自由な行為者としてのロボット

柴田正良（金沢大学人文学類）

１．　哲学者がロボットを作りたい理由（決定論的世界における自由な行為者）

応用哲学者がロボットを作ろうとするとき、その動機は何だろうか。動機はさまざまあるだろうが、最も単純にして純粋なものは、自分の哲学的主張をそれで確かめてみようとすることだろう。もちろん、哲学的主張の多くは通常の科学的・経験的主張を超えたものだから、それがロボット制作によって確かめられることは科学理論の確証ほどにもないはずだ。しかし、存在論や形而上学において物理主義的な傾向の主張の持ち主なら、つまり心や感情といった心的現象もすべて何らかの物理現象に依存して生ずると考えている哲学者たちなら、少なくとも自分たちの主張が説得力を強めるような仕方でロボットが制作されるのを見たいと思うだろう。一言にしていえば、それが理由で、応用哲学者（の一部）は自らロボット制作に参加したいと思うわけである。

例えば、人間である行為者は、ふつう自由な行為者として振るまい、その行為に責任を問われるという意味で＜自律した行為者＞である。つまり彼は、心神喪失者でも心神耗弱者でもなく、善悪の判断能力と行動の制御能力をもった行為者だと見なされる。そこで、もし(a)いかなる現象・存在も物理的な現象・存在によって生じ、しかも(b)ある種の高次機能は多くの種類の物理的素材によって多重に実現されるということがこの現実世界において真ならば、物理学・化学・電子工学などでよく用いられる素材を組み合わせて、われわれと同じような自由な行為者を＜自律したロボット＞として制作することが原理的には可能だろう。以下では、＜自由な行為者としてのロボット＞を制作するための哲学者のレシピを少し考えてみることにする。もしこうして制作されたロボットがわれわれから見ても自由な行為者にしか見えないなら、＜自由＞という概念それ自体の応用哲学的な解明が一定程度成功したということにもなるだろう。

まず、ある行為が行為者によって自由に行われる、とはどういうことだろうか？　伝統的な意味での自由概念の一つに、決定論的な世界においては自由な行為は不可能だ、とする考えがある。この世界では物理的事象には、いわゆる「自由な事象」は存在しないように見える。少なくともマクロな現象レベルでは、すべての物理的事象は、原因が決まればどのような結果が生じるかも法則的に決定されている。すると、われわれの意図や意志が脳過程に依存し、脳過程が物理的事象である限り、意図や意志もまた先行する原因によって決定されている他はないように思われる。したがって、＜自由である＞ということを「先行する原因によって完全には決定されていない」と解釈するなら、この意味での自由は決定論的世界では不可能であろう。

しかし、多くの哲学者が多くの箇所で論じているように、決定論的でない世界、つまり

非決定論的な世界においても、＜原因による決定性から逃れている＞という意味での自由は、われわれが＜自由であること＞によって理解しているもう一つの意味とは相容れないように思われる。その意味とは、われわれはほとんど常に「なそうと意図した（意志した）通りに行為する」ということだ。つまり、われわれの意図の出現は、ほとんど常に意図された内容の行為を引き起こし、とくに意図通りの行為が自分にできなかった場合でも、なぜできなかったかを説明する十分な原因がある、とわれわれは確信している。実は、決定論と相容れない＜自由＞は、多くを論ずるまでもなく、ほとんど行為者にとって有り難みのない、無内容な＜自由＞なのである。というのも、このような自由の下では、「空港に行こう」と意図したら「お風呂に入っていた」とか、「ラーメンを食べよう」と意図したら「洗濯機を回していた」といったことが、原因による結果の非決定性の度合いに応じて頻繁に生ずるからだ。このようなことを実は含意している＜自由＞の観念は、恐らく人類が長年の間保持してきたある種の誤解による幻想である。

そこで、後者、つまり意図通りにほぼ常に行為がなされるという意味での＜自由＞が原因（意図の出現）と結果（行為の遂行）の例外のないつながりを要求する限り、決定論的な世界においても、いや決定論的な世界においてこそ自由な行為は実現される、と考えるのが妥当であろう。たとえば、さまざまな動物たちの行動を考えてみよう。彼らに人間のような複雑な内容をもった意図が生じているかどうかは定かでないが、少なくとも、＜意図＞の原初的な形態としての未分化な＜欲求＞や＜信念＞に関する限り、捕食者の注意を雛からそらせるために怪我をした振り（擬傷行動）をするヒバリや、餌の隠し場所に確実に舞い戻るリスに、それらを帰属させるのはそう不自然なことではない。かれらがそうした内的な原因に駆られて、意図通りとしか思われない仕方で巧みに動き回るとき、われわれはそこに動物たちの＜自由な行為＞をごく自然に見出さないだろうか。人間の場合でも、発達心理学などによる正確な特徴づけはどうあれ、乳幼児から大人に成長する過程のどこかに、いわゆる「自由意志」という特別な能力が獲得されるといった出来事を想定するのはきわめて困難である。つまり、動物と人間との連続性の文脈で考えるなら、決定論的なこの現実世界に現に「自由に行為するものたち」が存在する限り、「自由な行為ロボット」を制作する上での障害は、人類だけが形而上学的に特殊な＜自由＞を所持しているのかどうか、といった昔の倫理学的問題にはないことになるだろう。それなら、われわれは、たんに動物と同じレベルで自律した振る舞いをするロボットを作ればよいのだろうか？　もしあなたが動物たちと同じ程度の「自由な振る舞い」で満足するならばそうだろうが、やはりここには、動物たちと人間の行為者を分かつ大きな溝がある。それは、人間の場合、すべてではないにせよ多くの典型的な自由な行為においては「それをなそうとする意図」を明確に自他に示すことができるという点である。

つまり、自由な行為者としてのロボットの制作は、＜原因なき行為＞という奇跡の現象をロボット上に実現する必要はないとしても、意図的な行為者としてのロボットを制作する必要はあることになる。

２．　意図的行為者としてのロボット

ここで私は、デイヴィドソンの行為者性(agency)についての暫定的定義を導きの糸にしている。それはこう述べる。

「ある人がある出来事の行為者であるのは、彼がなしたことに関して、彼がそれを意図的になしたという文を真ならしめるような記述が存在するときであり、またそのときに限られる」[デイヴィドソン　一九九〇、六九]。

これは人間のなすあらゆる身体運動・身体動作のうち、しゃっくりや胃の蠕動運動などから「行為」を分かつためにまず「意図的行為」の概念に訴えている点が重要である。つまりこれは、ロボットを行為者に仕立てる際、われわれはまず行為ができるように彼を作ってから、その後で意図的行為ができるように「意図性」を彼に付け加える、といったヴァージョンアップはできない、ということを意味する。行為ができるためには、まず意図的行為ができなければならない。したがって、この観点からすれば、繊細かつ軽妙な運動能力を次々に身に付けている現在の様々なロボットは、もしそれらが意図的行為をなしているのでないとしたら、単なる「機械的動作」をなしているにすぎないということになるだろう。

では、ある行為が意図的行為であるとはどういうことだろうか。この点で、現代哲学がアンスコムの基本的分析（基準）より大した前進をなしたとは私には思われない。アンスコムによれば、意図的行為とは、「ある意味で用いられる「なぜ？」という問いが受け入れられるような行為」である［アンスコム　一九八四　一七]。もちろんアンスコムは、「なぜ？」という問いが行為の理由を問うものであることを承知している。したがって、この問いが行為者によって肯定的に答えられる場合、彼は自分の行為をなした理由を与えていることになり、そのとき彼は、その問われた行為記述の下での行為を自分が意図的になしていることを認めているのだ。たとえば逆に、自分がコーヒーを捨てるつもりで、それと知らずにカップの中身であるお茶を捨てたとき、彼は、「なぜあなたはお茶を捨てたのですか？」という問いを受け入れることができない。彼はこのとき「お茶を捨てる」という意図的行為を行っていない(「カップの中身を捨てる」という意図的行為は行ったとしても)。

では、行為の理由とは何か？　ここでも私は、デイヴィドソンの行為の因果説の要点に基本的に従う。「行為の主たる理由は行為の原因である」[デイヴィドソン　一九九〇、一五]。そして、行為の原因としてわれわれが最もなじみ深く理解しているものは、素朴心理学で言うところの欲求と信念である。まさにわれわれは、通常の行為理解の場面では、「今、オムレツが食べたい」という欲求と「今、タマゴを割らなければオムレツは食べられない」という信念に動かされて、直ちにタマゴを割るのである。

さてそこで、これまでの意図的行為に関するストーリーをロボットの身になってまとめてみよう。ロボットにとっても、ある「欲求」d1 を持つということは、外部刺激と過去の

内部状態から自分に自然に生じてくる過程だ。おそらくいつの時点でも、ロボットにはそのような自然な欲求 d1～dn が複数個生じているだろう。それらは、「現在の状況 s では、d1 を実現するためには行為動作 a1 をしなければならない」、あるいは「現在の状況 s では、d2 を実現するためには行為動作 a2 をしなければならない」といった信念群 b1～bn と一緒になって、身体動作 a1～an を行うための原因となるだろう。これらは極めて機械論的なプロセスであって、この段階ではロボットは、アンスコムの意図的行為の基準を満たすことはできない。しかし、それを満たすために必要なのは「意識」などといった高級な仕掛けではない。必要なのは「表象」と「選択」と、それらを内容とする「第一階の自己」である。欲求と信念に相当する内部状態が身体動作を引き起こすとき、まず、それらの状態がロボットのある種の内部表現において分類されていなければならない。その表現は行為原因の「表象」であり、それらがロボットの「第一階の自己」にとっての行為理由となる。その際、複数の互いに競合する＜欲求・信念＞のペアが行為原因の候補としてロボット内部の、たとえば「動機」といった表象ボックス内に生じているのが普通であろうが、そのうちの一つのペアが「選択」され、いわば「意図」といったボックスに送り込まれるだろう。ロボットの行動は、「意図」ボックスに送り込まれた＜欲求・信念＞を原因とし、それを実現するメカニズムに従って展開されることになる。

３．　進化の文脈における自由な意図的行為

前節での極めてラフなストーリーがすべて＜原因-結果＞の因果連鎖によって生じていることに注意しよう。たとえ、そこにいわゆる「サイコロを振る」ような確率過程を介在させたとしても、その因果的な性格に変わりはない。言い換えれば、ここで時に応じて生じてくる「欲求」も「信念」も、ロボットの第一階の自己にとっては、われわれと同じように、自分が与り知らぬ因果過程によって自らに生じてくるものにすぎない。では、「選択」はどうか？　ここにこそ、ロボットには不可能だが人間には可能な「自由な行為」の根拠があるのではないか？　しかし、そうではない。結局のところ「選択」も、＜欲求・信念＞ペアの複数の選択肢の中から、何らかの因果的メカニズムによって一つを選び出すプロセスに他ならない。成熟した人間の「選択」の場合、各々の選択肢に対する評価・重み付けが過去の経験や、帰結に対する仮想的想像や、手段の合理性に関する計算などによって行われ、その比較の結果、一つの選択肢が「外的な何かに強制されることなく」選ばれる。もちろん、推論や評価や反事実的仮想などの能力をまだリッチに持っていないロボットにとっては、この「選択」のプロセスは比較的単純であろう。しかし重要なことは、このとき、ロボットの第一階の自己にとって、選択の因果的メカニズムは「表象」の背後に隠されたままであり、表層レベルにある「表象」こそが自己にとっての行為理由を形成する、という点である。つまり、この種のロボットに行為理由を尋ねたなら、彼は、第一階の自己の内容たる「意図」ボックス」内部の「表象」を答えとして返すだろう。

しかし、一体なぜ人間は自らの行為の原因を改めて「表象」のレベルで様々な行為理由として分類し、それによって自分たちの行為を正当化したり、合理化したりする必要があったのだろうか。ロボット制作の観点から見れば、ロボットの身体動作を引き起こす原因の連鎖の各結節点をあえて「表象」として抽象的に内部表現し、それを外部に行為理由として表出可能とするような仕掛けは、さほど必要ないように思われる。たとえ「表象」に対する操作の重要性をどれほど高く見積もったとしても、その「表象」を「自己知」の内容とする必要まではないはずだ。むしろ、行為の原因を行為理由として分類し、それを行為の説明に用いる必要があったのは、他者とのスムーズな協同作業が集団生活において必須であった人間にとってではなかっただろうか。逆に言えば、集団としての協同作業が必須でない動物たちや、作業の手順や役割分担などの細かな分節化ができない動物たちにとっては、アンスコム基準が求めるような行為理由の伝達や共有は不必要であるか、あるいは不可能であろう。ただ人間だけが、集団としての協同行動を行うために、まず自らの行動原因を行為理由として分類・分節化し、その内容を外的に示すことによって他者による行為予測を可能にさせ、そのようにして他者による意図理解を自分の意図的行為の遂行に役立てる必要があったであろう。したがって、誤解を恐れずに言えば、もし人間が他者との綿密な協同作業を一切行わない存在であったなら、行為の原因を行為理由として捉え返しもしなかっただろうし、そもそも自らの心的状態に対する現在のような内省もなかっただろうし、アンスコム基準を満たすような意図的行為も、したがって「行為」も存在しなかったであろう。サールの言う集団的志向性(collective intentionality)が本当に彼の言うように個人的志向性に還元できない「生物的に原初的な現象」なのかどうかはともかく［Searle 1995 24]、仲間たちとの意図の共有は、激しい鳴き声や突然の疾走による動物たちの本能的な行動にせよ、人間の場合の言語による意図の伝達にせよ、どちらも進化における生存上の必要性から生じたように思われる。ただ両者に、意図の分節化された「内容」がどれほど細かくやり取りされるかに関して、決定的とも言える違いがあるのも確かだ。つまり人間の場合には、より肌理の細かな他者との協同作業の必要性が、アンスコム基準を満たすような意図的行為を生みだしたように思われるのだ。そして、その本質的な原型なら、すでにスケッチしたようにロボットにおいて十分に実現可能であるだろう。

だが、それにしても行為に際して人間が持つ「自由である」という実感をどう考えたらいいのだろうか。それすらもロボットは持つことができるのだろうか。この問題と、なぜ「自由だ」という実感を人間は進化の途上で持ってしまったのかという問題について、私はまだ自分で納得のいくストーリーを描くことができないでいる。その本筋の周辺をうろついているばかりだ。おそらくそれは、「論証」というより「物語り」といったものになるはずなのだが･･･

参考文献：

１）和文献・論文

金野武司･柴田正良　二〇一二　「回帰的意図理解をめざす共同注意ロボット」『科学哲学』四四巻二号（掲載予定）

柴田正良　二〇〇八　「感情のクオリアと可能世界」長滝祥二・柴田正良・美濃正編『感情とクオリアの謎』昭和堂、三－三〇頁。

柴田正良　二〇〇九　「幻想としての自由意志と責任の帰属可能性」『日本倫理学会第六〇回大会報告集』日本倫理学会、二四－二八。

柴田正良　二〇一〇　「自由であるという実感（これは幻想ではない）」『倫理学年報』第五九集、日本倫理学会、五一－五五。

柴田正良　二〇一〇　「行動ロボットとAI」、「感情ロボット」海保博之・松原望監修『感情と思考の科学事典』朝倉書店、三一四－三一五頁、三一八－三一九頁。

２）和文献・単行本

柴田正良　二〇〇一　『ロボットの心』講談社現代新書

アンスコム　一九八四　菅豊彦訳『インテンション』勁草書房。

デイヴィドソン　一九九〇　「行為・理由・原因」、「行為者性」服部裕幸・柴田正良訳『行為と出来事』勁草書房。

３）欧文献・論文

Shibata, Masayoshi, 2011 “Toward robot ethics through *the Ethics of Autism*”, in J. L. Krichmar and H. Wagatsuma(eds.), *Neuromorphic and Brain-Based Robots,* Cambridge University Press, 345-361.

４）欧文献・単行本

Searle, John R. 1995, *The Construction of Social Reality*, Free Press.