

## 2. エッセイ集

# 3 機械の中の道徳

久木田水生 (名古屋大学)

### 道徳的な機械を作る試み

近年、道徳的に判断し行為する人工のエージェントを作る試みが行われており、「機械倫理」、「人工道徳」などと呼ばれている。その背景にあるのは、自律性の高い人工のエージェントがオープンな環境で不特定の人間と接するような状況では、それが人間の感情や健康や財産を害さないように振る舞う必要があるという認識である。

たとえば Michael Anderson らの開発しているシステムは介護士が直面する倫理的なジレンマ状況において、実際に専門家が下す判断の事例から規則を導出するというもので、手法としては帰納論理プログラミングを使っている<sup>1)</sup>。ここから分かるように、彼らのプロジェクトは「人間が従っている規則を明示化し、それを機械に実行させる」というアプローチをとっている。つまり彼らのアプローチは古典的な記号的人工知能、論理的人工知能のパラダイムに沿っている。

論理的人工知能のパラダイムは人工道徳に従事する研究者の多くに共有されている。その理由の1つは西洋哲学における伝統的な道徳概念である。西洋哲学では人間の知性や理性は身体や感情とは切り離されたもので、かつ個別の状況に依存しない普遍性を持つものとされてきた。そして西洋の倫理学の主流派は道徳的判断も知性や理性によってなされるものとする。そのため倫理を機械的に実装するには論理的アプローチを使うのが自然なことだったのである。

しかし実際に人間の道徳はアルゴリズムに沿った計算のようなものなのだろうか。20世紀の後半に人工知能の発展が明らかにしたことは、人間の知能は何も論理的なものばかりではなくもっと多様な要素からなり、多様なやり方によって実現されている、

ということだった。そしてこれは合理的な推論や計算を知性のコアと考えていた西洋哲学の伝統的な常識を揺るがすことになった。道徳についても同じようなことが起こりはしないだろうか。

### 科学が明らかにする道徳性

近年、動物行動学、神経生理学、心理学などの分野において、道徳性についての研究が盛んに行われている。そしてそこで得られた結果は、従来の理性中心主義的な理解とは異なった道徳の姿を示唆している。特に顕著なのは道徳的意思決定における感情の重要性である。

2つ例を挙げよう。神経科学者の Antonio R. Damasio が報告しているある患者は、脳腫瘍の手術の後、日常的な意思決定に困難を覚えるようになったという。記憶や推論の能力には問題がなかったが、よく調べると彼には感情を感じる能力が失われていることが分かった<sup>2)</sup>。ここからは感情が意思決定において重要な役割を担っているという可能性が示唆される。これが正しいとすれば、道徳的な意思決定においても感情は重要な役割を果たしている可能性が高い。

神経経済学を提唱している Paul J. Zak によれば、私たちの道徳性はオキシトシンやテストステロンなどのホルモンのレベルと密接にかかわっている<sup>3)</sup>。オキシトシンのレベルが上がることは他者への信頼、寛容さ、協力性を増大させる。一方テストステロンは敵意や攻撃性を増大させる。オキシトシンのレベルは相手から自分への信頼を示されたり、相手と親密なコンタクトを取ったりすることで上昇する。そこで道徳的な行動を促進するために Zak が推奨する「処方箋」は「1日8回のハグ」である。

## Moral mod Science (道徳を科学で割った余り)

科学が道徳性についてさまざまな新しい知見をもたらす一方で、人工知能とロボット工学においては道徳性を機械に実装する試みが行われている。しかしながら今のところこの2つは互いに交わずに進められている。道徳を機械に実装する試みが、限られた状況でそして純粋に実用的な目的で行われる限りでは、それでも良い。しかし道徳的な機械を作ることによって人間の道徳性を明らかにすることを目指すならば、道徳性についての近年の科学的知見を反映した設計を考えることが必要である。

そこで私はさまざまな分野の科学者、倫理学者、人工知能研究者の協力を仰いで2016年度から「Moral mod Science」という研究プロジェクトを立ち上げた。自然科学において得られた道徳性についての新しい知見に基づいて道徳概念をアップデートし、それを反映した人工的なシステムを作ろうというのが目論見である。実用的なシステムを作るといよりはシステムを作ることによって道徳性の理解に役立てることができれば良いと考えている。

道徳は中々に捉えがたいもので、何を道徳的に良いとみなすかは文化や個人によって異なっている。しかし科学と人工知能の発展が道徳についてのより良い理解を与えてくれるかもしれない。これは20世紀の後半に知能について起こったことと平行である(図-1参照)。私たちはこれが人間の道徳性についてすべてを明らかにするとは期待していない。

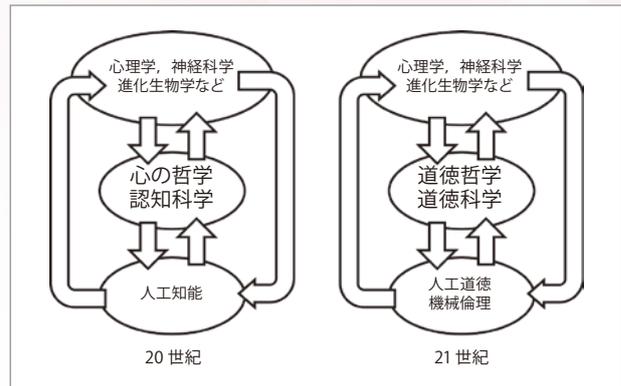


図-1 人工知能と人工道徳

道徳性には科学で説明しきれずに残されるものがあるだろう。それが何かを知り、それをどう扱えば良いのかを考えたい。「Moral mod Science (道徳を科学で割った余り)」という言葉はそのことを示しているのである。

### 参考文献

- 1) Anderson, M. and Anderson, S. L. : Machine Ethics : Creating an Ethical Intelligent Agent : AI Magazines, 28(4), pp.15-26 (2007).
- 2) Damasio, A. R. 著, 田中三彦 訳: デカルトの誤り, ちくま学芸文庫 (2010).
- 3) Zak, P. J. 著, 柴田裕之 訳: 経済は「競争」では繁栄しない—信頼ホルモン「オキシトシン」が解き明かす愛と共感の神経経済学, ダイヤモンド社 (2013).

(2016年6月30日受付)

久木田水生 minao.kukita@is.nagoya-u.ac.jp

2005年に京都大学で博士号(文学)を取得。2014年より名古屋大学大学院情報科学研究科准教授。研究会「AIR:人工知能が浸透する社会を考える」、「ロボットの応用哲学」メンバ。専門は哲学・倫理学。