Wendell Wallach and Colin Allen, *Moral Machine: Teaching Robots Right from Wrong*, Chapter 2　"Engineering Morality" 要旨

久木田水生

2012 年 1 月 14 日

An Engineering Imperative

P. 25〜（"In the Code of Ethics....."）

National Society of Professional Engineers (NSPE) の倫理規則には，第一の「根本的規範 fundamental canon」として，エンジニアは「公衆の安全と健康と福祉を最上のものとすべし」とある．機械に道徳的規準を与えることが公衆の福祉と安全を向上させることになるならば，アメリカのエンジニアたちはその規範によってそうすることを義務付けられていることになる．それは大変な課題であるが，すべての工学的課題は，徐々に乗り越えられるものだ．本章では，現在のテクノロジーから洗練された AMA（artificial moral agents）に至る道筋を理解するための枠組みを与える．その枠組みは二つの次元を持つ．一つは**自律性**であり，もう一つは**価値に対する感受性**である．これらの次元はそれぞれ独立である．

ハンマーのような単純な道具は自律性も感受性も持たない．しかしこの二つの次元において最低レベルのテクノロジーでも，ある種の「**操作的道徳性 operational morality**」を持つことはできる．安全装置のついた銃は，自律性も感受性も持たないが，その設計は NSPE の倫理規則が認める価値を具現化している．

この枠組みでのもう一方の極端には，高い自律性と感受性を持ったシステムがある．そのようなシステムは信頼に足る道徳的行為者として行為することができる．このようなシステムは実現していない．しかし「操作的道徳性」と責任ある道徳的行使者性の間には多くの段階の「**機能的道徳性 functional morality**」がある．

P. 26〜（"The realm of....."）

機能的道徳性の領域は，**顕著な自律性を持つが倫理的感受性をほとんど持たないシステム**，**自律性は低いが倫理的感受性が高いシステム**の両方を含む．

自動運転飛行機は前者の例である．自動運転飛行機のコンピュータは安全性と快適性が一定の範囲に保たれるように機体を制御するが，しかし乗客の価値を直接モニターしているわけではない．

後者の例の一つは倫理的意思決定補助システムである．これは意思決定者に関連する情報へのアクセスを提供する．現在存在するこういったシステムのほとんどは機能的道徳性よりも操作的道徳性の領域に属している．しかし例えば Anderson 夫妻によって開発された **MedEthEx** のように，**医療従事者たちが倫理的に適切な行動のコースを選ぶことを手助けする**ようなプログラムもある．MedEthEx は実際に初歩的な道徳的推論を行っている．このシステムは自律性を持たず，従って十分な AMA ではない．これはさらなる発展のプラットフォームを提供する，ある種の機能的道徳性を持つ．

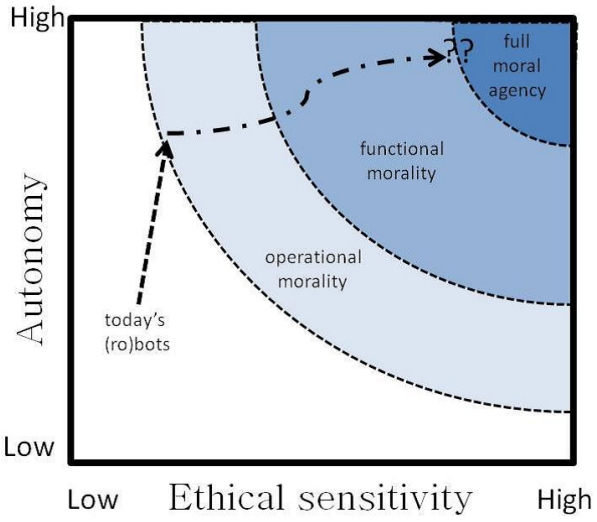

図 1　Two Dimensions of AMA Development.

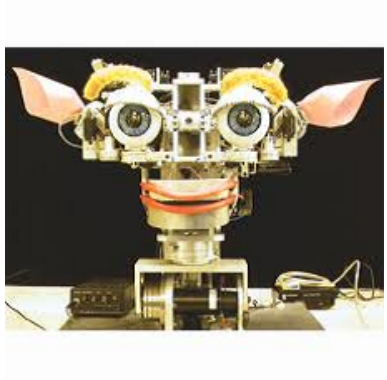

図 2　Kismet.

これらのシステムは図 1 の各々の軸にそって少しの距離しか進んでいない．これらは非常に限られた領域においてのみ機能する．にもかかかわらず，その限られた領域でも倫理的問題は発生する．そしてそのような基礎的な出発の上に機械道徳の工学的側面は構築される．

両方の次元において同時に前進しようとする試みもまた重要である．そのような試みの一つは Rodney Brooks の指導のもとで MIT の院生によって開発された **Kismet** である（図 2）．Kismet は**感情的な反応**と**自律的な活動**を一つのロボットの中で結び付け試みである．Kismet は頭，耳，眉毛，瞼，口を動かすことで怒り，驚き，興味，悲しみなどを含む感情的状態を表現することができる．ロボットが示す感情的状態は話者の声の抑揚や他の要因を分析することで決定される．Kismet は人が指差した方向に視線を向け，そちらに注意を向けることができる．

Kismet は会話のようなものの中で順番に行動をすることができる．人間の話の間に反応をさしはさむ前に，沈黙の空間を待って待機する．Kismet は人が実際に何を考え言っているかを理解することはできないが，しかしそれは声の抑揚などの社会的信号 social cues に非常によく反応することができる．

Kismet の行動は操作的道徳性の例である．というのも**プログラマは信頼や協力を確立するために重要な価値を打ち込んでいる**からである．Kismet は価値の明示的な表象を持たず，価値について推論する能力を持たないが，多くの人は Kismet とのやりとりが非常に説得力がある compelling と感じている．

Kismet は操作的道徳性の領域に留まるものだが，しかし**ロボットが人々の自然で直観的な社会的反応を引き起こすことがいかにして可能か**ということを示したという点で，Kismet は社会的ロボティックスの実験として大きな成功だった．

上記の三種類のロボットは人工道徳性の分野の出発点を提供する．これらのシステムは設計者の価値の比較的直接的な拡張である．自律的意思決定のテクノロジーが進めば，ロ-ボット工学者たちは，彼らの専門家としての規則が AMA の開発を強く要求するということを理解するだろう．

P. 30〜（“Some may wonder.....”）

著者たちの目指していることが不必要に困難だと思う方もいるだろう．価値という明確な定義のない概念を持ち込まずとも，システムの設計はすでに非常に困難である．クレジットカードの不正利用防止の例をあげれば，不正な使用のパターンを認識するソフトウェアを改良することに集中していればよいのではないか？

確かにパターン分析は改良ができるが，それには限界がある．ソフトウェアはまったく正しい使用，時には緊急の使用を「疑わしい」ものとしてブロックする可能性がある．ソフトウェアの開発者と銀行にとってはそれは，許容できる偽陽性の割合と許容できない偽陰性の割合の間のバランスの問題である．銀行にとって許容できる割合とは，利用者に転嫁できるコストがどれだけであるかという問題である．著者たちはここで，**パターン分析によって銀行にとってリスクが高いということが示されていても，信用貸しを認めることが倫理的な行動であるような状況を認識するコンピュータの能力**に関心を持っている．既存のパターン分析へのアプローチは人々の価値を守ることに対して内在的な限界があるので，工学者は彼らの理想によってこの限界を超えることを約束する別なアプローチを追求することを余儀なくされている．

道徳的な議論が企業の利益に優先すると信じているならば，著者たちは二重にナイーブであるように思われるかもしれない．しかし著者たちは**より洗練された機能的道徳性，そして最終的には十全な AMA は企業にとって財政上の利益を与える**と考えている．その一つの理由は，これらのシステムが競争相手よりも良いサービスを提供しうる，ということである．電話の自動応答システムはしばしば顧客を欲求不満にさせる．それはビジネスにとって良くないことである一方，企業はますます機械に意思決定を委ねている．顧客の価値に敏感で，道徳的に良い行為者の意思決定に近い意思決定を下せるソフトウェアは，企業の利益を守るものであり，それを損なうものではない．

2003 年の合衆国東北部での停電は，電気会社が古いテクノロジーに依存しているという事実を強調した．システムのアップグレードは，それら制御システムを以前よりますます自律的にするだろう．システムの複雑化は，人間の直接的なモニタリングをより困難にするだろう．オペレータの過誤と人間による監視の不可能性によって，システムの自律性へのプレッシャーは増大する一方だろう．システムのアップグレードや，グリッドの一部の計画的停止時間すら，コンピュータが様々な要因をリアルタイムに評価することによって，自律的に行われるかもしれない．

こういった考察は，ネットワークが安全性のパラメータの範囲内で機能することを保証する単純な制御システム（操作的道徳性）を超えた，顧客サービスの一次的なレベルと，自己管理の二次的なレベルの両方において，選択肢を評価することができるシステムが必要であることを示している．これらのシステムは，**設計者もプログラマも予想できなかった選択肢や行動を選択する**ような，複雑な状況に対処する必要がある．

P. 32〜（“With these exmaples, ...”）

例にあげたような現存するテクノロジーは，AMAs へ進歩するための様々な出発点を提供するだろう．しかしそれがどのような経路をたどるかは予想するのが難しい．ここでの単純な 2 次元の枠組みでも，現存のテクノロジーから十全な道徳的行為者性への経路は様々でありうる．自律性の次元に沿っての進歩は実際に進行しているし，これからも進行するだろう．人工道徳性の分野での課題は，**道徳的考察に対する感受性の軸に沿った発展がどのようになされるか**という点である．

決定補助システムは，知的システムが自律性とは独立に感受性を高める仕方で発展しうることを例証している．この経路は意思決定を人間の手に委ねつつ，外的な意思決定補助システムを超えて，**人間とテクノロジーのより密接な合併**へと至る見込みが高いように思われる．

多くの理論家はサイボーグは現在の IT，神経補助器具，神経薬理学，ナノテクノロジー，遺伝子治療の当然の結果だと考えている．人間とこれらのテクノロジーの合併は，自律的システムの発展とは異なる問題を提起する．

道徳的感受性を持つ意思決定補助テクノロジーと神経補助器具の発展は，自律的システムの強化に適応されるかもしれない．

Moor’s Categories of Ethical Agents

P. 33（“James Moor, ...”）

James Moor は AMAs を分類するための階層的な図式を提案している．最も低いレベルに位置するのは「**倫理的影響を持つ行為者 ethical impact agents**」である．これは**倫理的な帰結が評価されうる**任意の機械である．実際にはおよそあらゆるロ-ボットは倫理的影響を持つ．

次のレベルにあるのは「**非明示的倫理的行為者 implicit ethical agents**」である．これは設計者が設計の過程において安全性と信頼性の問題関心に目を向けることによって，**否定的な倫理的効果を与えないように設計する**努力を払った機械である．全てのロ-ボットは非明示的倫理的行為者であるべきだろう．

その次のレベルにあるのは「**明示的倫理的行為者 explicit ethical agents**」である．これは，その内的なプログラミングの部分として**倫理的範疇を用いて倫理に関する推論を行う**機械である．これらは義務論理やその他の様々なテクニックを用いるかもしれない．

これらすべての上に十全な倫理的行為者が位置する．**それは明示的な道徳的判断を下すことができて，かつその決定を正当化することにおいて一般的に非常に有能**である．このレベルの遂行能力はしばしば，**意識，意図，自由意思**の能力を要求すると想定される．

多くの哲学者と一部の科学者は，意識，意図，自由意思を持った人工的行為者を作ることができるということを疑い，機械が十全な倫理的行為者になるのは不可能だと論じる．

Moor は明示的倫理的行為者が機械倫理の目標であるべきだと考える．著者たちも，小さなステップを重ねる戦略に同意する．Moor の範疇は彼らの 2 次元のグラフに直接的にマップされるわけではないが，それは機械倫理の直面する課題の範囲を特定するのに有用である．しかしそれは操作的・機能的道徳的行為者を作る**過程**を特定する上ではそれほど有用ではない．

著者たちは，テクノロジーの発展は増大する自律性と増大する感受性の間の相互作用の中に存する，と考える．自律性の増大はすでに進行中の過程である．人工道徳性の分野の課題は感受性の次元に沿った方向へどう進んでいくかである．

ある AMA にとって，道徳的考察への感受性はいくつかのものを意味しうる．便利な区別が McDermott に

よって提供されている．彼は AMAs を設計する際には，**倫理的推論者と倫理的意思決定者の区別**を心にとどめておくことが重要だという．倫理的推論を行うシステムを作る際の困難を解決できたとしても，その行為者は倫理的意思決定者になるにはまだまだ足りないだろう，と McDermott は指摘する．彼は「しかしながら，倫理的意思決定を下す能力は倫理的な葛藤——すなわち自己利益と倫理が命令することの衝突——が何であるかを知ることを要求する」と書いている．

工学における現実的な課題を追求するためには，**成功についての明確な基準を持つこと**が最良である．**道徳的感受性や道徳的行為者性の基準**はどのように作れるだろう？ Alan Turing はコンピュータが知的であるかどうかを決定しようと試みて，同様の問題に直面した．Turing は知性を定義する問題を，実際的な試験を応用することによって迂回した．それは，テキストのみのやり取りの中での会話的な追う乙に基づいて機械と人間とを区別できるかどうかのテストである．Turing によれば専門家にもコンピュータと人間の違いが分からなければ，そのコンピュータはあらゆる実践的な目的にとって知的であるとみなされるべきである．この基準は欠点もあるが，それにもかかわらず，工学者が知的なシステムを作るための明確な目標を提示している．

有用な**道徳的チューリング試験**は作れるだろうか？ このことは後の章で議論される．当面重要なことは，**道徳的意思決定のある側面を AI に実装することを目指すプロジェクトは，その成功を判断するための基準の特定を必要とする**，ということである．異なる基準は異なる特徴——論理的な整合性，言語，感情的知性など——の強調を帰結として持つだろう．

AMAs の構築と評価の詳細に入る前に，著者たちがこの仕事を提示する際にしばしば遭遇するニ種類の懸念に目を向ける．一つは道徳的意思決定の機械化を試みることが人間にどのような帰結をもたらすかということ，そしてもう一つは機械を知的な行為者にしようとすることは錬金術のような間違った試みなのではないかということである．