ソーシャルロボットの倫理における仮想的行為者性概念の可能性と限界
The possibilities and limitations of virtual moral agency in ethics of social robots

水上拓哉

Abstract

This paper addresses the problems of the moral status of social robots by examining Coeckelbergh's concepts of "virtual moral agency" and "virtual moral responsibility." We first briefly summarize the discussion of morality in robot ethics and then present Coeckelbergh's approach. Coeckelbergh's "relational turn" is useful in analyzing social robots because it is impossible to understand the moral competence of social robots by only analyzing the robots themselves. However, the concept of "virtual moral agency" creates conceptual confusion in the ascription of moral responsibility.

(1) 研究テーマ

ロボット倫理学においてはしばしばロボットの道徳的身分、すなわちロボットは道徳的行為者であるか否かという問題に焦点が当てられてきた。本論文では、その中でも社会的役割を担うことが期待されるソーシャルロボットに対象を限定し、その道徳的身分を考える上で M. Coeckelbergh が提案する「仮想的道徳的行為者性（virtual moral agency)」という概念がもたらす貢献とその限界について批判的に検討する。

(2) 研究の背景・先行研究

2.1. ロボット倫理学における行為者性をめぐる混乱

ロボットの引き起こす倫理的問題について取り組む研究領域にロボット倫理学(Lin et al. 2014, 2017)がある。これまでのロボット倫理学において特に重点的に論じられてきたのは「ロボットは道徳的行為者（moral agency）たりうるか」という問題、すなわちロボットの道徳的身分（moral status）についての議論であった。一般的に私たちが道徳的責任を帰属させるのは、何らかの観点において十分な道徳的行為者性をもち、その行為者のすることやしたことの正しさを問えるような存在だけである。だから、ロボットの道徳

的価値や責任について考察する上でもその道徳的身分が問われなければならない、というのが基本的な構図である(Coeckelbergh 2009a, p.218)。

ロボットの道徳的行為者性について、これまでどのような論争が展開されてきたかについては Gunkel(2012)や Sullins(2006)が詳しい。基本的には、ロボットの（将来的なものも含む）能力が人間のそれと比較検討され、ロボットが何らかの水準で道徳的行為者といえるかどうかが議論されてきた。あるものが行為者であるための必要条件のラインナップについて統一的な見解は存在しないが、Himma(2009)が「標準的見解」として提出しているように、少なくとも自由意志や意図の有無が問題になることが多い。

2.2. ロボットの道徳的能力――Coeckelbergh の関係論的転回

初期のロボット倫理学においては、ロボットが何らかの意味で自由であり、また意図をもっているかが議論の中心であった。このとき問題となってきたのは、ロボットが実際にこれらの性質を有しているかどうかである。自由意志や意図といった心的状態を何らかの意味で拡張し、広い意味でこれらの性質をもっていると主張する立場もあるが、そのような立場でもこの「実在指向」は同様である。この考えに基づき、初期のロボット倫理学においてはロボット側に分析の焦点が当てられ、ロボットが本当に何を考えているのか、何らかの意味で心的状態を実際にもっているのかが議論されたきたのである。

一方、この実在への指向を拒絶し、アピアランスに基づいたロボット倫理学を展開しているのが技術哲学者の M. Coeckelbergh である。Coeckelbergh は、ロボットが（その実在とは独立的に）倫理的場面において人間・非人間の境界を容易に乗り越える存在だと考える。つまり、彼にとってロボットは、想像力をかきたてる、感傷的でアピアランスドリブンな存在なのだ(Coeckelbergh 2009a, p.218)。現在のロボットが、人間と同じ意味で心をもった存在だということは難しいが、私たちはしばしば、そういったロボットがあたかも心を有しているかのように振る舞うことがある。そして、それが人間のロボットに対する生き方なのであれば、私たちの倫理的配慮はロボットではなくむしろ人間、すなわち「私たち」が何を考え、感じ、想像するのかに向けられるべきだ、というわけだ(Coeckelbergh 2009a, p.219)。

2.3. 仮想的な道徳的行為者性

Coeckelbergh の提案するロボット倫理学の方法論的転回は、いわば人間中心的かつ関係論的な転回である。では、このアプローチにおいてロボットの道徳的行為者性や道徳的責任の問題はどのように再構成されるのだろうか。

Coeckelbergh は別の論文で、ロボットの道徳的身分を理解するために「仮想的な（virtual）」道徳的行為者性という概念を提案している(Coeckelbergh 2009b)。

ここで Coeckelbergh が問題視するのは、ロボットの道徳的身分の帰属が、自由や意識といった証明困難な条件に依存して行われているということだ。もし仮に技術的人工物が心的状態や意識をもつことが原理的に可能だとしても、それは、個別具体的な人工物がそのような属性をもっているのだと証明する方法があることを含意しない。そしてそれゆえに、道徳的行為者性の標準的見解は苦境に陥るのである(Coeckelbergh 2009b, p.182)。

そこで Coeckelbergh は、この道徳的行為者性と「他者の心」の問題の関係性を、ロボットの文脈で積極的に利用することを提案する。すなわち、私たちは人間同士の道徳的行為者性および道徳的責任の帰属において、実際に特定の人間が心的状態や自由意志を持ち合わせているか否かを確かめていないのだから、ロボットにおいても同様のアプローチが取れるのではないか、と主張するのだ。そして、実在とは独立した形で、ロボットがユーザに対してどのように現れるのか、その経験に基づいて帰属できるような行為者性について言及するために Coeckelbergh が作り出したのが、「仮想的道徳的行為者性（virtual moral agency）」の概念である。

Coeckelbergh の仮想的道徳的行為者性の概念は、いわゆる行動主義（behaviorism）にコミットするものではない。Coeckelbergh は、私たちが心的状態と呼んでいるものが公に観察できる行動パターンに過ぎない、という主張に同意も否定もしないからだ(Coeckelbergh 2009a, pp.219-220)。心についてあくまでも不可知論的な立場を取りつつ、ロボットが与える経験に基づき道徳的行為者性概念を再構築する。それによって Coeckelbergh は、ロボットの道徳的身分に関する厄介な問題を避けつつ、従来の倫理学の用語を用いてロボットの道徳的価値を説明しようとしているのだ。

(3) 筆者の主張

3.1. ロボットが「ソーシャル」になるとき――仮想的行為者性の可能性

Coeckelbergh の関係論的転回、およびそれに基づく道徳的行為者性の再構築は、ソーシャルロボットの倫理の概念的基礎を与える可能性を秘めている。というのも、ソーシャルロボットの「ソーシャル」性は、ユーザが相互行為の中で見出していくものであり、その道徳的影響力についても人間側の想像力が中心的要素となるからである。

私たちにはコンピュータを、わずかな社会的合図（social que）だけで社会

的存在として扱ってしまう傾向性がある。これは今日、「メディアの等式（media equation）」と呼ばれるものである(Reeves and Nass 1996)。つまり私たちは、コンピュータをその内実ではなく、アピアランスに基づいて自分自身と同一視してしまうのだ。

道徳的行為者性の標準的見解においては、ロボットが特定の能力を有しているかどうかが問題になってきた。だが、ソーシャルロボットにおいてはまさにこの「能力」というものが当該のシステムのみで分析しきれない場合がしばしばある。あえて頼りないロボットを作ることで人間の優しさを引き出し、共生的関係を作り出そうとする岡田美智男の「弱いロボット」の研究（岡田 2012）はその好例だといえよう。ソーシャルロボットが道徳的行為者性の必要条件を満たしているのか否かが関係論的に見いだされるものならば、道徳的行為者性そのものもシステムのみに帰属させられるものではない、ということになる。

3.2. 仮想的な責任帰属の実践とは――仮想的行為者性の限界

仮想的行為者性の概念は、ソーシャルロボットの道徳的身分を考える上で有効である。では、道徳的行為者性の標準的見解が想定していたような責任帰属の実践に立ち返るとき、仮想的行為者性の概念はどのような示唆を与えるのだろうか。

Coeckelbergh は仮想的道徳的行為者性をもつロボットに問う責任も同様に仮想的なものでよいと主張し、これを「仮想的道徳的責任（virtual moral responsibility）」と呼んでいる。つまり、道徳的行為者性がアピアランスのみに基づいて帰属できたように、道徳的責任についてもアピアランスに基づき仮想的なレベルで問うことができると主張するのである。このとき、Coeckelbergh は、ロボットにおいては Himma（2009）がいうところの「不快な心的状態」、つまり苦痛を感じたり後悔しているように見えるのであれば、仮想的道徳的責任の帰属としては十分だと述べている(Coeckelbergh 2009b, p.185)。

だが、この仮想的責任の内実は仮想的行為者性に比べていささか表層的なものに見える。以下では Coeckelbergh の仮想的道徳的責任についての議論を確認し、その概念がなぜ表層的なものに留まっているのかを分析する。

3.2.1. 仮想的行為者性を認める基準の曖昧さ

Coeckelbergh が指摘するように、たしかに私たちの道徳的行為者性や道徳的責任の帰属の実践は、実在へのアクセスをせずに行っているといえるかも

しれない。たとえば、ある殺人事件の犯人の責任について考えるときには、その犯人の心についての探求に基づいてその答えを出そうとはしない。より日常的な場面でも、私たちはいわゆる「他者の心」の問題についていちいち悩むことはせず、実際に心をもつ存在として扱っているに過ぎない。

だが、この私たちの道徳的実践をそのままロボットに適用できるとは限らない。上記のような実践は、あくまで人間同士において（ひとまず）うまくいっているものではあるが、そこにロボットという技術的人工物を入れても問題がないとは限らない。

すでに言及したように、私たちは特定の状況の下では、ソーシャルロボットに対して容易に行為者性を帰属させてしまう。メディアの等式に関するNassの研究は、コンピュータがユーザに対して、実態よりも高度なものとして現れる可能性を示唆してきた。また、NeffとNagyは、マイクロソフトのチャットボット“Tay”がユーザから攻撃的な発話を学習して炎上した問題について、言説分析を行った(Neff and Nagy 2016)。調査は事件当時Tayについて言及したツイートを収集し分類するという単純なものではあるが、NeffとNagyは、ユーザの反応としては、Tayを人間の行動の犠牲者として扱うものと、Tayを（いわゆる）人工知能の潜在的な脅威として扱うものがあると分析している。Tayのプログラムに実装されている機能は、人間の心的能力と比べれば極めて単純なものに過ぎないが、それでも私たちは、特定の状況においては、そのような存在に行為者性を帰属させたくなるのだ。

Coeckelberghは仮想的行為者性の定義において、ロボットが行為者として現れる時間的範囲について明確に規定していない。メディアの等式がいうところの「等式」は、いわば錯覚のようなものであり、長期的な関係の構築においてもその等式が保たれ続けるとは限らない。したがって、メディアの等式が示唆するような事実から、私たちが行っているような、「他者の心」の問題を回避する形での行為者性の帰属の正当化に繋げるというのは飛躍がある。

Gerdesは、Coeckelberghの仮想的な道徳的行為者性の概念について、このような概念拡張を導入することは、長期的な視野においては人間同士の関係性を損なう危険をはらんでいると批判している(Gerdes 2016)。先述したように、ロボットが行為者性の必要条件を有しているように見えるためのアピアランスの達成がそこまで高いハードルではないことを考えれば、この危険性はより現実的なものになるだろう。

3.2.2. 「仮想的」を付けることによる責任帰属の構造の崩れ

そしてこのように人間の経験に基づいて行為者性を拡張することは、もと

もとの「標準的見解」で想定されていたような責任帰属の実践において、厄介な概念的混乱を招くことになる。

標準的見解においては、道徳的行為者であることは道徳的責任が問えるための必要条件であった。もともと責任帰属の問題を考える上でのひとつの中心的な基準が道徳的行為者性という概念なのだ。

では、「仮想的」な水準においてはどうだろうか。仮想的道徳的行為者だといえるロボットが仮想的道徳的責任をももっている、といえるとは限らないと思われる。ここで確認したいのは、道徳的行為者としてみなされるアピアランスと道徳的責任をもつものとしてみなされるアピアランスの基準に乖離がある、ということだ。すでに述べたように、仮想的道徳的行為者性の帰属については、ユーザ個々人がロボットのアピアランスをどのように経験するかに依存しており、あるロボットが短期的に道徳的行為者のように見えてしまう可能性はある。だが、このような水準のロボットは仮想的道徳的責任をもつともいえるのだろうか。おそらく Coeckelbergh の立場においてはもつ、ということになるのだろう。Coeckelbergh にとって仮想的道徳的責任が要請するのは、責任帰属対象が苦しんでいるようなアピアランスであった。この水準であれば、道徳的行為者のように見えるロボットが道徳的責任をもっているようにも見える、という構図を作ることはできそうだ。

しかしながら、Coeckelbergh がここで想定している「責任」は、極めて限られた意味での責任でしかない。反省しているように見える、苦しくなっているように見えるという要素は、たしかに私たち人間の責任帰属の実践においても重要ではあるが、そのすべてではない。したがって、仮想的道徳的責任の内実は（仮想的道徳的行為者の概念と比べ）本来の道徳的責任の内実からかけ離れたものになってしまっているのだと思われる。

これに対して、ここでいう仮想的道徳的責任は、あくまでロボットの文脈での責任概念の定義を試みるものであって、その文脈においては処罰に対して苦痛を感じるアピアランスだけで十分なのだ、という反論はありうる。だが、その場合、そのような責任の定義がロボットの文脈で正当化できる根拠を提出しなければならないし、少なくとも Coeckelbergh(2009b)はその議論を十分にしていない。

Coeckelbergh がここで想定している責任概念は、あくまで事後的な処罰に指向したものである。その一方で、そのロボットが将来的にその振る舞いを改善させる、という側面には言及されていない。だが、「責任」の定義においてはむしろ、そのような側面が重視されることも少なくない。

ここでは一例として少年犯罪と少年法における責任の関係を取り上げる。

秋本(2015)によれば、少年法上の責任概念は、再非行可能性と矯正可能性（成長発達可能性）という、非行少年の要保護性が主軸となっている。つまり、ここでの「責任」は、非行や犯罪をなした少年たちの行為に対応した事後的な責任非難を意味せず、処罰や制裁、非難としての法的責任を問うているわけではないのだ。また、髙内(2003)が指摘するように、少年法上の法的責任とは「保護処分を受けるべき地位」として解釈されている。すなわち、（過去の行為の非難としてではなく）将来的な非行反復可能性、また保護処分による矯正教育を通して成長発達を遂げる可能性に基づいて与えられるのがここでの「責任」なのだ。

もしロボットに対して（少年犯罪のように）苦痛を感じるよりもむしろ矯正・成長発達に関するアピアランスを求めるということが、「ロボットの責任」の意味として適切だということになれば、反省する素振りを見せるだけのロボットは、仮想的責任すら帰属できない存在だということになるだろう。

Coeckelbergh は道徳性についての標準的見解である「道徳的行為者→道徳的責任」という構図を議論の出発点としている。そして、ロボットの道徳的身分の考察という文脈でアピアランスの重要性を指摘し、仮想的な水準で行為者性の再定義を行っている。すなわち、ここで「標準的見解」の構図は、「仮想的道徳的行為者→仮想的道徳的責任」という形で拡張されている。しかし、これまで検討してきたように、この拡張版の構図が元の構図の関係性を保つことができるのは、極めて限定的な意味での「責任」においてのみであった。

したがって、もし仮想的道徳的行為者性をロボットの道徳的身分の考察の基礎に据えるのであれば、そこから導かれる仮想的道徳的責任の内実はどのように定義されるべきなのか、という規範的な問いにも取り組まなければならないだろう。そして、もし仮にロボットの文脈においても責任概念は苦痛のアピアランスを超えるものが要請されるのであれば、仮想的行為者性を帰属できる水準と仮想的道徳的責任を帰属できる水準は大きく乖離したものになる。その場合、もともと想定していた「標準的見解」の構造は保たれていないのだから、そのような段階のロボットの道徳的身分を、道徳的行為者性の概念を用いて説明しようとすることの妥当性は疑わしいものになるだろう。

(4) 今後の展望

私たちがソーシャルロボットを人間と近い社会的存在であるかのように扱うという心的傾向は、たしかにソーシャルロボットの道徳的身分を考える上で無視できるものではない。しかし、ロボットを自分と同一視するユーザが

いたとしても、そのユーザが「本当に」そのロボットを同一視しているとは限らない。現在のソーシャルロボットの文脈では、むしろユーザは騙されているのではなく、むしろアピアランスを手がかりに、社会的存在である「体で」扱うという場合が多い。この場合、ソーシャルロボットのユーザがもつ感情は、手品ショーやテーマパークのキャラクター・グリーティングで生じるような「虚構的な」感情の一種として考えられるだろう。このように考えたとき、Coeckelbergh の仮想的行為者性の概念は、K. Walton や M. Ryan などの虚構論を取り入れつつ、ある種の虚構的な行為者性として再構築可能かもしれない。この点については今後詳細に検討していきたい。

(5) 参考文献

Coeckelbergh, M. (2009a). "Personal robots, appearance, and human good: A methodological reflection on roboethics." *International Journal of Social Robotics*, *1*(3), 217–221.

Coeckelbergh, M. (2009b). "Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents." *AI & society*, *24*(2), 181–189.

Gerdes, A. (2016). "The issue of moral consideration in robot ethics." *ACM SIGCAS Computers and Society*, 45(3), 274-279.

Gunkel, D. J. (2012). *The machine question: critical perspectives on AI, robots, and ethics*. MIT Press.

Himma, K. E. (2009). "Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?" *Ethics and Information Technology*, *11*(1), 19–29.

Lin, P., Abney, K., & Bekey, G. A. (2014). *Robot ethics: the ethical and social implications of robotics*. The MIT Press.

Lin, P., Abney, K., & Jenkins, R. (2017). *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford University Press.

Neff, G., & Nagy, P. (2016). "Talking to bots: Symbiotic agency and the case of Tay." *International Journal of Communication*, *10*, 17.

Reeves, B. & Nass, C. (1996). *The media equation: how people treat computers, television, and new media like real people and places*, Cambridge University Press.

Sullins, J. P. (2006). "When is a robot a moral agent?" *Machine ethics*, 151–160.

岡田美智男. (2012). 『弱いロボット（シリーズ　ケアをひらく）』. 医学書院.

秋本光陽. (2015). 「犯罪少年の『責任』はいかにして組織されうるか」 『ソシオロゴス』, *39*, 191–210.

髙内寿夫. (2003). 「現行少年法における『責任』概念について」 『法政理論』, *35*(4), 75–113.

（東京大学）