われわれは自律的ロボットの行為に責任を負いうるか
Are we responsible for autonomous robots' acts?

姜雪菲

Abstract

Robots are now becoming more and more autonomous. If completely autonomous robots are eventually invented, are we responsible for those robots' acts? If not, it will be the case that there exists nobody responsible for those robots, which is called "the responsibility gap." However, some philosophers say the responsibility gap cannot happen in the first place, and we still are responsible for autonomous robots because of several reasons. In this paper, I argue against these claims; that is, I try to defend the occurrence of the responsibility gap. Then, I would like to suggest my own solution to that problem.

(1) 研究テーマ

ロボット倫理は応用倫理の一部門である。概して、ロボット倫理とは、ロボットの利用によって生じる倫理的な問題を扱う分野である。様々な問題の中で、近年非常に関心を集めているのは、自律的ロボットの行為に対して誰が道徳的責任を負うべきなのかという問題である。もし将来、機械学習などの技術によって完全に自律的なロボットが実現されれば、我々はそのようなロボットの行為に対して責任を負いうる適切な人間を見つけられない状況に陥る可能性がある。この懸念は多くの場合、「責任ギャップ」と呼ばれている。一方、「責任ギャップ」の発生を否定し、そのような問題は生じない、あるいは少なくとも必ず発生するとは限らないと主張する立場がある。本稿では、こうした責任ギャップの発生に否定的な立場の議論を整理して応答を行う。

本稿の構成は以下のとおりである。まず研究の背景において、「責任ギャップ」とはどのような倫理的問題なのか、さらにそれがいかに生じるのかを説明する。そして、この問題に対して様々な否定的な立場の主張を先行研究として提示する。次に筆者の主張で、提示したそれぞれの立場に対して応答を行い、「責任ギャップ」は十分に発生しうる問題であり、それに対して解決策を講じることは重要な課題であると結論付ける。最後に、今後の展望において、一つの可能な解決策を提示する。

(2) 研究の背景・先行研究

ロボットは次第に自律性を獲得する方向へと発展している。完全に自律的なロボットがまだ実現されていない現段階では、プログラマーや操作者など、監督や制御の役割を果たしている人たちこそロボットの行為に道徳的責任を負うべきだと考えている。しかし、もし将来的に完全に自律的なロボットが実現されれば、ロボットの行為に対する道徳的責任を、上記の人々に帰すのは困難になる。なぜなら、自律的ロボットが独力で意思決定ができるようになれば、そのようなロボットの行為を制御し予測することは困難になるからである。それゆえ、我々は責任ギャップに直面する。以下では、この問題をより明確にするために、まず道徳的責任とはどういうことなのかについて説明し、道徳的責任はどのような条件に基づいて、ある特定の人間に帰属するのかを確認することにする。

概して、道徳的責任を負うとは、「ある人の行為が正と不正、徳と悪徳、善と悪などの用語で評価されることができ、そしてその人が自分の行為とその行為によって導く結果に対して非難や賞賛を受けることもできることを意味している」（Mckenna and Pereboom 2016）。従来、道徳的責任の対象は成熟した人間に限定されている。動物や幼い子どもは道徳的配慮の対象にはなるが、自ら責任を負うための十分な能力を持っていないため、我々は彼らを道徳的行為者とはみなさない。

では、我々はどのような条件の下で、ある人の特定の行為に対して、それを非難（賞賛）に値すると考えるのだろうか。基本的には、その人が自分の行為をコントロールすることができて、かつ自分が何を行っているのかを知らなければならないことが要求される（Talbert 2019）。これらはそれぞれコントロール条件と認識的条件として規定される。例えば、ある人が悪い催眠術師によって強引に催眠術にかけられ、術師の指示に従って詐欺を実行したとしよう。この場合、催眠された人は悪いことを行ったが、催眠されている状態のため、自分の行為をコントロールすることができないし(コントロール条件の欠如)、しかも自分が何をしているのかに関しても十分な知識をもたないので(認識的条件の欠如)、自分の行為とその結果に関して責任を要求されることはないだろう。

上記の二つの条件のもとで、自律的ロボットが利用される場面を考えよう。例えば、自律武器システム（AWS）が司令官から命令を受けて、自ら敵軍に攻撃を行うとする。それらは人間の兵士と同じように、具体的な攻撃の対象を自分で判断し、どのように攻撃を行うのかも自分で決める。すなわち、AWSは戦場においてはほとんど人間の兵士と同じような役割を果たしている。AWSがこのような能力を持つのは、主に実装される学習技術によるものであ

る。このような技術はAWSの自律性を支えるが、AWSの本来の開発者たちも含めて、AWSが学習技術を通して習得する新たな情報、そしてその情報を用いてAWSが下す判断に関して、それを予測したり、それに干渉したりすることはできない。このとき、開発者や司令官の本来の意図や命令から外れて、AWSが有害な結果を引き起こしたならば、誰が道徳的責任を負うべきなのだろうか。

責任ギャップが発生しうると考えている論者たちは、このような場面では責任を負いうる適切な対象はいないと主張する（Matthias 2004, Sparrow 2007, Danaher 2016）。例えば、ロバート・スパローは、AWSが引き起こした損害に対して、プログラマー、司令官、あるいはAWS自体が責任を負う可能性を検討したが、責任を負うべき対象は存在しないと結論づけ、そのような帰結を招くAWSの開発自体を控えるべきだと主張するに至った(Sparrow 2007)。

一方、責任ギャップの発生を否定する論者たちは、それぞれの理由に基づき、自律的ロボットによって生じる損害に対して責任を負う対象を特定することは可能だと主張し、責任ギャップをめぐる議論に異を唱える。筆者の見解では、こうした試みはおおよそ二つの方向性に分類できる。第一の方向性は、たとえロボットが自律的になっても、依然としてそれに対する何らかのコントロールをもつ人間が存在するため、責任ギャップは生じないというものである。第二の方向性は、自律的ロボットの行為はある程度、または全面的に人間に依存するので、責任は関連する人たちが負うべきだという考えである。前者としては、Himmelreich（2019）やSchulzke（2013）が挙げられ、後者としてはNyholm（2018）やRobillard（2018）が挙げられる。以下では、それぞれの主張の要点を示そう。

(2)-1 自律的ロボットに対しても何らかのコントロールをもつ対象がいる

ヨハネス・ヒンメルライヒは、たとえAWSが完全に自律的になったとしても、上位にいる司令官は、命令を下すことによって最終的な結果に対しある程度のコントロールをもつことができるので、結果に対して責任を負うべきであると主張する。ここでヒンメルライヒは、たとえ不適切な目標への攻撃といった、望ましくない結果を実際に引き起こすのがAWSであるとしても、その発生は司令官たちが下した命令まで遡ることができると述べる。たとえ司令官らがAWSの特定の攻撃や、その攻撃に伴う具体的な結果に対して直接コントロールすることができなくとも、誰かが攻撃されるという確率的な結果、そして不適切な目標に危害を加えるリスクに対してはコントロー

ルすることができるからである。加えて、戦争で、本来の意図と外れた結果が発生すること自体は一般に予見可能であるため、司令官は AWS が機能不全などによって、望ましくない結果をもたらす可能性があることは認識できる。したがって、司令官らはコントロール条件だけでなく、認識的条件も満たすことになり、AWS によって引き起こした損害に対して道徳的責任を負いうることになる。ゆえに、ヒンメルライヒは、たとえロボットが完全に自律的になっても、責任ギャップは生じないと結論づけるのである。

同様にして、マーカス・シュルツケは、戦争中のＡＷＳ使用に焦点を当て、AWS の行為に対しては責任を負う適切な対象が存在し、責任ギャップは生じないと主張した。しかし、ヒンメルライヒとは異なり、シュルツケは責任帰属の対象には司令官だけではなく、AWS の開発者たちも含まれると主張する。なぜなら、以下で述べるように、司令官も開発者もＡＷＳに制限を加える対象である点では、共通する振る舞いをもつからである。

シュルツケによると、司令官の道徳的責任は、軍事的ヒエラルキーというシステムによって定められるものである。戦争に関するすべての意思決定は、ヒエラルキーの上位にいる人たちが行い、下位にいる兵士たちは実行できる行動の範囲が限定され、特定の行動のみが許可または禁止される。このようにして、兵士の自由は厳しく制限されているため、兵士は、指示された通りに遂行した自分の行為に対して責任を負わされることはほとんどない。代わりに、責任は意思決定者たちに分配されている。ゆえに、将来的に人間の兵士が AWS に置き換えられたとしても、同様のシステムによって、責任はヒエラルキーの上位で意思決定をつかさどる人間たちに帰属することになる。

さらに、シュルツケは AWS の開発者も責任を負うべきであると主張する。シュルツケによれば、ロボットの操作は初期の設定に強く依存しているため、ロボットが損害を引き起こしうるのは、開発者が適切な制約を加えることに失敗しているからである。AWS が新たな情報を学習することができるにもかかわらず、それらの総合的な行動の範囲は初期の設定によって制限されるため、それらの能力と傾向は開発者まで遡ることができる。ゆえに AWS が誤った学習を開始しないよう制限を加えないことは、開発者たちの過失であるとシュルツケが述べた。

つまりシュルツケから見ると、司令官や開発者は自律的ロボットに制限を加えることができるため、もし損害が生じれば、彼らが適切な制限を設定することに失敗したことに対して責任を負うべきである。それゆえ、自律的ロボットの行為に対しては責任を負う対象が存在し、責任ギャップは生じない。

(2)-2 ロボットの行為者性を否定する

スヴェン・ニーホルムとマイケル・ロビラードは、ロボットが完全な行為者性を持たないという点で共通の見解をもつ。ただし、ニーホルムはロボットがある程度の行為者性を持つことを認める一方、ロビラードはロボットが全く行為者性を持たないと考えている。

自動運転車や軍事ロボットなどの自動化されたシステムの行為者性について、ニーホルムは人間―ロボットの協同という形で理解すべきであると提案する。なぜなら第一に、これらの自動化されたシステムの動きは特定のルールや原則に規制され、しかもそれらの動作に干渉することができる権威者が存在するからである。例えば自動運転車の場合、自動運転車が独力で自分の行き先を決めることはできない。すなわち、目的地は人間によって設定されているため、自動化されたシステムの自律性とは単に人間が設定した目標に向けて実行する動作のことにすぎない。この意味で、それらは独力で行為を行うとは言えず、必ず誰かほかの対象と一緒に行動しなければならないといえる。ゆえに、たとえ自動化されたシステムがほとんどの作業を自分で行ったとしても、それらの行為は他の対象による監督に依存しているので、それらの行為に対して責任を負うべきなのは、監督の役割を果たしている人間であるとニーホルムは主張する。

ロビラードはさらに強くロボットの自律的な行為者性を否定する。ロビラートによれば、ロボットの行為者性は企業や政府などと類比的に考えられるべきである。すなわち、ロボットは単に「物理的な形で例示化された社会的に構築された組織にすぎない」（P.713）ため、それらのあらゆる行為には人間の本来の意図と決定を超えることはない。そして責任はこの組織を構成する人間たちが負うべきである。もちろんここで集団が意思決定を行うという主張自体に対して、つまり集団主義（collectivism）に対して反論を加えることはできるが、ロビラードはその種の反論は既に議論されている問題であって、AWS に特有の問題ではないとして退ける（P.713）。

以上のように、ニーホルムとロビラードはともに、ロボットの行為に対する責任は関連する人間に帰属すべきであり、責任ギャップは発生しないと結論づけている。

(3) 筆者の主張

私の考えでは、責任ギャップに対する上記の二つのタイプの反論は、いずれも責任ギャップの発生を完全に排除することに失敗している。すなわち、私の考えでは責任ギャップは依然として生じうる問題である。

まず、自律的ロボットに対するコントロールすることができる対象がいるという主張する論者たちに対して、以下の点を指摘して応答したい。

第一に、AWS の司令官だけが、最終の結果に対してコントロールをもち、一般的な予測もできる、というわけではない。例えば AWS の開発者も AWS を開発するかどうかを決めることができるという点では、最終の結果に対してコントロール持っているといえる。このような漠然とした基準に基づいて責任を負うべき対象を決めることは、関連のない対象へと責任の範囲を過剰に拡大してしまう恐れがある。

第二に、確かに現在は軍事ヒエラルキーの上位の対象が下位の行為を統率し、責任を負うというシステムが機能しているかもしれない。しかし、このようなシステムがずっと維持され続けるかどうかは誰も保証することができない。ルチアーノ・フロリディ（2017）は、技術だけからなるループを三次技術と呼ばれ、それが実現されれば人間のユーザーは、「もはやループの中には存在せず、せいぜいそれに接しているだけである」（P.38）と考えている。このことは、AWS に対する軍事ヒエラルキーに対しても当てはまりうる。すなわち、軍事システムの全体がロボットや AI によってなされる状況は想定可能である。もしそのような状況になれば、統率役のロボットに責任を帰属させない限り、同様のプロセスで自律的ロボットの行為に対する責任を誰かに帰属することはできない。

第三に、たとえ司令官や開発者がロボットに制限を加えることができるとしても、すべて可能な結果を予測することはできない以上、あらゆる望ましくない結果の発生を防ぐことは不可能であると思われる。自律的ロボットに加える制限には限界があるのであり、限界を超えた事象に対して、制限を加えたものに全ての責任を負わせることは不合理である。

以上のように、ヒンメルライヒとシュルツケは自律的な行為者の行為が行為者自身以外に帰属するケースやシステムを持ち出して、自律的ロボットにも適用しようとしているが、両者の想定はいずれも不十分である。すなわち、両者は、自律的なロボットの出現によって、標準的な責任帰属条件が完全に適用不可能になる可能性に、十分な注意を払っていない。

次に、自律的ロボットの行為者性を否定する論者たちの主張について、以下の点から応答しよう。

第一に、確かに、自動運転車や軍事ロボットなどの自動化したシステムの最終的な目標は人間によって設定される場合が多い。しかし、たとえロボットがそれ自身では目標を持たないとしても、設定された目標に向けて自らで動作する過程では、どのようにタスクを達成するのかについて、自身で判断

を下し、具体的な行為を実行することができると想定されている。つまり、この過程において、自律的ロボットは従来人間の役割を完全に代替することができるようになる（Coeckelbergh 2016)。そうであれば、ロボットの動作を監督し、権威のある人が実際の状況でロボットの一連の動作全てに介入することはないと思われる。そして、自律的ロボットが独自の意思決定を経て、最終的な目標から逸脱し、損害等、意外な状況が発生させることは十分にあり得る。少なくともこのような状況に対しては、責任ギャップは依然として発生しうると考えられる。

第二に、個人でも、集団でも、意図されていない結果に対してすべての責任を持たせることは不適切である。例えば、人間の兵士が意図的に不適切な対象を攻撃する場合、司令官にそれに対する責任があるが、兵士自身に全く責任がないとはいえない。同じように、自律的ロボットの開発または利用において、ロボットが全く意図せざる行為を行った場合でも、司令官や開発者が道徳的に非難されることは妥当ではないと私には思われる。

以上のように、責任ギャップの発生を否定する論者たちの主張はいずれも不十分であって、決定的ではない。すなわち、責任ギャップは依然として発生しうる問題であり、それに対する対処法を考えることは重要である。

(4) 今後の展望

責任ギャップが発生しうることが認められるならば、次の課題は、それが深刻な影響を及ぼす前に、予め解決策を講じることである。ここではその解決に向けて、一つの方法を提案したい。それは、自律的ロボット自身に道徳的責任を帰属することによって、責任ギャップを埋めるというものである。ここで、ロボットに責任を持たせるという提案自体は既に提出されている（Bechtel 1985, Stahl 2006)。しかし、彼らはいずれもロボット自身が道徳的責任を負うことができるように明確な基準を記述できていないという問題を抱えている。

では、自律的ロボットはどのような条件を満たせば、道徳的行為者とみなされるのだろうか。標準的な枠組みのもとでは、道徳的行為者は成熟した人間に限定されるため、道徳的行為者性の条件はそれに応じて、自由意志や人間性などであると考えられてきた。しかし、これらの条件は、特定の行為を個人に帰属するという道徳的責任の一つの側面だけを重視するものである。こうした条件は、道徳的責任が、行為者と道徳的コミュニティ(moral community)における他のメンバーとの関係において果たす役割について、十分に表していない(Watson 2004)。ゆえに、我々は自由意志や理性など人間

特有の性質に頼らずに、道徳的行為者であることの条件を再考する必要がある。そこで、精神をもたない行為者が道徳的行為者となるための条件を整備することが、本論で論じてきた問題のもっとも見込みのある解決策であると私は考える。

(5) 参考文献

1. Bechtel, W. (1985) “Attributing responsibility to computer systems.” *Metaphilosophy* 16 (4):296-306.
2. Coeckelbergh, M. (2016) “Responsibility and the Moral Phenomenology of Using Self-Driving Cars.” *Applied Artificial Intelligence* 30:8, 748-757
3. Danaher, J. (2016) “Robots, Law and the Retribution Gap.” *Ethics and Information Technology* 18 (4):299–309.
4. Himmelreich, J. (2019) “Responsibility for Killer Robots.” *Ethical Theory and Moral Practice* 22 (3):731-747.
5. Matthias, A. (2004) “The responsibility gap: Ascribing responsibility for the actions of learning automata.” *Ethics and Information Technology* , 6: 175–183.
6. McKenna, M. & Pereboom, D. (2015) *Free Will: A Contemporary Introduction* Routledge..
7. Nyholm, S. (2018) “Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci.” *Science and Engineering Ethics* 24 (4):1201-1219.
8. Robillard, M. (2018) “No Such Thing as Killer Robots” *Journal of Applied Philosophy* 35 (4):705-717.
9. Sparrow, R. (2007) “Killer Robots.” *Journal of Applied Philosophy*, Vol. 24, No. 1.
10. Schulzke, M. (2013) “Autonomous Weapons and Distributed Responsibility.” *Philosophy and Technology* 26 (2):203-219.
11. Stahl, B. (2006) “Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency.” *Ethics and Information Technology* 8 (4):205-213.
12. Watson, G. (2004) “Two faces of responsibility.” *Agency and Answerability: Selected Essays*, 260-288.
13. フロリディ, L, 2017, 春木良且・犬束敦史訳者代表, 『第四の革命―情報

圏(インフォスフィア)が現実をつくりかえる』 新曜社.

（大阪大学）