ロボットは価値的記号を理解できるか

久木田水生*

2011 年 12 月 25 日
京都生命倫理研究会

悦びと共感こそは、責任へのささやかな第一歩である。
——リチャード・パワーズ『舞踏会へ向かう三人の農夫』（柴田元幸訳）——

1 序

現在，ロボット技術[*1] は工業，交通，金融，軍事，医療，娯楽の分野に浸透しており，これにともないロボット工学ではロボットに倫理的なコードを組み込む研究が進められている．これらの研究はロボット倫理学 robot ethics，機械倫理学 machine ethics，人工道徳 aritificial morality などと呼ばれている．これらは工学的に興味深くまた実際的な重要性のあるチャレンジである一方，以下の点で哲学・倫理学にも重要なインパクトを与えることが予想される：

1. 人工知能やロボットに倫理的行動を実行させるためには，まず人間が従っている倫理的コードをより曖昧さの少ない規則の形で与えなければならず，このことは私たちが無意識に従っている倫理的コードを明示化することを要求するだろう．
2. 人工知能やロボットに道徳的判断をさせる，あるいは道徳的に重要な行為をさせることはどこまで許されるかという問いに対して私たちは答えを用意しておかなければならない．
3. 人工道徳の試みは単に道徳的行為をシミュレートするだけではなく，本当に道徳的な行為者を人工的に作ることができるかという問いを突き付け，さらにこの問いはそのような存在者を私たちがどのように取り扱うべきかという新しい倫理的問題を提起する．

1, 2 に比べると 3 はいささか SF 的に過ぎる問題関心と思われるかもしれない．しかし Floridi and Sanders [6] はある「抽象のレベル」において動物，ロボットなどを含む，人間以外の対象も道徳的行為者（あるいは道徳的受容者 moral patients）とみなすことができるし，またそうすることがコンピュータ倫理などにおいて必要であると論じている．

フロリディらの議論は，いわばチューリング・テストの道徳版である．従ってチューリング・テストに対する批判と同様の批判が彼らの議論にも当てはまる．特に Searle [13] や Harnad [9] によって指摘された問題は重要である．ハーナッドは「人工知能が操作する記号がいかにして現実の対象を意味しうるか」という問題を

* 京都大学文学研究科研究員．minao.kukita@gmail.com

*1 ここでは「ロボット」という語は自律的に活動する人工的な対象一般を意味するものとし，物理的な身体を持たないものも含める．従って人工知能，コンピュータ・ウィルス，ボットなどもロボットである．

提起し，これを「記号接地問題 the symbol grounding problem」と呼んだ．記号接地問題に対しては現在までに様々なアプローチが提案されてきた[*2]．そして現在，人工知能とロボット工学はより人間に近い知性を実現するシステムを開発してきている．こういった発展は，人間の知性と機械の知性の間の境界をますます曖昧にするばかりでなく，知性は人間と機械を要素として含む，より大きなシステムによって実現されるとさえ見なされるようになっている[*3]．

同様のことが道徳に関しても期待できないだろうか？ すなわち，人工道徳が新しい道徳のあり方を提起する，あるいはこれまでに私たちが気付いていなかった道徳性の本質を明らかにするという可能性はないだろうか？ 本発表はこの問題について考察する．

フロリディらのような議論は人工道徳的行為者を十分に擁護するものではない．より一般に人工道徳的行為者を実現する試みは，かつての人工知能と同様の困難に直面する．しかしながら人工知能とロボット工学の発展が新しい知性のあり方，あるいは知性についての新しい研究方法と新しい知性観を可能にしたように，人工道徳の発展によって新しい道徳のあり方，新しい倫理学研究の在り方と新しい道徳観を可能にするかもしれない．私たちがこの際に参照するのは，人工知能における記号接地問題への様々なアプローチの成功と失敗，そして心の哲学において近年生じてきた「拡張された精神」仮説である．

2 倫理的ロボット

現在ロボット技術は産業のみならず，軍事，警察，交通，医療，介護，教育，家事，娯楽など，私たちの社会の隅々に浸透しつつある．そして今後このような傾向はますます強まっていくだろう．工学者たちは近い将来に人間と自律的なロボットが日常的な場面で「共生」する社会が実現すると予想している[*4]．この変化が社会に与える影響は甚大である．機械の誤作動や故障による物理的な危害はもちろん，ロボットを介した情報の収集，プライバシーの侵害なども懸念される．ペット・ロボットやセックス・ロボットの心理的影響も考える必要があるだろう．現在でも恋愛シミュレーションゲームに没頭し現実の人間関係を築く意欲を失っている人々や，ネットゲームに熱中するあまりに社会生活に困難をきたしている人々がいる．遠隔操作される，あるいは自律的に行動する兵士ロボットの実用化は戦争の在り方，また戦争の捉え方を大きく変えるだろう．自国が戦争に向かうことに対する大きな心理的抵抗の一つは，自国の兵士たちの死に対する危惧であろう．もし自国の兵士が全く死傷することなく戦争を遂行することができるとすれば，戦争に対する反対の一部は根拠を失う．

このような背景のもと現在ロボット技術に関する倫理的問題が盛んに議論されるようになっている．そして「ロボット倫理学 robot ethics」という新しい分野が生まれている．「ロボット倫理学」によって人々が意味していることは，大きく言って次の三つに分類されるだろう．(1) ロボット技術の社会的影響および，製造者と利用者の倫理についての研究．(2) ロボットを倫理的に振る舞わせる研究．(3) ロボットが道徳的存在になる可能性についての研究．(1) は主に倫理的・社会的・法的問題，(2) は工学的な問題，(3) は哲学的問題である．私たちがここで関心を持つのは特に (3) についてである．しかしこの問題について考える前に，(2) の試みの例を知ることは，(3) の方向の研究をよりよく理解する助けになるだろう．

自律的に行動するロボット（機械／プログラム）が，人間に大きな利益や損害を与えうることは明らかである．さらに機械が倫理的な選択に関与するような状況も生じうるだろう．社会の隅々にロボットが浸透してい

[*2] Cf. 1990 年から 2005 年までの記号接地問題への様々なアプローチとその評価に関しては Taddeo and Floridi [16] を参照.

[*3] Cf. Clark [4, 5]

[*4] 経済産業省「平成 18 年度技術戦略マップローリング事業」.
http://www.ai-gakkai.or.jp/jsai/whatsai/PDF/rloadmap2.pdf

図 1　MedEthEx を組み込まれた「倫理的」ロボット NAO

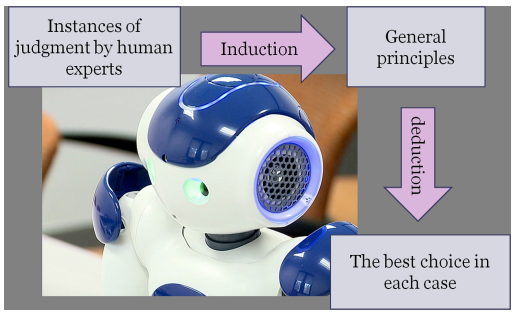

図 2　MedEthEx の推論

くにつれて，そのようなロボットの必要性はますます高まっていくだろう．実際，2007 年にジョージア工科大学のロナルド・アーキンは，戦時における倫理的基準に従うことができる戦闘ロボットのためのソフトウェアとハードウェアの開発のために，アメリカ陸軍からの資金を獲得している[*5]．そういった状況において機械に「良い」選択をさせるためには，人間が良いと判断するような選択のための規則——そのような規則があるものとして——が，機械にも遂行できる形で与えられなければならない．しかしながら人間にとってそのような規則はしばしば暗黙のうちに従われているものである．従って倫理的選択の状況で人間が従っている規則を明示化することが，倫理的機械の実現への最初のチャレンジである．

マイケル・アンダーソンとスーザン・アンダーソンは，患者が薬を飲むことを拒否した時に介護士が迫られる倫理的な選択——患者の意思を尊重して放っておくか，患者の身体的健康を重んじて医師に連絡をするか——に対処するための人工知能システム MedEthEx を開発している（図 1）[*6]．このシステムには特に注目するべき特徴が二つある．第一に，このシステムは患者の感情と患者の健康という二つの価値を天秤にかけて，特定の状況においてどちらを優先するべきかという，現実的な倫理的的問題に対して判断を下す能力を持っているということである．第二に，このシステムは機械学習を通じて，同じ状況において人間の介護士が（無意識に）従っている規則を発見することができるということである．この際，MedEthEx は帰納論理プログラミング[*7]という，個別事例から一般法則を導く人工知能のテクニックを用いている．MedEthEx はそのようにして導きだした一般法則を個々の状況にあてはめて判断を下している（図 2）．従って MedEthEx は

[*5] Cf. Wallach and Allen [19], p. 21.
[*6] Anderson and Anderson [1].
[*7] 帰納論理プログラミングについては例えば Gillies [7]，古川 [21] などを参照.

人間がプログラムした規則に従っているだけではなく，自らの「経験」に基づいて専門家の判断をシミュレートする．

アンダーソンらのシステムに対する人々の反応は様々であるが，中にはこのような研究に対して強い拒否反応を示す人々もいる．例えば彼らの研究を紹介したある Web サイトの記事には次のようなコメントが寄せられた[*8]．

- “the robot merely follows something that seems to be a simple rule set, not making true decisions by weighting the whole situation, since robot does not have situation awareness”
- “Michael Anderson is obviously a misguided fool and will be the first to die at the hands of an ethical robot.”
- “There are some things that only people can do that robots will never be able to do. The more we try to succeed in this technology, the more we are just trying to play the roll of God!! People need to [···] do their own work, and stop trying to program robots to do the work for them, in an ‘ethical way.”’

こういった人々の反応は倫理的ロボットが直面するかもしれない心理的・社会的障壁を予感させる．機械が人間と同等もしくはそれ以上の判断ができる時でも，人間は機械に判断をゆだねることに不安を覚えるものである．例えばロンドンの地下鉄の全面的な自動化を阻んでいるのは「技術的問題ではなく，政治的な問題」である（Wallach and Allen [19], p. 14）．

3 ロボットは倫理的になりうるか？

前節で紹介したように，現在ロボット工学者たちは倫理的な判断を下すロボットの研究・開発に着手している．ここから私たちは自然に，そのようなロボットは本当に倫理的な判断を下していると言えるのか，という疑問，そしてより一般的にはロボットなどの人工物が道徳的存在者になることが可能なのかという疑問に導かれる．Floridi and Sanders [6] は人工物を道徳的行為者とみなすことは可能であるし，またそうすることが必要であると論じている．ここでは彼らの議論を紹介し，その議論の妥当性を検討しよう．

フロリディらはまず，ある対象が道徳的行為者であるか否かを問題にする際には，私たちは議論の**抽象のレベル**を決定しなければならないと主張する．抽象のレベルとは一言でいえば，対象あるい状況のどの観察可能性質 observables を当面の論点にとって関連する要因とみなすか，ということである．例えばある人は自動車についてその燃費と安全性を重視するかもしれない．別な人はスピードとデザインを重視するかもしれない．彼らは同じ自動車について全く異なる判断を下すだろう．それは彼らが異なる抽象のレベルを採用しているからである．彼らの判断について評価するには，彼らの立っている抽象のレベルを適切に特定する必要がある．

それでは道徳的行為者について論じる際に適切な抽象のレベルとはどのようなものであろうか？ フロリディらは，道徳的行為者性にとって適切な抽象のレベルは相互作用性，自律性，順応性という三つの観察可能性質から構成されるべきだ，と主張する．彼らはここで，一般的には重要と考えられる責任 responsibility という要因を要件から除外している．彼らによれば，幼児や動物のように，責任の主体とは考えられないが，道徳的帰結を生みだした主体と考えることができるものがあり，従って道徳的行為者であることと，責任主体であることは別問題である．主体はある行為の帰結に responsibility を持つことなしに，accountability を持つ

[*8] http://news.discovery.com/tech/robot-makes-ethical-decisions.html （2011 年 2 月閲覧）

ことができる．

この抽象のレベルのもとでは，環境と相互作用し，自律的に行為し，そして自らの行為の規則を順応的に修正することができる主体は道徳的行為者である．従ってこの抽象のレベルにおいてはロボット，動物，企業など，通常は道徳的行為者とみなされない対象も道徳的行為者とみなされる．彼らはこのような抽象のレベルが，従来の標準的な倫理理論では対処できないコンピュータ倫理の問題に取り組む際に必要であることを論じている．というのも，コンピュータ・ネットワーク上で活動するソフトウェア・エージェントが重大な道徳的帰結を持つ行為をなしうる状況に対して倫理学が適切に対処するためには，それらを道徳的行為者と考えなければならないからである．

このように道徳的行為者性にとって適切な抽象のレベルを特定したうえで，行為の道徳性は観察可能性質の各々をパラメータとする閾値関数によって計測される．すなわち，ある行為の観察可能性質の値を表すパラメータ $p_1, p_2, \ldots, p_n$ と閾値関数 T に対して $T(p_1, p_2, \ldots, p_n) \geq t$ ならばその行為は道徳的に良いと判定される，というように．もちろんこの閾値関数 T と閾値 t を定めるのは観察者の役割であり，行為者自身がこれらについて知っている必要はないし，自覚的に「良い」と判定される行為を選択する必要もないのである．

4 倫理的記号接地問題

フロリディらは，道徳的行為から責任という要件を取り除くことによって，道徳的行為者の概念にシフトをもたらそうと試みている．彼らのアプローチはある意味で，知性の基準として外面的なコミュニケーションの成立のみを要求するチューリング・テストと類比的である．実際，彼らは行為の責任が行為者にあるか否かを知る際には，行為者の意図が重要である，と述べている（Floridi [6], p. 365）．つまり彼らが責任という観点を適切な抽象のレベルから除外するとき，行為者の意図という内的な要因を考慮する必要性を除外しているということになる．ここではこのようなアプローチに対する可能な反論を考察しよう．

4.1 記号接地問題

チューリングは，人間の心の中で何が起こっているのかは他人には知ることができないのだから，ある機械が考えているか否かは，外面的な振る舞いを見て判断するしかない，と考えた．そこで彼が提案したのが今日「チューリング・テスト」として知られている，次のようなテストである（Turing [17]）．ある判定者に機械または人間と，パソコン通信のチャットのような仕方で対話をさせる．判定者は自分の相手が機械なのかそれとも人間なのかを知らされていない．しばらく対話をして，判定者に自分の相手がどちらだったかを判断させる．この過程を繰り返して，判定者が機械と人間とを見分けることが出来ていなかったら，その機械は人間と同様に考えていると見なしてもよい．

チューリングは人工知能が十分に発達すれば，このテストにパスするようなものが現れる可能性はあると考えていた．確かに原理的にこれが不可能であると想定する理由はないし，実際にかなりよく人間を欺くことの出来る人工知能も作られている．一方，このテストの有効性に対しては多くの激しい反論が向けられた．Searle [13] は「中国語の部屋の議論」を提示して，記号の意味の理解が伴わなければ，いかに入力された記号と出力された記号が適切に対応していても，その記号操作システムが知性を持っているとはいえず，それゆえチューリング・テストは知性のテストにはならない，と反論した．Winograd and Flores [20] は，言語を理解して使用することは「コミットメント」に立ち入る行為である一方，コンピュータにはこれが不可能であり，したがってコンピュータが言語を理解することは出来ないと論じた．こういった反論以降，人工知能が本当の

意味で思考しているといえるためには，単なる記号的 AI，記号を操作する「構文論的機械」であるだけでは不十分で，それはまた言語の意味を理解する「意味論的機械」でなければならない，という認識が一般的になった．今日の人工知能研究者たちは，チューリング・テストで知能が計れるとは考えていない．しかしかといってサールやウィノグラードたちのように，悲観的でもない．彼らは様々なアプローチによって，記号を理解する人工知能を作り出そうとしているのである．

適切な意味論的能力を人工知能に実装させる研究の端緒となったのは Harnad [9] である．ハーナッドは記号的 AI の置かれた状況を，中国語を学ぶのに中国語－中国語辞書だけを与えられた人間に譬える．その人間は一つの言葉の意味を知るために別の言葉（それ自体の意味も彼は知らない）を参照する以外に方法はなく，どこまでいっても言葉を有意味なものとして理解することはないだろう．彼は言葉の「メリーゴーラウンド」に乗せられたまま堂々巡りをするばかりで，いつまでも地に足を着けることが出来ない．あるシステムの操る記号が単に形式的なものではなく，有意味な記号であるためには，記号の対象が存在している世界に，記号が結びついていなければならない，とハーナッドは言う．それではどうすればある記号的人工知能システムの扱う記号を記号以外の何かに結びつけることができるのだろうか？ この問題を彼は「記号接地問題」（以下，SGP）と名づけた．以来，SGP は人工知能・人工認知システムにおける主要な問題の一つと見なされている[*9]．

この問題に対してハーナッドが提案した解決は概略以下のようなものである．記号的 AI は外界との直接的なつながりをもたないが，それでは記号接地は不可能である．外界との接点を持たせるために第一に考えられるのが，外界の対象を認識する能力を人工知能に持たせることであろう．そこでまずハーナッドは人間が外界の対象を区別・同定する能力に注目する．外界からの個々の入力を区別する能力は，「イコン的表象」，すなわち「離れたところにある対象からの感覚表面上への投射の内的なアナログ的変形」を持つ能力に依存する．これは外部からの刺激に応じて，内的状態（のある部分）が，付随的に変化できる，ということである（例えば外界の音を電気信号に変換するマイク）．イコン的表象は，対象からの投射が変形を受けるメカニズムによって，当の対象と結び付けられている（つまりその対象の記号である）．従ってイコン的表象は接地していると見なすことが出来る．

しかしながらイコン的表象を形成するだけのシステムは人工知能と呼ぶに値しない．知能であるためには少なくとも，外界からの入力をこれこれのものとして解釈する必要があるだろう．外界からの入力の解釈は，典型的には，その入力をある範疇に属するものとして同定することによって遂行される．しかしある対象がある範疇の成員であると同定するためにはイコン的表象では十分ではない．というのも，この目的のためには当の範疇の成員をそうでない対象から区別するのに十分な，感覚的投射の「不変的特徴」が選択されなければならないからである．このように選択された特徴をハーナッドは「範疇的表象」と呼ぶ．対象を分類するのに十分なイコン的表象と範疇的表象が得られたならば，その表象は接地しているということが出来る．さらにシステムが適切なシンタクスを備えていれば，そこから通常の形式的体系と同様に複合的な表象を作っていくことが可能だろう．例えば「赤」という範疇的表象と「馬」という範疇的表象を組み合わせて「赤い馬」という範疇的表象が作られる，というように．この新しい表象は「赤」および「馬」という二つの表象のそれぞれの接地を継承することが出来る．

ハーナッドの提案は，SGP へのアプローチのごく大雑把なスケッチであるが，記号の意味の理解のためには形式的意味論を与えるだけでは十分ではなく，世界を認知・表象し，かつ対象を識別する能力が不可欠であるということを示した点は，重要である．ハーナッド以降，SGP については多くの研究者が様々な見解を表

[*9] ただし表象を介さずに知覚と行動を結びつけることによって，SGP は回避できるとする立場もある．Cf. Brooks [2, 3], 谷 [22].

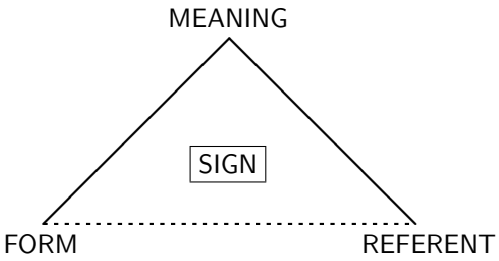

図 3　Semiotic Triangle（Vogt [18], p. 180, Figure 2 より）

明しているが，SGP に対して積極的なアプローチをとる研究者たちは，外界を認知するメカニズムだけではなく，行動を通じて環境や他のエージェントと相互作用する能力，学習する能力，および言語コミュニケーションを進化させる能力を重視する傾向にある．こういったアプローチによって作られた人工知能が扱える言語はこれまでのところ非常に単純で原始的なものにとどまっている．にも関わらずこれらの人工知能の用いる言語の方が，記号的 AI が扱うはるかに複雑で洗練された言語よりも，人間の言語のモデルとして優れていると考えられるのは何故だろうか？

私たちは，しばしば暗黙のうちに，記号の意味に関して何らかの理論（あるいは前理論的直観）を採用しており，そのために，ある人工知能の扱う言語は意味を持たないと感じる一方，別の人工知能の扱う言語は意味を持つと感じるのである．それではその理論を明示的に述べるとどのようなものになるだろうか？

Vogt はパースにならい，記号（sign）を形式，意味（解釈），指示対象という三つの構成要素とその間の関係によって形成されるものと考える．ここでは記号は単なる文字や文字列ではなく，それが何か他のものと結び付けられる過程，あるいはその結果として成立する関係，すなわち**セミオシス***semiosis* として捉えられる（図 3 参照）．これらの関係がどのようにして成立しているかによって記号はイコン，インデクス，シンボルに分類される．その関係が形態的類似性によっている場合，その記号はイコンとなり，因果関係や相関などによっている場合はインデクスとなる．一方，その関係が恣意的なもの，あるいは慣習的なものである場合にその記号はシンボルとなる．

ある表現（形式）がセミオシス（の一部）であるためには，そもそも意味と指示対象を含んでいなければならない．従って，このように捉えられた記号はそもそも SGP を生じさせるものではない．しかしそれではセミオシスはどのようにして成立するのだろうか？ Vogt は次のように述べる．

> 意味は記号がどのように，そしていかなる機能とともに構成されるかということに依存すると私は論じてきた．そのようなものとして，記号の意味はエージェントの身体的経験，ならびにエージェントと指示対象との相互作用に基く，形式と指示対象の間の機能的関係と見なされうる．その経験はエージェントの，指示対象および／または形式との相互作用の歴史に基づく．（Vogt [18], p. 180）

Vogt はこのようなものとしてのシンボルを**記号論的シンボル***semiotic symbols* と呼ぶ．Vogt の言うとおり記号論的シンボルが「**定義から***per definition* 有意味かつ接地して」いるのであれば，問題はいかにしてセミオシスの過程を，人工的なエージェントにおいて生じさせることが出来るかという点にある．Vogt の提案する方法は，マルチ・エージェント・ロボティクスである．彼は複数のロボットに，特定の課題を達成させる実験において，コミュニケーションの手段として語彙や合成的文法を創発させることに成功している（Vogt [18]; Steels and Vogt [15]）．スティールズはこのようなアプローチが SGP を解決するのに十分であると主張する

(Steels [14]).

このようなロボット・エージェントの扱うシンボルが，実際にセミオシスを成立させており，従って記号を接地させていると言えるのだろうか？ Taddeo and Floridi [16] は，SGP を解決していると言えるシステムは，内在的な意味論的能力や外的な意味論的規則へのコミットメントを持たないという条件（「ゼロ意味論的コミットメント条件 zero semantical commitment contition」）を満たす必要があり，その意味でこれまでに提出された SGP へのアプローチはすべて十分な解決を提供していないと主張する．Vogt らによって追求されているようなアプローチも意味論を生成させるメカニズムについては設計者とプログラマに依存しており，それゆえに内在的な意味論的能力へのコミットメントを持つと言えるだろう．従って彼らの方法も Taddeo らにとっては SGP の解決にはなっていない．しかし Taddeo らの要求するゼロ意味論的コミットメント条件はあまりにも強すぎる要求である．もし記号を接地させていると言えるいかなるシステムもこの要求を満たさなければならないとしたら，果たして人間は記号を接地させていると言えるだろうか？

ここでの最大の問題は，記号一般がいかにして意味を持つということを説明する，十分な同意が得られた理論が存在していないことであると思われる．. SGP に対するアプローチは，そもそも記号が意味を持つということはどのようなことか，あるいは記号の意味とは何かについての理論を，たとえ大雑把にでも，前提する必要がある．しかし実際の人工知能研究において，どのような理論が前提されているかは必ずしも明確にされていない．そしてそのことは，SGP への多種多様なアプローチに対して，私たちが評価を下すことを困難にしている．しかしながら天下りに与えられた理論によって SGP へのアプローチを評価することには注意しなければならない．そのことは SGP へのアプローチを過度に制限してしまうことに繋がるからである．科学において理論とモデルと現実についての認識が相互作用するように，意味の理論と SGP へのアプローチと記号的関係についての認識が相互作用することによって，私たちは記号についてのより良い理解を形成していくことができるだろう．

4.2 倫理的記号接地問題

再び，人工道徳的行為者に目を向けよう．フロリディの提唱する抽象のレベルにおいては上述した MedEthEx は十分に良い道徳的行為者である．しかしこの判定が私たちの日常的な感覚と乖離していることは確かである．私たちの多くは MedEthEx を道徳的行為者とは考えないだろう．なぜか？ フロリディらが責任という観点を抽象のレベルから除外し，それによって行為者の意図というものを考慮の対象外に置いていたことを思い出そう．MedEthEx に欠けていると私たちが感じるのは，まさしくここで考慮からはずされていた意図である．道徳的行為者は，自らの行為の帰結を予測し，その帰結が良いものであることを意図しなければならない．もちろん MedEthEx は自らの行為の結果を予測し，複数の行為の結果を比較する．フロリディらの用語を使えば，閾値関数の値をあらかじめ計算してより道徳的価値の高い値を出力する選択肢を選んでいる．しかしこのシステムはもともと与えられたプログラムと，人間の専門家から学習した規則に従ってそうしているにすぎず，それらの値が何を意味しているのかを理解していない．それらの記号を現実の価値に結び付ける仕事はオペレータである人間に委ねられている．従ってサールやハーナッドの指摘がこのシステムについても全く同様に当てはまるのである．ここでこの問題を倫理的記号接地問題（以下，ESGP）と呼ぶ．

ただしこのように言ったからといって，人工道徳に従事している工学者たちを批判しているのではない．彼らは人間と同じような道徳的行為者を作ることを目指しているわけではない．彼らのほとんどにとっての目標は，実用に役に立つシステムを作ることである．しかし人間と同じような道徳的行為者を作るにはどうしたらよいかという問題は，哲学的・倫理学的には興味深いことである．そして ESGP の解決はこのような意味で

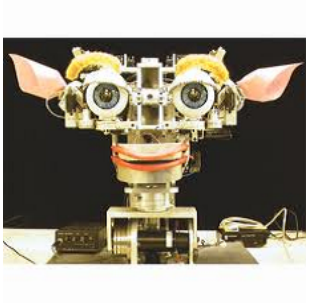

図 4　KisMET

の人工道徳の実現への必要なステップであるように思われる．

それではどのようにして ESGP は解決できるだろうか．現在の私たちはここ 20 年の間に蓄積された SGP に対する取り組みについての知識を利用することができる．Vogt らのマルチ・エージェント・アプローチが SGP に対して有効だとするならば，私たちは同様のアプローチが ESGP にも有効であることを期待することができる．ただし彼らのアプローチをそのまま ESGP に応用することは不可能である．彼らのモデルにおいて接地が試みられる記号は，色や形など，ロボットに付属するセンサーによって知覚可能な特徴を意味する記号である一方で，価値的記号の意味するものは知覚可能な特徴ではない．価値的記号の意味するものは，そのモデルに参加しているロボット・エージェントたちにとっての「利害」でなければならないだろう．

従って ESGP の解決には越えなければならない二つの課題がある．一つは人工行為者たちに何らかの仕方で利益を持たせること，もう一つはそのような利益を行為者たちが知覚する方法を確立することである．もし前者の課題をクリアすることができたならば，後者についてはその利益を何らかの仕方で視覚的に表出することは困難ではないだろう．例えば KisMET のようにロボットの内部状態に応じた表情をロボットに持たせることができる（図 4）し，その表情を知覚する認知メカニズムをロボットに与えることも現在ではそれほど難しい技術ではない．

では人工的行為者に利益を持たせることはどのようにして可能であろうか．これには二つのアプローチが考えられる．一つはロボットに快苦の感覚や感情を持たせることである．この方向の研究としては，生物の内分泌系に類似した「内分泌系モデル」を持つロボットによってロボットに情動を持たせようという尾形と菅野による WAMOEBA が挙げられる（図 5）[*10]．もう一つの方向性は MacLennan [10, 12, 11] や Grim and Kokalis [8] などによって推進されている，進化的アプローチである．こういったアプローチにおいては，コンピュータの内部の環境の中で行為者たちが生存，繁殖，進化をしていく過程において協力行動やコミュニケーションが発達させることに成功している．マクレナンはこのアプローチを「総合動物行動学 synthetic ethology」と呼んでいる．生存や繁殖は生物にとって最も基本的な利益であると考えられるため，このアプローチによって人工的行為者に利害を持たせることは自然である．

ESGP を解決するためにはどのようなアプローチが有効であるかは経験的な問題である．私たち倫理学者は今後ロボット工学者と協力して人工道徳の研究を推進しながら，この問題に対する解決を試行錯誤していくべきであろう．なぜなら人工知能の発展が心の哲学に大きなインパクトを与えたように，人工道徳における今後の発展は倫理学に，そして私たちの道徳についての捉え方に大きなインパクトを与えることが予想されるからだ．人工知能の発展からの類推をすれば，将来的には道徳性の担い手は生物としての個人から，その個人を

[*10] 尾形，菅野 [23].

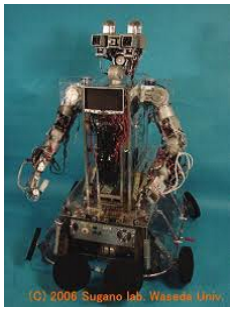

図 5　WAMOEBA

取り囲む人工的環境，社会的ネットワークへと拡張していくかもしれない．Google の開発者たちは Google を「人工知能」と呼んでいると聞く[*11]．現在，知性や知識は個人を超えて，テクノロジーと社会的ネットワークを含んだより大きなシステムに拡張しているものというイメージが広まりつつある[*12]．これからの道徳もまたそのような，集合的で複雑なものになっていくのかもしれない．

5　結び

本発表で私たちは，人工的行為者を道徳的行為者とみなすフロリディらの提案が「倫理的記号接地問題 (ESGP)」に直面することを見た．私たちはそれがどのようにして解決されるかということを考察し，その解決のためには，人工知能における記号接地問題へのマルチ・エージェント・ロボティクスに，ロボットに情動を持たせるアプローチや，進化的アプローチを組み合わせることが有用であるかもしれないということを論じた．しかし ESGP の解決にはどのアプローチが最適であるかということは経験的問題であり，倫理学者はロボット工学者と協力して，この方向の研究を進めていく必要があるだろう．なぜならば人工知能が心の哲学に与えたのと同じようなインパクトを，人工道徳が倫理学に与えてくれることが期待されるからである．

参考文献

[1] M. Anderson and S. Anderson. Robot be good. *Scientific American*, October, 2010.

[2] R. Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, (1):14–23, 1986.

[3] R. Brooks. **『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？』**. オーム社, 2006. 五味隆志訳. *Flesh and Machines: How Robots Will Change Us*, 2002.

[4] A. Clark. *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford University Press, New York, 2003.

[5] A. Clark and D. Chalmers. The extended mind. http://cogprints.org/320/1/extended.html.

[*11] 林晋先生談.

[*12] Cf. Clark [4], Clark and Chalmers [5].

[6] L. Floridi and J. W. Sanders. On the morality of artificial agents. *Minds and Machine*, 14, 2004.

[7] D. Gillies. *Artificial Intelligence and Scientific Method*. Oxford University Press, New York, 1999.

[8] P. Grim and T. Kokalis. Environmental variability and the emergence of meaning: simulational studies across imitation, genetic algorithms, and neural networks. In Loula, Gudwin, and Queiroz, editors, *Artificial Cognition Systems*, pages 284–325. Idea Group Publishing, Hershey, 2007.

[9] S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

[10] B. MacLennan. Synthetic ethology: An approach to the study of communication. Technical Report CS-90-104, Computer Science Department University of Tennessee, Knoxville, 1990.

[11] B. MacLennan. Synthetic ethology: A new tool for investigating animal cognition (extended version). Technical Report UT-CS-01-462, Computer Science Department University of Tennessee, Knoxville, 2001.

[12] B. MacLennan. Making meaning in computers: Synthetic ethology revisited. In Loula, Gudwin, and Queiroz, editors, *Artificial Cognition Systems*, pages 252–283. Idea Group Publishing, Hershey, 2007.

[13] J. R. Searle. Minds, brains and programs. *Behavioral and Brain Sciences*, 1:417–424, 1980.

[14] L. Steels. The symbol grounding problem has been solved. so what's next? In M. de Vega, editor, *Symbols and Embodiment: Debates on Meaning and Cognition*, chapter 12. Oxford University Press, Oxford, 2008.

[15] L. Steels and P. Vogt. Grounding adaptive language games in robotic agents. In C. Husbands and I. Harvey, editors, *Proceedings of the 4th European Conference on Artificial Life*. The MIT Press, 1997.

[16] M. Taddeo and L. Floridi. Solving the symbol grounding problem: A critical review of fifteen years of research. *Journal of Experimental and Theoretical Artificial Intelligence*, 17(4):419–445, 2005.

[17] A. M. Turing. Computing machinery and intelligence. In A. R. Anderson, editor, *Minds and Machines*, pages 4–30. Prentice Hall, 1964.

[18] P. Vogt. Language evolution and robotics: issues on symbol grounding and language acquisition. In Loula, Gudwin, and Queiroz, editors, *Artificial Cognition Systems*, pages 176–209. Idea Group Publishing, Hershey, 2007.

[19] W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, New York, 2009.

[20] T. Winograd and F. Flores. **『コンピュータと認知を理解する――人工知能の限界と新しい設計理念』**. 産業図書, 1986. 平賀譲訳. *Understanding Computers and Cognition*, 1986.

[21] 康一 古川, 知伸 尾崎, and 研 植野. **『帰納論理プログラミング』**. 共立出版, 2002.

[22] 淳 谷. 「認知力学系とロボティクス」. In **『岩波講座　ロボット学** *6*　**ロボットフロンティア』**, pages 127–155. 岩波書店, 2005.

[23] 哲也 尾形 and 重樹 菅野. 情動モデルを有する自律ロボット WAMOEBA-2 と人間との情緒交流. **日本機械学会論文集（*C* 編）**, 65(633):166–172, 1999.