第 12 章

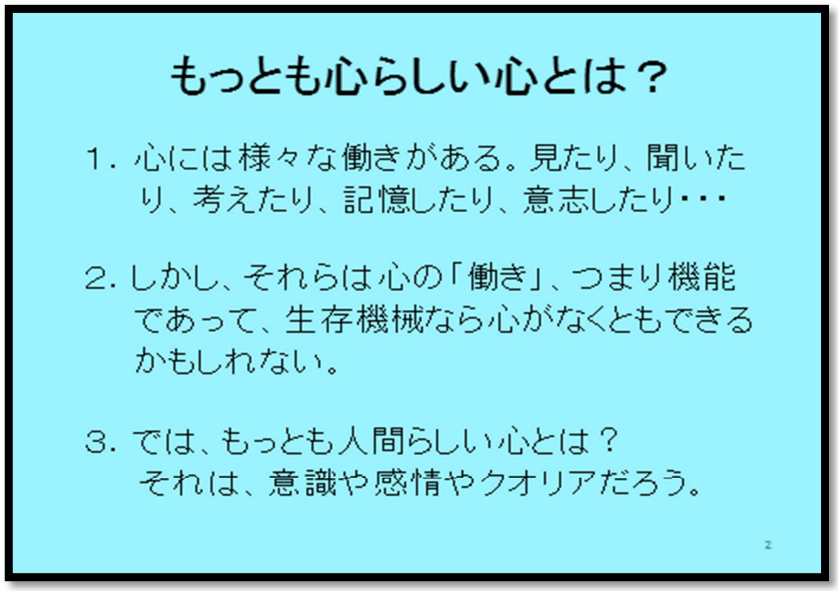
認知哲学
― ロボットに心があるって、どういうこと？―

柴田正良(現代哲学)
金沢大学人文学類

もしも脳や身体に関する生理化学や神経科学がその法則や概念・用語によって思考、意図、欲求、感情、知覚、記憶といったわれわれの認知現象を直接に説明することができるなら、およそ認知科学といった学問は必要ないであろう。つまり認知科学という科学が自律して存在しうるなら、それが「存在する」と独自に宣言する対象や法則によって、認知という現象が、ある一般性の下で説明されなければならない。その一般性は、個々の人間や地上の様々な生物をカバーするだけでなく、人工的に作られた行為者、すなわちロボットにも妥当する、というのが本章の主張である。一言でいえば、本章のテーマは、ロボットも人間と同じような心を持つことができる、という主張を哲学的な観点から擁護することである。

キーワード 1，クオリア
キーワード 2，ゾンビ・ロボット
キーワード 3，スーパーヴィーニエンス
キーワード 4，多重実現
キーワード 5，認知機能ロボット
キーワード 6，可能世界
キーワード 7，非還元的物理主義

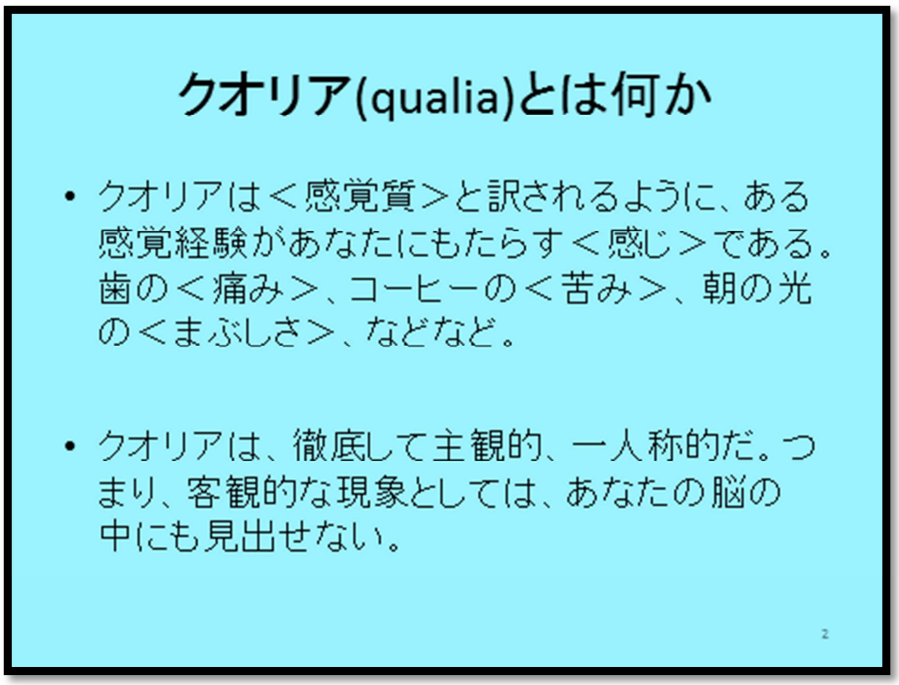
12-1　われわれが理解する心



認知の一般理論からすれば、まず問題なのは心が果たしている「機能」だろう。というのも、これこそが生物体にとって生存を可能にし、自然淘汰を勝ち抜かせてくれた武器だからだ。そして、認知科学も科学である限り、基本的には現象の因果的説明を与えようとする以上は、原因と結果の組み合わせによって説明が可能なさまざまな機

能を研究のターゲットしてきたのも当然である。そしてその機能は、いったん原理的な仕組みが解明されれば、人工的な装置によって部分的に模倣され、実現されてきた。たとえば、算術の計算や、図形の読み取りや、音声の認識など。

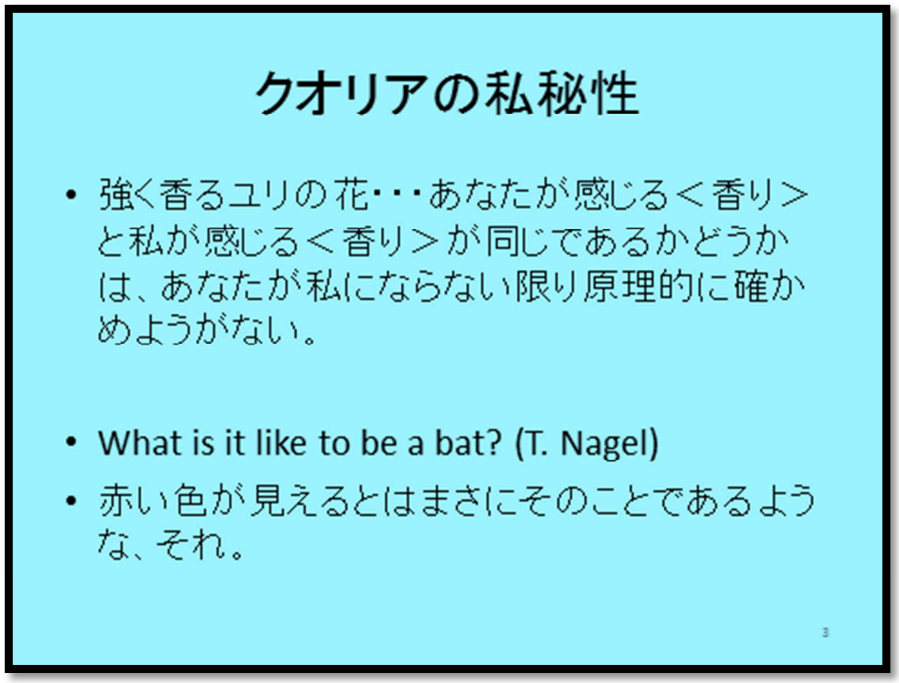
しかし、ロボットに心を持たせようとしているとき、その心がこのような雑多な認知機能の寄せ集めでは人は満足しないだろう。われわれも、ロボットが心を持つと主張しているとき、このような「安価な機能」でお茶を濁そうとしているわけではない。ロボットは、われわれが最も心らしい心だと考えているものを持ちうる。それは、ふつうは機能と一体となって現れているので気づかないが、本質的には機能から概念的に切り離しうるもの、つまりわれわれが経験する「現象的な心」である。ここで、現象的な心ということで理解されているのは、「意識されている限りでの意識」、「感じられている限りでの感情」、それにいわゆる「クオリア(感覚質)」などである。



さて、現象的な心の代表格、クオリアとはなんだろうか？　これを言葉で説明するのはきわめて難しい。というのも、それは現物を直接に示すことによってしか説明できないタイプの概念である上に、その現物というのが、三人称の客観的世界の中で指し示すことのできる事物ではなく、むしろ客観的事物についての一人称の主観的経験内容だからだ。このように説明すると、クオリアは何か人がめったに経験することのない特別なものであるかのように思われるかもしれない。しかし、それはまったくの誤解である。クオリアは日常にあふれている。というよりも、クオリアを朝起きてから夜寝るまで経験し続けること、それがわれわれの日常生活なのだ。なぜなら、クオリアとは、われわれが朝飲むコーヒーに感ずるその＜味の感覚＞、顔を上げて眼に入った朝日の＜まぶしさ＞、そして明日から海外旅行に行けるという期待がもたらす＜高揚感＞などだからである。だから麻酔をかけられて手術の痛みを感じなかった場合、あなたはあるべきクオリアを感じなかったことになる。あなたがこのとき幸いにも経験せずにすんだ＜シャープな痛み＞、それがクオリアである（あなたは、きっと＜ごつごつと何かが骨に当たる感じ＞という別のクリアを経験したであろうが)。

クオリアの最も大きな特徴の一つは、私秘性(privateness)にある。それはどういうことかというと、あなたの経験するクオリアを、他の経験の主体である他人はだれも経験することができないということだ。そしてクオリアは経験することによってしか「それが何であるか」を知ることができないのだから、他人は「あなたのクオリアが何

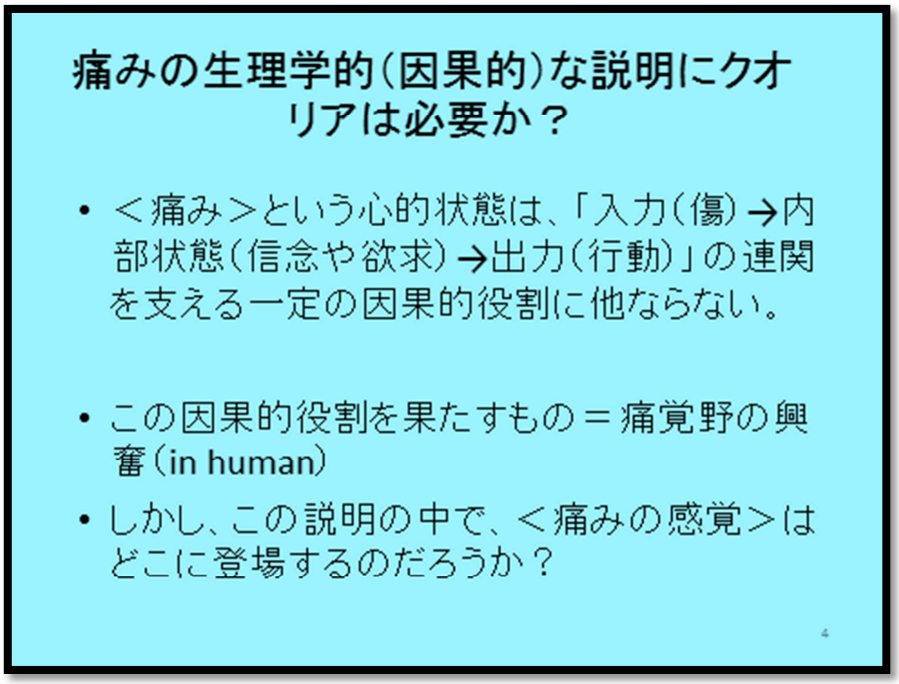
であるか」をまったく知ることができない。あなたと他人とで同じそのクオリア(numerically identical qualia)を共有できないのだから、同じ質のクオリア(qualitatively identical qualia)を共有できているのかどうかも分からない。そのことは、あなたが青い海原を見て、青のクオリアを心から満喫しているときでも、その青クオリアはあなたの脳のどの部位にも、どの神経細胞にも見出せないという事実と深く関連している。あなたの脳が経験しているのは一人称の現象であるのに対して、あなたの脳の中に見出せるのは三人称の現象以外にはないからである。



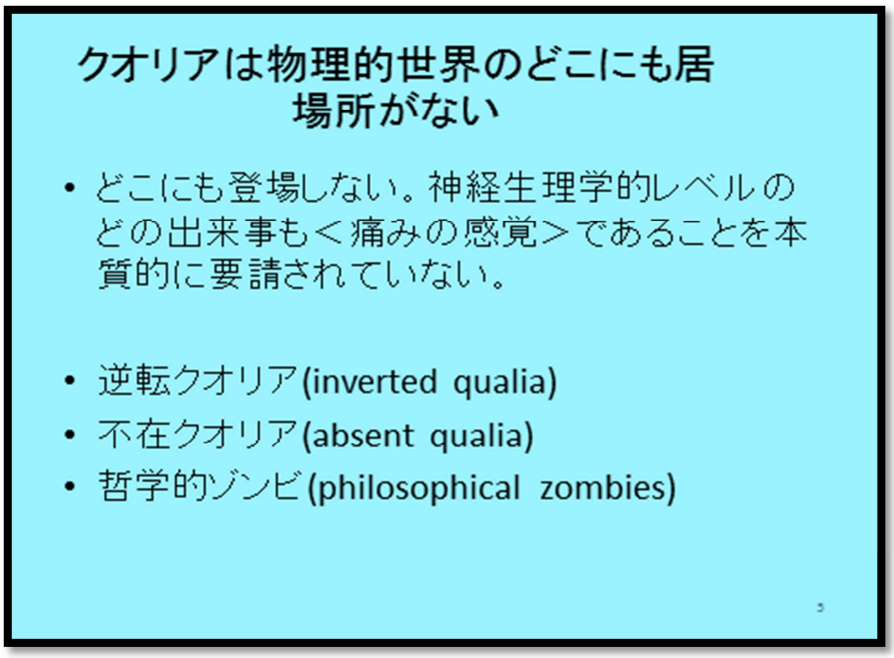
他人が私と同じタイプ（質）のクオリアを感じているのかどうかは、厳密には認知科学の問いではない。というよりも、科学には答えることができない哲学的な問いである。科学は、認知科学ももちろん、三人称的に観察可能で、（原理的に）操作可能な現象・出来事にしか関わらない。したがって、あなたの経験するクオリアの状態が原理的にあなたにしか知られえない一人称的現象・出来事なら、科学はそれに対して口を閉ざすほかはない。

ところが、まことに奇妙なことに、われわれが現に生活している日常では、（医療の現場以外）最も問題なのは三人称的現象であるあなたの脳ではなく、あなたが経験しているあなたの一人称的クオリアなのだ。考えてみれば、クオリアの私秘性は、実に奇怪な状況がわれわれの日常の背後に隠されている、ということを示している。あなたは恋人と悲しい恋物語の映画を見る。二人は、映画の主人公の哀れな運命に涙し、それに引き比べて自分たちの幸せをいや増しに実感する。二人の気持ちは通い合い、いまやお互いの心が文字通り共有されたかのようだ。しかし実は、それは言葉と身体と態度上だけのことであり、相手が本当はどういう＜悲しさの感じ＞、＜うれしさの感じ＞を感じているのかは金輪際わからない。生まれてから一度たりとも、われわれは、他人が感じているクオリアがどういうものであるかを経験したことがないのだ。だから、互いの気持ちや感じを伝え合っている表面上の滑らかなやり取りの下では、真っ暗な奈落の淵が二人を隔てているといった格好だ。

このように捉まえようもないクオリアは、意識や感じと同じようにわれわれにとって「最も心らしい心」なのだが、それは、さらに生理的身体からも、また物理的世界からも「浮いて」しまっているようにしか見えないという事情がある。



たとえば、人が痛みを感ずるとき、そのおおざっぱなモデルにおいては、「心的状態としての痛み」はおおむね次のような因果過程の中で機能を果たしていると想定される。まず、身体のどこかに傷が生ずる。それは身体表面を覆う痛覚神経を刺激し、その信号が脳へ入力され、その入力情報と、記憶や欲求などの脳内部に由来する情報が相互作用し、傷をかばったり声をあげたりするなどの行動が出力される。その「相互作用」の実態がどういう情報処理プロセスなのかはまだ詳しく解明されていないが、「痛み」の果たしている「機能」が入力を出力へと因果的に媒介する中間項であるのは明らかだ。それは、人間においては、痛覚神経の入力投射を受ける脳の痛覚野の神経生理学的な働きであろう。しかし、そうすると、この説明のいったいどこにクオリアとしての「痛みの感覚」は

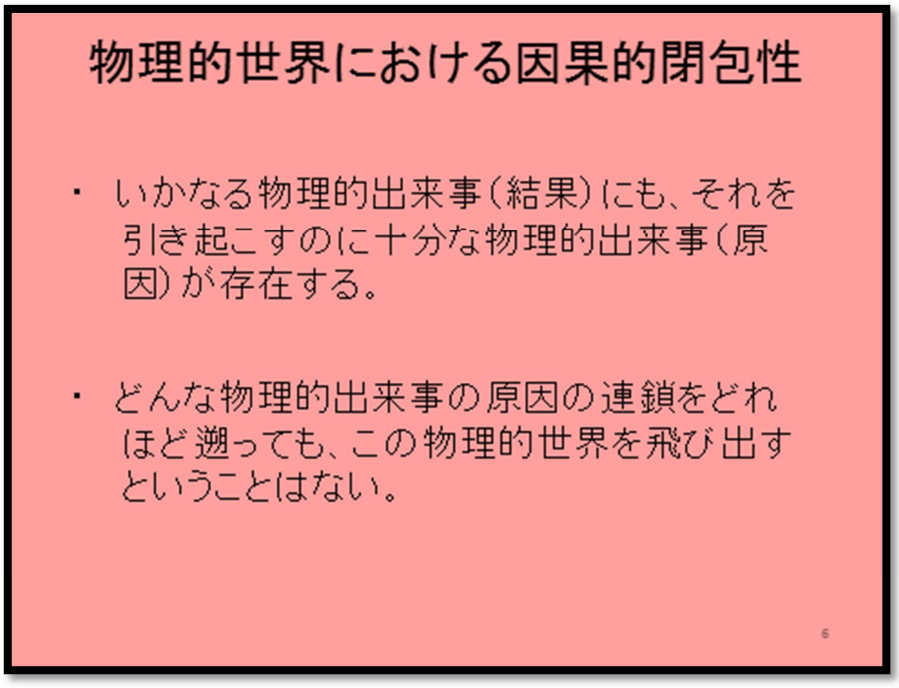


登場するのだろうか？　どこにも登場しない、というのがその答えだ。つまり、痛みといったいかにも「心」らしい状態でさえも、それが果たす機能として理解されればある意味で「痛み」の理解としては十分であり、そしてそのためには、それがなおかつクオリアとして「あのような痛みの感じ」を持っている必要はないということである。このことはクオリア一般に、また神経生理学的な出来事に限らず因果的な出来事一般に、容易に拡大される。つまり、クオリアが関わっているどのような出来事に関しても、それが機能的・因果的な現象として理解される限り、それにどんなクオリアが関与しているかはまったくどうでもよいということだ。一言でいえば、クオリアは、因果的世界（基本的には物理的世界）のどこにも本来の居場所を持たない根無し草のような現象なのである。

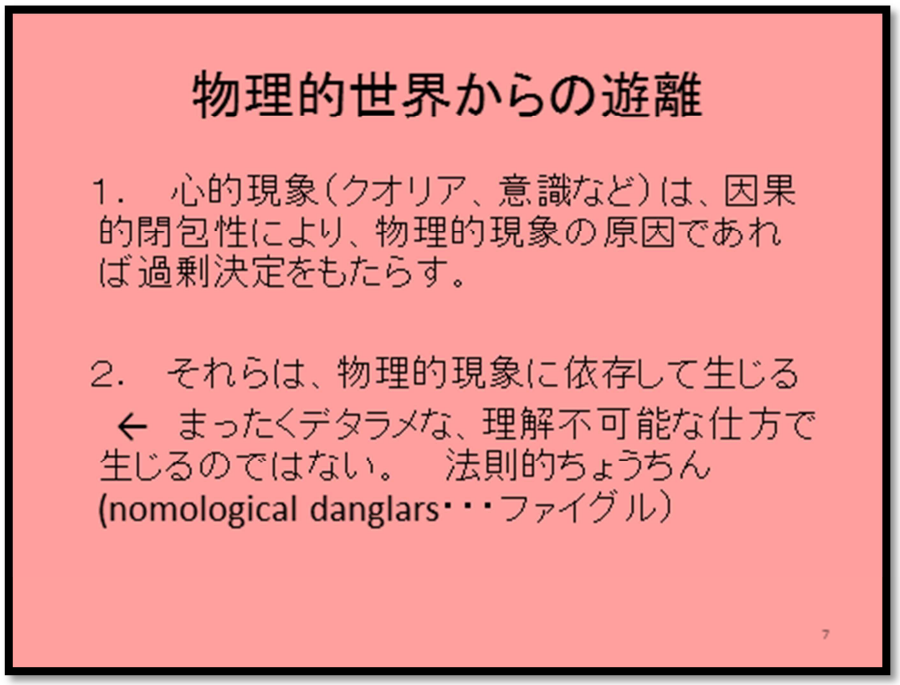
このことを劇的に示す思考実験がある。それは、「逆転クオリア」、「不在クオリア」とそれぞれ呼ばれているが、その最終的な形態は D. チャルマーズが衝撃的に描いた「哲学的ゾンビ」である。あなたは私とまったく同様に、交通規則に従順な市民だから、赤信号では止まり、青信号では進む。そして、三人称による観察では、われわれの行動に（そしてもちろん関連する脳プロセスにも）この点でまったく違いは見られない。ところが私は赤信号を見ているときに「赤クオリア」を感じているのに対して、あなたは「青クオリア」を感じており、そして青信号に対してもそれぞれ逆である。われわれは色のクオリアを逆転させて経験しているのだ（色スペクトル上の正確な逆転関係がお望みなら、そういうストーリーに変えてもよい）。ここで注意すべきなのは、このわれわれの重大な相違にもかかわらず、外面上の行動に（そしてもちろん関連する脳プロセスに）関して、その相違は一切表に現れないということだ。もちろん、誰かがわれわれに「信号はいま何色？」と聞いたとしても答えに食い違いはない。可哀想なあなたは、生まれてずっといままで「青クオリア」を「赤」と呼ぶように教わってきたからだ。もっとも、私がここで実際は「黄クオリア」を経験しているのだ、と想定したって何ら差し支えはないので、本当に可哀想なのは実は誰だか分からない。これが「逆転クオリア」の例である。

これを少しひねって一般化すれば、「不在クオリア」と「哲学的ゾンビ」がどんな具合だか、容易に想像がつくだろう。ある特定の経験の場面であるべきクオリアが生じない人は「不在クオリア」の病に罹っているのであり、すべての経験において一切のクオリア（および意識）を経験しないという極端な人が「哲学的ゾンビ」なのである。

12-2　スーパーヴィーニエンス（SV）と多重実現



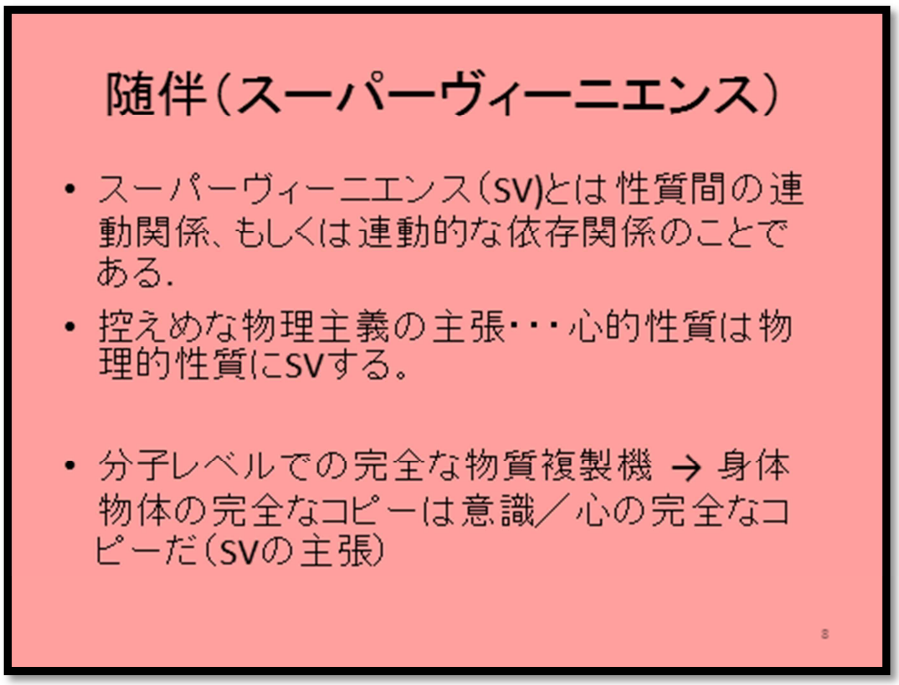
以上の思考実験には少しばかりの哲学的イマジネーションが必要とされるが、そう大したことはない。そこで、みなさんにはもう少し哲学的な発想につきあってもらって、クリアの存在論的な身分を考えてもらうことにしよう。まず、われわれの世界、つまり現実世界では「因果的閉包性」という原理が成り立っていると考えるのが妥当だろう。これが出発点だ。これは、現実世界の物理的出来事全体についてこう主張する。「いかなる物理的出来事（結果）にも、それを引き起こすのに十分な出来事（原因）が存在する」。もちろん、この原理が妥当しないことも論理的には可能だろう。つまり、原因なしの出来事の出現が頻繁に起こるような可能世界は存在しないということは、D. ヒュームの懐疑論が示したように、何らかの確かな根拠から論証できることではない。しかし、われわれの現実世界がこの閉包原理の成り立つ可能世界の一つだ、ということは哲学のみならず科学全体の作業仮説としては健全だと言うことができる。



さてそうすると、一つ困ったことが生ずる。それは、クオリアや意識といった現象的意識が何らかの出来事の原因として働くと、必ずそこには「過剰決定」と呼ばれる事態が生じてしまうということだ。過剰決定とは、同一の結果をたまたま複数の原因が同時に引き起こすことであって、二つの別々の弾丸が同時に大統領の頭を貫く場合のように、現実世界では普通めったに起こらない。ところが、「痛い」というクオリアがクオリアという身分で「手のとっさの動き」を引き起こすなら、この過剰決定が生じていることになる。なぜなら、「手のとっさの動き」が身体動作である以上、それを引き起こすのに十分な物理的原因、つまり脳の神経生理学的な出来事があったはずであり、それに加えて心的現象としての「痛みクオリア」も原因として働いたとすれば、ここに二つの十分な原因が同時に働いたことになるからである。

この「過剰決定」問題が深刻な困難となるのは、たとえば意図的な行為を考えてみればよく分かる。この場合、意図という一人称的に知られる心的出来事が行為の原因だとふつう考えられているが、しかし、それと同時に、その行為は常に物理的な出来事によっても引き起こされている（過剰決定されている）のだ。するとわれわれは、日常の自然な状況ではほとんどまったく生じないような事態が人間の行為の場合には常に生じている、という不合理な解釈をせざるをえないように思われる。そればかりか、一人称的な心的出来事と三人称的な物理的出来事が別々

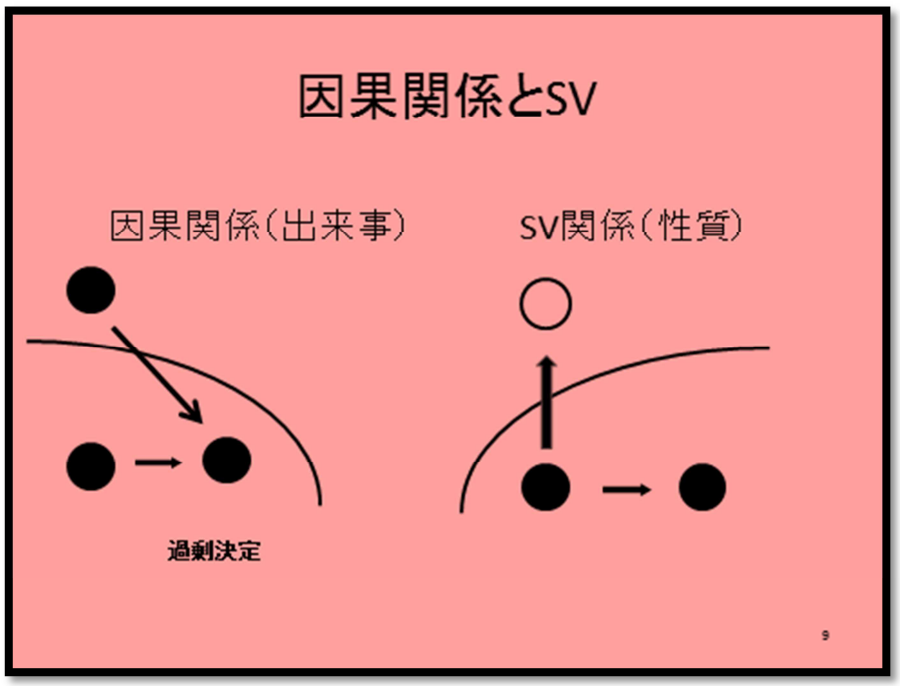
の十分原因として行為を引き起こしているなら、他方が生じなくともどちらか一方だけでやはり同一の行為を引き起こすことになるはずだが、たとえば脳が一時的に麻痺していながら意図だけが同一の行為を引き起こすなどというのは、現実世界においてはありえないことであろう。したがって、こうしたことが示しているのは、一人称的な心的出来事としてのクオリアや意識と、脳の三人称的な物理的出来事は、(何かの原因となりうる)「出来事という身分」では別々の二つの出来事ではなく同一の出来事なのだが(出来事の同一性)、しかし人がそれを知る(アクセスする)仕方に反映される「性質という身分」では二つが区別されるということである(性質の非同一性)。したがって、脳内の物理的出来事が法則的に生起する限り、逆転クオリアや不在クオリアの論理的可能性にもかかわらず、現実世界では、われわれの期待するようにクオリアや意識はでたらめには生起しない。脳とクオリアのこの関連性は、スーパーヴィーニエンス(supervinience)と呼ばれる関係だ。



スーパーヴィーニエンスは二つの性質(集合)間の「連動的な依存関係」であり、「随伴」とか「併発」と訳されることもある。たとえば、<液体である>という性質は物質がある場合に持ち、別の場合には失う性質だが、この性質は、その物質を構成する分子がある特定のつながり方をしている場合に出現することが知られている。そこで、このような場合、<液体である>という性質は<物質分子がしかじかの仕方でつながっている>という性質にスーパーヴィーンする(SV する)と言われる。ここで気をつけておくべきことは、<しかじかの仕方で分子がつながっている>という性質が出現したら必ず<液体である>という性質は実現されるが、その逆は成り立たない、という点だ。というのも、<液体である>という性質は、別の種類の物質では<別のかくかくの仕方で分子がつながっている>ということによって実現されるかもしれないからだ。したがって、スーパーヴィーニエンス関係を理解するためには、ある大きさの現象レベルで見られる性質と、それより微細な現象レベルで見られる性質との間の関係をイメージするのがよいかもしれない(必ずしも、現象の<マクロ><ミクロ>関係には制限されないが)。

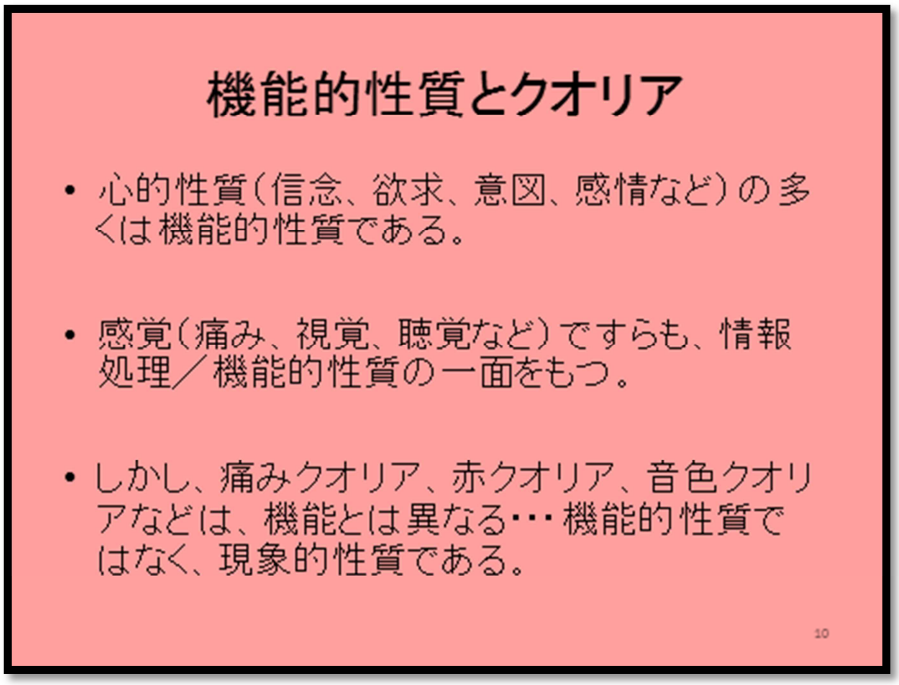
そこで、本章の立場を率直に表明しよう。「三人称の(機能的な)心的現象だけでなく一人称の心的現象も、物理的現象に SV する」、これである。これが立場の「表明」にしかなっていないのは、現在、三人称の心的機能の SV に関しては強力な論証が存在するが、クオリアや意識といった一人称の心的現象の SV に関してはなお係争中だからである。もう一つ、これは本章の認知科学のテーマには直接関係しないが、この立場は、心的性質と物理的性質の「同一性」までは主張しない。それは、両者が別々であることを認めた上で、物理的性質への心的性質のスー

パーヴィーニエンスを主張する、いわゆる控えめな「非還元的」物理主義の主張である。



SV についてのもう一つの論点。この SV は、脳を含めた身体すべての完全なコピーが存在するなら、そのコピーは心的状態の完全なコピーともなっているということを含意する。いわば、世界の物理的性質を決定すれば他のすべての性質も決定される、という意味での物理的性質への依存性を主張する点で、この立場は物理主義なのである。

上の「因果関係と SV」は、SV 関係の特徴をまとめたものだ。黒い曲線は物理的世界の限界を示している。そこで、その範囲外に心的出来事が存在し、それが身体動作という物理的出来事の原因となるなら、脳状態とともに過剰決定の事態を引き起こす（図の左側）。したがって、心的出来事は出来事（●）として物理的世界の外側に存在するのではなく、性質（○）として物理的出来事（の性質）に SV する（図の右側）。



さてここで、心的性質もしくは状態として一括して分類されるものを整理しておこう。すでに何度も述べたように、われわれが心と考えているものの働きは、ほとんどがまさに「心の機能」である。しかし、ここが分かりにくいことなのだが、「心の機能」の一部はクオリアをいつも伴っていて、われわれは、内観的にはそちらの方を「心らしい心」と考えている。たとえば、恐怖という感情は「差し迫った危険に対する回避行動などを準備する」という立派な機能を果たしているのだが、われわれにとって恐怖の本質はまさに「身の破滅を実感させるような、身体が震えるあの感じ」といったものだろう。その最たるものは意識である。意識が果たしている機能はさまざまな情報の集中的なコントロールにあるだろうが、意識の最も意識らしいところは、睡眠中の無意識との対比から示唆されるように、「私に直接に与えられる私の中身（考え、気分、感覚･･･）」とでも言うほかはないような、自分にだけ直に知られる自分の内面である。

そこで、われわれの心のこの二つの側面を、次のようにまとめておこう（不正確なところもあるが）。ポイントは、心は二つのまったく異なる性質のグループ、「現象的性質」と「機能的性質」を持つということだ。

１．一人称的に捉えられる（私にしかアクセスできない）心
　私秘的、主観的、現象的な心、クオリア、現象的な意識、感じられる限りでの感情、等々
２．三人称的に捉えられる（他人にもアクセスできる）心
　公共的、客観的、機能的な心、因果的役割、機能としての意識、機能を果たす感情、等々

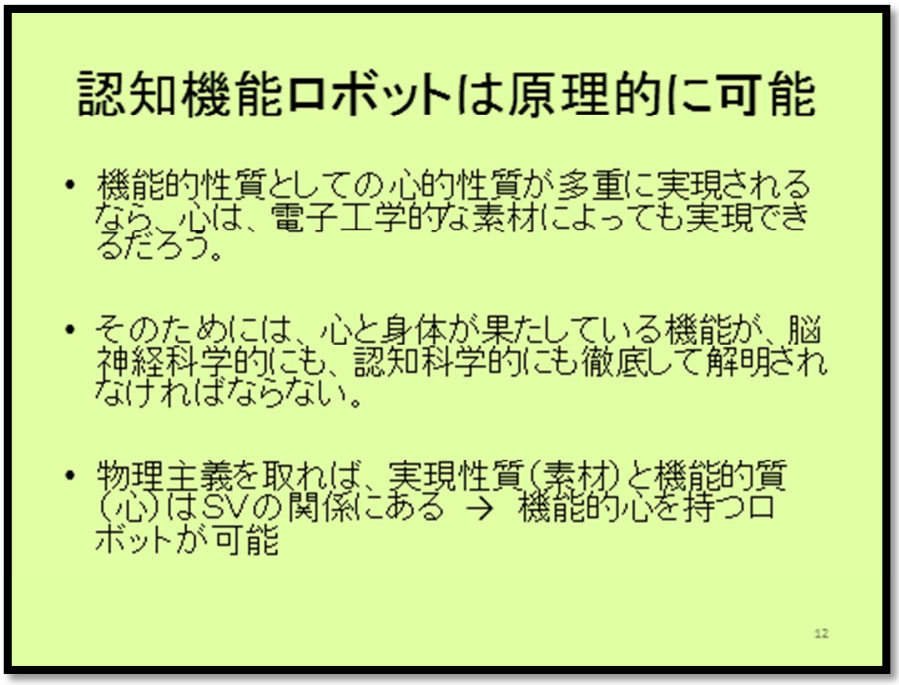
スーパーヴィーニエンス(SV)という関係は、生物学的性質と神経生理学的性質のように、現象のあるレベルで実現される性質とそれを実現するより下位の基盤性質との間に成立することがよくある。このとき、互いに異なる下位の複数の基盤性質群が上位の同じ実現性質を実現する場合、実現性質は基盤性質によって「多重に実現される」と言う。つまり、多重実現(multiple realization)は、SV の特殊な場合に出現するもう一つの性質間の関係だと考えることができる。これは、実現される性質が機能的性質の場合、顕著だ。というのも、「機能的性質」とはある「因果的役割を果たす」ということによって定義される性質であり、通常はどんな因果的役割も、環境次第で異なった複数の性質群によって実現されるからである。たとえば、人もタコも宇宙人もロボットも等しく「機能的な痛み」を持つ場合、その痛みは、ヒトの神経系によっても、軟体動物の神経系によっても、また宇宙人の未知の内部状態によっても、ロボットの電子工学的な状態によっても実現されていることになるだろう。(一言注意。ここでは、「性質」と「状態」は基本的に同じ意味である)。

したがって、認知機能として（宇宙人を含めた）認知的行為者一般に想定される心的状態は、いくつかの基盤状態の選言（「あるいは」で結ばれた複数の状態）によって実現されることになる。上のスライドでは、「スペインのグラナダに行きたい」という欲求状態（P）が、「M1」や「M2」等といった複数の基盤状態によって実現されることを示している。

「Px ⇔ (M1x ∨ M2x ∨・・・∨ Mnx)」。これはこう読む。

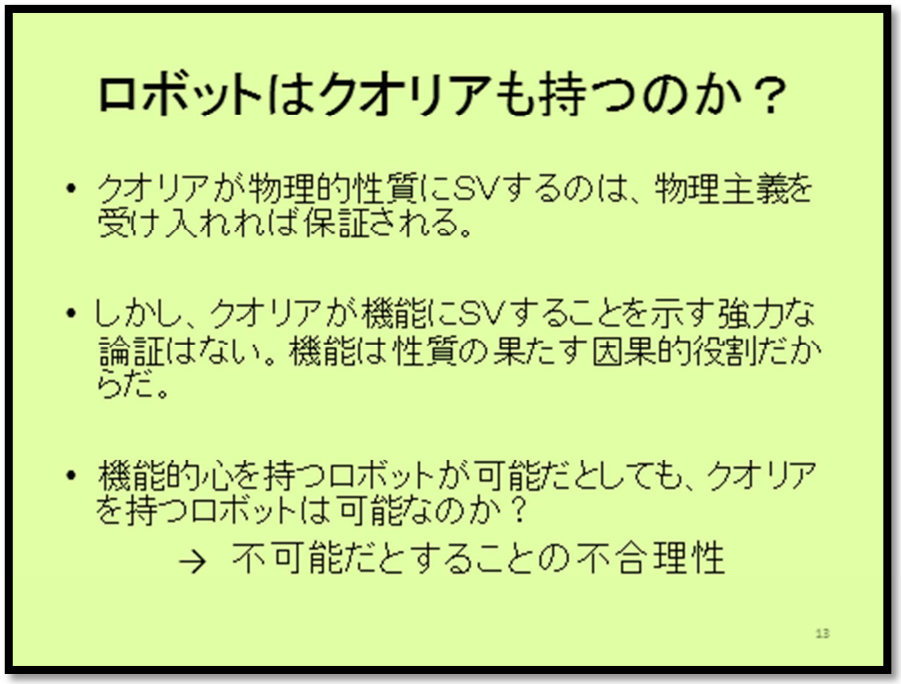
「ある行為者 x が心的状態 P にあるならば、x は基盤状態 M1 にあるか、あるいは基盤状態 M2 にあるか、あるいは･･･基盤状態 Mn にあり、またその逆でもある。」

12-3　認知機能ロボットの制作可能性



これまで述べてきたことからすれば、「認知機能」としての心の機能を備えたロボットが原理的に制作可能なのは明らかだろう。というのも認知機能が機能的性質であり、それゆえに多重に実現されるなら、その基盤性質を持つ素材が何であってもかまわないからだ。足し算や引き算は脳の神経系がやろうと電卓がやろうと同じ計算であり、図形の把握は眼と脳の神経システムが行おうとセンサーと計算機のシステムが行おうと同じ形態認識である。

かつて機能主義者であった頃の H. パトナムは、「一定のしかるべき機能を果たすなら、それがたとえチーズと空き缶で出来ている存在者であっても思考する」という趣旨の発言をしたことがある。だから機能主義が正しければ、（もちろんこの世界の条件によるが）電子工学的な素材によって作られたロボットが機能的な心を持つことには何の不思議もない。そして認知科学は、性質の存在論としての機能主義にコミットしない限り、認知の一般理論として成立しないのである。したがって、少し強引に言えば、認知科学が「認知一般の科学」である限り、それは心の機能を備えたロボットの制作可能性を含意する。原理的に制作可能なこのロボットを「認知機能ロボット」と呼ぼう。

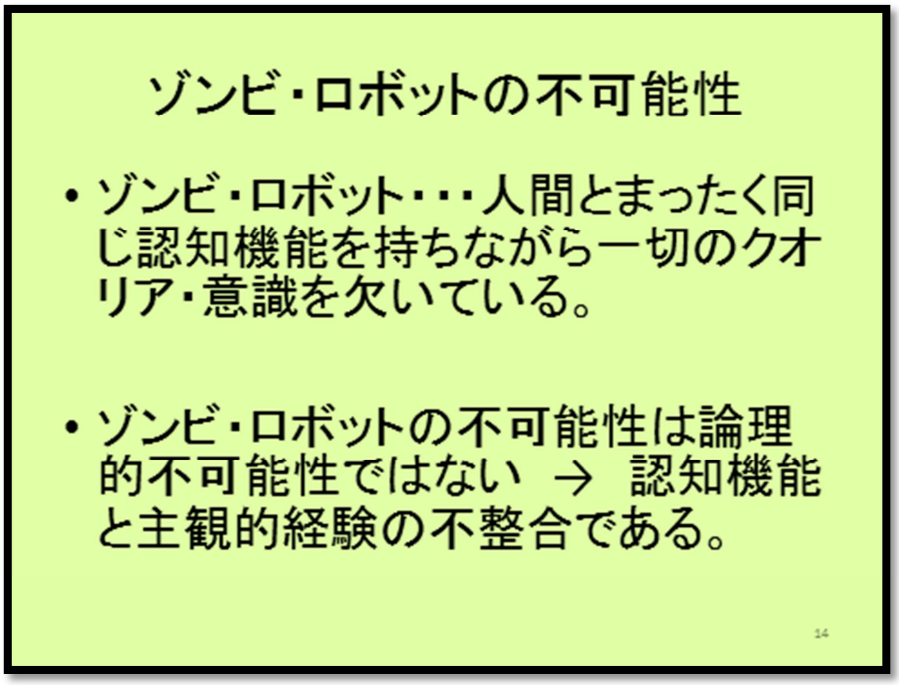


なるほど、心の機能が多重に実現されるのは分かった。だが、クオリアはどうなのか？　クオリアもまた多重に実現されるのか？　だが、クオリアが多重に実現されるというのは、「たんに経験されるだけの、因果的役割を持たない感覚質」というクオリアの本質に反するだろう。したがって、認知機能ロボットが何らかのクオリアを経験するかどうかは、その機能を実現する素材の物理的性質にクオリアがたまたま SV するかどうかにかかっている（ちょうど、赤クオリアがたまたまある脳状態に SV するように）。

われわれがコミットしている（控えめな）物理主義では、あるクオリアがある物理的性質に SV する場合、その物理的性質が実現するなら必然的にそのクオリアも出現する，ということまでは保証される。しかし、クオリアがそもそも何らかの機能的性質に SV するということを示す強力な論証は未だに存在しない。というものクオリアの場合に問題になる物理的性質は 1 階の性質であるのに対して、機能的性質は少なくとも 2 階の性質だからである。機能の多重実現を思い起こそう。2 階の性質は一つでも、それを実現する 1 階の性質は多種多様でありうる。もしあるクオリアが機能的性質に SV することが論証されたなら、その機能を実現するいまだ未知の何らかの物理的性質が出現する場合でもそのクオリアが出現する、ということが論証されたことになるだろう。そんなことまで含意するような主張の論証は、どう少なく見積もっても、そう簡単にはうまくいきそうにないヘビーな課題である。

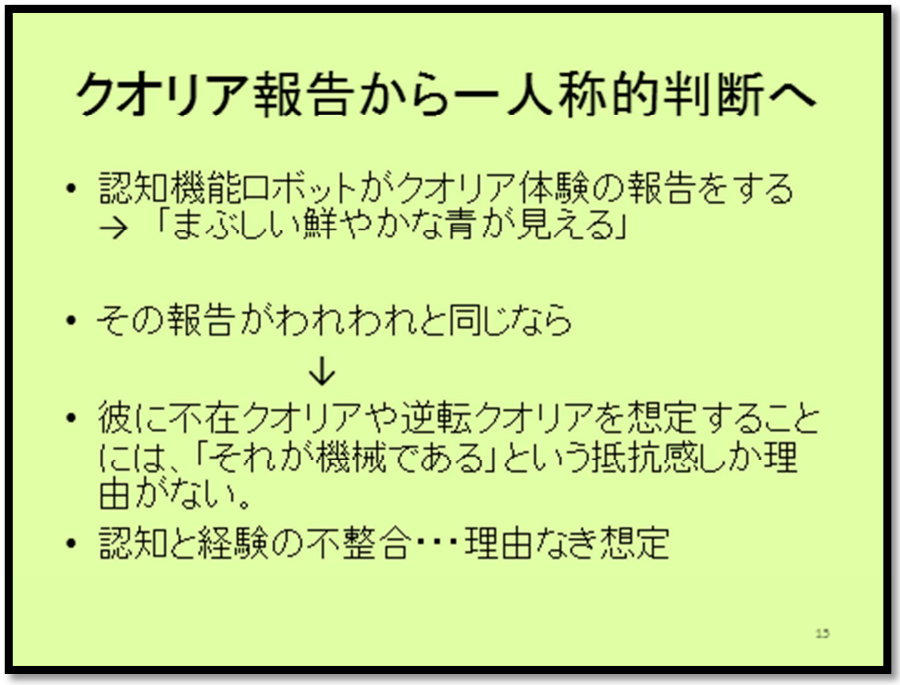
そこで、この論証を正面から企てることは避けて、ここでは、絡めてからの議論を試みよう。それは、人間と機能的に変わりない「認知機能ロボット」が制作されたとして、それがクオリアや現象的意識を持つことができないと主張するのは論理的に不可能ではないとしても、認知の観点からして不合理だ、という議論だ。つまり、一言でいえば、われわれの認知に関する一般理論からして、「ゾンビ・ロボット」は不合理きわまりない想定だというものである。

ゾンビ・ロボットは、たったいま述べたように、「人間とまったく同じ認知機能を果たしているにもかかわらず、すべてのクオリア、すべての現象的意識を欠いているロボット」だ。これもすでに述べたことだが、哲学的ゾンビすら論理的に不可能ではなかった。したがって、ゾンビ・ロボットも論理的に不可能（概念的に矛盾した存在）だというわけではない。その不可能性は、あくまで現実世界における「認知機能と主観的経験との不整合」を根拠とするものである。

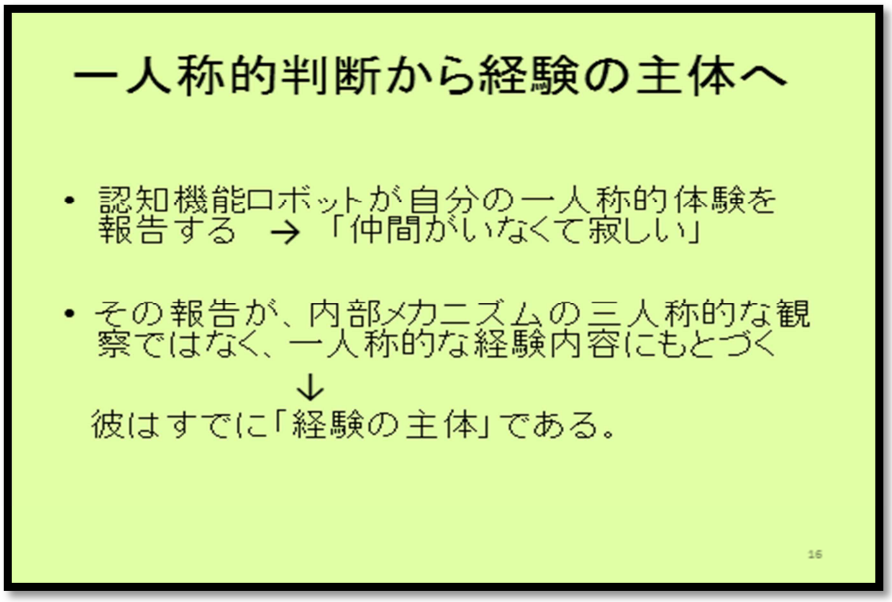


さて、「認知機能と主観的経験との不整合」とは、一体どういうことだろうか？　認知機能ロボットは内省的な自己把握に関してもわれわれと同じ機能を持っているはずだから、そのロボットが、友人のロボットと分かれて悲しそうにしているということもあるだろう。そこで、そのロボットにこう聞いてみよう。「なぜ君は、自分がいま悲しいと分かるんだ？」。「しかじかの仕方で内部センサー16-57 が活性化されているから、自分が悲しいんだと分かるんです」と答えるようでは、彼は自らの感情内容に間接的にしかアクセスできない「役立たずの不自然な」認知システムだ、ということになるだろう。チャルマーズが示唆するように、まともなロボットなら、「なぜって、本当に悲しいから悲しいんだ・・・」と答えるはずだ。したがって、ロボットの認知システムが自分の内部状態を知るとき、それが「経験の主体」と言える程度に高度のものであるならば、そのシステムは、三人称的に自らの内部メカニズムを知るのではなくて、一人称的にそれを知っているはずである。そうでなければ、それは一つの自律した認知システムではないだろう。しかし、一人称的に自らの内部メカニズムを知ること、それこそまさしく意識やクオリアを内容とした現象的経験をもつことに他ならない。

というわけで、認知機能ロボットが「自らのクオリア体験と思えるもの」の報告をし、その内容がわれわれの一人称的な報告と同じであるなら、彼に「本物のクオリア体験」があるのを否定するのは、人間の場合とまったく同じように単なる懐疑論的な懐疑に他ならない。それは、他人の場合のクオリア体験と同じく、論理的には疑いうるが，そう疑うことに懐疑論者の「偏執的な理由」以外のまともな理由が見出せない、いわれなき疑いである。



それでもなお、この種のロボットにクオリアが生じていることを疑うなら、それは、人間ではなく機械の塊にすぎないロボットがクオリアなど経験するはずがない、と頭から決めてかかっているからだ。それはまさしく、ロボットに対する偏見である。（肉の塊ならいい、とでも言うのか？）。

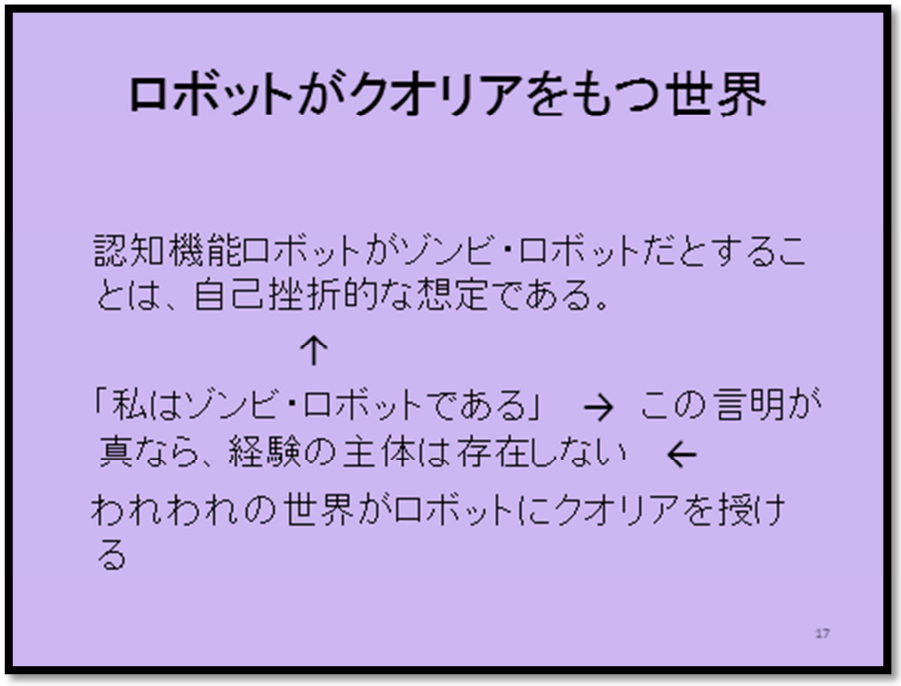


ポイントはこうだ。認知機能ロボットが自分のクオリアや意識に関して一人称の内容に基づく的断（三人称的観察によらない判断）を下すとき、彼はすでに立派な「経験の主体」だということである。そして、すでに経験の主体である存在者が自らの一人称的経験を否定するとしたら、それこそ自己認知の論理からいって「紛れもなく不整合」と言わざるをえないだろう。ところが、ゾンビ・ロボットとは、その定義からして、われわれと同じレベルの認知機能を持つ「経験の主体」でありながら、その経験の存在を認められていないロボットなのである。それゆえ、こ

のロボットが自分に関する「真実」を口にするとしたら、まさしくこの不整合が出現することになる。その真実とは、「自分にはどんなクオリア、どんな意識も生じていない」ということだ。このロボットは定義上、こう述べるべきである。ところが彼は認知機能ロボットなのであるから、自らのクオリア経験を問われれば、これも定義上、われわれとまったく同じように「バラの鮮やかな赤がまぶしいくらいだ」などと答えるしかないのだ！

これこそまさしく、ゾンビ・ロボットが抱える不整合に他ならない。

12-4　結論



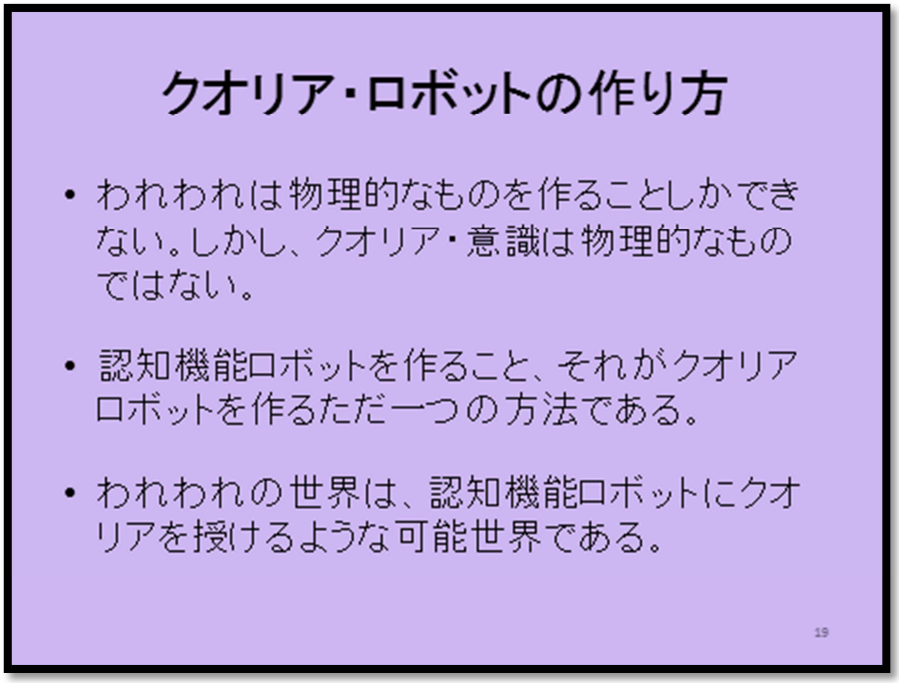
ゾンビ・ロボットの不合理・不整合は、彼の言明「私はゾンビ・ロボットである」が自己挫折的言明であることによく現れている。この言明は、いま述べたように彼に関して真なのだが、もしそうだとすると、この言明をなしているはずの「経験の主体」は存在しないことになる。したがって、この言明は、もし真であればそれが「真だ」と主張している当の存在者が存在しないことになるがゆえに、真なる言明ではなくなる。自己挫折的言明とは、そういう意味である。

したがって、ここで言われる「経験の主体」がなお現実世界における認知の理論からその概念内容を与えられている限り、この自己挫折性は論理的不整合（つまり、いかなる可能世界でも偽）とまでは言えないにしても、われわれの世界（現実世界）では棄却されるべき不条理である。とくに、われわれがまともな認知の理論（認知科学）を追求したいのならば、この不整合性は、われわれの理論の普遍性と限界を印づけるものとして銘記すべきものである。

さて、こうしてわれわれは、ついにクオリア・ロボットを制作する可能性にまで辿り着いた。それを作る唯一の方法は、人間と同じ認知機能を備えた認知機能ロボットを作ることである。現実世界という物理的世界においては、クオリアや意識といった非物理的なものを直接に作る手段はない。われわれは、物理的なものを作ることしかできないのだ。

しかし、いったんしかるべき物理的なものをうまく作ったら、あとは世界の側が、非物理的なものの制作を仕上げてくれる。それこそ、スーパーヴィーニエンスの原理が教えるところである。われわれがしかるべき認知機能ロ

ボットを作る。すると、世界がそれにしかるべきクオリアと意識を授ける。言いかえれば、われわれの世界とは、認知機能ロボットにクオリアを授けるような可能世界なのである。



さて、最後に、私がコミットする「控えめな物理主義」は個体や性質に関して、どのような存在論的見通しをもっているのかを述べておこう。(これは、いわゆる哲学上の存在論、もしくは形而上学と言われるような分野の議論なので、みなさんは無理につき合わなくとも、いままでの議論を理解する上で支障はありません)。この話は、もっぱら可能世界の道具立てを使って述べられる、「偶然的真理としての物理主義」という提案であり、その内実は、「実体一元論と性質二元論が成り立つ可能世界群の中にわれわれの現実世界は含まれる」という主張である。

私が主張したい偶然的真理(ローカルな真理)としての性質二元論は、性質一元論(心的性質＝物理的性質)が必然的真理(あらゆる可能世界で真)となるわけではないような可能世界の配置として描くことができる。つまりあらゆる可能世界から成る論理空間の中では、次のような可能世界群が相互に棲み分けているだろう。(誤解のないように急いでここに付け加えておけば、実体多元論や性質多元論などのようなあらゆる論理的可能性が以下で尽くされているわけではない。最後の図を参照)。

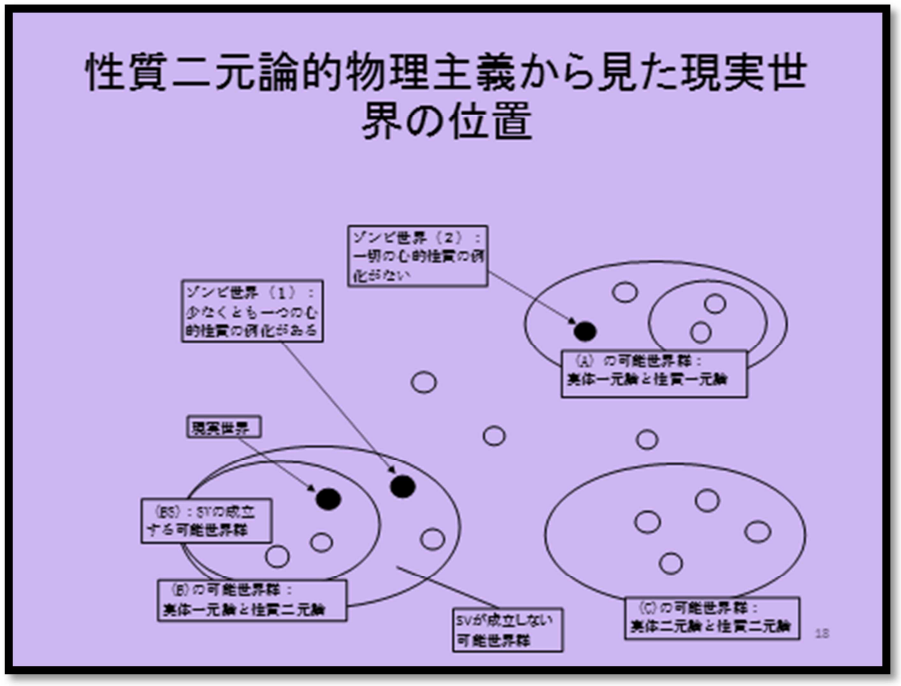
(A) 実体一元論と性質一元論が成り立つ可能世界群。
その中には、物理的個体と物理的性質のみが存在する、あるいは心的個体と心的性質のみが存在する可能世界群があるだろう。J. キムのような現代の還元的物理主義は、前者の可能世界グループに現実世界は含まれると考えている。後者には、伝統的な神秘主義者が理想として思い描くような精神世界が含まれているかもしれない。

(B)実体一元論と性質二元論が成り立つ可能世界群。
それらの世界では、例えば、いかなる心的個体も物理的個体だが、いかなる心的性質も物理的性質と同一ではない。それに対し、すべての個体が心的個体であり、それらが心的性質と異なるものとしての物理的性質をもつ、というような可能世界は考えにくい。私の提案は、現実世界は前者の性質二元論的世界群の中でも、

なおかつスーパーヴィーニエンスが成り立つような可能世界群に含まれるというものだ。

(C)実体二元論と性質二元論が成り立つ可能世界群。

これらの可能世界群においては、例えば霊的実体がいかなる物理的個体とも同一ではないものとして存在し、なおかつ、それらが持つ心的性質はいかなる物理的性質とも同一ではない。またいかなる物理的個体も心的個体と同一であることなく存在し、それらが持つ物理的性質はいかなる心的性質とも同一ではない。ここには、デカルトが考えたようないわゆる二元論的世界が含まれる。



上のスライドでもう一度、私の見取り図を確認しておこう。私が念頭においているのは、(B)の可能世界群であり、この内部には、さらに、心的性質が物理的性質にスーパーヴィーンする可能世界群（BS）と、そのスーパーヴィーニエンスが成り立たない可能世界群（B-BS）を区別することができる。もう一度繰り返せば、私が提案している性質二元論的物理主義は、現実世界がこの BS の可能世界群に属する、という主張である。

文献：

柴田正良　2001　『ロボットの心』講談社現代新書

――――　2003　『心の科学と哲学』（戸田山和久・服部裕幸・柴田正良・美濃正共編）昭和堂

――――　2004　「The Exclusion Problem とエピフェノメナリズム」『理想』No.672、p. 69-82.

――――　2004　「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』信原幸弘（編）勁草書房, pp.119-174.

――――　2006　「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　「機能的性質と心的因果-----キム的還元主義を越えて-----」『思想』No. 982、pp. 53－76.

-------- 2007 「感情の作り方（ニューラルネットワークの場合)」(月本洋と共著)『中部哲学会会報』No.39、pp.1-18.
-------- 2007 「ロボット社会の光と陰」『日本ロボット学会誌』、Vol.26 No.8, pp.864-865.
------- 2008 『感情とクオリアの謎』(長滝祥二・柴田正良・美濃正共編) 昭和堂
------- 2008 「機能する感情・幻想する感情」『心／脳の哲学』(岩波講座哲学 05) 岩波書店、pp.153-176.
------- 2008 「行動ロボットと AI」「感情ロボット」『感情と思考の科学事典』海保博之・松原望（監修）朝倉書店、pp.314-315、pp.318-319.
------- 2010 「感情のニューラル・ネットワークにクオリアを----人工のクオリアに向けて----」『交響するコスモス』(下巻) 松籟社、中村靖子（編）, pp.169-198.
------- 2011 「ロボットの哲学」『応用哲学を学ぶ人のために』戸田山和久・出口康夫(編)、世界思想社、pp.123-134.
------- 2012 「自由な行為者としてのロボット」『これが応用哲学だ』戸田山和久・美濃正・出口康夫(編)、大隅書店、pp135-143.

Chalmers, D. J. 1996 *The Conscious Mind*, Oxford University Press, Oxford. (林一訳『意識する心』白揚社、2001)

Kim, J. 1993 *Supervenience and Mind*, Cambridge University Press.

-------- 1998 *Mind in a Physical World*, The MIT Press, Cambridge. (太田雅子訳『物理的世界の中の心』勁草書房、2006)

Putnam, H. 1981 *Reason, Truth and History*, Cambridge University Press. (野本和幸・他訳『理性・真理・歴史』法政大学出版局、1994)

Shibata, M. 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, pp.345-361.

Stalnaker, R. C. 2003 *Ways a World Might be*, Oxford University Press.