特集

「人間らしさ」と人工知能

橋本 文彦　大阪市立大学 大学院経済学研究科

人間らしさの定義

本論では、私自身が行ってきた、いくつかの被験者実験の結果を参照しながら、「人間らしさ」とは何かを問いかけ、翻って現在の人工知能あるいはロボット開発の研究に何らかの新たな視点を提供することを目指すものである。

さて、「人間らしさ」とはどのようなことを指すのか。また、人間が「心を持つ」とはどのようなことか。

私はこれまでに次のような定義を行ってきた。

「私が「心」と呼ぶとき、この語で意味するところのものは、私が「私は心を持っている」と発話する際に、この発話が直接に意味するところのものであり、この発話の聞き手が、これほどまでに曖昧な言明に対して、その意味を直ちに理解することが可能であることこそが、同時に聞き手にとってもまた「心」という語が意味するところのモノが保持されていると言うことの決定的な証左であると考える。私にとって他者である聞き手が私の発話を私が意図したとおりの意味で理解するか否かは必ずしも観察可能では無いと思われるが、それでもなお、聞き手が「心をもっているのであれば」、私がこの発話で意図していた意味とさほど相違ない意味で理解されていると考える。」[(1)]

この「定義」はしかし、やや哲学的にすぎるきらいがあるかもしれない。

むしろ、「人間らしさ」を直接に定義するのでなく、「機械（＝人工知能・ロボット）はどのようになれば人間らしいと言えるのか？」と問い直し、人工知能研究の側からのアプローチを確認することで、人間らしさをより的確に捉えることが可能になるかもしれない。

一般に「人工知能」と呼ばれるものが目指す到達点としては次の二つの方向性がある。それは、

(A) 人間と同様の振る舞いが出来ること
(B) 人間よりも高度な専門的処理が出来ること

の二つである。

一見すると、(A) よりも (B) の人工知能の方が実現が難しそうであるが、実際には逆で、家電製品やエキスパートシステムなどの例からもわかる通り、すべての能力において人間を凌駕するわけではないが、部分的な能力についてはすでに実現している。高度に専門的で誤りの無い動作をすることは、現実の人間にとっては困難であるものの、コンピュータにとっては、むしろ (A) タイプの人工知能に比べて実現が容易であるために、著しく進歩した分野である。後述するように、少なくとも現時点では計算を時折間違えるコンピュータよりも、高速に演算を行うコンピュータを制作する方が簡単である。

(B) タイプの人工知能は、通常のコンピュータの演算能力や、材質の頑健性、if ～ then形式の論理の単純さなどだけが満たされていれば、心や意識の問題に踏み込む必要がないと考えられるために、「人間らしさ」を問おうとする観点から本論ではこれ以降、(B) タイプの人工知能は扱わず、(A) のタイプの人工知能のみを議論の対象としたい。

(A) タイプについてはさらに

(α) 姿かたちを人間に似せること
(β) 人間との関係や機能を人間に似せること

という二つのアプローチが存在する。

古くは「からくり人形」、新しくは大阪大学の石黒浩の「アンドロイド」研究などが前者 (α) 型アプローチの代表例であり、「チューリングテスト」が目指したものは後者 (β) 型アプローチの例である。

よく知られているように、「チューリングテスト」[(2)] は、機械または人間が、別室の人間Xとコミュニケーションを行い、Xから見て相手を機械か人間かの区別がつかないならば、その相手は「知性を持っている」とみなそうというテストである。

チューリングはこの論文で、機械が「知性」をもつための十分条件を提供しようとしているが、チューリングテスト自体は、「人間と区別されるか否か」が問題となっているため、知性とは関係のない、あるいは知性との関係が薄い人間の特徴を再現しなくてはチューリングテストにパスできない。

例えば、相手の言葉に気分を害するといった感情的な対応やタイプミスは、チューリングテストの枠組みでは、機械と人間を見分けるための重要な特徴となるが、タイプミスをする（＝「人間らしい」ので、チューリングテストをパスする可能性がある）ことが「知性的」であり、タイプミスをしない（＝「人間らしくない」ので、チューリングテストをパスしない）ことが「知性的でない」という判断になることには違和感が残

る。

あるいは、難しすぎる問題を短時間で容易に解けるような機械は、その時点で人間ではないと判断が可能であるために、チューリングテストをパスできない。

つまり、チューリングテストは、一般的な意味での「知性」を定義しているのではなく、むしろ「人間的らしさ」や「人間的な心」を要請しているものであると解すべきである。

つまり、チューリングテストでは、「ミスをしないこと」ではなく、むしろ「ミスをする」ことに「人間らしさ」が示される、とも読むことが出来る。

しかし、本当に「人間らしさ」を定義するのに、「知性」とは別の「感情的になる」「ミスをする」などを判断材料とするだけで良いのだろうか？

この点について、次の２。および３。で私自身の被験者実験を参照しながら検討したい。

合理性と非合理性

さて、それでは人間はどのくらい「非合理な」側面を示すのだろうか？あらためて問うてみる。

私は、ある種の確率場面において、人間がその確率を把握しながら「一見」非合理的に見える意思決定を行うことを被験者実験で確認してきた。その実験の概要と結果の意味するところを検討したい。

実験は以下のようなものであった。

コンピュータ画面に［1］または［0］のカードが呈示され、被験者には次に表示されるカードを予測することが要請された。的中するとポイントを獲得できるので、できるだけポイントが高くなるように予測することが求められた。カードは実際には［１］と［０］とが3：7の比率でランダムに呈示された。人間の被験者（＝ヒューマン・エージェント）は、各自の戦略で［1］と［0］を予測し、コンピュータ・プログラム（＝コンピュータ・エージェント）は、予測と結果を学習するアルゴリズムによって自身の利得が最大になるように動作した。

図1および図2にヒューマン・エージェントとコンピュータ・エージェントの典型的な振る舞いを示す。

ほとんどすべてのヒューマン・エージェントは、30％の割合でカード「1」を選択（カード「0」を70％選択）した。

このことは、ヒューマン・エージェントがカードの「1」と「0」が出る確率を、実験時間内に自らの経験によって理解することができていたことを意味している。

これに対して、すべてのコンピュータ・エージェントは、カード「1」を0％（カード「0」を100％）選択した。

これだけを見ると、コンピュータ・エージェントは、ヒューマン・エージェントに比べて確率事象への感受性が低いよう

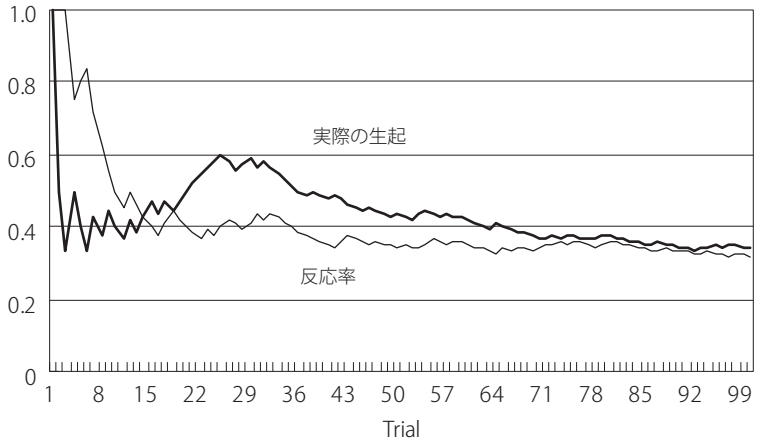

図１：典型的なヒューマン・エージェント

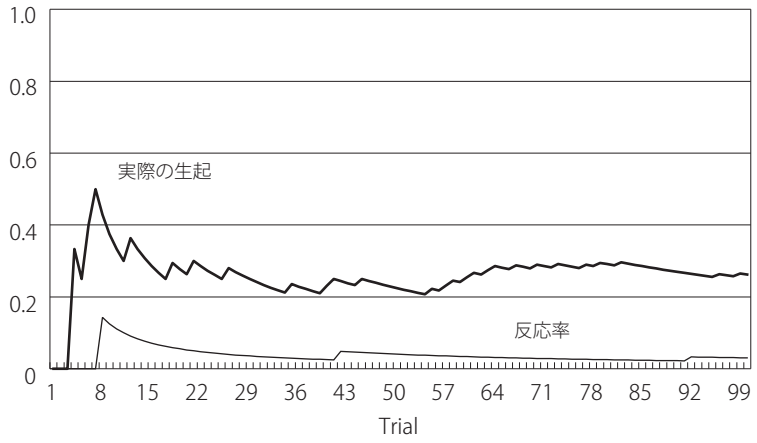

図２：典型的なコンピュータ・エージェント

に見える。

ところが、100試行によって獲得された点数（＝予測が実際のカードと一致した数）を比較すると、その差がはっきりとわかる。

すなわち、コンピュータ・エージェントの平均得点が、70ポイント程度であるのに比べて、ヒューマン・エージェントの平均得点は、57ポイント程度で、コンピュータ・エージェントの方が高い得点を獲得しているのである。

この理由は、以下のような計算によって直ちに明らかとなる。

エージェントたちに与えられた「1」と「0」のカードの確率は、30％対70％で、その順序は真にランダマイズ[3]されていた。

ヒューマン・エージェントは、「1」のカードが確率0.3であることを経験から理解し、それに対して0.3の割合で「1」カードを予測したが、これによって正答を得る確率は、実際には0.3×0.3＝0.09でしかない。同様に、「0」のカードが出る確率は0.7であるが、ヒューマン・エージェントが予測した「0」のカード0.7のうち、その予測が的中するのは0.7×0.7＝0.49である。したがって、ヒューマン・エージェントの期待得点は49＋9＝57ポイントということになる。

他方、コンピュータ・エージェントは、もっぱら「0」を選択したが、このうち実際にカード「0」が出る確率は70％であるために、コンピュータ・エージェントの期待得点は70ポイントとなり、ヒューマン・エージェントを上回るものとなった。

これによって、ヒューマン・エージェントは、事象の生起確率を正しくとらえているにもかかわらず、もっぱら一方の選択肢のみを選び続けたコンピュータ・エージェント[4]の方が平均的に高い得点を得ることとなった。

しかしながら、私はこのことをもって、「人間は合理的な最適行動をとれない」と結論づけようとは思わない。

被験者（＝ヒューマン・エージェント）は、提示された「1」と「0」のカードがランダムであると考えず、そこにに「何らかのルール」があると想定して、選択を行った。

もしも、実際にこのカード提示にルールがあるならば、被験者は選択を繰り返す中で、そのルールを発見し、すべての試行に正解を答えることができる時が来るかもしれない。

これに対して、今回の実験で用いられたコンピュータ・エージェントは、今回のように「ランダム」に提示されているカードに対しては、確かに最適行動であったが、「確率が0.5よりも高い方の選択肢のみをもっぱら選択する」という戦略では、このカード提示が何らかのルールに従っていた場合でも、いつまでもそのルールを見いだすことができない。

本実験は、「How Can Irrational Agents Survive?」というタイトルで発表されたものであるが、これは一見必ずしも合理的で最適行動をしていないように見えるエージェント（＝人間）が、それでも地球環境の中で生き延びてこられたのは、このような「ルール発見能力」とでも言える、「未来の最適化に向けた、現在の不合理性」をもっているからであると考えることができるのではないだろうか？

すなわち「人間らしさ」は、チューリングテストで見たような「わざとミス」「わざと非合理」のような振る舞いによって形成されているのではなく、あくまでも「世界でよりよく生き延びるために必要な振る舞い」から生じているのである。

ゲームと勝敗

人工知能は、最初はオセロゲーム、次にチェス、さらには将棋というゲームにおいて、人間の最強プレイヤーに勝利してきた。

このことは、「コンピュータが人間を凌駕した」ことを意味するのだろうか？

答えは、「ある意味においてはYes」であるが、本論で検討してきた「人間らしさ」という意味においてはNoである。なぜか。私自身が行った被験者実験の例を挙げて確認したい。

ある種の形や名前をもつ対象から、人はどのようなゲームを創り出すのだろうか？

ここでは、詳細な手続き等は省略[5]し、結果の概略を確認する。

この実験では、環境の中にすでに存在している「モノ」に対してではなく、ゲームのルール作りという場面を与えることで、人工的に作られた「モノ」に対して、複数の被験者がどのようにして「共通のルールを生み出す」のかを探った。

実験では二人一組の被験者が、7種類の（動物の形[6]または数字を描いた）駒と（7×7マスの）ボードを与えられ、各駒に許される動きと、相手の駒とぶつかった際の駒の強さという属性（本論ではこれらをあわせて「ルール」と呼ぶ）を定めることを求められた。被験者は二人で相談をしながら、それぞれの駒に動きと強さの属性を定めていった。また、被験者らは、ルール設定の後に、実際にそのルールに従ってゲームを行うことを求められた。

このようにして、各駒に与えられたルールは、動物の形の駒を用いた実験Aにおいて、複数の被験者組でかなりの程度共通したものとなっており、駒の形や名前がその動きや強さに対してある種の意味づけを与えたことで複数の被験者組が共通のルールを作成したと考えられた。一方、実験Bの数字駒のほうは、「形」に差異が存在しないために、被験者組での共通性はそれほど強いものとはならなかったが、被験者は「他の駒との差異」を作り出そうとしていた。

実験Aでは、駒として用いられた動物の種類に依存した属性づけがなされることが多かった。

ただし注意すべきは、これらの属性が必ずしもその動物に必然的に結び付けられているのではないという点である。

各被験者組の経過を撮影したビデオによると、いくつかの被験者組は当初「ウマ」の駒に将棋の桂馬やチェスのナイトに相当する属性（斜め移動とジャンプ）を与えようとしたが、このボードゲームには「トリ」がいることを思い出した被験者は、その属性を「トリ」に与え、「ウマ」には、別の本来の動物として想起される属性を与えることに変更した。これによって、「ウマ」は、複数駒を移動はできるが、斜め方向よりもむしろ真っ直ぐに（水平方向と垂直方向に）移動するように定められた。

このことは、被験者が各駒に対して与えている属性が、駒に必然的に直結するものというよりは、他の駒との差異を構

成しつつ、それぞれの駒から想起される属性と矛盾しないように調整された結果であるということを示している。

各駒の強さに対して、「推移律」が破られていることも特筆に価する。すなわち、推移律が成立していれば、XがYよりも強く、YがZよりも強い場合には当然XがZよりも強い、ということになるはずであるが、被験者らが設定したルールではそのような関係にはなっていない。

ある被験者組（Subj_No. 12）の自由記入用紙に描かれた図を、図3に示した。このように、被験者組の多くは、（推移律でなく）循環型で「駒の強さ」を設定している。これは、第一にゲームを面白くするためであり、第二に現実の世界も特定の動物のみがいつでも強いわけではなく、循環的な関係にあることを考慮していると考えられる。

一方、数字駒を用いた実験Bでは、実験Aでの動物のような多くの属性をもつ駒を用いていないために、それぞれの駒はもともと「数値の大小」という属性しかもっていない。

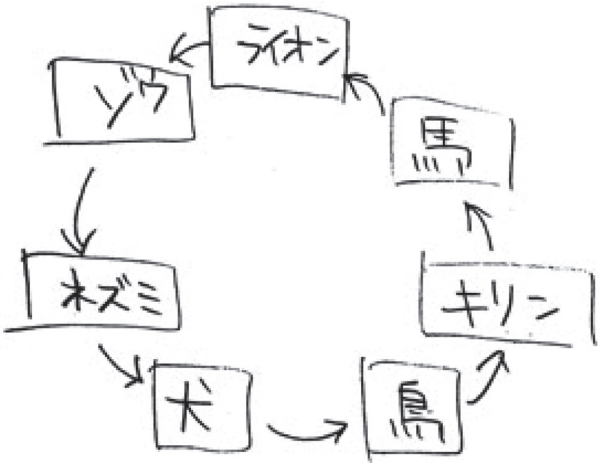

図3：被験者が描いた駒の強さイメージ

しかし、各被験者組は、「（一般に）大きな数字は小さな数字よりも強い」といった特徴のほかに、「偶数と奇数」や、（駒が七つあるので）「4を中心に対象の位置の駒は属性が似ている」などのあらたな制約条件を自らに課す傾向があった

これらの結果を合わせると、被験者たちが行ったのは、

1. 勝ち負けを決めるだけのための単純なゲームルールではなく、あえて複雑なルールを生成した。
2. コマに「見た目」の属性が備わっている場合、経験をもとに現実世界の属性をコマの属性に投企した。
3. 「見た目」の属性が存在しない場合（＝数字コマ）も、さまざまな観点からそれぞれの駒を差異化して、複雑なルールを作り上げた。

オセロ・チェスに続いて、将棋においてもコンピュータ・プログラムが人間の世界チャンピオンに勝利しつつあるが、人間の特徴は「ゲームに勝つ」ことではなく、「勝ち負けがつきにくいゲームを考案する」ことであることを、今回の実験は示唆していると私は考える。

コミュニケーションとシミュレーション

では、前章で見たような人間の「特徴」を再現することの出来るコンピュータ（＝人工知能）が登場すれば、人は直ちに「人間らしさ」を備えた人工知能が誕生した、と言えるのだろうか？さらに言えば、「人間と同様の心」を持っていると言えるのだろうか？

実はチューリングテストは、かなり早い時期に意外にも非常に簡単なコンピュータ・プログラム（＝イライザ[7]）がパスした[8]ことで、むしろこのテストが本当に「知性」の有無を判定できるのか否かが問題となった。

逆に言えば、イライザのような簡単なプログラムと対話した人たちは、そこに「人間らしさ」を感じたということである。

サールは「中国語の部屋」という思考実験によって、機械の「心」はあくまでもシミュレーションであって、実態ではないとした。

その問題の概略は以下の通りである。

ある部屋の外の窓口から、中国語で書かれたリクエストを入れると、適切な中国語で書かれた返答がなされる。しかしこの部屋の中の住人は、中国語を全く介さず、英語で書かれたマニュアルに従って、ある種の図形の並びに対して、定められた図形の並びを返しているだけである。部屋の外から見ると、（中国語のリクエストに対して、適切な中国語の返答が返ってくるので）、あたかもこの部屋の住人は中国語を理解しているかのように見えるかもしれないが、実際には彼は、英語のマニュアルにしたがって図形処理をしているだけなので、中国語を理解しているとは言えないだろう、というものである。

サールによれば、この「中国語の部屋」の状況は、コンピュータ・プログラムによる人工知能と同様である。

すなわち、コンピュータによる人工知能は、たとえ人間と区別のつかないコミュニケーションができて、そのコンピュータが会話を「理解」しているように見えたとしても、中国語の部屋の住人が差し出す返答と同様に、そのコンピュータの出力する会話は、その「内容」を実際に意味しているわけではなく、プログラムによって書かれた手順にしたがって、ある種の電気信号が入力されたら、それに対応した（とプログラムに書かれている）ある種の電気信号を出力しているだけである、というものである。

つまり、コンピュータは原理的に人間同士の会話と同様の「意味内容」をもつことができない、と彼は唱えた。

つまりサールによれば、「人間らしさ」は人間であることを直接には意味しない、ということになる。

もちろん、このサールの思考実験と主張については現在でも賛否両論があるが、主要な反論の一つは、人間の脳細胞のどの一部をとっても「心」を持っているわけではないが、それらが「全体として」心を持っているのは間違いない、というものである。

素材と時間

人間を対象とした研究では、余りに明らかなことであるが、人間に情報が与えられ、それが処理された後に反応へと至るまでには強い時間の制約がある。

生物学において、動物のサイズは、寿命や進化の速度に緊密に連関していることが知られている。

宇宙のどこか、サイズや時間刻みが地球上のそれとは全く異なる星で、金属やシリコンや岩石でできた生物（あるいはロボット）がもつ心は、地球上の人間と「コミュニケート」できる可能性があるだろうか？

私は「ない」と考える。

私は、本論のテーマについて、これまでに何度か検討を行ってきた。その最初の段階では、「心の形式化」は、将来において可能でありそれによって、機械は心をもつことができるであろうと考えた（1994）[9]。しかしながら、形式化可能＝人工知能が心をもつ、わけではなく、そこに「素材」依存の問題があるかもしれない点に言及したのが橋本（1996）[10]である。この後に私は、大阪市立大学の太城敬良らが行ってきた左右逆転眼鏡装着実験に参加し、また人工市場において、人とコンピュータが共存して取引をするという、U-Mart実験を行ってきた。

私が行ってきたこれら種々の実験の中で、人間がほぼ一定の大きさでほぼ一定の時間をかけて世界と関わりながら学習し、意思決定し、さらに行動していくことを見る[11]につけて、この「素材」に関する問題は、「素材が化学変化するための時間」の概念へと変化してきた。

つまり、あらゆる「素材」は、その素材（＝原因）に対応する心（＝結果）をもっている可能性がある。したがって、本論は「なぜ、人間だけが心をもつのか？」という問いに答えることはできない。人間以外のありとあらゆる物質は「心」をもつ可能性があると私は考えるからである。しかし他方で、そのような人間以外の物質がもつ「心」は、たとえそれがチューリングテストをパスしたとしても、その物質が本来的にもつ時間刻みの相違によって人間とは異なっていると考えるのである。

これらの物質からなる「他者」がチューリングテストを通過した際に私たちが感じた「違和感」は、私たち人間（の心・意識）にとって本来的である時間が、他の物質では本来的でなく、「遅延回路」等何らかの時間調整を行った「作り出された人間らしさ」であるために、そのコミュニケーションに「意味」や「意図」を感じることができないからなのである。

ここで改めて、より単純な例をあげると、人間の子どもは、一度に多数のこと[12]を教えても、それを一度にできるようになるわけではないし、「加速教育」できるわけでもない。環境の中で身体を成長させながら、けがをしたら治癒に時間を要し、数学の計算ができるためには何度か誤りながら正しい答えに行きついたりするのである。人間の場合には、その時間の刻みは一意的なものであり、加速したり、遅延したりすることは不可能なのである。

これに対して、チューリングテストに通過したこれまでのプログラムや、現在設計されているほとんどの知能ロボットは、この時間刻みを変更することが可能であるという点で、そこに「調整可能な範囲」があるために逆に、人間とは「心」を通い合わせることができないと私は考える[13]。

人間は、世界の物理的環境の中で、世界の情報を知覚し、それに対して身体と共に行動を行い、その行動の結果が脳を可塑的に変化させることで学習して、次の行動へと至っている。

この際に、人間を構成する物質の化学的性質による神経伝達の時間、脳が生理学的に変容するために要する時間、さらに人間が一定の大きさをもつことで物理的に定まる環境の時間的制約（例えば、ある物体が人間の背の高さから地上に落下するまでの時間は、人間と世界の全体のミニチュアや拡大版を作っても、再現されない）などによって、人間の成長速度（身体を含めた人間の状態遷移の速度）は、かなり狭い範囲でだけ相互の類似性が成立している[14]。

私は「機械は心をもちえない」とか「人間は心をもつものを人工的には作れない」と主張しているわけではない。変化・変容に要する時間刻みが人間を構成する素材と類似していて、それらの大きさが人間とよく似ているロボットが作られたなら、彼らと「違和感なく」コミュニケーションをすることは可能であろうし、人間によって「彼らは人間らしい心をもっている」とみなされることは可能であろう。この場合には、そのロボットが話す言葉は、単なる統語論的なものではなく、意味論としての領域に踏み込むことができるのだろう。

注

(1) 橋本（1994），p.533

(2) A. Turing（1950）

(3) 被験者は、「ランダマイズ」されていたことを知らない。

(4) コンピュータ・エージェントが、実際に1と0の生起確率をどのようにとらえていたのか、はもちろん不明であるが、1000試行の学習を行っているので、確率を計算することそのものは簡単であろう。

(5) 橋本（2008）

(6) 実験Aでは動物、実験Bでは数字の描かれた駒を用いた。

(7) J. Weizenbaumが1964年頃に開発した、精神科の医者の役割を演じるプログラムだが、基本的に相手の会話を繰り返すだけのものであった。

(8) 本来のチューリングテストでは、コミュニケーションのテーマは限定されていないが、イライザは、特定のテーマと場面設定を対象とするために、本来の意味でチューリングテストをパスしたというわけではないが。

(9) 橋本（1994），橋本（1995）

(10) 橋本（1996）

(11) 逆さ眼鏡実験では、環境への順応時間がきわめて重要な問題となり、また人工市場実験では、リアルタイム実験の場合と「加速」実験の場合とで（コンピュータ・エージェントの振る舞いに相違はなかったが）ヒューマン・エージェントの振る舞いには大きな差が観察された。

(12) 知識だけでなく、運動や動作も含めて。

(13) もっと踏み込むと、人間は自らの「死」に向かって不可逆的に一定の時間間隔で向かっていくが、現在の知能ロボットやプログラムはそうではない、という問題も出てくるが、「死」は本論が扱うにはいささか荷が重すぎるために、この観点については、この註にとどめる。

(14) つまり、物理環境自体が全く異なる場合（例えば重力の異なる別の惑星など）に生物がいたとしても、地球上の人類

は、この異星人を相手に「心」を感じてコミュニケーションをとることは困難であろう。

参考文献

Hashimoto, F. (2005). How can irrational agents survive? WEHIA.
橋本文彦（1994）．機械は心をもちうるのか．心理学評論，Vol. 37，533-544.
橋本文彦（1995）．哲学・AIにおける心身問題．サイコロジスト・レポート．
橋本文彦（1996）．脳はなぜ思考するのか．東北哲学会年報，Vol. 12，86-87.
橋本文彦（2012）．機械の身体と人間の身体，機械の心と人間の心．思索，Vol. 45，207-232.
橋本文彦・中塚寛史（2008）．ボードゲームのルール形成に見るアフォーダンスと他者との調整．人工知能学会・知識ベースシステム研究会，51-58.
石黒浩（2009）．ロボットとは何か―人の心を映す鏡―．講談社．
前野隆司（2010）．脳はなぜ「心」を作ったのか．ちくま書房．
Searle, J. (1984). *Minds, Brains, and Science*. Harvard Univercity Press.
柴田正良（2001）．ロボットの心―7つの哲学物語―．講談社．
Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, Vol. 59, 433-460.