テストから経験へ：チューリング・テストと鉄棒ロボット※

浦 上 大 輔

近年、「シンギュラリティ」という言葉をよく耳にすると思います。いわゆる人工知能が人間と同等以上の「知能」を身につける日がやってくるという話です。ある予測によると、その日は数十年以内にやってくるそうです。そのような予測の根拠は大まかにいうと、「コンピュータの計算速度と記憶容量が現在と同じペースで増大し続けると数十年後に人間の脳を上回る」というわけです。この種の予測を耳にする度に、私は『哲学探究』（ウィトゲンシュタイン 1953）の冒頭の引用を思い出します。

「総じて進歩というものには、実際よりも遙かに偉大に見えてしまうという一面がある」
ネストロイ

シンギュラリティのような楽観的というよりは素朴な、あるいは「線形」な予測がまことしやかに流行している背景には、「何が難しいか」あるいは「何がわかっていないか」ということを我々はよくわかっていないということがあるのだと思います。言い換えると、そもそも「知能とは何か」ということがよくわかっていないのだと思います（「何が難しいか」がわからない、あるいは記述できないということが「知能」の本質に関係していると思われます)。

そこで今日は、「知能とは何か」あるいは「機械は考えることができるか」というテーマについて、古典的な議論を教科書的に振り返りつつ、郡司ペギオ幸夫先生の天才的な？議論を足掛かりにして、私がここ数年研究している「鉄棒ロボット」の強化学習の話につながるように議論を展開していこうと思います。これらのテーマについて、まだまだ考えるべきことがたくさんあるということと、しかもそれは哲学的な議論に留まるものではなく、実験やシミュレーションといった手法を使って科学やあるいは芸術の領域を巻き込むような「経験的」な展開が可能なのだということをお伝えできればと思います。

１．チューリング・テストと中国語の部屋

「機械は考えることができるか」。この問いに正確に答えるためには「機械」と「考える（＝知能がある）」という言葉を定義する必要があるかもしれません。哲学的な議論は往々にして言葉の定義に多くの労力を割きます。一方、チューリング（1950）は、そのような言葉の定義を試みる代わりに別の問題に置き換えて検討することを提案しています。それがいわゆる「チューリング・テスト」です。よく知られているチューリング・テストとは次の

※本稿は早稲田大学及び防衛大学校で行った特別講義の内容を文章として書き起こしたものです。

ようなものです（図１）。

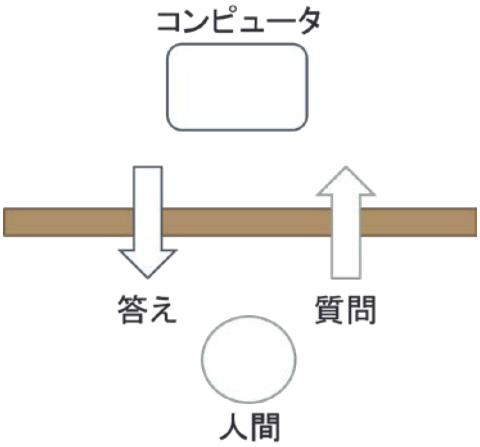

図1．チューリン・グテスト

- ある人間が質問し、コンピュータが回答する。
- 質問者とコンピュータは壁で隔てられていて、質問者は壁の向こうにコンピュータがいるのか人間がいるのかを知らない。
- 質問者は質問し、回答によってのみ壁の向こうにコンピュータがいるのか人間がいるのかを判断する。
- 壁の向こうにコンピュータがいるのか、人間がいるのか、質問者が判断がつかないならばコンピュータは知能をもっているとしてよい。

チューリング・テストは一見素朴ですが、もっともらしくも思え、よく考えると確かにこれ以外の方法はないようにも思えます。実はチューリング自身が提案したものはこれとは少し違います。その違いが我々の議論を大きく転回するのですが、それについては本稿の終盤で述べます。ここではチューリング・テストを上記のようなもとしておきます。チューリングは、チューリング・テストを提案した論文の中で想定される反論を列挙して一つ一つ再反論しています。それらは平易に書かれていますが、今日までのチューリング・テストをめぐる議論のほとんどを先取りしていてとても説得的です。「知能とは何か」といった問題に関心がある方は、まずはチューリングの論文を読まれることお勧めします。

チューリング・テストをめぐる議論において必ずと言ってよいほど言及されるのが、サール（1983）の「中国語の部屋」です。中国語の部屋とは次のようなものです（図２）。

- 中国語をまったく理解していない英国人＝サールがある部屋の中にいる。

- その部屋には、中国語で書かれた例文集と英語で書かれた規則集があり、この規則集には中国語の質問に対してどの例文で答えればよいかが全て書かれている。
- 部屋の外から中国語で質問をすると、サールは規則集にしたがって例文集から中国語で回答することができる。
- このとき、質問者は部屋の中の人＝サールが中国語を理解していると判断するが、実際にはサールは中国語を理解していない。

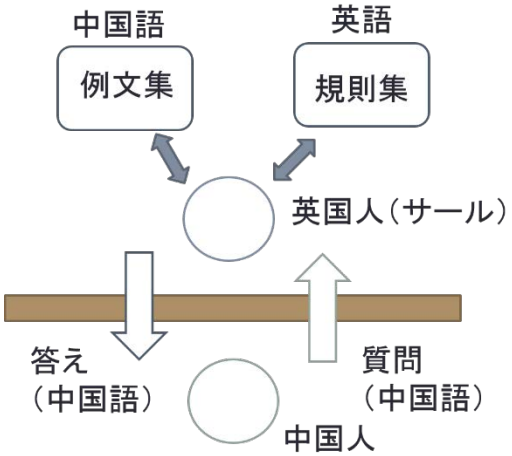

図2. 中国語の部屋

以上のように、質問と回答のみに基づくテストでは、実際には中国語を理解していない人を理解していると誤って判断する場合がありえるのでチューリング・テストは有効でない、というのがサールの主張です。サールの主張は十分に説得的でしょうか？ サールの主張に批判的な考えを持つ人はもちろんのこと、サールの結論には同意する人でさえ、サールの議論は不十分に感じるのではないでしょうか？ 一方で、サールの中国語の部屋は、多くの議論を喚起したという点においては大いに評価されるべきでしょう。

さて、中国語の部屋に対するチューリング・テストを擁護する側からの再反論で、代表的なものは次のようなものです（※サールをめぐる議論については服部（2003a）の論文を参考にさせて頂いている）。

- 中国語を理解しているのはサールではなく、サールと例文集と規則集を含む「部屋全体」である。

このような議論は「システム論」と呼ばれます。システムとは何か？ とりあえずここでは、「入力と出力を持つもの」としておきます。システム論の主張は強力で、これに反論するの

は容易ではないと感じるかもしれません。そのように感じるのは、我々がシステムという見方に慣れているからでしょう。サール自身も、想定される反論の第一にシステム論を挙げています。また、すべての反論は結局のところシステム論と似たようなものになるだろう、と予想しています。そして、サールの反論（再々反論）は以下のようなものです。

- 例文集や規則集を１人の人（サール）の頭の中に「押し込める」ことはできるはずだ。だとしても、サールは中国語を理解していない。それにも関わらず、中国語を話すということは想定できる。

ここで、「押し込める」とは中国語の例文集は画像として丸暗記し、英語で書かれた規則集も暗記するということです。そのような場合は中国語を理解しているとは言えない、というのがサールの主張です。このようにマニュアルを押し込められて中国語を理解せずに中国語を話す人を「サールのデーモン」と呼んでいいでしょう。サールのデーモンは、あまりにＳＦ的でまっとうな議論ではないと感じるかもしれません。そもそも、サールのデーモンは中国語を理解していると言っていいのではないか？ そのような疑問も湧いてきます。

これまでの議論を整理します。サールの議論は結局のところ次のようなものです。

- 入出力による知能の判断は間違っている。なぜなら、知能がなくても適切な入出力は想定できるからだ。

このような主張はほとんどトートロジー（同語反復）です。主張の前提と結論が同じことを意味しています。また、「中国語の部屋のような想定は本当に可能か？」ということにも議論の余地がありそうです。その点を検証したのが「足し算の部屋」なのですが、それは後ほど述べます。一方、システム論が意味しているところを整理すると次のようなものです。

- 入出力による知能の判断は擁護される。なぜなら、適切な全体を想定すればそこには入出力しかないからだ。

この主張もほとんどトートロジーです。また、ここでいう「適切な全体」を設定することが可能か、という問題があります。それは「フレーム問題」という大問題に関係します。フレーム問題とは、ある状況において考慮すべき事柄の範囲＝フレームを決めようとしても際限がないことをいいます。中国語の部屋のような状況では、例えば部屋の明かりです。部屋の中の人が例文集や規則集を読むためには明かりが必要で、部屋の中に蛍光灯などを用意しなければならないでしょう。蛍光灯に電気を供給するためには、部屋の外から電線を引く必要があります。電線に電気を供給するためには、発電所が稼働している必要があります。

発電所を稼働するためにはそのエネルギー源となる石油や核燃料が必要で、くわえて地震や津波などの自然状況や、さらには地域の人々の市民感情や政府の方針といった社会状況も関係してきて云々、というように際限がありません。システム論が想定する適切な全体とは一見素朴のように思えて、その内実はとても強い要請が含まれているのです。システム論の立場に立てば、それは定義上可能であるということになります。

サールもシステム論もどちらもトートロジーで議論が全くかみ合っていないようですが、システム論の方により大きな間違いがあると思います。なぜなら、システム論は結局のところ、「チューリング・テストは知能の定義である」と主張していることになるからです。しかし、そもそもチューリング・テストは知能とは何かを定義する代わりに提案されたものです。一方、サールが中国語の部屋で主張したことは、彼自身の論証の是非はともかく、「テストそれ自体」と「テストされるもの」は違うということです。ここで、テストそれ自体とは「入力と出力のペアの集まり」で、テストされるものはとは「知能」のことです。チューリング・テストは知能の定義ではなく、あくまでもテストであるということに留意するなら、「入力と出力のペアの集まり＝知能」ではありません。この限りではサールは全く正しいと言えます。チューリング・テストがテストとして有効かどかうかは、また別の問題です。いずれにしてもチューリング・テストは知能の定義ではない、ということは我々の議論において極めて重要です。

２．足し算の部屋

チューリング・テストが知能の「定義」ではなく文字通り「テスト」であるならば、それは経験的に有効か否かということは、問われてよい問題のように思えます。逆にいうと、「中国語の部屋」のような想定は実際に可能か？、ということです。レベック（2009）による「足し算の部屋」はこのことを、つまり「行動（入力と出力）だけをまねることは可能か？」ということを、計算量や記憶容量の観点から検討したものです。「足し算の部屋」とは次のようなものです（※足し算の部屋については中島（2011）の論文を参考にさせて頂きました）。

- タスク：１０桁の数を２０個足す。
- 足し算の計算ができない人間と足し算のマニュアルを想定。
- 人間がマニュアルを丸暗記しても「足し算を理解していない」といえるようなマニュアルが作れるか？

さて、このような問題設定のもと、まずは一番素朴なマニュアル、全ての組合せを表にすることを考えます。それをマニュアルＡと呼ぶことにします。マニュアルＡは次のようなものです。

- 最初の数と同じ番号の章に行く。
- その章内で２番目の数と同じ番号の節に行く。
- その節内で３番目と同じ番号の副節に行く。
- これを２０個の数全部にわたって繰り返す。
- そこには最大１２桁の数が書かれている。
- その数を紙に写して部屋の外に返す。

このとき、「１番目の数に対応する１０の１０乗個の章が必要で、それぞれの章に１０の１０乗個の節が必要で、…」ということを計算すると、全部で１０の１０乗の２０乗＝１０の２００乗の項目が必要となり、それは宇宙に存在する分子の数（１０の１００乗個程度）より大きくなってしまいます。したがって、マニュアルＡを作成することは実際には不可能です。足し算という簡単な問題でさえマニュアルを作成することが不可能なので、中国語の部屋は当然不可能だというのがレベックの主張です。

たしかにマニュアルＡは不可能ではあるけれど、なんだかのデータ圧縮法を使えばマニュアルが作成できるかもしれないという反論はありそうです。そのようなマニュアルの例の１つは次のようなものです。

- １０桁の数を１桁ごとに分解。
- １桁の数の１０×１０の表を使う。
- 繰り上げについてもマニュアルで指示。

これをマニュアルＢと呼ぶことします。マニュアルＢの記憶容量は微々たるもので、それは実現可能でしょう。しかし、マニュアルＢを丸暗記した人は足し算を理解していないと言えるでしょうか？　レベックは、そのような人は正に足し算を理解しているのだと主張します。なぜなら、マニュアルＢは我々が小学校で習う足し算の方法そのものだからです。レベックはその他にも幾つかの圧縮方法を検討し、１０の２００乗の記述を実際に格納できるサイズに圧縮するためには、何らかの形で足し算の本質を表現した「アルゴリズム」を用いるほかはないと結論しています。

足し算の部屋に対する反論を考えてみましょう。１つ思いつくことは、「足し算の理解を判断するために１０桁の数を２０個足す必要があるのか」ということです。我々は足し算を理解していますが、実際に１０桁の数を２０個足したことがある人はいないでしょう。それでも足し算を理解しているとされるのは、小学校で行われているテスト程度の問題で足し算の理解をはかるのは十分であると考えられているからです。そして、小学校で行われているテスト程度の足し算、たとえば３～４桁であれば表を作ることは十分に可能です。また、

たしかに１０桁の数を２０個足す組合せの「全体」を表にすることはできませんが、実際にテストされるのはその一部分です。その「部分」については表を作ることが可能なのではないでしょうか？　これは中国語の部屋についても当てはまります。中国語の可能な例文の全ては足し算と同様に膨大な数でしょうが、実際に会話で使われている中国語の文はそれよりは遥かに小さい数で宇宙の分子の数ほどでないと想像されます。いずれにしても、ここでは部分と全体の関係が問われているということを心に留めておいてください。

さて、「テストそれ自体＝入出力」と「テストされるもの＝知能」が異なるという点においては、サールと足し算の部屋における主張は一致しています。それでは、足し算の部屋が主張するように、知能＝アルゴリズムでしょうか？　サールは同意しないでしょう。サールは知能に不可欠なものとして、「志向性」あるいは「意味」を挙げています。そして、志向性や意味は脳の神経生理が深く関係しているはずだから、システム論（チューリング・テスト）が暗に措定しているアルゴリズムとその実装という二元論は受け入れ難いというのがサールの結論です。私はサールの議論にかなり好意的ですが、「神経生理を支える物理学や化学もまたシステム論の一形態である」という点を付け加えて強調しておきたいと思います。このへんの議論を補足するために、「コネクショニズム」や「分散表象」などをめぐる議論にふれたいと思います。

３．コネクショニズム、ロボット論、トータル・チューリング・テスト

知能はアルゴリズムではない。人間の脳にアルゴリズムはない。そのように考える立場の一つがコネクショニズムです（ただし、後で見るようにサールとコネクショニズムはかなり異なります）。コネクショニズムとは、人間の認知活動（知能）をニューラルネット、つまりニューロンモデルの結合（＝コネクション）で実装しようという立場です。ニューラルネットは画像認識などに優れ、近年、その発展形であるディープランニングが人間の認識能力に迫る能力を示して話題になっています。知能とは何かといった哲学的な議論の文脈おいては、コネクショニズムは「分散表象」（ラメルハート 1986）という概念と結びついています。分散表象は「古典的計算主義」と呼ばれる立場に対立するものとして提案されたものです。それは、心のモデルを作るか（＝古典的計算主義）、脳のモデルを作るか（＝コネクショニズム）といった対立でもあります。

コネクショニズムについて詳しく論じる前に、まずは古典的計算主義について説明します。古典的計算主義とは、人間の思考を「表象」とその「変化操作」とみなす立場です。ただし、自らの立場を古典的計算主義だと宣言した人たちがいるわけではなく、それを批判する立場の人たちによって名付けられたものです。ここで、「表象」とは記号や言語のようなものです。たとえば、現実のリンゴを見たときに心の中に浮かぶ何かです。あるいは、目の前にリンゴがなくても「リンゴが食べたいなぁ」と思ったときに浮かぶ何かです。それは

リンゴのアイコンのようなものかもしれませんし、「リンゴ」という言葉かもしれません。いずれにしても、古典的計算主義では「人間の心の中に世界の事物に対応する何か」が存在して、その変換操作が思考という訳です。たとえば、「リンゴ」という表象から「食べ物」という表象を導く操作です。より端的にいうと、「１＋１」という表象から「２」という表象を「計算する」ことが思考だというわけです。表象の変換操作のことをアルゴリズムと言い換えてもいいでしょう。古典的計算主義は、チューリング・テストを擁護する立場に近いと考えられそうです。ただし、システム論よりは心の中の変換操作に関心があります。一方で、心の中と現実との対応関係、「リンゴ」という表象と現実のリンゴとの関係については素朴です。両者をどのように対応づけるかという問題には、技術的にも哲学的にも困難が存在し、いわゆる「記号接地問題」として知られています（図３）。

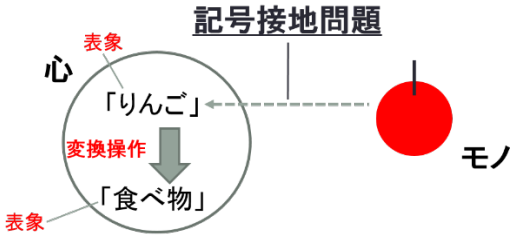

図3. 古典的計算主義

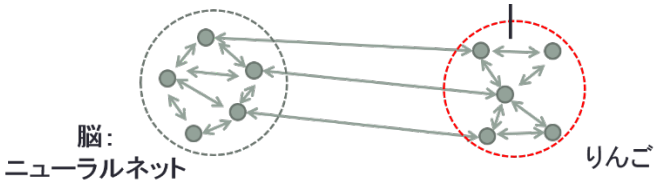

図4. コネクショニズム

一方、コネクショニズムでは古典的計算主義が想定するような表象は実在しない、と考えます。そこに在るのは、モノとモノの「相互作用」であると考えます。人間の認知活動は、脳を構成する神経細胞とリンゴを構成する分子の相互作用であるというわけです（図４）。そもそも記号（＝表象）が存在しないので、記号接地問題に悩まされることもありません。一方で、リンゴの認知は成立しているわけで、それを成立させているは脳の全体としての状態であり、それを「分散表象」と呼びます。コネクショニズムでは脳のモデルとしてニューラルネットを想定しているので、分散表象とは具体的には「複数のニューロンの発火状態を表すベクトル」を意味します。ここで、「リンゴ」に対応するニューロン（いわゆるおばあ

ちゃん細胞）が存在するわけでないことが重要です。ニューロンのネットワークが全体として様々な状態をとり、それら内のある状態がリンゴに対応し、別の状態がバナナに対応します。つまり、リンゴの記憶（＝表象）は、ネットワーク全体に「分散」しているわけです。

コネクションズムからの批判の対象は古典的計算主義だけにとどまらず、サールの中国語の部屋にも及びます。つまり、「コネクショニズムが提唱するシステム＝ニューラルネットでは、中国語の部屋にあるようなデータ（中国語の例文集＝表象）やプログラム（英語の規則集＝変換操作）といったようなものは存在しない」ので、中国語の部屋の議論はコネクショニズムには当てはまらないというわけです。これに対して、サールは「中国語のジム」と呼ばれる議論で反論しています（※コネクショニズムをめぐる議論ついては服部（2003b）の論文を参考にさせて頂きました）。

- 英語のみを話す人が多数いるホール（ジム）を想像する。
- この人々は規則集に従って、ニューラルネットにおけるノードとシナプス（ニューロンとその結合）と同じ作業を遂行する。
- このジムにいる人は１人として中国語を理解しないが、適当に調整すれば、このジムは中国語の質問に対して正しく答えることができる。

サールの議論の執拗さには感心します。しかし、この反論はコネクションズムにとってはまったくの見当違いで、「中国語を理解しているのはジムにいる個々の人ではなく，ジム全体（＝分散表象）である」ということになります。そして、その主張はシステム論と一致することになります。

コネクショニズムはシステム論の一種なのでしょうか？　たしかに、ニューラルネットも全体としては一つのシステムです。しかし、そのことを認めると、「ニューラルネットは（入出力だけをみれば）古典的計算主義の実現形態の一つでしかない」ということになります。ニューラルネットも結局のところ、なんだかのアルゴリズムを実現しているわけで、知能＝アルゴリズムという点においては、コネクショニズムと古典的計算主義に違いはないというわけです。また、ニューラルネット自体は数理モデルなので、それは記号や数式、つまり表象によって構成されています。個々のニューロンは古典的計算主義が想定するような計算をおこなっており、その結合も計算です。つまり、ニューラルネットはプログラム可能なアルゴリズムで実現されています。ニューラルネットは外側から見ても（入出力だけに着目しても）、内部を見ても（個々のニューロンとその結合に着目しても）一種のアルゴリズムであるということになります。このような批判は、チューリング・テストをめぐる議論の主要な一つである「ロボット論」へとつながっていきます。

ロボット論では、チューリング・テストと中国語の部屋をめぐる議論の欠点は入出力をテキストベースに限定したことにあると考えます。これまでの議論では特に明言しません

でしたが、チューリング・テストにおける入出力はチャットのようなものを想定していました。実際、チャット上で人間と見分けがつかないコンピュータはかなり存在するようですが、それらのコンピュータが人間並みの知能があるとは言えないようです。ロボット論では、そのようなテキストベースのやりとりだけでは中国語を理解している、あるいは知能があるとは言えないと考えます。代わりに、ロボット論は以下のように主張します。

- コンピュータに感覚器官と運動器官に相当するものを取り付けて、世界に対して自ら働きかけることができるようなシステム、つまりロボットを想定する。そのようなロボットが人間の行動と区別つかないなら、そのロボットには知能がある。

ロボット論はシステム論の拡張であると考えられます。システム論には「部屋全体」とは何かといった問題があり、それは「フレーム問題」という大問題につながることは既に述べたとおりです。言い換えると、システム論が想定する「適切な全体」とは入出力をテキストベースに限定することであり、それはテストする側が「記号接地問題」をあらかじめ解決しておくことを意味します。ロボット論は、そのようなシステム論の弱点をみずから解決するようなシステム＝ロボットを想定し、そのようなロボットであれば知能があると考えてよいだろうと主張します。

はたして、ロボット論が主張するようなロボットを作ることができるのでしょうか？現在のところ、そのようなロボットは作られていないようです。そもそも、ロボットが人間と区別がつかないようにするためには、ロボットをどこまで作り込めばよいのでしょうか？ 文字通り意味で考えますと、ロボットの外見も重要になってきそうです。表情や仕草は人間のコミュニケーションにとって重要ですし、現存するロボットのほとんどは外見ですぐにロボットであるとばれてしまいます。これらに対して、「トータル・チューリング・テスト」（ホーランド 1991, 石黒 2011）と呼ばれるものが提案されています。トータル・チューリング・テストでは、人間とそっくりのアンドロイドをつくって人間と比較します。チューリング・テストにおけるテキストベースの対話を、人間のもつ全てのモダリティ（判断と感じ方）に拡張しようという試みです。トータル・チューリング・テストの試み自体は興味深いものです。しかし、ロボットをどこまで作り込めばいいのかという問題には際限ないように思えます。人間のもつ全てのモダリティにテストを拡張するということは、肌触りや臭いも全く同じにする必要があります。そのためには、おそらく分子レベルで人間とそっくりにする必要があるでしょう。そのように分子レベルで人間とそっくりのアンドロイドを作ってテストしようという考えは、「トータル・トータル・チューリング・テスト」と呼ばれます。トータル・トータル・チューリング・テストでは、テストのやり方によっては、ロボットの外見だけではなく中身も分子レベルで人間とそっくりにする必要があるでしょう。しかし、外見も中身も分子レベルで人間とまったく同じアンドロイドは正に人間であって、それは「機械」とは言えないのではないでしょうか？　もちろん、人間は一種の機械だ

という考えはあるでしょう。いずれにしても、ここでは機械の「定義」と問われてしまっています。しかし、そもそもチューリング・テストは機械や知能を定義する代わりに提案されたものです。その点に留意するのであれば、トータル・チューリング・テストやトータル・トータル・チューリング・テストは、チューリング・テストの本来の狙いから逸れてしまっていると言わざるをえないでしょう。また、人間と分子レベルで同じアンドロイドを作るという議論は、機械と何かという論点にくわえて、「知能を理解するとはどういうことか」という論点にも関係してきます。人間とまったく同じアンドロイドができたとしても、そのアンドロイド＝人間が人間と同じように話したり行動したりするのは当然であって、知能とは何かを理解することに何一つ寄与しないように思えます。事実、人間とまったく同じアンドロイド＝人間は世界にあふれていますが、知能とは何かは未だに議論が尽きないわけです。

同様の批判は、コネクショニズムにも当てはまります。ニューラルネットは脳のモデルですが、それはどこまで作り込めばいいのでしょうか？ 先に、ニューラルネット自体は数理モデルであるから結局のところはアルゴリズム、つまり表象とその変換操作で構成されているので古典的計算主義と本質的な違いはないといった批判がありうると述べました。このような批判に反論する方法は二つあるようです。一つは、ニューラルネットは脳のあるレベルの近似的モデルであって、脳の分子レベルでの数理モデルを作成すればその批判は当てはまらないというものです。もうひとつは、知能を実現する分散的な計算をデジタルコンピュータによってではなく、アナログ回路やロボットなどの身体それ自体によって実現しようというものです。サブサンプション・アーキテクチャ（ブルックス 1986）はその代表的な成功例です。どちらも表象というものを徹底的に排除しようという点においては一致しています。前者は数理モデルで、後者はモノで実現しようという点が違うようですが、脳の挙動を分子レベルで計算する、あるいは脳と同じ機能を実現するためには、どちらも結局のところ脳それ自体をもってくるしかないようにも思えます。そして、脳それ自体をもってきても知能とは何かを理解することに何一つ寄与しないことは、人間とまったく同じアンドロイドの場合と同様です。

「表象」という概念は常に批判の対象になってきました。一方で、表象という概念が議論の俎上に何度も上がるのには、次のような事情があるからだと思います。

- およそ表象（記号あるいは言語）とその変換操作といったものを抜きにして人間の認知活動（知能）を説明しても、それは人間の認知活動（知能）を理解したことには到底ならない。

この指摘は我々の議論にとって極めて重要です。ここでは、知能に対する理解の方法が問われています。知能を理解するためには、素朴な、いわゆる還元主義では不十分ではないかと

いうわけです。コネクショニズムの場合、分散表象という概念は表象を批判的に吟味するという意味においては有効であったと思われます。しかし、その有効性は批判の対象である表象とニューラルネットがセットになって議論が展開している間でのみ発揮されたのであって、表象をニューラルネットによって還元するに議論が至ると、知能の理解はさらに先へと隠れてしまったように思われます。そのような議論に至るのではなく、分散表象という表象として中途半端な状態の中途半端さをより徹底することが、知能の理解につながるのではないでしょうか？ 私が以前に太田宏之さんとの共同研究で提案した「概念形成型ニューラルネット」は、そのような発想に基づくものです（Uragami 2014a）。その詳細を解説することは別の機会としますが、興味のある方は論文を読んでみてください。

コネクショニズムをめぐる議論はこのへんで終わりにしようと思いますが、最後に二つのことについて補足しておきます。一つはアフォーダンスについて。アフォーダンスとは、生物に行動を促す情報（＝表象）のようなものです。リンゴを例にしますと、アフォーダンスとは「食べられる」という性質のようなもので、それを知覚した生物に「食べるという行動」を促します。このとき、アフォーダンスは知覚されるリンゴ（モノ・環境）の側に備わっているとされる点が重要です。古典的計算主義では、リンゴの認知においてリンゴの表象は人間（生物）の心の中にあって、心の中の変換操作を経て「食べ物」という表象が導き出されるのでした。一方、アフォーダンスは表象のようなものですが、それは心の外に実在する点と行動を直接的に促す点が特徴です。アフォーダンスは、記号接地問題や思考と行動のギャップを解消しようという概念です。それは、コネクショニズムなどとはちょうど逆側からのアプローチで表象を批判的に解体するものです。そのような意味で、アフォーダンスは興味深い概念です。しかし、そのアプローチを徹底すると、コネクショニズムが行きつく先と裏表の関係でぴったりと一致します。それは一種の物理主義になります。このへんの事情についての詳しい議論は私と郡司ペギオ幸夫先生の共著論文（浦上 2009）を参照していただければと思います。ここでは、アフォーダンスによるアプローチはコネクショニズムと同様に「知能の理解のあり様が問われることになる」ということだけは述べておきたいと思います。

もう一つは分子レベルでの理解について。コネクショニズムなどの議論の行きつく先が分子レベルでの理解であったことはこれまでに論じてきたとおりです。それは、知能を「計算」とみなす立場から「相互作用」とみなす立場への移動が行きつく先です。計算をモノとモノとの相互作用に還元する。そのような還元主義は知能の理解として不十分ではないかという議論については既にふれました。それにくわえて、そもそも相互作用の記述は計算ではないかという批判も重要です。分子の相互作用は微分方程式によって記述されることが一般的です。微分方程式は一種の計算であり、システム論です。したがって、システム論に当てはまる批判は微分方程式に基づく分子レベルでの理解にも当てはまります。これまでの議論でみてきたように、我々の議論はシステム論を批判することが焦点の一つとな

っています。次節以降の議論を先取りすると、チューリング・テストをシステム論から知能の理解へと転回するために、テストではなく「経験」を志向することになります。次節以降の我々の議論はチューリング・テストのレベルでなされますが、分子レベルでも同様の議論が可能であることは述べておきたいと思います。そして、分子レベルで「経験」を志向することは正に松野孝一朗先生が先駆的に実行してきたことであり、それは「内部観測」と呼ばれています。

4．テストと経験

前節までででは、チューリング・テストをめぐる議論を教科書的に紹介してきました。紹介した様々な立場や主張のほとんどは、議論が行き詰っているように思えます。一方で、これらの議論を通して明らかになったこともあります。その中でもこれから議論にとって大事なことをまとめると以下のようになります。

- チューリング・テストは知能や機械の「定義」ではなく「テスト」である。
- テストでは、部分（テストそれ自体＝入出力）と全体（テストの対象＝知能）の関係が問われる。
- 理解の方法が問われている。

これらを踏まえて、我々の議論はいよいよ本題へと入っていきます。その足掛かりとして上記の二つ目の論点である部分と全体の関係について、幾つかの思考実験によって事態を明確化したいと思います。

まず始めに、「確率的な部屋」とでも呼べそうな状況を考えます。それは次のようなのです。

- 部屋の中には乱数発生装置があり、それを使って入力に対してランダムに出力する。
- この部屋はある確率でチューリング・テストに合格するが、知能はない。

もちろん、確率的な部屋がチューリング・テストに合格する確率は極めて小さいでしょう。統計的には無視できるかもしれません。しかし、チューリング・テストは統計的に正しいに過ぎないのでしょうか？ 統計的な正しさというのは、知能とは何かといった哲学的な議論にとってはずいぶんと居心地の悪いもののように思えます。しかも、ここで問題になっていることは確率や統計に限定されることではありません。そのことは次のような、より極端場合を考えると明確になります。それは「カンニングの部屋」とでも呼べそうなものです。

- チューリング・テストをある人間に行い，合格と判断した（当然のことだが）。そのと

きのテストパターンを記録した表を作成する。

- その表を備えた部屋は「同じ」チューリング・テストに合格するが、知能はない。

カンニングの部屋に対する対策として真っ先に思いつくことは、先におこなったチューリング・テストとは別のテストをするというものです（別のテストパターンを使用する)。しかし、別のテストをするということは、そもそも先におこなったチューリング・テストを否定することにならないでしょうか？　カンニングの部屋と人間、どちらも同じテストに合格したという事実のみを判断の根拠にするべきで、それ以外のこと、たとえばテストの順番などはチューリング・テストの範囲外のことのようにも思えます。

確率的な部屋とカンニングの部屋に共通して問題となっていることは、「個々のチューリング・テスト」と「チューリング・テストの全体」は違うということです。ここでは部分と全体の関係が正面から問われています。では、両者をつなぐような「一般的なチューリング・テスト」は可能でしょうか？　もしそのようなものが定義されるのであれば、次のような事態が考えられると思われます。それはカンニングの部屋を一般化したようなもので、「チューリング・テストの部屋」とでも呼べそうなものです。

- 個々のチューリング・テストが一般性を持つためには、テストパターンは表あるいはアルゴリズムで定義されていなければならい
- その表あるいはアルゴリズムを備えた部屋はチューリング・テストに合格するが、知能はない

この部屋には知能があると認めてよいだろう、という反論があるかもしれません。しかし、この部屋に知能があるのであれば、チューリング・テスト自体が知能の定義になってしまいます。それは知能の定義を回避しようというチューリング・テストの本来の主旨に反します。ここには、チューリング・テストを定義するためにはチューリング・テストに合格するアルゴリズムが必要になるという「自己言及的」な構造が存在します。それは、知能について考える際には不可避的に直面する問題のように思えます。そもそも、我々の知能はそれ自身の全てを理解することができるのか？ そこには「記述不可能」なものが関係しているように思えます。そして、知能の本質は、記述可能な領域ではなく記述不可能な領域にこそあるのではないしょうか？ 今日の講演の冒頭で、シンギュラリティを批判的に取り上げて、人工知能にとって「何が難しいか」あるいは「何がわかっていないか」ということがわかっていないと述べましたが、それは我々が愚かだからではなく、「知能とは何か」を考える際には本質的な問題であるように思えます。なぜなら、知能それ自体が知能の全体について考えるとき、そこには自己言及的で記述不可能な領域が関係してくることは不可避なことであると予想できるからです。

チューリング・テストは知能の記述不可能な領域を、テスト＝「対話」という手法によ

ってなんとか経験的に理解しようという試みであるとも考えられます。それは、精神分析が対話によって無意識をあぶり出そうとすることに似ているかもしれません。対話は、問う側と答える側の非対称性を利用して、自己言及的な方法ではアクセスすることが困難な領域に光を当てるものです。先ほど我々はチューリング・テストの自己言及性を確認しましたが、そこではテストにおける非対称性をあえて無視していました。テストする側とテストされる側には非対称性があります。例えば、因数分解や方程式です。因数分解や方程式の答えを見つけることは大変ですが、ある答えが合っているかどうかをチェックすることはそれほど大変ではありません。そこには非対称性があります。チューリング・テストにおいてこの非対称性は本質的な役割を果しえるか？ それは我々の知的活動の根本的な態度に関係する問題です。一般的に、数学や科学では非対称性をなんとか対称な双対図式に回収しようとします。例えば先に例として挙げた方程式の場合は、ガロア理論という双対図式が知られています。郡司ペギオ幸夫先生がしばしば言及するアジャンクションとは、このような双対図式を抽象的に一般化したものです。郡司先生はアジャンクションを「脱構築」して、アジャンクションによって捨象されたもの＝「潜在するもの」を炙り出そうとしてきました。他には、柄谷行人は言語的なコミュニケーションに「教える／学ぶ」という非対称性を強調する観点を持ち込むことによって、そこに潜在するものを焦点化しています（柄谷 1992）。

実は、チューリング自身が提案した（本当の）チューリング・テストは、テストにおける非対称性をより強調したものになっています。以下では、それを「本当のチューリング・テスト」と呼ぶことにします。それは次のようなものです（図５）。

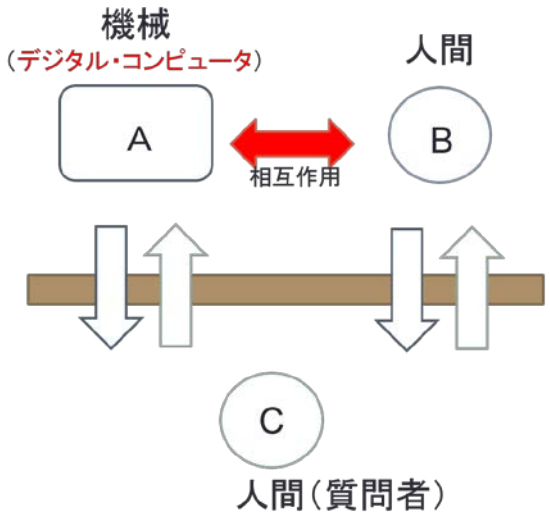

図5. 本当のチューリング・テスト

- 機械（A）と人間（B）と質問者（C）の３名でおこなう。
- 質問者（C）は他の２名（AとB）から隔離された部屋にいる（回答は紙に書くか、タイプライターでタイプする)。
- 質問者（C）の目的は、AとBのどちらが機械でどちらが人間かを判断すること。
- Aの目的は、質問者（C）に間違った判断をさせること。
- Bの目的は、質問者（C）を助けること。

チューリング自身はこれを「イミテーション・ゲーム」と呼んでいます（「ゲーム」という呼び方はウィトゲンシュタインの「言語ゲーム」を連想させます)。なぜチューリングはこのように入り組んだテストあるいはゲームを提案したのでしょうか？ 先に述べたようなチューリング・テスト（「普通のチューリング・テスト」と呼ぶことにします）では不十分だと考えたのでしょうか？ チューリング・テストをめぐるほとんどの議論は、普通のチューリング・テストについて議論されています。そこでは、本当のチューリング・テストは無視されています。一方、郡司先生は両者の違いに着目し、本当のチューリング・テストは「テスト」を積極的に解体し「経験」へと転回するものであると指摘しています（郡司2012)。そのことが書かれている郡司先生の論文はチューリング・テストの生誕１００年を記念した論文集に掲載されていますが、その論文集に掲載されている１０以上の論文のうちで、このことを正面から指摘しているのは郡司先生の論文だけでした。もちろん、チューリング自身の意図は本人に聞かなければわかりません。しかし、チューリング自身の意図がどうであろうと、チューリング・テストがこのような形で提案されたことに意味があるように思えます。その意味を積極的に評価するならば、チューリング・テストはその起源においてはそれ自体を乗り越えるものを内包していたということになります。

普通のチューリング・テストと本当のチューリング・テストの違いは、テストと経験の違いを際立たせるものです。その点を整理します。通常、およそテストと呼ばれるものは２者間で行われます。普通のチューリング・テストもそうです。テストする側はあるモデルに基づいて入力（問題）を選出し、テストされる側の出力（回答）によってそのモデルを変更します。モデルと入出力が一致してモデルを変更する必要がなくなるとテストは終了し、ある判断がなされます。テストの終了は、非対称性が解消されたことを意味します。一方、本当のチューリング・テストは３者間で行われます。この第３者が２者間でのテストが終了することを妨げるので、判断は常に一時的なものでしかありえません。また、テストされる側には人間と機械という本質的に異質なもの（＝非対称なもの）が含まれるので（チューリングはテストの対象となる機械を「デジタルコンピュータ」と明言しています)、テストする側は両者を混同しつつ分離するしかない状況にあります。このような状況は、もはや「テスト」ではなく「経験」としか呼びようがないものです。機械が知能を持つかどうかは結局のところ、テストによって理解される問題ではなく経験的に理解される問題です。しかし、経験的に理解するとは経験する者が勝手に決めることではありません。むしろ逆です。テスト

では、テストする者がなんだかのモデルや判断基準にしたがって判断します。テストする者＝観測者が勝手に決めています（＝外部観測）。一方、経験とは、おのずから／自ら立ち上がるものです。ある種の抗えなさは、経験の内部にいる者（＝内部観測者）においてはじめて生じるものです。

「知能の理解」＝経験であり、「知能」＝経験です。ここでは、「経験」という言葉をより厳密に使用ししようとしています。経験は我々にとって身近なものですが、それを理解することはたやすいことではありません。経験とは何か？　その答えは一言で言い当てるようなものではなく、我々の議論の全体を通して少しずつ理解されるものです。あるいは内部観測をめぐる議論の全体を通して明らかにしようとしているものです。とりあえずその特徴のいくつかを述べます。まず、経験には終わりがない（経験の終わりは「死」を意味しますが、それは正に異質なものとの接続です）。経験に終わりがないということは以下の２つのことを内包します。

１．経験は行為につながる。
２．経験に基づく判断は改められうる。

「１」と「２」は不可分ですが、判断と行為は異質なものです。そのような意味でも経験とは異質なものの混同と分離が共立する過程です。そして、経験のこのような側面を強調するような仕組みを積極的に導入することが「経験＝理解」へとつながる道です。本当のチューリング・テストは壁の向こうにアナログな人間とデジタルなコンピュータという定義上異質なものを併置することにより、テストにおける非対称性を強調してテストを終わりのないものにしています。それは正に「テスト」を発展的に解体して「経験」へと導くもので、その手法は「脱構築」とよばれものにぴったりと合致します。脱構築とは「形式」の向こう側にある「経験」を理解するための迂遠な方法です。経験（の理解）こそが難しい。それは、ただ経験すれば言い訳でありません。経験を理解するためには異質なものの混同／分離を積極的に導入すること、それが本節の結論です。次節ではそれを「経験的に」試みた私の研究を紹介しようと思います。

５．バンディッド問題から鉄棒ロボットへ

人工知能の古典的なテスト課題としてｎ本腕バンディッド問題というものがあります。バンディッドとはスロットマシンの腕のことで、複数のスロットマシンが並んでいる状況を考えます。ここでは簡単化のためスロットマシンは２つ（ＡとＢ）だけとします（図６）。また、どちらも当たりが出ると報酬が１もらえ、はずれの場合の報酬は０とします。ただし、当たりが出る確率はそれぞれで異なり、エージェント（スロットマシンをやる人）はその確率を知らないとします。状況を整理すると以下のとおりです。

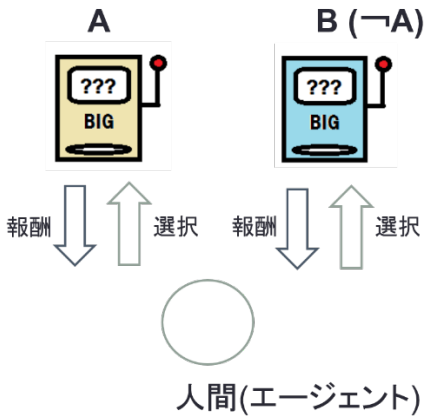

図6. バンディット問題

- 当たり（報酬：１）or はずれ（報酬：０）
- それぞれ異なる確率で当たりが出る
- エージェントはあらかじめ当たりが出る確率を知らない

このとき、エージェントはどちらのスロットマシンが当たりが出やすいかを調べつつ、なるべく多くの報酬を得るためにはどうすればよいかを考えます。例えば、AとBをそれぞれ何回か引いてみて、そのときに当たりが出た回数とはずれが出た回数を表１のように記録します。そして、AとBのどちらが当たりが出やすいかを評価するために、Aを引いたときに当たりが出た確率とBを引いたときに当たりが出た確率、いわゆる条件付き確率をそれぞれ次式のように計算します。

表1. 共起頻度

報酬	A	B
1	a	c
0	b	d

$$P(1|A) = \frac{a}{a+b} \ , \ P(1|B) = \frac{c}{c+d}$$

この両者を比較して大きい方がより当たりが出やすいと判断し、以後はそちらのみを選択し続けるようにします。このような方策はもっとらしいのですが、悩ましい問題があります。それは、AとBをそれぞれ何回ぐらい引いてから判断すればいいかということです。判断の正確さを求めるのであれば、両方とも多くの回数を引いてから判断すればいいことになります。しかし、それでは当たりが出る確率が低い方も多くの回数引くことになり、無駄が多

くなります。したがって、なるべく早く判断した方が良いわけですが、早く判断すると判断が間違っている可能性が高まります。例えば何回か両方を引いた後にAの方がよいと判断してAを引き続けたとします。しかし、Bは最初の何回かでたまたま当たりが出なかっただけであって、本当はAよりも当たりが出る確率が高いかもしれません。つまり現在の知識を利用した判断を優先するか、それとも判断を保留してより多くの知識を得ることを優先するかという悩ましい状況です。このような状況は「探索と知識利用のトレードオフ」と呼ばれています。探索と知識利用のトレードオフはバンディッド問題に限られたものではなく現実の世界においてもよくある様相で、たとえば動物がえさ場を選択する場合などが当てはまると考えられます。

ここで、バンディッド問題（図６）と本当のチューリング・テスト（図５）を比べると、両者は似ていることに気が付きます。まず、２つの対象を並行して調べる点が同じです。バンディッド問題における２つのスロットマシンが、本当のチューリング・テストにおける機械と人間に対応します。本当のチューリング・テストにおいては機械と人間が異質なものであることが重要でしたが、バンディッド問題における２つのスロットマシンは互い異質なものとは言えないかもしれません。しかし、２つのスロットマシンをAと￢A（Aでないもの）と解釈するのであれば、両者は異質でレベルの違うものと考えられます。スロットマシンが２台ではなく多数ある状況を考えるとAと￢Aの違いはより際立ちます。さらに現実の世界ではスロットマシンは無限にあると考えられるので、拡大して解釈すると￢Aはフレーム問題をも担うものと考えることができます。もう一つ重要なこととして、本当のチューリング・テストにおいては２つのテスト対象が相互作用します。機械と人間は互い相手のふるまいを見て真似をしたり違いを際立たせたりすることができます。一方、バンディッド問題では２つのスロットマシンの当たりが出る確率は独立に設定されています。したがって両者に相互作用はありませんが、現実的な状況では両者に関連がある場合が普通です。例えば、えさ場Aにえさがない場合にえさが￢Aに移動したと考えることは自然なことです。実際に人間はそのように考えることが知られていて、そのような傾向性は対称性バイアスや相互排他性バイアスと呼ばれています（より一般的には認知バイアスといいます）。

篠原修二さんは、人間の認知や判断に伴う対称性バイアスと相互排他性バイアスを数理モデルとして定式化しています（篠原 2007）。そのモデルは LS モデルと呼ばれています（LS モデルという呼称は高橋達二さんによるものです）。表１のように行動の選択とその結果が与えられているとき、「Aならば当たり（報酬：１）」に対する信頼度を LS モデルでは以下のような式で計算します。

$$LS(1|A) = \frac{a + b \cdot d/(b+d)}{a + b \cdot d/(b+d) + b + a \cdot c/(a+c)}$$

先に述べた条件付き確率$P(1|A)$の計算式と比較すると、c やd が含まれていることが特徴です。このことはAを評価する際にB（¬A）の結果も考慮していることを意味します。LSモデルは実際の人間の傾向性とよく一致することが実験によって明らかになっています。また、LSモデルにしたがって判断を行うエージェントは、バンディッド問題において優秀な成績をおさめることが確認されています。

バンディッド問題はどちらの当たりが出る確率が高いかを判断するだけの問題ではないところが重要でした。より多くの報酬を得るためにはなるべく早く判断して、判断の結果を行動に反映させる必要があります。一方で行動の結果によっては判断を改める必要もあります。そのような意味でバンディッド問題では判断と行動が不可分になっています。LSモデルはAと¬Aを混同することによって、素早く判断することとその判断を改めることを両立させています。その結果、LSモデルは探索と知識利用のトレードオフを解消することが可能となるのです。以上をまとめると、バンディッド問題とそれを解く LS モデルは「判断が行為につながる」、「判断は改めるられる」、「それらが異質なもの混同と分離によって実現されている」という様相を含みます。そして、そのような様相は我々が「経験」と呼ぶものです。このような意味でバンディッド問題とそれを解くLSモデルは本当のチューリング・テストの簡易版になっていて、それはシミュレーションによって実証可能なかたちで「経験」を理解するための方法となっています。

シミュレーションではなく現実の世界ではどうか？　私が高橋達二さんと共同でおこなっている鉄棒ロボットの研究は、バンディッド問題とLSモデルを実環境で動くロボットに拡張したものです（Uragami 2014b, 2016）。我々の議論のクライマックスとして、これからその研究について紹介します。

バンディッド問題では一回の行動選択に対して報酬がもらえるかどうかかが決まりましたが、より現実に即した状況では、いくつかの行動を連続しておこなった後にその一連の行動が良いか悪いかがわかる場合があります。例えば迷路の探索では、ゴールについてはじめて以前の選択が正しかったことがわかります。他には人間やロボットが身体動作もそうです。例えば「ボールを投げる」という動作も、足を踏み出すことや腰をひねるといった予備動作によってより速くボールを投げることが可能となります。人工知能の一領域である強化学習は、このような一連の行動選択を学習する方法です。強化学習の手法の一つに Q 学習というものがあります。Q学習のアルゴリズムは簡単なものですが、先ほど述べたような一連の行動選択を学習することができます。Q 学習はロボットの行動学習などの複雑な課題に適用することが理論的にはできます。しかし、現実のロボットに適用すると、学習回数が膨大に必要であることや環境の確実性の影響などが問題となります。

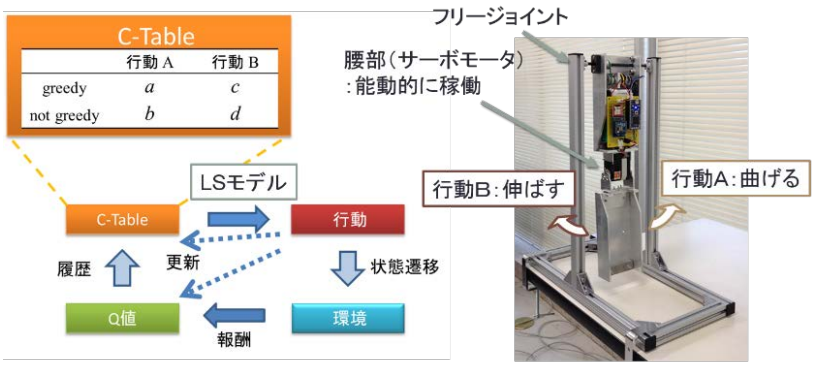

図7. LS-Qの学習アルゴリズム(左)と鉄棒ロボット(右)

図8. Q学習とLS-Qの学習結果の比較

そのような問題を解消するものとして、我々は LS-Q という学習手法を提案しています。LS-Q は Q 学習に LS モデルを応用したものです（図 7 左)。そして、LS-Q の学習能力を評価するための学習対象として鉄棒ロボットを製作しました（図 7 右)。鉄棒ロボットは鉄棒とロボットの結合部はフリーになっていて、腰部の関節のみが能動的に稼働することが特徴です。腰部をタイミング良く曲げ伸ばしすることよって、ロボットをなるべく高く振り上げることや回転することが課題です。報酬はロボットの先端の位置が高くなるほど大きくなるように設定しました。このとき、ロボットの腰部を曲げた状態から伸ばすと一時的に報酬が小さくなります。しかし、より報酬が大きい状態に移行するためには、曲げ伸ばしを繰り返して反動をつける必要があります。いわゆる「損して得とれ」という状況で、鉄棒ロボットは強化学習が解決すべき課題の典型的な例となっています。先ほど述べたように、Q 学習はこのような状況に理論的には適応可能です。しかし、現実的には状態認識の不完全さなどの理由により学習は成功しません。Q学習によって学習された行動は、ロボットの先

端を振り上げた状態で停止してしまいます（図８上）。一方、LS-Q によって学習された行動では、曲げ伸ばしを繰り返して回転運動を実現しています（図８下）。

現実のロボットではなんだかのかたちで記号接地問題が存在します。この鉄棒ロボットでは、状態分割を粗くすることによってそれをあえて強調した学習環境になっています。それが原因となって通常のQ学習では行動が停止してしまいます（判断を改められない）。一方、LS-Q では素早く判断を改めて別の行動をおこないます（フレームの外へ出る）。LS-Q のこのような能力は、ある行動Aと他の行動¬Aを混同する LS モデルの特性に由来するものです。LS-Q による鉄棒ロボットの学習実験は、異質なものの混同による判断の継続という意味において、「経験」を実験していると言えるかもしれません。

＊＊＊

最後に、我々の議論の全体をまとめておきたいと思います。我々の議論は、「定義」「テスト」「経験」という３つのキーワードをめぐって進められてきました（図９）。定義は哲学の、テストは科学の、経験は芸術の領域であると考えられます。普通のチューリング・テストは定義とテストの間に関係するような議論です。一方、本当のチューリング・テストはテストと経験の間に関係するような議論であると言えるでしょう。鉄棒ロボットの研究もテストと経験の間に関係するようなことを実験によって試みたものです。そのねらいが十分に果たされたかどうかはわかりませんが、そのような試みを徹底しておこなうことが知能とは何かといったことを理解する／経験することにつながっている、というのが我々の議論の結論です。それは哲学、科学、芸術の領域を横断するような試みです。

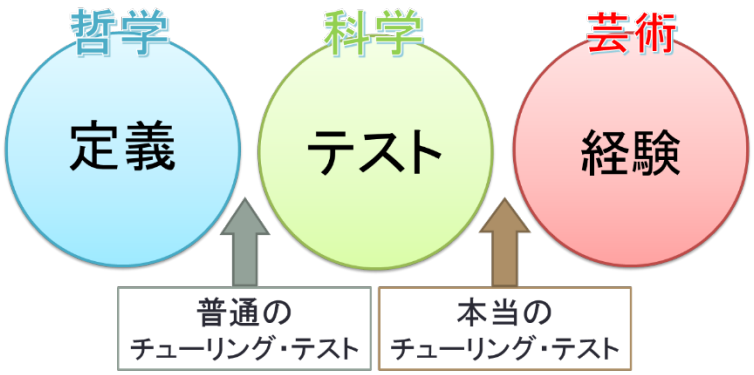

図9．全体のまとめ

参考文献

- ウィトゲンシュタイン（1953),『哲学探究』, 藤本隆志（翻訳）, ウィトゲンシュタイン全集 8, 1976.
- アラン・チューリング（1950),“計算機械と知性”, , 現代思想 １１月号臨時増刊号 総特集「チューリング」, pp.8-38, 2012.
- ジョン・サール（1983),“心・脳・プログラム”,『マインズ・アイ［下]』第 22 章, p.178-222, 阪急コミュニケーションズ, 1992.
- 服部裕幸（2003a),“AI 研究および認知科学に対するジョン・サールの批判”, 人文・社会科学編 『アカデミア』 第 77 号, pp.81-103, 南山大学, 2003.
- レベック, H. J.（2009）,“Is It enough to get the behavior right?,” Proc. of IJCAI-09, Pasadena, CA, 2009.
- 中島秀行（2011),“中国語の部屋再考”, 人工知能学会誌 26 巻 1 号 特集「チューリングテストを再び考える」, pp.45-49, 2011.
- ラメルハート（1986),『PDP モデル―認知科学とニューロン回路網の探索』, 甘利俊一(翻訳) 産業図書, 1989.
- 服部裕幸（2003b),“「分散表象」は認知の説明にはたしてやくにたつのか？”,『こころの科学と哲学…コネクショニズムの可能性』, pp.29-51, 昭和堂 2003.
- ホーランド（1991),“Other bodies, other minds: A machine incarnation of an old philosohical problem, ” Minds and Machines, Vol. 1, No. 1, pp. 43-54, 1991.
- 石黒浩（2011),“アンドロイドによるトータルチューリングテストの可能性”, 人工知能学会誌 26 巻 1 号 特集「チューリングテストを再び考える」, pp.50-54, 2011.
- ブルックス（1986),『ブルックスの知能ロボット論―なぜ MIT のロボットは前進し続けるのか?』 五味隆志(翻訳), オーム社, 2006.
- Uragami, D., Ohta, H., “Multilayered neural network with structural lateral inhibition for incremental learning and conceptualization,” Biosystems, Volume 118, Pages 8-16, 2014. (2014a)
- 浦上大輔, 郡司ペギオ幸夫（2009),“地理把握における動的双対性－不定性を伴う内包・外延対とアフォーダンス－”, 生態心理学研究 Vol.3(1), pp.45-56, 2009.
- 柄谷行人（1992),『探究 I 』, 講談社学術文庫, 1992.
- 郡司ペギオ幸夫（2012),“チューリングのバイオロジカルな拡張”, 現代思想 11 月号臨時増刊号 総特集「チューリング」, pp.128-149, 2012.
- 篠原修二, 田口亮, 桂田浩一, 新田恒雄（2007）, 因果性に基づく信念形成モデルと N 本腕バンディット問題へ応用, 人工知能学会論文誌 22 巻 1 号 G, pp.58-68, 2007.
- Uragami, D., Takahashi, T., Matsuo, Y., “Cognitively inspired reinforcement learning architecture and its application to giant-swing motion control,” BioSystems, 116, 1-9, 2014. (2014b)

- Uragami, D., Kohno, Y., Takahashi, T., Matsuo, Y., "Robotic Action Acquisition with Cognitive Biases in Coarse-grained State Space," BioSystems 145, pp. 41-52, 2016.