Wendell Wallach and Colin Allen, *Moral Machine: Teaching Robots Right from Wrong*, Chapter 5 "Top-down morality" 要旨

久木田水生

2012 年 7 月 27 日

Putting Ethical Theories to Work

P. 83–("What does the engineer.....")

AMAs の設計にとっては、トップダウンの、つまり理論駆動型のアプローチが適切であるように思われるかもしれない。その理由の一つは、理論は包括的な解決を与えてくれる、ということである。著者たちは倫理規則を形式的な意思決定アルゴリズムとして実装できる見通しは暗い、と考えるが、にもかかわらず人間は実際にトップダウンルールに訴えて行動を決定し正当化する。AMAs の設計者は道徳性のこの側面を考慮しなければならない。

人工道徳に対するトップダウンのアプローチとは、アルゴリズム化することができる一連の規則を持つということである。これらの規則は任意のものでよい。これは道徳の「戒律（commandment）」モデルである。戒律モデルの課題は、規則同士の衝突に対処することである。衝突の問題に対処するために、一部の哲学者たちは、特定の個別的規則が由来する、より一般的で抽象的な原理を見つけ出そうとしてきた。別の哲学者たちはトップダウン規則をヒューリスティックとして理解しつつ、倫理規則が包括的な意思決定手続きを提供するという考えを拒否する。しかし倫理規則が行動の規範であるのかヒューリスティックであるかという議論はここでは論じられない。

ここでは特定のトップダウン理論を実装するための計算論的要件がどのようなものであるかに焦点を当てる。それはアルゴリズムのためのタスクの特定として適切なのか？ そうでないとすれば、そのことは AMAs を作ると言うプロジェクトにとって何を意味するのか？

倫理的推論は一つの一般原則のもとにもたらされると考える学派には、**功利主義**と**義務論**という対立する二つのものがある。

功利主義は、道徳性とは究極的には世界中の効用を最大化することに関係していると考える、**帰結主義**の一種である。功利主義には、行為を道徳的評価の対象とする**行為功利主義**と、行為を生み出す規則を対象とする**規則功利主義**の二種類がある。まずは前者について考える。功利主義的 AMAs は重い計算的要求に直面する。というのもそれは行為の道徳性を評価するために、その選択の帰結の多くを考えなければならないからである。行為者にとっての問題は、ある尺度の効用を最大化するようにどのように様々な行為の帰結を決定するかである。設計者にとっての問題は、帰結とその最終的な効用を決定するメカニズムをどのように作るかという

ことである。

義務論は、倫理の核にあるのは義務だと考える立場である。ここでは権利と義務は表裏一体のものと理解される。一般に義務あるいは権利のいかなるリストも内部衝突の問題を抱える可能性がある。それを解決する一つの方法は、それらの義務をより高次の原理に従属させることである。例えばカントは、すべての正当な道徳的義務は**定言命法**という一つの原則に基礎づけられると考えた。

義務論的アプローチをとる人工的行為者にとって重要なのは、規則（あるいは規則の決め方）を知ることと、特定の課題に規則を適用する方法を持つことである。また特定の規則の妥当性について整合的に反省することができることが望ましいが、これは高望みだろう。設計者は状況が規則の適用を要求するときに、その規則が有効化されていることを保証する方法を見つけ出し、規則の衝突に対処するためのアーキテクチャーを定式化する必要がある。

功利主義と義務論のアプローチに共通するのは、どちらも理論をリアルタイムで適用するために必要な情報のすべてを収集し比較することが何らかのコンピュータにできるかという問題である。この問題は特に帰結主義的アプローチにとって信仰である。というのも行為の帰結には本質的に時空的限界がないからである。

道徳についての一般的な議論は義務と効用だけを問題にするわけではなく、人格もまた問題にされる。この第三の要素はアリストテレスに起源が求められ、現在では「**徳倫理**」として知られる。徳倫理理論家は、道徳的に良い行いは良い人格を育てること（cultivation）から生じるのであり、そしてそれは特定の徳を実現することに存する。AMAs の設計への徳倫理の負うようについては第 8 章で論じる。

Is An Omniscient Computer Needed?

P. 86–（“The eighteenth-century British philosopher Jeremy Bentham.....”）

ジェレミー・ベンサムらは道徳を客観的な土台に乗せたいと考え、状況を量的に評価する方法を思い描いた。それは行為から生じる善と害に数字を割り当てるということである。効用の量的計測は単純な意思決定規則を可能にする。すなわち、結果的に全体の効用が最も高くなる行為を選択せよ、というものである。

数学的に扱える量を扱うという点で、功利主義は AMAs にとって唯一魅力的な種類の倫理理論を提供すると思われるかもしれない。しかし実際に功利主義的 AMAs を作るためには何が必要だろうか。ジェイムズ・ギプスは帰結主義的ロボットの計算的要件を次のように素描した。

1. 世界の状況を記述する方法
2. 可能な行為を生み出す方法
3. 現在の状況である行為が行われた時に、そこから帰結する状況を予測する方法
4. 良さと望ましさの観点で状況を評価する方法

これはアルゴリズムの特定というにはまだまだ不十分であるが、しかし関連するサブタスクを特定するための有用な枠組みを提供する。実際に功利的推論を実装するには、それぞれのサブタスクについて、それをどう設計するかを決定しなければならない。

最後のものから考えよう。ベンサムとミルはこの点で同意していないことは良く知られている。ベンサムはどんな種類の快もそれ自体として他の種類の快より優れているということはないと考えた。他の論者はすべての快が同じ尺度で評価されると考えるのはナンセンスだと考える。ある善を得るため、あるいはある害を避けるために、人がどれだけの金額を払う意志があるかによってその価値を計るという方法が示唆されることがあるが、多くの人々にとって道徳的価値をお金の価値と同一視することは極めて不適切である。

効用に数を割り当てることができたとすれば、コンピュータは帰結主義の理論の応用に向いていると思われるだろう。一方で、現在の利益と将来の危害を比較し計る、あるいはその逆、現実の利益と危害を潜在的なリスクと利益を計る適切な計算可能な評価関数を作ることは困難な問題である。

ギプスの記述した他のサブタスクについてもそれを完結させるために集めなければいけない情報の種類を考えると、計算の複雑さが明らかになる。

第一のサブタスクについて。状況の関連する要素は、道徳性に参加する対象の範囲（ここには人間、動物あるいは生態系も含まれうるが、それらはおそらく異なるウェイトを持つだろう）によって異なる。いずれにせよ倫理的に重要なあらゆる主題についての状況を記述するために必要とされるデータの集まりの大きさだけでも、途方に暮れてしまう。バーナード・ウィリアムズは、そのためには「すべての人の選好を知り、それらをまとめる『全知』の、善良な観察者——世界エージェントと呼ばれるかもしれない——」が必要だどうと論じている。

第二のサブタスクも要素の範囲に影響を受ける。道徳的に重要な事実がより多様であればあるほど、プログラミングがより精密（fine-grained）でなければならない。

第三のサブタスクに対処するためには、アルゴリズムの設計者は少なくとも次の二つの大きな問題に答えなければならない。すなわち、どの未来の分岐が計算されるべきか、そして遠い将来の結果は割り引きされる（discounted）べきか？

最初の質問に関して。すべての行為は無限に多くの二次的な影響を及ぼす。そのすべてを計算することは不可能である。さらに二次的な影響は非常に遠い影響を及ぼす（バタフライ効果）。さらに不完全な情報に基づく予測の問題もある。これらは天気予報に共通の問題である。しかしこれらの問題があるからといって天気予報をやめる理由にはならない。天気予報で使われる費等の便利な方法はいくつかのコンピュータモデルの平均をとるという方法である。「効用予報」も同じように複数のアプローチをとれるかもしれない。

二番目の質問に関して。「慈善は家庭から」、「地球規模で考え、局地的に行動せよ」ということわざは、倫理的行為は知覚にいる人々や場所に対する関係に基礎づけられる、という考えを表わしている。空間について言えることは時間についても言えるだろう。遠い将来の帰結は直後の帰結ほど強く人々を引き付けない。これが倫理的に適切な態度かどうかはともかく、AMAs がトップダウンの帰結主義的原理によって、倫理的に重要できる仕方で行為するべきものならば、将来のそして遠くの帰結を割り引きする何らかの方法は必要だろう。

ギプスの第四のサブタスクに関して、功利主義者の中でも、異なる主体の快や満足を異なった仕方で計るかどうかについて、意見が割れている。一つの方法は、集められる限りの主観的効用のランク付けを集めて、計量の公式をそれらに適用し、AMAs の選択と行為が満足のいくものであるように思われるまでその公式を調整し続ける、という方法である。もちろんこれにはリアルタイムで効用の主観的評価を集めることに関する深刻な困難がある。

功利主義的 AMAs を無限の計算から守るためには、ギプスの四つのサブタスクを達成する有効な戦略が必要である。計算を終わらせることの難しさは、計算という行為そのものの倫理的な帰結があるということである。助けを必要としている人がいるのに、意思決定に時間がかかりすぎてその人を助ける機会を失うとすれば、その意思決定の過程は機能していない。

この問題は人間にも共通する。人間はハーバート・サイモンが「限定された合理性（bounded rationality）」と呼ぶものを実践している。それは人の合理的意思決定における非常に限られた一連の考察を含む。問題はより制限された計算システムが、十分な道徳的行為者に慣れるのかどうかである。

サイモンと共同研究者のアレン・ニューウェルは AI において**ヒューリスティック**——機能的な近似、すなわち「経験則」——を使用したパイオニアである。おそらく同様の役割をする道徳的ヒューリスティックを開

発することができるだろう。

規則功利主義はある種のヒューリスティックアプローチとみなすことができる。規則は個々の行為の帰結のすべてを計算することを回避することを可能にする。問題は、その規則がどこから来るのかということである。最初は専門家が同意する規則をシステムにプログラムしても良い。しかし規則の適用そのものが功利主義的に正当化されなければならないので、規則も定期的に再評価されなければならない。洗練された AMA にはそのような評価を行う能力も要求されるかもしれない。初期の AMAs がそのような能力を持つことはないだろう。規則が専門家によって与えられるとするならば、規則功利主義はある種の戒律理論（commandment theory）として扱われうる。

Rules for Robots

P. 91–（“No discussion of top-down morality for robots.....”）

ロボットのためのトップダウン道徳性の議論はアシモフの三原則を無視するわけにはいかない。それは次のようなものである。

1. ロボットは人間を傷つけてはいけない、あるいは行動を採らないことによって人間に危害が及ぶことを許してはならない。
2. ロボットは人間から与えられた命令に従わなければならない。ただし第一原則に反する場合はその限りではない。
3. ロボットは自分の存在を守らなければならない。ただし第一第二原則に反する場合はその限りではない。

この三原則が確立したずっと後で、アシモフは次の第 0 原則を採用している。

> 第 0 原則：ロボットは人類（humanity）に危害を与えてはいけない。あるいは行動をとらないことによって人類に危害が及ぶことを許してはならない。

この原則は他の三原則のすべての優先される。

アシモフの原則はもちろんフィクションだが、AMAs に関しても興味深いアイディアを含む。それは AMAs の行動は人間の道徳についての通常の規則とは異なる基準に従うべきだ、というアイディアである。

アシモフのアイディアは前節でみた功利主義とはっきりと対照的である。功利主義は、なぜ、あるいは誰によってある行為が行われるか、ということに関心がない。義務論的な見方では、義務は行為者の特定の本性から直接生じている。

アシモフの原則は非常に明快なように思われるが、これらの単純な原則を実装することにさえ問題があるということは、アシモフや他の著者たちにとって明らかだった。例えば外科医が患者の体を切開しようとするとき、融通の利かない（literal-minded）ロボットはそれを邪魔するだろうか？ 採りうるあらゆる行動が人間に何らかの危害を与えるとき、ロボットはどうするべきだろうか？

与えられた規則が包括的でないとき、AMAs は未知の状況において機能不全に陥るだろう。現実世界の文脈では明快なように思われる規則が、遂行不可能であると分かることもある。規則の間に優先順位を付ける図式がなければ、規則同士の衝突が行き詰まり引き起こす。アシモフは原則の間に優先順位を付けているが、しかし一つの規則でさえ行き詰まりを引き起こすこともある。現実世界においては常に危害を避けることができるわけではないということを考えれば、危害を最小限にとどめるのがせいぜい望めることである（人間がこの

ようなロボットと共生することができるかは 12 章で扱われる問題である)。

規則ベースの AMAs は規則の衝突が起きた状況を管理するソフトウェア・アーキテクチャーを必要とする。IT コンサルタントのロジャー・クラークは、アシモフの小説からの教訓を振り返って、規則は道徳的ロボットを設計するための効果的な方法ではない、と結論してる。

包括的で衝突のない規則があったとしても、一つもしくは複数の規則を連続して適用したことで望ましくない結果が生じることもある。コンピュータは特にそのような、局所的には無矛盾だが結果が累積して矛盾を引き起こすような意思決定を犯しやすい。

十分に自律的な道徳的行為者は、倫理的なジレンマに陥った時に、身動きが取れなくなるべきではない、ということには容易に同意できるだろう。しかし AMAs がジレンマを解決するために人間に危害を加えても良い、ということには、誰でも同意できるだろうか。規則が破られなければならないことはある。規則ベースのアプローチで AMAs が規則を破ることを許すのであれば、いつそうするべきかについての極めて明示的な定式化が必要だろう。しかしそのようないかなる規準もまた別なジレンマを引き起こすといことは非常にありそうだ。

しかし義務論的な道徳規則は哲学者にとってだけでなく、道徳について一般人が論じる際にも重要な役割を果たしている。そういった規則は非常に具体的なもの(「汝盗むべからず」)から、非常に抽象的なもの(「他人が己になすことを欲することを他人になすべし」)まで様々である。具体的な規則と抽象的な規則では、その道徳性を計算する際に異なる困難を生じさせる。具体的な規則は単純な場合には適用が比較的容易であるが、より複雑な状況では、規則同士が、あるいは一つの規則でさえそれ自身と、衝突することがあり、はっきりとした指針を与えてくれない。そのような衝突を調整するためにはより抽象的な規則が必要であるように思われる。

Über-Rule Computing

P. 95–(“Kant’s categorical imperative.....”)

カントの定言命法、黄金律はより抽象的な義務論的な理論を代表している。それらは、どのような状況にも当てはまる一般的な原理を述べることで、衝突を回避することを試みる。定言命法は明示的に論理的な無矛盾性を保証するように設計されている。従って論理的な枠組みで機能するコンピューターにとって特に適切であると思われるかもしれない。その鍵となるアイディアは次のように述べられる。「ある格率を通じてあなたがそれを普遍的法則とすることができ、かつあなたがそれを意志する場合にのみ、その格率に基づいて行為しなさい」

この定言命法から AMAs の設計者は何を得ることができるだろうか。ありうる第一の近似は次のようなものだと著者たちは考える。すなわち、いくつかの選択肢を考慮するロ-ボットは、同じような他の行為者が対応する状況で同じように行為したならば、彼らの目標は達成されるのかどうかを調べなければいけない。このように解釈された定言命法は、行動を導く格率の道徳性を調べる形式的な道具として、AMAs によって使われるかもしれない。この道具を適用するためには、AMAs は次の三つの要素からなる、明示的かつ十全に述べられた実践的な推論の原理を必要とするだろう。すなわち目標、目標を達成することを提案するための行動の手段、その行動によって目標が達成されるような状況の言明、である。非常に強力なコンピュータは、これらの三つの要素が与えられれば、他のすべての行為者が同じ格率に従って行為したならば、その目標が阻止されるかどうかを決定するための分析あるいはシミュレーション・モデルを実行することができるかもしれない。

カントの推論を応用する AMAs はまた、人間とロボットの心理学について、そして世界における行為の結

果について、たくさんのことを知る必要があるだろう。

義務ベースのシステムにおいても規則を適用したことの帰結は重要である。多くの規則は、結局のところ、悪い結果を引き起こさないために採用されているのである。

具体的な規則が定言命法のような上位規則（über rules）と無矛盾であるかどうかを見分けることは極めて難しい。結局のところ、義務ベースのアプローチが直面する計算的問題は、帰結主義的システムが直面する問題に近づいていく。

義務論的理論は、それを実装した AMAs が、道徳的判断を必要とする任意の状況で正しく推論ができるのに十分、規則を理解することを要求する。これは規則がすべての状況で曖昧さのない指示を与えてくれればずっと容易だろう。しかしこのことには一見すると乗り越えがたい困難がある。規則を完全に明示的にするには、AMAs はすべての言葉の明確な定義を与えられている必要がある。例えばカントの定言命法の「普遍的な」という言葉の曖昧さを考えれば分かるように、これは決して簡単なことではない。にもかかわらず、曖昧な概念でも明確な応用を持つ場合はある。禿げであるというのは曖昧な概念だが、しかしピカード船長は実際にはげなのだ。最初は明確な場合に集中して、日常の道徳性の多くをトップダウン的に捉えることは可能かもしれない。最終的には、AMAs は倫理的な事例に関してトップダウン的に推論する能力を持つ必要があるだろう。

Top to Bottom

P. 97–（“The limitations of top-down approaches.....”）

著者たちの意見では、トップダウンのアプローチのいくつかの限界を複合的に考えると、AMAs が従うべきトップダウン規則の曖昧でない組み合わせを AMAs に与えることは実現不可能だと言う結論に至る。スーザン・アンダーソンとマイケル・アンダーソンは医療の「専門家」の判断に基づいた無矛盾なシステムを構築する試みを行っている。スーザン・アンダーソンは、医療の専門家は一般的に互いに同意しているという前提のもと、ひとまとまりの無矛盾な原則が現れると信じている。著者たちは AMAs が直面している課題は、人間の道徳的判断の、内在的に曖昧な本質を扱うことを学ぶことである、と考える。

そもそも人間はいかにしてこの曖昧さを扱っているのか。人間は経験から、注意深い行動と観察から、認知と感情と反省の等号から、実践的な知恵を身につけている。そのような知恵の必要性は、AMAs にも感情的能力が必要だということを意味するのだろうか？ そうかもしれない。この点は 10 章で扱われる。しかしその前に、道徳的行動は学習と進化から創発するものと見る、道徳性への「**ボトムアップ**」アプローチの強力さを論じる必要がある。