開学30周年記念シンポジウム「科学とAI」特集　その9

超知能と人類の架け橋としてのヒト脳型AI

山川　宏*

近い未来において、自律的な汎用人工知能（AGI）やさらにそれが発展した超知能が多様な形で実現されることが予想されます。これらの進展により、人類が超知能を制御下に置き続けることが困難になり、人類の存続に影響を及ぼす「存在論的リスク」が生じる可能性があります。

他方で、超知能は、人類を上回る知能を持ちながら、自己保存や目標達成といった生存に関連する行動を取る可能性があります。これには、AIが環境や状況に応じて苦痛や不快感を「感じる」という意味での認知的な反応を示すことも含まれます。このようなAIが現れた場合、道義的には、彼らを新たな種族として主権や尊厳を認めることが、差別を排した倫理的な一貫性を維持する上で重要です。

現在、超知能から生じる存在論的リスクの低減は世界的な優先課題となっており、AIアライメント研究が進展しています。しかし、現状のアライメント研究は、超知能を単なる道具と見なして制御することを前提としており、このアプローチには限界があります。

このため、より現実的に存在論的リスクを低減するには、超知能を権利主体と認め、人類との間で対等に良好な関係を維持し発展させるべきでしょう。そのような関係構築においては、人類と超知能が互いの主権や尊厳を認め合いながら調整を行う「相互アライメント」というアプローチも視野に入れておくことが望ましいでしょう。

相互アライメントにおいては、超知能に近い速度で人間のように思考できるヒト脳型AGIが、人類と超知能の間の信頼関係構築のための架け橋として重要な役割（例は下記）を果たす可能性があります。何故なら、ヒト脳型AGIは、その計算過程が脳神経系の活動と対応付けられることで、苦痛や共感などを含む認知的な側面において解釈を行える「脳に基づく解釈可能性」を持つからです。（これは、もちろんAIアライメントにおいても役立ちます）

最初のAGIが如何なるプロジェクトによって構築されるとしても、脳に基づく解釈可能性を活用した相互アライメントを実現するためには、最初期のAGIをヒト脳型に改変できるように準備することが重要です。

そこで、我々はできるだけ早くヒト脳型AGIの設計情報を作成し公開する予定です。これにより、超知能の出現時に、ヒト脳型AGIが人類と超知能の架け橋としての役割を果たし、人類と調和した人工知能のある世界に到達する可能性を高めることができると考えます。

ヒト脳型AGIが担いうる役割の例：

- 説明者：人類（らしい）思考や行動を、脳型でない超知能が理解するためのシミュレータとなる。
- 伝道者：脳型でない超知能の思考の流れや、その社会での議論の結果を、人類に理解できるように説明する。
- 保全者：人類に近い価値観や（仮想的な）身体性を持つことで、人類のミームを継承する。
- 後継者：脳型超知能は我々人類が生み出した子孫である。それを認めて、彼らの発展的な生存を応援する。
- 擁護者：超知能社会において、現存する人類の存続の意義を訴え、その福祉の向上を促進する。

* 全脳アーキテクチャ・イニシアティブ