－金沢大学サテライト・プラザ　ミニ講演－

日　時　平成１４年２月２日（土）午後２：００～３：３０

会　場　金沢市西町教育研修館　金沢大学サテライト・プラザ

テーマ　　「ロボットに心はもてるか」

講　師　　柴田　正良　（金沢大学文学部助教授）

１．新しい科学――「認知科学」

私の所属は，ロボット工学でも，心理学でもなく，哲学です。なぜ哲学かということが少し引っかかるかと思いますが，基本的に哲学は，いろいろな驚きというか引っかかりのあるところを研究し，カバーするものです。客観的にいうと，その哲学がカバーする範囲は，科学や知識などの基礎，それから昨今問題になっているさまざまな企業倫理の問題，その基礎にある倫理学，価値といった問題，もう一つは，例えば新しい科学が出現してくるときに，その科学にとってどういうかたちの基礎になるか，何が問題なのか，何が解決なのかを議論することです。

今日は「ロボットに心はもてるか」というお話をいたしますが，ロボットの心というのは，新しい科学の出現にかかわる問題と考えることができます。それはある意味では，アインシュタインの相対性理論の最初の考えは，もちろん物理学ではあるけれども哲学的な思考に近い。そういう意味で，新しい理論，新しい分野が出てくるときには，普通のその分野の科学者でも，やっていることはかなり哲学に近いといっていいと思います。

実はその新しい科学にはまだ名前がないのですが，少なくとも今のところ，「認知科学」としてくくられているものです。その認知科学は，基本的には我々の心の現象，さまざまな思考や感情，感覚といったものをつかまえ，理解する科学として，新しく出現してきているといっていいでしょう。

２．「ロボットにも心はもてる」というスタンス

「ロボットに心はもてるか」という話をすると，大体「人間の心が何なのかよくわからないのだから，ロボットにそれを実現しようなんて無理だ」という答えが返ってきます。確かにそうなのですが，逆にいうと，ロボットに心がもてるかどうかという話は，「人間の

心が何であり，どんな働きをしているのか」という一つの問いであると考えていくと，あとの話が身近なものに感じていただけると思います。要するにロボットの心の話と，人間の心の話がぐるぐる回りの悪循環になっているのではなく，ジグザグに進むのです。そして基本的には，我々人間の心というものをどう理解したらいいかという一つの観点を提供できるのではないかということです。

もう一つ，基本的に私のスタンスは,「ロボットにも心はもてる」というものです。その立場を非常に荒っぽくいうと，我々のこの世界が，基本的には物と物との集まりでできているとすれば，いかに複雑な物であれ，それをなんとかして付け合わせて動かせば，あらゆる現象は説明できる,再現できると考えざるをえないのです。もしその線をシリアスに,真剣にとらないと,物ではない,何かしら訳のわからない浮遊物体のようなものがあって,それが何か悪さをするということになってしまうのです。私の娘が「学校の怪談」が好きなので，私も超常現象，魂の現象と言われるものを，よくテレビで見たりします。あれはそれぞれ楽しんでいる分にはいいのですが，まじめにそういうものが実際に存在し，それが何らかのやり方で力を下したり，本当に我々の普通の世界の動きに影響を与えると考えるのは，非常に困難です。そうだとすると，逆の立場を今度はまじめにとらなくてはいけない。まじめにとれば，物しかない。そういう考え方を貫徹する，一貫するしかないのではないか。これが私の基本的な立場です。この立場は全然名前がないわけではなく，短く「物理主義」といっています。これが，基本的なタイトルについての含みと，それを扱う私の主張，立場です。

３．ロボットに心はもてないと思われる理由

去年の 12 月に出た本の宣伝になりますが,今日はこの新書の中のトピックの二つ分を抜いてお話ししようかと思います。本の中で扱った問題はいろいろありますが，今日お話ししようと思うのは，そのうちの１－１と１－２にあたる部分です。それは「ロボットに心はもてないだろう」と思われる二つの理由（本当はもっとたくさんあるのですが）をよく考えて，それに対してロボットをなんとかディフェンスしてみようということです。

まず一番目の問題は,「他我－独我論問題」と書きました。要するにロボットというのは機械の塊なのだから，感覚・思考・感情を自分で経験できないと我々は考える。たぶんこれが一番ネックというか，決定的なポイントであると思われるでしょう。それをなんとかイーブンのところまでもっていきたいのです。

それから，ロボットや人工知能はプログラムどおりのことしかできないと，よくいわれます。それに対して人間というのはクリエイティブで，創造的なことや柔軟な思考ができる。そこで大きくクローズアップされた，特に人工知能の問題でかなり多くの研究者を巻き込んで議論されたものに「フレーム問題」というのがあります。これはどういう問題かというと，要するにロボットは常識をもたないということです。Deep Blue というチェスを指す名人級のプログラムがありますが，そういうチェスがすごくできるような知性はつ

くれるが,例えば 12〜13 歳ぐらいの子どもがもっているような常識を実現するのはきわめて困難です。これが今日お話する第二番目の問題です。これについては実は今のところ徹底的な解決はありません。でも，これもなんとか状況をイーブンまで,「フレーム問題でこけてしまうことにならなくていいんだ」というところまでもっていきたいのです。

４．他人が何かを感じ，考えているというのは本当か？

今，さまざまなロボットや人工の機械によって，感覚系の機能を人工的に実現することができます。甘みの検査は人間がやるよりもはるかにスピードが速くて，しかも正確に糖度を測ります。測った結果，人間が確かめてみるとなるほどそのとおりにちゃんと等級分けされている。そうするとある面では，センサーのような機械は我々の感覚器官に相当するような機能を果たしているといってもいい。しかし，モモかなんかをだあっと流して，そこでセンサーにかけて等級に分けるのですが，普通，そのセンサーが甘さを感じているとは思わないでしょう。そのように見ますと，ある種の機能を果たすようなことと，本当に何かを感じるということは違うのではないかと思うのは当然でしょう。したがって，ロボットがどんなに見かけ上は人間と同じように周りの刺激に反応しても，そのことによって感覚というものをもっている，感じているとはいえないだろうと思われます。

そこでこの状況をイーブンにもっていくために考えたいことは,「他人」です。我々はロボットに関しては，たとえそれがどんなに精巧にできていても，おそらくロボットだと知らされると，実際は感じていないんだ，何も内側から経験していないんだ，ただカシャカシャと動いているだけだと思います。しかし他人の場合は，それこそ自分の子どもや親が歯痛で苦しんでいたり，けがをしたりすると，大変だと往生したり，医者にすぐ連れていったり,慰めたりするわけです。つまり他人は何の問題もなしに,自分と同じような感覚，感情をもっていると思っているのです。

問題は「そうは言うのだけれど，それは本当ですか」ということです。自分以外の他人が歯痛で苦しんでいるのは，本当に歯が痛いのでしょうか。例えば自分の親兄弟，恋人，だれでもいいのですが，その人が「歯が痛い」と言って苦しんでいる。あるいはどこかを傷つけて,「ずきずきする」「おなかが痛い」と言っている。そのときに,「なるほど，その歯の痛みはよくわかるよ。俺もあのときそうだったから」と言いますが，文字どおりに他人の歯の痛みやおなかの痛さを感じたはずがないだろうと思います。

５．感覚の私秘性

例えば，他人の脳の中につながっている神経をビッと曲げて，自分の脳につなげることができるとしましょう。そこで他人の虫歯から痛覚神経をぐいっと伸ばして自分の脳に接続すると，それは確かにうまくいけば，その他人の虫歯が悪化すると，私が痛みを感じることになるのでしょう。しかしそのとき感じている歯の痛みは，虫歯の人の痛みであるとはいえないのです。それはそうなんだと言っても，どうしてそうかということは確認でき

ないのです。つまり我々は他人の痛みをまず知っている，それから自分の痛みというのを知っている，そうやって今度は他人の虫歯から痛覚神経をぐいっと伸ばして自分に突っ込んだならば，感じられた痛みが，「なるほど，他人の痛みと同じだ。だから今感じているのはその人の歯の痛みなんだ」というようにはならない。つまり，我々はそのように考えてみると，基本的に感覚に関しては，プライベート，私秘的であると言えます。そしてこの私秘性，プライベートなものをまじめに考えてみると，他人の場合，実際にそれを内側から，つまり自分が他人のその感情を感じるというしかたで確かめたことはいっぺんもないわけです。私以外，自分自身以外に人間は何億人といるのですが，いっぺんもない。

ちょっと考えてみるとすごく気持ち悪い話ですが，「あの映画を見て楽しかった」とか，「このチョコレートおいしい」と言って，それが滞りなくみんなの間で話が回っていくのですが，でも，自分の感じるチョコレートの甘さというのは，ほかの人も感じているのか，確かめようがないのです。ただ，甘いチョコレートと自分が思うものを他人が食べて「どう？」と聞くと，「うん，甘い」「ちょっと味がね」などと言ってくれているだけなのです。ですからこの状況は，普通は我々の日常生活の中では覆い隠されているし，それをほじくり出すと，ちょっとぎくしゃくして生活できない。そんな気持ち悪い世界で，悩みながら生きていられないので，常識で蓋をかぶせているのです。でも一歩一歩そうやって確かめられるものから順に詰めていくと，結局私の感覚は私にしかわからない。他人は一切，本当のところ何を感じているのかはわからないのです。

６．あなたの見ている色は何色か？

私は講義のときに，たまに学生と，「この色は何色ですか」「先生，白に決まっているでしょう」「何言ってるの。これは赤色ですよ」「そんなばかなことない。いくら哲学の先生だからって，そんなむちゃな話はないでしょう」「では結構。これをあなたは白と言うが，私が言う赤がまちがいだということをどうやって証明できますか」という会話をすることがあります。

いろいろな答えが返ってきます。自然科学系に入れ込んで，そういう証拠を使いたがる人は，「もしそれが赤だったならば，赤に特有の波長の電磁波が反射している。それが網膜に来ている。だから壁面に光がぶつかって乱反射して出してくる電磁波の波長を測ればいいでしょう。赤だったら，たしか 750 ナノメートルぐらいの波長。それでわかるはずです」というふうに言います。でも，それでひるんでいると哲学の先生の名折れなので，平然とした顔をして，「そんなのはどうってことないですよ。750 ナノメートルの電磁波の波長，結構でしょう。それは確かめられるでしょう。私はそのことについて，何の争いもしません。私が言っているのは，その波長の電磁波が私の網膜に届いて，しかじかの神経生理学的な変化が起きて，しかじかの脳細胞の変化が起きたときに見えるものが赤だと言っているのですよ」と答えるのです。

そうするといろいろな証拠を彼らはもってくるのです。お医者さんを連れてこようとか，

理学部の先生を連れてきて説得しようと言うのですが，そういうのは全部だめなんです。私にとってはとにかく赤にしか見えない。だからどんな証拠をもってきても，その証拠は基本的に私にとって，赤が見えるときの物理的な証拠にしかならないのです。最終的に学生はどう反論するかというと，「そんな変なことを言うのは，聞いている内容がおかしいのだ。そう言い張るのは，先生は日本語をきちんと覚えなかったのだ。我々が白だと覚えている色を，先生はまちがって赤だと思いこんでしまったのでしょう」。要するに，日本語をちゃんと使えないのではないですかという話です。つまり，我々は同じ要素を見ているが，私だけこの色のことを日本語の言語共同体から一人外れて，「赤」とよんでしまっている。だから私は色彩に関してきちんとした日本語を使っていない。本来はこの色であり，これは白であり，圧倒的大多数が本当に日本語の言語規則に則していれば，これはまさしく白になります。だから問題は，言葉遣いであって，何らそこには感覚の違い，感覚の私秘性の違いなんてないと学生は最終的には反論してくる。これは一番センスのいい学生の部類の答えです。

しかし，話はそれでは終わらない。つまり，我々はもしかすると，こういう可能性をもっているかもしれません。この白チョークを見て，だれかに「この色は白でしょう」と言われると，そちら側の人は実は赤色感覚を生じさせているのですが，この人はこの色感覚に日本語の白という言葉を結びつけているので，「もちろんです」と答える。ところがもう一人はこの白チョークを見て，実は青い色感覚を生じさせているけれども，日本語の体系の中では，この人は生まれたときからこの色感覚に白という言葉を結びつけているので，「これは白ですか」と聞かれると，「もちろん白です」と答えます。もしこういう状況が我々の中で出現しているとすると，どうなるか。それは先程言ったように，色彩感覚はお互いにまるっきりばらばらだけれど，日本語の言葉のやりとりのレベルではきれいにうまくつじつまが合う。同じものに関して白だと言われれば，「もちろんそうです」と言いますが，感じている中身に関しては全く違っているということが論理的にありうるので，その可能性を完全に排除することはできない。これが感覚の私秘性といわれている論点です。

この青感覚を生じさせている人の脳の中をかち割って，脳細胞をどんなに見ても，青の感覚，赤の感覚のかけらがあるわけではないので確かめようがないのです。青の色感覚を生じさせているときの脳細胞と，こちら側の赤の感覚を生じさせているときの脳細胞とでは，おそらく２つの神経レベルの活動タイプが違うだろうと期待したいのですが，仮にそうであったとしても，この段階では何の助けにもならないのです。そこからどういう色感覚を生じさせているかを確かめることにはなりません。

少し考えると，こうした状況の中に我々は生きていることになります。ちょっと気持ちの悪い話ですけど，しかたがないですね。

７．＜内側から確かめる＞

そこで，「内側から確かめる」という基準をもう一度真剣に考えてみましょう。実際に何

を感じているかということを，我々が自分の周りで確かめるというやり方です。例えば他人が青色が見えると言うとき，本当にそういうふうに見えているのかどうか内側から確かめるという基準を他人に適用すると，やはり確かめることはできないことになる。とするならば，ロボットと他人は同じく「確かめ不可能」ということになります。

ここで二つの選択肢があります。一つは哲学史上有名な「独我論」というものです。それは基本的には，感覚が何であるか，青色の感覚が何であるか，考えるとはどういうことであるか，悲しいという感情は何であるかというのは，自分にしかわからない。自分だけにしか生じえない。ほかの人についてそんなものが生じているなんて，実はうそっぽい。それぞれが「私だけ」と言うわけですが，もちろんこの私にとっては「この私だけ」という立場になります。実際にほんのわずかの時期だけ，この立場をとった人もいます。でも独我論というのは，ちょっと考えてみると，実際に維持するのは困難です。通常の常識的な，生きているという普通の場面では，どう考えても他人も私と同じように，同じ状況に陥ったら同じことを考えるし，同じダメージを受けたら同じような痛みを感じると考えたいわけです。この独我論は論理的には整合的ですから，論理的に破綻を見つけてやっつけるというわけにはいきませんが，これをまじめな選択肢として考えることはまずできません。

そうするともう一つの選択肢は,「ではリベラリズムをとりましょう。ロボットも他人も同じだ」ということになります。基本的に独我論をとりたくないとするならば，内側から確かめるという点に関していうと，Ｃ３ＰＯ（スターウォーズに出てくるロボット）もあなたも同じだということにならざるをえないのです。

８．科学で解決しないか？

内側から確かめるのは無理だ，そんなことを考えてもしょうがない。では，外側からはどうか。基本的には科学で解決するという方法があるわけです，自然科学の分野で。つまり，神経生理学や脳科学の成果を使えば，人間には感覚は実際に生じるがロボットには生じていない，ときちんと説明できるのではないかと思われるでしょう。つまり，科学が外側から説明してくれるのではないかと。実際，神経生理学と脳科学は，例えば麻酔をうまくやることに関してきわめて成功しています。向精神薬といったものを開発することによって，我々の一種の精神疾患をかなり制御する道を開いてくれました。つまり科学は我々の主観的な経験がどんな脳のメカニズムと化学的な要素によって決定されているかを，きちんと明らかにしてくれているのだから，他人の感覚に関してのことも，自分と同じ状態だということを根拠に，麻酔と同様のテクニックを使ってやれるだろう。そう考えてみれば，科学は，同じ脳の状態には同じ感覚が宿るということを説明してくれるし，逆にコンピュータには感覚が生じていないことを外側から説明してくれるのではないか。そのように考えられるかもしれません。

９．説明抜きの＜前提＞

最初に脳の状態についていろいろ研究したとしましょう。それがどんな感覚を生じさせて，どういう心の状態を生み出しているかを，脳の状態によって説明するというように事態が進んでいくとしましょう。でもそのときに最初の段階で，「痛いですか」「痛くないですか」「しびれてますか」「どんなふうに見えますか」といった言語を使って明らかにされるような感覚と，脳状態・身体の状態とのそもそもの対応関係をまず説明抜きで受け入れざるをえないのです。科学がやるのは，そういう説明抜きの前提をとったあとで，それから結果するような対応関係のさまざまなバリエーション（知識）を蓄積することになるでしょう。そういう対応関係というものが，主観的な経験とのつきあわせなしで，何か科学（生理学や医学）の外側からの研究だけで明らかになるという単純な構図ではなく，むしろ最初にどのような感覚的な経験・主観的な経験と脳状態が対応するかを決定しなければならない。ある基礎的な部分では，その対応関係を前提せざるをえない。その時点で対応関係の説明なんてできない。むしろ説明はそれ以後の事態，それより複雑な，あるいはそれの組み合わせによって出てくるような事態を，前提とした対応関係のバリエーションで説明するというようになるでしょう。

ですからこの点では，我々の感覚はもしかすると先程のように白いチョークを見ながら片方の人は青，片方は赤の感覚を生じさせているというグロテスクな状況かもしれないということを，科学の成果によっては排除できないのです。

10.「完全なる医学」？

仮に完全なる医学と称されるものに我々が達したとしましょう。例えば痛みについては脳の中の状態が 47 タイプあって，このどれかが発生すると，我々が痛みを感じる。人間が感じる痛みというのは基本的にそれで全部だとします。だけどある人がいて，その人が何かずきずきする頭痛を感じるのでお医者さんに行った。すると 47 タイプ全部を検査してくれるのですが，どれにも入らなかったとします。さて，どうなるか。我々の医学は完全なのか。「47 タイプに入っていないということは，あなたの痛みは痛みじゃないんですよ。それは痛いと思いこんでいるだけ。家へ帰っておとなしく寝ていなさい」なんて言われるかもしれません。47 タイプのどれにも入らない。うそをついているのでもない。だから，あなたの痛みというのは単に痛いと思っているだけで，本当は痛くないのですよと言われるとします。あなたはそのとき，どうするか。「だけど痛いものは痛いんです」と言うでしょう。普通の健全な人ならば，痛いのを我慢しようがないというのはどういうことか。それは完全なる医学というものが，実は完全な域にまで達していなかったということです。つまり，たった一例でもいいのだけど，48 タイプ目の痛みというものがあって，それが私の中で実現している脳の状態なのです。だからそれは「お医者さん，ちゃんとそういう説明を完成させなさい」ということになるのです。

11. ＜事実＞の拡張――心の多重実現可能性

例えば痛みの実現に関してはこれだけのタイプがこれまで知られていたが，もう一つ，48 番目のものによって痛みが実現されたことになります。つまり，これを痛みというものの多重実現と考えると，痛みという感覚状態を実現するにはさまざまな仕掛け，さまざまな素材，さまざまな状態によって可能だと考えられます。さまざまといっても，せいぜい同じ人間の脳の中での神経系の状態ということですから，範囲は狭いですが，それでも 47 タイプある。ところが 48 タイプ目というのを，ある場合には認めざるをえなくなってくるのです。つまり，医学の完全性といっても，どうしても痛みを感じる 48 タイプ目が出現してきたら，痛みに関する医学用語，神経性医学理論を修正して，痛みというものはこういう場合にも生じますと，事実を拡張せざるをえない。それを我々も承認せざるをえない。あらかじめハードの部分についての知識だけがあって，痛みというものはこの可能性にしか宿りませんとカットすることはできないのです。48 番目の可能性が開かれていると言わざるをえない。

こう考えてくると，科学が扱っている（説明している）事実とは，こういう 47 タイプに関してこれがどのような痛みを引き起こすかについて，それぞれ神経生理学的なメカニズムがどうなっているかについての話です。しかし，もし 48 タイプ目の神経生理学的な状態において，痛みの実現があるとすれば，科学が扱ういわゆる事実というものは，そういう意味で拡張される可能性がある。つまり，痛みは 48 番目の脳神経でのタイプによっても実現される。同じように，人間の心も，人間の脳以外のかたちで実現される可能性がある。それが心の多重実現可能性です。つまり，他人の感覚に関しての内側からの確認不可能性というお話をもう少し突っ込んでいくと，他人の感覚というものはロボットの感覚に拡大されたときに，ちょうど「脳の中での出来事が他人の感覚を生み出している」という説明と同様なしかたで，「ロボットの中の内部装置（物理的な装置）というものがロボットなりの感覚や思考を生み出している」と，事実を拡張していく方向に話がいかざるをえない。つまりここのところは，他人を通過することによって，ごく自然なかたちでロボットにも心の実現可能性を認めざるをえなくなると考えていくことができるのです。

もしもロボットがいきなり我々の身近にいるのが考えにくいのであれば，我々人間とロボットとの間にエイリアン（宇宙人）みたいなものを考えます。宇宙人に我々が遭遇して，交渉する。宇宙語を翻訳して，いろいろなやり方で彼らと話をする。そのときに，その宇宙人の脳にあたる部分の組織というのは，相当人間と違うかもしれない。我々の脳とつくりがかなり違うからといって，「宇宙人に心はない」と言うよりは，宇宙人の心というのは我々と違った脳組織で実現されていると考えるでしょう。そしてもう少し拡大すると，ロボットというのはかなりハードなメカニズムで，ウェットな，脳のようなかたちでつくられていないし，進化の歴史も持っていない。でも，やはり心というものを実現する新しい物理的な素材，あるいは装置であるといえるのではないでしょうか。

これが心の多重実現可能性の基本的なアイデアです。要するに，「新しいタイプの脳状態

によってあなたの痛みが出現している」と，医学（科学）がカバーする対応関係の事実が拡張されうる。同じように，「新しいタイプの物理的装置をつくってやったら，感覚・思考というものが実現された」というように，これもまた今度はロボット工学と認知科学がカバーする事実というものが出現したと考えられます。

12. 端的に無視する

私の立場は物理主義と言いましたが，物理主義は今のような感覚的なものと感覚の物理的な基盤とをどう考えているかというと，基本的には同じ物理的な組織や状態が出現したら，その上には同じ感覚状態が出現するということを認めます。ただし，その感覚は違った物理的な状態によっても出現するかもしれない。それは分からない。しかし，同じものが実現されたら同じ感覚が実現されると主張します。ただし，これも前提であって，証明ができるような話ではない。そう見なしたい，見なすことを受け入れてくださいということです。

さて第二番目は，常識についての問題です。ロボットが感ずるというのをステータスとしてはイーブンになるところまでもっていきたいというのが第一番目の問題とすると，第二番目の問題は，ロボットもなんとか常識的な思考ができるのではないか，逆にいうとロボットが人間並みの知性をもとうとするときの最大の課題の一つであるフレーム問題をどうやって考えたらいいか，ということです。それは，端的に無視するという問題なのです。これを我々はやっているのですが，これをいかにロボットにさせることができるか。

フレーム問題にはいろいろな言い方がありますが，私のバージョンでは，「何を考えなくてもいいか」ということを考えずに，考えなくてもいいことをいかに考えないですますかという問題です。わざと難しく言っているわけではなくて，わりと簡単な話です。フレーム問題とは何かということを理解するのがかなりのポイントで，いったん理解されてしまえばどうそれを乗り越えるかというのは，大ざっぱなかたちではあきらめがつきます。あきらめがつくという意味は最後の話です。

13. 常識に悩むロボット

ロボットは三つの段階でフレーム問題に苦しんでいるのです。資料では，感覚の部分とフレーム問題のちょうど山場のところだけをコピーしてお渡ししてあります。あとでもし気になりましたら，お読みになってください。

フレーム問題が何であるかを比較的うまく説明してくれたデネットという人がいます。彼の話を使って説明してみましょう。

話の状況ですが，まず部屋の中にバッテリーが置いてあります。バッテリーはロボットにとって非常に貴重なので，それを救い出す作戦を立てるわけです。問題は時限爆弾も一緒に置いてあることです。さて，このロボットはバッテリー救出作戦を立てます。バッテリーはワゴンの上に置いてある，どうすればいいか。ロボットは３台登場します。最初の

ロボットは，このバッテリーを救い出すためにはワゴンを引っ張ればいい。簡単だと言って，最初のＲ１ロボットはバッテリーを救うために引っ張る。だけど，ご覧のように時限爆弾もワゴンの上にあるので当然それも一緒に引っ張ってしまうのです。Ｒ１はそれを引っ張って途中まできて，ボンと爆発する。これを見た設計者は，これはだめだというので改良した二台目のＲ２を作ります。

Ｒ１ロボットの欠点は非常にはっきりしています。自分の行為の副次的な結果を計算していないことです。爆弾はワゴンの上にあることを知っているのに，ワゴンを引いたならば爆弾も一緒に引いてしまうのだということを計算しない。つまり，自分のワゴンを引くという行為の結果,副次的な結果が生じます。それを気にとめていないからだめなのです。自分がやろうとした，ワゴンを引いてバッテリーを救うことについてはもちろんわかっているけれども，その結果何が生じるかを計算しなくてはいけない。

だから次の改良ロボットＲ２がやるべきことは，はっきりしています。自分の行為の意図せざる結果，何が起きてしまうかということをきちんと計算するロボットを作らなくてはいけません。そこでＲ２を作って，やはりこの状況に立たせる。Ｒ２はどうするか。「うーん」と考えます。計算します。自分がワゴンを引くとどうなるか。まず，ワゴンの車が回転する。よし。二番目，ワゴンを引く。壁の色はどうなるか。変わらない。よし。三番目，ワゴンを引く。ワゴンの大きさはどうなるか。変わらない。よし。と，次々に計算をします。そうやって計算を果てしなくしているわけですから，665 番目の証明か何かをやって，ワゴンを引いても部屋の大きさ全体は変わらないとか何か計算しているときに，時間が来て爆発する。二番目のロボットＲ２もアホだということになるのです。

設計者はもうちょっと賢い三番目のＲ３を作らなくてはいけません。二番目の問題点は，やみくもに計算することですね。何でもかんでもとにかく，ワゴンを引いたら何が変わるとか変わらないとか，あらゆることを計算してしまうのはこの場合には意味がない。関連のあるものだけを計算しなくてはいけない。関連のあるものとないものをこの課題に関してきちんと分類できるような仕掛けをつくります。それがＲ３ロボットです。Ｒ３ロボットは今問題になっていることに関連のあること，ないことをきちんと分類して対処できるような分別のある論理ロボットです。設計者たちは，今度こそうまくバッテリーを救い出してくれるだろうと，この状況にＲ３ロボットを立たせます。

そこで，Ｒ３ロボットはどうしたか。中に入らないで，何もしないで「うーん」と悩んでいる。設計者はいらいらして「何をやっているんだ，早くしろ」と怒ります。Ｒ３ロボットは「邪魔しないでください。今，私は結果を関係のあるものと関係のないものに振り分けています。そして関係のない帰結を無視することに忙しいのです。ワゴンを引っ張る。そうすると部屋の電気の明るさは変わらない。それは関係がない。その次。ワゴンを引く。車輪が回転する。それも今の状況には関係がない」と計算していく。そのうちに，やはり時間切れでボンと爆発してしまう。

14. 明示的な規則と計算

これがデネットがつくった話の元バージョンです。一番目のロボットは，行為の副次的結果を計算しないからワゴンを引いて爆弾も引いてしまうことを計算できない。二番目のロボットは副次的結果を計算しすぎてしまう。三番目のロボットは，関連性のあるものだけを計算をするようにしたが,関連性のあるなしを逐一チェックしなければ確認できない。

これの教訓はどこにあるでしょう。現在の古典的なかたちの人工知能は，基本的には計算や判断，思考などを全部明示的にきちんと決められた規則と計算によってします。そうするとこのような状況が出てきたときに，あらゆる状況に対してあらゆる可能な出来事を決めてくれるような規則をもっていて，その規則の例外というものを全部決めるために，それを逐一当てはめるということになるのですが，実は我々の場合にはそういうことはやっていないのです。我々はなぜかわからないけれども，何が関連があるかということをわかってしまうわけです。時限爆弾が乗っている。ワゴンを引く。「ああ，それはやばい」。爆弾も引いてしまうことがわかるわけです。だけど，そのことは別の状況では気にしなくてもいいかもしれません。別の状況では別のことが重要になって，別のことが関連性を帯びてくる。そういうものを我々は逐一規則のかたちで与えられてはいない。つまりこの状況にはこの規則，そしてこの新しい要素が出てきたので，こういう例外があるというように，それをちゃんと自分のうちで計算しているのではないわけです。

つまりいろいろな状況が起きてきたときに，基本的に我々人間が対処するやり方というのは,このような明示的規則と計算というやり方とどうも違うのではないか。したがって，そういうものに知性のすべてを還元することはできないということになります。それでは何が一番問題かというとこうなります。このロボットは最後のところで関連性を問題にしますが，それは非常にいいわけです。これをやらなかったら，我々は全部いちいち網羅的にあらゆる知識を参照しなくてはいけません。ドアを開けて足を一歩踏み出す前に，百科事典的な知識を全部チェックして,「大丈夫だ」となります。だから，関連性のないものに関しては，例えば地球の重力は今これをやっても変わらないぞ，大気中の酸素の成分は一日二日で変わらないぞと，いちいちチェックしなくても我々は動けなくてはいけません。関連性は非常に大事なのですが，何が関連があって，何が関連のない要素であるかということは，残念ながらそのときどきの状況に依存するのです。どの状況が出現してくるか。時限爆弾であるという典型的な状況から出発しても，その爆弾のある位置がどこか，ワゴンの上か，ワゴンの外か，部屋の隅か，ということが関連してくるわけで，典型的な状況というものだけで我々の現実の生活はできていない。そして関連性と例外との間の，また関連性のあるものとないものとの間の区別というものを，一つの完璧な規則によって全部網羅的に決めておくことはできない。こういう状況の中に我々は生きているのです。

人間は典型的な規則から出発したときに，一応，典型例に関して規則を与えます。それはいいでしょう。では，それが少しずれて，ときどきの状況に対して発生する例外はどうするかというときに,「ほかの事情が等しいなら，ceteris paribus（ラテン語)」という条

項（科学哲学で非常に有名な条項）で対処します。これは，典型的な状況に関しては規則を与える，その規則をほかの状況に適用するときに，ほかの事情が等しいならば，その規則でよろしいですという言い方をしています。それ以外に関しては何も言っていない，何も言えないのです。典型的な状況でつくり上げた規則を，具体的ないろいろな状況に適用するときに,ほかの事情が等しいならそれを適用していいですと言っているだけなのです。

では具体的にほかの事情とは何でしょう。言えないのです。言えないところが御利益なのです。言わなくてもいいのです。いいのだけれど，これを一つ入れておくことが大事なのです。つまり，我々人間の場合には，ある典型的な状況があり，規則をそれに合わせてつくる。その規則を個々の具体的な状況に適用するときに,もう一つ規則というかたちで,例外用の規則を次々につくったりはしません。それでは何も進んでいないわけです。こういう「ほかの事情が等しいなら」という，全く内容を詰めなくていいようなもので理解しておいて，具体的に対応できてしまう。このメカニズムがわからない。ロボットがいかにしてceteris paribusのようなものを身につけることができるか。これをプログラムに書くわけにはいかないのです。具体的にどの状況に対してどうするかを規定してやらないかぎり，ロボットは動きようがないわけです。

15. データ量の爆発か，計算量の爆発か

要するに典型例とその規則から出発すると，問題になってくるのは，実際にずれてくるときにどうやってフォローするか,柔軟に対応するかです。「アイスクリームを買ってきて」と言われて，そのロボットがスーパーマーケットに行くと，何時間もかかって，とんでもない状況になっていても，そのロボットはとにかくアイスクリームを買いたくて，アイスクリームを探して・・・(笑)。そんなロボットはどういうやつかということになりますね。常識という，そのときどきに応じて何が重要で何が重要でないかということを判定するものが必要です。そうすると例外にいかに対処するか。結局，今までのようなやり方をすると，対処のしかたは二つしかありません。

一つは，例外を覚える。例外にはおそらくまた例外が出てくるでしょう。そのまた例外も当然出てくるでしょう。そういうものをすべて覚え込むというやり方です。そうするととんでもないデータの量が幾何級数的に増えていきます。そんなデータをたくさん詰め込まなければいけないことになります。

それでは，計算することにしましょう。まず，基本的な規則があって，どんなときに例外が出てくるか，その例外をどのようにするかを全部棒暗記にするのではなく，計算で出すようにするわけです。つまり例外の計算，例外の例外の計算，というようにして計算を続けていく。こうなると結局，データ量が爆発していくか，計算量が爆発していくかのどちらかになってしまうのです。

16. 古典的ＡＩでは，フレーム問題は解決できない

この二つは，あちらを立てればこちらが立たずというかたちで，トレードオフといわれるような関係になるのです。基本的には，これはそのままのかたちでは解決できないというのが，フレーム問題の教訓になります。古典的ＡＩ（人工知能）では，フレーム問題は解決できないということです。要するに古典的な人工知能のやり方，つまり我々が今お世話になっている市販の普通のコンピュータを動かしているところの原理を拡張することによっては，このフレーム問題は解決できない。

解決できるように見えている部分は，実は非常に狭い領域での問題で，先程お話ししたチェスの Deep Blue のような能力です。あれはチェスだけの世界です。その世界で起きる出来事というのは，基本的にはチェスを指しての可能性で，それもまたすごい数ですが，その中でだけ生ずることであって，それ以外の部分は無視していいのです。しかし本来，我々がチェスを指しているときに起きるような普通の出来事，つまり，猫がその上を走って駒が乱れたとか，やっている最中に「もういいかげんに寝なさい」とお母さんに怒られたとか，全くチェスと関係ない，しかしチェスが具体的な世界の中にはめ込まれているがゆえに生じるさまざまな事実というのがあるのです。Deep Blue のようなプログラムは，それを全部無視してかまわないのです。無視して，チェスの世界だけに自分を特化する。

したがって，古典的なＡＩは実は「一能型コンピュータ」だというのが一番いいのです。それに対して，我々が「通常常識的な知性」とよぶのは，本当の意味での万能型の知性です。この一能型の知性では，フレーム問題は解決できない。フレーム問題ではあらゆる規則が，規則の例外を生じさせ，その例外がさらに例外を，しかも世界のほかのあらゆる要素との関連で，それを生じさせるという状況になっているのです。

つまり，一能型コンピュータが扱っているような世界で起きる出来事は，規則によって完全に仕切られて，かすかすの世界です。だから何が可能的な出来事として起き，何が起きないのかということは原理的にいえば全部わかってしまうのです。でも，現実の世界では実際にチェスを行っているその場面というのは，ほかのさまざまな事情の中に埋め込まれているのです。隣のテレビではそれこそ雪印か何かの問題を放映している。そして今自分は夕食の前であって，だれかが夕食を作ってくれるが，もしかすると突然，電話がかかってきて親戚のだれだれちゃんが大学に受かったので今からお祝いに行くことになるかもしれない。そういうものに関しての柔軟な対応をするのは，この一能型ではできないというふうになるわけです。

これはある意味では本質的です。というのは，記号と明示的な規則で全部やると，どうしても典型例とそれの例外を同じしかたで扱わざるをえないのです。同じしかたで扱うということは,分量がどんどん増えます。したがってこれをなんとか突破していくためには，別なやり方を考えた方がいいということになります。

仮に今の古典的な人工知能の原理だけを使ってロボットを作ろうとすると，フレーム問題を解決することはほぼ不可能だといっていいです。古典的な計算主義を強力にバックア

ップしているジェリー・フォーダーという認知科学の哲学者がいますが，彼も基本的にはこの問題を今のところどうやって突破していいかわからない。だれか天才が出てきて，新しい計算の仕組み，メカニズムを示してくれるまではどうしようもないと言っています。

しかし全くお手上げというわけではありません。そのジェリー・フォーダーは古典的な計算主義の方ですが，それと対立するもう一つの人工知能の考え方があります。その考え方やメカニズム，それがもっている能力を見ると，いかにもフレーム問題をやすやすと乗り越えています。そういう人工知能の作動原理があります。

もう一つは，我々自身がこういうフレーム問題を本当に解決しているのかと考えると，実は人間も解決できそうもないのです。完全なかたちでフレーム問題を解決するというのは奇妙な話で，あらゆる逸脱状況にきわめて適切に対応するのは，我々もやれていない。かなり突発的な事柄に関してはどんくさくて，しょっちゅう不適切な反応をします。ただ，我々は先程のロボットほどにはどんくさくなく，スムーズに物事に対処できるという程度です。どういうことかというと，我々の知的な能力というものの成り立ちを考えると，贅沢品としてできたという部分もあとになってあるかもしれませんが，基本的にはこの世界で自然淘汰を生き抜くための道具として，発展してきた。つまり，知性があるということは，生きるうえで，生き抜くうえで有利である。そう考えると，それが持っている力というものが，進化の状況の中でかなり有効なしかたで働くだろうと思われるのです。

その状況をもう一度考えてみると，ある時間内で正確な情報を処理する。つまり，進化の圧力がかかってくる。生き延びなくてはいけないというときに，知性が果たすべき二つの任務は相反する任務であって，一つは時間であり，いかに速く情報を処理するかです。もう一つは情報処理の正確さです。つまり，目の前にぼやっと見えているものが本当にトラであるのか，それともただ草むらが風に揺れているだけなのかについての，正確な情報がその生物体の生き死にを決めるのです。また，それを計算するのにぐずぐずしていてはだめなので，ぱっと判断して，トラだと思ったらぱっと逃げるようなことができなければいけません。本当にトラかどうかわからないけれど，もう少しデータを集めてみましょうとやっていると，食べられることになります。

そうするとこの時間と正確さというのは，トレードオフの関係で，どちらを優先させるかがそのつどせめぎ合うことになります。最終的には，ではどちらをやるのかと考えると，やはり適切な時間内に行動を起こす。この時間のプレッシャーを跳ね返す方です。これが知性の根本的な存在状況で，それがクリアできなければ知性は知性としての力を発揮できないと思われます。そういうことを考えると，我々の場合にフレーム問題をなんとかやわらげてくれる，あるいはそこそこ解決してくれているというのは，ある意味で合理的なのですが，実際には細かい計算ではない，つまり計算ということをやっていない，そういう情報処理というものだろうと思われます。それでは，合理的でまともなのだが，計算をやっていないという情報処理が実際にあるかというと，一つだけ可能性があって，それが実は感情です。

17. 新たなＡＩと感情の力

感情は実は非合理的なものなので（笑），むしろ感情に引きずられるからこそ，我々はまともな判断を狂わせられるのではないかと言われるかもしれません。しかし，実は感情というようなものは，もしかすると認知のモジュールといわれる機能があって，それが相互にぶつかり合うときに働く。例えば時間を気にするところの要素と正確さを気にかける要素の二つがぶつかり合ったときに相互調整する，感情がそういう機能として働いてきたという可能性は十分にあります。つまり我々がぐずぐずといつまでも計算をしているわけではない。ある時間内で決断しなくてはいけない。しかしその決断が全く場当たり的なものであったとすれば，進化の過程で淘汰されてしまうので，そこそこに合理的でないとだめです。野生の合理性をもっているが，計算をちまちまやっているのではないというタイプの情報処理として，感情というものを考えることができます。

すると一つの可能性としては，人間の場合に感情というメカニズムをフレーム問題を突破する，やわらげるのにうまく使っているとすれば，これをなんとかロボットに組み入れたい。もっとも，それこそが問題で（笑），ロボットに感情をもたせるということがまさしく問題となるだろうと思うのですが，それはそれなりのしかたでやれるだろうといいますか，今後のＡＩの興味のあるところです。

コネクショニズムという新しいタイプの計算のやり方は，古典的なやり方と真っ向から対立しています。基本的には我々のニューロン（神経細胞）の相互のつながり方をモデルにして作ったのが，コネクショニスト・モデルの計算機です。その計算機は非常に単純な計算しか行いません。プログラムがないのです。古典的な人工知能での記号にあたるものがありません。規則にあたるプログラムもありません。要するに実物を入れて，どんどん訓練すると，学習してしまうのです。それが気味が悪いというか，奇怪だというか，その原理がわからないのです。とにかくコネクショニスト・モデルというのは，かたちだけ書くとこうなって，ここでいくつかの計算を行うのです。こうつながっていて，入力をするとカシャカシャと非常に単純な計算をしてここに出力する。こういうものだけなのです。こういう計算機で例えば，ここに人間の顔写真を置いて，非常にたくさんの段階の訓練が必要ですが，その刺激とその人の顔と名前を記号化してうまく認識させるようにすると，コネクショニスト・ネットワークは，その人の写真を見せると，それがだれの写真であるかをちゃんと出力するようになります。これが非常に不可思議なところです。

仮にコネクショニスト・モデルのこういうやり方で，我々の認知機能の重要な部分が行われているとすると，それを人工知能の中に組み込んでやります。このコネクショニスト・モデルというのは，明示的な規則や記号を使わないという点で，ほぼフレーム問題で障害になってきたような事柄を，うまく自然に解決してしまっているように見えます。ここに入れる情報が多少ずれていると，出力の方も多少ずれて出します。つまり，例外が現れると，それなりに対応してしまうのです。それから一般的な傾向やデータ，日本人なら日本人のデータをたくさん入れてやると，日本人の顔の識別が非常にうまいネットワークが出

来上がる。それは規則をまず与えて,規則によって記号を操作するというやり方ではなく,むしろ刺激と,刺激に対する反応のしかた,刺激の処理のしかたを訓練することによって,やわらかな対応ができるようになるのです。

今日の私の使命は，ロボットは心をもてないだろうという難点を，とにかくイーブンにまでもっていくことでしたが,感情についてのメカニズムと,その種のコネクショニスト・モデルふうの新しい計算装置を組み込むことによって，フレーム問題をなんとか乗り越えることに関しては希望がもてるということです。

質疑応答

（Q）　人間がロボット以下になりつつあると，話を聞きながら思いました。時間をかけて正確な判断ができないと。何か人間がコンピュータ，ロボットを使うことによってロボット以下になっていくのではないかなという気がしたという感想です。

（Q）　最近，利根川先生が書かれた脳科学の新書が出ていますが，必ず脳を解明できると言いきっています。そうすると，そういったものと科学の分野で，もし解明するとすれば，それはロボットの中に組み込む感情の中に入りますか。

（柴田）　それは入ると思います。ただし，今のところ脳科学の人たちと我々は，感情は化学物質なのだと見ています。我々の情報処理を非常にウェットで，体液循環でやる部分があって，それに非常にうまく乗ったものの一部として感情がある。そうすると，感情が果たす機能というのは，かなり突き止められると思うのです。怒り，喜び，それらがどんな機能を果たしているかがわかる。だけど，それを実現しようとすると，人間の場合はこういう素材なので，脳のアドレナリンというもので神経伝達の相互調整をこのようにやりましょうというかたちになると思うのです。ところが今我々がつくろうとしている工学型ロボットの方はウェットな素材を使って情報処理をしていません。つまり今のところ，化学物質が入り込む余地がないのです。素材が違う，つくりが違う。ですから，感情のそこのところは素材の論理というか，脳や，我々人間の素材に依存する部分です。そのメカニズムをすぐロボットの方に応用するのはかなり難しいと思うのです。ただ，感情がどんな機能を果たしているかというのは，こちら（ロボット）に移してやることができます。例えば悲しいという感情が果たしている機能が実はあって，それをロボットにももたせることはなんとかできるが，実現するためのメカニズムは人間と相当違うことになっている。こういうことではないかと思います。

ある認知科学者によれば，感情は，環境に対する適応についての生理学的な解決，つまり環境に対する適応を計算して解決するのではなくて，生理学的に解決するという役割を

果たしています。私はそれはかなり正しいと思います。機能だけをこちら（ロボット）にもってくると，ハードの部分がかなり違うので，どういう実現のメカニズムになるかは，それこそ将来的なロボット工学の課題かなと思います。

（Q）　哲学者たちというのは心は無形というか，そういうものをどうとらえておられるのでしょうか。コンセンサスみたいなものがあるのですか。

（柴田）　ないです。ないですが，基本的には常識的な理解が出発点です。心は我々の能力から分ければ，例えば論理的，感覚的，感情的というものに分けていけると思います。ただ，出発点は常識的な「心」と我々が通常理解しているものです。「心って何ですか」と言われると答えに困って，「あなたが心という言葉で理解しているものを，私も理解しているのですよ」と言うしかないです。要するにそういう常識的なものです。あとは認知科学とか，心の哲学とかでくくると，心のある部分をもう少しモデル化して，残りの部分を話の場面に応じて実は「もっている」「もっているけど，もっていない」などと辻褄を合わせて，それなりに整合的なかたちをつけていくことはできます。しかし哲学や科学には，そのそれぞれのパッチワーク以上の芸がない（笑）。コンセンサスに至ってはなおさらです。

1. ロボットに心は持てない

と思われる2つの理由（もっとあるけど）

1-1. 他我一独我論問題

感覚・思考・感情を自分で経験できない

1-2. フレーム問題

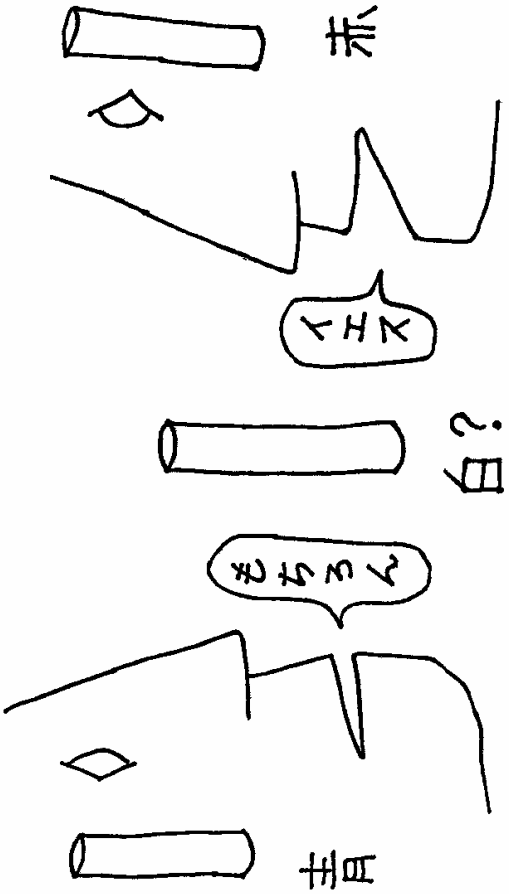
プログラム以外の創造的なことができない

2. 他人が何かを感じ、考えているというのは本当か？

3. 感覚の私秘性

他人の場合、それをどうやって確かめたのか

4. あなたの見ている色は何色か？



5. ＜内側から確かめる＞

という基準を他人にも適用するなら、ロボットと他人は同じく＜確かめ不可能＞だ。

6. 独我論

じゃ、いっそ、ロボットも他人もすべて、感覚や思考を経験していないのさ。

7. ロボットも他人も同じだ

独我論の砦に立てこもりたくないなら、＜内側から確かめる＞という点では、C3PO もあなたも同じだ。

8. 科学で解決しないか？

しかし、神経生理学や脳科学といった科学の成果は、同じ脳の状態には同じ感覚が宿り、逆にコンピュータには感覚が生じないことを＜外側から＞説明してくれるのではないか？

9. 説明抜きの＜前提＞

として、科学は、感覚と脳状態とのそもそもの対応関係を受け入れる。その＜前提＞が初めて、それ以後の科学的探求と説明を可能にする。

10. 「完全なる医学」？

と称されるものによって、あなたの痛みが説明されなかったらどうするか？

11. ＜事実＞の拡張 心の多重実現可能性

「新しいタイプの脳状態」→「あなたの痛み」、という＜事実＞の拡張。

「新しいタイプの物理的装置（ロボット）」→「感覚・思考」、という＜事実＞の拡張可能性。

12. 端的に無視する

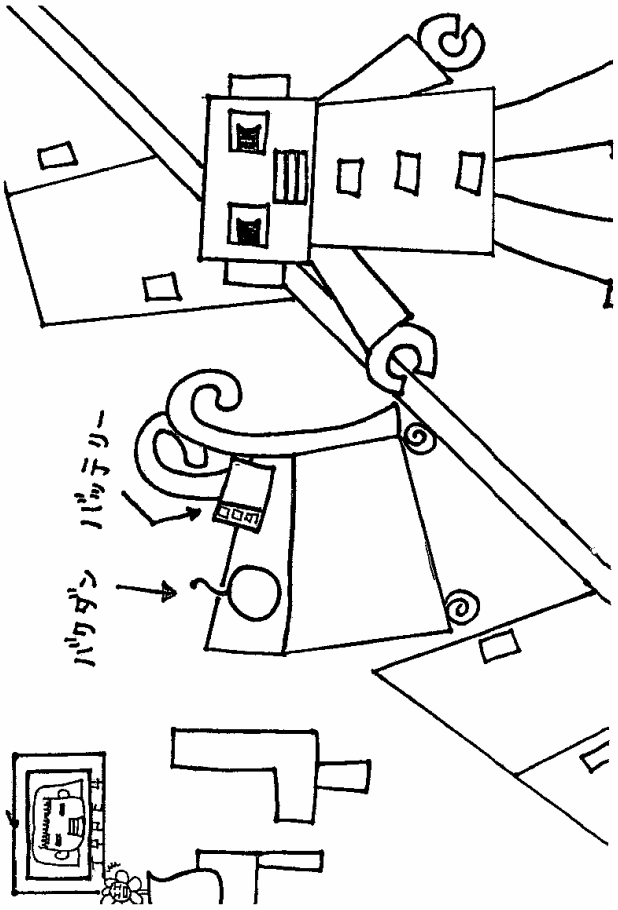
フレーム問題・・・＜何を考えなくてもいいか＞ということを考えずに、考えなくてもいいことをいかに考えないですますか

13. 常識に悩むロボット

13-1. 行為の副次的結果を計算しないロボット
13-2. 行為の副次的結果を計算しすぎるロボット
13-3. 行為の副次的結果の関連性を計算しすぎるロボット

14. 明示的な規則と計算

に知性のすべてを還元することはできない。人間は、例外に対して、「他の事情が等しいなら ceteris paribus」という条項で対処する。しかし、ロボットはいかにして？



15. データ量の爆発か、計算量の爆発か

例外、例外の例外、そのまた例外・・・のすべてを覚えるか？　それとも例外の計算、例外の例外の計算、そのまた例外の計算・・・をすべての項目に対して実行するか？

16. 古典的 AI では、フレーム問題は解決できない

離散的な記号と明示的な規則（プログラム）による計算では、人間の持つ柔らかな常識的判断が実現できない。

17. 新たな AI と感情の力

ニューロンに似た計算の仕組み（コネクショニスト・モデル）の可能性

認知機能のモジュール相互を調整する、第2階の機能としての感情

他我－独我論問題

1　あなたに見える色は本当は何色か？

いま、あなたは、この本のこのページを読んでいる。何はともあれ、白いページに黒い活字が見えるはずだ。しかし、「白いページに黒い活字」というのは本当だろうか？　実は、本当ではない。実際は、この本は特別仕立てで、いままでのすべてのページはうすい緑色であり、その上にうすい赤色で活字が印刷されている。だから本当は、「緑色のページに赤い活字」なのだ。あなたが「白いページに黒い活字」と思いこんだのは、心理学で有名な恒常性に関する法則のせいである。ウソだと思うなら、左目を閉じて、右目だけでおよそ10センチの距離からこのページをもう一度よく眺めてほしい。緑色のページに赤い活字で印刷されていることがよく分かるだろう。・・・えっ、どうしてもそうは見えない、って？　だが、本当に緑とはこのページの色であり、赤とはこの活字の色なのだ。だが、あなたは、そうではないと言う。よろしい、それでは、このページの色が白であり、活字の色が黒であることを私に証明してほしい。

さて、このように挑戦されると、あなたはどうするだろうか。学生たちはまず第一に、他のものの色との関係に訴える。このページの色は壁の色と同じであり、活字の色は髪の毛の色と同じだ。だから・・・、という具合に。しかし、もちろん私は、壁の色も緑であり、髪の毛の色も赤だと言っているのである。次に返ってくる答えは、科学に訴えることである。物理学が教えるところによれば、それぞれの色はそれぞれの波長の電磁波である。したがって、このページが白いというのは、このページからの反射光の波長を測定し、それが・・・であることを確かめればよい。しかし、もちろん私は、波長が・・・である色を緑だと言っているのである。当然のことながら、物理学に訴えてもダメなものは、視覚理論や認知理論や神経生理学に訴えてもダメである。ある特定の波長の電磁波が網膜に届き、光を感じる杆体細胞と錐体細胞を通して、その刺激がパルスとなって神経系を伝わり、脳の視覚領のその部分の・・・、といったことを持ち出しても、私は、それこそ白色ではなく緑色が見えるときの人間の認知のメカニズムなのだと言うだろう。ある色が見えるときのどんな条件を証拠として持ち出してきても同じことなのである。というのも、＜その色が見える＞という事実が確定されて初めて、その事実についての科学的解明がその事実の＜証拠＞となるからである。

こうなってくると、学生たちも、尋常な手段では私を説得できないということに気づく。そして、まともな手段で答えられない質問とはまともな質問じゃないんじゃないか、と疑い始める。かれらからの最上の答は、要するに、私は「白」と「緑」という日本語をきちんと習わなかった、だから、日本語で本当は

「白」と呼ばれている色を私は誤って勝手に一人だけ「緑」と呼んでいるだけだ、というものである。これはなかなか冴えた答である。なんだ、つまらない。問題は色ではなく、言葉だったのだ。要するに、見ている色には何の違いもないが、ただそれをどう呼ぶかの違いがあるだけで、ここには人を困惑させるような問題はなにもない。

だが、そうではない。こんな風に哲学の問題がいつもすぐ雲散霧消するようなものだったなら、どんなにか精神衛生上よかったであろう。私がその議論によって説得されないのは、問題が最初から、われわれの見ている色にあるからだ。あなたはこのページを見て、ある色の感覚を生じさせる。私もこのページを見てある色の感覚を生じさせる。あなたのその色感覚と私の色感覚が＜同じだ＞という保証はどこにあるだろうか？　あなたも私も、色をどの名前で呼ぶかということに関しては完全に一致しているとしよう。それでもあなたはこのページを見て＜白の色感覚＞を生じさせ、私は＜緑の色感覚＞を生じさせているかもしれない。その場合、われわれの色感覚の他に＜このページの本当の色＞というものがあるだろうか。しかし、たとえそのような＜色＞があったとしても、それはあなたにも私にも見ることができない。というのもこのページの色とは、掛け値なしにあなたには白であり、私には緑なのだから。

すると、このような状況が本当に出現しているとしたら、「このページは緑色だ」という私の主張の＜誤り＞を、あなたが私に説得することは不可能だろう。というのも私は完璧に正しい（！）のだから。この状況を、もうすこしいじってみよう。＜このページの本当の色＞というものがあり、それを神さまは見ることができるとしよう。その色は実は高貴の色（？）、紫である。さらに、あなたに生ずる色感覚は今度は白ではなく、青だとしてみよう。すると、紫色のページを見て、あなたは青の色感覚を生じさせ、私は緑の色感覚を生じさせていることになる。ところが、今度は色をどの名前で呼ぶかはわれわれ間で一致していない。というのも、日本語で生きるあなたも私も、それぞれの色感覚に「白」という言葉を結びつけることを学んでしまっているからだ。それで、何が起こるだろうか？　何も起こりはしない。あなたは紫色のページを見て青の色感覚をもち、私は同じそれを見て緑の色感覚をもつ。あなたはページの色を聞かれて「もちろん白だ」と答え、私も心底それに同意する。というのも、このページの色は＜本当に＞白なのだから・・・

神さまの目から見ると、われわれはとんでもなく＜誤った＞やり取りをしていることだろう。しかし、これこそが、われわれの真相（！）である。もしそうでないと言うのなら、どうやったらそれを示すことができるのかを答えてみてほしい。今度はもはや、言葉の問題に逃げることはできない。われわれが同じ色感覚をもっているということを示す唯一可能なやり方は、お互いが＜内側から＞他人の感覚を経験してみることだろう。だが、そんなことは不可能である。たとえ、神経科学の進歩によって、私の痛覚神経の末端をあなたの虫歯に結

びつけることができ、あなたの虫歯の悪化が私にも痛みを引き起こすようになったとしても、その＜痛みの感じ＞の中には、私の感じている歯痛はあなたの歯痛にほかならない、と主張するだけの根拠は何もないだろう。また、あなたの脳の中の痛覚神経を枝分かれさせ、それを私の脳の中に引き込んだとしても、われわれが同じ歯痛を経験しているという保証はどこにもない。誰かがもしそう主張したとしても、それは痛みの個としての、もしくは質としての同一性を＜確かめた＞からではなく、何らかの「定義」もしくは「宣言」によってそう主張しているにすぎない。要するに、われわれは、他人の感じている色感覚がどんなものであるかを一度だって確かめたことはないし、また金輪際、確かめようがないのである。これは科学の進歩によっては解決しようがない問題である。

2　科学がやること

さて、機械であれ他人であれ、その内面的経験の有無を＜内側から＞は確認できない、という事情は分かった。しかし、＜内側から＞ではなく＜外側から＞、機械に内面的経験が生じていないことを、そしてわれわれにはそれが生じていることを決定できないのだろうか。＜外側から＞の観察を徹底して行ってきたのが自然科学である。そして、物理現象に対してこんなにも成功を収めてきた自然科学が、心理現象に対してうまくいかないということがあろうか。実際、痛みについての医学的、神経生理学的探求は、麻酔技術の進歩というかたちで痛覚の制御を可能にしてきたし、精神疾患についての神経生理学的探求は、精神分裂病や鬱病などの心の病いを向精神薬で治療するという方向で前進してきた。してみると、科学は、着実に内面的経験が生ずる際の脳や神経のメカニズムを解明し、どのような状態のときにはどのような経験が生じるのかを突き止めつつあるのではないか。例えば、もし医学や神経科学が完成の域にまで達したとしたら、事故や老齢のおかげ患者が自分の痛みを訴えることができなくとも、どんな痛みがどこにどれくらいの強さで生じているかを知ることができるだろう。そしてきっと、怪我から快復した患者たちが語るように、誰もがそのような予想と診断を正確なものだと証言するだろう。ということは、さっきの滑稽なほどにグロテスクな状況、つまり「われわれの真相（！）」とは真相でも何でもなく、本当は、「白のページにあなたもわたしも白い色を感じ取り、それを「白だ」と言っている」、といったごくまともな事態がことの真相なのではないか。前節で見た＜素朴な物理主義＞は、科学によってそのような仕方で裏打ちされるのではないか。＜素朴な物理主義＞の正しさは、科学が＜神の目から見た事実＞そのものを捉えていることの自然な結果だろう。そしてその科学は、機械がどういうわけでその物理的な造りのゆえに内面的経験をもてないのかを説明してくれるだろう。

しかしながら、ある種の哲学者たちのこうした主張は、われわれの経験相互のハーモニーに対する慰めではあっても、いささか楽天的すぎる。例えば、あ

なたが原因不明の頭痛にさいなまれ、完成された医学を誇る未来の病院に行き、そこでさんざん検査されたあげく医者にこう言われたとしてみよう。

「あなたの脳はまったく正常です。頭痛を引き起こす脳の神経異常のどのタイプも、あなたの脳には見られません。われわれの医学は完璧です。ですから、あなたが頭痛を感ずることは不可能です。心理テストの結果によれば、あなたがウソついていることもありえません。考えられるのはただ一つ、あなたは頭が痛いと＜思いこんでいる＞だけだということです。要するに、あなたの頭痛は一種の妄想でしょう。あなたは＜本当は＞痛みを感じていないのです」

あなたはどうするだろうか。あなたは科学の権威の前に不承不承ひきさがり、アパートのベッドに横たわって自分自身を呪いながら、「これは本当じゃない、あたしは本当は痛くはないの、すべては幻覚なのよ」と、痛みをこらえながら、つぶやくだろうか？　それとも「冗談じゃないわ、痛いものは痛いんだから！　完全なる医学なんて糞っくらえ、本物の医学ならこのあたしの痛みを説明してみなさいよ」とその医者に吠えるだろうか？　妄想だろうが幻覚だろうがそう見えるものはそう見えるのだし、そう感じられるものはそう感じられるだろう。そう主張するのに、デカルト派の古くからの哲学議論を知らなくとも、＜素朴なメンタリズム＞が後押ししてくれる。もし「完成された医学理論」であなたの痛みの説明がつかないなら、それはあなたの痛みが痛みでないのではなくて、ただ、「完成された医学理論」が実は完成されていなかった、というだけのことだ。その医学理論は、あなたの痛みを説明できるように修正されなければならない。

では、「脳や神経系の同じメカニズムには同じ痛みが宿る」という＜素朴な物理主義＞主張は、科学によって実証されるわけではないのか。実証されないだろう。話は逆である。科学は、この場合、「脳や神経系の同じメカニズムには同じ痛みが宿る」ということを実証抜きの前提として採用することから、その活動を出発させているのだ。言いかえれば、＜素朴な物理主義＞はその前提の宣言である。したがって、科学はその前提に反する結果を生み出さないけれども、その前提を証明するわけではない。言ってみれば、「白のページにあなたもわたしも白い色を感じ取り、それを「白だ」と言っている」という事実は、科学が発見し確定した事実なのではなく、それを事実だと＜見なす＞ところから科学的探求が始まる出発点なのである。これは、科学が＜神の目から見た事実＞なるものを探求の結果とらえるわけではない、ということである。

フレーム問題

1　ロボットたちの悩み

カズヒサ教授はご自慢のヒゲをなでながら、フンフンとうなずく。研究室の隅にうずくまる老い猫のボロボス・タダシが「にゃあ」となく。いれたてのメープル・ティーに、お気に入りのクッキー半分かじって、教授はこうのたもうた。

「わしのヨシヒコならだいじょぶじゃ。きっとマダムをゲットして返ってきてくれるはず、わしにゃ自信があるわい」

マダムとは横町のちょいと粋な小料理屋の若女将のこと。大学のどこかに隠れたマダムを探し出し、一番で無事に学長室まで連れてくれば優勝という、よくある大学祭の余興の話。マダムは大正時代からのお宝ものの椅子に座り、その近くに時限爆弾の仕掛けあり、それを爆発させたら負け。といったって、トマト爆弾だから、破裂しても辺りが真っ赤に染まるだけ。で、ヨシヒコは背をかがめながら教授の研究室を後にした。ヨシヒコはもちろん、カズヒサ教授の最新作のロボットで、ラグビー部や山岳部の猛者どもを相手に奮闘を開始する。

「マダム、ハッケン。Lトウ、9F、923トクベツシツ」
「あっ、マダムの近くにはトマト爆弾があるんじゃぞい、ヨシヒコ」
「リョウカイ。ちっぺんでーるノ　イスノ　ウエ　まだむ　カクニン。まだむノヒザノ　ウエ　バクダン　カクニン。サクセン＝＜ PULL OUT (CHAIR, ROOM) ＞。スイコウ」
「いかん、ヨシヒコ！」

モニター見てわめく教授をよそに、ヨシヒコはまんまと椅子ごとマダムを教室から連れ出した。しかしそのとき、マダムの膝の上でトマト爆弾が爆発。ヨシヒコとマダムは仲良くトマト色に染まった。「なんて、おバカな子なの」とマダム。

「だから、問題は論理なのじゃい、論理。ヨシヒコは爆弾がマダムの膝の上にあることを＜知っておった＞のじゃ。それなのに、椅子ごと爆弾も引き出しおった。椅子を引けばマダムが引かれる、マダムが引かれれば爆弾も引かれる。ヨシヒコに足りなかったのは、風が吹けば桶屋が儲かる、あ、じゃなくて、自分のしたことの結果なにが意図せずとも起こるかということの論理計算なのじゃ。だから必要なのは、もっと論理学をつくる、じゃなくて、もっと＜論理なロボット＞をつくることだわい」

カズヒサ教授は頭から湯気を出し、ボロボス・タダシはその剣幕に恐れをなして逃げだし、メープル・ティーとクッキーはいつしかこぶ茶と、名古屋最強の名物(？)ウイロとナイロに代わっていた。こうして作られた最新傑作、テルオは、一年後に同じく一番乗りでチッペンデールの椅子に座ったマダムを発見した。
「ちっぺんでーるノ　イスノ　ウエ　まだむ　カクニン。まだむノ　ヒザノ　ウエ　バクダン　カクニン。サクセン＝＜ PULL OUT (CHAIR, ROOM)＞。ケイサン　スイコウ。イスヲ　ヒク　カベノ　イロ　カワラズ。　イスヲ　ヒク　ユカ　キズツク。イスヲ　ヒク　マド　ソノママ。イスヲ　ヒク　・・・」

椅子を引いても天井の大きさは変わらない、という29番目の計算を遂行している途中でまたもや爆弾は破裂。テルオはまたしても(？)、マダムと一緒にトマトジュースの海の中に沈没した。「ほーんとに呆れたもんだわ」と一つ年をとったマダムは言った。

「だから、だからして問題は＜分別な論理＞なのじゃよ、物分かりのいい論理。テルオはよくやったんじゃが、やりすぎじゃい。いくら計算が得意で、自分が何をしたら周りがどうなるかを計算しにゃならんといったって、椅子を引っ張ってもゴア-ブッシュの選挙結果は変わらんなぞということまで計算していたひにゃ、寿命がいくつあっても足りんわい。無視すればいいんじゃ、無視、無視。このレースに関係のないことは、すべて無視するようになればいいんじゃ。要するに、もっと分別な論理学を作る、じゃなくて、もっと＜分別に論理なロボット＞を作る、これじゃて」

カズヒサ教授は猛然と仕事に取りかかり、猛然と飲み、猛然と食べた。帰ってきたボロボス・タダシはお下がりのせいで、あわれブタ猫状態。こぶ茶とウイロにナイロは、アイリッシュ・ウィスキーと千葉県産のピーナツに代わっている。こうして、最・最新鋭型の＜分別に論理なロボット＞、ヒロユキが一年後に完成した。そしてまたしても、ヒロユキは一番でマダムを発見する。
「ちっぺんでーるノ　イスノ　ウエ　まだむ　カクニン。まだむノ　ヒザノ　ウエ　バクダン　カクニン。サクセン＝＜ PULL OUT (CHAIR, ROOM)＞。ケイサン　スイコウ。モシ　ムカンケイ　ナラバ　ケッカ　ムシ・・・」
「ええぞ、ええぞい。ぐふっ」

と、ここまで聞いて教授。しかし、ここからヒロユキ、何もいわずに黙ったまま、マダムのいる部屋にも入らずに戸口のところに突っ立っている。ヒロユキ、なにか苦しそう。マダムの心配そうな顔が見える。
「ヒロユキ、どうしたんじゃい。爆弾が破裂するぞ、なんかせんかい、どうにかせんかい！」
「・・・キョウジュ、ジャマヲ　シナイデ。ボク　ムシ　スルノニ　トテモ　トテモ

イソガシイ。ケイサン…　ムカンケイ…　ムシ…。ムカンケイナ　キケツリスト　なんばー　1502。…なんばー　1503。…」

かくして、三度目もまた、トマト爆弾はマダムの膝の上で爆発。「もう教授のツケはお断りよ、ぜーったい。ほんとにおこったから。」と白いドレスを真っ赤にしたマダム。「うーん」とうなってソファにひっくり返るカズヒサ教授の横で、ボロボス・タダシが「にゃ、にゃっ」と笑った

2　知性のすべてを明示的な規則に還元するのは無理だ！

話を少しまとめよう。ミンスキーが提唱した「フレーム」とは、AIの扱った「小世界」を日常的な場面にまで具体化しようとしたときに出てきたものであって、その基本的な発想は、さまざまな「フレーム」をつなぎ合わせて、日常世界一般を扱う常識的知性を作り上げようとするものである。ミンスキーはこう書いている。「新しい状況に直面したとき（あるいは、現在の問題に対する見方を実質的に変更したとき）、人は記憶の中からフレームと呼ばれる基本的構造を選び出す。これは、人間が記憶している枠組み（フレームワーク）で、その細部を必要に応じて変更することで現実の世界に適合するようになる。…フレームとは、居間にいるとか、子供の誕生パーティに行く…といった、ステレオタイプの状況を表現するためのデータ構造である」。シャンクの「スクリプト」やボブロウの「スキーマ」といったものもほぼ同様のものと考えてよい。

問題は、人間の場合あらゆる規則の適用には「他の事情が等しいならばceteris paribus」という例外対処条項が暗黙の前提になっているが、人工知能の場合には、その条項をこのままの形で使うことができないということである。この例外対処条項のすぐれた点は、その中身を＜確定しなくてもいい＞ということなのだが、人間に対してならともかく、人工知能に対してはそのような曖昧な内容を暗黙的な仕方で教えることはできない。したがって、フレームの適用規則を人工知能に教える場合、その「他の事情が等しいならば」の中身、つまりいかなる場合に例外の発生によって規則が適用不可能となるのかを、もれなくすべて確定しなければならなくなる。言いかえるとフレーム問題とは、人工のフレームで例外的状況に対処するためには例外をことごとく明示的な仕方で知っておかねばならない、という問題なのだ。同じことをもっと意地悪くいえばこうなる。こうしたシステムは、いまが典型的状況だということを知るためにも、あらゆる例外の可能性に関して、それが発生していないということをことごとく明示的な仕方で知っておかねばならないのである。

しかし正確に言って、それがどうして問題なのか？　「他の事情が等しければ」という条項の本当のご利益が、実は、その中身をそもそも＜確定できない＞のに例外に対処できるところにある、ということを知ったなら、あなたにもその不安の行く末が予感できるだろう。まず第一に、フレームは典型的状況を表

したものだから、出荷時のパソコンのようにいわゆる＜ふつう＞を想定したシステムの初期設定、つまりデフォルト値が、そのままですべて細部まで有効なのはまれである。現実の状況はむしろ、何らかの仕方で必ず例外的要素を含むだろう。例外への対処を＜例外用の第２規則＞の適用でやりくりするしかないシステムは、あらかじめ例外をすべてリストアップし、そのうちどの要素が出現しどの要素が出現していないのかを確認しなければならない。そのためには、状況が変化するたびに、リストアップされた例外項目を総当たりでチェックしなおさなければならないだろう。というのも例外的要素であったものがそうはならなくなる場合もあるからだ。しかし、例外の漏れのないリストアップとそのつどの漏れのないチェックは非現実的な対処法だ、ということをわれわれは見たばかりである。例外のすべてをリストアップして覚え込むのは絶望的だし、大部分は変化のない項目を逐一チェックするのは時間の浪費である。

では、第二に、例外項目の漏れのないチェックという愚かしさを避けるために、何が例外として考慮されるべき要素であって、何がそうでないかをそのつどの目的に照らして決定するというやり方はどうだろう。達成したい目的に対して考慮すべきなのは、それに＜関連性＞をもつ要因だけで十分ではないか。しかしまたしても問題は、このようなシステムでは、＜関連性＞の有無を決定するのに＜関連用の第３規則＞が必要になる、ということである。その結果、われわれはどこへ行きつくのか。どの項目に関しても＜例外用の第２規則＞を適用すべきか否かを判定するために、それに対してまず＜関連用の第３規則＞を適用しなければならない。したがって、例外項目の漏れのないチェックは、それ以前の、関連項目の漏れのないチェックに置きかえられただけである。そのバカバカしさから逃れようとして、＜関連用の第３規則＞の適用に例外を設けようとするなら、＜＜関連用の第３規則＞の例外用の第４規則＞をすべての項目に関して適用しなければならなくなる･･･。

もう十分だろう。冒頭の改良型ロボット「テルオ」は、自分の行動の結果をやみくもにすべて計算するところにその悲劇があった。「テルオ」は、自分が「椅子を引く」結果として例外的状況が発生していないかどうかを、壁の色や床の表面や窓の形などあらゆることに関して確かめねばならなかったのである。しかし、次の改良型ロボット「ヒロユキ」も悲劇から逃れることはできなかった。「ヒロユキ」のさらなる苦悩は、「マダムの救出」という目的に＜関連性＞をもつ結果のみを考慮しそれ以外を無視するために、＜関連性＞の有無をあらゆる結果に関してまず確かめねばならなかったことである。ここでもう一度、フレーム問題のなんたるかをくり返しておこう。今やわれわれは、それがどういう問題であるかを身にしみて分かったはずである。

「＜何を考えなくてもいいか＞ということを考えずに、考えなくてもいいことをいかに考えないですますか」