人工知能における「自律性」とはなにか

山﨑 かれん（Karen Yamazaki）
東京大学大学院総合文化研究科

「人工知能」というワードを打ち込み、インターネット上を検索しよう。技術革新を期待する記事と並んで、危険性を喧伝し、人類への脅威を煽るような言説が目に入る。SF 映画に出てくるような、優れた「人工」知能を持つロボットを想像しているのだろうか。彼らが、われわれ人間を襲ってくるとでも？

「自律性」というのが、そうしたときに言われるひとつの標語である。人工的に生まれた主体が自律的に動くことが問題にされているのだ。一方で、現在の人工知能研究では、その自律性の付与を目指す開発がいくつかの理由から実際に行われている。しかし、開発の現場においても、危険を唱える言説においても、「自律性」の意味は明らかでないように思われる。そこで本発表では、人工知能における自律性がどのようなものであるかを検討したい。

まずは人工知能研究の現場で、自律性という言葉がどのように使われているかを見てみる。一口に人工知能研究と言っても当然ながらそれは多岐にわたり、網羅的な言及は難しいが、主体としての人工知能のふるまい方に注目した研究にかんして特徴を抽出することを試みたい。次に、対照として哲学的議論のうちの「人格的自律性（personal autonomy）」を挙げる。人格的自律性は主に人間の自律性について議論されるときに言われる概念であり、自己決定にまつわるものである。

自律性概念の概観から、自律性という言葉を使う際に二つの観点があることを指摘する。一つ目を「ふるまいの自律性」と呼びたい。これは、主体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。ただし、ふるまいの自律性は、どれだけ制御を受けているかという程度の問題から評価される。二つ目は「心的な自律性」と名付けておく。こちらの立場は主体自身の意図や目的選択といった心的状態がふるまいを生み出しているという点から自律性を定める。

この二つの自律性は、違う意味をもっているのにも関わらず、なぜ同じ「自律性」という言葉で表されるような事態になっているのだろうか。それは人工知能研究で使われる「自律性」という言葉の出自によると私は考えている。人工知能の開発目的はさまざまであるが、人間を参考にしてその機能を作り出すということは多く行われている。自律性についても、意識的にせよ、無意識的にせよ、人間のもつ自律性を参考にして作られようとしているのではないだろうか。そして、自律性の使用に揺れがあるのは、そのときにどのような自律性の特徴を取り出しているのかに違いがあるからだ。つまり、人間の自律性にかんして、ふるまいの自律性は外部からの制御に着目し、心的な自律性はふるまいを生み出す主体の心的状態に着

目しているのである。

ふるまいの自律性と心的な自律性はいかなる関係にあるのだろうか。心的な自律性では、欲求や意図といった心的状態がその主体のふるまいを生み出す。これは、心的状態がふるまいを制御していると言ってよく、外部からの制御によらないふるまいの自律性を実現する。では、ここで問題にしたいのは、心的な自律性なしに、外部からの制御にまったくよらないような、「完全な」ふるまいの自律性は、実現されるのだろうかということである。本発表では、「完全な」ふるまいの自律性には心的な自律性が必要であり、心的な自律性なしには「完全な」ふるまいの自律性は実現できないだろうと結論する。

ここまでの自律性概念の検討を通じて得た洞察から、それぞれの開発目的に応じた人工知能の「自律性」を考えたい。「完全な」ふるまいの自律性は心的な自律性なしには実現されえないと上で述べたが、一方では、ふるまいの自律性は制御の程度から定められるとも先に述べた。さまざまな研究の方向性がある人工知能研究において、一意に「自律性」を定めることは乱暴である。環境の変化に対して柔軟に行動することができるなどといった実用性を高める上では、使用目的に応じたさまざまな程度のふるまいの自律性を与えるという方向性を考えることは有効である。しかし、人間のような知能を実現するであるとか、人工知能を構築することで人間の知能を理解しようという場合ではふるまいの自律性だけでは足りないのである。

最後に、本発表では詳しく言及しないが、人間のような自律性（ここでは人格的自律性を想定している）を人工知能で実現するためには、本発表で述べた心的な自律性の定義ではまだ不十分であるということも添えておく。