# Wendell Wallach and Colin Allen, *Moral Machine: Teaching Robots Right from Wrong*, Chapter 4 "Can (ro)bots really be moral?" 要旨

久木田水生

2012 年 3 月 16 日

## Careworthy Technology

**P. 55**–(""Soldiers Bond with Battlefield Robots" declared.....")

兵士たちは戦場で使われるロボットに絆を感じ，それらに対して道徳的配慮をするようになっている．その逆はありうるのだろうか？ ロボットが人間に配慮することは可能か？

多くの人は，人間の最も重要な関係を規定し，人間の倫理的規範を形作るのは，**理解 understanding と感情 emotions** であり，ロボットが本当の意味でこれらを持つことはない，と考える．これらの能力は何なのだろうか？ どのようにしてそれらについて科学的な知識が得られるのか？誰もこの質問に答えることはできないし，人工道徳がこれの質問への答えに依存することもない．しかしこれらの質問に答えることができない理由を考えることは，実際的な問題に対するアプローチを形作る役に立つ．

この章では，人工知能の限界についての論争が人工道徳の実践的な目的に影響を与えない，ということを論じる．それどころか，**AMAs を作る実践的な課題に取り組むことは，倫理そのものの本性についての存在論的・認識論的疑問をよりよく理解するのに貢献する**と，著者たちは考える．

## Artificial Intelligence: The Very Idea

**P. 56**–("The mysteries of matter,.....")

精神についての科学的研究には主に二つのアプローチがある．一つは脳についての生物学的な研究によるアプローチである．もう一つは**情報処理**という精神の中心的な性格をより一般的に追及するアプローチである．初期の AI の提唱者たちは，この文脈において精神は科学という安全な足場に立脚させることができる，と考えた．カーネギーメロン大学の Allen Newell と Herbart Simon は「物理的な記号体系は，一般的な知的行為にとって必要十分な方法を持っている」と宣言した．彼らが予見する認知心理学は，人間知性に含まれる記号操作を明らかにすることによって，計算がいかに知性にとって必要であるかを示す科学であった．また AI の課題は，いかなる知的能力もコンピュータにプログラムできることを確証することで,，記号計算の十分性を示すことであった．

現在のコンピュータはまだ人間の脳に比べてはるかに単純である．しかしテクノロジーは指数関数的に進歩している．Ray Kurzweil は，2020 年には人間の脳に匹敵するデスクトップ・コンピュータが入手可能になる

だろう，と予測する．2029 年ごろには，機械の知性が，すべての人間の知性を合わせたものを凌駕し，そのとき人類は**特異点 Singularity** に突入する，と論じる．特異点とは，変化があまりにも過激すぎて，もはや予測不可能になる時点を特徴づける．

Kurzweil は John Searle が「**強い AI**」と呼ぶものの推進者である．これは適切にプログラムされたコンピュータは精神になる，という見解である．1980 年に Searle は，「**中国語の部屋**」の思考実験によって，**プログラムに従うコンピュータによって実行された形式的な記号の操作は，決して知的な理解を生み出さない**，と論じた．

Searle の議論は，著者たちの人工道徳に対するアプローチもまた同様に望みがないことを示している，と論じる懐疑主義者もいる．しかし著者たちは，哲学的な反論によって，倫理的意志決定に対するよりよい計算的解答を得る試みをやめることはない，と考える．しかしながら，著者たちが予見しているシステムの本性や地位に関して哲学的な問題がある，ということは認識しなければならない．

本当の道徳的行為者に要求されるものは何か？ ある人は**意識**を強調し，ある人は**自由意志**を重視し，ある人は**道徳的責任**を重視する．

## Could a (Ro)bot Be a Real Moral Agent?

**P. 58**–（“The discussion about “real” moral agency.....”）

「現実の」道徳的行為者についての議論は，それが AMAs の中に作りつける必要のある能力を示唆してくれるのであれば，有益である．

Searle の議論において，ただの記号操作の体系と「真の理解 genuine understanding」は，行動的な違いがないにも関わらず，区別される．Searle の議論では，現実の道徳的行為者と行動的に区別のできない AMAs の可能性は排除されない．したがって彼の意識，意図的理解についての概念は，いかにしてロ-ボットに倫理的に行動させるかという実践的な問題にとってはまったく無関係である．

René Descartes の形而上学にとって，機械の精神という考えは馬鹿げている．彼は精神と物質をまったく異なる実体と考え，物質に知性的属性が宿ることはあり得ない，と論じた．Descartes が正しければ AMAs の見通しはやや暗くなる．

しかし Descartes は，このことに対してきちんとした議論はしていない．物質的対象の能力についての彼の理解は 17 世紀の科学によって限定されている．しかし彼の同時代人，Thomas Hobbes は「精神は有機的身体の特定の部分の動きに過ぎないだろう」と信じていた．また今日では二元論の擁護としては，Descartes の見解はそれほどの重要性はない．しかし現代でもプログラムされたシリコンによっては達成できない能力を生み出す特別なものが人間の脳にはある，と主張する人々はいる．この見解を証明あるいは反証することは，現代の科学でもいまだに不可能である．

現在のコンピュータ技術によって開発されているそのような特別な性質の一つは自由意志，もう一つは意識的な理解である．

## The Ethics of Deterministic Systems

**P. 59**–（“Unless suffering from compulsion.....”）

人間が行動をとる際に感じる自由は何に由来するのだろうか？ そしてそれは倫理的行為にとって必要なのであろうか？

人間の自由意志はしばしば，科学的によっては定義できないものとして，いくぶん神秘的に受け取られている．Daniel Dennet はこのような「魔術的な」自由意志概念を拒否する．彼は，**複数の選択肢を考慮し，そのうちの一つを選択する能力**だけが人間の持つ自由意志だとする．自由意志とはそのようなものであるとしよう．人間の自由意志には魔術的な要素があるという議論は，AMAs を作ることに対して曖昧な反論しか与えないし，またそのような要素は AMAs を作るという工学的課題には応用できるものでもない．

このことは AMAs プロジェクトの価値を減じない．Gary Kasparov を破った Deep Blue の能力にとって，魔術的要素の欠如は問題にならない．Deep Blue は決定論的システムであるが，しかしそれはある面では行為者の資格を持つ．Luciano Floridi と J. W. Sanders は人工的行為者にとって重要な三つの中心的特徴を次のように特定した：

- **相互作用性**：刺激に対して状態を変化させることで反応する．つまり行為者と環境はお互いに作用しあう．
- **自律性**：刺激なしに状態を変える能力．これによって，ある程度システムが複雑になり，また環境からある程度切り離される．
- **適応性**：状態を変化させる「遷移規則」を変化させる能力．すなわち，行為者は経験に強く依存しながら，作動する仕方を学習するものとみなされる．

Deep Blue はある程度，相互作用的・自律的であるが，適応性は欠いている．学習的な要素をシステムに加えれば，適応性は高まるだろうが，今日利用できる学習アルゴリズムは十分というには程遠い．

**P. 60**–（“Chess isn’t ethics, of course.....”）

道徳的行為者としての人間の経験にとって中心的な特徴は，**利己的行動と利他的行動**に引き裂かれる感覚である．このテンションが自由の可能性を開く．

決定論的システムに対して，どのように倫理が生じるだろう．著者たちはサイバネティシスト，Heinz von Foerster によって示唆されるような選択のうちにその可能性を見出す．現存するロ-ボットは倫理的規則を受動的に伝導するのみではなく，それ自体が現存する**道徳的エコロジーの中で他の行為者と相互作用する**．爆弾探知ロボットに対する兵士の懸念は，たとえば彼がロボットの生存を犬などの生存と比較してどのくらい重視するかという新しい倫理の可能性を示す．しかしそういった可能性は倫理的行為者の設計にも影響を与える．例えば兵士がロボットに近づいて相互作用する傾向を測定して，それを利用して彼がロボットを保護するように行為するかを見積もることは現在のテクノロジーの範囲に収まることである．特定の任務に関しては，ロボットに対してある態度をもって接する人を選んで繋がりを持つ必要があるかもしれない．このようにして人間とロ-ボットの関係は相互的でありうる．この相互的な構造の中で行動する行為者は何であれ，自分の目標と他人の目標の間の葛藤に直面するだろう．ある目標の追求が他人への危害を含む時，倫理的な問題が生じる．**より多くの選択肢がシステムに与えられ，システムがそれらを評価するとき，葛藤が生じる可能性がより高くなる**．

著者たちは，より洗練された道徳的行為者とは**異なる視点が異なる選好の順位を生み出すことを認識している行為者**である，と提案する．AMAs の設計は，倫理学の領域に存在する自由の程度を許容しなければならない．

Von Foerster にとって，単に何を選ぶべきかということだけが問題なのではなく，**可能な選択肢を拡大すること**も倫理学にとって最も重要である．Dennett によれば，選択肢の拡大は「自由の進化」によってもたらされる．この言葉で彼が意味しているのは，進化によって人間は複数の選択肢を考慮し，複数の結果を予見す

る能力を得た，ということである．

選択肢の拡大は倫理的機械の進化と発展にとっても重要である．Chris Lang は，探索ベースの学習コンピュータにとって選択肢を拡大することは，人間に友好的な道徳的行為者として行為する可能性の高いシステムを生み出すことになるだろう，と提案する．この楽観的見解は，機械学習による「戦略の合理的探索は，新しいアイディアに遭遇する割合を最大化することを意味し，そしてそれはさらに**そのシステムが参加している集団の多様性と相互作用の割合を最大化する**ことを意味する」という理解に基づいている．Lang によれば，このアプローチは「世界に中の自由を最大化すること」を意味するのであり，それは通常「生命の保持と民衆に権力を与えること」を含む．

彼も von Foerster のように，**選択を最大化することが道徳的行為者性にとっての鍵だ**と考えているということは注目するべきだ．

決定論的システムが**本当に**道徳的行為者になれるかどうかは答えの出る問題ではない．しかし行為のための決定論的な枠組みの中でも，倫理はオープン・エンドの選択を含む．道徳的環境に新しい行為者が入ってくると，選択肢は拡大し，帰結は増大する．人間的な道徳的文脈でよいパフォーマンスをあげるためには，AMAs は複数の選択肢を評価し，異なる評価の視点を考慮する必要があるだろう．AMAs は行動することによって，現存する道徳的エコロジーにフィードバックを与え，そしてそれを変化させる．しかし人間と AMAs との結びつきが彼らのの実際の道徳的能力を凌駕するとしても，洗練された AMAs が道徳的エコロジーを嘆かわしい方向に変えることにならないことが望まれる．これはひょっとしたら誤った希望かもしれない．それでも道徳的自由は，決定論的システムと両立可能であるか否かに関わらず，AMAs の設計にとって重要である．

## Understanding and Consciousness

**P. 63**–（“Like free will, .....”）

どのような種類の理解をロ-ボットは持てるだろうか？ それは AMAs を作るのに十分なのだろうか？ AMAs には意識が必要なのだろうか？ 意識を持たないシステムが道徳的行為者とみなされうるのだろうか？ 法的にも哲学的にも，道徳的行為者性は道徳的責任と同等のものとして扱われ，自分がしていることを理解せず意識していない個人に帰せられることはない．人工道徳について聞いた人が真っ先に飛びつくのがロ-ボットの権利と責任の問題である．しかし著者たちの関心は AMAs を作ることに成功した後の下流部門の問題ではなく，上流部門での，システムの倫理的判断を下す能力における意識と理解の役割にある．

### *Understanding*

**P. 63**–（“Searle’s Chinese room is.....”）

Searle の中国語の部屋の議論以来，AI 研究者のほとんどはチューリング・テストをパスする会話の能力だけに焦点を当てるのではなく，ロボット工学に対する「マルチモーダル」なアプローチをとる．そのようなシステムは聴覚，視覚，触覚を同時に処理し，行為と発話を同時に学習する，そのようなシステムが学習する語は，このようにして，ロボット自身の行為と，観察される他の行為者の行為において「接地させられ」る．このとき真の人間的な理解とロボットの理解の間のギャップはそれほど重要ではなくなる，と多くの研究者は考える．

Searle が想定したような世界から切り離され，身体を持たない記号操作システムでは真の言語理解は実質的に不可能である．現実の認知システムは，物理的に身体化され physically embodied，物理的対象と社会的行

為者からなる世界の中に置かれている situated in the world of physical objects and social agents．このようなシステムによって使用される語，概念，記号は，そのシステムと対象，そして他の行為者との相互作用の中で接地させられる．

Rodney Brooks の仕事に端を発する**身体化された認知の理論**は，科学的ロボットと商業的ロボット双方のの発展に革命を起こした．Brooks のロボットはすべての行動を調整する中枢のプロセッサーを利用するのではなく，各部分が独立に制御される設計になっている．このようにして彼は，押されることに順応でき，様々な地形の中で動くことができる，安定したロボットを作り出した．ロボットの**物理的な身体化**と**環境への埋め込み**について真剣に取り組むことで，Brooks は**一連の比較的単純な局所的過程が全体としてより複雑な行動を創発する**ことを示した．Brooks は異なる行動的能力を階層的に組み合わせることでより洗練されたロボット・システムを作り上げることを提案している．彼は，**システム全体に指令を発する中枢制御機器なしで，調和的な行動が達成できる**ことを示した．

Brooks は彼のアプローチを「包摂アーキテクチャ subsumption architecture」あるいは「行動ベース・ロボット工学 behabior-based robotics」と呼ぶ．環境からの合図に反応して基本的な行動を遂行するようにロボットを作る，というのがその発想である．ここでは，ある時点で包摂的階層のどの部分がロボットの活動を制御するかを決定する上で，環境が中心的な役割を果たす．包摂アーキテクチャの可能性は，単純で低レベルの仕事を遂行するサブシステムの間の相互作用から順応的行動がいかにして創発するかという点にある．言い換えると，複雑な動物や人間において，特定の課題を遂行する比較的単純な構成要素が全体として複雑な行動能力と高レベルの認知的機能を持つように見える．

**身体化された認知の理論**は，世界の中でいかに行動するかについて推論するために必要なすべての詳細を含んだ完全な世界の内的表象，モデルあるいはシミュレーションを脳が作り出している，という見解に対するオルタナティブとして登場した．認知に対する古典的なアプローチでは，いちいちの行動を決定するために，世界のモデルを構成する内的な記号を脳が操作する．しかしこのアプローチに基づいて設計されたロボットは非常に脆弱であり，例えば予期せず押されると，それに反応して内的表象を更新する前に倒れてしまう．自らが作り出した仮想現実に生きているロボットは，現実そのものに即座に連続的に反応するロボットにはかなわない．

AMAs の設計にとっては，身体化された認知と世界の内的モデルとの間の関係についてもっと理解されなければならない．認知の身体化と世界への埋め込みの重要性を理解することは，二つの重要な洞察を与える．第一に，行為者が必要とする情報の多くはすでに環境の中に備え付けられているのかもしれない．第二に，何らかの理解を伴うように思われる，物理的・社会的環境に対する人々の反応が，彼らの身体や感覚の構造と設計に多くを負っており，それによって彼らは意識的な思考や反省をほとんど（あるいはまったく）必要とせずにほとんどの反応を生み出せるのかもしれない．

判断や理解のどの道徳的側面が身体化と埋め込みに依存するのだろうか．人間にとっては道徳的行動の多くは，関係している人々の変化する必要，価値，期待に合うように，リアルタイムで社会的状況に適合することに関係している．AMAs も同様に，彼らの関係の中に置かれなければならないだろう．その関係は常に進化していくだろう．AMAs 自身が多くの活動の領域の中で，新しい課題を乗り越える過程の積極的な参加者になるだろう．

ではこの議論の中で「理解」は何を意味するのか？もしそれが社会的・物理的環境の中で適切かつ順応的に反応する能力を意味するのであれば，私たちは適切に身体化され埋め込まれたコンピュータがそのような反応を示さないと考える理由はない．すでにすべての感覚的モダリティを使ってシステムを利用できるようにする，“enactive” 人間とコンピュータのインターフェースを開発している工学者もいる．このようなシステムが

洗練されるにつれ，すべての理解がシステムのデジタル・プログラムの部分にのみ位置しているのか，という問いはますますどうでもよくなる．

## *Consciousness*

**P. 66**–（“Understanding is sometimes equated with.....”）

理解は時に意識を同列に論じられる．そして意識もまた魔術的な含みを持つ言葉である．言葉は覚醒と睡眠の間の違いを特徴づけるために，あるいは様々な高次の認知機能を捉えるために使われる．

人間とは異なる種類の経験は認識論的な問題を提起する．人間は蝙蝠であるということがどういうことかを推測するしかない．同様にコンピュータが何を感じるかは人間の知りえないことである．

意識しているという神秘的な経験を人間精神の非物理的側面（魂，霊魂，超自然的実体）に帰す人々もいる．意識は物質の普遍的な性質だ，と考える人もいる．意識は情報処理，ニューロンのネットワーク組織，あるいは神経系の基本的な神経生理学的性質という観点から理解されなければならない，と考える人々もいる．その中間には，精神が客観的な情報処理あるいは神経の観点で説明できるかどうかはともかく，それは脳の観察可能あるいは測定可能な性質と密接に相関している，と考える人々がいる．晩年の Francis Crick は同僚の Cristof Koch とともに，意識に対応する神経的相関物を探求した．哲学者はここから異なる教訓を引き出した．Patricia Churchland は，意識を構成する個々のシステムを科学が理解するようになるにつれて，意識を理解することの問題はなくなっていくだろう，と主張する．他にも，David Chalmers と Colin McGinn のように，意識と脳の相関を発見することは科学的には価値があるが，それは意識的経験の現象学的側面を説明しない，と主張するものもいる．

意識を理解するためにニューロンを探るのは正しいのだろうか？ Searle のように，他にもっと良い場所はないと論じる人もいるだろう．少なくとも人間の場合，ニューロンが意識を生み出していることは知られている．しかし計算によって人工的な意識を生み出そうと試みることは，人間による飛行の初期の試みのようなものであるかもしれない．人間が飛行するのに鳥をまねるのが最良の方法ではない．飛行とは機能的な性質である．それは様々な異なる方法で実現されうる．意識についても，その重要な性質は**機能的に理解するのが最良である**．コンピュータは人間と同じ仕方で意識を持たないかもしれないが，同様の能力を持っているかのように機能するように設計されうるかもしれない．

機械の意識は AI における一部門として発展している．Aleksander は意識の要件は 5 つの領域をカバーする公理に分割されると提案する．それは，自己の感覚，想像，注意の集中，将来の計画，そして感情である．Owen Holland と Rod Goodman は身体化されたロボットに少しずつスキルを加えていく，ボトムアップ的なアプローチをとる．彼らは，このプロセスがやがてロボットの内的表象とそれ自身の行動を生じさせ，それが意識的現象を引き起こすだろう，と信じている．Stan Franklin は，**人間が意識を持つことで可能になっている課題の多くを，システムのアーキテクチャとメカニズムによって人工的行為者が遂行できているならば，その行為者は機能的に意識を持っているのだ**，と提案した．

機械の意識の分野が成功するか否かは時間だけが教えてくれる．一部の哲学者は，現象的意識は機能的同等性以上のものがあり，人間の意識と関連付けられる課題を遂行するコンピュータの成功によっては満足されない，と主張するだろう．しかし観察可能な行動にとって何の違いも生み出さない何かとしての意識は AMAs の開発にとってはどうでもよい．**AMAs の設計に関する実際的な問題にとって問題となりうるのは，行動の機能的同等性のみである**．

## What AMAs Still Can't Do

**P. 68**–(“Armchair arguments.....”)

後に見るように，現在の AI，人工生命，そしてロボット工学の水準は，AMAs の設計における興味深い実験を始めるのに十分である．

しばらくの間は，コンピュータが理解できること，意識できることは，人間よりも限定的であるだろうと考えた方が安全である．そしてこの限界はニュアンスを感じ取り，敏感な判断を下す能力に影響を与えるだろう．研究するべき問題は，**人間のような理解と意識を欠いたシステムにはアクセスできない，道徳的に重要な情報があるかどうか**である．

人間の理解や意識は，特定の課題に対する解決として，進化によって創発した．それは必ずしもその課題に対処するための唯一の方法ではない．しかしこれらの問題は非常に洗練された AMAs の開発が容易ではないことを示唆する．

## Assessing AMAs

**P. 69**–(“Engineering, more than philosophy, .....”)

課題を明確に特定することが工学には必要である．しかし AMAs の課題は何であろうか？様々な行為の道徳性について，人々は同意してしない．正しい理論的アプローチは何かということについて，倫理学者は同意していない．

機械の知性に対するチューリング・テストは，哲学的問題に対する工学者の解決だった．人工道徳の分野において，道徳的チューリング・テスト（MTT）が同様の役割を果たせるだろうか？ Colin Allen，Gary Varner，Jason Zinser はこの問題を考察して，批判的なコメントをした．ここでそれを見直そう．もともとのチューリング・テストと同様に，いかなる MTT も完全な評価ツールではない．しかしそのようなテストの限界についてよく考えることは，AMAs の評価にとって何が重要である可能性があるかということを明らかにする役に立つ．

MTT を使うことの一つの利点は，特定の倫理的問題についての不和を迂回するということである．機械の結論が質問者と違っても，それに対する適切な理由を挙げることができれば，質問者はそれを道徳的行為者とみなすかもしれない．

しかしこのように道徳的推論と正当化に焦点を当てることは適切でないかもしれない．正当化の重要性に関して，倫理理論同士の食い違いがある．カントの見解では推論の過程が行為の道徳性の本質的な要素である．アリストテレスの徳理論では，良い性質に起因する習慣の結果としての正しい行為が重要視される．功利主義では，行為の道徳性はその動機とは独立であり，行為の原因や正当化よりも行為の結果が重視される．多くの人は子供や犬なども道徳的行為者であると主張することで，カントの見解を拒否するかもしれない．

これらの違いに答えて，Allen と同僚は別な MTT についても考慮した．その MTT では「質問者」はひと組の実際の道徳的に重大な行為の記述や例を見せられる．質問者の役割はロボットを特定することである．人間が倫理的でなく振る舞うことで機械と区別されるならば，このアプローチは問題がある．これは「どちらの行為者が AMAs かわかりますか？」と尋ねる代わりに，「どちらの行為者がより不道徳ですか？」と尋ねることで対処できる．Allen らはこれを比較的 MTT（cMTT）と呼び，そして AMAs にとっての成功は，常により道徳的であると判断されることだ，と示唆した．

しかしこの基準は低すぎるかもしれない．倫理とは，人々がすることではなく，人々がするべきことに関わ

るものである．したがって実際の人間との比較は不適切であるかもしれない．加えて，機械のある行為の遂行全体が，部分的に道徳的に誤った行為を含むことをしても，全体として人間の行為よりも勝っていれば，機械は cMTT に合格するかもしれない．しかし人々は機械のそのような過ちを許容しないかもしれない．人は人間の道徳的な過ちには寛容だが，機械にはより不寛容である可能性が高い．

人間と機械の根本的な違いを認識するのは重要である．人間は生化学的プラットフォームから進化した．その推論能力は感情的な脳から創発した．一方，AI は現在のところ，論理的なプラットフォームの上で開発されている．これは道徳的な課題に反応することに関して，コンピュータが人間に対して持つ優位を示唆するかもしれない．コンピュータは人間よりも多様な可能性を計算できるかもしれない．人間によって下される決定は完全に合理的なものではなく，人間は少数の選択肢しか考慮せず，そして良いと思う最初の選択肢に落ち着いてしまう．さらにコンピュータによる道徳的な決定は感情によって邪魔されることがない．もっともコンピュータに感情的なメカニズムを導入すればこの限りではない．このようなメカニズムが AMAs の設計にとって有益であるかもしれないということは 10 章で検討される．7 章で論じる進化的なボトムアップアプローチからは，欲のようなものが創発するかもしれない．しかしそのような貪欲なコンピュータは，それらは名声や権力やセックスよりも，エネルギーや情報を欲する可能性が高いだろう．

これらの要因が，コンピュータが人間よりも高い水準を満たす能力を持つことを意味するのであれば，そのような水準はどこから来るのか？道徳についての諸理論は一つの答えに同意しない．またそれらは簡単にアルゴリズムに翻訳されることもない．しかし著者たちは理論を実践に翻訳することはロ-ボット工学者にとっても倫理学者にとっても有用だと信じている．