人工知能と人間，どちらが合理的？

立命館大学総合心理学部 教授
服部雅史（はっとり　まさし）

Profile―服部雅史
1996年，北海道大学大学院文学研究科行動科学専攻博士後期課程単位取得退学。博士（文学）。2016年より現職。専門は認知心理学，思考心理学。著書は『基礎から学ぶ認知心理学』（共著，有斐閣），『思考と推論』（監訳，北大路書房）など。

「散歩に出かけるために乗合馬車に乗った。その階段に足を触れたその瞬間，（中略）突然わたくしがフックス関数を定義するに用いた変数は非ユークリッド幾何学の変換とまったく同じである，という考えがうかんで来た。馬車内に座るや否や，やりかけていた会話をつづけたため時がなく，検証を試みることをしなかったが，しかしわたくしは即座に完全に確信をもっていた。」（ポアンカレ，1908/1953, p.58）

いつの日か，AI（人工知能）がポアンカレのような創造的な仕事をすることができるだろうか。囲碁AIは人間を超えたが，数学でそれが実現する日は来るのか。

数学は演繹である。つまり，前提を正しいとしたときに正しく導かれることがらだけで構成される体系である。本来，こうした演算はコンピュータの得意領域である。しかし，数学が演繹的体系であっても，演繹だけで数学を「作る」ことはできない。原理的には網羅的探索で必勝できるはずの囲碁が，実際の勝負で勝つにはさまざまな工夫が必要なのと似ている。

では，創造性に何が必要か。ポアンカレは，数学の証明に審美的感受性が不可欠と考えた。つまり，発見には，斬新な発想と正しく評価できる「眼」が大切ということであろう。また彼は，意識的で努力を要する認知過程と，無意識的で自動的な過程の両方が必要である旨を先駆的に論じている。以下では，認知過程と合理性の概念を基軸として，人工知能と人間の「賢さ」について考えたい。

AlphaGoに足りないもの

いま，ディープラーニングが熱い。音声・画像認識や囲碁など，これまでできなかったことがコンピュータでできるようになっているのを見るにつけ，Googleの宣伝効果を割り引いても，これが画期的なAI技術であることは疑う余地がない。世界最強の棋士を破っただけでなく，定石から外れた「棋士の理解を超える着手の連続」（朝日新聞，2017.6.2）が見られる事実からも，囲碁AIはもはや完全に人間を超えたと言えよう。

しかし気になることがある。プロ棋士の大橋拓文氏が「AlphaGoはよい手を打つがそれを言語化できない」という旨のことを話されていた（日本認知科学会第34回大会招待講演，2017.9.14）。人間は手に「意味」を見出して打ち進めるが，AIにはそれがない。妙手の意味を他者に伝えることができないのである。

これは，人間にとっては当たり前の「メタレベルの解釈」がAIにないからである。ポアンカレは自分がよいアイデアを思いついた状況を振り返って，それに意味づけをしたり，自分自身が考えた過程を分析したりした。この過程がAIにはない。この点が，人間とAIの決定的な違いではないか。

もちろん，それはルーチンを組み込めばよいだけだという反論はあるだろう。実際，Googleは，画像の認識のみならず，画像からその説明文を自動的に生成するImages To Textを開発している。しかし，これは何か違う気がする。おそらく，対象レベルの認知とメタレベルの認知が，ほぼ同時に自動的に発動し，しかも両者が不可分な過程として存在することが重要なのではないだろうか。

簡単だが難しいこと

メタ認知とは，認知についての認知を指す。これは対象レベルの認知よりも一段上の認知を指し，無意識的認知の上に位置する意識的認知のさらに上に位置づけられる。しかし，私は，メタ認知活動の多くはむしろ潜在的（無意識的）で自動的な性質を持つと考えている。さらに言えば，ひょっとすると，潜在的メタ認知がなければ本当の知性は創発しないのではないだろうか。

たとえば，私たちの雑談を考えてみよう。雑談中に話題はどんどん変わる。話題に応じて発話内容を決めるわけだが，思いついた内容がすべて適切というわけではない。相手の好みや傾向を考え，反応を予想しながら内容を選定する。

相手についての知識（0次的信念）だけでなく，相手が自分をどう見ているか（1次的信念），相手について自分が何を知っていると相手が思っているか（2次的信念）も話の内容の適切さに影響する。

会話中は，メタ認知がフル活動する。本人ははっきり意識しないが，相手の表情だけでなく，相手としゃべる自分の様子や，相手に映る自分の表情がどうであるかについても，きっとモニタしているはずである。また，相手の表情が曇ったら，さりげなく話題を変えるといったコントロールも重要となる。多かれ少なかれ人間なら誰もが行っているが，これらのことをAIに教えるのは容易ではない。

目標は一つではない

もう一つ重要な側面は，目標多重性である。通常，雑談に明確な目的はないが，会話をすることの前提として複数の欲求を想定することができる。たとえば，相手と仲良くなりたいとか，自分の気持ちを理解して欲しいといった欲求である。こうした目標は時に競合する。自分のことばかり話していると，相手は退屈するかもしれない。最終的に出力する行動は一つなので，複数の目標の統合のしかたが常に問題となる。

多重目標の統合というタスクは

複雑にみえるが，実は人間は案外すんなりやっているのかもしれない。ここでは，その考えの根拠となる現象を一つだけ挙げておく。

ある実験で，「心臓が健康だと，冷水耐性が高く，平均寿命も長い」と教示すると，これ以上冷水に腕をつけていられないと感じるまでの時間が35パーセント伸びた（Quattrone & Tversky, 1984）。参加者は，心臓タイプが冷水耐性と寿命の共通原因であることを知っていたので，これは因果推論の誤りである。しかし，自己欺瞞が自己効力感を高めて幸福感を増すのなら，それを目標の一つとすることには意味がある。

この実験は，三つのことを示唆している。第1に，人は自ら設定する目標を実験室に持ち込むこと，第2に，そうして多重化した目標を統合する認知処理が自動的・無意識的に発生すること，第3に，一見「不合理」に見える行動は，目標の多重性を考慮すると必ずしも不合理ではないことである。第3の点は，合理性の意味，そして，AIと人間の違いを考えるための論点を提供する。

合理性と意識と魂

私たちの認知にはバイアスがあり，それが時には，不合理なエラーの原因となる。そうしたエラーは，課題で要求される目標に自ら持ち込んだ目標を混入させることによって起こると考えてみよう。そこで問題となるのは，果たして，そうすることは「誤り」であるかどうかである。

単一の目標に焦点化することは，コンピュータの得意技である。しかし，そうして硬直化したシステムは，二つの干草の間でどちらに行くか迷って餓死したビュリダンのロバになる危険性がある。ポアンカレは，よいアイデアを思いついても，すぐに正しさを確かめずに馬車の中で雑談を続けた。考えが行き詰まったら散歩に出かける。話を面白くするために話題を飛躍させる。こうしたコントロールができなければ本当に知的であるとは言えない。

もし，碁を打つAIが対局の最中に碁を中断して雑談を始めたら，それはエラーかもしれない。しかし，その時こそ，いわゆる「シンギュラリティ」がもたらす脅威に対するホーキング博士の警告（Independent紙，2014.5.1）について，少しは真剣に考え始めてもよいかもしれない。さらに言えば，そうした知的なシステムは，結局，「意識」を必要とするのではないかと私は考えている。

多重目標の統合において重要な点は，目標の多くの要素に意識が関係していることである。ここで，意識には，自分自身をかけがえのないものとみなして尊ぶことを可能にするという特別な機能がある（ハンフリー，2011/2012）と仮定してみよう。すると，自意識を介した魂のウェルビーイングという多重化された目標の達成は，生物としての知性に不可欠なのかもしれないという思いが頭をよぎる。生の尊重は，死の脅威と表裏一体である。ならば，真のAIは，『2001年宇宙の旅』のHAL 9000のように，やはり自らの終焉に心を乱すのではないだろうか。

文　献

N. ハンフリー／柴田裕之（訳）（2012）『ソウルダスト：「意識」という魅惑の幻想』紀伊國屋書店

H. ポアンカレ／吉田洋一（訳）（1953）『科學と方法』岩波書店

Quattrone, G. A. & Tversky, A. (1984) *Journal of Personality and Social Psychology, 46*, 237-248.