では欠落している二つの要素があることを指摘した。第1に自我、第2に意識・無意識である。

本稿では、それらの２要素を動作可能なシステムとして組み込んで実装する上で、有用と思われる指導原理が、二つあることを指摘した。第1に、自我と意識・無意識の両方について、仏教哲学で知られる唯識論である。すなわちそこでの末那識（まなしき）は自我に対応すること、また阿頼耶識（あらやしき）は意識・無意識に対応することである。第2に、画像処理で抽象化レイヤーを導入した上で、機械学習させるというI-JEPAの提案が、本著者が提案している思考（すなわち抽象化）のアーキテクチャに対応することを指摘し、この方法で人間により近い、思考モジュールとその統合的構造を設計出来るという可能性である。

仮に「人格を持つAIエージェントの自我と意識」に対する設計指針が、ここまでの議論で明確になったとするならば、次の課題は実装段階である。これは今後の課題として残されており、鋭意試みたい。

参考文献

1. 山川宏, “ポストシンギュラリティ共生学にむけて”, 人工知能学会第二種研究会資料 2024 (AGI-027), 249-256, 2024-08-03.
2. 山川宏, “NAIAビジョンの提案”, 人工知能学会第二種研究会資料 2025 (AGI-029-03), 2025-03-17.
3. 水野義之, ”情報社会における「情報」の発展モデル”, 日本社会情報学会第24回全国大会研究発表論文集, pp.184-187 (2009).
4. Y.Mizuno, “Modeling of human intelligence applied to general education of informatics in AI era”, AXIES 2020.
5. 水野義之, ”AI（人工知能）の理解を目的とする「人間知能」のモデル化提案と情報教育の改善”、現代社会研究科論集：京都女子大学大学院現代社会研究科紀要、第15号), pp.77-87 (2021).
6. 水野義之, ”「人間知能」のモデル化におけるAI(人工知能)特に言語モデルGPT-3の位置付け”、AXIES 2021.
7. 水野義之, ”AI人材の情報倫理教育におけるインフォームド・コンセントを基盤とした能動的学修”, 私立大学情報教育協会 教育イノベーション大会 (2019).
8. 水野義之, “文科系大学におけるICT教育を再興する～アクティブ・ラーニング(AL)から人工知能(AI)の時代へ”、阿部勘一編著『ICT教育再考～文科系大学におけるICT教育の現状と課題』, noa出版 (2020).
9. 水野義之, ”AI/LLM(大規模言語モデル)の時代における機械・人間の「理解モデル」の提案”、AXIES 2023.
10. 水野義之, “AIっていったい誰なのよ：いまのAIは「アホ」なのか?─たかがAI, されどAI”, RAD-IT21 WEBマガジン, https://rad-it21.com/ai/mizuno20180611/ (2018).
11. 人工知能学会監修, 三宅陽一郎, 清田陽司, 大内孝子, 共編, “人工知能と哲学と 四つの問い", Ohmsha, 2024.
12. M.Assran et.al., “Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture”, https://doi.org/10.48550/arXiv.2301.08243 (2024).