【解 説】

機械は心を持てるか

和歌山大学
瀧　寛和　*taki@sys.wakayama-u.ac.jp*

明治大学
石川 幹人　*ishikawa@kisc.meiji.ac.jp*

岐阜大学
伊藤　昭　*ai@elf.info.gifu-u.ac.jp*

松下電器産業（株）
岡　夏樹　*oka@mrit.mei.co.jp*

この解説では，人の心や認知的な心をそのまま論じるのではなく，工学的に機械の心をどのように捉え，どのような仕組みとして実現すれば，有用な機械を構築できるかを考えてみる．

近年，簡単な判断・学習のできる知能ロボットや二足歩行の可能な人型ロボットが身近なものとなってきた．ホンダのASIMOに代表されるヒューマノイドには，まだ，高度な知能は搭載されていないが，心を持ったロボット・鉄腕アトムを連想させる．また，エンタテインメントロボットの代表であるAIBOは，基本的な欲求や感情のモデルを備え，発達に似せた段階的な変化が組み込まれており，ペットのように振る舞うことができる．人や動物に似たロボットが身近な存在となり，我々の社会の一部を構成し始めている．

さらに，多くの情報端末や情報処理装置が我々の身の周りに浸透してきている．情報家電や携帯電話がその代表的なものであるが，これらの情報機器は，広い意味で人の身体の一部として機能したり，人と人のコミュニケーションにも介在して，人が機械と対話しているような錯覚さえ生んだりしている．社会のあらゆるものが情報化し，ネットワークで結合されるようになり，人は情報化された社会基盤の中で暮らすようになってきた．このような情報機器に取り囲まれている社会で生きていくには，人が機械に慣れるだけでなく，機械をより人に近づける必要性が生まれてくる．そこで，人と機械の関係が見直される時代が到来したといっても過言ではない．

鉄腕アトムのような心を持った，より人に近い機械を実現していくには，どのような技術が必要になってくるのかを，その歴史と技術動向からみていきたい．

―機械の知能―

機械の心について，少し過去に遡って，計算機分野での議論を追ってみることにする．「機械が心を持つか」の議論は，最初は「機械が知能を持つか」というかたちの議論として現れた．

当初，知能とは何かの定義が問題となったが，これは，ダートマス会議（1956年）においてミンスキーらの著名な人工知能研究者が提示した，「記号処理による高度な計算」の定義で一応の方針が示された．

機械が知能を持つかどうかの議論では，「チューリングテスト」と「中国語の部屋」の問題が有名である．チューリングテストでは，機械と文字端末で通信し，人の出す質問の応答が，機械であるか人であるかの区別ができなければ，その機械は知的である（知能を持つ）と判定する．同様に，「中国語の部屋」では，漢字の札で質問し，その返事（漢字）を見て，中国語が理解されているかどうかを判定するものである．この議論では，この部屋に，漢字が読めない外国人が漢字の対応表を見て，意味が分からなくてもその表に従って，妥当な漢字札を選ぶことができる場合と，実際に中国人がいる場合との違いが判定できないことが示された．つまり，表面的な観測では，機械が知能を持っているかを判別できないのではないか，記号計算だけでは真の理解が実現できないのではないかなどが議論された．

とはいえ現在は，機械が記号計算という，ある種の知能を持っていることは広く認知されているようである．エキスパートシステムやワープロのAI辞書，インターネットのホームページを訳す機械翻訳などに，人工知能の技術が利用されていることは，事実である．では，なぜ，機械の知能が認知されるに至るのであろうか．これは，内部の複雑な仕組みを公開していることで，機能的に知能が実現されている事実を認識できるからといえよう．

一見，問題の少なそうな「知能」という概念についても，このような輻輳した議論がある．ましてや「心」に至っては，一歩間違えると底なしの議論となってしまう．

―機械に心を感じる―

人は，明らかに心を持っていない機械に対しても，心を感じることがある．このことを実際のプログラムとして具体的に示したのは，ワイゼンバウムである．彼は，上記のチューリングテストのように，英文で対話するプログラムELIZAを開発した．このELIZAは，セラピストのように振る舞い，人の入力文の一部を書き換えて，ほとんど機械的に単純な応答をするものである．このプログラムの対話に関心を示し，機械の応答文をセラピストのカウンセリングとして判断する人

図-1　対話型高齢者支援ロボット「タマ・クマ」

も少なくなかった．ELIZAは，心はもとより，知能も持たない機械を，知的な存在と感じてしまう問題点を提示した．

一方で，機械の心とは何かが明確にならずとも，人が機械に心の存在を感じる仕組みを研究し，その機能を役立てることができる．寂しさを紛らわす効果や，癒しの効果として，すでに利用されている．

エンタテインメントロボットの役割：エンタテインメントロボットは，買主（飼主）とのインタラクションで買主の覚えさせたい内容を学習する．そのことから，買主は，自分のペットとしてその個性（学習内容）を認めることで，より親しみを感じる．買主は，ロボットであるペットに感情移入をしやすくなる．エンタテインメントロボットは，限られた学習内容と組込みの知能により，買主の働きかけに反応するが，その反応が正しくなくても，それを受け取る側が人であるから，勝手な解釈を行い，ペットに心を感じてしまう．正常に反応すれば，賢いと思い，正しくない反応であれば，機嫌が悪いと判断してしまう．このペットロボットの代表がロボットらしい姿をしたSONYのAIBOである．猫に風貌が近いオムロンのネコロは，猫のぬいぐるみを着たような姿で本物の猫と間違うほどである．ペット分野では，心を持った機械の前身として，より心の存在を感じさせる機能の開発が進みそうである．

福祉的なロボットの役割：従来，福祉関係のロボットといえば，障害者の機能の補助（視覚，聴覚，肢体機能）のセンサや運搬機能を持ったロボットが主流であった．しかし，これらのロボットは，杖の代わりや補聴機の代わり，手や足の代わりであって，心を持つ機械の対象とはならなかった．松下電器のタマ（**図-1**）（現在は，池田市実証実験対話支援ロボット「ワンダー」，松下老人ホーム対話支援ロボット「こうちゃん」に進化）は，高齢者の家に置かれ，役所とインターネットでつながれている．役所からの福祉情報の端末としての機能もあるが，音声認識の機能も備え，老人の話し相手としての役割も果たす．老人とタマの会話は，正しく反応した会話となることもあるが，誤った反応も多い．しかし，老人はよき話し相手として，タマを認識している．タマは，本当の心は持たないが，親しみを感じ感情移入ができる存在となっている．

ー心を持った機械の社会への受け入れー

心を持った機械が人間社会の中に出現した場合について，考えておかなくてはならない．現在は，人（および動物の一部）のみが心を持つと信じられているが，機械が心を持てば，人の存在の優位性が揺らぐかもしれない．人は機械の存在に脅威を感じることになるだろう．

欧米人は人と機械の間にある溝を大きく意識している．映画「ターミネータ2」のターミネータや映画「2001年宇宙の旅」のHAL9000には，人のような心はないが，高い知能を持ち人に危害を加える．しかし，映画「ターミネータ2」では，最後には，人同士のように，人とターミネータがインタラクションできるようになり，「人がなぜ泣くのかが分かるような気がする」までに成長する．映画「アンドリューNDR114」では，アンドロイドの行動がプログラムされたものであるのか，それとも自発的行動であるのかが問題の1つとして提起されている．学習により得た行動や自発的な行動のできるアンドリューが，人として認められるまでの物語である．物語では，人として認められる条件に，命の有限性が必要であった点は，面白い．人権のように心を持つロボットのロボット権についての議論は必要であろう．

映画「A.I」（原作：スーパートイズ）では，知能や心を持ったロボットであっても人間社会に受け入れられない問題を描いている．「人でないものは，決して本物（本当の人）になれないが，それでよいのか」という疑問を投げかけている．これもロボット権を考えさせる作品である．

日本では，どうであろうか．鉄腕アトムの話には，ロボットが人に差別され，悩む話が多く扱われている．アトムはロボットであるが，日本人は，アトムの勇気・優しさ・思いやりに疑問を持っていない．おそらく，産業用ロボットがいち早く工場に普及したように，機械が人らしい心を持っても日本の社会では受け入れられやすいと考えられる．

「機械の心」の論点

ここで，少し論点を整理してみよう．歴史を振り返れば，心は，哲学的，心理学的，医学生理学的，物理学的な観点から，多角的に説明されてきた．その過程

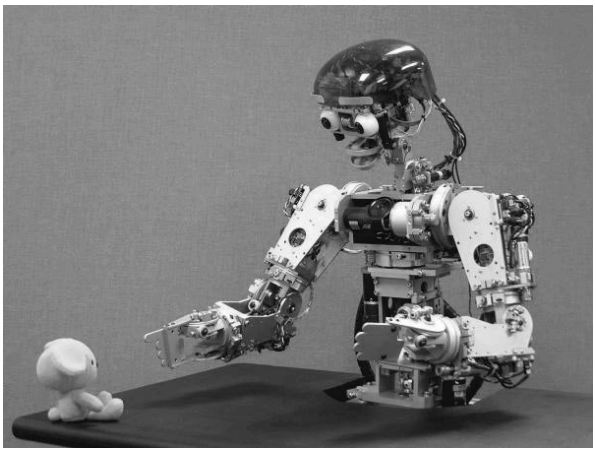

図-2　心の発達の研究を目指すインファノイド

で心は，さまざまに異なるものとして扱われてきた．行動としての心(行動主義)，脳としての心(心脳同一説)，コンピュータとしての心（機能主義）などである．機械の心を考えるときは，その実現対象となる心をどのように捉えるかが重要である．我々が実現したい心とは何であるのか，もう一度検討してみる必要がある．我々の理解の対象となる心は，言語学的な「人称」という概念を持ち込んで，次の3つに大別できる．

ー心の理解の観点ー

1人称的理解：「自ら自覚する心」を目指す

内観という方法で主観的体験を記述し，その構造を明らかにするのが目標である．現象学哲学や，かつての意識心理学は，この観点からの心の理解を目指していた．しかし，内的世界を独自に追求するあまり，物的世界は存在しない（観念論）とか，存在しても心の世界とは無関係である（二元論）とみなされがちとなる．1人称的心が機械上に実現できるとする前提には，物的世界は心的世界と一体である世界観（拡張された物理主義や中立的一元論など）が必要であろう[1)]．

機械の心の実現対象として，この1人称的心を取り上げるのは難点がある．自覚する「私の心」が機械上に実現される様子を想像するのは難しい．ただ，「私」が機械に接続されることによって「私の心」が変容する体験を通して，1人称的心の構造や機能を明らかにしていく道はあるだろう．それにしても内省的分析は個人差が大きいとか，研究を積み重ねるには，倫理的問題が大きすぎるといった問題が残される．

2人称的理解：「相手が持つ心」を目指す

対話を通して得られた相手の反応や，参与的に観察される相手の行為をもとに，共感的に相手の心を理解するのが目標である．こうした心の理解は，社会的状況におかれた行動主体の研究やコミュニケーションの研究と接続している．

2人称的心は，機械による実現対象として有望である．我々が他者に心を感じるのと同様に，機械に心を感じるようになればよい．同時にそれは，機械が人間社会の一員として適切に振る舞うことにも貢献する．社会的行動を実現するには，さらに他者の心を想定し，その行動を予測して，協調行動を行う必要もある．

現在，機械がこうした心を形成するのに，人の心の発達過程が参考になるという仮説のもとに，鋭意研究が進められている．幼児と親の相互作用（インタラクション）の模倣（シミュレーション）について，数々の研究がある．MITのCOGプロジェクトや通信総研のインファノイド（図-2）などの研究では，人の頭部（さらには腕も実装された）ロボットにより，人の視線を追跡させたり，人の表情を模倣させたりすることに成功している．さらには，他者の興味に関心を示したり，他者の注目している対象からその意図を読み取ったりする機能も研究されている．

3人称的理解：「他人の心」を目指す

外在的な位置から距離をおいて観察することで，他人が持つ心を理解するのが目的である．これは客観的な法則として，心の存在を理解することでもある．そのためには，心が存在するための必要十分な条件を，還元主義的な観点から，解剖学的，生理学的に追求する方法論が有力視される．

現時点では，3人称的心の機械上での実現は難しい．その実現される心とは何であるかが明確化されていないからである．富士山を登る機械を実現するのは純粋に工学的問題であるが，3人称的心を持つ機械の実現にはそれ以前の「心の定義」問題がある．現在は，3人称的心を支える知見を，脳の部位と機能の関係から探る段階であるともいえよう．古くは，脳に障害のある患者の障害部分と患者の行動との関係を，医学的・心理学的に分析していたが，最近では，各種の脳の部位の活動を測定する機器（MRI，脳磁計，PET）を利用して進められている．1人称的に自覚する心的状態（音を聞くとき，物をつかもうとするとき，悲しいとき，うれしいとき，計算しているときなど）と，脳の活動部位の関係を調べることで，脳地図や脳の連合的活動がかなり明らかにされつつある．これらの成果は，将来，3人称的心を客観的に定義する手がかりになるだろう．

また，心を生むメカニズムを脳の要素に求めずに，脳神経（神経細胞とシナップス）や神経に栄養を送るグリア細胞の構成物からなる神経回路ネットワーク（ニューラルネットワーク）として，脳を，モデル化することも進められている．ニューラルネットワークは，人工知能の研究として，パーセプトロンから始まり，パターン認識や学習の研究にも利用され，家電製品のセン

サ処理や制御にも応用されている．ここから2人称的な機械の心が生まれてくれば（まだ生まれていないが），それを根拠に，3人称的心を客観的に示すモデルも現実のものになろう．

以上から，工学的に機械上に実現される心は，当面，2人称的観点で捉えていくのが妥当であると結論づけられる．医学・生理学が，1人称的に自覚する心と，それに対応する脳現象との相関を明らかにし，工学が，社会的に実現される2人称的心を支える機械モデルを構築することで，ゆくゆくは3人称的な心の定義が確立される見通しが期待できる．

ー心を持つ機械の工学的実現法ー

工学的に機械の心を実現するには，2人称的心の機能的側面を考える必要がある．ここでは，その側面として，「感情」「自発性」「意識」の各点を取り上げてみよう．

●感情：感情表現の研究

感情を表現する顔，動作は広く研究されている．最近では，機械への働きかけ（言葉など）と感情の関係モデル（叱られると悲しさの度合い上昇）の研究へも展開している．しかし，表面的な反応が中心で，人のように複雑な感情表現までは行き着いていない．社会的なコミュニケーションにも適用できる深いレベルの感情行動の研究が望まれる．

顔の表情の研究では，CGや顔の要素（眉毛，目，口など）の変形で表情を表す研究がある，また，感情のモデルを考える研究では，猫型ロボット「ネコロ」では，「感情生成機構」を持ち，猫の感情を表現できるようになっている．

感情表現のできる機械とのインタラクションがなされれば，人はその機械の感情表現の裏に，心を感じとれる可能性が出てくる．

●自発性：自主性や自律性の研究

行動目標を自ら決定する機能を持つロボットやエージェントの研究は，2人称的心の実現につながる．しかし，既存の研究では，ロボットの行動に明確な目標が与えられている場合が多い．目標の副目標は自分で選ぶ，目標の集合から選ぶ，特定の刺激で目標が選ばれるなどである（たとえば，電池の蓄電量が下がると，自ら充電する）．目標を自発的に生成するメカニズムを深く研究することが望まれる．これには，機械が置かれている環境の状況を認識して，その環境の問題と解決策を総合的に判断して，目標を設定したり，現在実行中の目標の上位目標を推定して，現在の目標を他の目標に入れ替えたりするなど，目標そのものを決定する，メタな問題解決の仕組みの研究が望まれる[2)]．

こうした研究が進めば，いちいち命令したり，指示を与えたりしなくとも，必要とされる仕事や問題解決を自ら行ってくれる機械の実現につながる．しかし，人の望む以外の行動も行う可能性を否定できない．

●意識：行動を意識して遂行する機械の研究

目標が与えられたロボットやエージェントに，「意識的に行動しているか」をたずねると，その組み込まれた目標を答えることが可能である．しかし，これでは意識していることにはならない．複雑な行動を行っているときに，自らの行動を合理的に判断・認識して，行動に修正を加えることができる機械でないと，心を持つとは認識されないだろう．機械と対話している人に対して，その人の意図していること（その人の意識的行動）を理解できる能力も必要である．

通常の機械は，まさに，入力に反応して，出力（行動）を短絡的に（多少の演算はあるが）応答しているだけであり，その行動について意識（目的意識）しているわけでなく，もちろん，行動の本来の意味を理解しているわけでもない．

ブルックスの表象なしの知能機械の研究では，単純な行動の組合せから複雑な機能が創発されることを目指しているが，この枠組みでは，基本的には意識的行動の実現は難しい．行動を主体とした（単純な行動の組合せ）機械で意識機能を実現するには，行動を観測・制御するメタな機能が必要となる．

状況と自分の行動を意識する機械が実現されれば，人との対話において，人の意識しているものと，機械の意識しているものの差を判断でき，より適切な対応ができる可能性がある．たとえば，駅の切符販売業務などの機械化に応用すれば，顧客の意図するもの（「目的地」「早く到着したい」）と，窓口で提供しようとするもの（「経由地の入力と金額提示」）の違いを判断して，「経由地を入力してください」ではなく，本当の意図を考えて，「近い方の経路ですか？　金額が安い経路ですか？」と助言できる．

機械の心への取組み

ー物理的な身体なしの機械は心を持てるかー

この問いに対して多くの読者はNoと答えると思うが，ここではYesである2つの可能性を検討する．

1. 仮想環境の共有
2. あるがままの身体性

●仮想環境の共有

認知科学の最近の潮流では，機械が心を持つためには身体性（embodiment）が必要であるとされるが，身体性が何を意味するかについては必ずしも合意されているわけではない．ここでは，身体性の本質は物理的

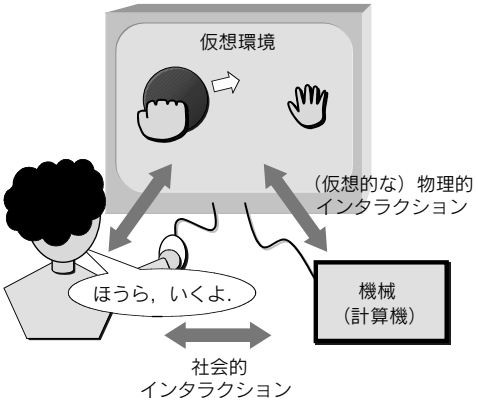

図-3　心を持つような振る舞いを機械に獲得させるための環境設定例

な身体を持つことではなく，「人」「機械」「環境」の3項間にインタラクションが存在することであると考える．**図-3**に示すように，機械（計算機）と仮想環境との間に（仮想的な）物理的インタラクションを持たせ，その仮想環境を共有した人との間に社会的インタラクションを持たせることにより，心を持つようなふるまいを機械に獲得させることが可能であろうと考えている[3]．

機械は，物理的／社会的インタラクションに基づいて強化学習や模倣を行うことにより状況に応じた適切な振る舞いを獲得できると考える．特に，快いコミュニケーションの成立を強化信号とした強化学習により，円滑で豊かなコミュニケーションが生じる行動の仕方を獲得できるだろう．また，コミュニケーションに基づく自他の行動の結果満足が得られた経験から，言葉の機能（要求，注意喚起，情報提供，情報請求等）や言葉の意味の学習が行えるだろう．

物理的な身体を持たないことの欠点として，自然な社会的インタラクションや自然な共同注意が生じるようにするための工夫が必要なこと，物理的な身体に起因する制約（自分の行動の影響は時空間的にその行動と近接して観測されることが多いこと等）が利用できないことが挙げられるが，いずれも心を持たせることに対する致命的な欠点にはならないと考える．

●あるがままの身体性

通常，計算機は身体を持っているとは見なされず，したがって，心を持つことはできないと論じられることが多い．しかしながら，たとえばパソコンのキーボードは皮膚に密に並んだ60個あまりの触覚センサであると見なすこともできる．そうすると，パソコンが（背中を掻いてもらっているように）「もうちょっと右．あ，そこそこ．」というような発話をする情景を想像できよう．映画「Toy Story 2」（Disney/Pixar）において，さまざまなおもちゃがそれぞれの身体に応じたインタラクションを行っていたことを思い起こす読者もいるだろう．

人とコミュニケーションするためには，人に近い身体を持つことが必要であると主張されることがあるが，ペットとのコミュニケーションがそれなりに成立することから考えても，人と異なる身体を持つものであっても，インタラクションの可能なチャンネルに応じた心を持てる可能性があると思う．

ー心の機能仮説ー

本来，機械自体にとっては，感情，意志，思想といった心的概念は意味を持たない．機械は本来，それが何をするものか，すなわち機能によってのみ意味づけられる．したがって，機械にとって心に意味があるとすれば，心は何らかの機能を意味するものでなければならない．同様に人にとっても，生物種としての人が，心を進化の中で獲得したのだとすれば，そこに何らかの「生き延びに有用な機能」が存在したと期待できるであろう．このようにして，次の心の定義の必要条件に達する．「心とは，人が進化の過程で獲得したある種の情報処理機能，またはその実装様式である」

では，心とはどのような機能であろうか．それをどのように探索していったらよいのか．我々は心の定義を明確には知らないが，幸いに人や物に「心を感じる」ことはできる．そこで，我々が心の機能の候補と考えるものを機械に実装してみて，その機械に我々が心を感じることができるのなら，それこそが心の機能である，というのが我々の接近法である．この2人称的接近法により，心の定義を実証科学の土俵に持ち込めると考えている．現在，我々が追求している心の機能の候補は，「合目的的でありながら，他者から予測されない（または予測を裏切る）行動の生成」である．他者から予測されないということは，競争の場では明らかに有用な機能であり，心の定義の必要条件を満たしている．しかしながら，いざこれを実装しようとすると，合目的性と予測拒否とは相反するものであることに気づく．ランダムな行動では合目的性を達成できない．一方，現在の囲碁や将棋のプログラムのように「最善手」に拘りすぎると，自分よりも計算能力で勝る相手に，容易に自己の行動を予測されてしまうことになる．

適当に自由度を残して合目的的行動を生成することは，行動の場で（実時間で）学習するプログラム（強化学習はその枠組みの1つ）により実現可能である．現在，この方向での研究が精力的に行われているが，パラメタの調整を越えるような，（新しい技能の獲得など）本質的な部分でのプログラムの変更を実時間で獲得するシステムは存在しない．その理由は，1つは学習時間の遅さであり，もう1つは実世界で動作させるコストである．標準的な強化学習のアルゴリズムでは，簡単なタスクを学習するのにも膨大な試行（学習）時間が必要で

ある．遺伝的プログラミングは新しい機能の獲得方法としてシミュレーション世界では（また，オフラインの学習では）魅力的ではあるが，1台の優秀なロボットを得るために，百万台のロボットをランダムに作ってテストしてみようとは誰も思わないであろう．生物の世界では，突然変異率を上げることで，種としての生き延びを図るものもある．しかしながら人は，ランダムにプランを生成して，イチかバチかに命をかける，なんてことはしない．そうするにはこれまでの投資が大きすぎるからである．とはいえ，合理的（機械的）に推論できる範囲にとどまっていては，新しい状況に対処する創造性は生まれ得ない．安全で効率的な学習（自己のプログラムの変更）にはどうすればよいのか，これが第1の問題である．

予測の拒否（行動の自由の獲得といってもよい）により生じるもう1つの問題は，仲間とのコミュニケーションの崩壊である．人のコミュニケーションは言語を使用する場合を含めて，「心を読む」ことがその中心となる．ここで「心を読む」とは，（言語発話，表情も含めて）相手の外部に現れた行動から，（相手の内部状態を推測し）将来の行動を予測することである．我々が親しい人に「いい天気だね」といったときに，多くの場合それは気象についての情報提供というよりは，我々の内部状態の伝達という意味が大きいであろう．ところが，予測を裏切る行動の生成能力は，相手が我々の行動を予測できないこと，言い換えれば我々が相手に自己の内部状態を伝達する手段を失うことになる．

我々は現在，合目的的でありながら「予測の拒否」の機能を持ったシステムの設計，評価を行っているところである．その1つの獲得目標は「嘘をつく能力を持ったエージェント同士における有意味なコミュニケーションの実現」である．正統のゲーム理論では，そのようなことは論理的に不可能であると放棄している．もう1つは，「行動生成における恣意的な制約を用いた探索空間の削減」である．いずれもこれまでの発想では不可能といわれそうな課題である．このようなアルゴリズムの開発自体も，工学として非常に面白いものである．しかしながら，我々の興味の中心は認知的観点であり，このような機能を機械に実装してみたときどのようなことが生じるのか，人が「心」を感じるような何かが生じないか，というものである[4]．

―クオリアの進化とその機械的実現―

1人称的に自覚する心の特徴に「クオリア」がある．クオリアとは，痛みの感覚や赤みの感覚，確固たる現実の存在感などに代表される感覚である．クオリアは，我々皆に備わった原始的感覚でもある．そしてクオリアの存在は，心的世界の独自性を主張する思想の大きなよりどころとなっている．ところが，「心とは進化の過程で獲得された機能である」と考える自然主義的立場に立てば，クオリアでさえも，進化の文脈におく必要がある（たとえば，デネットはこのスタンスをとっている）．すなわち，環境への競合的適応の結果，我々人を始めとした一部の動物にクオリアが備わっている，と考えるべきである．そこで，クオリアの進化的意義を考察し，2人称的心を実現する工学へのヒントがないか検討してみる．

すでに，茂木は『脳とクオリア』（日経サイエンス社）の中で，脳の神経回路からクオリアが生まれる原理の一端を提唱している．このアプローチは，1人称的心を手がかりに，3人称的な心の定義を模索するものである．それに対しここでは，2人称的心の実現にもクオリアの検討が貢献することを示したい．

知覚心理学者のグレゴリーは，クオリアは「現在を知らせる」という役割をすると述べている[5]．グレゴリーの説は，クオリアが，実世界の中で臨機応変に生きぬいてきた動物が必然的に身につけてきた機能であると示唆する．世界について考えるという高度な知能が発達した動物は，内的なイメージや思考が豊富になるあまり，現実への対応がおろそかになる．そこで，何かにつまずいた痛みや，敵が襲ってきた恐怖感，そして空腹感など，現実的にすばやい対処が必要な情報を，思考の中心部分に対して際立たせて提示する機構が必要となった．それがクオリアであるというのだ．

クオリアが「現在を知らせる」機能を果しているとすれば，クオリア（あるいはそれに相当する機能）がない動物，あるいは機械は，この現実世界の中で適切に対処するのが難しいと想像される．この考えの背景では，クオリアを持つ認知主体に，認知機能のモジュール性，感情の進化的役割，意識の単一性があることが想定されている．人はまさにそうした認知主体である．ガードナーの『個性を生かす多重知能の理論』（新曜社）では，人の知能は少なくとも8つの異なる種類があるとされるし，ミンスキーの『心の社会』（産業図書）では，人の課題達成の背後にいくつもの機能モジュールが動作していると指摘されている．感情は，古くは最も合理的でないものとして，理性の対極におかれたのだが，フランクの『オデッセウスの鎖』（サイエンス社）では，感情は一時の不合理さを越えて，長期的な（あるいは社会集団全体としての）利益を合理的に追求する手段として，進化的な意義があるとされる．また，脳生理学者のダマシオは『生存する脳』（講談社）の中で，合理的な思考の達成には感情が大きな役割を果していると主張している．

心の中で，中核的な思考を担っているのが意識である．意識は，その生物体に唯一の存在であり，その行

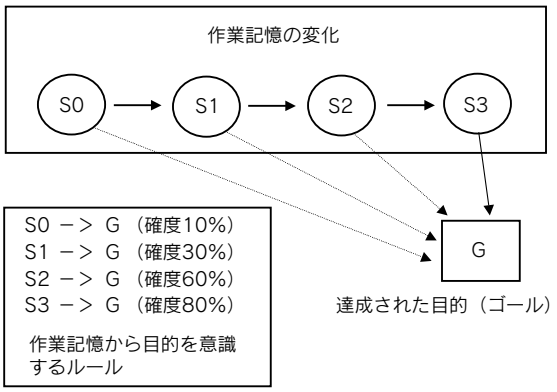

図-4　作業記憶と目的の関係付け

動の自由と責任の根源でもある．よって意識は単一でなければならない．複数の意識があれば，行動の体系性は生まれない．まさに「船頭多くして船山に登る」である．しかし，その意識は，心の中心にいるものの，認知機能のモジュール性により，心の中のすべてを把握できる状態にあるわけではない．そこで各モジュールは，クオリアとして，意識にそのモジュールの処理結果の重要性を伝えねばならない．意識は，経験から得た知識とともに，そのクオリアを手がかりにして，物理的行動や社会的行為を決定することで，結果として自己の生存確率を高めているのである．

機械に2人称的なクオリアを実装することは，上述したような人の認知システムを機械上に形成し，人に備わっているクオリアと同等の重要度判断機構，そして，意識と同等の，実時間性を重視した思考の中央集約機構を設けることであろう．考えてみると，最近のロボットやエージェントの開発の中には，すでにこうした路線のうえにあるものも多い．人においてクオリアがどのような働きをしているかを把握し，それと同等の機能を実現しようと試みることが，人にとって違和感のない機械を作る近道ではなかろうか．

ー自己状態観測と学習による意識の創発ー

エージェントが他のエージェントとの相互作用で，自己の推論状態を認識し，他のエージェントの推論状態を推定することで，自己・他者の意識が創発する仕組みについて考察する．

意識は，2つの場面（意味）で利用されている，1つは，意識的行動というように，あらかじめ目的意識を持って行動する場合である．これは，自発性，自主性を持った行動である．もう1つは，自らの行動や心理状態を観測し，認識していることである．言い換えると，自分の行動状態を把握して，その目的を説明できることである．

人の場合も，目的意識を持って行動する場合と，無意識に行っている行動を意識レベルで知覚して合理的な説明を加える場合がある[6]．ここでは，行動主義的に作られた刺激反応型ルールベースのエージェントが，最初は，環境に無意識に反応しているだけであるが，学習により，自己の行動を意識できるようになる方式について説明する[7]．

●自己の内部状態の観測

ソフトウェアが，ソフトウェアの挙動を観測する仕組みは，リフレクションと呼ばれている．オペレーティングシステムでは，アプリケーションプログラムのメモリへのアクセス違反やプログラムの暴走を監視して，その作用を制限している．リフレクションは，プログラムを観測する仕組みであるが，アプリケーションプログラムの詳細な行動に説明を加えることはできない．しかし，エージェントの推論状態を観測するには，必要な仕組みである．

ルールベースのエージェントでは，現在の推論状態を観測するには，2つの手段が考えられる．1つは，推論に利用されたルールの連鎖であり，これは，従来から推論結果に至る推論過程の説明に利用されている．もう1つは，作業記憶（Working Memory）内の事実や推論途中に導かれた中間結果である．前者は，推論途中のルールの連鎖が与えられるが，過去に類似の推論のルールの連鎖（類似の説明木）があれば，結論の予測に利用できる．また，作業記憶の状態のスナップショットから，どのような帰結に到達できるかを予測できる．当然，エージェントの持つ知識に変更がなければ，同一の作業記憶（外部環境の状態のコピーを含む：センシング情報も同じとする）であれば，同一の帰結が導かれるのは自明である．すなわち，エージェントがある目的を達成したときには，目的達成以前の作業記憶の内容は，その目的達成に到る作業記憶過程である．作業記憶からどのような目的に思考が向かっているかを推定できる（図-4）．

●複数目的への対応

エージェントが行動する場合に，複数の副目的を達成しながら，真の目的に向けて行動している．たとえば，サッカーロボットを例題に考えると，「ボールの発見を目的とする行動」「敵を探し，回避することを目的とする行動」「見方を探し，パスの可能性の確認を目的とする行動」「サッカーゴールを探し，シュートを目的とする行動」である．これらのすべての目的に関する情報が作業記憶には，重なって存在する．ある時点の作業記憶に対して，複数の目的に関連する情報（事実）が存在する，この状態で特定の目的に関連する情報だけを選択する必要がある．これには，上記のルールの連鎖の情報（演繹推論で推論過程情報）を利用することで，特定の目的に関係する作業記憶情報を選別できる．

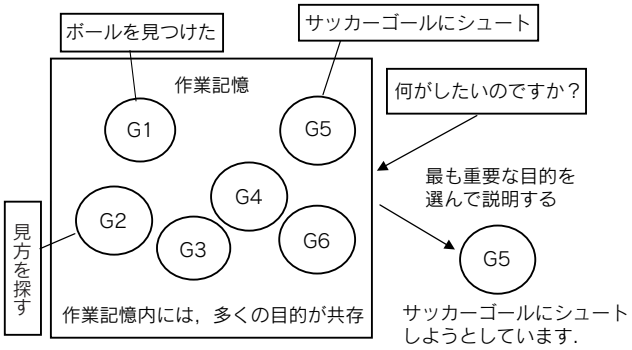

図-5　意識すべき目的の選択

次に，ある状態のエージェントに，「何をしようとしていますか？」と質問すると，「ボールを探している」「サッカーゴールを探している」「敵の位置を計算している」など複数の意識している目的を説明されても意味をなさない．最も重要な目的，あるいは最も注力している目的を説明できる必要がある．ルールや作業記憶の情報の中で，現在注力中の作業と利用された（追加された）最新の情報から主目的を選択できる．これは，複数の知識源がブラックボードモデルを用いて推論を行う場合での競合解消（次に実行するルールを1つだけ選択する作業）に似ている（図-5）．

●他者の行動の予測

人が人や高等動物に感情移入ができるのは，その生物の気持ちになって考えられるからである．実際には，他人の心を読めるわけでないが，他者の心理状態を自分の心に当てはめているに他ならない．エージェントにおいても同様の機能を持たせることが可能である．ただし，他者となる別の個体のエージェント（心を読まれるエージェント）が，元のエージェント（心を読むエージェント）と同じルールで動作しているときに正しい予測が可能となる．しかし，現実には，エージェントは個々独立の知識と経験を学習することが可能であり，その保有するルールには差がある．たとえば，サッカーロボットなどでは，相手チームと行動ルールが異なっているのは当然である．相手のルールをその行動から帰納的に学習して，他者モデルを作ることが必要である．ただ，他人に心が読まれないようにする機能や他人が予測し得ない行動をとるエージェントが出現した場合には，行動の予測はさらに困難となる．

ま と め

機械に心を持たせるには，多くの研究課題があり，いろいろな分野の研究者との協力が必要である．ここでは，情報処理と人工知能に関連した研究項目について，前述の議論をまとめることにする．

ー機械の心の機能的要件を考える（認知的）ー

従来研究は，人の心の解明に注力してきたが，工学的な観点での心の定義があってもおもしろい．あるいは，人の心の一部のモデル化であっても，十分である．さらに，拡大解釈して，心が生まれる環境や条件の研究も重要である．エンタテインメントロボットに，人が心を感じるモデルの研究もさらに発展させられる．クオリアについても具体的モデルの提案が考えられる．

たとえば，次のような心のモデル研究が想定できる．

- 人が機械に心を感じる仕組み
- 心の部分モデル
  - ・感情生成モデル，感情抑制モデル，感情と動作のモデル，感情と表情のモデルなどの感情モデル
  - ・意識モデル，自己認識のモデル，他者の心の推定モデル
  - ・自発的行動のモデル，目的生成モデル，上位の目的の推定モデル
  - ・心の発生モデル，創発モデル
  - ・無意識と意識の相互作用のモデル
- 心の社会のモデル（ミンスキーの心の社会の発展）

ー機械の心の機能実現（工学的）ー

工学的技術による「機械の心」の構成には，まず，実現すべき機能と手段を考える必要がある．

モデルの実装：上述の心のモデルの実装方式を考える必要がある．人工知能の知識表現やロボットのサブサンプションアーキテクチャ，制御理論で利用される状態表現，統計モデルなどの仕組みを利用してモデル化を試みる必要がある．

発現の仕組みの研究ツール：各モデルがどのように心として機能するかを考える必要がある．顔による感情表現であれば，顔表現のCGやロボットが必要であり，ペットのしぐさやロボットの行動による心の表現であれば，プログラミング可能なペットロボット（AIBOは無線LANで制御できるようになっている）や，プログラミング可能な腕を備えて動くロボット（Robovie）などを利用する必要がある．しかし，ロボットや人のような身体を持たない既存の機械に心を持たせるのであれば，その形や構成に合った発現の仕組みが必要となる．

学習・進化の仕組み：心の機能仮説に基づく機械や

心の進化を検証するにも，人工知能の各種学習機能（帰納的学習，説明に基づく学習，強化学習など）や進化計算（遺伝的アルゴリズムや遺伝的プログラミング）が研究のスタートとしては利用できる．複雑な心の実現には，作りつけのプログラムだけでなく，学習を利用するのが有効である．

ー何の役に立つのか（ニーズの研究）ー

機械に心を持たせることで，どのようなメリットがあるかを考えることで，産業応用や社会生活への応用が期待できる．

•気持ちの分かる機械：

人が対応する窓口では，利用者の意図することが伝わりやすい．これらの機械が利用者の気持ちが分かれば，より利用しやすい機械が提供できることになる．また，窓口業務だけでなく，福祉分野の応対でも，利用者の気持ちが分かることで，優しさを感じさせる「思いやり機械」に進化できる．

•やる気のある機械：

機械は黙々と命令に従い作業を実行するが，命令がなければ，停止してしまう．そこで，有能なビジネスマンのように，自ら進んで，仕事を見つけだし，問題を解決できる「やる気のある機械」ができる可能性がある．自発的・自主的・自律的なロボットでは，いろいろな問題をいち早く発見し解決してくれるかもしれない．

•世話好きな機械：

介護や福祉，あるいは，情報家電の分野で，家庭の中で利用する機械を想定すると，その家の住人（ロボットの持ち主，ペットロボットの買主）の気持ちを察して，いろいろと世話を焼いてくれる機械やロボットが生まれてくるかもしれない．世話好きを越して，お節介なロボットになると問題である．

ー社会的影響（社会学的）ー

•自発行動による社会や人への負の効果と責任：

心を持つロボットや機械が，自発的行動で事件を起こしたり，人に危害を加えたりしたときは，誰が（何が）責任を負うべきであろうか．現状では，ロボットが責任を負うのは難しいため，製造物責任や所有者責任（ペットの買主）の責任となりそうだが，自発的行動に，他者が責任を負うべきであるかの議論が必要である．情報処理学会誌（2001年11月号）のほっとタイム「超機械知能たち」の翻訳記事にも，「超知能機械をどう教育すれば人に危害を加えないか」の議論がなされている．

強化学習機能を組み込むことにより，個々のユーザと価値観を共有するように育つロボットは，将来，技術的には実現可能になると思われるが，このロボットが何を正しいことだと判断するかは，それを教えるユーザに依存するため，悪用される危険性を伴う．ロボットは従来の受動的な道具とは異なり，悪用を企てた人を上回る能力を持ち得るし，さらには，自らと同じ物を産出する能力さえ持ち得るものであるから，価値観を伝えることができる機能を持たせた場合の危険性はきわめて大きい．たとえば，原爆を作る能力がない人でも，兵器を保有することを善とする価値観のロボットを育てることは可能かもしれない．

この種の問題に対しては，たとえば殺人できないような良心回路をロボットに組み込んでおけばよいという意見もあるが，「教育によっては上書きできない，望ましい価値判断基準とその適用条件」をあらかじめすべて書き下しておくことはできない（フレーム問題）．

•機械の地位：

人権に相当するロボット権は必要か．どのような責任と義務を与えて，どのような権利を保証するのかの議論が必要である．ロボットが収入を得る存在になれば，税金を納めてもらうことになるだろう．

•廃棄基準（死亡認定）や修理基準（病気の判定）：

壊れたロボットは，どのような条件で廃棄できるのか．ロボットの心が傷つくような処理はできないと考えられるが，では，廃棄することができないことになるのか，ロボットの安楽死は可能か，など悩ましい問題が多い．

•所有権の移動：

心を持つ機械の所有者は存在するのかという問題がある．心があっても所有されるべきか．また，所有者がロボットを購入したとき，販売したときは，人身売買と同様な問題にならないかの議論が必要である．

機械の心の研究は，情報処理研究の1つのグランドチャレンジになるに違いないが，社会的な影響も考えつつ研究を進めなければならない．

参考文献

1) McGinn, C.: The Mysterious Flame, Basic Books（1999）. 石川幹人ほか（訳）: 意識の＜神秘＞は解明できるか, 青土社（2001）.
2) Taki, H. and Hori, S. et al.: Environment Generation Intelligence Framework for Humanoid, Proc. of 1998 Japan-USA Symposium on Flexible Automation, Vol.2, pp.445-449（1998）.
3) Oka, N., Morikawa, K., Komatsu, T., Suzuki, K., Hiraki, K., Ueda, K. and Omori, T.: Embodiment without a Physical Body, Workshop on Developmental Embodied Cognition, Edinburgh, Scotland, UK（2001）.
4) 伊藤　昭: コミュニケーションは心ー「心の理論」と他者理解のモデル, 岡田他編, 別冊bit, 身体性とコンピュータ, pp.269-283, 共立出版（July 2000）.
5) 石川幹人ほか（編著）: 心とは何か～心理学と諸科学との対話, 北大路書房（2001）.
6) 下條信輔: サブリミナル・マインド, 中公新書（1996）.
7) 濫, 松田, 堀, 安部: 自己行動のリフレクションによる他者の行動・意識推定モデル, 人工知能学会, 第55回知識ベースシステム研究会資料（to appear）（2002）.

（平成13年11月28日受付）