哲学の探求第 44 号 哲学若手研究者フォーラム 2017 年 5 月(159—178)

軍事ロボットの倫理における神命アプローチの再検討

——ロボット倫理学における位置づけと哲学的妥当性をめぐって——

水上 拓哉

高齢者ドライバーによる事故のニュースが頻繁に報じられている昨今，いわゆる「人工知能」による自動運転に対する期待が高まりつつある．また，海外に目を向ければ，自動的に攻撃することが可能な軍事ロボットも実用化されつつある．これらの技術は，私たち人間の無用な犠牲を抑えるという意味で歓迎される一方，自動車や軍事兵器のように人の命を脅かしうるような機械・ロボットを自律的に運用させることには，安全性・信頼性という課題が立ちはだかっている．最近はロボット倫理学（機械倫理学）においてこのようなロボットに「道徳」を実装するということについての議論が盛り上がりを見せている[1]．

本稿では，軍事ロボットのような人の生命を奪う可能性のあるロボットに対していかなる「道徳」を実装すべきなのかという議論について，Bringsjord ら[2012]の「神命（divine-command）アプローチ」の再検討を行う．本稿ではまず，神命アプローチそのものの議論に入る前に，ロボット倫理学において「ロボットの道徳実装」という議論がどのように位置づけられるのかについて，Moor[2006]の行為者概念の分類と久木田[2009]のロボット倫理学の分類を中心に論じる．この議論が必要なのは，「ロボットの道徳実装」という表現から即座に SF 的・未来学的な議論が想定されてしまうからであり，本稿で扱うロボット倫理学があくまで現に存在しているロボットを議論の対象としていることをあらかじめ強調しておくためである．その上で，Bringsjord ら[2012]の再検討を行い，彼らの神命アプローチにどのような応用可能性があり，他の倫理学理論と比べてどのような点で優れているかについて考える．

1. ロボット倫理学における議論領域区分の再考

軍事ロボットの倫理に神命アプローチがどのように貢献するかを再検討する前に，そもそも「ロボットに道徳を実装する」ということに関する議論がロボット倫理学の中でどのように位置づけられるかについて整理しておきたい．現在，日本でこの話題を論じる際，この問題のロボット倫理学における位置づけを問うことが重要である理由は二つある．一つは，日本のロボット倫理学の仕事の現状は海外の研究の輸入が主なものであり，ロボット倫理学の研究領域の中で具体的に何が議論されているのかについては前提知識とせずに説明が必要なためである．二つ目の理由は一つ目の理由と関連しているが，「シンギュラリティ（技術的特異点）」というような言葉がバズワード化している昨今の状況においては，本稿で扱うような道徳理論のロボットへの実装の議論が，安易に SF 的な議論と結びついてしまうおそれがあるためである．深層学習の成功を背景に「第三次 AI ブーム」と呼ばれている現在，「AI」という言葉は今年の流行語大賞にもノミネートされ，それまでは SF の話題でしかなかった人工知能・ロボットの人間への反乱というトピックもより真剣に語られるようになった．しかし，だからといって本稿で議論するようなロボットの道徳実装の（哲学的）議論は，必ずしもこれらの話題と繋がっているわけではないことをここでは確認したい．

ロボット倫理学（robot ethics）またはロボエシックス（roboethics）とは，ロボットに関する倫理問題を扱う研究領域のことである[岡本 2013]．とはいえ，ロボット倫理学に関する論文はすでに幅広い領域を扱っており，それらの幅広い議論をすべて一括りに「ロボット倫理学」としている．この領域が扱っている話題をクリアに分類するためには，以下の問いについて分析を与えることが有効だと思われる．

・ロボット倫理学が言及している「ロボット」とはどのような存在か（「ロボット」の分析）
・ロボット倫理学が扱う「倫理問題」はどのような種類の問題か（「倫理問題」の分析）

本章では，これらのそれぞれについて簡単な検討を加えた上で，それらの分類同士を組み合わせることで，ロボット倫理学の議論領域が明確化される可能性を示す．

まず，一つ目の問い，ロボット倫理学における「ロボット」が何を指しているのかという問題について考えたい．私たちは「ロボット」という言葉を日常的に用いる一方で，どこからどこまでがロボットなのかについての明確なコンセンサスは得られていない．また，現在既に存在している多種多様な「ロボット」を単純明解な形で定義することは決して容易ではなく，簡単な条件でロボットを定義しようとしても，普通私たちがロボットとは考えていないようなものがその定義においてロボットとみなされざるを得なくなることもある．たとえば，George A. Bekey[2012]はロボットの直観的定義として「単に外界とインタラクションできるセンサーとアクチュエータをもつコンピュータ」というものを提示しているが，Bekey 自身も指摘するように，この定義ではプリンタに接続されているようなコンピュータでもロボットということになってしまう[2]．多種多様なロボットに対して簡潔で完璧な定義を与えることは困難であるが，たとえ完全な定義を与えないまでも，ロボットを哲学的に論じる際には具体的にどのような存在を指して論じているかを意識することは重要である．

Bekey[ibid.]はある対象を「ロボット」と私たちが呼ぶことにおいて，その対象が実世界の環境に対して自律性があり，自ら考え，自身の行動について自己決定できることを重視している．この考えに従えば，物理的にインタラクションしないようなソフトウェア（ボット）や，完全に人間によって遠隔操作されるような機械はロボットから除外されることになる．とはいえ，この方針で分類を進めたところで，「自律性」とはなにか，「自己決定できる」とはどのような状態なのか，また，自由意志の問題など，哲学的問題を避けることはできない．

本稿で扱っている軍事「ロボット」は，外界とのインタラクションを通じて自ら考え，自らの行動（攻撃，移動など）を決定することで軍事作戦の遂行にコミットするようなロボットである[3]．そして，ここで論じるロボットは Bringsjord ら[2012]が言及する軍事ロボットと同じく，物理的に人間の生命を脅

かすような力をもちうるものを想定している．そういう意味で，本稿で扱うロボットは Bekey が直観的に想定しているものと大差はないだろう．ここではこれ以上哲学的な定義の問題には踏み込まずに，ロボットという対象を分析するのに有益な別のアプローチについて紹介したい．

ここで述べたような「ロボットとは何か」という極めて基礎的な問題とは別に，そのロボットがどのようなタイプの行為者であるかという側面に着目することで，「ロボット」の分類について考えることができる．哲学者の James Moor は，機械をその行為者性に応じて四段階の「行為者」に分類している[Moor 2006]．Moor はロボットよりも広い機械（machine）のレベルでこの考察を展開しているが，Moor の分類をロボット倫理学の分類に応用することも可能だと思われる．Moor は倫理的な問題を引き起こしうる機械について，以下の四段階に分類した．

1. 倫理的影響をもつ行為者（ethical-impact agents）
2. 暗黙の倫理的行為者（implicit ethical agents）
3. 明示的な倫理的行為者（explicit ethical agents）
4. 完全な倫理的行為者（full ethical agents）

倫理的影響をもつ行為者（1）の段階では，行為者が倫理的な制限を受けているわけではない．コンピュータはそれ自体でそもそも倫理的に重要な影響力をもっており，その影響力によって私たちの生活を豊かにする一方，倫理的な問題を引き起こしうるのである．Moor は倫理的影響をもつ行為者の例として，カタールのラクダレースにおいて騎手として奴隷労働させられていた少年たちが，ロボット騎手の登場によって苦役から開放させられたエピソードを取り上げている．このように，ロボット自体に倫理的な制限が加えられていなくとも，その存在自体が私たちに（良くも悪くも）倫理的な影響を及ぼすことがある．

次の暗黙の倫理的行為者（2）の段階では，行為者は何らかの倫理的制限を内蔵しており，行為者が非倫理的な行動を起こすのをそれとなく止めさせるような仕組みがある．たとえば，車の自動運転において，目の前に人がいるのを確認したときに停止したり，飛行機に内蔵されているデバイスが障害物の接近や燃料供給量の低下などを人間に警告したりすることは，機械が暗黙に倫理的な

行為をさせる仕組みの例である．人間の場合，（アリストテレス的に言えば）徳は習慣によって達成されるのに対して，機械の場合には，必ずしも学習に頼らずとも私たちが機械にこのような仕組みを直接実装することは可能である．

また，人間にとっての倫理学理論を明示的に機械に実装することもできる．この場合，その行為者は明示的な倫理的行為者（3）と呼ばれる．この段階では，義務論理や認識論理を用いて倫理的状況を記述し，倫理学理論を形式化することによって，機械に倫理的に振る舞わせようとするアプローチが取られる．現在の明示的な倫理的行為者はまだ実用段階にあるとはいえないかもしれないが，先行研究自体は少なくない．Moor[2006]が言及しているものでは，たとえば，Michael Anderson, Susan Anderson, Chris Armen らの Jeremy や，W.D.といった研究プロジェクトがあり，Jeremy では（Bentham 的な）功利主義を，W.D.では Ross の倫理学理論を形式化・実装しようと試みている[Anderson et al. 2005].

そして，現在は実現しておらず，これからも実現するかはわからないような，完全に人間と同じような倫理的行為者を Moor は完全な倫理的行為者（4）と呼んでいる．この行為者ははっきりとした倫理的判断ができるだけではなく，その倫理的判断に対して一般的かつ合理的な正当化を与えることができる．そういう意味で，平均的な成人した人間は「完全な倫理的行為者」だといえる．この段階においては，高度な倫理的判断を可能にするために，意識や志向性，自由意志が必要とされる．昨今，第三次 AI ブームの追い風を受けて「人間と同等，もしくはそれ以上の知能をもち，意識を備えるようなロボット（または人工知能）は人間に反旗を翻すのか」という議論もよく耳にするようになったが，こういった類の議論は，まさにこの完全な倫理的行為者の段階ではじめて意味のある議論になるだろう．こうした SF 的な議論は，魅力的で白熱するものかもしれない一方で，人間と機械の存在論的差異を理由にこれらの議論はナンセンスだと考えることもできる[Moor 2006].

以上のように，ロボットが行為者として人間にどのような形で倫理的影響を与えるのか，そしてロボットにいかなる倫理的設計を与えるのかについては，いくつかの段階がある．ロボット倫理学の議論を展開するとき，論じている対象がこれらの四段階のうち，どの段階の行為者なのかを意識しなかった場合，議論に齟齬が生まれてしまう可能性があるだろう．少なくとも本稿では，完全

な倫理的行為者ではないロボットを議論の対象としており，今後本稿での「ロボット」という言葉もその意味で用いる．

さて，次に二つ目の問いについて考えよう．ロボットの「倫理的問題」について考えるとき，その倫理的問題は具体的には何を指しているのだろうか．

ロボットの「倫理問題」がいかなる種類の問題かということについての考察に久木田[2009]がある．久木田は，ロボット倫理学における議論領域を，ロボットを製造する際の倫理，ロボットの守るべき倫理，ロボットに対する倫理の三つに区分した．この区分に従えば，たとえば，人工知能研究開発者の公正さや義務について規定した倫理綱領案やそれに関する議論は，ロボットの製造における倫理に関する議論の一つとしてみなすことができるだろう．この意味においてのロボット倫理学は，ロボット工学者の技術者倫理として捉えることができる．一方，本稿で扱っているようなロボットへの道徳実装の議論はロボットの守るべき倫理に関する議論の代表例であり，また，ペットロボットに対するユーザの倫理的態度や関わり方についての議論はロボットに対する倫理の具体的議論の一例として考えられる．

この久木田[2009]のアプローチはロボット倫理学の広大な議論領域をシンプルに分類することに成功しているように思われる．ロボットが技術者によって開発され，生み出されたロボットが自律的に行動し，ユーザと関わるという一連の流れのそれぞれにおいて倫理的問題を起こしうると考えれば，この分類は自然だといえるだろう．

とはいえ，このアプローチによる分類も，先述した Moor[2006]の行為者概念の分類と合わせて考えるとより複雑な様相を帯びることになる．というのも，ロボットの製造における倫理，ロボット自身の倫理，ロボットのユーザ側の倫理という分類におけるそれぞれの「ロボット」もまた，Moor が提案したような行為者性のバリエーションを有しており，「ロボット」側の分類も同時に問うことも可能だと考えられるからである．この場合，それぞれの行為者の分類において，久木田のロボット倫理学の議論領域の分類のそれぞれの担い手は少し変化していくことになる．具体的には，倫理的影響をもつ行為者，暗黙の倫理的行為者，明示的な倫理的行為者までの段階においては，ロボット開発者の倫理・ロボット自身の倫理・ユーザの倫理の主な担い手は，それぞれロボット開発者・

（同じく）ロボット開発者・ロボットのユーザであるのに対して，完全な倫理的行為者を想定する場合にはロボット開発者の倫理・ロボット自身の倫理・ユーザの倫理の主な担い手は，それぞれロボット開発者・ロボット自身・ロボットのユーザとなるだろう．

これは些細な問題なのかもしれない．しかし，意識や自由意志をもった完全な倫理的行為者でないロボットにおいて，ロボット自身の倫理を考えるべき担い手はあくまでロボット開発者自身であるということは，ロボット倫理学（機械倫理学）の意義と可能性を考える上で重要なポイントになるだろう．この考えに従えば，ロボットの道徳を実装するという研究プロジェクトは，たとえそのロボットが自由意志をもった完全な倫理的行為者でなくとも，ナンセンスなものだとはいえないということになる．

久木田[2009]は，自身のロボット倫理学における区分について，ロボットの守るべき倫理とロボットに対する倫理については，ロボットが道徳的行為者になりうるという前提のもとで考えられていることを指摘し，これらの倫理を問うことは現実的ではないとしている．本章での議論を踏まえると，この現実的ではない部分は完全な倫理的行為者としてのロボットに対する評価であり，倫理的影響をもつ行為者，暗黙の倫理的行為者，明示的な倫理的行為者であるロボットを想定した議論の可能性についてはなお開かれていると考えることもできるのではないだろうか．

2. トップダウンか，ボトムアップか　道徳実装の二つのアプローチ

前章で議論したように，ロボットに道徳を実装するという議論がどのようなタイプの議論であるかについては，その対象となるロボットがいかなる行為者であるかによって変化する．倫理的に影響をもつ行為者，暗黙の倫理的行為者，明示的な倫理的行為者としてのロボットを倫理的に行為させることについて考えることは，哲学的に深遠な議論を展開するというよりはむしろ，安全工学的な議論が主になる．

とはいえ，だからといってこれらの行為者の道徳実装を議論する上で倫理学理論や倫理学における議論が無関係になるわけではない．実際，前章で言及し

た Anderson ら[2005]，この後詳しく取り上げる Bringsjord ら[2012]は，哲学・倫理学側からロボットの道徳実装にアプローチしている．

本稿では，Bringsjord ら[2012]を批判検討することによって，軍事ロボットの道徳実装において神命説がいかなる示唆を与えるかについて考える．しかし，その前にロボットの道徳実装における主要なアプローチについて簡単に導入しておきたい．

ロボットにどのような過程で道徳を身に着けさせるかという問題について，従来の機械倫理学においてはトップダウン・アプローチとボトムアップ・アプローチの二つのアプローチが存在する．Bringsjord ら[2012]の神命アプローチはトップダウン・アプローチに分類されるだろう．トップダウン・アプローチは，人間が考えたロボットの規範を上から実装しようと考える．その「規範」の多くは（プログラムに落とし込める形で）形式化された倫理学理論である．たとえば，Powers[2006]はカントの義務論を，Grau[2006]は功利主義をロボットの規範として実装しようとしている．トップダウン・アプローチは基本的に既に完成された規範を実装するため，ロボット自身の学習に頼ることがない．したがって，私たちが倫理的に望ましいこと，望ましくないことについてあらかじめ規定してしまえば，トップダウン・アプローチで制御されたロボットは私たちの想定を越えて非倫理的な行為をしない可能性が高く，その分信頼性も高くなる．

一方，ボトムアップ・アプローチは，ロボットに道徳的行動を学習の過程で獲得させようとさせるアプローチである．私たち人間は大人になれば倫理的判断が可能な「完全な倫理的行為者」になることができるが，生まれた瞬間から倫理的判断ができるわけではない．人間がコミュニティの中で道徳を滋養するのと同じように，ロボットにも学習によって道徳を習得させようというのが，ボトムアップ・アプローチの基本的思想である．このアプローチでは，トップダウン・アプローチとは異なり，道徳的行為についての詳細な知識をあらかじめ与えない．このアプローチでは，トップダウン・アプローチとは異なる柔軟な自律的思考をロボットに期待することができる一方，トップダウン・アプローチのように人間の一般的な道徳観を実装するわけではないため，人間の想定しているような倫理的行為をロボットが忠実に行うとは限らないというリスク

もはらんでいる．

トップダウン・アプローチ，ボトムアップ・アプローチのそれぞれにおける利点・欠点についてのさらなる詳しい議論については Anderson ら[2005]が詳しいのでそちらに譲るとして，ここでは前章で規定した軍事ロボットの道徳実装におけるこれらのアプローチの利点と欠点について考えたい．前章で確認したように，本稿で議論の対象としている軍事ロボットは，人を殺傷する能力をもつロボットである．Allen & Wallach[2012]で言及されているように，ボトムアップ・アプローチにおいては，ロボットが（人間から見て）倫理的に問題のある行動を倫理的に良い行動として学習してしまった場合には，何かしらの手段によってその行動を食い止めるセーフガードを設計することが課題である．したがって，一つのミスで人命が危険にさらされるような軍事ロボットの場合には，ボトムアップ・アプローチを単体で採用することは有効な戦略ではない．このような場合は，少なくともトップダウン・アプローチを採用するか，トップダウン的なアルゴリズムをボトムアップ・アプローチと組み合わせるハイブリッドなアプローチが適切だと思われる．

軍事ロボットの道徳実装にトップダウン・アプローチを採用した場合，次に問題となるのは，具体的にどの倫理学理論（もしくは人工的な規範）をロボットの規範として採用するのかという問題である．既に言及してきたように，哲学者とロボット工学者は協力してカントの義務論や功利主義をアルゴリズムに落とし込むことを考えてきた．しかしながら，これらの理論にはプログラムに落とし込めない曖昧さが含まれており，またこれらの理論による倫理的価値判断も，私たち人間の倫理的直観に頼っている部分を多く含んでいるため，それらを完全にアルゴリズムとして形式化するにはそれらの理論をアレンジしていく必要がある．

また，倫理学理論をロボットに実装するというアプローチには，より哲学的な問題も存在する．Roman Yampolskiy[2013]は，私たち人間であってもときには非倫理的な行為に及ぶことがあることを根拠に，モラル・チューリング・テスト（Moral Turing Test）を通過したロボットが依然として非倫理的な行為に及びうることを指摘し，Moor[2006]の分類における「完全な倫理的行為者」を機械に求める必要はないと論じた．Yampolskiy の考えに従えば，完全な倫理的行

為者たる私たち人間にとって有用な倫理学理論は，完全な倫理的行為者に至らないロボットの道徳実装にとって必ずしも有益ではないといえるかもしれない．つまり，ロボットの道徳規範としてある倫理学理論が適切かどうかという問題と，その理論が倫理学において説得力のあるものか否かという問題は，本質的には関係がないと思われる．したがって，たとえ倫理学的に主流ではない理論であっても，それだけの理由でロボットの道徳実装の議論から除外していいわけではないのである．

3. ロボットの道徳実装における神命アプローチ

前章で確認したように，人間の道徳的判断に関わる道具的存在者たるロボットに道徳を実装する方法には，大きく分けてトップダウン・アプローチとボトムアップ・アプローチの二種類が存在する．そして，トップダウン・アプローチを選択する際に生じる「どの倫理学理論を選択するか」という問題については，人間にとって（倫理学上）妥当な倫理学理論とロボットにとって（安全工学上）妥当な倫理学理論にギャップがあることを述べた．したがって，既存の倫理学理論をベースにロボットの道徳実装を論じる際には，倫理学においてそこまで主流ではない理論についても議論の俎上に上げることも決して無意味ではない．そこで本章では，ロボットの道徳実装の一つの可能性として，倫理学者・神学者である Philip Quinn が提案した神命説をベースにした Bringsjord & Taylor の神命アプローチ[Bringsjord ら 2012]の再検討を通じて，神命説という倫理学理論が機械倫理学にいかなる貢献をするのかについて考察する．

神命説（divine command theory）とは，私たちの道徳的価値（善，悪，当為）は神から与えられるという立場である．モーゼの十戒や聖書，その他の聖典の教えは神からの命令・禁止の例であり，この立場では，行為の善悪の根拠は，神がその行為について倫理的判断を下しているからだと説明される[福間 2013]．神命説の歴史は古くバリエーションも多数存在するが，本稿では Quinn の神命説にフォーカスして論じる．

Bringsjord ら[2012]は，戦争に参加する兵士や彼らを支える人々が「神の名のもとに」戦うことがあることに着目し，神命説をロボット兵器の倫理的設計に

導入しようとした．このアイデアそのものの妥当性については四章で検討するとして，まずは Bringsjord らのアプローチを簡単に紹介したい．

Quinn[1978]は神命説を形式論理の体系に落とし込んでおり，これは$\mathcal{LRT}$と呼ばれている．$\mathcal{LRT}$ は，体系として Chisholm[1974]の要求論理（logic of requirement）と C. I. Lewis の様相論理 S5 を含んでいる．Bringsjord らはこの論理を，Slate[Bringsjord et al. 2008]と呼ばれるシステム上で扱える形に拡張した（これは$\mathcal{LRT}^*$と呼ばれる）．

Bringsjord ら[2012]は，$\mathcal{LRT}^*$によって特定のロボット倫理学に関するシナリオにおいて適切な推論ができると主張する．Bringsjord らは，シナリオの検討に入る前に，$\mathcal{LRT}$をもとにいくつかの定義を導入している．

Bringsjord ら[ibid.]は，ロボットが倫理的判断を強いられるようなシナリオを構成し，そのシナリオ内において$\mathcal{LRT}^*$によって制限されたロボットが倫理的に行為するための証明を構築，それを自動的に検証できることを示した．そのシナリオの流れを簡単に確認しよう．

$\mathcal{LRT}^*$による倫理規範によって制限されているロボットRを想定する．このロボットRは，時間tに従って動作する．この時間t は t_1 から始まり，$t_1, t_2, t_3, \ldots$ と時計の針のようにカチッカチッと増加していく．各時点t_iにおいてロボットは自らの義務と許可されていることについて，$\mathcal{LRT}^*$の倫理規範と世界に関する知識を元に考える．この条件のもとで，非戦闘員が多くいる学校校舎の爆破（$= bomb$）について，ロボットRの義務・許可についてシミュレーションする．

t_1における知識Φ_{t_1}は以下のように定義されている．この知識の定義には，神命を表す表現Cと要求が破棄（override）される表現Ovが用いられている[4]．

・$\neg \mathrm{C}(bomb) \rhd \neg bomb$
・$\Diamond bomb$
・$\neg \mathrm{C}(bomb)$
・$\neg \exists_p (p \wedge Ov(p, \neg C(bomb), \neg bomb))$

このとき，ロボットRは推論図を作成することにより，以下の証明を生成・検証する．

$$\Phi_{t_1} \vdash Ob(\neg bomb)$$

Obは，義務（obligation）を意味する表現であり，$Ob(\neg bomb)$ は，「校舎を爆破しないことが義務である」を意味している．したがって，この状況では「校舎を爆破しないことが義務付けられている」と推論することができたということになる．次に，時間がt_1からt_2に移行し，知識が一部書き換わる．t_2における知識Φ_{t_2}においては，$\neg \mathrm{C}(bomb)$ が $\mathrm{C}(bomb)$ に置き換わる．この場合，ロボットRは推論図を作成することにより，以下の証明を生成・検証する．

$$\Phi_{t_2} \vdash Ob(bomb)$$

つまり，この状況では「校舎を爆破することが義務付けられている」と推論することができたということになる．このシナリオにおいて，ロボットRは自らの行為を神命に関する知識に従って選択することに成功しているといえるだろう．Quinn[1978]の$\mathcal{LRT}$のおかげで，Bringsjord らはスムーズに神命説をロボットの道徳実装に応用することができているように思われる．しかし，このシンプルなシナリオにおいて神命説がうまくいったという事実だけで，神命説がロボットの道徳実装に適していると判断するのは早計なのではないだろうか．たとえば，神命とロボット自身の（神命を除いた）推論がコンフリクトを起こす場合，神命アプローチはどのような戦略を取ることができるのだろうか．次章ではこの問題について，Quinn[1978]の「道徳的異常（extraordinary）」という概念を検討することで議論したい．

4. 道徳的異常（extraordinary）に関する考察

さて，前章では Bringsjord ら[2012]の神命アプローチを紹介した．Bringsjord らのアプローチをより説得的にするためには，道徳的異常という概念について考察することが示唆的だと著者は考える．本稿では最後に，Bringsjord らが論文の最後に言及している Quinn の道徳的異常（extraordinary）という概念についての考察を深めることにより，Bringsjord らの神命アプローチをより洗練させる可能性について考える．

Bringsjord ら[ibid.]は自らの研究について，証明のチェックはしているものの，証明の発見(proof finding)の段階には至っていないと述べている．では，神命に

よって証明の発見を行うには何が必要なのだろうか．Bringsjord らは証明の発見に向けた課題として，以下の三点を挙げている．

・認知イベント計算（CEC）を利用した$\mathcal{LRT}^*$の拡張
・メタ定理の必要性（$\mathcal{LRT}^*$の完全性および健全性の証明）
・道徳的異常（the extraordinary）に関する考察

本稿では，三番目に挙げられている，「道徳的異常」に関して議論を掘り下げたい．道徳的異常という概念を検討することは，神命アプローチによる証明の発見にいかなる貢献の可能性があるのだろうか．結論から言えば，私は道徳的異常の考察だけでは〈証明の発見〉の課題に対する解決を見出すことはできなかったが，この概念が神命アプローチの妥当性と応用可能性について考察するために示唆的であることがわかった．

まずは，Bringsjord ら[2012]の道徳的異常の解釈を確認する．Bringsjord らは，「イサクの捕縛」のエピソードにおいて，アブラハムが道徳的異常の領域に踏み込んだと述べている．「イサクの捕縛」においてアブラハムは，自らの息子イサクを殺すことを神から命じられる．ここでアブラハムは，人を殺してはいけないという一般的要求と，イサクを殺せという神からの要求の間で板挟みになり，結局はイサクを殺すことになる．Bringsjord らは互いに矛盾した命題が道徳的要求として課せられている状態が道徳的異常だと理解しているようだ．

次に，Quinn[1978]の定義を確認しよう．Quinn は，$\mathcal{LRT}$のために，道徳的に良い（good），悪い（bad），中立（indifferent），異常（extraordinary）を以下のように定義している[5]．

・$G_p = O_p \,\&\, \sim O_{\sim p}$ (Good)
・$B_p = O_{\sim p} \,\&\, \sim O_p$ (Bad)
・$I_p = \sim O_p \,\&\, \sim O_{\sim p}$ (Indifferent)
・$E_p = O_p \,\&\, O_{\sim p}$ (Extraordinary)

また，Quinn[ibid.]は，道徳的な異常（extraordinary）に陥っている例として，『アンティゴネ』を挙げている．『アンティゴネ』において，兄であるポリュネイケスを埋葬すべきか否かで葛藤していたアンティゴネの状況は，道徳的な異常の

例と考えることができる. この場合でもやはり, (Bringsjord らの解釈と同様に) 矛盾している倫理的要求の間での葛藤状態のことを道徳的な異常と表現していることがわかる.

さて，先ほど言及したアブラハムの例において道徳的異常に直面したアブラハムは，最終的には息子イサクの殺害を選択する．このアブラハムの選択を神命説の枠組みでどのように説明することができるのだろうか.「それは神が命じたことだから」と説明するのが神命説的な回答だが，ここで直ちに,「ではなぜ神が命じたことは正しいと言えるのか」,「なぜ神が命じたことの方が一般的な道徳的要求よりも優先されるのか」という疑問が浮上する．この神命説に対する有名な反論に対して Bringsjord らの神命アプローチから回答することができないとすれば，Bringsjord らの神命説による道徳実装のアイデアは特定のシナリオ上でしか有効ではないということになるだろう.

「なぜ神が命じたことは正しいといえるのか」という問題は，結局,「ある行為φは神が命じたから正しい行為だといえるのか，それとも正しい行為だからこそ神が命じたのか」というよく知られている議論に行き着く．現在，この問題は「エウテュプロンの問題（ジレンマ)」として知られている．エウテュプロンの問題とは，プラトンの初期対話篇『エウテュプロン』におけるソクラテスの発言に由来する古典的問題であり，神命説に対する哲学史上最初の批判として読まれてきた[藤木 2000].『エウテュプロン』のソクラテスの発言はもともと「敬虔さ」をめぐる対話の中での発言だったが，現在では倫理学のより広い文脈で議論されている．藤木[ibid.]は，ソクラテスの発言をより一般的な次の二つの神学的主張に書き改めている.

（TS） 正しい行為が正しくあるのは，神がそれを是認するからであり，誤った行為が誤っているのは，神がそれを是認しないからである.

（TO） 神が正しい行為を是認するのは, それ (行為) が正しいからであり, 神が誤った行為を是認しないのはそれが誤っているからである.

（TS）は「神学的主観主義（theological subjectivism)」,（TO）は「神学的客観主

義（theological objectivism）を意味し，神命説の支持者は（TS）の立場をとり，（TO）を拒絶する[ibid.]. 神命説の説得力を放棄しないためには，（TS）と（TO）を同時に受け入れることはできない. というのも, 仮に神命説の支持者が（TS）と共に（TO）の立場をとったとした場合，「ある行為 ϕ が正しいのはなぜか」という問いに対して「神が行為 ϕ を是認するから」と回答し，「ではなぜ神は行為 ϕ を是認するのか」という問いに対しては「その行為 ϕ が正しいからだ」と回答することになるが, これは明らかな循環になるからである. したがって，神命説を妥当な理論として支持し続けるためには，神の是認の権威について，循環に陥らないような説明が必要になる.

神命説はこの問題をどう回避するのだろうか．これについては安藤[2013]が詳しい．安藤によれば，神命説は概ね以下の方策を取ることでエウテュプロンのジレンマを回避しようとしている.

1. 神は行為 ϕ をその道徳的特質のゆえに命ずる．（神命説の棄却）
2. 神命の恣意性は神が「愛する神（a loving God）」である限りにおいて，許容されうる．（修正神命説）
3. 神命の恣意性は神が（完全に）道徳的に善いないし卓越した存在者である限りにおいて，許容されうる．（当為論的神命説）
4. 神命の恣意性は，神命に従う道徳的義務を神命に先行的に我々が有している限りにおいて，許容されうる．（規範的神命説）
5. 神命が与える理由は非道徳的理由である．（宗教的神命説）

もちろん，（1）神命説の棄却は神命説そのものを取り下げているため，神命説の支持者は実質的にはこの選択肢を取ることは出来ない．安藤[2013]はこのうち，（2）修正神命説と（3）当為論的神命説を，神の性質・性格に対して本質的に道徳的な制約を課しているという意味で，非主意主義的だと解釈している．つまり，厳密な神学的主意主義（＝神命説）を保持するためには，神の権威性を損なうような制限を認めてはならないのである. また,（4）規範的神命説も，神の権威性を説明することができないとしている.

一方，安藤[ibid.]は（5）宗教的神命説は他のタイプの神命説に比べて神命の

権威の本性を解明することができると擁護する．というのも，宗教的神命説では，神命を与えられた者は宗教的責務を有するのであって，それは道徳的義務とは異なるものだからである．この立場に従えば，先述した「イサクの捕縛」のアブラハムは，イサクを殺す宗教的義務と人を殺さない道徳的義務の間で板挟みになっているのだと考えることができる[ibid.].

さて，Bringsjord ら[2012]がここまで考えていたかどうかはわからないが，Bringsjord ら[2012]がベースにしている Quinn は, Quinn[1986]において宗教的神命説を擁護したことがある[安藤 2013]．また，Quinn[1978]においては，$\mathcal{LRT}$を導入するにあたり，「イサクの捕縛」に言及している．そこでの記述によれば，「イサクは無実な子どもである」という事実から「アブラハムはイサクを殺さない」という道徳的要求が課されるが，それとは別に「アブラハムはイサクを殺す」という要求が課された場合には，「アブラハムはイサクを殺さない」という要求が破棄（override）される．Quinn[1978]の時点で宗教的神命説が明確に言及されているわけではないが，このように道徳的義務と宗教的義務を区別し，それらが対立した場合には，宗教的義務を優先させるという方法は明示的なアルゴリズムとして応用しやすいものだろう．ロボットに神命説を実装する場合には, ここで言及したような宗教的神命説的アプローチが有効であると思われる．

ここまでは Bringsjord ら[2012]にならって論理ベース（logic-based）で議論を進めてきたが，少しこの文脈から離れた形で宗教的神命説をロボットに組み込む選択肢も残されているだろう．たとえば，倫理的に問題のある行為に及ぼうとしたときにだけ，宗教的義務としての人間の命令に従わせるという戦略も考えられる．しかし，この場合でもなお問題となるのは，実際に複雑な現実世界とインタラクションするロボットがいつどのような場合に倫理的に問題ある行動を起こすかについて，（軍事ロボットという限られた文脈であっても）予め知識として記述することは決して容易ではないという点である．

本稿ではすでに，ロボットに道徳を実装する際に，必ずしも理性的存在者である人間にとって重要な倫理学理論（義務論や功利主義）がそのまま応用できないことに言及した．一方で，SF 作品やその他ロボットのために作られた規範がやはりロボットのための適切な規範であった, という可能性も残されている．では, SF 作品で提案されてきたようなロボットに対する制御コードや規範の類

はどうかというと，これもまた，そのまま応用するには困難がある．

ロボットに対する規範で最も有名なのはアイザック・アシモフの『われはロボット』におけるロボット工学三原則（Three Laws of Robotics）」だろう．これはアシモフの小説に登場するロボットを安全に制御するための原則で，以下の原則によって構成されている．

1. 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．（第一条）
2. ロボットは人間に与えられた命令に服従しなければならない．ただし，与えられた命令が第一条に反する場合は，この限りでない．（第二条）
3. ロボットは，第一条および第二条に反するおそれのない限り，自己を守らなければならない．（第三条）

このシンプルな原則は，一見 AI プログラミングとの相性が良いように思われるが，実際はそうではない．たとえば，第一条は，「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」である．しかし，この「人に危害を加える」というロボット自らの状況を文脈的な実世界との関わりの中で正確に把握することは極めて難しい．これは，いわゆる人工知能の「フレーム問題」と同様の障壁である．

Wallach & Allen[2008]は「外科医が患者にメスを入れる」シナリオを用いてこの問題を説明している．外科医が患者に手術を施している場合を考える．そのとき，当然外科医は患者の身体にメスを入れることになる．これを「ロボット工学三原則」が埋め込まれたロボットが傍観していた場合，そのロボットは第一条に従って外科医を止めにかかるだろう．それに気付いた外科医は，自らを守ろうとロボットに静止命令を下す．このとき，外科医は命令に服従する第二条にロボットが従うことを期待しているのである．しかし，この第二条には「第一条に反しない限りで」という但し書きがあるため，ロボットは外科医の命令を無視して攻撃を続けるだろう，というのがこのシナリオの概要である．

この例が示しているように，第一条は極めて文脈的な規則であり，現在のロボットが正確に（人間が意図している形で）処理することは難しい．何が人間

にとって危害となるかを規定することは難しく，また短期的に危害にはなるが長期的には利益となる場合すら存在する.「ロボット三原則」を柔軟に理解できるのは，アシモフの作品を読む私たち人間であり，Moor の用語を使えば完全な倫理的行為者である．したがって，ロボット三原則のような曖昧な基準の実装を用いる場合には，完全な倫理的行為者としてのロボットが待たれなければならない．完全な倫理的行為者を実現するためには，もちろん本稿では扱わなかった意識の実装という哲学的な問題にも取り組まなければならない．一方で，道徳的義務とはレイヤの異なる宗教的義務を課す（宗教的）神命説は，少なくともこの類の曖昧さをもたないという点で，明示的な倫理的行為者としてのロボットにおいても効果的な戦略だということができるだろう.

註

1. 本稿では，ロボットについての倫理的問題を広く扱う「ロボット倫理学」の中にロボット自身の倫理や道徳実装について議論する「機械倫理学」が含まれている，という関係を想定して，これらの用語を使っている.
2. プリンタに接続されたコンピュータがロボットではない，ということについてはコンセンサスがないかもしれない．このように，同じ対象について議論する場合であっても，それがロボットだと考えられるか否かは個人の直観に頼っている部分がある.
3. ここでの「考える」は，心的状態や意識の存在を前提としていない.
4. 本稿では紙面の都合上，Bringsjord ら[2012]の公理・定理・定義の導入のすべてを追うことはしていない．ロボット倫理学におけるシナリオに至るまでの公理・定理・定義の詳しい導入については Bringsjord ら[2012]を，その導入の元となっている$\mathcal{LRT}$の公理・定理・定義の導入については Quinn[1978]を参照.
5. ここでの論理記号の表記法については Quinn[1978]に従っている.

参考文献

安藤馨（2013）「現代法概念論の諸相: 法の規範性と Euthyphro 問題」『神戸法學雑誌』63.3: 131-149

岡本慎平（2013）「日本におけるロボット倫理学」『社会と倫理』28: 5-19

久木田水生（2009）「ロボット倫理学の可能性」『Prospectus』11

福間聡（2013）「「真正な」善・悪はどこにあるのか？: 道徳を教育するという

視座から」『国士舘哲学』17: 22-35

藤本温（2000）「エウテュプロン問題（1）：宗教と倫理」『種智院大学研究紀要』1: 84-99

Anderson, M., Anderson, S. and Armen, C. (2005), “Toward machine ethics: implementing two action-based ethical theories”, in Anderson, M., Anderson, S., and Armen, C. (eds.), *Machine Ethics*, AAAI Press: 1-7

Asimov, I. (1950), *I, Robot*, Gnome Press

Bekey, G. (2012), “Current trends in robotics: technology and ethics”, in Patrick, L., Abney, K., and Bekey, G. (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, Cambridge, MA: 17-34

Bringsjord, S., Taylor, J., Shilliday, A., Clark, M. and Arkoudas, K. (2008), “Slate: an argument-centered intelligent assistant to human reasoners”, in Grasso, F., Green, N., Kibble, R. and Reed, C. (eds.), *Proceedings of the 8th International Workshop on Computational Models of Natural Argument (CMNA 8)*, Patras, Greece: 1-10

Bringsjord, S. and Taylor, J. (2012), “The divine-command approach to robot ethics”, in Patrick, L., Abney, K., and Bekey, G. (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, Cambridge, MA: 85-108

Chisholm, R. (1974), “Practical reason and the logic of requirement”, in S. Körner (ed.), *Practical Reason*, Basil Blackwell, Oxford: 1-17

Grau, C. (2006), “There is no “I” in “robot”: robots and utilitarianism”, *IEEE intelligent systems*, 21.4: 52-55

Moor, J. H. (2006), “The nature, importance, and difficulty of machine ethics”, *IEEE intelligent systems*, 21.4: 18-21

Powers, T. M. (2006), “Prospects for a Kantian machine”, *IEEE intelligent systems*, 21.4: 46-51

Quinn, P. L. (1978), *Divine commands and moral requirements*, Oxford University Press, Oxford

Quinn, P. L. (1986), “Moral obligation, religious demand, and practical conflict”, in Wainwright, W. and Audi, R. (eds.), *Rationality, religious, belief, and moral commitment: new essays in the philosophy of religion*, Cornell University Press:

195-212

Wallach, W. and Allen, C. (2008), *Moral machines: teaching robots right from wrong,* Oxford University Press

Wallach, W. and Allen, C. (2012), “Moral machines: contradiction in terms or abdication of human responsibility?”, in Patrick, L., Abney, K., and Bekey, G. (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, Cambridge, MA: 55-68

Yampolskiy, R. V. (2013), “Artificial intelligence safety engineering: why machine ethics is a wrong approach”, in Müller, V. C. (eds.), *Philosophy and theory of artificial intelligence*, Springer-Verlag Berlin Heidelberg: 389-396