「ロボットの感情」論争

—ポール・ジフ vs. J.J.C.スマート—

岡　本　慎　平

（広島大学助教）

1．はじめに

ロボットやA.I.（人工知能）が社会的に普及するにつれ、それらと人間のインタラクションが多様な倫理問題を引き起こすことが明白になってきた。こうした（そう呼んでよければ）ロボット倫理学の問題群には「ロボットに対する倫理」も含まれる。例えば、人間に非常によく似たシステムを構築しようとして「人工知能に苦痛を感じさせる」ことを目指す研究があり、誰の目から見ても苦痛を感じているとしか感じられない反応を出力する人工知能ができたと想定しよう。このとき、そこで苦痛を感じている人工知能を道徳的配慮の対象とみなし、動物実験と似た規制を必要とすることはありうるだろうか？　少なくとも「苦痛を感じるかどうか」を配慮の条件とみなすタイプの帰結主義者（つまり功利主義者）であれば、即座に「否」と答えることは控えるべきだろう。

さて、今から遡ること半世紀以上前、人工知能が研究分野として成立したばかりの1950年代末に話を移そう。当時、アラン・チューリングの提案した「チューリング・テスト」は人間と機械（ロボット）を区別する有益な基準になりうるか否かが、大きな哲学的関心を呼んでいた。そうした大小様々な論争の中に、「たとえロボットが人間と変わらない応答を示し、そこに真の知性を認めたとしても、依然としてロボットは『感情』を持てない」という主張をめぐる論争があった。1950年代という、今から考えれば極めて初期段階の人工知

能をめぐる論争であるものの、この論争を振り返ることは現代でも有益である。というのも、論争の発端となったポール・ジフの主張は、特定の技術的前提に基づいたものではなく、むしろ「ロボット」や「感情」といった言葉の意味に関する議論だからである。

2. 1950年代の人工知能研究

論争に入る前に、少し時代背景と主要な登場人物を紹介しておこう。この論争が起こった1950年代後半は、人工知能の歴史においてその「成立期」であるとか、最初の「黄金時代」だったと言われることが多い。そのようにみなされる契機を二点だけ紹介しておこう。

一点目は、数学者アラン・チューリングの論文「計算機械と知能」(1950年)の発表にある。チューリングはこの論文で「機械は考えることができるか？」という問いを考察し、被験者が、複数の対話相手(そのうち一人は計算機械で、他は人間である)と会話を進める中で、対話相手のうちどれが機械なのか識別できなければ、その計算機械は「考えうる」とみなしてかまわないという基準——いわゆるチューリング・テスト——を提案し、人工知能を工学的なリサーチプログラムの一つとして成立させた。

二点目は、1956年の夏に開催されたいわゆる「ダートマス会議」である。ジョン・マッカーシーの主導によって開催されたこの会議では、約2ヶ月にわたり、後に「人工知能」と呼ばれるようになる領域の研究集会が断続的に開かれた。もちろん、「人工知能」という言葉自体、この会議で「知能機械、特に知能的なコンピュータ・プログラムを構築するための科学と工学(The science and engineering of making intelligent machines, especially intelligent computer programs)」としてマッカーシーにより提案された言葉である(Gonsalves 2017: 229)。

会議の報告の中でもとりわけ有名なのは、アレン・ニューウェルとハーバート・サイモンら発表した「Logical Theorist」である。これは、数学の定理を証明することが出来、「人工知能」と呼びうる史上初のプログラムとみなされている。ニューウェルらに限らず、この会議の出席者たちがその後の人工知能研究をリードしていくことになった。

当初の人工知能研究は、数学の定理の証明など、人間の知的能力のなかでも「推論」や「記号操作」の側面に向けられたものだったが、いずれは他の側面についても機械による再現が可能だろうと期待されていた。ポール・ジフによる「ロボットの感情」論文が提出されたのはちょうどこの時期、1959年である。

ここでポール・ジフがどのような人物なのかを紹介しておこう。彼は、今日では人工知能やロボットに関する哲学を論じた人物とみなされることはほとんどない。彼の名を現在見ることが多いのは「芸術の哲学」、いわゆる美学である。例えば、「ロボットの感情」とほぼ同時期に(1958年)、彼は「芸術批評の理由」という論文を公表し、芸術において評価の理由となるものを考察している。こうした業績は現在でも少なからず参照されており、例えば数年前に『分析美学基礎論文集』(2015年)の一遍として翻訳されたことは記憶に新しい。

美学とロボットの哲学の間には、一見すると何の繋がりもないように見えるかもしれない。しかし、彼の「ロボットの感情」論文は、日常言語の意味論的分析に基づいて「ロボット」と「感情」という言葉が指す意味を明らかにしようとするものであり、美学の議論と無関係ではない(むしろ方法論としては同じものだとすら言える)。とはいえ、「ロボットの感情」に関する彼の議論がわかりやすいものかというと、そうでもない。むしろ、後に見ていく批判でも分かるように、彼の議論はいくつもの暗黙の前提に依拠したものであり、難解ですらある。

それでは、この論文で何が主張されたのかを見ていこう。

3.「ロボットの感情」

この論文における彼の主張を一言でまとめれば、「ロボットは感情を持ちえない(A Robot could not have feelings)」、あるいは「ロボットはロボットのような振る舞いをする(A Robot would behave like a robot)」だろう。このような結論に至った議論を確認していきたい。

この論文でジフが念頭に置いている「ロボット」は、当時の特定の技術的制約に基づいた具体的な人工知能やロボットではない。むしろ彼が考えているのは、外見の点でも、動作の点でも、発話の点でも、あらゆる点で人間と見分けがつかない、未来のロボットである。そして彼の議論は、「人間とまったく見分けがつかないロボットは」という主語に、「感情を持つ」という述語を帰属させることが適切かどうかという、意味論的な分析へと進んでいく。

ジフの議論によれば、「前提により、ロボットはメカニズムであり、有機物ではなく、生き物(living creatures)ではない」。そしてこの前提に基づけば、「壊れたロボット(broken-down robot)」は存在しうるが、「死んだロボット(dead one)」は存在しえないことになる(Ziff 1959: 64)。なぜなら、「死んだ」という形容詞は――「バッテリーが死んだ」といった慣用表現を除けば――生き物以外に当てはめることが不適切な言葉だからである。

ジフは、「死んだ○○」が生き物にしか当てはまらないのと同じことが、「○○は疲れを感じている(… feels tired)」のような表現にも当てはまると主張する。例えば、我々は特定の人間を指して「彼女は疲れを感じている」と主張することが出来る。しかし、「そのロボットは疲れを感じている(the robot feels tired)」と主張することは出来ない。なぜなら、前提により「ロボット」の指示対象は「メカニズム」であり、「生き物」ではないにもかかわらず、「○○は疲れを感じている」という述語は「生き物」にしか述定できない表現だからである。したがって、もし「ロボットは疲れを感じている」と言ってしまえば、

主語(「ロボット」)に当てはめることの出来ない述語(疲れを感じている)を当てはめていることになるため、矛盾となる。したがって、「ロボットは感情を持っている」という表現は意味をなさない、たとえ意味をなすとすれば単なる暗喩(メタファー)としてである、というのがジフの主張の要点である。

同じような問題は、感情以外の言葉でも当然発生する。例えば「行為」である。我々は人間の動作を様々な言葉を使って表現する。しかし、比喩ではなく文字通りに、行為を指す動詞を「ロボット」に当てはめることはできない。なぜなら、主語である「ロボット」は、その動詞を文字通りに当てはめてよい対象(人間)ではないからである。

> ロボットは計算する(calculate)かもしれないが、文字通りに推理する(reason)ことはない。おそらくロボットはものを取る(take things)だろうが、文字通りにそれを借りる(borrow them)ことはない。ロボットは人を殺す(kill)だろうが、文字通り殺害する(murder)ことはない。謝罪の声を出すかもしれないが、文字通り謝罪することはない。これらは、人間だけが実行できる行為である。そして前提により、ロボットは人間ではない。(Ziff 1959: 65)

ジフは様々な行為を挙げているが、それらを「文字通りに(literally)」当てはめてよいのは人間だけだと強調している。ジフによれば、「ロボットが疲れを感じている」という発言は、「石が疲れを感じている」とか「17という数字が疲れを感じている」とまったく同じように、言わばカテゴリー錯誤を犯しているのである。

また論文の後半では、チューリング・テストを突破して「確かに感情を持っている」かのように人々から判断されるロボットが作られたとしても、問題の

解決にならないとジフは主張する。ここでジフは、ロボットとよく似た例として「俳優の演技」を挙げる。

> あなたと私が、とある俳優の家にお邪魔したとしよう。彼は悲嘆に暮れる男の役を練習している。悲嘆に暮れる男となった彼は我々の存在に気付いていない。彼の演技は見事なものだ。彼が俳優であり役の練習をしていることを、私は知っているがあなたは知らない。あなたが「なぜ彼はあんなに悲しんでいるんだ？」と問うと、私は「彼は悲しんでないよ」と答える。するとあなたはこう言うだろう。「彼は悲嘆に暮れているぞ。彼を見ろよ！君がそんなことを言った根拠を見せてくれ！」もちろん、そこに見せるべきものなど何もない。(Ziff 1959: 66)

俳優の見事な演技によって、文脈を知らない人には「彼は文字通り悲しんでいる」と見えたとしても、その俳優を知っている人にとっては、その俳優が文字通りの悲しみを感じていないことは自明である。ロボットにも同じことが言える。ジフによれば、チューリング・テストによって分かるのは、「文脈を知らない人にそのロボットがどのように見えるのか」だけである。「対話相手がロボットだと知らない人」には感情を持っているように見えても、「対話相手がロボットだと知っている人」には依然として感情を持っているようには見えないだろう、とジフは言う。なぜなら、ロボットの心がプログラマに設定されたプログラムに添って作動するものである以上、ロボットが何かを意味するとしても「蓄音機のレコード」が再生する音が何かを意味すること以上のものは意味しえないからである(Ziff 1959: 68)。

したがってジフは、たとえチューリング・テストを突破して、だれの目から見ても人間とまったく同じように振る舞い、人間と同じように感情を持ってい

るように見えるロボットが出来上がったとしても、それが文字通りの感情を持つことはないと結論付ける。なぜなら、「○○は感情を持っている」という述語は、生き物だけに適用される言葉であり、ロボットに当てはめることはカテゴリー錯誤だからである。

4．スマートによる批判

ジフの論文は、当時の哲学者たちには主として批判的に受け止められた。例えばキース・ガンダーソンは、ジフの議論は苦痛に対する動物の反応を一種の「時計と同様のメカニズム」にすぎないと考え、動物が感情を持ちうるという主張を否定した17世紀のデカルトやラ・メトリの主張を約300年後にそっくりそのまま繰り返しただけにすぎないと論じている。デカルトやラ・メトリは、動物は人間とは異なり魂を持たないため、動物は「文字通りに」は感情を持つことはありえないと主張した。しかし現在では、そうした見解は動物に対する当時の無理解を象徴する意見として取り上げられることがほとんどである。ジフの主張も、ラ・メトリが動物に対して持っていた偏見と同じような偏見をロボットに向けているだけなのではないか。

もちろん、ジフの議論にそうした「保守的」な偏見を見て取ることは出来る。しかし本節では、ジフに反応した哲学者の中でも最初期に批判をおこなったJJ.C.スマートを取り上げよう。彼の批判は、ジフが暗黙に前提していたものを明らかにしてくれるからである。

まず、スマートについて簡単に紹介しておこう。彼はイギリスのケンブリッジで学者の一族の中に生まれ、オックスフォード大学で学んだ後、オーストラリアで活動した哲学者である。この論文の当時はアデレード大学に在籍していた。また、心の哲学では「心的対象は脳神経のメカニズムと同一である」という一種の物理主義(「心脳同一説」と呼ばれる)の擁護者として、倫理学では行

為功利主義の支持者として(バーナード・ウィリアムズとの共著『功利主義：賛成論と反対論』で、賛成側の論陣を張ったことは特に有名である)、現在でもよく知られている。さて、彼のジフへの指摘は、どちらかといえば前者、心の哲学に関連するものである。彼の批判では「心脳同一説」へのコミットがはっきりと示されているが、批判そのものはそのテーゼとは無関係である。

スマートはジフの論文がAnalysis誌に掲載された翌々月に、早くも同雑誌に「ロボットに対するジフ教授の見解について(Professor Ziff on Robots)」という短い文章を寄稿し、ジフの議論を三点にわたって批判した(ちなみに同号には弟のニニアン・スマートも同じようにジフを批判する論文を寄稿している)。一点目は、「生き物」と「ロボット」を区別する基準が明確ではない、という点である。おそらくジフは「ロボットはプログラムされたものであるため、生き物ではない」という前提を暗黙のうちに置いているが、これが適切な基準であるとは言い切れない。例えば、スマートが出す反例はなんと聖書である。仮に、『創世記』で描かれた人類誕生の物語が真実の歴史なのだとしたら、アダムとイヴは「ロボット」だということになる。なぜなら神学の教えと生物学によって判明した真理を組み合わせれば、神がアダムとイヴに、遺伝子や「遺伝情報を記録する機能を持ったDNA」という形で、心を含めた人間としての特徴を「プログラムした」と考えざるをえないからである。もちろん(聖書が真理を教えてくれているのなら)、我々はアダムとイヴの子孫かもしれないが、我々自身は神によってプログラムされたわけではないだろう。しかし当時フォン＝ノイマンが提唱していた自己複製メカニズムというアイデアに従えば、オリジナルの機械が子孫となる機械を生み出し、その子孫に「自然選択によって進化し、オリジナルの機械には無かった特徴と能力を発展させる」(Smart 1959: 117)ことは可能である。さらに、そうして生み出された機械が既存の生き物と同じ程度の複雑さを持っているのであれば、もはや生き物とロボットの違いは存在し

ないだろう、とスマートは主張する。

もちろん、このスマートの主張は聖書に加えて、「心脳同一説」というスマート自身の(異論の余地が多分にあると知られている)心の哲学についての主張を前提にしている。彼は「なぜなら私は、生き物は非常に複雑な心理的・化学的メカニズムに過ぎないという物理主義的テーゼを受け入れるにやぶさかではないからである」(Smart 1959: 117)と付け加えている。しかし、批判の要点はそこではない。問題は、少なくともジフは「生き物はどのような点でメカニズムと異なるのか」という点を不明瞭なままにしており、それを明示しないまま「メカニズムは感情を持てない」と主張しても論点先取になるという点にある。

二つ目の批判点は、たとえ「『○○は生き物である』という主張は『○○は人工物(artifact)ではない』を含意する」が真であったとしても、「『○○は感情を持つ』というという主張は『○○は生き物である』を含意する」という根拠にはならない、という点である。「『○○は生き物である』という主張は『○○は人工物(artifact)ではない』を含意する」という主張から論理的に出てくるのは、「『○○は人工物である』は『○○は生き物ではない』を含意する」という対偶だけである。ジフのように、そこからさらに、「『○○は人工物である』は『○○は感情を持ちえない』を含意する」と言うためには、「感情を持ちうるのは生き物だけである」という別の前提を追加しなければならない。しかしこれも、少なくとも心はメカニズムであると理解するスマートにとっては受け入れがたい主張である。そしてスマートにとって受け入れがたいということは、少なくとも論証抜きで前提してよい主張ではない、ということにもなる。スマートによれば、「人工物は感情を持ちえない」という主張は当時の技術的制約に過ぎないものであり、将来にわたってずっとそうであるとは限らない。

よって我々は、事実として、今の時点では、「これは感情を持っている」

ならば「これは生き物である」と演繹しても、問題になることはない。ところが、これが論理的包含関係である必要はない。おそらく将来、我々は反例を発見することになるだろう。(Smart 1959: 117-8)

結局、スマート自身は反例を見ることなくこの世を去った——2010年没——ことや、現在の我々でもそのようなものを見たことがないという点は差し引こう。重要なのは、これが論証抜きに自明な前提ではなく、論証を必要とするものだという点にある。

最後の批判点は、「感情を持っているように見える」と「感情を持っている」との間に大きな違いは存在しない、という点である。スマートは、結局ジフはロボットのプログラムを蓄音機と同じような決められた手順を踏むだけの機械とみなしているのではないか、と批判する。しかし、当時考えられていたアイデアだけに限定しても、人工知能はもっと複雑で、自分自身で目標を設定できるほど自律的なものが目指されていた。スマートが例に挙げるのは、チューリングの「計算機械と知能」で提示された学習機械のアイデアである。仮に、子供が自分の目的を発見し能力を身につけていくのと同じような仕方で、ロボットも学習が可能になったとしよう。その時、もはやそれが考えたり感じたりするという主張は、比喩ではなく文字通りの意味で、ロボットに当てはめても構わないはずだ、とスマートは主張する。

そのロボットは、会議に出席し、人間の哲学者がするのと全く同じように発達して、哲学者になることすらあるかもしれない。このことは、それが言わんとすることを意味している(it *meant* what it said)、と我々が言うべきではないのはなぜだろう？(Smart 1959:118)

少なくとも、ロボットは感情を持たないとするジフの主張は、「ロボット」と「生き物」の間の違いは重要な違いだという、本来証明すべき事柄を前提に置いてしまっている。

5. おわりに

スマート自身のコミットメントはさておき、ジフの議論が論点先取になっている、というスマートの指摘はおそらく正しい。しかしジフが暗黙のうちに置いてしまった前提は、論争から50年経った現在でも人々の「直観」として残り続けているように思われる。それは、デカルトやラ・メトリの時代から数百年経った現在でも、彼らと同じような主張によって――当時と比べて遥かに説得力が弱まっているとはいえ――「動物の感情」への配慮がたびたび論争の火種になることと無関係なものではない。

我々の直観は時代によって大きく変わる。倫理学がその後押しをすることも少なくない。しかし、変化の速度という点で言えば、決して楽観視できるものではない。おそらく将来、ロボットと人間の違いが――それこそ動物と人間の違いと同程度に――道徳的に重要な相違点ではないとみなされるほど人々の直観が変化するには、大きな時間がかかると思われる。

参考文献

Gonsalves, Tad, 2017 "The Summers and Winters of Artificial Intelligence" in Encyclopedia of Information Science and Technology, Fourth Edition

Gunderson, Keith, 1968 "Robots, Consciousness, and Programmed Behaviour" in the British Journal for the Philosophy of Science, vol. 19, No. 2, pp. 109-122.

Turing, Alan M., 1950 "Computing Machinery and Intelligence" Mind, vol. LIX, no. 236:433-60.

Smart, J.J.C., 1959, "Professor Ziff on Robots" Analysis vol. 19, issue 5: pp. 117-8.

Ziff, Paul, 1959, "The Feeling of Robots" Analysis vol. 19, issue 3: pp. 64-68.

"The Feeling of Robots" debate: Paul Ziff vs. J.J.C. Smart

Shimpei OKAMOTO

Assistant Professor, Hiroshima University

More than half a century ago, the philosopher and theorist of aesthetics Paul Ziff claimed that robots were inherently unable to have feelings. Although Ziff's argument elicited objections of various philosophers, it has received little attention in recent years.

The significant developments in contemporary artificial intelligence (AI) and robotics research since Ziff's writing could be argued to have resolved this controversy. However, Ziff's argument is not dependent on a specific technological premise. In addition, other writers have posed similar objections to the notion of robot ethics. Therefore, it may be valuable to reflect on the "The feeling of robots" debate.

In the current article, I first describe the relationship between AI and philosophy at the time of Ziff's writing, and introduce the protagonist of the debate. Second, I discuss the thesis that robots are unable to have feelings, as presented in Ziff's work. Third, I examine J. J. C. Smart's argument against Ziff's thesis.