Wendell Wallach and Colin Allen, *Moral Machine: Teaching Robots Right from Wrong*, Chapter 1　"Why machine morality?" 要旨

久木田水生

Trolley Car Drivers and Robot Engineers

P. 13〜（"A runaway trolley..."）

「トロリー事件」は倫理学の授業ではおなじみのシナリオである．トロリーが分岐点に差し掛かっており，片方の分岐の先に 5 人の作業員が，もう片方の分岐の先には 1 人の作業員がいる．このまま進めば 5 人を殺すことになるがスイッチを切り替えれば 1 人を殺すことになる．このとき運転手はスイッチを切り替えるべきか？ 別のシナリオでは，行為者は運転手ではなく傍で見ている立場で，巨体の男を橋から線路の上に突き落とすことで 5 人の作業員を救えるとする．死者の数は同じでも，多くの人は後者のケースの方がはるかに反論されやすいと感じる．心理学者と神経学者の実験では，後者の方が脳の感情を処理する中枢に激しい反応を引き起こすことを明らかにした．このような実験は，善悪に関する哲学的な問題に答えるわけではないが，倫理的問題に対する人々の反応の複雑さを示している．

P. 14〜（"Given the advent..."）

自動化された「運転手のいない」鉄道システムが実現できる現在では，これは人口道徳的行為者（AMA）の問題にもなりうるだろうか？ 鉄道網の複雑性が増大するにつれてトロリー事件と同様のジレンマが生じる可能性は高まる．

エンジニアは，鉄道システムが人間よりも安全であると主張する．ロンドンの地下鉄では 40 年以上も前に無人電車の試験を行っている．しかしその当時このシステムは仕事を奪われると考える鉄道労働者と，安全性に確信が持てない乗客からの政治的な抵抗にあった．現在では電車はコンピュータによって運転されている．人間が「監督」的役割を果たすために車両に乗っているはいるけれども．多くの乗客は緊急の時には人間の運転手の方がより柔軟に対処できると考えるが，コペンハーゲン地下鉄の安全責任者の Morten Sondergaard は「自動列車は安全だし，タイムテーブルが変化する速さのせいで自動列車の方がいざという時にはより柔軟だ」という．

しかしながら人々はなお，どんなプログラミングの範囲も超えた危機的状況があり，そこでは人間の判断の方がよいだろうと考えている．そのような状況の中には関連する判断が倫理的考察を含むような状況があるだろう．自動運転システムはもちろん倫理には無関心である．**ソフトウェア・エンジニアは彼らのソフトウェア・システムが倫理的な次元を表現できるように，それらを強化することができるだろうか，あるいはそうするべきだろうか？** この質問に対しては**人口道徳の領域で何が可能であるか**をよりよく理解しない限りは適切に答えられない．

コンピュータ革命は，自動化への異存を促進しており，自律的システムはますます様々な決定を担うようになっており，それらの中には倫理的な帰結を含むものもある．人の命や幸福を倫理的に無知なシステムにゆだねて，人はどれほど安心していられるのだろう．

P. 15〜（"Driverless trains..."）

運転手のいない列車はすでにある．線路に太った人間を投げ入れて 5 人の命を救うロ-ボットは当分実現しないだろう．一方で，テロを防ぐための監視システムは線路だけでなく，橋やトンネルや道路にも及んでいる．乗客の顔を調べて，既知のテロリストについてのデータベースと照会する空港監視システムは開発中であ

る．これらのシステムは異常な活動が見られたら人間の監視員に警告を発するのであるが，監視員がその行動を調べて対処するのに十分な時間がない時には，システムが列車の進路を変えたり空港の一部を封鎖したりするような可能性も容易に想像できる．

自動運転の列車が，一方の分岐に 5 人の作業員がおり，もう一方の分岐には子供がいることを認識できるとしよう．システムはこの情報を考慮に入れて決定を下すべきだろうか？ 自動化されたシステムが利用できる情報が豊かになるほど，それが直面するジレンマは複雑化する．**コンピュータが行動の帰結をどれほど深く考慮することを人々は望むだろうか？**

エンジニアはしばしば，ロ-ボットが困難な状況に遭遇した時は，ただ停止して人間が問題を解決するのを待つべきだと考える．倫理学には行為の方が行為しないことよりも責められるに値すると考える長い伝統がある．Responsibility と liability の問題については本書の最後でも扱う．ここで重要なのは **AMA の設計者は単純に行為しないことで良い行為の代用をすることを選ぶことはできない**，ということである．

Good and Bad Artificial Agent?

P. 16〜（“Autonomous systems are...”）

自律的システムは倫理的なもの，良いものになるだろうかと問うとき，ここでの「良い」は単に道具的に良い——ある特定の目的に相対的に良い——ということではない．道具的な良さは設計者やユーザーが持つ特定の目的に照らして測られる．自律的システムに求められる良い行いの種類はそれほど容易には特定されない．この意味での良さについて語るとき，私たちは倫理学の領域に入っている．

人工的行為者が害を与えうるというだけでは，その人工的行為者が倫理学の領域に入ってくるということにはならない．¥bf 道徳的行為者は，彼らの行動が引き起こしうる害や彼らは怠りうる義務に照らして，自分の行いをモニターし統制するものである．AMA にも同じことが求められる．これを達するには二通りのやりかたがある．一つは**プログラマーが行動の可能な進行を予測し，AMA が使われる状況の範囲で望ましい結果につながる規則を与える**ことである．もう一つは**情報を集め，行動の帰結を予測し，課題に対する反応をカスタマイズすることができるより開放的なシステム open-ended system を作る**ことである．

機械が本当に倫理的になれるか（あるいは本当に自律的になれるか）と問うときには常に，いかにして人工的行為者に道徳的行為者であるかのように行為させるか，という工学的な難問が残される．多目的機械が設計者や所有者の手を離れて動作し，現実のあるいはバーチャルの世界で柔軟に反応するようプログラムされながら，信頼できるものになるには，それらの行動が適切な規範を満たしているという確信が必要である．これは伝統的な製造物の安全性を超えている，自律的システムが危害を最小限にするべきものならば，そのシステムはまた可能な有害な帰結を「認知するもの」でなければならず，またこの「知識」にもとづいて行動を選択しなければならない．

Present-Day Cases

P. 17〜（“Science fiction...”）

自律的システムはすでに日常的な活動において倫理学の視野の中に入っている．Colin Allen がテキサスからカリフォルニアに車で移動したときに，ガソリンスタンドで給油をしようとしたらクレジットカードが使用できなかった．カード会社のコンピュータが，自宅から 2000 マイル離れたところで，そこに至るまで使われた痕跡なく，カードが使われるのが怪しいと判断したのだった．これは自律的システムが，人間にとって有益

または有害になりうる行動をとった出来事の例である．しかしこのコンピュータが倫理的な行動をしたというわけではない．このコンピュータがとった行動の倫理的な意義は，それにプログラムされた規則に内在する価値から派生している．その価値がカードの持ち主や会社に対して引き起こす不便を正当化するとは言えるかもしれない．しかし顧客は，システムは経済的な損益以上のことを考慮してほしいと感じるかもしれない．

2003 年，合衆国の東側とカナダで 1000 万人以上の人々が停電による被害をこうむった．これはクリーブランドで過熱した送電線が木に垂れ下がって起こった電力の急上昇によって引き起こされた．この出来事がたちまち，コンピュータによる電力の遮断の連鎖を引き起こし，8 つの州とカナダにまで及んだ．これは Blaster というコンピュータウィルスによってコンピュータがオペレーターに情報を伝えていなかったために起きた．Mikko Hyppönen の分析によれば電力ネットワーク上のコンピュータの一台か二台がウィルスに感染していればセンサーから電力オペレーターにリアルタイムのデータが送られるのを妨害し，オペレーターのエラーにつながる．

人間のエラーが避けられず，システムの状態すべてを把握するのが難しい場合は，自動化へのプレッシャーは増大する．不確実な状況では，**自律的システムがリスクと価値の重みを測ることが必要である**．自律的システムの広範な使用は，どの価値を重視するべきかという問いを差し迫ったものにする．デジタル時代には著作権などについての新しい考え方が生じ，またこれまでになかった犯罪も生まれる．このようなあたらしい価値や犯罪は新しい規制を生む．人々はその規制が AMA にも組み込まれることを望むだろう．

Ethical Killing Machines?

P. 20〜（“If the foregoing examples...”）

遠隔操作される乗り物（remotely operated vehicles; ROVs）はすでに軍事的に利用されている．しかしロボット工学の軍事利用は ROV にとどまらない．巡航ミサイルのような半自律的ロボットシステムがすでに爆弾を運んでいる．軍はまた爆弾処理や監視のために設計された半自律的ロボットを利用している．

軍事利用のためのロボットをつくることはやめるべきだという考える人もいるが，それに対しては，兵士や警官の命を救うという理由で反論があるだろう．もし戦闘ロボットに賛成する議論が勝てば，これらのロ-ボット，そしてすべてのロ-ボットの応用に必要な，組み込み式の倫理的制約について考えなければならない．実際にジョージア工科大学のロボット工学者 Ronald Arkin は，2007 年にアメリカ陸軍から資金を得て，戦闘ロボットが戦時における倫理基準に従うことができるようにする，ハードウェアとソフトウェアの開発を始めている．しかしロボットに戦時における倫理基準を守らせるのは困難な課題であり，戦争に利用される信じられないほど進んだロボット兵器システムの開発に，はるかに後れを取っている．

Imminent Dangers

P. 21〜（“The possibility of a human disaster...”）

ロ-ボット・システムが社会のあらゆる側面にますます埋め込まれるようになっているので，本当の潜在的な危害は，予期せぬ出来事の組み合わせから生じる可能性が最も高い．

専門家はアメリカの電力配電網は，古いソフトとハードに依存しており，ハッカーによるテロ攻撃に対して脆弱であると指摘してきた．脆弱なソフトとハードの多くが，より洗練された自動的システムに更新されている．これによって電力配電網はますますコンピュータによる制御システムの決定に依存することなった．経験したことのない状況において，これらの決定がどのように働くかを予想することは誰にもできない．

配電網の管理者は産業と一般の電力要求と，基本的なサービスを維持する必要とのバランスをとらなければならない．彼らは，節電中および使用電力が上昇しているときには，誰に回す電力をカットするかを決めなければならない．この決定には価値判断が含まれる．システムの自律性が増すにつれて，これらの判断は人間のオペレーターの手を離れていく．**不確実な状況において決定をガイドするべきシステムが，関連する価値に対して盲目であったならば，大災害を引き起こす可能性が高い**.

P. 22〜（“Even today...”）

今日でもコンピュータ・システムの行動は個別的に見れば小さいが，しかしそれを集積してみれば深刻である．Google の研究所の所長 Peter Norvig が言うには，医療ミスで亡くなる人は毎日 100 人から 200 人いて，しかもその多くがコンピュータに関係している．

今日のロ-ボットによって引き起こされている危害の多くは，欠陥のある部品やまずい設計に帰せられるものである．それ以外の危害は，十分な安全措置を組み込まない，そのシステムが直面する状況のすべてを考慮しない，ソフトウェアのバグを除去しないなどといった，設計者のミスに帰せられる．安全性が確かめられていないシステムを売り出そうとする，あるいはフィールドで試験しようとすることや，予想されていない複雑な状況に対処する際に，システムに誤って頼ることもまた危害を引き起こしうる．しかしながら部品の欠陥，不十分な設計，不適切なシステム，そしてコンピュータによる選択の明示的な評価の間の区別はますますつけにくくなっている．予測されなかった災害の後で，人は頼りにしたロ-ボットの不十分さのみを発見するかもしれない．

企業の重役たちはしばしば，倫理的な制約はコストを上げ，生産を阻害すると心配する．新しいテクノロジーに対する大衆の理解は，そのリスクに対するいわれのない不安によって妨害される．**しかし道徳的意思決定の能力を持つことによって AMA は，そうでなければリスクが高すぎると思われるかもしれない文脈で使われ，新しい応用の可能性を広げ，テクノロジーのリスクを軽減することができる**．システムが洗練され，異なる文脈や環境で自律的に機能する能力が広がるにつれて，それらが自分自身の倫理的サブルーティンを持つことがより重要になる．MIT の Affective Computing Group のリーダーの Rosalind Picard はこう言っている．「機械の自由度がおおきければ，それ分その機械は道徳的基準を持つ必要がある」