道徳的行為者としてのロボットの可能性…人類とロボットの共生の時代に向けて

The Possibility of Robots as Moral Agents : Toward the Age of Co－existence of Humans and Robots

柴田正良　金沢大学名誉教授

1. なぜ哲学はロボットに関わるのか？

たぶん，哲学と聞くと多くの人は，ロボットに最も縁遠い，徴臭い図書館の書棚に鎮座する「…著作集」や「…全集」のようなものを想像するに違いない．また，ギリシャの昔から有名な世界の哲学者たちの文献を，ひたすら読み続ける研究者たちの姿を．

しかし，実はそれは哲学についての誤ったイメージである．古代からの哲学書の研究は，「哲学」ではなく「哲学史」のお勉強である[1]．「哲学」はアリストテレスと生物学，カントと古典物理学，ヘーゲルと社会学，ジョン・ロックと心理学，そして最近では J．フォーダーと認知科学のように，その時代の最先端科学の誕生に自ら関与したり，その科学が人類にもたらす新たな変革を探索したりしてきた．そして，その最先端科学が「通常科学」にまで成長し，世間に受け入れられるようになると，哲学はそれを科学者の手に委ね，新たな未知の科学の最前線へと向かっていったのである．その意味で，哲学は常に「現代哲学」であり，そのイメージはタマネギの皮むきのように，科学の生誕とともに１枚ずつ剥がれる皮の芯にある現在形の「問い」の営みであって，決して出来合いの書物の中に答えを探そうとする文献解釈ではない．

もっとも人工知能やロボットは，たまたま現代の最先端の科学技術だという時代特性だけでなく，哲学にとっては，古代から続く大テーマの一つだという面も合わせもっている．それは，物と心という世界の二つの構成要素の関係を問う存在論である．思い切り大ざっぱに言えば，この存在論の第１段階では魂もしくは精神といった心が実在性において優位に立ち，物はそれに付随して存在する「おまけ」にすぎない．例えばプラトンにおいては，真なる実在はイデア（観念や想念そのもの）であって，われわれ人間が知る具体的事物は感覚世界に映し出されたその「影」である．それが時代を下ったデカルトの哲学では，心と物の地位は拮抗し，両者をつな

ぐために神の手に助けを求めざるをえなかった．これが第 2 段階，いわゆるデカルトの二元論である．そして現代の第 3 段階では，第 1 段階とは逆に物が心に対して優位に立ち，心や意識の存在と働きは脳という物体に付随してのみ出現し，物レベルの出来事や性質が心レベルの出来事や性質をすべて決定すると主張する物理主義が大勢となっている．こうして見るなら，ロボットというか，ロボットの心がなぜ哲学の問題となるかが理解されるであろう[2].

2. 他人の心…他我認識

「ロボットは心をもつようになれるか？」，この問いは，ロボットの哲学の核心に直接グサリと刺さる.「ドラえもんや鉄腕アトムを見れば，もてるようになるのが当たり前だって分かるよ」. これでは何とも小学生か科学の素人の答えで話にならない，と思うかもしれないが，実はどっこい，この答えには後で示すような深い真実が隠されている. それをつかまえるには，われわれ人間の間で他人の心がどのように認識されているのかを考えてみるのが一番よい.

哲学者たちは，割と古くから，他人の心の内容をどのように確実に知ることができるのか，あるいは，そもそも他人にも自分と同様の心が存在するとどうして言えるのか，ということに悩んできた．というのも，他人の心はあまりにもありふれていて，朝起きてから夜寝るまでの間それから離れたこともないほどなのに，改めて「では他人の心とは何か？」と問われると，誰もが答えに窮してしまうからである. 妻が示した許しの眼差しは，妻の心そのものだろうか？　バスの中で子どもを叱る母の厳しい声は，母の心そのものだろうか？　また，コーヒーを手渡す娘の手の温もりは，娘の心そのものだろうか？　いずれも否であろう．それらは心の表れであって，心はむしろそれらの物理的な事象の背後にある何かであるように思われる．しかし，背後のどこに？　どのようなものとして？

こうして哲学者たちは，他我認識・他我問題という迷路に入り込むことになった．生理学や生物学の進歩は助けにならない．先ほどの現象の背後にある身体組織やそれらを統括する脳をどれほど詳しく観察しても，見出されるのは心そのものではなく，もう一つの物理的存在である血液や神経細胞や，ホルモンやニューロンや，最終的には分子や原子である．しかし，一般にはどうみても，他人に心があるということ，つまり心の実在性は疑うべくもない事実だ．それどころか，生理学や分子生物学など存在しない

遥か昔から，人々にとって他人の心は立派に存在してきたのである．

このジレンマに直面して，哲学者たちの反応はいくつかに分かれた．心の実在性を否定して，いわゆる「心」というのは「心があるかのように語る言い回しや説明上の虚構にすぎず，本当に存在するのは生理学的な組織やその働きだ」とする道具主義や消去主義の人々[3]がいる一方で，「心を物質の一種として捜すのは誤りで，脳や身体の機能として考えれば，それらを大ざっぱに捉えたものとして実際に存在する」と主張する機能主義や還元主義の人々もいる[4]．それらの中間を含めた様々なヴァリエーションの中で，どれが　　はいまだに決着がついていないが，もう一つ，他我論に独特な厄介な主張がある．それは独我論(solipsism)とよばれる．そもそも問題の発端は，他人の心が，いろいろな形での表れとは別に，心そのものとして直接に知ることができないという点にあった．だから他人にも自分と同じような心が本当にあるのかどうかは分からない，というひどく懐疑的な反応も理解できるだろう．その際，自分に心があることは，少しも疑われていない．いや，むしろ自分が直接に知っているのは自分の心だけであって，実は心とは自分だけにしか存在していないのではないか．こうして，いかにも常識外れの主張である独我論が成立する．それによれば，他人に心があるように見えるのはただ自分が勝手にそう思っているだけのことにすぎない．本当は，世界にはただ一つの心，自分の心しか存在しない．

この極端な主張は，しかしロボットの心を考えるうえで極めて示唆に富んでいる．というのも，「機械仕掛けのロボットはただ心がある〈ふり〉をしているだけではないか」，と非難されたとき，われわれは，「状況は他人の場合とどこが違うのか？」と反論し，さらに「もしあなたが独我論者でないならロボットにだけ心を拒むのは首尾一貫していない」と突っ込みを入れることができるからである．そもそも，われわれは日頃こともなげに他人の心の状態について語るけれども，一度としてそれを「じかに知った」ことなどないのだ．「ああ，歯が痛いんだね，分かるよ」とか，「彼の悲しい気持ちがひしひしと感じられる」とか，「あなたが感動したのはあの強烈な色彩なんですね」とか言うけれども，われわれは彼らが経験した痛みも，悲しみも，色彩の感覚も自分のものとして経験したことはない．というよりも，経験することが原理的にできないのだ．われわれは，他人，他者の心の内容を文字どおりには経験することができない．これは，少し大げさに言えば，われわれの世界に課された（形而上学的な）独我論的状況である．もちろんロボットの心ですらも，この世界においてはこの状況の中にある他はない．

3. 一人称的世界とクオリア

あなたに十年来の友人がいるとしよう．ある日，ひどく気の合ったこの友人と横断歩道を渡っていると，あろうことか横から来た暴走車に友人がはねられ，道路脇に吹っ飛ばされた彼の頭蓋骨が真っ二つに割れてしまった．何という惨劇！　しかし友人の頭蓋からは脳漿やちぎれた脳細胞ではなく，IC チップやリード線やガラス片がこぼれている．彼は人間ではなくロボットだったのだ！！　このときあなたは，大いに嘆くだろうが，「自分がロボットだなんて一言もいわなかったな，オレを長いこと瞞しやがって」と，その友人を罵ることだけはしないだろう．

この小話は親ロボット派（？）の哲学者によるものだが，他者が心をもっていることにとって内部構造などはいかに関係ないかがよく示されている．要するに，大事なのは脳の中身ではなく外見や行動だ．実際，子どもたちにとって，ディズニー映画に出てくるようなコーヒーカップや木や岩が喋り，歌い，ときにはダンスさえするのは当たり前で，それらは子どもたちの本当の友人である．哲学的にこの種の考えを洗練させたものが，いわゆる行動主義である[5]．それは一定の説得力をもつが，心にはその持ち主固有の内面的経験があり，他人にもそのような内面世界がある，というわれわれの強固な常識を犠牲にしている．ロボットにしても，中身が空っぽで文字どおりの操り人形であったなら，たとえ一時は人をその気にさせても，やがて「自前の心をもたないただの人形」だとみなされるだろう．したがって，「心ある振る舞い」を生み出すには何らかの内部状態が必要だが，それは人間にのみ可能なのだろうか？

この内部状態，一般に心理状態とよばれるものには，哲学者がこだわる一つの区別がある．それは，信念や欲求や判断といった心理学的機能と，「感覚質」と訳されるクオリア（qualia）に代表されるような意識状態との区別であり，実はこの区別にこそ，ロボット哲学の最大の難問が潜んでいる．しかし，これからそれにアタックする前に，これまで一本調子で議論してきた「心の問題」の背景を示しておくほうがよいだろう（皆さんに飽きられないためにも）．

以下に列挙するのは，認知科学ならぬ認知哲学の観点から，AI やロボットに関していまだ完全には解かれていないと思われる問題群である（これらがすべてではないが）．

(1) AI／ロボットは，統語論的操作に加えて，意味論的
理解をすることができるのか？[6]
(2)「思考の言語」は存在するのか？　そもそも心的計算の対象である心的表象は存在するのか？　典型的な対立では，フオーダー VS. チャーチランド[7][8].
(3)ヒトの脳の計算方式は何か？　AI／ロボットはそれと同じ方式を採りうるのか／採るべきか？　例えば，古典的計算主義 VS.コネクショニズム[9].
(4) いわゆる「フレーム問題」は解決されたのか？　AI／ロボットは，専門領域ではなく，むしろ常識がもつ知識（信念）の「蓄積と活性化」能力をもちうるのか？　もちうるとしたら，どのようなメカニズムによってなのか？[10]
(5) AI／ロボットは，モジュールとしての複数機能からの情報をいかにして統合しうるのか？　中央演算処理の謎[11].
(6) AI／ロボットにおける学習は，いかなるタイプの機械学習が最適か？それは，ニューラルネットワークの重みづけ学習と同一か？
(7) AI／ロボットは，どのようにして，またどの程度，帰納的推論のような一般化能力をもちうるのか？
(8) 感情の機能とは一体何か？　それを，AI／ロボットにいかにして実装するのか？　また実装すべきか[12][13]
(9) 心理学的な意味での意識，つまり機能としての三人称的意識は，AI／ロボットでいかにして実現できるのか？
(10)「私」にのみ経験される一人称的意識，つまり現象的意識は，AI／ロボットにおいていかにして可能か？　また，AI／ロボットはクオリアを経験できるのか？[14]
(11) AI／ロボットは，いかにして価値判断，例えば道徳判断や美的判断をすることができるか？[15]

以上の各項目について詳しい説明は省くが,現在のわれわれの議論が(10)番以降に位置していることを見ておいていただきたい．実はそれらは，ロボット制作の技術的・工学的難題がきちんと定式化され適切に解決されたとしても，なおその外側に残る問題である．こうした哲学問題をうまく理解するには，上の（9）と（10）の違いに相当する以下の区別が決定的に重要になってくる．それは,「心の状態」ということによって，三人称的にあるいは客観的に観察・確認可能な状態と，その心的状態を経験する当人にしか分からない徹底して私秘的（private）で主観的な一人称的状態との区

別である．ところが，この 2 つはふつう区別されずに混同されており，その結果，後者のもつ真に哲学的な問題が見えにくくなっている．前者は心理学や生理学や脳神経科学が捉える因果的な「機能」であって，三人称的な世界の中の物理的事実として確定可能である．他方，後者は，その内実を他人が知ることは原理的に不可能な，一人称的な「感覚／意識」状態である．例えば，あなたが大海原の青さに感動しているとき，その海の色を他人も見ることはできるが，あなたがどんな色を感じているのかは他人には決して分からない．鮮烈な青を経験しているときの視覚野の興奮をあなたの脳の中に見出したとしても，それは青色をしてはいないし，ましてやあなたの感じている青色そのものでもない．そもそも各自に経験されている（はずの）色が，どんな風に当人に感じられているかはその当人にしか分からないのだ．私に生ずる一人称的経験は，いかなる他者も踏み込むことが原理的にできない一人称的世界を形成する．これが経験に関する独我論的状況であり，独我論を採ろうと採るまいと，われわれはそこから逃れることができない．

機能とクオリア（感覚）のこの二重奏を「痛み」の例で確認してみよう．たまたま自分が弄んでいたナイフであなたがかなり深く手を切ったとしよう．あなたは「いてて！」とか言いながら，応急処置をし，病院への運転を私に頼んだ．医者は傷を見て，素早く消毒と傷の縫合を行った．このとき，「痛い」という心理学的状態は，「身体の損傷→痛覚神経の興奮→信念や知識の発動→応急処置以降の行動」という一連の出来事のうち「痛覚神経の興奮」に対応するだろう．それは，前後の身体行動の橋渡しの役割を果たしている．つまり，一連の因果的出来事の連鎖の中で一つの機能を果たすものとして，三人称的に同定可能である．もちろん，あなたはこの時なによりも「強い痛み」を感じていた．しかし，あなたの感じた痛みはどこにあったのか？　あなたの痛覚神経の興奮は痛みの感覚そのものではない．たとえあなたが自分の痛覚神経の興奮をどうにかして自分で見たとしても，そこに痛みの感覚は観察できないだろう．それどころか，先ほどの心理生理学的・神経学的説明のどこにも，「痛みの感覚」は登場しないのである．それは，「痛みの感覚」に言及しなくとも先ほどの説明が完結している，ということだ．出来事の流れ全体をエイリアンの科学者が宇宙から観察したら，「痛みの感覚」があろうがなかろうが「すべては因果的に起こるべくして起こった」と報告するだろう．つまり一言でいえば，機能は感覚（クオリア）を必要としていない．これはわれわれの物理的世界が因果的に閉じているせいである[16]．この因果的閉包性もほかの哲学的テーゼと同様に経験的には立証できないが，それによれば，世界のどんな物理的出

来事にもそれを引き起こすのに十分な物理的原因が存在する．したがって，痛みに関する先の「心理学的・神経学的・身体行動的」な一連の出来事にとって，痛みの感覚，痛みクオリアは余計な代物である．言いかえれば，痛みクオリアはこれらの出来事のどれと対応していても構わないのだ．

ここには交差しえない 2 つの出来事・事象のレベルがあって，一つは最終的には物理的な説明に還元される出来事，つまり因果的閉包性の内部に位置する出来事であり，もう一つは，クオリアや「意識される限りでの意識」など，まさに経験の主体だけに主観的に，私秘的に生ずる出来事である．この 2 つのレベルの間には，哲学者が言う「説明ギャップ」が存在する．つまり片方のレベルの説明をいくら精緻に仕上げても，そこから他方のレベルの説明を導き出すことはできない．そこでわれわれも混乱を避けるために，一貫して，物理的世界に属する意識と心的機能を「心理学的意識」,「心理学的機能」とよび，経験主体の内面的世界に属する意識と感覚を「現象的意識」,「クオリア」とよぶことにしよう．

すると直ちに，三人称的な物理世界の出来事（機能）と一人称的な内面世界の出来事（クオリア）とがぴったりと対応している，ということの方が不思議に思われるだろう．両者がズレる可能世界だってある，どころかこの現実世界でもすでにズレているかもしれない，とわれわれを挑発したチャーマーズという哲学者がいる．彼の議論に従えば，３つのズレ，つまり逆転クオリア（inverted qualia），不在クオリア（absent qualia），哲学的ゾンビ（philosophical zombies）が論理的に可能である[17]．機能とクオリアが相互に完全に独立なら，物理的かつ機能的に同型の 2 人の人物が赤信号で同じく車のブレーキを踏むとしても，一方が信号に「赤クオリア」を感じ，他方は「緑クオリア」だったということも可能である．それでも 2 人の行動の間にまったく違いがないどころか,「何色が見えた？」という問いに対して 2 人とも「赤」と答えるだろう（逆転クオリア）．電磁波の特定の波長に対してどの色クオリアが対応してもいいのだから，ある人にはまったく色クオリアが生じないということも可能だろう（不在クオリア）．すると，この話の総仕上げとして，また別の人にはクオリアのすべて，および現象的意識の一切が欠けているということすらありうるだろう，外面的にはほかの人と何も変わらないのに，その人の内面生活は何もない．哲学的ゾンビとよばれるゆえんである．チャーマーズは，いまわれわれの中にこのゾンビがいるかもしれないが，誰も誰がゾンビであるかは分からない，と言う．この挑戦を退けるには，スーパーヴィーニエンス（付随生起）原理[18][19]という哲学的テーゼを新たに呼び出さなければならないが，それはさておきロボットに戻ろう．

他人にさえあるかどうか検証のしようがないクオリアや現象的意識を，ロボットは一体もつことができるのだろうか？　つまり，あれやこれやのロボットが一人称的内面世界をもつことは可能なのだろうか？　あるいはすでにもっている，とでも言うのだろうか？

もちろん，われわれの状況はそれほど「何でもあり」というわけではなさそうだ．一人称的世界の出来事と三人称的世界の出来事は法則的に対応している（スーパーヴィーニエンスが成立している），という実感がある．つまり，一人称的世界を三人称的世界のどんな事物でももっている，というほどクオリアや現象的意識は「暴走」していないように思われる．であるから，ロボットが一人称的世界をもつなら，それは人間の場合と同様に，その物理的基盤として心理学的意識や心理学的機能やセンサ機能が実現されている必要があるだろう．それらを少なくとも人間並みに実現するのにどれほどのロボット工学的なブレイクスルーが求められるかは分からないが，ロボットの哲学としては，それらがすべて成功したという前提に立っている．しかし，そのロボット工学の偉業が達成されたとしても，それは，ロボットたちに現象的意識やクオリアを生じさせる技術的保証とはならない．つまり一人称的世界をロボットにもたせるための確実な（あるいは自然法則的な，と言うべき）保証は存在しないのだ．ロボット制作者たちは当該のロボットをつくる．あとは世界任せだ．つまりわれわれの現実世界がどういったタイプの可能世界であるかによるしかない．だが，これはそれほど捨てたもんじゃない．姿形はどうあれ，そのロボットはあなたと同じように周囲の出来事に反応し，自分の気持ちすら正確に語る．要するに，あなたの隣にいる他人と同じだ．これ以上の何を，あなたは望むというのだろうか？

4. 自律ロボットと道徳的行為者としての可能性

さて，ロボットの哲学にとっては，先の（1）～（11）のような問題がすべて解決されたとしてもなお残る最後の問題がある．それは，いかにしてロボットとわれわれが共生していくか，という問題である．

かつて大森荘蔵はロボットの訴えをこう活写した．「ロボットに『本当に痛いのか』と尋ねればもちろんのこと，『間抜けたことを言うな，痛いったら痛いんだ』と答えるだろう（そしてその夜日記に，差別待遇を受けて心が痛んだ，と記すかもしれない）」[20]．われわれが相手にするロボットは，こういうロボットである．このレベルに達していないロボットは，もちろ

ん，われわれの身の回りにあふれている．というか，それが現在でも普通であろう．要するにそれらは，人の役に立つロボット，つまり道具としてのロボットだ．有名なアシモフの「ロボット 3 原則」が適用対象と考えていたようなロボットである．曰く，（i）ロボットは人間に危害を与えてはならず，また（ii）人間の命令に従わなければならないが，（iii）上の 2 つに反していない限り自己を守らなければならない．

こうしたロボットは，結局のところ「誰かのためのロボット」であって，「自分自身のためのロボット」ではない．変な言い方に聞こえたかもしれないが，道徳哲学で有名なカントの言い回しを借りればこういうことだ．人間は「それ自体が他のために存在するのではなく，それ自体のために存在するような価値の源泉」，つまり内在的価値そのものをもっている．それに対してロボットは，ほかの何かのためにのみ存在する道具的価値しかもたない．

われわれ人間が共生のパートナーとして選ぶのは，道具としてのロボットではないだろう．彼らとの知的に見えるどんな会話も，ある程度の長い付き合いには耐えられない．彼らに欠けているのは，見かけ上の知性ではなく，彼らそれぞれの背後にある一人称的な内面世界だ．それはわれわれの現在の（ある種の）共生の相手，猫や犬を見れば分かる．犬や猫でさえ，ある種の一人称的な内面世界をもっている．時として彼らに感ずる「見通しがたさ」や「勝手気ままさ」こそが，「従順さ」以上に，彼らをしてわれわれのペットたらしめているのだ．だからロボットが完全にわれわれの言いなりだったなら，それはわれわれの共生の相手ではない．

一言でいえば，共生の相手たるロボットは「自律性」をもたねばならない．ここでの自律性とは，「自らの内部状態から自分に最善と思われる判断を下し，それに従って行為することが可能であり，外部からのいかなる命令や規則も疑い，拒絶することができる」ということに他ならない．まさに，ある意味でアシモフ的原則の真逆である．自律性をもったロボット，自律ロボットは，自らの内部状態として，具体的には欲求と信念をもつだろう．そして，それらを内容とする「態度」をもち，それに基づいた意図的行為をなすという点で，最小限の素朴心理学的メカニズム（folkpsychological mechanism）をもつ[21]．ロボットの一人称的世界という意味での内部状態が，ほかの誰にとっても文字どおりには「侵入不可」であり，「共有不可」であることを思い出そう．これは，たとえこの世界でクオリアや現象的意識が三人称的世界の出来事に法則的に依存して生ずるとしても変わりがない．したがって，自律ロボットはほかの誰とも代えようがない一人称的視点から三人称的世界に臨み，その世界の中で行為する，

ここに生ずる代替不可能性こそ行為の「責任と自由」の基点である．というのも，通常の憶測レベルの「本人にしか分からない行為理由」の背後には，他者に対するこの絶対の「侵入不可能性」と「共有不可能性」が存在し，それによって，他者がどうあってもその行為の肩代わりをなしえないという限界がすべての行為者に課されるからである．この限界は，三人称的・心理学的行為理解にとっては，どれほど探索と説明を精緻にしても追い払うことのできない一種の「光源のような闇」であるが，行為世界においては，行為の責任と自由を生み出す原理である．

しかし，すべてをプログラムされたロボットに自由な行為が可能なのだろうか？　ここで，「自律性」の概念も，またそれが可能とする「責任」と「自由」の概念も，本来は程度を許すということを銘記すべきである．われわれの現実世界は，少なくとも量子レベル以上のサイズの物体に関しては決定論的な自然法則が支配する物理世界だと考えられる．そして「行為」は，量子レベルの出来事ではない．これが，ロボットどころか，人間を含めたすべての「行為者」の前提条件だ．まず，物理的な因果連鎖をまったく免れた「自由意志」，「自由な行為」という概念の空虚さを確認しよう．そんな自由意志は，行為者にとって迷惑きわまりない突発性の病気のようなものである．トイレに行こうと思ったら自分が空港行きのタクシーに乗っていたり，まったく何の前兆も感じていないのに突然お墓に行こうと思ってしまったり，さらにそれらが無原因の現象としてしょっちゅう生じたりしたら，われわれの行為世界は「ぐちゃぐちゃ」である．つまり，およそ「行為」という概念が崩壊してしまっているであろう．結局のところ，自由な行為とそれに課される責任は，決定論的な因果連鎖の世界においてのみ意味をなすのである[22]．

そのように見るならば，犬や猫も，また人間の幼児や認知症の老人も，ある程度は自由な行為者であるし，場合によってはそれなりの責任を問われうる．ロボットがあらかじめプログラムされた部分をもつように，かれらも本能という生得的なプログラムをもっている．要するに，「自律」や「自由」や「責任」や「行為」といった概念は，物理学や神経生理学や心理学といった科学的概念ではないし，それらにキレイに還元することもできない，言わばわれわれの「創作的概念」なのである．誤解を恐れずに言えば，ある種の「幻想」と言ってもよいだろう[23]．さらに先走るなら，倫理や道徳にまつわる概念もそうである．「私が悪いんじゃない，そのときの私の電子回路が悪いんです」と言うロボットは，「私が悪いんじゃない，そのときの私の脳が悪いんです」と自分の薬物摂取を言い訳にする人間と同様に．法廷でどう裁かれるかはわれわれが彼らの行為をどう見るかにかかってい

る．責任が創作的概念だという意味は，「どう見るべきか」の最終根拠が科学ではなく，われわれの態度や決定の内にある，ということなのだ．

これまでの議論をまとめるとこうなる．「何らかの仕方で自己接続する特異な物理システム」という存在者，すなわち「行為者」という一人称的特異点が三人称的物理世界に出現するのを現実世界は許容する．現実世界はそのような可能世界の一つであり，そこでは巧みにつくられた自律ロボットもまた，そうした行為者の一人になりうる．ただし，「行為者」としてのロボットがどういう問題をわれわれの社会にもたらすかは，このことから直ちに出てくるわけではない（実は，これこそ，21 世紀の哲学・倫理学が取り組むべき最大の課題の一つである）．

その答えに接近するために哲学的行為論（行為の哲学）のややこしい議論[24]に深入りするのは止めて，むしろ，なぜロボットが道徳的行為者として存在すべきなのかを述べて，この拙い解説を締めくくることにしよう．

私事で恐縮だが，大学在任中に 2 つの高校の新聞部から，「自分たちとロボットや AI の未来について」インタビューを受けたことがある．彼らは，概して真面目に，ロボットがもたらす自分たち人類への脅威を心配していた．というのも，未来のロボットはそれこそ知力も体力も人間より上であり，人間より丈夫で，過酷な環境にも耐えられるからである．しかし，そうした有能なロボットがもしも人間に対して悪意や敵意をもつようになったら，どうなるのだろう？　すでに多くの SF 小説や SF 映画で描かれてきたような，ロボットや巨大コンピュータによる人間支配，人間殲滅が生じないだろうか．つまりロボットによる人類の破滅が，高校生たちの心配の種なのであった．

これまでの議論の中には，こうした事態を不可能とする理屈はない．そこで，むしろ積極的に人類の破滅を回避する 2 つのシナリオを考えてみよう．一つは，自律ロボットが一人称的視点をもつパートナーとなる手前で，ロボット開発を禁止することであろう．この場合，自律ロボットは，封印された人類の技術として博物館入りになるだろう．しかし，これはほぼ確実に失敗する．人類の歴史を振り返ってみるまでもなく，人間は，違法だろうが何だろうがロボット開発を決して止めないだろう．われわれは好奇心によって今の人類になり，これからも好奇心とともに生きるだろう．たとえそれが身の破滅を招くとしても…．

シナリオの第二は，自律ロボットとともに，一つの同じ道徳共同体（共同社会）を創ることである．そこで人間とロボットは，同等のメンバーとして原則的に同じ権利と義務をもつことになるだろう．「同じ道徳共同体」というのがどれほど異様に聞こえようと，将来たまたま遭遇するかもしれ

ないエイリアン（宇宙人）たちと共存・共生の道を歩まざるをえない場合を考えてみるなら，むしろそれが不可避であることが分かる．ロボットたちとの道徳共同体の中でわれわれは，いま現在の幼児や認知症の老人のように，能力的には「劣った」メンバーかもしれない．しかし，この道徳共同体が創れなければ，ロボットたちとの共生の道はない．もちろん，これはわれわれが新しく創る道徳共同体であって，そのためには，道徳の概念や道徳規則もまたわれわれが新しく創らねばならない．人類のこれまでの知恵が試される
のはここだ．

かくして，道徳的行為者としてのロボットの可能性とは，実は，人類存続の可能性なのである[25]．

参　考　文　献

[1] 柴田正良：“哲学史の「お勉強」から哲学研究へ”，哲学・人間学論叢，no. 7，pp.83-89，2016.

[2] 柴田正良：ロボットの心　7つの哲学物語．講談社現代新書，2001.

[3] D. デネット：志向姿勢の哲学･･･人は人の行動を読めるのか？　君島・河田（訳），白揚社，1996.

[4] J. A. Fodor：*A Theory of Content and Other Essays*. MIT Press, 1992.

[5] G. ライル：心の概念　坂本・井上・服部（訳），みすず書房，1987.

[6] J. サール：心・脳・科学．第二章，土屋（訳），岩波書店，1993.

[7] J. A. Fodor：*The Language of Thought*. Harvard University Press, 1975.

[8] P. チャーチランド：認知哲学―脳科学から心の哲学へ．信原・宮島（訳），産業図書，1997.

[9] J. L. マクレラン，D. E. ラメルハート：PDPモデル―認知科学とニューロン回路網の探索．甘利（監訳），産業図書，1988.

[10] 柴田正良：“ロボットがフレーム問題に悩まなくなる日”，シリーズ心の哲学：ロボット篇．信原（編），勁草書房，pp.119-174，2004.

[11] J. A. Fodor：*The Mind Doesn't Work That Way*. The MIT Press. 2000.

[12] 柴田正良：“機能する感情・幻想する感情”，心／脳の哲学（岩波講座哲学05）．村田（編），pp.153-176，岩波書店，2008.

[13] 柴田正良：“感情ロボット”，感情と思考の科学事典．海保・松原（監修），pp.318-319，朝倉書店，2010.

[14] T. ネーゲル：コウモリであるとはどのようなことか．永井（訳），勁

草書房, 1989.
[15] 柴田正良:人間がロボットと共生する日…ロボットの心から人類の道徳まで…, 第 61 回香料・テルペンおよび精油化学に関する討論会, 2017. http://siva.w3.kanazawa-u.ac.jp/image/TEAC170910.pdf
[16] J. キム：物理世界のなかの心. 太田（訳), 勁草書房, 2006.
[17] D. J. チャーマース：意識する心. 林（訳), 白揚社, 2001.
[18] J. Kim：*Supervenience and Mind*. Cambridge U. P., 1993.
[19] 柴田正良：“よみがえったソクラテス　物理主義と心的因果の問題を理解するために”, 思想, no.982, pp.4-15, 2006.
[20] 大森荘蔵：流れとよどみ. 産業図書, pp.67-68, 1981.
[21] 柴田正良：“行動ロボットと AI”, 感情と思考の科学事典. 海保・松原（監修), pp.314-315, 朝倉書店, 2010.
[22] 柴田正良：“自由な行為者としてのロボット”, これが応用哲学だ！, 戸田山・美濃・出口（編), pp. 135-143, 大隅書店, 2012.
[23] 柴田正良：“幻想としての自由意志と責任の帰属可能性”, 日本倫理学会第 60 回大会報告集, pp. 24-28, 2009.
[24] D. ディヴィドソン:行為と出来事.服部・柴田(訳),勁草書房,1989.
[25] M. Shibata: “Toward robot ethics through the Ethics of Autism,” *Neuromorphic and Brain-Based Robots*. J. L. Krichmar, H. Wagatsuma (eds.), pp.345-361, Cambridge U. P., 2011.