Wendell Wallach and Colin Allen, *Moral Machine: Teaching Robots Right from Wrong*, Chapter 3 "Does humanity want computers making moral decisions?" 要旨

久木田水生

2012 年 2 月 8 日

Fear and Fascination

P. 37〜（"We've informally polled....."）

著者たちが非公式に AMAs（artificial moral agents）が望ましいかどうかについてアンケートを取ったところ、結果は二つに分かれた。多くの人々は AMAs が必要かつ不可避と考える一方、他の人々は AMAs は先進テクノロジーについての不安感を一層強めると言う。

洗練されたテクノロジーによって作り出される不安を、なお一層洗練されたテクノロジーによって取り除こうという考えにはパラドキシカルなところがある。テクノロジーの魅力とそれが引き起こす不安の間には何らかのテンションがある。この不安には二つの源泉がある。一つは良くある、テクノロジーの進歩が人間の手に負えなくなるという未来派の危惧である。もう一つは**テクノロジーが人間自身について何を明らかにするのだろうか**という心配である。

人間とテクノロジーとの深い関係はときどき「道具の作り手としての人間」という言葉によって表される。原始人が石をとってそれを道具や武器に作り替えたとき、人間とテクノロジーの共進化の歴史が始まった。今日の子供はコンピュータなどのテクノロジーない世界を想像することはできない。

人間にとってのテクノロジーの重要性が、テクノロジーの哲学の主要なテーマである。哲学者たちはテクノロジーが人間の文化において果たしてきた役割に——その結果として生じるコストと利益も含めて——注目する。人間は器具や機械なしでは生きていけないので、ある程度人間を定義する道具はまた、**人間の生活を制御し、そして人間の自律性を低めている**ものとみなすことができる。

P. 38〜（"Two notions of value....."）

ここには二つの価値の概念が関係している。一方では**外的価値**がある。これはテクノロジーが公共の福祉にとって役立つか否かということに関係している。他方では**内的価値**がある。これはテクノロジーがいかにして人間であることの意味を形作るか、ということに関係している。Andy Clark が言うように私たちは「生まれながらのサイボーグ」であり、それらと文字通り合併してしまうほど易々とテクノロジーを積み込んでいるのか、それともテクノロジーはズボンのように簡単に履き替えることができるものなのか？ **人間の自律とテク**

ノロジーに対する依存の間には哲学的なテンションがある。

新しいテクノロジーはまた人間の能力のみならず、**人間の性格や意識**を変えてしまうという側面もある。

テクノロジーの哲学は、テクノロジーの発達した社会おける**人間の自由や尊厳**についての問題を提起してきた。高度に工業化した社会では人間は単調な仕事ばかりをさせられるようになるのだろうか？新しいテクノロジーの発達は、人間が制御できないかもしれない過程が進行するという危惧を引き起こす。テクノ哲学者たちの多くは、テクノロジーに対する楽観主義者に対する反対の重りとして、テクノロジーの進歩に対して批判的な態度をとる。

古いタイプのテクノロジーの哲学は大体において反応的 reactive であるが、新しいテクノロジーの哲学者たちはより**先向行動的 proactive** である。彼らは、設計の過程に工学者が持ちこむ価値について自覚的であるように工学者を促そうとし、そしてテクノロジーの設計と実装に反応するだけでなく、それらに影響を与えたいと考えている。哲学者の Helen Nissenbaum はこれを「工学的能動主義 engineering activism」と呼ぶ。

工学者の中には価値に関する問題を「ソフト」過ぎると敬遠する向きもあるが、彼女は工学的能動主義を、人類に役立つ「価値の側に立って擁護する」必要性として正当化する。

人工道徳の分野はこの能動主義のアプローチを共有する。それは根本的に**人類の福祉を高める価値をテクノロジーに組み込むこと**に関心を持つ。工学者たちも無意識のうちに彼らの価値を道具の設計に組み込んではいるのだが、工学者たちは近年まで、その価値をどのようにテクノロジーに非明示的に埋め込むかということを考慮しなかった。工学者が彼らの仕事の内的そして外的な倫理的次元に注意を払うようになったのは Nissenbaum のような哲学者の重要な功績である。

人工的行為者の行為の中に非明示的に含まれる道徳性は、単に工学倫理の問題ではない。現代のコンピュータの複雑さを考えると、工学者はシステムが新しい状況でどう行動するかを予測できない。一つの機械の設計には何百もの工学者が関わっている。異なる企業、研究所、設計チームが、最終的な製品を構成する個々のハードウェアやソフトウェアに取り組んでいる。コンピュータ・システムのモジュラー化された設計は、システムが新しい入力の複雑な流れとどのように相互作用し、どのようにそれに反応しているかを完全に把握している一人の人間、一つのグループはいない、ということを意味する。人工道徳の目標は、**操作的道徳性を形作る際の設計者の価値の役割を強調することから、システム自体に明示的な道徳的推論と意思決定の能力を持たせることへ**と、工学的能動主義を拡張することである。

理想的な AMAs は、選択し行為する際に外的価値と内的価値の両方を考慮するようなものであるが、当面はロ-ボットが外的な危害を加えないことを保証することが強調されるだろう。内的価値へ注意を払うのは主として、システムを設計する工学者と、新しいテクノロジーを採用するか拒否するかを選ぶ社会とユーザであるだろう。

Delegating Responsibility for Decisions to a Computer

P. 40〜（“We have suggested.....”）

AMAs を開発することの負の側面はあるだろうか。人間の尊厳と責任に対する影響に関する心配が、比較的限定された形の人工知能システムに関してもただちに生じる。

機械道徳への最初のアプローチは、意思決定者を支援するソフトウェアという形になるだろう。しかしそのような支援ツールは、自分の批判的思考に代わって機械の出力を利用するユーザの松葉づえになる危険がある。社会科学者の Batya Friedman と Peter Kahn は決定支援ツール（DSTs）に関するこの懸念を提起している。

DSTs によって、人間の意思決定者が道徳的責任を放棄し、人間が DSTs に制御されることになる可能性がある、と彼らは考えている。

しかしなぜこれが悪いことなのだろうか？ Friedman と Kahn はこの点について明確ではない。彼らは責任ある計算技術は、十分に意識的な行為者がすべての決定に責任を持つことを必要とする、と考えているようである。確かに生死に関わるような重大な決定の場合はそうかもしれない。しかし人間による直接的な監視が実際上困難な領域がたくさんある。このような文脈では、責任ある計算技術は、考慮されている問題の倫理的に重要な特徴を考慮し、そしてそれに対して反応するプログラムを持つことを意味する。

彼らは ICU での DST の採用を例に挙げて彼らの懸念を説明する。彼らが注目するのは APACHE という ICU での治療に利用される決定支援システムである。彼らは救命医療のスタッフが APACHE の推奨に自動的に従って行動するようになり、熟達した医師でさえも APACHE の権威に逆らうことが難しくなるかもしれないかもしれない、という。

APACHE があることは、医師の自律性の低下を導くのだろうか？ この質問に確固とした答を与えるのは難しい。しかしもし APACHE の助けを借りた方が、借りないよりもよい治療ができるのであれば、医師の自律性が減じることがそれほど悪いことだろうか？ Friedman と Kahn の重視する危険を認識することは重要ではあるが、このような心配は思弁よりも経験的な研究に基づくべきである。

Friedman と Kahn はまたいつか APACHE が人間の意思決定者の直接的な行動なしに、患者の生命維持装置を停止するのに利用されるかもしれないという見通しを検討している。この考察は警戒を要するが、しかし彼らがこの懸念を提起してから 15 年ほどたったが、生死にかかわる決定を機械が完全に制御するような可能性は少しも現実に近づいてはいない。このような大きな一歩が踏み出される前には、利用できるソフトウェアは現在想像されているソフトウェアの倫理的感受性をはるかにしのぐ洗練のレベルに達する必要があるだろう。より倫理に重点をおいたシステムは患者の生存の可能性のみならず、治療が患者や家族の願望に反していないか、治療の予測される結果が認容できる QOL を提供するかなどの問題を考慮することが期待される。これらの問題は通常、十全かつオープンな患者と医師などとの対話のなかで最もよく答えられる。

しかし場合によっては患者が対話できる状態にないこともある。このような末期的な状態の患者の選好を、コンピュータは患者の家族と同じくらい正しく予想できるだろうか。ある調査では、その治療によって過去に 1 %の患者が通常の意識を取り戻すことができている場合には患者はその治療を望む、という単純なルールに基づいて患者の選好を予測するプログラムは、患者の家族や友人と全く同じくらい正確に患者の願望を予測することができた。そのほかの情報を利用することで、ソフトウェアが人間よりも顕著に良い成績を上げることも可能だろう。だとすると遺書を書かずに意思表明ができない状態に陥った時、人は集中治療についての判断を家族よりも機械に下してもらうことを好むかもしれない。

患者が完全に意思表示できない状態でないときには、DSTs を乱用しないように注意しなければならない。文脈に関わりなく、意思決定支援はデフォールトで意思決定になるべきではない。しかしこれはより AMAs に近い DSTs を作ることを妨げるものではない。DSTs に倫理的な判断を下す能力を持たせることは、そうでない DSTs が間違った判定をする可能性をより低くするかもしれない。

Pulling the Wool

P. 43〜（“Two notions of value.....”）

1944 年に Fritz Heider と Mary-Ann Simmel は実験によって、人間がいかに自然に、生きているように動く物に擬人的な性質を帰属させるかを示した。

人間的なスキルをテクノロジーに持たせることは、人間とコンピュータなどとのインタラクションを容易にするというかなりの証拠もある。

MIT の Affective Computing Lab はシステムがユーザのいらだちを認識して、それに対応できるような方法を実験している。また MIT の Humanoid Robotics Group は人間の基本的な社会的ジェスチャーを読み取り、人間的な社会的合図で答えるように設計されたシステムを研究するグループもある。非常にレベルの低い機械的な社会的メカニズムでも、それが実際に生きていて、実際に社会的インタラクションに従事しているのだと説得的に感じさせることができる。

こういった人間的な特徴や動きがテクノロジーとのインタラクションをより容易に快適にするのは確かだが、テクノロジーがどこまで人間的になれるか、そしてなるべきかという点については不確実なところがある。日本のロボット工学者、森政弘は、ロボットが人間的な特徴や動きを身につけるにつれて人々はより共感的・快適に感じるが、ロボットがあまりに人間的になりすぎると、それらに対して不快に感じるようになる、という理論を立てた。森はこの快さの下降を「不気味の谷」と呼んだ（ここにはアンドロイドがさらにより人間に近づけば人々は否定的感情を克服することができるという前提がある）。

ロボットの設計者たちは不気味の谷に対して異なる反応を示した。石黒浩はそれを乗り越えるべき課題として見て、可能な限り人間に近く見えるアンドロイドを設計するという目標を持っている。他のロボット工学者は、不気味の谷は、効果的なロボットはいくつかの人間的特徴を持つが、人間であるふりをしないロボットであるということを示唆するものと受け止めた。

Kismet のような、基本的な社会的合図を認識して、社会的ジェスチャーで返す能力はすでに、おもちゃやサービス・ロボットを設計し販売することに興味を持つ企業によって流用されている。消費者はしかしこれを倫理的問題と思わないだろう。それでもこのようなテクノロジーによって顕在化する擬人的な反応は、意図されない有害な、あるいは非倫理的な活動を覆い隠す可能性がある。社会学者 Sherry Turkle は実験として養護老人ホームにロボットの人形を持ち込んだ。入居者たちがロボットの人形に対して持つようになった愛着は驚くほどだった、と Turkle は言う。

人間の感情的要求への反応として社会的ロボットを作ることが望ましいかどうかは難しい問題である。

Friedman と Kahn はテクノロジーを擬人化する傾向の別な深刻な倫理的問題点を指摘する。現在のテクノロジーは、人間の道徳的行為者に要求されるような知性や意図をまったく持っていない。このような行為者性を機械に帰属させることは危険である。それは**人間が責任を放棄する可能性**を示唆する。企業はユーザに慎重になるよう訓練する必要がある。

1980 年代後半の、核兵器を搭載した Trident 潜水艦の導入は、冷戦時代に地球を深刻な危機に陥れた要因の一つである。この船は、核兵器が発射されてから相手国に届くまでの 10 分の壁を破り、指導者がレーダーに現れた映像が攻撃なのか無害なのかを評価し、意思決定を下す時間を奪った。これによってソ連は、コンピュータにデータの分析をさせ、コンピュータに報復措置を始動させることを余儀なくされた。人類の未来は危うく 1980 年代のソ連のコンピュータ・テクノロジーに委ねられるところだった。幸運なことに兵器競争は終わった。しかし今日でも、生死の問題をコンピュータの手に委ねる人は、現在のテクノロジーの限界を理解していないのである。

P. 46〜（“It remains unclear.....”）

AI システムが世界の隅々に浸透するにつれて、人々がその限界をよりよく理解するようになっていくかどうかは明らかではない。例えば、人々は、人間がコンピュータよりも優れている分野を正しく理解するだろうか、それとも人間が機械よりも劣っていると感じるだろうか？機械的・反復的作業に関してはコンピュータは

非常によく機能するということは一般的に人間にとってありがたいことだと認識されている。

しかしコンピュータが人間よりもうまく、道徳的考慮、創造的仕事、そのたの複雑な課題をこなすようになったら、ある種の劣等感が生じ、自尊心が傷つけられるかもしれない。創造的能力を持つ機械は人間がその才能を試す動機を奪うかもしれない。しかし著者たちはこのような見方は誤りだと考える。知性や、芸術や、運動などに特別秀でていない子供も、素晴らしいことを成し遂げるように効果的に動機づけられることは可能である。洗練された機械の存在する世界における社会の課題は、人々の希望を養うことだ。

楽天家は人間に匹敵し、人間に勝る機械が将来登場すると考える。このような機械は人間の尊厳に対する打撃になるだろう。しかし著者たちは高次の精神的機能を人工的なシステムに実現することは、不可能ではないにしても、非常に難しく、その難しさはかえって人間がかくも素晴らしい生物であるかを強調することになると考えている。

Soldiers, Sex Toys, and Slaves

P. 47〜（“Might accepting robot.....”）

ロボットを人間の生活の中に受け入れることは人間が大切にしている価値を減じ、人々の人間性を劣化させるだろうか？ 最も成功したロボット工学者の Ronald Arkin は、「Bombs, Bonding, and Bondage」という言葉で、人間とロボットのインタラクションの三つの主要な形——兵士としてロボット、友人としてのロボット、奴隷としてのロボット——によって提起される社会的な関心を表現した。

人類はロボットの兵士を望むだろうか。私たちはすでにクルーズ・ミサイル、遠隔操作車両、戦場ロボットを持っている。2006 年のニュースでは、何百もの「PackBot 戦略的可動ロボットがイラクとアフガニスタンの市街戦で」配備されていると報じられた。

アメリカではロボット研究の多くは防衛省によって資金を与えられており、武装したロボットの開発という長期的な目標に数十億ドルが費やされる予定である。

最初は武装したロボットに殺人をさせる決定には人間のオペレーターが関わることになっているとしても、いつまでもそうとは限らない。実際に自律的な戦闘ロボットの開発が Defense Advanced Reserch Projects Agency（DARPA）によって進められており、2010 年には戦場に配備されることになっている。

著者たちの知る限りでは完全に自律的な、銃や爆弾をもったシステムはまだ世に出ていない。しかしこのようなシステムの利点はシンプルで抗いがたい。ロボットは戦闘における人間の需要を減らし、兵士たちの命を救う。また遠隔操作されるシステムも、現在のものは非常に複雑で多くのオペレーターを必要とする。自律的なロボット技術によってこれらの人員を削減することは、戦争において大きな利点を持つ。

戦闘ロボットが殺す許可を与えられたならば、ある特定の人間を殺すことが正当化されるか否かについてのリアルタイムの決定が必要である。**いつ、どこで、そして誰に対して致死的な力を行使することが許されるかについての道徳的判断なしでは、戦闘ロボットが受容できない危害を引き起こす可能性を減じる方法はない。**

人間が擬人化する傾向は、戦闘ロボットにも及ぶ。兵士たちはロボットとの間に繋がりを感じるようになる。

P. 48〜（“An entirely.....”）

社会的ロボット（人間とインタラクションするよう設計されたロボット）は、人間の心理を利用して、その使用しやすさを高め、製品に対する感情的な繋がりを感じさせようとする。

テクノロジーの発達はポルノ産業への応用によって推進されてきた長い歴史があり、ロボット工学の分野も

その例外ではない。他のテクノロジーのポルノ産業への応用と同様、女性の搾取や反社会的な行動の育成に関する深刻な問題が生じる。しかしロボット兵士についての議論と同様に、ここにも二つの側面がある。性的パートナーの代理として機能するロボットのアバターは、おそらくある種の「安全なセックス」を提供する。しかし性的なおもちゃとして使われるロボットが異常な反社会的行動を引き起こすことを示唆する逸話的な証拠はあるだろう。将来の研究がこのことを確証するかもしれない。

孤独の問題を解決するためのロボットの使用は、一時的な性的満足のための使用よりもさらに深刻である。研究者たちは非言語的および言語的なヒントから感情を読みとり、あたかも共感をしているかのような幻想を作り出すロ-ボットに取り組んでいる。2007 年の著作 *Love and Sex with Robots: The Evolution of Human-Robot Relationships* において、著者の David Levy は現在の研究の動向から、将来は人間とロボットとの長期のパートナーシップ、あるいは結婚すら実現するだろうと主張している。しかしロボットとの感情的な繋がりを深めることは、悪質な設計者や、将来的には半知性的なロボットに、ナイーブなユーザを搾取する機会を与えるだろう。もちろん人間とロボットの間の性的振る舞いについて、様々な共同体がどのようなものを倫理的とみなすか、という問題もある。少なくとも社会は洗練された友達ロボットが社会に与える帰結に取り組む用意をしておくべきだ。

P. 49〜（“A long-standing attraction of robots.....”）

人類は数千年の間、互いを望まない労役に従事させてきた。奴隷制の廃止は最近の事柄であり、またいまだに脆弱な道徳的原理である。Arkin は、ロボットを奴隷として使用することが、現在では公式に奴隷制度を廃止している社会において、その制度を有効な選択肢として復活させることになるかどうか考え、またこのことが人間を奴隷にすることを再び合法化する、あるいは人間を怠惰にするかどうかを考えている。

ロボットを労働に使用することは産業ロボット工学において、あるいは掃除機のような商業的な応用において、すでに確立されている。日本ロボット協会は、これからの数年のうちに、老人や障害者の介護をするサービス・ロボットを開発する、という目標を定めている。サービス・ロボットがよりかわいく愛敬のある、人間のようなあるいはペットのような外見をしていれば、その使いやすさと魅力はより高まるだろう。さらに人間と、サイボーグと、ロボットの間の区別があいまいになるにつれて、奴隷制への障壁もますます低くなるのは当然である。

それからまたロボットが最終的に感覚や感情をもち、知性、意識、自己理解を持つようになるだろうという、未来派の懸念がある。痛みを感じるロボットは、乱用をやめるよう人間に命令する権利を持つだろうか？高度な知性を持つロボットは働かないと主張する自由を持つだろうか？あるいはそういった証拠があっても、人間はそれらを、本当の感情や高次の精神的機能、意識を持たない劣等生物であると主張しつつづけるだろうか。

近い将来にはほとんどの仕事は、細分された様々な器具に埋め込まれたテクノロジーによって遂行されるだろう。ロボットのヘルパーが実質的に不可視であり、人格や感情を持たないときには、人間の奴隷制の不道徳性が疑われることはないだろう。しかし友達と奴隷の両方であるように設計された家庭用ロボットはすでに工学者の構想の中にある。

Can Technology Risks be Properly Assessed?

P. 50〜（“Our discussion so far.....”）

以上の議論において、先進的なロ-ボットの開発におけるいくつかの種類の社会的リスクが強調された。しかしそれらは**正確にどのくらいリスキーなのか？**

新しいテクノロジーのリスク評価は科学には程遠い。リスク評価には予測できないことがつきものである。

リスク評価のプロセスを実行することの価値は、予見できる利益と予見できるリスクを重みづけすることにある。これを実行しないことには、人々は目立った利益や否定的要因に対して不当な重要性を与えがちである。リスクの特定は、リスクのマネジメントを容易にする。テクノロジーの評価はまだ新しい分野であり、新しいテクノロジーの導入が、ただでさえ動的な社会的な文脈における変化にどのような影響を与えるかについての効果的なモデルを模索している。

リスク評価は二つの側面で、意思決定ロ-ボットを構築するプロジェクトにとって重要である。一つには、そのようなシステムの実装によって個人や社会に対して提示されるリスクがある。またもう一つにはリスクを評価するための道具は、様々な行為の選択肢の確率と帰結を予測することができるものであり、その限りにおいてその道具は、ある課題に対して選択できる反応に含まれるリスクを評価するためのロ-ボットに応用されるかもしれない。すなわち**リスクの分析は AMA が利用できる情報に基づいて、最善の行為を選択するために役立てることができる**。リスクを評価するために専門家が利用する、専門的な道具やテクニックはすでにコンピュータ化されている。これらは AMAs が、自分の行動の帰結を分析するためのプラットフォームを提供するかもしれない。

The Future

P. 51〜（“Nothing in life.....”）

人間は大規模な災害のリスクを高く見積もる傾向がある。AI に関する議論の背景には、AI がやがて人類を滅ぼしたいと望むような生物に進化するだろうという SF 的なシナリオがある。著者たちの現在の展望では、そのようなシナリオが実現するリスクは極めて低い。それが可能かどうかは判断するのが難しい。安定した AI を作るための土台がはっきりすれば、AI システムに適切な倫理的制約を組み込んで、人間を絶滅させるような可能性を消去することができるかもしれない。

未来派の極めて思弁的なファンタジーゆえに、AI から得られる利益を放棄するのはあまりに早計である。この生まれて間もない分野が進歩するにつれ、社会理論家、工学者、そして政治家は、パンドラの箱がすぐにでも開こうとしているのかどうかという問いについて考える機会を計画的に持つだろう。

人間は警戒していても過ちを犯すものだという考えは、未来派の危険に対処するのに有用な考えではない。この考えは、ある行為の帰結が知られていないが、大きなあるいは取り返しのつかない否定的帰結を引き起こす一定の可能性があると判断されるとき、その行動を避ける方が良い、という「用心の原則」としてしばしば定式化される。用心の原則の困難は、それをいつ持ちだすかについての基準を確立することに存する。1950 年代にあったロボットによる世界の乗っ取りに対する危惧ゆえに、過去 50 年のコンピュータ・テクノロジーの進歩を犠牲にしたいと思う人はほとんどいないだろう。後付けの知識なしには、どの危険が制御できない困難を表しているのかをいうことは難しい。だからといって警戒の必要性が軽視されていいということではない。

ここで提起された社会的問題は、AI の発達とともに生じてくるだろう懸念を強調している。しかし**これらの懸念は、人間が意思決定のできる、あるいは自律性を持つ AI を作ることをやめるべきだという結論を導くとは言い難い**。またどのような議論や証拠がそういった結論を支持するかも明らかではない。WHO によれば、交通事故は 15 歳から 44 歳までの負傷に関連した死因のトップであり、自動車事故は 1998 年に世界中で 1,170,694 件の死の原因になった。もしも 100 年前の人々が自動車がこれほどに破壊的になるということを知っていたら、彼らは自動車の開発をやめただろうか？ おそらくそうではないだろう。ほとんどの人は自動

車のもたらす恩恵はその破壊的な可能性よりも重要だと信じている。

著者たちは AI システムの破壊的な可能性に懸念を持っている。その懸念が人工道徳の分野を推進することへの彼らの興味を駆り立てている。かれらは社会批評家や未来派の人々がいま提起し血得る問題だけを理由に AI 研究を差し控えることには根拠がないと考える。AI の発達の危険がその報酬を上回るかどうかを折に触れて再評価する機会はあるだろう。一方で AMAs は、自律的なシステムによって提示されるリスクを効果的に管理する方法を探求するための重要な場所を提供する。それはまた道徳的行為者性そのものの本質を評価する場所も提供するのだｋが、これは次章の話題である。