新進研究者 Research Note

人工知能の自律性について

山﨑かれん

Abstract

Can artificial intelligence be autonomous like human agents? In this article, I adopt Frankfurt's and Bratman's views of "personal autonomy" as possessed by human agents and examine whether BDI agent can have it. BDI agent is an artificial agent in a field of artificial intelligence studies. The architecture of BDI agent focuses on mental states of human agents. I point out that BDI agent doesn't hold some features of personal autonomy, but it can be said that it will have "self-governing policy" which Bratman takes as one important element of personal autonomy.

(1) 研究テーマ

自らの行為を自ら決定するという行為者としての人間が持つ自律性を、人工的なエージェントが持ちうるかどうか検討する。

(2) 研究の背景・先行研究

昨今なにかと話題にのぼる人工知能についての言説の内では、「自律性」が一つの標語となっている。人工知能が自律性を獲得することで人間による制御が効かなくなることが不安だとか[1]、人工知能に汎用性を持たせるためには自律性を付与することが効果的だとか言われている（栗原ほか, 2017）。

しかし、人工知能の研究開発の現場や、人工知能開発について言及する研究を見ても、自律性という言葉の使用は多岐にわたる。そうした状況を指摘している研究として、認知科学者であるトム・フレーゼらのものがある。フレーゼらは、人工生命研究[2]において適切な自律性の形を提案するために、ロボティクスや人工知能研究、人工生命研究を参照して人工的な主体を構築する際の自律性を分析している。フレーゼらは、環境との間での安定した、柔軟な相互作用能力から特徴づけられる「行動の自律性」と、オートポイエーシス研究的な伝統から特徴づけられる「構成的な自律性」という二つの区

別がつけられることを指摘した。行動の自律性は、環境と相互作用しつつ、初期の設計や行動時の持続的な介入によることなく動作する能力や、さらにその上で目標を作り達成する能力、また行動の柔軟性が高いということなどから特徴づけられる。一方で構成的な自律性は、自己創出性などから特徴づけられる生命システムに注目したもので、本来的には生物にしか達成しえない、非常に制約の大きい自律性概念だと指摘されている。

今回は前者のような、主体の動作に主眼を置いた自律性を主題としたい。動作する主体、あるいは行為者を人工的に構築することや、そうして構築した人工的な行為者を用いて研究を行う人工知能研究の領域を「エージェント研究」という。エージェント研究にはいくつかの異なる領域があるが、その中に、世界の中で環境に応答し、判断して行動するエージェントを作る研究がある。エージェントの例としては、現実世界で活動する知能ロボットや、仮想的な世界で行動するゲームのキャラクターが挙げられる。特に自律性を強調したものを自律エージェントとも言うが、こうしたエージェントについて言われる自律性を詳しく見てみよう。

ゲームのプログラマであるマット･バックランドは､「『自律エージェント』という用語を、ある程度に自律的な動作を遂行するようなエージェントに関して用いる。例えば、道中で壁に出くわすとかいうような予期せぬ状況が起きたとき、自律エージェントはそれに対して反応し、適切に動作を生成することができる」（傍点は筆者）と著している。

人工知能に関する哲学的論考を多く行うヴィンセント・ミュラーは、主体の行動の柔軟性に着目し、それを与えるために自律性が求められるとした。ミュラーは、自律的な主体に関して、「主体 X は、X が主体 Y からの入力なしに目的を追い求める程度において、Y から自律的である」と特徴づける。

人間の行動決定過程についてのマイケル・ブラットマンの分析をもとに、主体の心的状態に焦点を当てて自律エージェントを構築しようする研究も行われている。BDI モデルに基づく BDI エージェントである。この研究では自律エージェントとは、人間と同じように目的を持ち、それを達成する方法を考えて、そのために行動するようなエージェントだとされている。BDI モデルは、心的状態のモデルとその時間的変化を用いて、複数の目的を並行して扱うことを可能にすることで、人間の合理的な行動決定を模倣し、自律エージェントを実現しようとする。

心的状態の中でも情動が行動を動機づけることに注目し、ロボットに情動システムを与えることで自律的な行動生成を目指す研究もある（錦田, 2016 など）。こうした研究では、ロボットの情動を作り出すことで、ロボットが

自ら考え、自らの意志によって行動するようになり、それこそが自律性の実現だと考えられている。

このようにして、人工的な行為者の自律性を見ていくと、自律性についてさらに二つの観点があることが見て取られる。一つめの観点は、その主体の行動が環境に対して柔軟であり、外部からの制御を受けずにふるまうという点から自律性を言うものである。たとえば、バックランドは自律的に動作すれば、自律エージェントだとした。ミュラーも、外部からの入力なしに目的を追うというところに自律性を定めた。二つめの観点は、その主体のふるまいが、まさにその主体自身の欲求や信念、意図、目的選択といった心的状態によって生み出されているという点から自律性を言うものである。BDI エージェントは人間の心的状態に着目していた。また、情動モデルを与えることで自律的な人工物を作るという構想も、情動が行動生成のための主体の思考や意思を生み出すことに期待している点で、情動や思考、意思といった心的状態が自律性には不可欠だと考えているだろう。つまり、この二つの自律性の違いは、自律エージェントの自律性を主体のふるまいから捉えるか、それとも主体の心的状態から捉えるかという観点の違いである。以降では、上で挙げたうちの二つめの観点を問題にする。

哲学には、行為者たる人間の重要な特性として自律性を分析した「人格的自律性（Personal Autonomy)」という考えがある。粗く言うと、人格的自律性では「自らの行為を自らの欲求や信念などによって決定できる」ときにその行為者は自律しているとされる。つまり人格的自律性によると、行為者の心的状態のあり方がその自律性を実現するのだ。

先に挙げた二つめの観点をとる上で、人格的自律性は人間という行為者の自律性についての適切な分析だと考えられよう。人工知能の開発目標には、「人間のような」知的ふるまいをする人工物を作る、というものがある[3]。そのために、人間という行為者の一つの特徴として人格的自律性があるとすると、それを人工知能あるいは人工的な行為者に持たせるということが考えられてもよいだろう。先に挙げた自律エージェントの自律性についての二つめの観点は、エージェントの自律性を構築するための手段として、人格的自律性が要求するような心的状態をエージェントに持たせなければならないと考える、エージェント設計上の立場であるとも言える。では、この立場で設計されたエージェントは、人格的自律性として分析されている自律性を本当に持つことができるのだろうか。ここからは、人格的自律性を人工的な行為者が持ちうるかどうかを問うていく[4]。

(3) 筆者の主張

まずは、問題にする人格的自律性がどのようなものかを明らかにしておきたい。人格的自律性の議論は 1970 年代のハリー・フランクファートやジェラルド・ドウォーキンの論考に端を発する。ここでは、その後の議論に広範な影響を与えるフランクファートの議論を取り上げ、そこから人格的自律性の骨子を理解する。そして、フランクファートの不足を補うものとしてブラットマンの説を導入し、今回検討する人格的自律性の概念を明確にしたい。

フランクファートによると、人間はなにかをする／しないといったことを欲するような「一階の欲求」に対して、その欲求を持つことがよいのかどうかを反省的に評価する態度を持つような行為者である。この反省的な自己評価を行う心的状態は「二階の欲求」とも呼ばれる。そして、その二階の欲求のうち、ある一階の欲求が自らの意志になる（＝実際に行為を引き起こす）ことを望む二階の欲求のことを「二階の意欲」と呼び、フランクファートは二階の意欲を持つものを人格として認めた。また、二階の意欲が成立している状態を、その人が一階の欲求に同化しているとも言う。この立場は高階の心的状態による自己統制の仕組みを明らかにすることで自律性の成立を説明し、階層的アプローチと呼ばれて広く影響を与えている。

しかし一方で、単に欲求の階層的な構造から自律性を説明するのでは、高階の欲求も単なる欲求にしかすぎないので、それだけでは行為者が一階の欲求を支持することを保証する権威を持ちえないだろう、という問題が起こる。フランクファートはそうした問題を避けるために、ただ二階の意欲が一階の欲求を欲するというだけではなくて、その人が二階の意欲を自らのものとして受け入れることも必要だと付け足した。これを満足と呼び、行為者が一階の欲求を欲する二階の意欲について反省を行ったあともなお、その二階の意欲を変更するという気が起こらないときに、その二階の意欲に満足していると言われる。そしてそのように二階の意欲に満足しているときに、行為者は一階の欲求に同化していると考えるのである。

しかし、ブラットマンはこのようなフランクファートの提案では十分ではないと考え、その論考をさらに発展させる形で分析を行う。ブラットマンによると、自律的な行為者の特徴として「計画すること」があり、計画によって複数の時点にまたがる行為の組織的な統一が可能になる。人間は時間的な広がりの中で活動する社会的な行為者だが、計画を持つことで行為の直前にその手段について考える必要がなくなり、また個人内部での活動の調整と他人との間の活動の調整が可能になるのだ。そしてブラットマンは、計画を構成する一つの要素として、「意図」という心的状態に、ある重要な独自の役割

を与えている。このような意図は比較的特定された状況においてある種の仕方で行為するという、行為に対してのコミットメントである。意図の中には、あるタイプの状況が生じるたびに一定の行為を行うこと（車に乗るときはシートベルトを締める、とか）にコミットする、より一般的な意図がある。これは「方針」と呼ばれる。

このような方針は、行為に関するものから一階の欲求に関するものへと拡張され、一階の欲求に関するものは特に「自己統制的方針」と呼ばれる。ブラットマンは自己統制的方針の例として、チョコレートの誘惑に負けないようにしよう、とか、陽気であろうとすることに努めよう、といったことを挙げている。自己統制的方針は、実践的推論においてある欲求を行為の理由として扱うことを可能にするものであり、このことが欲求に対する行為者の支持を意味するのである。そして、ブラットマンはここに「満足」についての彼なりの考察を組み込む。ある人が持つ自己統制的方針に、異議申し立てをする別の自己統制的方針がないということが、その人が自己統制的方針に対して満足しているという状態である。行為者が自分の満足する自己統制的方針で欲求を肯定的に評価し、その欲求を実践的推論における行為の理由として扱うことで、その自己統制的方針がその欲求に対する行為者の支持を保証する権威になると考えるのだ。

ブラットマン自身もほのめかしているが、他にも自律的な行為者の特徴づけとして適切な性質はあるだろう。しかし、少なくともここまで見てきた考察は自律的な行為者が持つ重要な特徴を捉えていると考えられよう。以下では、今まで見てきたフランクファートとブラットマンの考察から、人格的自律性を考える。つまり、高階の態度が一階の欲求を反省するという欲求の階層的な構造の上で、長期的な視野を持った自己統制的方針が一階の欲求を肯定的に評価し、それを実践的推論における行為の理由とすることで自己統制的な行為をすることのできる主体を、人格的自律性を有する主体だと言おう。

人工的な行為者はこのような人格的自律性を持つことができるのだろうか。しかし人工的な行為者一般について問うことは難しいため、ここでは具体的な例として先にも挙げた BDI エージェントを分析の対象とする。BDI エージェントは、ブラットマンの意図と計画についての論考を参考にし、人間の心的状態に焦点を当ててエージェントの行為をモデル化した BDI モデルに基づくものであり、その意味でもここでの分析対象として興味深い。

ブラットマンは、人間が合理的に行為するときには、信念（Belief)、欲求（Desire)、意図（Intention）という三つの心的状態が深く関与していると考えた。BDI モデルはその分析を下敷きにして自律エージェントを構築しよ

うとする、工学的な設計指針である。BDIエージェントは工学的なモデル化を経て実現された三つの心的状態とその時間的変化を反映させて意思決定を行う。特にBDIエージェントは自ら欲求を持たないため、BDIモデルにおける欲求は外部から与えられる目標となっている。

BDIエージェントのとる動作は、次のようにまとめられる[5]。

① 与えられた目標（欲求）を読み込む。
② 実践的推論によって目標を達成する手段を定めて、その手段を実行するという計画を、意図として形成し、それを保持する。
③ 環境についての信念を得て、実行の前提となる条件が満たされれば、その意図が実行しようとしている手段が遂行される。
④ 意図はそれらの手段を順次実行に移させる。またその手段が、BDIエージェントが直接遂行できない複雑な動作である場合、その手段を副目標として設定し、副目標を達成するための副計画を立てる。
⑤ 一度形成された意図はある程度の持続性をもち、これによって、目標達成に向けた動作が一貫したものとなる。

今のところBDIエージェントが完全な形での人格的自律性を持ちえないことは明らかであるように思われるが、何が不足しているのかを具体的に挙げてみたい。まず、BDIエージェントには、欲求を自ら生み出す機構が欠けていることは明らかであろう。BDIエージェントにとって欲求とは、設計者や使用者から与えられた目標であり、BDIエージェントはそれを反省、評価せずに、その目標を達成するための行為を実行しようとする。それゆえ、BDIエージェントは欲求を反省、評価するような機構も持っていない。

ただし、BDIエージェントは与えられた目標を達成するための手段を講じて、その手段を実行する意図を形成することで、人間のような自律的な行為者の一側面を捉えるように作られている。こうした意味では、「計画すること」という自律的な行為者の一側面を掬うものになっている。

目標を実行しようという意図が、BDIエージェントが直接実行できる動作よりも複雑な手段を必要とするとき、BDIエージェントはその手段の実行自体を新たな目標（副目標）とする。そしてBDIエージェントはその副目標を達成するために、さらに手段を講じてそれを実行するという意図を形成する。

BDIエージェントがはじめに持つ目標（欲求）は、設計者や使用者から与えられる。一方、与えられた目標を達成するための副目標は、自ら立てた目標（欲求）である。副目標を立てるとき、そこではその副目標が与えられた

目標を達成するための適切な目標たりうるか、という評価や反省がなされるはずである。つまり副目標に対して、それが自分に与えられた目標を達成しようという意図に沿うかものなのかどうかという評価､反省がなされるのだ。

ここにおいて､自己統制的方針を BDI エージェントが持つことになると考えてもよいのではないか。長期にわたる複雑な計画を通してはじめて達成されるような目標を BDI エージェントに与えたとき､その目標を首尾よく遂行するために一般的な意図、つまり方針を立てなければならないだろう。この方針の中には自己統制的方針も含まれよう。なぜなら、自分の立てた副目標（欲求）が、与えられた目標を実現するために適切であるかどうかを評価、反省し、かつ、その副目標を実践的推論における行為の理由とすることを可能にするような方針（すなわち自己統制的方針）も必要になると考えられるからである。このことから、BDI エージェントは今回定めた人格的自律性が必要とする行為者の性質を、部分的にではあるが備えていると言えよう。

(4) 今後の展望

BDI エージェントは人間のような形で自ら欲求を持つことがないという点において、今回定めた人格的自律性を持つことにはなっていない。この問題に対処するための研究として二つの方向を考えている。一つは人工物に自ら欲求を生み出す機構を与えるという方向、もう一つは自ら欲求を持たない人工物に自律性を認めるという今回とは異なる形の自律性概念の構築を行うという方向である。特に後者については、道徳的責任を帰せられる人工的主体の研究として倫理学の方面からも検討が試みられているようなので、そうした知見を取り入れつつ考察を深めたいと考えている。

[1] 長倉(2017) などで､こうした不安を煽る言説が批判的に指摘されている。

[2] 人工生命研究は、ロボットやソフトウェア上のモデルによって、生命システムを人工的に模倣するものである。人工知能と人工生命という双方の研究領域は、どこまで共通しているかを指摘することは難しいが、人工的な行為者の自律性を考える上ではフロースらの分析が役立つと考えられる。

[3] 中島 (2013) など、このシリーズでは人間の知能の解明が人工知能研究の目的の一つだと繰り返し述べられている。そのために､「人間のような」知的ふるまいをする人工物の構築が目指されていると考えられるだろう。

[4] この種の考察として、たとえば Schmidt et al. (2006) は以下の本文で述べるようなフランクファート的な欲求の階層構造を人工知能が持つかどうかを問うている。

[5] 新出（2010）のまとめを参考にした。

(5) 主要参考文献

Bratman, Michael (1987). *Intention, Plans, and Practical Reason*, Harvard University Press.(門脇俊介＆高橋久一郎（訳）(1994).『意図と行為 合理性、計画、実践的推論』, 産業図書.)

Bratman, Michael (2000). "Reflection, Planning, and Temporally Extended Agency," *The Philosophical Review*, 109(1), 35-61.

Buckland, Mat (2004). *Programming Game AI by Example*, Jones & Bartlett Learning.

Frankfurt, Harry (1971). "Freedom of the Will and the Concept of Person," *The Journal of Philosophy*, 68(1), 5-20.

Froese, Tom et al. (2007). "Autonomy: a review and a reappraisal," In: e Costa, F. Almeida et al. (eds.) *Proceedings of the 9th European Conference on Artificial Life*, Springer-Verlag, 1-13.

Müller, Vincent (2012). "Autonomous cognitive systems in real-world environments: Less control, more flexibility and better interaction," *Cognitive Computation*, 4(3), 212-15.

Rao, Anand et al. (1995). "BDI Agents: From Theory to Practice," *Proceedings of the First International Conference on Multiagent Systems,* 312-319.

Schmidt, Colin et al. (2006). "Robots, Dennett and the autonomous: a terminological investigation," Minds and Machines, 16(1), 73-80.

Taylor, James (eds.) (2005). *PERSONAL AUTONOMY New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*, Cambridge University Press.

栗原聡ほか (2017).「汎用 AI 実現のための鍵となる自律性とマルチモーダル性についての考察」, 第 31 回人工知能学会大会発表原稿.

新出尚之 (2010).「自律エージェントの論理モデル」,『人工知能学会誌』, 25(3), 419-428.

長倉克枝(2017).「『人工知能脅威論』が覆い隠す、本当の問題は何か？――日仏の研究者が議論」, ハフィントンポスト日本版, http://www.huffingtonpost.jp/katsue-nagakura/ai-ethics_b_16904442.html, (2017/11/26 参照).

中島秀之(2013).「レクチャーシリーズ『人工知能とは』第一回 人工知能とは(1)」,『人工知能学会誌』, 28(1), 139-143.

錦田将悟 (2016).「自律ロボットのための情動に基づく意思決定システムに関する研究」, 山口大学大学院理工学研究科博士論文.
人工知能学会 (編) (2017).『人工知能学大事典』, 共立出版.

(東京大学)