ＡＩと自由意志

櫻井成一朗

１　はじめに

現在、深層学習により開花した人工知能（ＡＩ）によって世界はＡＩ社会を迎えようとしている。日本では、二〇二〇年、レベル三の自動運転自動車が解禁され、世界初のレベル三認定を受けた市販車も登場した。レベル三の自動運転では、緊急時には運転者の対応が求められるが、レベル四やレベル五になればＡＩが運転の主体となる。近

い将来に、レベル四やレベル五の自動運転が可能になるとすれば、自動運転を制御するＡＩの責任について検討せざるを得ない時期が到来することになる。ＡＩ社会においては、自動運転自動車に限らず、ＡＩを搭載した自律機械の導入が進んでいくことは間違いない。自律機械が社会に浸透すれば、あたかも人間のように会話する自律機械の責任問題が生じることが予想される。自由意志をＡＩの責任として基礎づけるのであれば、ＡＩが決定性機械であることから定義上ＡＩは自由意志をもちえない。自由意志をもたないはずのＡＩが、深層学習がもたらした自然言語処理の進歩によって、二〇世紀にはＳＦの中だけで存在していた、コンピュータと人間が見かけ上音声で会話できるようになっている。従来地上で言葉を使えるのは人間だけに限られていたのに、ＡＩ社会においては、ＡＩも人間と同様に言葉を見かけ上駆使してしまうのである。本論文では、巧みに言葉を使うＡＩの自由意志について様々な観点から検討し、自由意志をもたないＡＩの責任について、法哲学および社会情報学の観点から議論し、ＡＩの判断が社会規範を遵守することを示す必要があることを示す。

２　ＡＩと自由意志の定義

ＡＩの定義は、研究者の間で統一されたものはなく、研究者毎に異なる定義を採用していると言われる。本論文では、ＡＩの自由意志を検討するためにＡＩの定義を明確にする必要があるので、ＡＩをコンピュータ上で実現されるソフトウェアであると定義する。コンピュータとは、入力に対して有限時間内で計算を終了し、結果を出力する機械、すなわちチューリング機械である。言い換えれば、ＡＩはチューリング計算可能関数であると定義することに他

ならない。チューリング計算可能であるとは、アルゴリズムが存在することであるから、知能のアルゴリズムを解明することがＡＩの目標なのである。

ＡＩには、大きく分けて**強いＡＩ**と**弱いＡＩ**の二つがある。強いＡＩとは心を実現するＡＩであり、強いＡＩ以外を弱いＡＩと呼び、心を実現したコンピュータは存在しないので、現在までに実現されているＡＩはすべて弱いＡＩである。最近では、汎用人工知能（artificial general intelligence）[1]と呼ばれる研究が活発に行われるようになっているが、ＡＧＩが強いＡＩであるかどうかは研究者により分かれている。ＡＧＩとの対比として、従来の弱いＡＩのことを特化型ＡＩと呼ぶこともあるが、現在のすべてのＡＩは特化型ＡＩでもある。

チューリング機械とは、無限の記憶容量を与えられ、有限ステップで計算する決定性機械であり、決定性機械とは、内部状態の遷移が入力によって確定している機械である。スーパーコンピュータも量子コンピュータも、一般のコンピュータよりも計算速度は圧倒的に高速であるものの、決定性チューリング機械であることには違いがない。理論上、内部状態の遷移が入力によって確定していない、非決定性を導入した非決定性チューリング機械を想定することもでき、非決定性チューリング機械の要素技術を開発したという報告[2]もある。ただし、もし非決定性チューリング機械が実現してしまうと、コンピュータ科学の未解決問題で、多くの研究者がＰ≠ＮＰだろうと支持している、Ｐ≠ＮＰ問題が否定的に解決されてしまうので、非決定性チューリング機械は実現できないだろうとも言われている。いずれにしても、非決定性チューリング機械が実現されていない現状では、すべてのＡＩの動作は決定的ということになる。

自由意志とは、哲学者の宇都宮芳明名誉教授によれば、「人間のうちにあって、自然の因果の必然によって規定さ

れずに自発的に発動する能力が自由意志である。」[3]とある。この定義は人間の能力に限定されているので、ＡＩの自由意志を検討するにあたり、本論文では、自然の因果の必然によって規定されずに自発的に発動する能力が自由意志であると定義する。この自由意志の定義を採用しても、ＡＩがチューリング計算可能関数であるとすれば、ＡＩによるすべての決定は初期状態（入力）によって確定し、アルゴリズムを通じた因果の必然によって出力が確定するので、ＡＩには自由意志は存在しないことになってしまう。基礎情報学の立場から西垣通名誉教授も「ある存在がこれから実行する行為を、（推定できても）完全に予測できないことは、その存在が自由意志を持つことの必要条件である。」[4]とし、ＡＩの出力が予測可能であることからＡＩの自由意志の不存在を主張されている。

人間の自由意志の存否を問えないという立場からは、ウェンデル・ウォラック博士とコリン・アレン博士は「決定論的システムを本物の道徳的行為者と考えられるのかどうかという問題は、人間が本当に自由意志をもつのかどうかという問題と同様に答えが出ない。もし本物の道徳的行為者に関するあなたの考えが、自由意志という「魔術的」観念を含むものなら、人間がそれをもっているとも言い切れない」[5]と主張される。両博士によれば、決定論的システムであるＡＩが自由意志をもつのかどうかという問題には答えが出ないことになる。

また、人間の自由意志については、大庭健教授は「私たちが考え、迷い、決断するとき、私たちの脳神経系で起こっている個々のプロセスは因果的に決まっている、と。だからといって自由意志は存在しない、という結論がただちに出るわけではない。」[6]と指摘し、ミクロな状態変化が決定的であったとしても、自己組織化のようなプロセスを通してマクロな状態変化として現れるのが自由意志であると主張される。さらに、大庭教授は「マクロな状態の変化は、環境に対応しながら、ある機能をうまく充足する変化にもなっている。」[7]と指摘され、ＡＩにおいても、マクロ

な状態変化として自由意志が存在できる可能性が示唆される。ＡＩによる創発については、一九九〇年代に環境からの報酬によって新しい機能を創発する強化学習が注目されていたが、報酬や状態空間の設計の難しさから最近では強化学習に深層学習を採り入れた深層強化学習が注目されている。[8] 但し、仮に疑似的な自由意志が創発されるようになったとしても、疑似的な自由意志による判断が決定的であれば、それを果たして自由と呼べるかどうかという問題が残されることになる。

一方、物理学においては素粒子の自由意志を許容する自由意志定理がある。物理学者の筒井泉博士は素粒子の自由意志について、「二〇〇六年にコンウェイとコッヘンによって示された（さらに二〇〇九年に改定版が出された）もので彼ら自身によって「自由意志定理」と名付けられている。ここで自由意志は非決定性の意味で使われており、したがって、この定理は実質的に、世界は非決定論的であるということを主張するものになっている。[9]」と紹介されている。素粒子の世界は古典力学ではなく量子力学に従うが、量子力学には非決定性が現れることから、素粒子が自由意志をもつことが論理的に示せるというのである。しかしながら、多くの人間が素粒子のような無生物に対しても自由意志の存在を許容することは難しいであろう。

ＡＩが自由意志をもたないとしても、ＡＩスピーカーやＡＩによる自動翻訳機を通して、日常的にＡＩと会話をするようになると、人間のＡＩに対する見方が変化し、ＡＩの自由意志の存在を信じるようになるかもしれない。次節では、ＡＩによる自然言語処理について紹介する。

3　AIと自然言語処理

二〇世紀のAIは、実用的な問題を扱えないことから、第二次ブームの終了後長い冬の時代を余儀なくされていた。AIと同じく一九五〇年代に登場した人工神経回路網は、入力と出力の関係を学習するために、入力から出力が算出できる人工神経回路網の重みを学習する方法であり、学習時間は必要であるが、一旦学習してしまえば、入力から瞬時に出力を算出できる。一九八〇年代には、入力層と出力層の他に隠れ層を導入することで認識能力が向上し、AI同様にブームを迎えたが、性能向上が頭打ちになってしまった。それからさらに二〇年以上経過して、層の数を四層以上に増やし、アルゴリズムを工夫したところ、人間の能力を超える認識能力を発揮するようになり、人工神経回路網を用いた機械学習は深層学習（ディープラーニング）として着目されるようになった。現在の第三次AIブームを牽引しているのが深層学習であることは間違いない。深層学習は、その高性能さゆえに、画像認識や音声認識等のパターン認識においては標準的ツールとして定着した。

深層学習は主にパターン認識に応用されていたが、音声認識以外の自然言語処理にも応用されるようになると、自然言語処理は飛躍的に性能が向上した。深層学習は入力と出力の関連付けを学習するので、入力を符号ベクトルに変換する深層学習器（エンコーダー）と符号ベクトルから出力に変換する深層学習器（デコーダー）の二つの深層学習器を組み合わせること（トランスフォーマーと呼ばれる）で自然言語処理に応用されている。従来は、形態素解析・構文解析を経て、意味処理の後、出力文を生成するのに、ルールベースが用いられていたが、現在の自然言語処理では、従来の記号処理を省き、エンコーダー・デコーダーを用いて、符号化ベクトルを介し、入力から出力が計算される。

入力を日本語の文章とし、出力を英語の文章とすれば、機械翻訳システムを実現でき、入力を質問文とし、出力を応答文とすれば、質問応答システムを実現できる。注意しなければならないのは、現在の自然言語処理システムは記号列変換システムに過ぎないことである。もし符号ベクトルが言葉の意味を正確に捉えているならば、ＡＩが意味を理解していると見ることもできるであろうが、符号ベクトルは大量の文書データを情報圧縮したデータに過ぎないので、符号ベクトルが意味を正確に捉えているとみなすことは難しく、ＡＩが言葉を理解しているとは言えない。言葉を理解しないＡＩであっても、日常会話程度であれば、深層学習に基づく市販の携帯型機械翻訳端末でも十分に有用であるし、動画共有サイトにＡＩスピーカーの動画を掲載している人が少なくないことから、ＡＩスピーカーとの会話を楽しんでいる人も少なくないであろう。音声を使ってコンピュータと会話できるようになるのは、二〇世紀にはＳＦの中でしか登場しなかったが、今や誰もが身に着けているスマートフォンを通してコンピュータと会話できるのである。

従来のルールベースの自然言語処理や確率・統計に基づく自然言語処理と比較すれば、深層学習による自然言語処理能力の進歩の速度は著しい。自然言語理解のベンチマーク[10]によれば、人間の標準的成績を超える自然言語処理システムが既に一三件登場している。より難しいベンチマーク[11]では、人間を超えてはいないものの、人間の点数に迫るGoogle の自然言語処理システム T5[12]がある。T5もトランスフォーマーに基づく自然言語処理システムである。また、ＡＩを開発するOpenAI が開発したGPT-3[13]は人間の作成した文章であるかのような文章を生成することができる。GPT-3は、公開が躊躇われるほどの文章を出力したGPT-2の後継であり、ＡＰＩが公開され、クラウドサービスとして利用できる。ソフトウェア的には、GPT-2と基本的に変わらないが、深層学習の層の数を九七層に増やし、パ

ラメータの数を一七五〇億パラメータに拡大することで、性能は大きく向上している。GPT-3に対してチューリングテストを実施した報告[14]もあるが、チューリングテスト合格に近づいたと見ることはできても、まだチューリングテストに合格したとは言えない。チューリングテストは、チューリング機械のチューリング博士が提案した知性を測定するテストで、チューリングテストに合格するということは、人間との区別がつかないということを意味する。GPT-3の限界[15]も指摘されており、現状では、まだまだ人間には及ばないとしても、GPT-3がＡＩであることを知られずに人間と議論するなど[16]、ＡＧＩに接近したと考える意見[17]もある。

４　ＡＩに実現できていないこと

二〇二〇年ノーベル物理学賞を受賞したロジャー・ペンローズ教授は、著書「皇帝の新しい心」[18]の中で知性には意識が不可欠であるから、知性は実現できないという強いＡＩ批判を行った。脳科学者の茂木健一郎博士は、この批判を「ペンローズは、さらに、人間の知性には、アルゴリズム以上の要素、すなわち非計算的要素があるとする。そして、このような非計算的要素は、意識によって支えられるとする。」[19]と要約された。ペンローズ教授の批判の中心は、知性には「意識」が不可欠であり、意識が計算可能ではないと仮定すると、コンピュータには知性は実現できないことになる。ＡＩ研究の先駆者の一人だったジョン・マッカーシー教授は、意識の問題には触れずに、知性の実現についてご自身の常識推論研究を例に挙げてペンローズ教授の批判に反論した[20]。すなわち、推論技術の高度化によって知性が実現可能だと主張されるのである。二〇世紀のＡＩは、ＡＩによる法的推論にも関係する、計算可能な推論、と

りわけ論理プログラミングに基づく研究は発展したが、人工意識についてはほとんど研究が進んでいなかった。実際、ペンローズ教授の批判から三〇年経過した現在でも、ＡＩにおける意識の実現には至っていない。

ペンローズ教授がＡＩにないものとして指摘するのは、意識だけではない。茂木博士の言葉を借りれば、「ロジャー・ペンローズ（Roger Penrose）は、知性、とりわけ「理解」（understanding）の本質は意識にこそあると主張する[21]」。すなわち、ペンローズ教授の主張によれば、ＡＩは意識をもたないのであるから、意識をもたないＡＩは理解などするはずがないということである。ＡＩが理解しないという主張はペンローズ教授だけではない。たとえば、人工知能プロジェクト「ロボットは東大に入れるか」プロジェクトディレクタを務めた新井紀子教授は、「「太郎は花子が好きだ」という文は、まさにそのとおりの意味で、何か他のものに還元することはできません。「花子は太郎に好かれている」と受け身に変換したり、「Taro loves Hanako」と英語に翻訳できたりしたからと言って、意味を理解していることにはなりません。人間ならば誰もがわかる「その通りの意味」をＡＩに教える道具は、少なくとも数学にはありません。そして、繰り返し申し上げているように、コンピューター上で動くソフトウェアにすぎないＡＩは徹頭徹尾数学だけでできているのです。[22]」」新井教授が指摘するように、意味を解明する数学が開発されない限り、コンピュータが意味を理解することはないということである。機械翻訳の品質が向上しようとも、自然言語の応答の品質が向上しようとも、確率・統計的に尤もらしい記号列を出力しているだけなのである。

次節では、言葉を巧みに使いこなしても、理解はしないＡＩが責任主体となり得るのかについて議論する。

5　自由意志と責任

刑法の責任論の基礎としての意思自由論には、山中啓一教授[23]によれば、①非決定論、②決定論、③相対的非決定論およびやわらかな決定論、④不可知論と擬制、⑤規範的要請説がある。①非決定論に立てば、自由意志の存在と両立できるが、②の決定論に立てば、自由意志の存在と矛盾することになる。このため、これが哲学の問題として大きな問題とされてきた。法哲学者の碧海純一名誉教授は、「この、哲学の極めて古くそして最も根本的な論議された問題に深入りすることは困難なので、ここでは、一つの点にだけふれておこう。それは、いわゆる「決定論」と「意志自由論」とは必ずしも矛盾しないという点である。[24]」と両立説に立たられたことがわかる。

瀧川裕英教授は、決定論と自由意志の問題を責任実践の立場から論じ、「決定論問題は、社会実践と切り離された単なる形而上学的問題ではないということである。ここで、今までの論程を簡単に振り返っておこう。まず、決定論問題の二つの問題の内、機械論問題の焦点問題は、「行為の究極的な原因が行為者の内にあるという態度をとるのか、行為者の外にあるという態度をとるのか」という問題であった。そして、行為の究極的な原因が行為者の内にあるという態度ができるためには、行為者に理由能力があること、換言すれば穏当な理由反応性が成立していることが必要であるとされた。その穏当な理由反応性が成立するか否かは、可能的理由について第三者の了解可能性が存在するか否かにかかっており、要するに社会的に決定されることが論証された[25]」と結論づけている。瀧川教授によれば、自由意志の存否とは独立に、行為者であるＡＩに理由能力があれば、ＡＩの行為の原因がＡＩの内にあるという態度を取れることになる。そして、瀧川教授は「責任実践は、問責者と答責者の間のコミュニケーションであり、そのために

行為の客観的帰結でも行為者の主観的意図でもなく、行為者の客観的意図が問題となるのである」[26]とコミュニケーションの重要性を指摘されており、行為の原因がＡＩにあるといえるためには、ＡＩが問責者とのコミュニケーションを通してＡＩの意図を示せるかどうかに帰着されることになる。残念ながら特化型ＡＩには意図をもたせることはできないが、ＡＧＩは自らのゴールを設定できるであろうから、ＡＧＩが実現されれば意図のようなものをもたせることができるようになるかもしれない。

河島茂生准教授は「理性に基づき「自分で自分のことを決定する」という思考の一種が個人の自律性と一般に呼ばれているものに等しい。ネオ・サイバネティクスの見地に立てば、個人の尊厳を基礎づけてきた精神の自律性は、ラディカル・オートノミーという枠組み全体のなかでのきわめて特殊な一形態である。前記したように、この心理レベルの自己決定は近代社会における個人の選択の自由をもたらす。しかし、自己決定の存在基盤はゆらぎはじめている。したがって、それよりも広い意味でのネオ・サイバネティクスにおける自律性概念に基づいて根拠づけなければ、今後のＡＩ社会のなかで個人の自律性は見えなくなってしまうだろう。（中略）人間がオートポイエティック・システムの集合体であり自分で意味を作り出している存在であるがゆえに、そこに存在意義があると見なすのである」[27]と、意味を作り出せるのは人間だけであるという、人間の存在意義を主張される。これに対して、西垣名誉教授は、ＡＩの責任についての別の見方として、「行為者の内面的な自由意思という難問はいったん括弧に入れ、外部から見たときの行為者の選択の自由の有無を責任概念と結びつけようというものだ。端的にいうと、行為者が他の選択をおこなうことが可能なのに、ある選択を実行して誤判断をしたとき、その責任を問う、というものである。もしその選択肢しかなければ責任を逃れることができる。」[28]その上で、両立説を採用した場合には、「ＡＩロボットが選択肢

をもつか否かという問いに関わってくるのだ」[29]と指摘され、「両立論に基づくなら、ＡＩの普及とともにその統一的な責任概念が不明確になり、種々の矛盾が出現してくる。したがって、仮に両立論を採用するにしても、自律系と他律系の相違をふまえて問題の構造を明確に把握し、きちんと整理しなくてはならない」[30]と問題を提示される。この問題を解決するために、「広義の自律性とは、のちに詳しくのべるが、他者の指令をまったく受けずに行動することを可能とする、下等な生物でももっているような原理的な特性である。また、狭義の自律性とは社会的な自律性であり、人間という道徳的主体と同等、あるいはそれ以上に的確な判断をくだし、社会的な責任をとれる主体のもつ特性とする。前者を「理論的自立性（theoretical autonomy）」、後者を「実践的自律性（practical autonomy）」と呼ぶ。[31]」と自律性を定義される。すなわち、もしＡＩが実践的自律性を有する存在となれれば、ＡＩに社会的責任をとらせることができるのである。西垣名誉教授は、「倫理は「社会規範／行動／道徳観」からとらえられることになる」[32]と主張されるので、瀧川教授のいう理由能力として、選択肢に対するＡＩの判断が社会規範を遵守していることを説明できれば良いことになる。

６　おわりに

深層学習はＡＩの新しい応用を切り開き、ＡＩ社会を推し進める原動力となっている。深層学習を自然言語処理に応用した結果、自然言語処理の性能は飛躍的に向上し、性能は人間に匹敵するようになった。従来とは桁違いの大容量の言語データを用いて学習した結果、性能は向上したものの、所詮は記号列の統計的に出現させているだけで意味

を理解しているわけではない。しかしながら、現在、自然言語処理技術は全世界の技術者・科学者が競っており、その進歩の速さは正に秒進分歩である。その結果、自由意志のないＡＩが言葉だけはあたかも人間のように駆使してしまうのである。ＡＩを道徳的主体とみなせるようにするためには、選択肢におけるＡＩの判断が社会規範を遵守していると人間に説明できるようになることが必要となるのである。

（1）ＡＧＩについては、例えば以下を参照のこと。
山川宏、市瀬龍太郎、井上智洋「汎用人工知能が技術的特異点を巻き起こす」電子情報通信学会誌、Vol. 98, No.3（2015）.

（2）Currin A, Korovin K, Ababi M, Roper K, Kell DB, Day PJ, King RD. 2017.
Computing exponentially faster: implementing a non-deterministic universal Turing machine using DNA. J. R. Soc. Interface 14: 20160990. http://dx.doi.org/10.1098/rsif.2016.0990.

（3）宇都宮芳明「自由意志」Japan Knowledge 版日本大百科全書。

（4）西垣通・河島茂生『ＡＩ倫理―人工知能は「責任」をとれるのか』五七頁（中央公論新社、電子書籍版、二〇一九）。

（5）ウェンデル・ウォラック、クリン・アレン：『ロボットに倫理を教える』（岡本慎平・久木田水生訳、名古屋大学出版会、初版、二〇一九）八四頁。

（6）大庭健『「責任」ってなに?』（講談社現代新書、二〇〇五）六八頁。

（7）大庭・前掲注（6）七〇頁。

（8）新井幸代、石川翔太、中田勇介、北里勇樹「強化学習における脱創発志向の潮流 試行錯誤～見まね～目的理解へ」人工知能、三三巻二号（二〇一八）一七〇頁以下。

（9）筒井泉『量子力学の反常識と素粒子の自由意志』、（岩波書店、電子書籍版、二〇一五）一三〇頁。

（10）https://gluebenchmark.com/leaderboard (2020. 11. 15)

（11）https://super.gluebenchmark.com/leaderboard (2020. 11. 15)

（12） Collin Raffel, et.al., Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, arXiv: 1910. 10683, (2020. 11. 15).

（13） Khari Johnson, OpenAI debuts gigantic GPT-3 language model with 175 billion parameters, https://venturebeat.com/2020/05/29/openai-debuts-gigantic-gpt-3-language-model-with-175-billion-parameters/, (2020. 11. 15).

（14） Kevin Lacker, Kevin Lacker's blog, https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html (2020. 11. 15).

（15） Theodore F. Claypool, New AI Tool GPT-3 Ascends to New Peaks, But Proves How Far We Still Need to Travel, The National Law Review, https://www.natlawreview.com/article/new-ai-tool-gpt-3-ascends-to-new-peaks-proves-how-far-we-still-need-to-travel, (2020. 11. 15).

（16） Kmem, GPT-3 Bot Posed as a Human on AskReddit for a week. https://www.kmeme.com/2020/10/gpt-3-bot-went-undetected-askreddit-for.html (2020. 11. 15).

（17） GIGAZINE〝「人間と見分けが付かないほど高精度な文章を生成するＡＩ「GPT-3」について哲学者らはどう考えているのか？…」〟 https://gigazine.net/news/20200803-philosopher-gpt-3/ (2020. 11. 15).

（18） ロジャー・ペンローズ『皇帝の新しい心　コンピュータ・心・物理法則』（林一訳、みすず書房、第四刷、一九九五）

（19） 一二四三頁。

（20） John McCarthy『Review: Roger Penrose, The emperor's new mind』, Bull. Amer. Math. Soc. (N.S.), Volume 23, Number 2 (1990), 606−616.

（21） 茂木健一郎『クオリアと人工意識』（講談社現代新書、電子書籍版、二〇二〇）四二頁。

（22） 新井紀子『ＡＩ vs. 教科書が読めない子どもたち』東洋経済新報社、電子書籍版 Ver.10〟（二〇一八）〟一四四頁。

（23） 山中敬一『刑法総論』成文堂、第三版（二〇一五）六二四頁以下。

（24） 碧海純一『新版法哲学概論』弘文堂、全訂第二版（一九九一）。

（25） 瀧川裕英『責任の意味と制度　負担から応答へ』（勁草書房、第一版第一刷、二〇〇三）一一三頁。

（26） 滝川・前掲注(25) 一五九頁。

（27） 河島茂生編著『ＡＩ時代の「自律性」』（勁草書房、第一版第一刷、二〇一九）二七頁。

(28) 西垣通『AI言論　神の支配と人間の自由』（講談社、電子版、二〇一八）一六八頁。
(29) 西垣・前掲注(28)一七一頁。
(30) 西垣・前掲注(28)一七三頁。
(31) 西垣・河島・前掲注(4)五二頁。
(32) 西垣・河島・前掲注(4)一一八頁。

（明治学院大学法学部教授）