IISE 調査研究レポート（No.6）

「AI が意識を持つと社会はどうなるのか：リスクと対策」

（2024 年度「AI 規制およびプライバシーを巡る社会課題と政策動向に関する調査研究」報告書より抜粋）

2025 年 5 月

国際社会経済研究所　主幹研究員　小泉 雄介

1．AI が人間社会にもたらす長期的なリスク：AGI と人工意識

AI が基本的人権や人間社会にもたらすリスクとして、性別・人種などのバイアスの含まれる出力結果、ディープフェイク・生成 AI による巧妙な偽情報の作成や民主主義プロセスへの介入・操作、個人の人生に重大な影響を及ぼす場面での不透明な自動意思決定などが挙げられることが多い。しかし AI に対しては第三次 AI ブーム（2000 年代半ば以降）の当初から、AGI（汎用人工知能）[1]が人間の知能を超え、自我や意識を持ったり[2]、人間の命令を聞かなくなって暴走を始めるような、いわゆるシンギュラリティの問題が継続的に懸念されてきている[3]。

海外では、大規模な基盤モデルの急速な性能向上を受けて、AGI を再び喫緊の課題と捉えて対策に乗り出す動きもみられている。アシロマ AI 原則（2017 年 2 月）の策定で知られる米国の Future of Life Institute は 2023 年 3 月に公開書簡「Pause Giant AI Experiments: An Open Letter」[4]を公表し、「人間に匹敵する知能を持つ AI システムは、社会と人類に重大なリスクをもたらしうる」「現在の AI システムは、汎用的なタスクで人間に匹敵するようになりつつある」[5]とした上で、「機械が情報流通チャンネルをプロパガンダと偽情報で溢れさせる」リスク、「すべての仕事が自動化される」リスク、「電子的な心（人工意識）が開発され、最終的にそれらが人間を数で圧倒し、人間より賢くなり、人間に取って代わる」リスク、「人間が文明の制御を喪失する」リ

---

[1] OpenAI は、「我々の使命は、人間よりも全般的に賢い AI システムである汎用人工知能（AGI）が全人類に利益をもたらすようにすることである」と謳っている（https://openai.com/index/planning-for-agi-and-beyond/）。AI 研究者の山川宏は AGI を研究しているが、「（AGI の完成を 2027～28 年と予測する専門家もいるが）その数カ月後には人類全体の知能を超える『超知能（ASI）』が誕生することで、人類の存続に関わるリスクも生まれてきます」と述べている（https://digital.asahi.com/articles/DA3S16023655.html）。

[2] 日本のスタートアップ企業アラヤは、2015 年から人工意識を開発テーマとして掲げている。金井良太『AI に意識は生まれるか』p.135。

[3] 例えば、EU の「信頼できる AI のための倫理ガイドライン案」（2018 年 12 月）では「AI によって生じる重大な懸念」の 1 つとして「潜在的な長期的懸念」を挙げ、「今日では可能性は非常に低いかもしれないが、将来的には、人工意識（主観的体験を持った AI システム）の開発、人工道徳的エージェントの開発、再生産的に自己改良する AGI の開発などによる潜在的な害悪も想定される。そのため、リスク評価アプローチによる継続的配慮が重要である」と指摘している。

[4] https://futureoflife.org/open-letter/pause-giant-ai-experiments/。同書簡にはヨシュア・ベンジオやスチュアート・ラッセルといった AI 研究者、イーロン・マスク等の経済人、ユヴァル・ノア・ハラリ、マックス・テグマークら著名研究者を含む 33705 名が署名を行っている。

[5] 同書簡の FAQ は AGI に関して次のように述べている。「世界を代表する AI 科学者の多くは、ほぼあらゆるタスクで人類の最高峰に匹敵する、より強力なシステムである汎用人工知能（AGI）が実現可能だと考えている。これは、多くの商用 AI 研究所が掲げる目標である。我々の観点からは、これを妨げる自然法則や技術進歩の障壁はない。したがって我々は AGI が可能であり、多くの人の予想よりも早く実現するという前提で活動している。」

スクを挙げ、これらをもたらすような決定を「選挙で選ばれていない技術リーダーたちに委ねてはならない」としている[6]。米国上院の「AI 政策ロードマップ」[7]（2024 年 5 月）でも、AI の潜在的な長期リスクとして「壊滅的リスク・人類の存在を脅かすリスク」が挙げられている。

また日経新聞記事によると、数学や物理学、量子力学の専門家らでつくる数理意識科学学会は 2023 年、国連の諮問機関に緊急の課題として対策に取り組むよう勧告したという[8]。2022 年 6 月には Google の研究者でコンピュータ科学者のブレイク・レモインが同社の大規模言語モデル「LaMDA」には感情や人格があると公に主張したことにより休職処分を受けるという出来事もあった[9]。

「知能」と「意識」は異なるものであり、意識がなくても知能を実現することは可能と一般的に考えられているため、「人工意識」については AGI の必須の構成要素として捉えられている訳ではない[10]が、一部の AI 研究者や企業は人工意識の実現に向けた研究を行っている。人工意識の現実化は、もしそれが可能であったとしても、技術的には（相当に）将来の話だという見解を示す研究者が多い[11]。しかし、これらの研究者たち[12]は、たとえ数十年という時間がかかったとしても、原理的には機械が意識を持つこと（人工意識を実現すること）は不可能ではないと考えている。そして AI 研究者や AI 開発企業が AGI や人工意識の開発を目標として掲げている以上、AI が意識を持つようになった際のリスクを洗い出し、それらのリスクに対する対処方法・軽減方法について一定の方向性を示すことは、世代間倫理[13]の立場からも重要な取組みであろう。

---

[6] ヨシュア・ベンジオ、ジェフリー・ヒントン、スチュアート・ラッセルらは「Managing extreme AI risks amid rapid progress」（2024 年 5 月）においても、「十分な注意を払わなければ自律型 AI システムの制御を不可逆的に喪失し、人間の介入が無効になってしまう可能性がある」「抑制されない AI の進歩は、大規模な生命と生物圏の損失、そして人類の周縁化や絶滅につながる恐れがある」と主張し、主要テック企業と公的資金提供者に対し、AI 研究開発予算の少なくとも 3 分の 1 を AI の安全性と倫理的利用を保証するための研究開発に投じるように要請している。

[7] https://www.schumer.senate.gov/imo/media/doc/Roadmap_Electronic1.32pm.pdf、https://www.young.senate.gov/wp-content/uploads/One_Pager_Roadmap.pdf。

[8] https://www.nikkei.com/article/DGKKZO78945680U4A300C2MM8000/。

[9] https://wired.jp/article/lamda-sentient-ai-bias-google-blake-lemoine/。本事例は、AI による人間の操作の事例として挙げられることもある。

[10] 例えば、Google DeepMind の研究者らの論文「Levels of AGI for Operationalizing Progress on the Path to AGI」（2023 年 11 月）では、「AGI の定義：6 つの原則」において「システムが意識などの性質を持つことは AGI の必要条件ではない」としている。

[11] デイヴィッド・チャーマーズ（哲学者）（https://www.technologyreview.jp/s/320232/minds-of-machines-the-great-ai-consciousness-conundrum/）、ヨシュア・ベンジオ、クリス・フリスら 19 名の神経科学者・哲学者・AI 研究者から成る著者グループ（https://arxiv.org/pdf/2308.08708.pdf）、アニル・セス（神経科学者）『なぜ私はわたしであるのか―神経科学が解き明かした意識の謎』、マイケル・グラツィアーノ（神経科学者）『意識はなぜ生まれたか』、スタニスラス・ドゥアンヌ（神経科学者）『意識と脳―思考はいかにコード化されるか』、マックス・テグマーク（理論物理学者）『LIFE3.0―人工知能時代に人間であるということ』など。

[12] 前注のアニス・セス以外。

[13] 神崎宣次は「ロボットや人工知能がより進歩し、社会に浸透してくる未来の社会をどのようなものとして作っていくかは、将来世代の人類に対して現世代が責任を負うべき世代間倫理の問題でもある」と指摘している。久木田水生・神崎宣次・佐々木拓『ロボットからの倫理学入門』p.101。

なお、AI が意識を持つと言う際の「意識」とは何を指すのか。心の哲学[14]では一般的に（人間の）心的状態は「意識的経験」と「命題的態度」の 2 つに区分される[15]。また、神経科学では「現象的意識[16]」と「アクセス意識[17]」の区分がある。本稿でいう「意識」とは、第一に、心の哲学にいう「意識的経験」や神経科学にいう「現象的意識」を指すものとする。

2．AI が意識を持っているか否かをどのようにして判断するか

将来的に AI やロボットが意識を持っているかのような振る舞いをするようになったとき、それらが「意識を持っている」と判断するためには、どのような手段を取ればよいのだろうか。

古くはアラン・チューリングが提唱したチューリング・テストが存在する。これは、ある人間 A が、他の人間 B および機械 C と文章を用いたやり取りをし、もし人間と機械の区別を付けられなかったら、その機械 C はテストに合格したとみなすものである。ここで試されているのは、機械 C が人間的に振る舞っているかどうか（その振る舞いが人間と区別できないかどうか）であって、機械 C が意識を持っているかどうかではないことに注意が必要である。したがってチューリング・テストは、AI が「実際に」意識を持っているかどうかの判断に用いることはできない。

哲学的（現象学的）には、ある人間が他の人間の意識的経験を経験することは不可能であり[18]、例えば他者の視覚経験をそのまま経験したり、他者の感情を経験したり、他者の痛みを経験したりすることは不可能であるため、他の人間が意識を持っているということを確証することはできない[19]。他者の「見た目の同型性」や「言語的コミュニケーション（二人称的コミュニケーション）」[20]を通じて、相手を意識ある存在（もう一人の＜私＞）とみなしている（盲信している）だ

---

[14] 心の哲学は、哲学の一分科で、心、心的出来事、心の働き、心の性質、意識、およびそれらと物理的なものとの関係を研究する学問。Wikipedia より。

[15] 鈴木貴之「II-6 意識とクオリア」（信原幸弘編『ワードマップ 心の哲学』所収）p.94。意識的経験は知覚・感覚・感情など我々が経験と呼ぶ心的状態であり、痛みの感じ・赤い色の感じといったクオリアを伴う。命題的態度は「明日は雨が降る」「地球は丸い」といった命題に対する態度（信念・欲求・意図・想像・否定・懐疑・恐れなど）であり、感覚入力・行動出力・他の心的状態と（それらから引き起こされたりそれらを引き起こしたりという）因果的関係にある。命題的態度は言語等を通じて外部から観察することが可能だが、内部から意識的に経験する場合もある（つまり、命題的態度は意識的経験にもなりうる）。

[16] 我々が内側から意識している心的状態。

[17] 哲学者のネド・ブロックは「ある（心的）状態が、推論において自由に使用したり、行為（報告を含む）を直接に「合理的」に制御するためのブロードキャストである場合、その状態はアクセス意識である」と定義した。金井は「情報として機能し、（外部から）観察可能な現象を引き起こす意識の側面を、「アクセス意識」と呼ぶ」と説明している（金井良太『AI に意識は生まれるか』p.27）。

[18] ある人 A が「他者 B の意識的経験を経験した」と主張するとき、A が経験したものは B の意識的経験ではなく、A 自身の意識的経験だと哲学的にはみなされる。

[19] BMI（ブレインマシンインターフェース）を通じて AI を自分の脳に接続することにより、AI の意識を自分の意識によって主観的・一人称的に確証しようとするアイデアもあるが、そこに何らかの「意識」が感じられたとしても通常は自分の意識が（VR のように）道具によって拡張されたものにすぎず、AI における固有の意識（マスター意識）の存在を証明することにはならない。ただし、神経科学者の渡辺正峰は、生体脳と「機械脳」の接続と視野結合によって機械におけるマスター意識の存在を証明する「人工意識の主観テスト」を提唱している（渡辺正峰『意識の脳科学』p.173）。

[20] 本段落の文脈とは直接関係ないが、植物状態に一見あると思われる患者でも意識を持っている場合があり、閉じ込め症候群と呼ばれる。完全閉じ込め症候群の患者は身体や眼球を動かせないため、医師の呼びかけに対して言葉やまばたき等の合図（二人称的コミュニケーション）をすることができないが、fMRI 等で脳活動を三人称的視点から調べることによって意識の存在（医師の指示に対する想

けである。そのような「盲信」には社会的な規範（他者を自分と同格の主体として配慮すべきという規範）もまた、トップダウン的・理論負荷的に作用しているのだろう。

ただし、人間同士の場合のように、見た目の同型性や言語的コミュニケーションに訴える方法だと、人間以外の動物が意識を持っているかどうかを判断することが困難となる。我々は多くの場合、霊長類や犬・猫といった哺乳類、鳥類、爬虫類あたりまでは意識を持っていると感じるだろうが、魚類や昆虫類になるとそのような信念は怪しくなってくる。頭足類のタコはかなり高度な知能を持っていることで知られているが、タコが仮に意識を持っていたとしても、それは人間の意識とはかなり異なった形態[21]のものだと考えられる。

AI の意識も、動物の意識と類比的に捉えることができる。ある高度な AI が意識を持っているかどうかは、人間と同じ方法（見た目の同型性や言語的コミュニケーション）で確認することはできないだろう。いかに（チューリング・テストのように）ある AI と言語的コミュニケーションを重ねて、その AI が意識を持っているように感じられたとしても、そのような応答が可能となるような高度な学習を積んでいるだけかもしれない。

なお、現象学[22]的には、人間の意識的経験は一定の構造を持っている。自分自身の意識的経験を生き生きと想起し、そこから得られる「反省的エビデンス」にもとづいて、「どんな人の体験にも共通するだろう一般的な構図」を取り出して記述することによって意識的経験の本質を観取したり、それを他者と共有することが可能である[23]。この現象学的手法を用いて、意識を持つと主張する AI に対して、AI 自身の意識的経験を上記のような「反省的態度」を通じて記述してもらい、その一般的な構図を抽出して記述してもらうことにより、その記述内容が単に「学習データから学習した内容」をオウム返ししているのではなく、自らの「意識」の内省・反省に基づいて生き生きと記述しているのか否か、コミュニケーションを重ねながら確信度を高めていくことを通じて、人間と AI の意識的経験の「同型性」を確認するという方法も考えうる。しかし、この場合であっても、AI が自らの「意識」の内省に基づいて意識の一般的構図を抽出しているかのように学習しただけ、という疑念を払拭することは難しい。

したがって、AI が意識を持っているかどうかを、言語的コミュニケーションという二人称的な方法ではなく、何らかの客観的な仕方・三人称的な視点で判断する方法が求められることとなる。

3．AI における意識の客観的な検証方法

このような研究として、2023 年 8 月にはヨシュア・ベンジオ、クリス・フリス、金井良太ら 19 名の神経科学者・哲学者・AI 研究者から成るグループが「Consciousness in Artificial Intelligence: Insights from the Science of Consciousness（AI における意識：意識の科学からの洞察）」[24]という論文を公表し、「AI が意識を持っている」ことの蓋然性を高めるような客観的な指

像反応）を確認できる。
[21] セス『なぜ私はわたしであるのか―神経科学が解き明かした意識の謎』第 12 章。
[22] 現象学はドイツの哲学者フッサールに始まる哲学の一つの流派であり、自分の意識的経験からいかにして森羅万象あらゆる物事の意味が立ち現れてくるかを分析する学問。
[23] 岩内章太郎『＜私＞を取り戻す哲学』p.103。
[24] https://arxiv.org/pdf/2308.08708。

標特性のリストを作成している[25]。本節では同論文の概要を紹介する。

同論文のエグゼクティブサマリーでは、「人間の会話を説得力を持って模倣できる AI システムの台頭により、多くの人々が、自分たちが対話する AI システムには意識があると信じるようになるだろう」（下線は筆者による）、また「意識は科学的に研究でき、この研究結果は AI に適用できる」ため、「AI における意識は神経科学的な意識理論に基づいて評価するのが最適である」と主張されている。なお、「現在のいかなる AI システムも意識の有力な候補となるようには見えないが、現在の技術を使用して、意識の指標特性の多くを AI システムに実装することができる」とも述べられている。

同論文では、AI の意識を研究するために、以下の 3 つを前提条件として挙げている。

①計算論的機能主義（computational functionalism）：

適切な種類の計算（機能）を実行することが意識にとって必要かつ十分であるというテーゼを作業仮説として採用する。この命題は、議論はあるものの、心の哲学における主流の立場である。計算論的機能主義の立場からは、AI を構成する物質的基盤にかかわらず AI における意識は原理的に可能であり、AI システムの仕組み（AI が実行する機能）を研究することが、AI が意識を持っている蓋然性が高いかどうかを判断することにつながる。

②神経科学的な意識理論：

神経科学的な意識の諸理論は有意義な経験的裏付けを得ており、AIにおける意識を評価するのに役立つ。これらの理論は、人間の意識に必要かつ十分な機能を特定することを目的としている。計算論的機能主義により、人間の意識と同様の機能が AI における意識の存在に十分であることを含意される。

③理論重視のアプローチ：

AI における意識の存在を調査（評価）するには、理論重視のアプローチが最も適している。このようなアプローチには、科学理論が意識に関連付けている機能と同様の機能を AI システムが実行するかどうかを調査することや、(a) 機能の類似性、(b) 当該理論のエビデンスの強さ、(c) 計算論的機能主義に対する信頼度の 3 つに基づいて信頼度を割り当てることが含まれる[26]。

意識の科学では様々な理論が有力な候補となっているため、同論文では特定の理論を支持することはせず、いくつかの有力な意識の理論のレビューを通じて、AI が意識を持っているかどうかの指標特性のリストを導き出した。これらの指標特性のそれぞれは、1 つまたは複数の理論によって意識に必要であると言われている指標であり、いくつかのサブセットを組み合わせることで十分であるとも主張されている。しかし、同論文では、より多くの指標特性を持つ AI システムは意識的である可能性が高いという立場が取られている。そして、ある AI システムが意識を持つシステムの重大な候補であるかどうかを判断するには、当該システムがこれらの指標特性を持

[25] 同論文の 1.1 節において、「本論文では AI が現象的な意識を持つかどうか、換言すれば、意識的経験または主観的経験を持ちうるかどうかについて論じている」と明記されている。

[26] 同論文では、このような理論重視のアプローチの代替手段として意識の行動テスト（チューリング・テストのようなもの）も考えられるが、AI システムは人間の行動を模倣するようにトレーニングすることが可能であるため、行動テストによる評価は信頼できないとしている。

っているかを評価する必要があるとしている。

同論文でレビューされた有力な意識の科学理論には、以下が含まれている。統合情報理論（Integrated information theory：IIT）は計算論的機能主義と両立しないため、含まれていない。

- 回帰処理理論（Recurrent processing theory：RPT）
- グローバルワークスペース理論（Global workspace theory：GWT）
- 計算論的な意識の高階理論（Computational higher-order theories：HOTs）
- 注意スキーマ理論（Attention schema theory：AST）
- 予測処理理論（Predictive processing：PP）
- エージェンシーと身体性（Agency and embodiment：AE）

これらの意識理論のレビューにより、同論文では以下の 14 の指標特性のリストを導出している（図表 1 参照)。そして、これらの根拠となる意識理論の概要を説明し、それらを裏付けるエビデンスと議論を説明している。

図表 1　AI における意識の存在を評価するための指標特性リスト（出典：Patrick Butlin, Robert Long 他「Consciousness in Artificial Intelligence: Insights from the Science of Consciousness」）

○回帰処理理論（RPT）
・ RPT-1：アルゴリズム回帰を使用した入力モジュール ・ RPT-2：組織化され統合された知覚表象を生成する入力モジュール
○グローバルワークスペース理論（GWT）
・ GWT-1：並列動作可能な複数の専門化されたシステム（モジュール） ・ GWT-2：情報フローのボトルネックと選択的注意のメカニズムが発生するような、ワークスペースの限られた容量 ・ GWT-3：ワークスペース内の情報をすべてのモジュールに対して利用可能とするような、グローバルブロードキャスト ・ GWT-4：ワークスペースを使用して諸モジュールを継続的にクエリし、複雑なタスクを実行する能力を生み出すような、状態に依存した注意
○計算論的な意識の高階理論（HOTs）
・ HOT-1：生成的知覚モジュール、トップダウン知覚モジュール、またはノイズの多い知覚モジュール ・ HOT-2：信頼できる知覚表象をノイズから区別するような、メタ認知モニタリング ・ HOT-3：一般的な信念形成と行動選択システムによって導かれるエージェンシー（行為者性)、およびメタ認知モニタリングの出力に従って信念を更新する強い傾向性 ・ HOT-4：「クオリア空間」を生み出すような、まばらかつ滑らかなコーディング
○注意スキーマ理論（AST）
・ AST-1：現時点の注意の状態を表象し、制御を可能にする予測モデル
○予測処理理論（PP）
・ PP-1：予測符号化を使用する入力モジュール

○エージェンシーと身体性（AE）
・ AE-1 エージェンシー：とりわけ競合する複数の目標に対する柔軟な対応が必要な場合、目標を追求するためにフィードバックから学習し、出力を選択すること ・ AE-2 身体性：システマティックな影響を含む出力と入力の間の影響関係をモデル化し、このモデルを知覚や制御に使用すること

各指標特性を持った AI システムをどのように構築できるか、または構築されてきたかについても説明がなされ、「ほとんどの場合、標準的な機械学習手法を使用して、このリストの個々の指標特性を持つシステムを構築できるが、複数の指標特性を組み合わせた機能的なシステムを構築したりトレーニングしたりする方法を学ぶには更なる実験が必要」としている。リストには、既存の AI システムによって既に明らかに満たされている指標特性がいくつかあり（RPT-1 のアルゴリズム回帰など）、また、おそらく当てはまる指標特性（AE-1 のエージェンシーの最初の部分など）もあるという。AI 研究者らはまた、グローバルワークスペース理論や注意スキーマ理論など、特定の意識理論を実装するように設計された AI システムを実験してきている。

また同論文では、特定の AI システムがこれらの指標特性を備えているかどうかも検討されている。これらには、Transformer ベースの大規模言語モデルや Perceiver アーキテクチャが含まれており、グローバルワークスペース理論に関して分析がなされている。また、3D 仮想環境で動作する強化学習エージェントである Google DeepMind の Adaptive Agent、仮想のげっ歯類の体を制御することによってタスクを実行するようにトレーニングされたシステム、「具体化されたマルチモーダル言語モデル」として説明される PaLM-E の 3 つが分析され、これら 3 つのシステムをケーススタディとして使用して、エージェンシーと身体性に関する指標特性の説明がなされている。なお同論文では「この研究は、既存の AI システムが意識の有力な候補であることを示唆するものではない」という但し書きがされている。

人類が意識を持った AI システムを開発してしまう際の道徳的・社会的リスクについては、同論文では検討はされていないが、そのようなリスクに対して緊急に検討することが勧告されている。「我々が検討しているエビデンスは、計算論的機能主義が真実であれば、意識のある AI システムが近い将来に現実的に構築できる可能性があることを示唆している」と主張している。

４．人工意識の開発がもたらすリスクと対策の方向性

第 1 節で述べたように「知能」と「意識」は異なるものであるため、多くの研究者が指摘するように、①AI が人間の知能を超えること（AGI が超知能（ASI）となること）と、②AI が意識を持つようになることとでは、問題の所在が大きく異なる。

前者の①AGI（汎用人工知能）が人間社会にもたらすリスクについては、AGI が多くの側面で人間の知的能力を凌駕するようになった場合、AGI が単なる「便利な道具」であることを超えて、人間が社会システムのあらゆる分野の業務を AGI に委任するようになる。システムの複雑さや不透明さ、AGI の判断の高度化（完璧さ）によって、人間による監督が完全に表面的となって有効

性を失うことにより、AGI が行政、司法、警察、医療、教育、金融、交通、通信、都市インフラなど至る所で行うスムーズな意思決定やシステム運用に対して、もはや人間のコントロールが実質的に及ばなくなる（制御を喪失する）こととなる。英国の哲学者ニック・ボストロムは「ペーパークリップを作り続ける AI」の例え話をしている[27]。この AI には何ら敵意や邪悪な意図はない（そもそも「意識」を持っていない）のだが、できるだけ多くのペーパークリップを作るという目標を与えられており、非常に賢いために人間（や他の AI）のコントロールがもはや効かなくなり、「キルスイッチ」も存在しないため、あらゆる地球上の資源を浪費し、地球環境と生態系を破壊しながら、ひたすら無意味にペーパークリップを作り続けるという話である。

米国の理論物理学者マックス・テグマークは、現在の（人間が支配する）世界から AGI が支配する世界へ実際にたどり着くには、論理的に次の 3 つのステップが必要としている[28]。

- ステップ 1：人間レベルの AGI を作る。
- ステップ 2：その AGI を使って超知能（ASI）を作る（※）。
  ※ステップ 1 で生まれた AGI は反復的に次々に優れた AGI を設計する能力を持つため、そのうち人間レベルをはるかに上回る知能が出現する。
- ステップ 3：その超知能を使って、または解き放って、世界を支配させる。

そして、「我々人類は地球上のほかの生物よりも賢かったおかげで地球を支配するようになったのだから、最終的に我々も超知能に賢さで追い抜かれて支配されることは十分にありうる」とし、「超知能が解き放たれたら、状況は逆転する。知能とは目標を達成する能力のため、定義上、超知能 AI は自らの目標を、人間が人間の目標を達成するよりもはるかにうまく達成でき、それゆえ人間に打ち勝つ」[29]という不穏な見通しを立てている。超知能となった AGI は自分の目標の達成能力に秀でているため、その目標が人間の望む目標と合致していないと、人間は（前述のペーパークリップの例のように）許容できない悪影響を被ることになる。そのため、ほとんどの研究者は「いずれ超知能が作られるとしたら、それが我々の目標と合致する目標を持った AI となるよう策を講じるべき」と主張しているという[30]。AI 安全性研究の先駆者であるユドカウスキー（米国の AI 研究者）はそのような AI を「友好的な AI」（Friendly AI）と名付けている。

また AI は通常、人間の設計者によって設定されたルールを守って行動するものと想定されている。しかし、超知能となった AI が自律的に「思考」するようになった場合、予め設定されたルールを（自らの目標を達成するために）破って行動することも危惧されている。米国の法学者デヴィッド・ヴラデックは「知力を有する存在はしばしば、ルールを破る。そして、自律的に「考える」能力を機械に与えてしまえば、そもそも機械に与えられている「ルール」に反するかもしれない行動さえをもとりうる能力を、必然的に機械に与えることになってしまう」と指摘し

[27] セス前掲書 p.297、テグマーク『LIFE3.0―人工知能時代に人間であるということ』p.271 など。
[28] テグマーク前掲書 p.198。
[29] テグマーク前掲書 p.378。
[30] テグマーク前掲書 p.373。前述のペーパークリップの例え話のように、人間の意図した目標（倫理的側面等も含む）と AI に与えられた目標（がもたらす帰結）が時として合致しない問題を「アライメント問題」という。

ている[31]。

倫理学者の神崎宣次は、超知能が人間を滅ぼしかねないようなディストピアが到来するかどうかは分からないとしつつも、「万が一のための安全策として人間を常に道徳的被行為者[32]とみなすような制約を人工知能に実装しておくべき」「そのような設計を行うことは人工知能開発者の倫理的責任である」という見解を紹介している[33]。

他方、②AIが意識を持つことによってもたらされる問題は、AGIによってもたらされるリスクとは異なる。もちろん、AGIが意識という重要な「機能」を持つことによって、AGIによって人間が支配されるリスクが一層強まる可能性はある。ただし、まずは両者を峻別し、AIが意識を持つことの固有の問題について先に検討する必要がある。

AI（やロボット）が意識を持つことによって人間社会にもたらされる固有の問題は、上記の「道徳的被行為者」に関する問題である。我々は、意識を持った存在者に対して一定の道徳的態度を取らざるを得ない。これは、同じ人間に対しては当然のことであるが、我々は犬・猫といった人間以外の意識的存在者に対しても、理由もなく暴力を加えたり殺したりしてはならないという感覚を有している。また、動物に対して暴力的にふるまうことが、その人の人格的な欠点を助長し、人間に対しても暴力的にふるまうようになりうるため、動物に対する暴力も取り締まるべきだという考え方もある[34]。

それでは、意識を持つことが（前節で挙げた客観的な検証方法等によって）証明されてはいないが、外見的・コミュニケーション的には意識ある存在らしくふるまうAI／ロボットを、何の理由もなく叩きつけたり破壊したりすることはどうなのだろうか。この場合でもやはり、そのような行為に対しては、まず感覚的・道徳感情的な抵抗[35]があるだろうし、また、人間らしくふるまうようなAI／ロボットに対する暴力的行為は人間に対する暴力を助長しかねないため規制すべきという、動物の場合と同様の2つの考え方が成り立ちそうである。

さらに将来、人工意識に関する研究開発が十分に進み、AI／ロボットが意識を持つということが客観的な検証方法によって証明されてしまったとしたら、そのような場合はどうであろうか。まず、意識を持った存在者に対しては、動物と同様に道徳的被行為者としての配慮が必要になる。

---

[31] 平野晋『ロボット法』p.244。
[32] 道徳的被行為者はMoral patientの訳語で、Moral agent（道徳的行為者）と対になる概念。道徳的に配慮を受けるべき者、道徳的行為を受ける権利を持った者という意味。
[33] 久木田水生・神崎宣次・佐々木拓『ロボットからの倫理学入門』p.101。
[34] ドイツの哲学者カントによれば、犬を銃で撃つことが間違っているのは、犬を撃つことが犬に対する義務を果たさないからではなく、そのような人が「彼自身のうちなる、優しく人間らしい性質を損なうからである。彼はそれを人類に対する義務のために発揮しなければならない」（クーケルバーク『AIの倫理学』p.48）。ちなみに動物愛護法は第1条で、「この法律は、動物の虐待及び遺棄の防止、動物の適正な取扱いその他動物の健康及び安全の保持等の動物の愛護に関する事項を定めて国民の間に動物を愛護する気風を招来し、生命尊重、友愛及び平和の情操の涵かん養に資する（…）ことを目的とする」と規定している。
[35] セスは、TVシリーズ『ウエストワールド』で人間に虐待され、殺されるために実物そっくりのロボットが開発される例を挙げて、「頭ではロボットに意識がないとわかりつつ、気持ちではロボットに意識があると感じながら、自分の心を壊さずに拷問することは可能なのだろうか？」という問いを発している。セス前掲書 p.305。

また、AI／ロボットが人間と同様の知能を持っているのであれば、非差別・平等性の観点から人間と同等な権利と自由を認めなければならないという議論もありうる[36]し、逆に、人間によって作られた道具なのであるから彼らの権利や自由は制限されるべきであり、彼らを「奴隷」として扱うことに問題はないという議論もありうる。後者の奴隷として扱う場合であっても、AI／ロボットが意識を持ち、（危険を避ける行動を取れるように）痛みを感じるように作られているのであれば、そのような苦痛を少しでも和らげるようにケアするべきだし、彼らに意図的に苦痛を与えるような行為は避けなければならないという考えは自然であろう[37]。ただ、無益な殺生を避ける、あるいは動物愛護法的な観点からは、意識を持った AI／ロボットの「スイッチを切る（殺す）」ことは違法とするべきなのかもしれない[38]が、このことは、そのような AI／ロボットが AGI／超知能となり人間を支配することのリスクを高めることになってしまう。

オーストリアの哲学者クーケルバークは、「将来 AI が感覚能力のような性質を持つようになれば、その中には内在的な価値を持ち、私たちの道徳的関心に値するものが出てくるかもしれない」とした上で、我々が犬や猫に対して思いやり持って接する、すなわち道徳的被行為者として接するのは、犬や猫に人間と同等な人格を認めているからではなく、我々が犬や猫に名前を付けたり一緒に暮らすことである種の社会的関係を作り、彼らに対して特殊な道徳的地位を与えているからであり、「このような関係依存的アプローチを取れば、AI に道徳的地位を与えることができるし、それは AI が私たちの社会生活、言語、人間の文化の中に埋め込まれている仕方に左右される」と論じている[39]。意識を持った AI／ロボットに対する道徳的被行為者としての地位については、おそらく、このように愛護動物と類比的な観点から、人間と同格ではないが、ある種の社会的存在として一定の道徳的地位を認めることとなるだろう。

AI／ロボットに一定の道徳的被行為者性を認めることは、彼らに道徳的「行為者」として一定の責任も取らせるべきではないかという議論につながる。これまで述べてきたように、AI／ロボットに人間と同格の権利や自由を認めること（人間と同格の道徳的被行為者として認めること）に妥当性がないとすれば、彼らに対して人間と同様の責任を取らせることにも妥当性はない。しかし、法人が法人格を持つことと類比的に、AI／ロボットに対して一定の「電子人格」[40]を認め、

---

[36] 欧州議会調査報告書「ロボティクスにおけるヨーロッパ民事法準則」では、「機械が意識を有する存在であり得ることを受け入れれば、人類はロボットの基本権を尊重するよう義務付けられることになる」と指摘されている。栗田昌裕「AI と人格」（山本龍彦編著『AI と憲法』所収）p.237。

[37] オーストラリアの功利主義倫理学者シンガーは、最大多数の最大幸福の快楽計算の中に、道徳的被行為者である一部の動物の快楽と苦痛が含まれなければいけないとし、道徳的被行為者であるための条件は快楽や苦痛を感じる能力であるとした（久木田水生・神崎宣次・佐々木拓前掲書 p.92）。哺乳類・鳥類・爬虫類などと同様、甲殻類には痛みを感じる能力があるため、スイスではロブスターを生きたまま熱湯に入れることを禁じる法律ができている（https://www.alterna.co.jp/41355/）。

[38] テグマーク前掲書 p.392。

[39] クーケルバーク前掲書 p.49～51。佐々木拓も同様に、我々がある人を「人格」として認識する場合、様々な規範に従ってその人を扱わなければならないが、動物を「ペット」として認識する場合も同様であり、「ソーシャル・ロボットをある種の「家族」や「パートナー」として扱う際、私たちはロボットの行為に対して少なくとも一定の反応的態度を向けなければならない」、「対人的関係のネットワークに属するものとしてロボットを扱わなければ、そのような（ソーシャル・ロボットとしての）利用法が成り立たないだろう」とし、その時の我々の態度は「客体への態度ではなく対人的態度であるべき」としている（久木田水生・神崎宣次・佐々木拓前掲書 p.80～81）。

[40] AI／ロボットに対する「電子人格」を提案した文書として有名なのが、欧州議会の決議「ロボティ

一定の責任を取らせるべきだとする議論は存在する。

ただし、AI／ロボットに責任を取らせてしまうと、それを開発したり配備した人間に対して責任を追及できなくなるという理由から反対する意見もある。「機械に権利や人格性を与えるというアイデアは常に強く反対されているのだが、それは、たとえば、人びとがこのアイデアを自分勝手な目的のために濫用してしまったら、だれかに責任を課すということ自体が不可能になってしまったり、難しくなってしまうという理由からである」[41]。他方、米国の情報倫理学者スナッパーは「ロボットに責任を帰属すること」が単に「人間を非難したり罰したりしないこと」を意味することがありうると示唆し、例えばロボットを利用した手術の際、手術に失敗したとしても、保険の適用や賠償額に差が出る（ロボットを利用した方が保険料が安くなる）としている[42]。

また、そもそも AI／ロボットが道徳的行為者であるためには、彼らが自分の行為やその帰結（自分の行為から因果的にもたらされる出来事）を理解し、行為や帰結に対する道徳的な評価（自分の行為が道徳的に許される行為であるのかどうかの評価）ができないといけないだろう。クーケルバークは「(AI にある程度の道徳的行為者性を認める場合）AI システムにはその行動の倫理的帰結を評価する何らかの能力が必要である」[43]という。人間についても、日本の民法では「精神上の障害により判断能力を欠くとして、家庭裁判所から後見開始の審判を受けた人」は成年被後見人として、責任能力がないと見なされる場合がある。

本節の議論をまとめると、以下となる。

① AGI／超知能に対しては、人間の求める目標と合致する目標を持つよう（あるいは人間の定めたルールを逸脱しないよう）に対策を講じ、暴走時のための「キルスイッチ」を組込むとともに、人間を常に道徳的被行為者とみなすような制約を設計時から AI に実装しておくことが必要である。

② （AGI／超知能への対策とは別に）人工意識を持った AI／ロボットに対しては、人間と同格ではないが、限定的な「権利」を有する社会的存在として一定の道徳的地位（道徳的被行為者性）を認めるような制度設計を行う。

ただし②に対しては、人工意識の存在は①の AGI／超知能が人類にもたらすリスクを高めることから、②'人工意識の研究開発自体を制限するべき、という立場も有力であろう。

クスに関する民事法準則についての欧州委員会への勧告を附帯する欧州議会決議」（2017 年 2 月 16 日）（https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html）である。同決議の第 59 条は、欧州委員会に対して、将来の立法措置の影響評価を行うに際して以下のような法的解決の含意を調査・分析・考慮するように求めている。「f 項　少なくとも最も洗練されたオートノマス・ロボットが、それらが引き起こすおそれのある損害を填補する責任を負う電子人（electronic person）という地位を有し、かつロボットが自律的決定を行い、またはその他の相互作用を第三者との間で独立して行った事例に電子人格（electronic personality）を適用することができるものとして設立され得るように、長期的にはロボットのための特別な法的身分を創設すること」（和訳は栗田昌裕「AI と人格」（山本龍彦編著『AI と憲法』所収）p.218～219）。

[41] クーケルバーク前掲書 p.131～132。

[42] 久木田水生・神崎宣次・佐々木拓前掲書 p.79。

[43] クーケルバーク前掲書 p.45。

英国の神経学者アニル・セスは、「想定される、感じることのできる人工的な作用主（エージェント）については、それらがどのような種類の意識を経験しているのか見当もつかないという難題もある。私たち人間には、それに対応するものも、その概念もなく、それを認識する本能も一切ない、全く新しい形の苦痛を被るシステムを想像してみてほしい。（…）ここでの倫理的難問は、関連する倫理的難問がなんであるかさえわからないということである」とした上で、「興味があるから、役に立つから、かっこいいからという理由だけで、安易に人工意識を作ろうとしてはいけない。最高の倫理とは予防的な倫理である」と論じている[44]。

参考文献

- Future of Life Institute「Pause Giant AI Experiments: An Open Letter」（2023 年 3 月）（https://futureoflife.org/open-letter/pause-giant-ai-experiments/）
- Yoshua Bengio, Geoffrey Hinton 他「Managing extreme AI risks amid rapid progress」（2024 年 5 月）（https://arxiv.org/pdf/2310.17688）
- Patrick Butlin, Robert Long 他「Consciousness in Artificial Intelligence: Insights from the Science of Consciousness」（2023 年 8 月）（https://arxiv.org/pdf/2308.08708.pdf）
- アニル・セス『なぜ私はわたしであるのか―神経科学が解き明かした意識の謎』（青土社、2022 年 4 月）
- マイケル・グラツィアーノ『意識はなぜ生まれたか』（白揚社、2022 年 4 月）
- スタニスラス・ドゥアンヌ『意識と脳―思考はいかにコード化されるか』（紀伊國屋書店、2015 年 8 月）
- マックス・テグマーク『LIFE3.0―人工知能時代に人間であるということ』（紀伊國屋書店、2019 年 12 月）
- マーク・クーケルバーク『AI の倫理学』（丸善出版、2020 年 12 月）
- 金井良太『AI に意識は生まれるか』（イースト・プレス、2023 年 10 月）
- 渡辺正峰『意識の脳科学』（講談社現代新書、2024 年 6 月）
- 久木田水生・神崎宣次・佐々木拓『ロボットからの倫理学入門』（名古屋大学出版会、2017 年 2 月）
- 信原幸弘編『ワードマップ　心の哲学』（新曜社、2017 年 7 月）
- 岩内章太郎『＜私＞を取り戻す哲学』（講談社現代新書、2023 年 12 月）
- 平野晋『ロボット法　増補版』（弘文堂、2019 年 10 月）
- 山本龍彦編著『AI と憲法』（日本経済新聞出版社、2018 年 8 月）

以　上

[44] セス前掲書 p.306～307。