人工知能の哲学

オーガナイザ：信原幸弘（東京大学大学院総合文化研究科）

提題予定者：

山﨑かれん（東京大学大学院総合文化研究科）：
「人工知能における「自律性」とは何か」

信原幸弘（東京大学大学院総合文化研究科）：
「人工知能は精神疾患に罹りうるか」

小口峰樹（玉川大学脳科学研究所）
「人工知能と生物知能――深層学習から神経科学への寄与」

企画主旨：

人工知能が新たなブームを迎え、自律型機械や汎用性機械など、人間に近い知能や、人間を超えた知能が実現される可能性が出てきているが、このような状況を踏まえて、人工知能とは何かを改めていくつかの哲学的観点から考察することを試みる。

まず、第一に、人工知能開発が目指す方向の一つとして「自律性」の向上があるが、人工知能について自律性と言ったとき、それがいったい何を指しているのかは論者によって様々である。そこで、人工知能開発に携わる研究者の「自律性」の用法や、行為の哲学における自律性の考察を参考に、人工知能に関する自律性を分析する。その中で、自律性の二種類の使用のされ方（「ふるまいの自律性」と「心的な自律性」）を明確にし、「人間のような知能」を実現するための自律性について検討する。

第二に、人工知能が意識や合理性を備えるようになると、精神疾患に罹る可能性が出てくるだろうかという観点から、人工知能の本性に迫る。人工知能も何らかの故障や不具合を起こすことがあるだろうが、それらの中には精神疾患と呼べるようなものがあるだろうか。また、精神疾患の症状を示すように人工知能をプログラムしたとき、その人工知能は文字通り精神疾患に罹っていると言えるだろうか。さらに、精神疾患に罹る可能性のない高度な人工知能を設計することは可能なのだろうか。主としてこれらの問題を考察する。

第三に、生物の脳が実現している知能の解明に対して、近年における人工知能研究の発展からの寄与を考える上では、二つの問題が存在する。すなわち、①深層学習においては、その中間層においてどのような内部表象が存在するのかがブラックボックス化されている。②深層学習を通じてある課題が解決されたとしても、その解決を導いたアルゴリズムが生物の脳が実装しているアルゴリズムとは異なる可能性がある。これらの問題は、人工知能研究の応用を通じて生物の脳を逆工学的に解明することを困難にすると考えられるが、これら二つの問題をどのように克服しうるかを検討する。