日本科学哲学会第 50 回年次大会

「人間のような」自律エージェントはどのようなものであるか

山﨑かれん（Yamazaki Karen）
所属 東京大学大学院

昨今の人工知能研究に際して、「自律性」はひとつのキーワードになっている。予測の難しい実世界において、あらかじめ組み込まれたプログラムだけで行動することは難しい。激しく変化する環境にあっても効果的に動作する機械の構築を目指して、自律性の付与が目指されている。そうした実用性の向上を目指した開発の方向がある一方で、人工知能研究には「人間のような知能を作る」という動機もある。そこでは、人間の知能に不可欠な性質だと考えられるような自律性を人工知能に持たせようという試みがなされている。

本発表では、まず後者の「人間のような」知能を実現する人工知能が持つべき人間のような自律性とはどのようなものかを考察する。その上で実際の人工知能研究を吟味し、そこではたして人間のような自律性が実現されようとしているのかどうかを精査する。まず、人間のような自律性であるが、これについては、「人格的自律（personal autonomy）」の議論をもとにそれを特徴づける。人格的自律については、1970 年代のフランクファートなどを皮切りに多くの論者によって議論されてきた。様々な立場があるが、いずれも行為者が自分の心的状態を自分のものとして自覚ないし承認することを核心的な要素として行為の自律性を定義している。

人工知能研究の一領域であるエージェント研究は、自ら判断し行動するような自律エージェントを構築しようとしているが、そのような自律エージェント研究では、まさに主体の心的状態に着目した研究が行われている。たとえばブラットマンの理論に基づいた BDI モデルの研究がそうである。BDI モデルは信念・願望・意図という心的状態に基づいて、目的を選択しそれを達成するための行動決定を行うエージェントを構築しようとする。

しかし、そのようなエージェントがその願望や意図によって行動決定できたとして、それで人間のような自律性を持っていると言えるのだろうか。人格的自律の議論から導き出された自律性の定義からすると、それだけでは不十分だと考えられる。人間のような自律性のためには、選択する目的がまさに自らの行動の目的であるという自覚、あるいは承認が必要になるのだ。人間のような自律エージェントの構築のためには、そうした自覚や承認を可能にする仕組みを付与することも考えなくてはならないだろう。したがって、現在の実際の人工知能研究では、まだ人間のような自律性を実現する試みがなされているとは言えないのである。