ロボットの心・意識を作る

慶應義塾大学　前野 隆司

（Bio Industry, Vol. 26, No. 9, 2009 年 9 月, pp.59-63）

１．はじめに

ロボットの心、中でも「意識」を作る、という課題は壮大で、まだ緒についてもいないともいわれる。確かに、茂木健一郎氏 [1]がいうように、クオリア（現象的な意識、意識の質感）の問題が解決されない限り、ロボットの意識（現象的な意識）は作りえないという見方も可能である。しかし、茂木氏との対談 [2]でも述べたように、私は「ロボットの心の作り方」という論文 [3]を書いて以来、茂木氏よりも若干楽観的な立場に立つ。読者を煙に巻く議論に聞こえるかもしれないが、それは、西洋的人間中心主義の呪縛から人類を解き放つひとつの考え方だと思っている。本稿では、そのような考え方について概説したい。

２．意識とクオリアの定義

意識について語る前に、まずは言葉の定義をもうすこし詳しくおさらいしておこう。まずは「クオリア」。クオリアとは、上にも述べたように、現象的な意識ないしは意識の質感と呼ばれる概念である。「たたかれた頭が痛い」ときについて考えてみよう。ロボットの頭をたたいたとき、その衝撃力ないしは加速度を計測し、その大きさとパターンに応じて、ロボットに「痛い」と言わせることは可能である。早稲田大の高西研のロボットはまさにそれを行うことができる。もしもロボットの外見や動作が人間そっくりであり、ロボットがリアルに痛そうに振舞ったなら、観測者は「ロボットが痛がっている」と感じるかもしれない。しかし、ロボットは本当に痛いのだろうか？既存の技術で作られたロボットは、人間が心の中に感じる「痛み」を本当に感じているとは思えない。なにしろ、ある条件ではある行動を取るというプログラム（あるいはニューラルネットワークの指令など）に応じて痛いかのように振舞っているだけであって、ロボットの頭脳の中に、人間が感じる「痛み」が沸きあがっているとは思えない。

心の哲学者は、このように、ロボットに人工的に「痛いかのように振舞わせる」ときの痛みを「機能的な痛み」と呼ぶ。心理学において、人間の痛みを客観的に計測する際の「痛み」も「機能的痛み」である。一方、自分にしかわからない、心の中に痛みが沸きあがってくる感じのことを、「現象的な痛み」ないしは「痛みのクオリア」と呼ぶ（「現象的な○○」と「○○のクオリア」は同義と考えてよい）。こちらが、哲学者が対象とする「痛み」である。「現象的な痛み」は「機能的な痛み」と違って主観的であり、自分にしかわからない。極端なことを言えば、他人が「痛い」というとき、その「痛み」というものが、あなたの「痛み」と同じような感じであるかどうかもわかりようがない。

たとえば、「赤い色」もそうである。「機能としての赤い色」は、波長いくつの電磁波のこと、と客観的に定義できるが、「現象としての赤い色」は個人的体験であって、他人の脳裏に同じ「あの赤い感じ」として感じられているかどうかはわからないし、調べようがない。これが「赤い色のクオリア」である。ロボットの場合も同様であり、ロボットに「機能としての赤い色」を検出させ報告させることはできるが、「現象としての赤い色」な

いしは「赤い色のクオリア」を体験させることは現状ではできない。

つまり、現象的な知覚は第一人称的であって他人に説明できない。痛みを辞書で引くと「肉体的な苦痛」、赤を辞書で引くと「血のような色」、と書かれているが、あなたにとっての「肉体的な苦痛」や「血のような色」が他人と同じであるかどうかについてはやはり言語では説明できない。いくら言葉を並べても「クオリア」を説明しきることはできないのである。

なお、「痛み」や「赤い色」のような知覚でなく、「うれしい感じ」のような内面から沸きあがるものもクオリアである。前者を感覚性クオリア、後者を志向性クオリアと呼ぶ場合もある。

つぎに「意識」とは何か。覚醒している状態のことを「意識がある」というが、ここで取り上げる「意識」はそれではない。「痛み」や「赤い色」や「うれしい感じ」が脳裏（あるいは眼前）に沸きあがってきたということ自体に注意を向け、第一人称的に体験すること自体を、「現象的な意識」と呼ぶ。

３．意識のクオリアは幻想か？

これまでに述べたように、ロボットの脳であるコンピュータに、機能的な「痛み」、「色」、「嬉しさ」を作り出すことはできる。しかし、現象的な「痛み」、「色」、「嬉しさ」と、それらが沸きあがってくる場としての現象的な「意識」を作ることは、今のところできていないと考えられる。理研の谷氏は、ニューラルネットワークに現象的な「意識」が創発すると主張しているが、残念ながらクオリアは第一人称的なので第三者的に確認することはできない。

では、ロボットに「意識」を持たせることはできないのだろうか？

ここで発想を転換してみよう。

人間には（現象的な）「意識」はあるのだろうか？

科学は、客観的に再現可能なものを対象とする（と、とりあえず科学を狭義に定義しておく。日本語では、「人文科学」というときのように広義の「科学」定義も存在するのだが）。

科学を狭義に捉えると、第一人称的な現象である「意識のクオリア」は科学の枠外ということになってしまう。したがって、人間が意識を持つかどうかは、科学の枠外の（人文科学の中の）哲学の議論になってしまう（実際、本稿は哲学的考察である）。したがって、そもそも人間が意識を持つかどうかを定義できないのに、ロボットに意識を持たせられるかどうかという議論を行うこと自体が無意味だと結論付けられる。

これで論証終わり、でもかまわないのだが、もう少し科学的解釈を欲する方もおられよう。そこで、近年の脳・神経科学、認知科学の知見は、「機能的意識」をどのように捕らえているのか、という視点から「現象的意識」を眺めてみたい。

拙著[4)5)6)]にも述べたが、人間における「意識の時間」は思いのほかいい加減であることが知られている。たとえば、私たちが「青い服を着た人」を見るとき、「青い服を着た人」を見たという視覚のクオリアは、見た瞬間に意識に上る。しかし、「青い服を着た人」を認識するために必要な脳の神経回路の計算時間は 0.5 秒だという[7)]。見てから 0.5 秒も経ってから認識するのでは、なんとも間延びしていて行動も遅延してしまう（まさに一昔前のロボットのようである）。したがって、本当は「青い服を着た人」だとわかったのは見た 0.5 秒後であったのに、見た瞬間にそのように感じたかのように、意識の時間はずらされていると考えるべきである。奇妙に思えるかもしれないが、錯視図形を見たときに空間がゆがんで感じられるのと同様、意識の時間はゆがんでいると考えるべきなのである。また、素朴に考えると、現象的な意識の上に湧き上がった「自由意志」であるように感じられる意思決定が、実は意識に先行する無意識的な過程で既に生成されている証拠が続々と見つかっている。「自分がまさに行った」と意識の上で感じるよりも零コンマ何秒か前に、無意識的な意思決定はなされているのである。

つまり、人間の意識は、「あたかもいまここにある」かのように感じるように私たちが作られているから「あたかもいまここにある」かのように感じるのであって、本当は「いまここに」はないのである。つまり、幻想である。逃げ水が、あたかも目の前にあるように感じられるにもかかわらず、本当はそこにはないのと同様である。「幻想」とは、本当はないものがあるように感じられることを指す。

で、あれば、前述の痛そうに振舞うロボットとどこが違うのであろうか。

前述のロボットは、本当はいまここに痛みのクオリアは存在しないにもかかわらず、痛みが「あたかもいまここにある」かのように振舞うように作られているから「あたかもいまここにある」かのように振舞っているのであった。ロボットの痛みのクオリアは幻想である。

したがって、「本当は今ここにはない」幻想が「あたかも今ここにある」かのように振舞うという点で、人間とロボットは大差ない。

人間のクオリアは、どのようにして作られるのかは不明であるが、脳のニューラルネットワークによって、あたかも今ここにあるかのように感じられるように作られている。

ロボットの「痛み」「赤い色」「幸せそうな感じ」は、コンピュータプログラムや人工ニューラルネットワークによって、あたかも今そこにあるかのように感じられるように作られている。

ただ、ロボットのほうは、いまだ、人間とは違って、『「痛み」「赤い色」「幸せ」を本当にリアルにいまここに感じていますよ』とは言わないという点が、両者の相違である。

したがって、ロボットのプログラムをもう少し巧緻化し、「あなたは本当にクオリアを感じていますか？」という質問に対し「クオリアを本当にリアルにいまここに感じています」といわせるだけのプログラムを作れれば、もはや人間との差は極めて小さいといわざるを得ない。

以上のように、ロボット工学者としてはいささか邪道だといわれるかもしれないが、私は、ロボット技術や AI 技術を高度化させることによってではなく、人間についての見方を変えることによって、ロボットの意識の問題は解決されると考えている。

このような問題構造は幸福の問題と似ている。

統計的に見ると、科学技術がいくら進歩し、世の中がいくら便利になっても、人間はそれに比例して幸福になったわけではない。このことは幸福感調査の結果が証明している。幸福か否かは、各人の幸福についてのみかたに依存する部分が少なくない。したがって、心の問題を考えるとき、科学的、第三人称的な捉え方からは解決できない問題を、哲学でいえば現象学、心理学では質的研究、医学では臨床医学、大学院教育では専門職大学院のように、第一人称的な自己の問題として解決する事が今後重要度を増すと考えられる。

4．ロボットの心（意識）を作ってもいいか？

話を変えて、ロボットの心（意識）を作ってもいいかどうかについて考えてみよう。仮定として、ロボットの意識を作れるようになった未来世界を考えていただきたい。作れるようになったものが現象的な意識なのか機能的な意識なのかという問題があるが、ここでは、いずれであれ、外から観察したときに明らかに意識を持っているように見える場合とする。現象的な意識を持っているかどうかを演繹的に証明する事は、人間の場合であっても不可能なのであるから。また、仮定として、そのロボットの知能は人間以上であり、自由意志も兼ね備えているものとする。

そうであるとすると、よくいわれているように、そのようなロボットは人類を侵略し征服してしまうかもしれないという点が危惧される。これに対し、ひとつの解決策は、ロボット工学 3 原則を適用する事であろう。周知の通り、ロボット工学 3 原則は、SF 作家アイザック・アシモフによるもの[8]で、以下のような内容である。

- 第一条 ロボットは人間に危害を加えてはならない。また、その危険を看過することによって、人間に危害を及ぼしてはならない。
- 第二条 ロボットは人間にあたえられた命令に服従しなければならない。ただし、あたえられた命令が、第一条に反する場合は、この限りでない。
- 第三条 ロボットは、前掲第一条および第二条に反するおそれのないかぎり、自己をまもらなければならない。

これは、明らかにかつての奴隷とアナロジカルである。つまり、人間以上の知能を持ち人間と同様に意識を持つロボットに対し、奴隷のような権利の制限を加える事は、明らかに人権（ロボット権）問題をはらむ。人権や動物愛護の考え方を拡張すると、必ずや、そのようなロボット権侵害は許されない時代が来ると考えるべきである。したがって、人道的見地から、ロボット工学 3 原則は破棄されるべきであるという結論が導かれる。

しかし、だからといって、ロボット工学 3 原則を適用しないとなると、人間への危害の危惧が避けられなくなってしまう。したがって、ロボット工学 3 原則が破棄される以上は、意識を持ち、同時に高い知能や自由意志を持ったロボットは作るべきではない、という結論が導かれる。

要するに、ロボット工学 3 原則に則ったロボットも、則らないロボットも、人類のためには作るべきではないという結論が導かれるのである。

実は、私はこの人間中心主義的な結論に不満である。人間にとって不利益だから、意識を持つ高度なロボットを作るべきではない、という論調だからである。

日本産野生種のトキは絶滅してしまったが、人間という種がいなければ、彼らは絶滅を免れたであろう。彼らを絶滅させた張本人が、今度はロボットによって自らが絶滅させられることを恐れるのは、一種のエゴである。

考えてみれば、環境問題というのも人間のエゴである。「地球を守ろう」というが、守りたいのは明らかに地球ではなく人類である。温暖化が進展し人類さえ絶滅すれば、そこには恐竜の絶滅後と同様、新たな生物種が栄える事は自明である。

つまり、ロボットの問題は、奴隷問題から環境問題まで、人類の愚かな歴史を象徴的に思い起こさせてくれる題材であるという点で興味深い。

５．おわりに

ロボットの心についての議論は人間中心主義の呪縛から人類を解き放つひとつの考え方であると言い放って書き始めたこの文章であったが、ご堪能いただけたであろうか。

人間中心主義を否定する前野の思想はどこへ行くのか、という疑問を持たれた方がおられるかもしれない。これは至ってまっとうな方向である。すなわち、現代哲学の到達点、ニヒリズムである。全ての事に本質的な価値や意味はない。あるいは、釈迦のいうところの空である。すなわち、日々是好日である。

参考文献

1) 茂木健一郎、脳とクオリア、日経サイエンス社、1997 年
2) 茂木健一郎、脳は天才だ、日経ビジネス人文庫、2008 年
3) 前野隆司, ロボットの心の作り方―受動意識仮説に基づく基本概念の提案―, 日本ロボット学会誌 23 巻 1 号，2005 年，pp. 51-62
4) 前野隆司，脳はなぜ「心」を作ったのか―「私」の謎を解く受動意識仮説，筑摩書房，2004 年
5) 前野隆司，脳の中の「私」はなぜ見つからないのか？―ロボティクス研究者が見た脳と心の思想史，技術評論社，2007 年
6) 前野隆司，錯覚する脳　―「おいしい」も「痛い」も幻想だった，筑摩書房，2007 年
7) リタ・カーター、脳と意識の地形図、原書房、2003 年
8) アイザック・アシモフ，小尾芙佐訳，われはロボット，早川書房，1983 年