人工的道徳的行為者への近年のアプローチの検討

岡　本　慎　平

（広島大学大学院）

はじめに

カレル・チャペックが戯曲『R.U.R』の中で「ロボット」という言葉を世に送り出してから約１世紀。アイザック・アシモフが『われはロボット』で自律機械が従うべき規範を、いわゆるロボット三原則を提唱してから約70年。現在のロボット技術は、彼らの時代では考えられなかったほど進歩した。人間とほとんど見分けのつかない外見のロボットの開発が報道されたり、自動的に床の掃除を行うロボットが市販されるなど、ロボットは様々な局面で我々の生活に身近な存在となり始めている。アメリカの政治学者ピーター・W・シンガーの著書『ロボット兵士の戦争』は、このような自律機械が引き起こしている問題、とりわけ戦場における自律兵器の現在の活躍について興味深い事例を多く示している。例えば、アイロボット社の開発したパックボットというロボットは円盤状の基本素体に様々なオプションを組み合わせることで、小銃を備えた兵士にも、応急救助を行う衛生兵にも早変わりするという。シンガーはこの著書の終章で、ロボット工学の進歩を次のようにまとめている。

> 瞬く間に、SFの素材にすぎなかった機械が、現在の戦場で這い、飛び、泳ぎ、発砲するようになっている。しかも、こうした機械はこれらの新技術の第一世代にすぎず、あなたがこのくだりを読んでいるころには、既に時代遅れになっているものもあるかもしれない。[1)]

このような、現実が空想を追い越すような進歩の中でロボットに実装されるべき規範としては、70年前の小説で提唱されただけの、しかも実装するためには多くの困難が付きまとう、あのロボット三原則だけでは心許ない。我々は、小説の素材としてではなく、現実の問題としてロボットの倫理を考えなければならない。そこには、もちろんこのような自律兵器の開発が倫理的に許容できるのか否かという工学倫理上の問題も、戦場で実際に動作する自律兵器のための規範、言わばロボットの正戦論とでも言うべきルールも必要となってくるだろう。

まずは、こういった危急の問題として要請されているロボット倫理学の展開を歴史的に追い、どのような観点からこの問題が提起されたのかを見ていこう。

1．ロボット倫理学の誕生と発展

20世紀末から既に、自律的に作動するロボットについて何らかの倫理規定を備えるべきではないのか、またその開発にあたって、何らかの倫理規定を備えるべきではないのかという議論自体は行われていた。しかし、明確にロボット倫理学、ないしはロボエシックス(Roboethics)が誕生するのは、2004年にイタリアのサレルモで開催された「ロボエシックスについての第一回国際シンポジウム(the First International Symposium of Roboethics)」以降である。このロボエシックスの目的は「ロボット工学に付随する倫理的問題について体系的な評定を提供する[2)]」ことである。2005年には、イタリアのロボット工学者らが中心となり、ヨーロッパロボット工学研究ネットワーク(EURON)内において、ロボット工学の規範について研究をすすめるプロジェクト、「EURONロボエシックス・アトリエ(EURON Roboethics Atelier)」が立ち上げられた。

このロボエシックス・アトリエとは、「ロボット工学者とともに、哲学者、

法学者、社会学者、人類学者、倫理学者など[3]」が世界中から参加した大規模なプロジェクトだった。計画の立役者はジャンマルコ・ベルージオ[4]というロボット工学者であり、数年に渡って様々な試みが練られてきた。日本でも早稲田大学が中心となり、ロボエシックス・アトリエとの連携による研究や、日本での研究大会の開催も行われてきた。その成果の一つが、ロボット工学についての一般的な研究方針、倫理規定を定めた「EURONロボエシックス・ロードマップ」であり[5]、それをさらに詳細にした『シュプリンガー・ロボット工学ハンドブック』のロボエシックスのパートである。日本や韓国を中心に概観しても2007年に草案が出された韓国産業資源部による「ロボット倫理憲章」や、同じく2007年の千葉大学ロボット憲章など、ロボット工学上の倫理規定が続けざまに行われるようになった。またそれと平行して哲学者の間でも、より概念的なレベルでロボットや機械の倫理についての議論が活発になりはじめてきた。この流れを受けて、英語圏では専門哲学者向けのロボット倫理学のリーディング論文集も既にいくつか編纂されている。

このようにロボットにまつわる倫理問題は、既に単なるSF的な絵空事ではない。特に現在最も緊急のものは、先に挙げたような自律兵器の問題である。例えばピーター・W・シンガーの『ロボット兵士の戦争』やアーミン・クリシュナンの『殺人ロボット：自律兵器の適法性と倫理性』などで詳述されているように、既にアメリカ軍は「自律的に」動作する兵器を投入することによって、自軍兵士の損害を防ぐ試みを実行している。また、実際に医師がその場に急行するのが困難な状況で、自動的に要救助者を発見し、応急処置を施すという「災害救助ロボット」というものも構想されている。こうした自律機械の研究において、これまでの技術者を対象とした倫理規定だけでなく、技術者が自律機械に実装する命令や制約自体もまた、重要な争点となっていった。

また、EURONロボエシックス・ロードマップにおいて提唱されていたロボッ

ト倫理の三つの形について簡単に述べておきたい。ベルージオらによって方針が定められたロボエシックスは、大きく分けて三つのカテゴリーから成っている。一つは現在の「技術者倫理」と呼ばれているものの延長線上である、人間がロボットを製造する際の倫理規定である。第二に、ロボットが人間に対して何らかの影響を与える行為を行う際の倫理である。これは自律的に動作するロボットに、どのような規範を導入すべきなのかという問題である。最後に、人間とロボットの関係の倫理である。言い換えれば、人間が「ロボットに対して」どのような態度を取れば良いのかという、法的規制・社会構想のレベルでの問題である。今後我々の社会に身近になっていくであろうロボットを、我々はどのように扱うべきなのか。EURONではこのように問題点が三種類に区分されていた。

本稿では、こういった実践のための理論的枠組として、哲学者たちが関心を持っている「人工的道徳的行為者」の問題について近年活発に議論されているいくつかの主張を検討する。先述のロボエシックスの区分から言えば第二の問題に深く関わるものであるが、第一の問題とも密接に関わっている。

そこで、以下のように検討を行いたい。まず第一に、この問題を定式化したコリン・アレンらの議論を追うことによって問題の位置を把握する。第二に、独自の情報倫理思想からこの問題をより概念的・存在論的なレベルで考察し、援護を試みているルチアーノ・フロリディらのアプローチを検討する。最後に、アレンやフロリディらが論じるような、機械を「行為者」として扱うこと自体の問題性を指摘し、「人工的道徳的行為者」という概念を構築せずとも、ロボットの道徳性を論じることが可能であると主張するデボラ・ジョンソンの立場を検討する。

2．なぜ機械の道徳なのか

前節で見たように、ロボット工学者主導による「ロボエシックス」の構築と並行して、哲学者たちは自律機械の道徳性について様々な考察を行なっていた。その一つが、「人工的道徳的行為者(Artificial Moral Agent, 以降AMA)」の概念的構築という試みである。ベルージオらの編纂したハンドブックから、この概念の説明を見てみると、次のように記されている。

> この概念〔=AMA〕に従えば、ロボットや人工的行為者は、道徳的状況に巻き込まれうる存在者のクラスへと拡張される。というのも、それらは道徳的受益者として(善悪のために行為されうる存在者として)、また道徳的行為者として(自由意志や精神状態や責任を示す必要はないが、善悪のために行動することが出来る存在者として)、把握されることが出来るからである。[6)]

人の手によって造られたロボットは自らの「意志」を持ち得るのか、言い換えれば極めて人間的である「道徳的行為」を行うことが出来るのかという問題、「強いAI」は実現可能なのかという問題は古くからSF作品で多く扱われ、現代でも人工知能や心の哲学の主要トピックとなっている。しかし、問題はロボットが完全に人間と同じような意味で「道徳的行為者」になり得るかという点にはない。

コンピュータが現状のままで、本当に「人間と同じような」道徳的行為者／受益者になりえると考えている論者はロボット倫理学の研究を行う人々の中にもほとんどいないだろう。また「強いAI」の開発が射程内に入っていると考えているロボット工学者がいたとしても、また仮に「強いAI」が構築出来たとしても、ヒューバード・ドレイファスの強い懸念を見れば分かるように、そ

れがはたして人と同じような心を持ち得るのかという問題が残り続ける。

しかし、ロボットの倫理学を考える上でこのような心の哲学の難問へと向かう必要はないだろう。というのも、ロボットが人と同じような身分を法的・社会的・あるいは振る舞いのレベルで得られるか否かに関わらず、自律的に動作するロボットが人に与える影響は、今日ますます重大になっているからである。言わば、災害救助犬が「困っている人を、我が身を顧みず助ける」という、ある意味で道徳的行為とみなしうるような振る舞いを行うのと同じレベルで、ロボットの「道徳的行為」を論じようというのが出発点である。

3．人工的道徳的行為者

AMAの問題は、人工知能はいかにして道徳的行為者となりえるのかという問題として90年代からいくつかの先行的な議論が展開されていた。しかし「AMA問題」としてこれを定式化し、その後の方向を定めたのは2000年のコリン・アレンらの論文「将来の人工的道徳的行為者のためのプロレゴメナ(以下プロレゴメナ)」である。この論文では、AMA研究の方向性が、二つの方向で示唆されることとなった。一つは自律機械にどのような規範を実装すべきかという「理論的アプローチ」であり、もう一つは自律機械が実際の行動の中で、どのように規範を学習していくのかという枠組みを、我々の道徳的行為の習得をモデル化することで設計しようという「モデル化アプローチ」である。

「プロレゴメナ」に即して話を始めていこう。コリン・アレンらは、「良い」機械とはどのよう機械なのかという問いから始める。例えば、「良いサーバーはHTMLコードを上手くアップロード出来るコンピュータであり、良いチェス・プログラムはチェスの勝負に勝つプログラム[7)]」である。このように、通常我々が「良い機械だ」と言う際、我々の目的をうまく達成できる機械を良い機械と呼んでいる。しかし、我々は物や道具を「道徳的に」良い、悪いと判断するこ

とはない。というのも、普通の道具は、我々から独立して、自らの意思を以て動作することはないからである。しかしここで問題とするのは、このような我々から独立して、自ら動作するような機械の問題である。「操作者の手を離れて自律的に動作した結果、人間に有用な行為を行う能力を持つ自律機械は、また人間に害をなす能力も持っている[8)]」。この「自律機械がなしうる危害」を防ぐための手段を講じるために、自律機械にとって、どのような行為が許され、どのような行為が許されないのかを判断するための一種の規範が必要となる。この問いが、従来の技術者倫理－制作する際の倫理－を超えて、「ロボットの」倫理学を構築することが必要である所以である。このような外部からの操作を加えず、内在的な規範と認知機能のみによって判断を下す自律機械を、暫定的に「人工的行為者(artificial agent)」と呼んでおこう。

AMAを構築する試みには、まず二つの大きな問題がある。一つは「道徳的行為者はどのような基準に従うべきか」という自律機械に実装すべき規範のレベルでの問題であり、もう一つは「そもそも道徳的行為者とは何か」というより概念的・存在論的な問題である。この二つの問題のうち、特に重要かつ難問となるのが「道徳的行為者とは何か」という問題である。例えば彼らはカントを持ち出し、「定言命法」は人間という道徳的行為者のみに対する命令であり、ロボットにそのまま適用することは出来ないと主張する。なぜなら、人間以外の理性的存在者－例えば神－の場合は、定言命法などを持ち出すまでもなく、当為がそのまま意志となるからである。自律機械も、人間か神かという分け方をした場合は神に近いだろう。というのも、人間につきものの理性と本能の葛藤というものが、機械には生じ得ないからである。だが、たとえもしそのような葛藤が生じ得る機械を作ることが可能であったとしても、そのような機械を設計し、実装する理由はない。またそれが実装されたとしても、それを現実の自律機械の使用場所(例えば地雷を撤去する戦場や、緊急災害時の応急処置)で

使用することはない。なぜなら、機械はあくまで人間の利益・目的のために設計され、人間に強いることが難しいような「面倒くさく、汚く、危険な(dull, dangerous, and dirty)[9]」仕事を任されるために使用されるのであり、そういった場面では、人間のような(human-like)道徳的存在者としての能力はむしろ不必要だからである。

機械に実装する規範のレベルの問題だとどうだろうか。規範倫理学は義務論や功利主義などに分断されているので、その原理に関して統一的見解と呼べるものがない。これらのうちのどれを自律機械の行動原理として採用するかを問題とすると、議論が止まってしまう。例えばカント的な枠組みを用いれば、行動に至る推論を重視する一方で、その行動の帰結を度外視せざるを得ず、功利主義的な枠組みを用いれば、その行動へと至る動機や目的が無視されることになる。こういった原理レベルの問題に加えて、日々の生活でも人々の倫理的問題についての意見は一致しない。このような統一見解のとれない問題について、一定の原理からの演繹的なアプローチを試みることは困難である。

そこでアレンらが提唱するのは、AMAの構築にあたって、チューリングテストの改変を用いることである。チューリングテストは、質問者が機械と人間を区別できなければ、その機械は自然言語を使用出来ている、というアラン・チューリングによって提唱された、代表的な人工知能のテストである。道徳的チューリングテストは、自律機械に対して倫理的な判断が出来ているかどうかを問う。より正確に言えば、ある状況下において一定の行為を選択した際に、その選択の理由を「道徳的な観点から」説明可能か否かを問題とする。そして一定回数の質問の後、質問者がその対話相手が機械なのか人間なのかを正しく特定できれば、その機械はテストに落第となる。機械が他の人間から特定されない場合、この機械はテストを突破したということになる。このテストを通過すれば、その機械は道徳的行為者である。これを彼らは比較道徳的チューリン

グテスト(cMTT)と呼ぶ[10]。このテストを突破することが出来れば、自律機械は我々人間と同等の「道徳的推論」を行うことが出来るので、AMAとしての資格を満たすことになり得るのではないかというのが、彼らの主張である。

しかしながら、cMTTをAMA一般の基準とするにはいくつか問題がある。第一に、我々は機械に対しては非常に高い基準を課すのだが、このテストでは基準が低すぎるという問題である。通常自律機械が投入される現場は「面倒くさく、汚く、危険な」仕事であり、その役割は人間のそれとは要求される程度が異なる。第二の問題は、人間自身の行動自体が、機械に要求される道徳性から程遠いというものである。我々人間は、日々多くの道徳的な過ちを犯す。しかし、ある程度は人間の過ちは寛容される。しかし機械の場合には、人間より多くのことが期待される。つまり、機械には失敗が許されないのである。このように実践段階で考察した場合、チューリング・テストによってロボットを道徳的行為者とみなすという方針は不適切なものである。したがって、自律機械の道徳的能力を試すために、チューリングテストを用いるべきではないだろう[11]。

4. トップ・ダウンかボトム・アップか

このように、機械に要求される道徳的基準が人間に要求されるものより根本から異なっているのであれば、チューリング・テストは有力な手立てとはならない。それでは、コンピューター科学者は自律機械に実装するべき規範をどこに根拠を見出せばいいのか。アレンらは、二つのアプローチを提唱している。まず、チューリング・テストが有効か否かはさておき、AMAが従うべき行為の評価基準を与えるという「理論的アプローチ(theoretic approach)」である。これは、トップ・ダウン型のアプローチであると言えるだろう。第二に、道徳的性格についての理論や、道徳的行為の学習や進化を用いて、基礎的な原理を前提とせずに我々道徳的行為者の「仮想的モデル」を構築しようと試みる「モ

デル化アプローチ(modelling approach)」である。順に検討していこう。

理論的アプローチは、自律機械の行動に際して基本的なフレームワークを与えるというアプローチである。例えば、アシモフのロボット三原則は明白にこの立場を取る。まずは、歴史的背景を探る意味でもアシモフの三原則を参照してみたい。アシモフが『われはロボット』で提唱した、ロボット工学の設計の際の基準は、この三つの基本的法則である。

第一条　ロボットは人間に危害を加えてはならない。また、その危険を看過することによって、人間に危害を及ぼしてはならない。

第二条　ロボットは人間にあたえられた命令に服従しなければならない。ただし、あたえられた命令が、第一条に反する場合は、この限りでない。

第三条　ロボットは、前掲第一条および第二条に反するおそれのないかぎり、自己をまもらなければならない。

このように、いくつかの原則によって自律機械の行動を規制しようというのが、このアプローチの特色である。しかしながら、アシモフの三原則は『われはロボット』でのロボットたちの反乱をみるまでもなく、不十分なものであると多くの論者が指摘している。これに代わる実装の試みとして、サンダーソン&サンダーソンが「MedEthEx」というロボットによって構築した「一見自明な義務(prima-facie duty)」を挙げることが出来るだろう。彼らは看護師という専門職の倫理的判断をモデルに、人々が実際に行う規範を帰納的に収集することで、医療ロボットの道徳原則として「自律の尊重、善行、非侵害(respect for autonomy, beneficence, and nonmaleficence)」という三つを提唱した。これを暫定的に「サンダーソンの三原則」と呼びたい。しかしこのサンダーソンの三原

則も、アレンらによれば実装するためには問題が多い[12]。例えば、これら個々の原則は、実際に運用した場合衝突する状況も多々ある。トップ・ダウン型の考察を行う場合、こうった規則同士の衝突を回避する手立てが必要となる。

理論的アプローチから自律機械に実装すべき規範を考察する場合、代表的な倫理学理論に訴えることも出来るだろう。一般的な規範倫理学の立場からこの実装モデルを提供しようとすれば、二つの立場が思い描かれる。言うまでもなく、功利主義と義務論である。この二種類の理論を根底に置けば、少なくとも原則の衝突を回避することは可能となる。功利主義に従えば、ロボットには「当該の場合に最大多数の最大幸福をなし得るように行為せよ」という命令を、より正確には規則功利主義の形で「しかじかの行為は最大多数の最大幸福につながるので行うべし」という規則を書き込むことで、道徳の原理に調停の役割を与えることが出来る。義務論的立場を取れば、ロボットには「定言命法」のような一定の推論プロセスに従った行為のみが正当化されうるという道徳原理を与えるだろう。

このようなトップダウン型のアプローチは常に原則の参照を要求し、人間以上に厳密な計算に基づく行動が一見可能であるように見える。しかし自律機械といえども全知全能ではない。規則功利主義も義務論も、例外事情や衝突する事例をどう扱うのかが問題となる。そのため、このアプローチを元にしてAMAに実装するモデルを考案していくことには問題が多い。それに対して、より実践に即したアプローチとしてアレンらが提唱するのが、モデル化アプローチである。

モデル化のアプローチは、人間にとっての理想的道徳行為者をモデル化することによって、AMAが適切な道徳的決定を行えるようにするというものである。言い換えれば、我々が個々の場合にどのように振る舞うのかという事例を積み上げることで、原則から天下り式に規範を組み立てるのではなく、より実

践に即した意思決定モデルを作ろうとするものである。『モラル・マシーン』ではこのアプローチは先述の「トップ・ダウン」に対して「ボトム・アップ型」のアプローチとして展開されている。

さてこのアプローチは、大きく分けて三つの方針が考えられている。まず第一に、「道徳的行為者を、行為者の振る舞いを導く特定の徳を持つもの」として捉えることで、AMAに「原則」ではなく、特定の行為者の傾向性、すなわち「徳」をプログラムするという徳倫理モデルである。第二に、AMAが実践の中で規範を学習し、その実践を繰り返すことで道徳的行為者として発展していくという学習型モデルである。複数の自律機械を一定の状況下に置き、そこでの相互作用を通じて道徳をロボットに「内面化させる」という方針であると言ってもいいだろう。そして第三に、例えば「ライフ・ゲーム」と呼ばれるような生存・繁殖・進化のプロセスを擬似的に再現するプログラムによって、AMAに協力行為やコミュニケーションの発達を促すことによって、社会性を身につけさせるという道徳的行為者の進化型モデルである。

より後年に提出された『モラル・マシーン』では、さらにこの２つのアプローチの限界を克服するために、両者のハイブリッド型のアプローチを行うべきだという方針が示された。トップ・ダウン的に原則を演繹するだけでなく、実際の行為においては道徳的行為者を様々なボトム・アップ型のモデル化によって補強すべきであるというのが、その主張である。これらの実装レベルの問題は、現在既にロボット工学者らによって実験や施工が試みられており、どの程度成功を収めるのかは未だ未知数であると言えるだろう。

5．情報哲学からのアプローチ

以上のようなアレンらによる実装段階までもを視野に入れた議論に対して、ルチアーノ・フロリディらは独自の情報倫理思想を軸にしたアプローチによっ

て、AMAをより概念的・存在論的な段階で正当化しようと試みている。フロリディの主張によれば、確かに自律機械は人間と同じような意味での「道徳的行為者」にはなり難い。しかし、ある一定の「抽象化レベル(Levels of Abstraction)」においては「道徳的行為者」としてみなすことが可能となる。

フロリディによれば、我々はある対象物を観察する時、話題の前提としている概念的システムに沿うようにその対象を見る。例としてフロリディは、コレクターであり購入希望者のアン、副業で修理屋を営むベン、経済学者のキャロルの三人が、ある「何か」を見て話し合っている場面を以下のように示している。

> アン：それは、窃盗防止装置が付いていて、使われない時はガレージに仕舞われていて、過去の所有者は一人だけのように見えます。
>
> ベン：そのエンジンは純正部品ではなくて、ボディの塗装は最近塗り替えられているけど、皮革の部分はとても擦り切れているように見える。
>
> キャロル：その古いエンジンは使い込まれていて、安定した市場価格を持つけれど、交換部品は高価であるように見える。[13]

アンは使用者としての立場から、ベンは修理工としての立場から、そしてキャロルは経済学者としての立場から、それぞれの「抽象化レベル」に沿って「それ」を見ているのである。このように、我々がXをYであると観察する時、その観察には一定の負荷がかかっている[14]。道徳的問題についても同様である。我々が何かを「道徳的行為者」とみなすときも、どのような抽象化レベルでその「何か」を「道徳的行為者」とみなしているのかをまず明らかにしなければならない。

フロリディによれば、確かに通常の倫理学で問題となる「道徳的行為者」に

は帰責の問題が付随する。ある人物が何らかの行為を行った時、その行為の道徳的評価が行為者へと与えられるのは、その行為の責任(responsibility)が行為者へと帰せられるからである。盗みを行った際にその窃盗犯が罰せられるのは、窃盗行為の責任が窃盗犯に帰せられるからであるし、溺れた子供を助けた人が道徳的に賞賛されるのは、救助行為の責任がその人物に帰せられるからである。この抽象化レベルにおいては、AMAは道徳的行為者とはなりえない。というのも何であれ機械が起こした結果に対して責任が生じる際に、帰責の対象となるのはその機械の製造者であり、あるいはその機械の使用者である。この意味で、自律機械を道徳的行為者とみなすのは、「道徳的行為者」という概念の誤用であるように思われる。この他にも様々な反論が考えられ、例えばフロリディ自身は、機械は単にプログラムに沿って動作するだけなので自分で設定した「ゴールを持たない」という「目的論的反論」や、機械は「志向的状態を持たない」という「志向的反論」、機械は道徳的行為者に必然的な「自由」を持たないといった反論が挙げられている。このどれもが、一見AMAにとって致命的なものとなり得るように思われる。

しかしながら、この通常のレベルとは異なったレベルの抽象化では、自律機械も道徳的行為者となり得るのだとフロリディは主張する。彼は、「小さな子供」や「救助犬(search-and-rescue dogs)」を道徳的行為者としてみなす抽象化レベルでは、機械であっても道徳的行為者として見なされるのだと主張している。子供が困っている人を助けたり、犬が人命救助を行う場合も、我々は彼らを道徳的行為者とみなしているのだと。「犬を道徳的善行の源泉として、それゆえ道徳的状況における決定的役割を演じる行為者として、したがって道徳的行為者として同定することに問題はない[15]」と彼は主張する。

同じように、ロボットのような人工的行為者もたとえそのロボットの行為の道徳的責任(moral responsibility)が設計者や使用者に帰属することになろうと

も、道徳的行為を行うことは出来る。というのも、ロボットと人間では道徳的行為者という概念の抽象化レベルが異なっているのである。言わば、「道徳的責任(moral responsibility)」抜きの道徳的行為者を概念的誤謬ではない形で定義することが出来るのである。

このようなフロリディの主張は、情報倫理学全体についての幅広い考察(彼自身の言葉を用いればマクロ倫理学(macroethics))に裏付けされたものである。しかし、彼の「一定の抽象化レベルにおける」道徳的行為者という考えが、少なくとも「そうではない抽象化レベル」で道徳性を論じる通常の倫理学理論と衝突するものであることは変わりないだろう。もしロボットの倫理学を考える際にこのような想定をしなくても済むのなら、それなしで済ますことがオッカムの存在論的倹約という立場からも望ましい。

6.「人工的道徳行為者」不要論

以上のようなあくまでAMAを「行為者」として扱おうというアレンやフロリディのアプローチに対して、そもそもロボットはそれ自体では「行為者」ではない、という立場に立った上で、ロボットに実装すべき規範を考察していこうという議論も展開されている。こういった立場の代表的な人物は、いくつかの論文でフロリディらを批判しているデボラ・ジョンソンである。

彼女によれば、アレンやフロリディらの前提を承認し、自律機械を「人工的道徳的行為者」を構築しようと試みること自体が危険である。というのも、自律機械、とりわけ帰納的に情報を収集し、自らの規範を構築していくようなソフトウェア(bot)は、自分自身の規則を生み出し、改変していく事も出来るからである。このような機械の自己教育に任せることは、その機械の判断は人間には理解出来ないものとなってしまう。そのような人間に扱うことの出来ない機械を、兵器、医療機器、あるいは金融の場に持ち込むことは、非常に危険で

ある。さらに自律機械をAMAをとみなすことは、少なくともそのような自律機械の開発者に対する一種の赦免となってしまう。たとえフロリディが主張するようにAMAには道徳的責任は帰せられず、そのAMAについての責任者である人物が帰責の対象になるのだと主張したとしても、あくまで行為者は「機械自体」となってしまい、実践レベルでの問題としては設計者の道徳的責任を軽微なものにしかねない。彼女は次のようにフロリディらを批判している。

> そのようなシステムを道徳的行為者として概念化することは、それらを設計、展開する人々を、その行いに対する責任から赦免することである。それは、大量の化学物質を海に投げ入れる人々の責任を、その化学物質が海水や藻とどのように反応を起こすのかを彼らが適切に知らなかったという理由で赦免するようなものである。[16)]

それ故に、機械は「道徳的存在者」ではあっても、「道徳的行為者」であるべきではないのだとジョンソンは主張する。彼女の着眼点は、あくまで工学倫理・情報倫理の延長線上としてロボットの倫理を扱うことで、「ロボットを設計する際の倫理」に「ロボットの行うべき倫理」を包含しようという点にある。

それでは、自律機械の道徳的身分をどのように扱えば良いのか。ジョンソンの対案は、自律機械を「設計者－自律機械－使用者」という三要素からなる行為者の一部分として扱うことである。たとえ自律的に動作するロボットであっても、そのロボットの設計には人間である設計者の志向性が関わっており、そのロボットを一定の目的のために使用する場合、その使用者の志向性が関係している。このように、あらゆる自律機械は、それが人工物(artifact)である限り、その機械のみを取り出して道徳性を判断することは適切ではない、というのがジョンソンの主張である。

> 三要素(triad)の三つの部分全て－使用者という人間、人工物、人工物の設計／製作者という人間が、志向性と効力(efficacy)を持つ。使用者は行為を開始するという効力を持ち、人工物は何であれその行為を実行するという効力を持ち、人工物設計者は人工物の効力を創造している。[17)]

たとえば、自律的に敵味方を区別して、敵兵が接近した場合にのみ爆発するという「自律的かつ独立的に振る舞う地雷」が、民間人である子供を誤って殺害してしまったとしよう。もしフロリディの立場ならば、一定の抽象化レベルにおいてはその地雷が道徳的行為者とみなされるだろう。そしてまた、その地雷は当然ながら道徳的な帰責の対象とはならないだろう。というのも、AMAとしての自律機械に要求される抽象化レベルは、道徳的帰責を含まないレベルでの行為者概念だからである。しかしながら、フロリディの立場を徹底すれば、では誰が道徳的責任を負うのかという問題が完全に考慮の外に置かれてしまうことになる。

ジョンソンもまた、自律的な地雷が確かに道徳的に非難の対象となる行為の一端を担ったことは認める。確かにその子供の殺害という行為は、地雷に備え付けられたコンピュータの自発的な判断によって引き起こされた爆発によるものである。「にもかかわらず、地雷は自然的対象物ではない。その独立性と必然性は、人間存在が目論み、発展させたものである[18)]」という点を強調し、ジョンソンは次のように述べている。「兵士も設計者も、その子供を殺すことを志向してはいなかった。しかし彼らの志向性は、地雷の場所と、なぜ、そしてどのようにその地雷が爆発したのかを説明する[19)]」。確かに、自律兵器によって何らかの損害を被ったと考えてみれば、我々の怒りは自律兵器と、その自律兵器を設計した者、そしてその自律兵器を自分に損害を被らせる形で運用した使

用者という三者に向かうということは、我々の直観にも合致するように思われる。

何であれ、人工物に対しては、その人工物の設計を行った人間と、その人工物を使用した人間という二種類の人間が必ず存在する。ある人工物が、人間に何らかの影響を与える行為を行った場合、その人工物は－それが完全に自律的であろうとも、なかろうとも－その行為の実行者である。しかしながら、その行為の実行は、人工物のみによって引き起こされたものではない。ジョンソンはこういった見解を、次のようにまとめている。

> コンピュータ・システム(そして他の人工物)が人間の道徳的行為者性の一部となりえるのは、それらが人間の道徳的行為への効力を提供する限りであり、それらが人間の道徳的行為者性の結果となり得る限りである。この意味で、コンピュータ・システムは道徳的存在者にはなり得るが、それだけでは道徳的行為者ではないのである。[20)]

まとめにかえて

以上、近年提示されてきたいくつかの人工的道徳的行為者についての見解を、アレンらによる定式化と、フロリディとジョンソンの対立という観点から検討した。フロリディとジョンソンの対立においては、私はジョンソンの見解に賛同したい。そのための理由はいくつかある。まず、我々は「人間に対する行為」を行う自律兵器のような危急の問題に対処する必要があるからである。そのためには、フロリディらのアプローチは包括的ではあるかもしれないが、迂遠に過ぎるように思われる。そして、彼らが前提する形而上学を受け入れない者に対してはあまり効力を持たないように思われる。またアレンらの試みは、人工知能のさらなる発展を前提としており、現状の問題に用いるにはあまりに距離

が隔たっている。それに対して、ジョンソンの見解は、これら抽象的レベルの問題から一定の距離を置き、現実問題としての自律機械の問題に対応することが可能であり、また自律機械の行為を、「設計者－人工物－使用者」という三者の協同的な行為とみなすことは、我々の直観にも合致するように思われる。

本稿では、AMA問題に限定して、近年のいくつかの研究を調査することを目的とした。むろん、ロボット倫理の構築のためには、このAMA問題だけでなく、法制的、社会的なさらなる問題に対しての倫理的判断を下す必要がある。たとえば自律兵器の問題にかぎったとしても、ロボットの「正戦論」とでも言うべきものが考察されるべきであるし、実際にピーター・アサーロなど、ロボットの正戦論を構想している研究者もいる。我々はこういった実践的考察と平行して、このような概念的アプローチもまたさらに深めていく必要があるだろう。

参考文献

C. Allen, G. Varner, and J. Zinser. [2000] Prolegomena to Any Future Artificial Moral Agent. J. Experimental and Teoretical Artificial Intelligence Vol. 12. No. 3 pp. 251-261

M. Anderson and S. L. Anderson. [2007] Machine Ethics : Creating an Ethical Intelligent Agent. A.I. Magazine. Vol. 28. No. 4. pp. 15-26.

P. Danielson [1992] "Artificial Morality: Virtuous Robots for Virtual Games" Routledge.

L. Floridi and J. W. Sanders. [2004] On the morality of artificial agents. Mind and Machine, Vol. 14. No. 3 pp. 349-379.

D. G. Johnson. [2006] Computer systems : Moral entities but not moral agents. Ethics and Information Technology. No. 8. pp. 195-204

D. G. Johnson and K. W. Miller. [2008] Un-making artificial moral agents. Ethics and Information Technology. No. 10. pp. 123-133.

A. Krishnan [2009] "Killer Robots: Legality and Ethicality of Autonomous Weapons" Ashgate.

P. Lin [2011] Introduction to Robot Ethics. In "Robot Ethics: The Ethical and Social Implications of Robotics" MIT Press.

P. W. Singer. [2009] "Wired for War: The Robotics Revolution and Conflict in the 21st Century" Penguin Press. 小林由香利訳[2010]『ロボット兵士の戦争』(NHK出版)

G. Veruggio [2006] "Euron Roboethics Roadmap"
G. Veruggio and Fiorella Operto [2008] "64. Roboethics: Social and Ethical Implications of Robotics" in "Springer Robotics Handbook." Springer.
W. Wallach and C. Allen [2009] "Moral Machines : Teaching Robots Right from Wrong" Oxford University Press.
久木田水生[2008]「ロボット倫理学の可能性」『PROSPECTUS』、第11号、1-10頁
ルチアーノ・フロリディ（西垣通訳）[2007]「情報倫理の本質と範囲」、西垣通・竹之内禎　編著訳『情報倫理の思想』(NTT出版)所収

註

1）Singer. [2009(2010)]　邦訳628頁
2）Veruggio. [2006] p. 616.
3）Veruggio. [2006] p. 615.
4）ロボエシックス(Roboethics)という言葉は、彼によれば2002年にベルージオ自身によって作られた造語である。したがって、この言葉は彼の主導したロボエシックス・アトリエに関連するものとしてのみ用いる。アレンらの異なるグループからの研究については、ロボット倫理学(Robot ethics)等と呼び分ける。
5）このロードマップでは、以下のような方針が定められた。
- ロボット倫理について、学者と利害関係者間の共通言語を発展させること。
- アイデアを生み出し、結合させるために、お互いの領域を学ぶこと。
- 異なる文化、宗教、信仰における主な倫理的パラダイムの一般的調査を発展させること。
- 異なる文化、宗教、信仰へと「適合された」倫理的ガイドラインのロゼッタストーンを定義すること。
- 特殊研究を始めること。

6）Veruggio and Operto [2008] p. 1503.
7）Allen, et. al. [2000] p. 251.
8）op. cit.
9）Lin. [2011] p. 4.
10）Allen, et. al. [2000] pp. 254-5.
11）これらの問題点は、アレンら自身が提起したものであり、現段階では解決が難しいと思われる。現に2009年の『モラル・マシーン』においても彼は、プロレゴメナで提唱したcMTTは問題点が多すぎると主張している(Wallach and Allen, [2009] pp. 206-7.)。
12）Wallach and Allen. [2009] p. 97.

13) Floridi&Sanders.[2004]p. 354.

14) フロリディ自身は明言していないものの、これは科学哲学の文脈においては「観察の理論負荷性」と呼ばれるものと同様の主張であるように思われる。

15) Floridi&Sanders. [2004] p. 368.

16) Johnson&Miller. [2008] p. 131.

17) Johnson. [2006] p. 202.

18) Johnson. [2006] pp. 202-3.

19) Johnson. [2006] p. 203.

20) Johnson. [2006] p. 131.