社会と倫理　第28号　2013年　p.67−79

特　集　ロボット・社会・倫理

ロボット倫理学の基礎：責任とコントロール

佐々木　拓

近年のめざましい情報技術の発展により、ロボットの性能は格段の進歩を見せた。従来は人の手に頼らざるをえなかった、繊細だが単純な作業をこなすロボットは現在では実用化の段階にある。このような流れの中で、ロボットの振る舞いや学習の仕方が人間に類似するにつれて、ロボットの行動に対しても、人間同様に責任を帰属したいという気持ちが生じてもおかしくはない。また、1980年代にはすでに、コンピュータ自体に法的責任を帰属するための哲学的な議論が始まっている（例えば、医療現場でのコンピュータに対する責任帰属についてはスナッパーの議論(1) が好例である）。

しかし、このような発想には奇妙さを感じるのが現在では一般的だろう。とりわけ道徳的「責任」とは人間およびその行為に独特なものであり、人間以外の動物や、ましては器物であるロボットの行動に対して帰属されるものではないと考える人は多いように思われる。このような人は、法的責任の帰属がある種の取り決めとして可能だとしても、その根底には何らかの道徳的議論が必要だと考えるだろう。

本論文の目的は、ロボットのような非人間的対象への責任帰属を議論するための土台となるような責任理論とそれに関連する重要な概念を紹介することにある。責任論には、機械論的世界観と責任の両立/非両立を論じる枠組みがあり、その中で両立を認める理論の多くは〈非人間的対象への責任帰属〉という本論文のテーマと比較的相性がよい。本論では、このような立場の中から代表的で、本特集のテーマに寄与しうるような理論を解説することで、より根本的な議論のための材料を提供する。

さて、この目的のためには、まず「責任」という語の内実を確認する必要がある。また、議論を進めるにあたって、責任を見る視点を2つのレベルに区分することが有益である。そこで、次節では責任理論の紹介に先立って、本論文で扱う責任の種類の限定と責任のレベルの区別を説明する。それをふまえて、第2節以降でそれぞれの責任に対応した帰属理論を論じよう。

（1）Snapper 1998を参照。

1. 責任の種類とレベルの区別

一口に「責任」と言ってもその種類と内実は多様である。法的責任と道徳的責任とは区別される必要があるだろうし、「教師」や「親」といった役割に求められる責任と、特定の行為に事後的に割り当てられる刑罰や非難は区別されなければならない(2)。本稿のターゲットになるのは道徳的な事後的責任である（理論の解説を行う第2節以降では、断りがある場合を除いて「責任」の意味をこれに限定する）。すなわち、何らかの行為が行われた後に（例えば、嘘をついた）、その行為に対してなされる道徳的な賞賛・非難（この場合では、嘘をついたことに対する非難）が本稿で扱う責任の中心的なものだということである。

このような責任の概念にはいくつかの関連する概念が結びついており、両者を区別しておくことも以降の理解に役立つだろう。まず1つめは「行為者性」agencyである。行為者性とは、ある行為が〈特定の行為者のもの〉だという、いわば行為そのものの帰属を示す性質である。われわれは行為の帰属と責任の帰属とを区別して考えることが可能であるが（例えば、小学生の自分の子供が学校の窓ガラスを割った場合、「行為の責任は親である自分にある」と言うと同時に、「その行為をしたのは子供である」と言うことは一般的に受け入れ可能であろう）、多くの場合、ある行為の責任を特定の行為者に帰属するためにはその行為が〈当の行為者のもの〉であることが必要になる。換言するなら、ある行為（もしくは状況）が生じたことに誰かが（とりわけ道徳的に）責任を負うためには、その行為が何らかの仕方でその人に依拠していなければならない。この仕方については様々な見解があろうが、最もわれわれにとってなじみ深いのは、「コントロール」という依拠関係であろう。これが責任と区別されるべきもう1つの概念である。この概念を用いるなら、〈ある行為が行為者に依拠している〉ということは、〈行為者がその行為をコントロールできた〉ということであり、このコントロールゆえに行為者は当該行為に対する行為者性を帰属され、ひいては行為（およびその帰結）に対して責任を負う、ということになる。

さて、本節で重要なのは、このコントロールの見方には2つのレベルがあり、それぞれに対応した責任のレベルがあるということである。まず1つめは「局所的コントロール」とそれに対応する「局所的責任」である(3)。局所的コントロールとは、一般的な説明を与えるなら、〈特定の時点、特定の状況での行為に、行為者のもつ特定の能力もしくは行為の過程を反映させる

（2）英語のresponsibilityという言葉には、事後的責任としては刑罰や非難の他に報償や賞賛も含まれる（これらはまとめて「サンクション」と呼ばれることもある）。日本語の語感にはなじみにくいが、本稿でも事後的責任としては正と負の双方を意味するものとする。

（3）これらの用語はSmilansky 2000による。Fischer & Ravizza 1998, Kane 1996などでも類似の概念・用語が使用されるが、この概念と後に言及する究極的コントロールおよび責任を明確に対照的に論じたのはスミランスキーの功績の1つである。

能力〉と述べることができる。道徳的責任というテーマからはやや外れるが、理解の助けとして、スポーツ選手の功績を例に考えてみよう。この場合、〈ある競技者が「局所的コントロール」をもつ〉ということは、〈競技者がその時点で身につけている体力や精神力、技術をパフォーマンスに反映させうる状態にあった〉ということを意味する。そして、局所的コントロールを備えた状態でなされたパフォーマンスに対して「局所的責任」が帰属される。具体例として、オリンピック陸上競技で優勝が期待されている走幅跳びの選手を考えてみよう。この選手は大会で練習通り（もしくはそれ以上）のジャンプを跳べる能力をもつ限りで自らのジャンプに対して「局所的コントロール」を備え、ジャンプの結果に対して「局所的責任」が帰属される。すなわち、ジャンプが上手くいけば、例えば、金メダルの授与、世界記録の認定といった栄誉や観客の拍手による賞賛などに与り、ジャンプに失敗するなら、予選落ちの不名誉や観客や関係者の落胆といった非難を味わうことになる。

このことは、局所的コントロールの存在が局所的責任にとっての「行為者性」を担保していることを意味している。それは、ジャンプの結果にその選手のコントロールを外れた要因が強く働いた場合には、ある種の功績は選手に帰属されないことからもわかるだろう。例えば、追い風が一定の基準を超えた場合には、その時のジャンプが世界記録を達成しても、公式記録として認定されない。また、何らかの事情で局所的コントロールが失われた場合、行為者は（正であれ負であれ）行為の責任を免れる場合がある。例えば、大会直前にインフルエンザにかかり、当日はひどく体力と集中力を欠いて練習通りのジャンプが跳べなかったとしよう。この場合、ジャンプへのコントロールを失っているという理由で、コントロールを備えていたならば受けるはずの非難は選手には帰属されないかもしれない。このように、外的な状況による局所的コントロールの制限は責任免除の抗弁として機能する場合がある。

しかしながら、人によっては「体調管理も技術のうちだ」と言って、その選手を責めるかもしれない。また、このような（意地悪な）人は、選手が局所的コントロールを備えているとみなされる場合でさえ、悪い結果に対して「練習が足りない」「根性がない」などと言って非難するかもしれない。このような非難は一見局所的責任と同じような非難に聞こえるかもしれないが、異なる責任として区別されなければならない。というのも、このような非難においては、ジャンプに対する局所的コントロールだけでなく、〈局所的コントロールに対するコントロール〉を選手に要求しているからである。このような非難をする人は、単にジャンプに対する局所的コントロールだけでなく、局所的コントロールに反映されるべき身体的・精神的状態および能力、その他の技術に対するコントロールもまた選手が備えているはずだと想定していることだろう。そして、例えば事前の練習や競技会を通じて培われるこれらの状態・能力に対するコントロールがあってはじめて、大会当日のジャンプに対して功績が帰属されるのだと考えるかもしれない。このような、局所的コントロールと区別された、〈局所的コントロールに対するコントロール〉を「究極的コントロール」と呼ぼう。そして、〈局所的コントロールに対す

る責任〉を「究極的責任」と呼ぶことにしよう[4]。究極的責任の視点に立つならば、何らかの行為に対して本当の意味で行為者性を負うためには、行為者には究極的コントロールがなければならないし、究極的コントロールがあってはじめて、局所的な責任が行為者に帰属されることになる。

2. 局所的責任の帰属理論：ストローソンとフィッシャー＆ラヴィッツァ

2.1. ストローソンの反応的心情論

ロボットへの功績の帰属は比較的想像しやすいように思われる。というのも、われわれがオリンピック選手のパフォーマンスを評価するのと類似の仕方で、ロボットのパフォーマンスを評価することは日常的にありそうなことだからである。しかし、ロボットにある種の功績を帰属することが可能だとしても、ここで問題になるのは、ロボットに対して道徳的責任を帰属することができるかどうかであろう。そこで、とりわけ局所的責任の帰属を扱う責任論の中で道徳的責任の主体がどのように考えられているかを紹介しよう。

局所的責任を扱う論者の中で、P. ストローソンの議論は最も著名で影響力をもったものの1つと言うことができる。彼が「自由と怒り」（Strawson 1962）の中で提案した「反応的心情」[5] reactive attitudeとしての責任の捉え方は、この論文の主目的を超えて、多くの論者に影響を与えた[6]。さて、彼によれば、反応的心情とは「危害を受けた人やよいことをされた人々がもつ心情・反応であり、感謝、憤慨、許し、愛、そして傷心のような心情」である（op. cit.: 75[7]）。すなわち、他者との交流の中で、相手から何らかの危害や善行を受けた際に、その行為の与え手に対して受け手がいだく心情が反応的心情と呼ばれる[8]。このことからうかがえるのは、われわれのもつある心情が「道徳的」とみなされるには、他者との相互交流が必要だということである。成田はこのことを「対人的関係のネットワーク」と呼んでいる[9]。

しかしながら、この対人的関係は人間以外のものとの関係と区別されなければならない。その基準は何だろうか。ストローソンは「一般的に、われわれとこのような［対人的］関係にある人々に対しては、われわれはある程度の善意や尊敬を要求する」（op. cit.: 76[10]）と述べており、

（4）これらの定義はさしあたりのものである。厳密な考察は第3節で行われる。

（5）「反応的心情」という訳語は成田2004による。

（6）ストローソン説の影響力についてはMcKenna & Russell 2008を参照。

（7）ページ数は再版による。

（8）反応的心情はストローソンが「対人的心情」と呼ぶものの一種であり、これには他に立場交換的心情と自己応答的心情が含まれる。Strawson 1962: 84−5参照。

（9）成田2004第二章参照。

（10）［　］は筆者による補足。以下の引用でも同様。

反応的心情において重要なのは「他人の行為（中略）に反映されているわれわれへの心情が、善意、愛情、尊重という態度なのか、それとも軽べつ、無視、悪意なのか」だとしている（ibid.）。すなわち、対人的関係のネットワークとは、互いに一定程度の善意や尊敬を要求しあう者同士が形成するネットワークのことだと言える。そして、行為に反映されるべき善意や尊敬の要求の程度に応じて、様々な種類の反応的心情が生じるというのが、ストローソンの反応的心情論という考え方なのである。

対人的関係のネットワークに属することが道徳的責任の必要条件だとするならば、ネットワークの成員となるための資格とは何かという問いが生じるかもしれない。これに対してストローソンの議論から読み取ることができるのは「一定の善意や尊敬の要求に応えることができる」ぐらいが限度である。そこで、この問題をより深く考察するためにフィッシャーとラヴィッツァの理論に当たることにする。

2.2. フィッシャーとラヴィッツァの理由反応性説

フィッシャーとラヴッツァが著書『責任とコントロール』[11]で展開した責任論は1990年代責任論の1つの到達点だったと言って過言ではないだろう。彼らの最も重要な貢献の1つは、責任論に大きな転換をもたらしたことである。それは、彼ら以前の学説では責任帰属の条件が行為の構成要素（例えば、「熟慮の上で行為する」「自由意志から行為する」など）に求められていたのに対して、彼らは行為者の能力に焦点を当て、行為者の能力と行為との関係に帰属の条件を求めた点にある。このため、彼らの考察は自然に〈適切な道徳的責任主体〉の条件に向かうため、彼らの理説は先の問いに関する、より深化された議論と読むことができる[12]。

さて、彼らによれば、ある行為についての道徳的責任を行為者に帰属できるのは、行為者が行為に対して「誘導的コントロール」guidance controlを備えている場合に限られる（RC: 33, 170）。そして行為者が行為に対する誘導的コントロールをもつということは次の2つの条件によって説明される。（1）適度な理由反応性reason-responsibilityをもつ行為者の身体的・精神的メカニズムから行為が生じており、かつ（2）そのメカニズムが行為者自身のものである（RC: 170, 207）。ここで、誘導的コントロールは局所的コントロールの一種であるため、彼らの考えは局所的責任帰属の理論と解することができる[13]。

（11） Fischer & Ravizza 1998。以下、RCと略記する。

（12） フィッシャーとラヴィッツァ自身もこの連続性を認めている。RC: 5–8参照。

（13） フィッシャーとラヴィッツァはコントロールのレベルについて、「誘導的コントロール」と「統制的コントロール」regulative controlという用語を使用する。これらのうち後者は究極的コントロールの一種だと解することができる。RC: 31参照。

2.2.1. 適度な理由反応性

理由反応性という概念は、現代責任論では責任能力の中核の1つとみなされているが、この概念を詳細に分析したこともまた、フィッシャーとラヴィッツァの功績の1つに数えることができる。さて、彼らによれば、(1) の条件は2つの部分をもっている。1つは行為者が「適度な理由反応性」のあるメカニズムを備えていることであり（1a)、もう1つはそのメカニズムから逸脱的でない仕方で行為が生じていることである（1b)。

（1a）で言及される理由反応性とは、〈ある行動をする理由がある場合にその行動をする〉という性質である[14]。そして、この理由反応性が適度であるとは、理由反応性が強くも弱くもないことを意味する。理由反応性が強いとは、〈ある状況で、ある行動をする十分な理由がある場合に、行為者は必ずその理由を十分と認識し、その行動を選択する〉ということである（RC: 41)。この条件は帰責の条件としては強すぎる。というのも、この条件の下では、十分な理由がある場合にその理由を認識しながらも行動しないというだけで、行為者は帰責の条件としての理由反応性を失ってしまうからである。対して、理由反応性が弱い場合には、〈ある行動をとる理由がある場合に、行為者がその行動をとる反実仮想的なシナリオをいくつか想像できる〉だけでよい（RC: 44–5)。フィッシャーとラヴィッツァの例を借りるなら、ある女性が仕事の締め切りが迫っているにもかかわらず、バスケットボールの試合を見に行きたいと思っていたとする。彼女は意志が弱いので、その試合を見に行ってしまった。しかし、ここで仮に試合のチケットが1000ドルだったとしたら、彼女は試合には行かず、家で原稿を書いただろう。これに類似した反実仮想的なシナリオが1つでも考えられるのであれば、彼女には弱い理由反応性があるということになる（RC: 45)。

弱い理由反応性は道徳的責任の必要条件ではあるが（RC: 45, 81)、十分ではない。そこで適度な理由反応性が要請されるわけだが、これを説明するには理由反応性をさらに2つの部分に区別する必要がある。1つは理由を理解し、評価する能力である「理由受容性」acceptivity to reasonsと、もう1つは理由に対応した行為を選択・実行する能力である「理由対応性」reactivity to reasonsである。理由反応性が適度であるためには、理由受容性に合理的な規則性と道徳的理由の認識が必要であり、また理由対応性に最低限1つの理由を行動に移行させるだけの能力が要請される。これらの要請を見て行く過程で、なぜ弱い理由反応性が帰責条件としては緩すぎるのかも理解されるだろう。

理由受容性に合理的な規則性が必要な理由は次の事例の奇妙さによって理解されるだろう。締め切りを破ってバスケットボールの試合に行く女性の例を改変しよう。チケットが1000ドルするという理由が試合に行かない十分な理由であることは変わらない。しかし、ここでチケットが2000ドル、3000ドル、4000ドルであることが彼女にとって十分な理由にならないと

（14）フィッシャーとラヴィッツァにならい、「行動」という語で本稿では作為と不作為の両方を意味することにする。RC: 7参照。

したらどうだろう。ここでは〈チケットが1000ドル〉の他、彼女が試合に行かないシナリオがいくつかあるので、彼女は弱い理由反応性を備えている。しかし、彼女が責任主体として適切だと言えるだろうか。この事例に奇妙さを覚える人は、例えば〈チケットが1000ドルである〉ことを試合に行かない十分な理由とする人は、〈チケットが1000ドル以上のいずれの値段であっても試合に行かない〉というように、十分な理由に理解可能な規則性がなければならないことを認めるだろう。これが、理由反応性が適度でなければならない1つめの理由である。

理由受容性に関するもう1つの条件は、受容される理由に道徳的理由が含まれていることである。例えば、「利己的な理由がなければ約束を守らないような子供やサイコパス」は、弱い理由反応性をもつとはいえ、適切な意味で道徳的な責任主体とは呼べない（RC: 76）。先の例を用いるなら、バスケットボールの試合を見に行こうとしている女性には、例えば「約束を守ることは道徳的な義務だ」などといった道徳に関わる理由が認識されていなければならないということである。ここで「道徳的」の意味が要求されるかもしれないが、フィッシャーとラヴィッツァは多くの説明をしていない。言及されているのは、道徳的理由は自愛の思慮から生じる理由（自己利益的理由）とは区別されるという点と、道徳的理由には共同体の存在が前提されているという点くらいである（RC: 76–7）。

では、理由対応性についてはどのような条件づけがなされているのだろうか。ある行動への理由を認識するものの、それにまったく動機づけられない人は帰責の適切な条件を欠いているように見える。また、認識された理由のすべてに対する対応性を要求するのは過大であろう。この問題に対するフィッシャーとラヴッツァの答えは、「ある行動について、われわれが最低限1つの理由に対応できるなら、それはその行動に関するあらゆる理由に対応できることを意味する」というものである（RC: 74）。したがって、先の例では、「編集者から催促の電話があったら試合には行かない」という理由への対応性が彼女にあると認められるなら、それは同時に「約束を守る」という道徳的な理由への対応性もまた彼女に備わっていることになる。理由受容性は理解可能な形で規則的であり、かつ道徳的理由を含むものでなければならないが、対応性に関しては最低限1つの理由に対応できさえすればよい。さらに、弱い理由反応性の議論をふまえるなら、理由への対応は反実仮想的な可能世界の想像ができるという非常に弱いもので構わない。

理由反応性の議論について注意が必要なのは、適度な理由反応性は行為者のもつメカニズムに備わるものだという点である。ここでメカニズムが意味するのは、行動の産出に関わる身体の物理的構造および心理的能力と傾向性である。例えば、今私の右手に麻痺や物理的拘束がなく、かつ昨日の記憶を思い出せるだけの脳神経系をもっていれば、それは私が昨日の出来事を右手で日記に書くという行動の身体上のメカニズムを備えているということである。また、心理的なメカニズムとしてすぐ思い至るのは、行動の理由を受容するための熟慮の能力であるが、癖や習慣、本能といった無意識的な傾向性もここに含まれる（ibid.: 85–6, 215–6）。したがって、〈行為者のメカニズムに適度な理由反応性が備わっている〉とは、熟慮などの心理的能力によっ

て行動のいくつかの理由に気付くことができ、癖や習慣といった無意識的傾向性を前提にしても、最低限1つの理由に従って行動を実行できる身体的メカニズムがあるということになる。

ここまでの議論が本論文のテーマに親和的なのは、責任の帰属の条件が行為の構成要素ではなく、メカニズムと行為の関係におかれている点である。フィッシャーとラヴィッツァの理論では、道徳的に責任ある行為は、必ずしも熟慮というプロセスを必要としないし、意識的に生み出される必要もない。すなわち、精神的行為プロセスというものを必要としない。重要なのは、行為者が現状で備えるメカニズムを前提した時に、行動に関する（道徳的な理由を含めた）いくつかの理由にアクセスでき、それが実行に移される可能世界が想像可能だということにある(15)。

そして、このようなメカニズムが（1b）を満たすなら、理由反応性に関する条件は満たされることになる。（1b）とは〈行為者のメカニズムから逸脱的でない仕方で行為が生じていること〉というものであった。彼らは「逸脱的な仕方」を具体例で説明する。例えば、強力な洗脳や、催眠、依存性の強い薬物、脳の直接的操作、脳の損傷、精神疾患、脅迫などによって特定の行動が生み出される場合に、行為者は誘導的コントロールを失うことがある（RC: 35–6）。また、外的強制（身体の拘束や監禁、押されるといった物理的強制など）もこのような事例として考えられる。今言及した要素は「責任を台無しにする要因」responsibility-undermining factorsとして、逸脱的な行為の導出の範型とされる（RC: 36）。しかし、なぜこれらの要因が「逸脱的」とみなされるのか。彼らによれば、それはこれらの要因が理由反応性を失わせるからである（RC: 37）。例えば、特定の行動を指示する強力な洗脳や催眠は、現実に行われた行動以外の行動への理由受容性を行為者から奪うかもしれない。また、脅迫や物理的強制は理由受容性を阻害しないかもしれないが、理由対応性を奪うかもしれない。これらの要因が作用している場合、理由の存在は行為者の行動に何の影響も与えない。ゆえに、理由反応性を備えた行為者のメカニズムから自然に生じた行為とはみなされないのである。

以上の点をふまえるなら、ロボットであっても、道徳的理由を含めた複数の理由にアクセスでき、かつ状況の違いによっては実際に行った行動以外の行動をとるケースが1つでも想像可能なメカニズムを備えており、その行動が外部からの介入がない状態で生じている場合には、そのロボットの行動は条件（1）を満たしていると言うことができる。

2.2.2. メカニズムを所有する

条件（2）に移ろう。これは、ある行為者が道徳的責任主体であるためには、行為者のメカニズムが〈行為者自身のもの〉でなければならない、という条件であった。これは行為者のメカニズムに「責任を引き受ける態度」taking responsibilityが備わっていなければならないこと

（15）ここで、理由反応性が可能世界の想像で十分だという考えには、彼らの機械論的な世界像が反映されている。RC: 51–4参照。

を意味する（RC: 207）。そしてこの態度は次の3つの信念（もしくはそれに基づいた傾向性）の連言から説明される。（2a）自らを世界に対する因果的効力をもつ行為者だとみなすこと、（2b）自らを道徳的賞賛・非難の対象だとみなすこと、（2c）条件（2a）と（2b）を適切な証拠に基づいて形成すること、の3つである（RC: 210–3, 村上2010：211）。

この態度およびそれを構成する信念や傾向性は、子供の道徳教育が例にあげられているように、自らの意志によって引き受けるものというよりは、むしろ教育など環境によって形成されるものである（RC: 217–8）。自分のしたことについて誉められたり怒られたりすることで、子供は自分の行動が何かの帰結を生み出すことを知り、また、場合によっては自分の行動のためにある種の賞賛や非難を受けることを学ぶ。そして自らの体験を通じてこれらの信念を身につけることで（2c）が満たされる。このことが本論文のテーマにとって好都合なのは、熟慮や自己反省といった意図的活動がこの態度の獲得には必ずしも必要ないとされている点である（RC: 214）。これは帰責条件の焦点を能力にシフトさせたフィッシャーとラヴィッツァにとっては当然のことである。彼らにとって重要なのはメカニズムであって、責任の引き受けに関する条件は（2a）と（2b）の信念が適切な仕方でメカニズムに備わっているだけで十分満たされる。ということは、「自らを道徳的賞賛・非難の対象だとみなす」という（2b）の条件は、道徳的理由を認識し、非難や賞賛を受けた際にはそれに対応した行動をとったり、以後の自身の行動を変化させたりする傾向性がメカニズムに備わっていることで十分だと言える。そして信念を傾向性の観点から捉えることは、（2a）と（2b）の条件をロボットが満たすことをいっそう容易にするだろう。ただ（2c）をロボットが満たすことが一体何を意味するのかには一考が必要と思われる。

メカニズムの点から見ると、局所的責任をロボットに帰属するハードルは比較的低いように思われる。とはいえ、われわれ人間については、何らかの行動に対して責任を負うには局所的コントロールの存在だけでは不十分で、コントロールを背後で支える能力やメカニズムへのコントロールが必要なのだ、と考える人はいるだろうし、ロボットの場合にはこの傾向が一層強いように感じられる。この種の人々は責任を究極的視点から捉えており、責任の帰属のために究極的コントロールを要請している。

3. 究極的責任の帰属理論：スミランスキーの幻想主義

行動に対する究極的コントロールとは、局所的コントロールに対するコントロールだと先に述べた。となると、究極的コントロールとは、局所的コントロールに反映されるべきメカニズムに対するコントロールに他ならない。これには、特定の身体的能力や構造、そして特定の心的能力や傾向性（熟慮の能力や性格・価値観といった動機づけに関わる傾向性、癖や習慣などの無意識的傾向性など）を身につけることに関するコントロールが含まれる。先の走幅跳びの選手の例を思い出してほしい。究極的視点から見るならば、その選手がオリンピックでのジャ

ンプの功績を評価されるのは、ジャンプを成功させたという結果のためだけではない。技術や精神力、体力の修練を通じて、そのジャンプのためのメカニズムを形成したという過程が存在してこそ、本当の意味でジャンプの功績が評価されるのである。そして、この点で、例えば走幅跳びの世界記録を出せるよう設計されたロボットと、練習を積み重ねて世界記録を樹立した人間の競技者との功績には違いがあるのだと主張されるかもしれない。この筋の主張をとるなら、（功績を含めた）個別行動への責任の帰属は、行動を生み出した自分を自分の手で形成してはじめて正当化される。

責任論の一部の立場はこの主張を支持する。例えば、この分野で多くのアンソロジーの編集を務めているR. ケインは、「自己形成行為」という将来の性格や動機づけの傾向性を形成する行為があってはじめて人は道徳的責任主体たりうると考える(16)。ケインの考えを適用するなら、たとえあるロボットが自己学習プログラムを備えていたとしても、その学習プログラム自体を自分で作ることはできない。それゆえにロボットには本当の意味では行為者性が帰属できず、結果道徳的責任主体とは認められない、ということになるだろう。つまり、ロボットは自らの行動への究極的なコントロールをもたないがために、行動の責任を引き受けることができないというわけである。

この点は一見して決定的のように思われる。しかしながら、コントロール概念を介した局所的責任と究極的責任の依存関係はそれほど自明ではない。この依存関係を認めず、局所的コントロールさえあれば局所的責任は認められる（そして究極的責任というものは存在しない）と考える立場もある(17)。この依存関係の是非は重要な問題ではあるが、本論文の紙幅で扱うには大きすぎる問題である。そこで、以下ではこの依存関係を認める責任理論の中からスミランスキーの理論を幾分の再構成を加えつつ紹介しよう。

スミランスキーの議論にはある奇妙さが存在する。彼は〈帰責には究極的コントロールが必要だ〉という前提に立った上で、〈われわれは究極的コントロールをもちえない〉と主張する。通常、この主張は責任への懐疑論に行き着くのだが、ここで彼は責任の実在と重要性をも強調する。本稿では、彼がこの（奇妙な）一連の主張に至る過程を理解するために、彼の展開する2つの議論を解説しよう。その1つは究極的不正義の議論であり、もう1つは究極的不正義を解消するための幻想主義という考えである。

さて、ここまで究極的コントロールは「局所的コントロールに対するコントロール」であると述べてきたが、これは正確ではない。ケインやスミランスキーの意見に従うなら、今述べた「コントロールが本当の意味で行為者自身のものでなければならない」という主張が、究極的コントロールの必要性を訴える議論には含まれている。そして、この議論の背景には、責任帰属には「行為者性の究極的創造」という考えがある。つまり、責任を問われている行動やメカニズ

（16）Kane 1996, 2002. ケインの主張については佐々木2005も参照。

（17）ストローソンおよび、フィッシャーとラヴッツァの理論はこの立場の1つと考えられる。

ムの生成過程の中に、自らの手による行為者性があってはじめてそれらの責任が正当化される、という訳である。〈コントロールに対するコントロール〉を考える際、この「行為者性の究極的創造」はその後のあらゆる行動とメカニズムに行為者性とコントロール性を与える源となる。しかしながら、現実的に考えればそのような行為者性など存在しない。それは以下の例で示される無限背進の議論によって明らかになるだろう。

先の競技者の例で考えてみよう。オリンピックでのジャンプに対するコントロールに対してさらなるコントロールを求めるということは、それに必要な技術を身につけることに関するコントロールがこれに当たる。この技術は多くの場合練習（および事前の競技会）によって身につけると考えられるので、これを〈練習（という行動群）に関するコントロール〉と呼んで差し支えないだろう。これはある種の局所的コントロールである。したがって、ここには練習に反映されるべきメカニズムがあり、それが行動としての練習に反映されるという構造が存在する。しかし、競技会におけるジャンプへの功績帰属のために、ジャンプへのコントロール（C_1）とメカニズム（M_1）だけでなく、練習に関するコントロール（C_2）をも要請するのであれば、練習へ反映されるメカニズム（M_2）形成に対してさらなるコントロール（C_3）とメカニズム（M_3）を要求することには問題がない。またこのようなさかのぼりをここで打ち切るのは恣意的であろう。したがって、M_3の形成に対してC_4とM_4が要請され、M_4の形成に対してC_5とM_5が要求されなければならない。このさかのぼりを続けて行くと最終的には産まれた時に与えられたメカニズムに行き着くが、その間にある無数のメカニズム形成に関してすら行為者自身がコントロールできない要素（すなわち行為者のもつメカニズム外からの影響）が数多く存在することがすぐに判明する。例えばコーチの価値観や指導方針に関して、選手はコントロールをもたないかもしれない。また、練習に必要な体力やモチベーションの形成に関しても、さかのぼりを続ける末には偶然的な要因や、生得的な要因に行き着くかもしれない。このような議論を断ち切るために、先行する要因から一切影響を受けない行動を選択できる能力を仮に想定しうるとしても、事前の自分の性格や身体構造との関係を一切無視した決断が果たして自分の決断であり、自分のコントロール下にあると言えるかは、はなはだ疑問である。結局のところ、われわれは純粋な意味でのコントロールや行為者性はもちえないのであり、さかのぼりを続ければ続けるほど、自分のメカニズムに対するコントロールと行為者性の不在が明らかになるという結果になる。したがってわれわれは究極的コントロールに道徳的責任帰属を基礎づける限り、あらゆる責任の帰属はある種の不正義となってしまう。これが究極的不正義という問題である(18)。

この究極的不正義に対して、スミランスキーは責任帰属という実践に対するわれわれのコミットメントの強さを対比させる。先の議論が正しいならわれわれは究極的不正義を認めざる

（18）スミランスキーはこの議論についてはG. Strawsonの議論を援用する。Strawson 1994を参照。また、Smilansky 2002: 491、佐々木 2010：96も参照。

をえないにも関わらず、責任帰属という実践を放棄できない。この一見矛盾した態度を説明するために彼は「幻想」という概念を導入する。これが「幻想主義」である。彼によれば、われわれは責任帰属という実践を維持するために、「（例えばケインの自己形成行為のような）究極的コントロールをもっている」という誤った信念を抱いている。そして、この種の偽なる信念を抱いているという事実によって、なぜわれわれが究極的不正義の議論を前にしても責任帰属実践を変わらず維持し続けられるのかが説明されるのである（Smilansky 2002: 498–9, 佐々木 2010：98）。言い換えるなら、責任の帰属のために、われわれは実質的な（形而上学的な）究極的コントロールを必要としない。むしろ、「究極的コントロールをもっている」という認知的な感覚をもつことで、コントロールのさかのぼりを（恣意的に）中断することが重要である。そして、このような認識に加え、他人の行動を同様の仕方で見る傾向性があれば、その人は十分責任主体としてみなされるのである。

注意すべきことに、幻想主義は形而上学的な説明理論であって、規範的理論ではない。すなわち、われわれに欺瞞的に生きることや、無理矢理ある種の幻想を抱く責務が生じるということはこの理論には含意されていない（Smilansky 2002: 497）。幻想はむしろ、自覚のないままにそれを抱き続けるよう動機づけられているようなものなのである（Smilansky 2000: 146–7）。ストローソン風に言うなら、責任に関連する幻想は反応的心情としての責任と同様に、われわれの本性に「徹底的なほど深く根ざしている」（Strawson 1962: 68）ために、本性上容易に捨て去ることができないものだと言えよう。

スミランスキーの議論は非常にラディカルであり、批判も多い。しかし、彼の理論が責任論の中で一定の位置を占めていることも事実である[19]。本論文の目的としては、「人間の行為に関してさえ究極的なコントロールは求めえない」という主張を、ある程度確立された責任論の中に確認できれば十分である。

おわりに

前2節での議論をふまえるなら、善意と尊敬の反映と理解できる行動が可能な、適度な理由反応性をもったメカニズムが備わっている限りで、（100％のとは言えないものの）ある種の局所的責任や功績についてはロボットに帰属可能のように思われる。また、第3節の議論をふまえるなら、本当の意味での究極的なコントロールを要求するのは人間にすら不可能で、コントロールと行為者性のさかのぼりをどこかで打ち切るような幻想が必要である。そしてさかのぼりを打ち切ることで局所的な責任が担保されるのなら、同じ理屈がロボットの局所的責任にも適用される余地が生まれるだろう。

（19）彼の理論は*The Oxford Handbook of Free Will*においてハード・デターミニズムという立場の有力な理論として紹介されている。

しかしここで、「そもそもロボットに責任を帰属するとはどのようなことなのか」という根本的な問いが発せられるかもしれない。これは非常に重要である一方で、やはり本論文で論じるには大きすぎる問いである。とはいえ、先人から何らかの示唆を得ることはできる。例えば、冒頭で触れたスナッパーはコンピュータへの責任帰属について、「[人間は] 誰も責任を問われない」という選択肢を無視してはならないと警告している（Snapper 1998邦訳：68）。この示唆は大きい。例えば、局所的責任を行動に対する責任とメカニズムに対する責任とに区別し、前者をロボットに帰属することで、例えばロボットを使用する人間からその分の責任を免除する（そして、メカニズムに対する責任はメーカーや設計者に帰属される）という選択肢も考えられる。

このような責任概念そのものに関わる問題も含め、ロボットへの責任帰属についてはまだまだ論じなければならないことは山積している。しかし、必要な議論の多くは人間に対する責任帰属理論が共通して答える必要のあるものかもしれない。

参考文献

Fischer, J. M. & Ravizza, M. 1998. *Responsibility and control: a theory of moral responsibility*. Cambridge University Press, (RC)

Kane, R. 1996. *The Significance of Free Will*. Oxford University Press.

―― 2002. New Directions for an Ancient Problem. *Free Will*, R. Kane ed. Blackwell Publishers Ltd.: 222−248

McKenna, M. & Russell, P. eds. 2008. *Free Will and Reactive Attitudes: Perspectives on P. F. Strawson's "Freedom and Resentment"*. Ashgate Publishing Limited.

Smilansky, S. 2000. *Free Will and Illusion*. Oxford University Press

―― 2002. Free Will, Fundamental Dualism, and The Centrality of Illusion. *The Oxford Handbook of Free Will*, 2nd edition, R. Kane ed. Oxford Universitiy Press: 487−505

Snapper, J. 1998. Responsibility for computer-based decisions in health care. *Ethics, Computing, and Medicine: Informatics and the Transformation of Health Care*, K. W. Goodman ed. Cambridge University Press: 43−56, 邦訳「医療におけるコンピュータに基づいた決定に対する責任」（佐々木拓訳）、『医療IT化と生命倫理：情報ネットワーク社会における医療現場の変容』、世界思想社、2009：65−83

Strawson, G. 1994. Possibility of Moral Responsibility. *Philosophical Studies*, 75: 5−24

Strawson, P. 1962. Freedom and Resentment. *Proceedings of the British Academy*, Vol. 48: 1−25, (rpr. *Free Will*, 2nd. edition, Gary Watson ed. Oxford University Press, 2003: 72−93)、邦訳「自由と怒り」（法野谷俊哉訳）、『自由と行為の哲学』（門脇俊介・野矢茂樹監訳）、春秋社、2010：31−80

佐々木拓. 2005.「生き方が責任を作る：『もうひとつの可能性』再考」、『実践哲学研究』、28：21−44

―― 2010.「自由意志の非認知主義的解釈の可能性―スミランスキーの幻想主義とその補完―」、『倫理学研究』、40：93−104

成田和信. 2004.『責任と自由』、勁草書房

村上友一. 2010.「行為者性と道徳的責任―フィッシャーとラヴィッツァの責任論―」、『倫理学年報』、59：203−216