技術の媒介理論による AI システムの責任ギャップ問題再考

―理論的 AI 倫理と設計の接続―

秋葉　豊（Yutaka Akiba）

名古屋大学大学院情報学研究科

昨今の生成 AI の台頭により、様々な分野において AI システムが変革をもたらすことが現実的になりつつある。特に自動運転や自律型致死兵器システム（LAWS）等の、作動に人間が介在しない自律的な AI システムは、既存の慣習を大きく変えかねないため注目を集めている。このような自律的な AI システムがもたらす懸念の 1 つとして、責任ギャップ（Responsibility Gap）と呼ばれる問題がある。これは、自律的な AI システムが、それ自体がもたらす帰結に対して因果的に責任を負うものの、道徳的・法的に責任を負うことはできないため、責任を負う主体が不在になる可能性がある、という問題(e.g. Danaher, 2022)である。責任ギャップにはいくつか種類があるが、時間軸に焦点を当てると、既に起きてしまった出来事に関する後ろ向き（Backward-looking）責任ギャップと、将来的に起こり得る出来事に関する前向き（Forward-looking）責任ギャップが存在する（Nyholm, 2023）。

責任ギャップ問題に対する解決策の検討は、後ろ向き責任ギャップの回避を目的とし、任意の人間が責任を負うための必要条件を提案することが現在主流となっている。AI システムがもたらす帰結に対して、直接的な仕方で人間が責任を負うことは困難であっても、間接的な仕方で人間が責任を負うことが可能だと考えられているのである。例えば AI システムの設計要件として、任意の人間の重要な道徳的理由と動作環境における重要な要素の両方にシステムが応答することと、システムの作動がもたらす帰結の原因を因果的に任意の人間まで常にたどることができる可能性を保証することが満たされることによって、任意の人間が責任を負うことができるという提案がある（Santoni de Sio, & Van den Hoven, 2018）。また自律的とされている AI システムが実際のところ、常に人間により監督されており、むしろ人間の関与なしには AI システムが作動することは不可能であるため、AI システムがもたらす帰結に究極的には人間が責任を負うことができるという指摘もある（Nyholm, 2018）。

上記の対応策の議論にはある前提がある。それは、AI システムは部分的な行為者性を発揮するのみであり、道徳的行為者たりえず、あくまで人間によって用いられる道具にすぎないとみなされていることである。発表者はこの前提が後ろ向き責任ギャップへの対応、すなわち既に起きてしまった出来事の責任の所在の特定には有用であるが、前向き責任ギャップへの対応、すなわち将来的に起こり得る出来事の責任を任意の人間に負わせるためには不十分であると考える。というのも、AI システムによって将来的に生じうる不利益を回避するという視点が欠けているからである。代替の前提として、AI システムが単なる道具に留まらず、人間の認識・行為に働き掛けうるという媒介的技術

観を提案する。この前提に拠って立つことによって、将来的に起こり得る出来事に対して、責任を負う任意の人間をサポートする AI システムというイメージを描くことができる。そして責任ギャップ問題の議論を責任ある研究・イノベーションの議論と接続することが可能になる。

本発表では 2 つのアプローチから前向き責任ギャップへの対応を試みる。1 つ目のアプローチは、アクターネットワーク理論やポスト現象学等を含む技術の媒介理論である。技術の媒介理論は、人間の認識や行為を技術が媒介すると考え、人間の道徳的な慣習においても技術が大きく影響を与えている（e.g. フェルベーク, 2014）と捉える。責任についても、人間と技術の連合体において立ち現れてくる（Hanson, 2009）と捉えられるため、現在の AI システムが人間の道徳的な慣習に与える影響を技術の媒介理論は如実に表現することができる。そして人間が責任を負うという行為が AI システムによって媒介されていると捉えることによって、設計を単に満たすべき要件としてだけではなく、より積極的・能動的に人間に責任を負わせるアーキテクチャにするという方向性を切り開くことができる。

もう 1 つのアプローチは、Ethics by Design（以下 EbD と略）である。EbD とは、AI システムに自身の決定の倫理的側面について推論する能力を与える方法・アルゴリズム・ツールに関する設計思想である（e.g. Dignum et al, 2018）。EbD は規制とは異なる仕方で倫理的な AI システムを実装する方法として注目されており、プライバシーや公平性等について実践が行われている。EbD を用いることによって、責任について考慮する（≠責任主体）AI システムが、使用において人間の責任をサポートする設計へのヒントを示すことができる。

参考文献

フェルベーク、ピーター＝ポール、(2015)、『技術の道徳化―事物の道徳性を理解し設計する』、鈴木俊洋訳、法政大学出版局。

Danaher, J. (2022). Tragic choices and the virtue of techno-Responsibility gaps. *Philosophy & Technology*, 35(2), 26.

Dignum, V. et al. (2018). Ethics by design: Necessity or curse?. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 60-66.

Hanson, F. A. (2009). Beyond the skin bag: On the moral responsibility of extended agencies. *Ethics and information technology*, *11*, 91-99.

Nyholm, S. (2018). Attributing agency to automated systems: On human-robot collaborations and responsibility-loci. *Science and Engineering Ethics*, 24, 1201–1219.

Nyholm, S. (2023). *This is technology ethics: An introduction*. John Wiley & Sons.

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5, 15.