2408HumanModeling.pdf

2024.8 ブログ：「ＡＩを用いた詳細価値観モデリング」を読んで、の詳細
（→ http://www.1968start.com/M/blog/index3.html#2408b ）

「ＡＩを用いた詳細価値観モデリング」を読んで

中所武司

■このエッセイのきっかけ

下記の人工知能学会誌の解説論文について、冒頭の「認知モデルは…計算モデル…」に興味を持ち、読んでみた。

・「生成 AI 時代における認知のモデリング」特集
人は人の理解をやめるか
― AI を用いた詳細価値観モデリングと精密行動予測の可能性 ―
人工知能, Vol. 39, No. 2, pp. 230-239 (March. 2024)

■内容の要約とコメント（→★）

1. はじめに

・認知モデルは、人間の外界の理解，予測，意思決定についての計算モデルである．
本稿では、主に他者モデルと AI の関係について述べるが，その前に、
現象を数理的に理解，説明するという応用的価値が問われる現状について述べる．

・ニュートンは落下現象を体系的に説明した最初の人（**生体ニューラルネットワーク**）である．
物体は重心に全質量が集中し大きさをもたない質点である，という非現実的な仮定を置き、
モデルが複雑になるのを避けるため、摩擦や空気抵抗はないものとする．

・カルマンフィルタは，放物運動を予測する場合，ニュートンの運動の法則を基本として，
ノイズやバイアスを考慮したモデルを用いることで，より現実的な予測を可能にする．

・しかし，野球選手は、経験と練習を通じて**生体ニューラルネットワーク**の中に学習された
運動モデルを用いて落下地点を予測している．

・近年，**人工ニューラルネットワーク**が、人を凌駕する予測能力を獲得しつつあり，
予測が重視される工学では，数理モデルの価値が薄れるかもしれない．

→★「**生体ニューラルネットワーク**」と「**人工ニューラルネットワーク**」の表現が興味深い。
後者は前者を大胆に単純化したモデルであるが、数理モデルに比べれば近い関係ではある。

・本稿では，この現状を踏まえ，モデルを用いて他者を理解することの意義について述べる．
合理的行為者の生成モデルを用いた心的状態推論の計算モデルである**ベイジアン心の理論**の

研究をもとに，AI が個人の効用関数を精緻に反映したインプリシットな生成モデルをもち，
人の振舞いを正確に予測できるようになる可能性について述べる．

→★この解説のタイトル「人は人の理解をやめるか」とは、
　他人の理解をＡＩが肩代わりするようになるかもしれない、ということかな。

2. 他者モデル

・他者モデルは、他者の振舞いの理解と予測に用いられるモデルである．

・**目的論的因果**では，事物が終端状態の達成のためにあると考えるので、
終端状態は心的な表象となり、心的状態である「目的」が人の行動を駆動していると捉える．

・観察した行為からその原因となる目的を推論する**仮説推論**（abduction）では、
観察された事象から最も可能性の高い原因や説明を導き出す．

・目的の推論では，観察している状況と合理性が，選択される目的の妥当性を測る基準となる．
行為の原因を目的に帰属させることで，新規な状況における行為の予測が可能になる．

・心の理論が推論対象とする心的状態は，
（1）世界の状態に対する認識である信念や知識と，
（2）世界の状態に対する価値である願望，目的，意図，選好である．

・信念と知識は，客観的な事実や現実に関する個人がもつ見解や認識，理解である．
エージェントは，部分的にしか世界を観測できないうえに，恣意的な情報の取捨選択や希望，
バイアス，センサの特性によって，事実と異なる認識をもつ．

→★この解説には「エージェント」という記述が２０回以上出現するが、明確な定義はない。
　著者紹介にある「ヒューマンエージェントインタラクション」の WIKI での説明の中には、
　以下の定義がある。おそらくこの分野の専門用語と思われる：

（参考：引用元）ヒューマンエージェントインタラクション - Wikipedia
『エージェントとは、
コンピュータのソフトウェアか（ロボットのような）ハードウェアかによらず、
一定程度自律的に学習・推論する能力をもった主体のことを指す。つまり、
知的エージェントあるいは自律エージェント のことを意味する。
またエージェントは人と人の間を媒介する存在（mediator）としての役割・機能も含む』

・子供の心の理論の発達をテストするために用いられる誤信念課題―サリー・アン課題では，
他者の部分観測性（限られた視点）に基づいた，「誤った」信念帰属がテストされる．
サリーがボールをバスケットに入れて部屋を出た後，アンがそのボールを箱に移動させる
シーンを見せ，戻ってきたサリーがボールがどこにあると信じているかを問うという課題で，
ボールがある箱ではなく，サリーが誤って信じているバスケットと答えるのが正解である．

・GPT-4（2023 年 6 月のバージョン）がサリー・アン課題の 20 のバリエーションに対して 60％正解できることが，16 の厳密なテストプロンプトを用いて示されている．

・もう一つの心の状態である，願望，目的，意図，選好は**効用関数**をもとに定義される（図 1）．効用関数を精密に同定することは他者の理解と行動予測のための重要タスクとなる．

・合理性の仮定は行動説明の核心である。

3. 心の理論の計算モデル

3･1 行動観察による選好と信念の同時推論

・近年，観察した行動から行動主体のもつ心的状態を推論する計算論的モデルが提案されている．部分観測マルコフ決定過程**エージェント**の生成モデルを用いたベイズ推論であるが、環境の完全な状態を観察できず，センサからの観測を通じて環境の知識と信念を得る．与えられた知識，信念および自己の効用関数により最適な行動を選択する．

・隠れ状態である心的状態の推論は次のベイズ推論の式で表現できる．

$$p(m \mid a, x) \propto p(a \mid m, x)\, p(m) \quad (1)$$

＊m は直接観察できない信念，願望，意図，選好などの**心的状態**，
＊x は客観的に**観察可能な状況**，
＊a は**観察可能な行為者の動作**や表情出力である．

・この式は，もし人がどのような状況でどのような心的状態がどのような行動を出力するかの知識 $p(a \mid m, x)$ をもっていれば，事前分布を仮定することで，ベイズ推論によって，ある状況 **x** で相手の行動 **a** を観察したとき，相手の心的状態 **m** を推論できることを意味する．

・ある文献では，人が観測した行動のみから他者の価値と信念を同時推論できること，および，ベイズ推論が心の理論の計算モデルとして妥当であることを示した．

【その文献で用いられたタスク】
観察者は A の行動を観察し，A が行動開始時に Y の位置にどのフードトラックが停車していると信じていたかと，A の食べ物の選好を回答することを求められる．
＊X の位置に寿司のフードトラック，
＊Y の位置にハンバーガーのフードトラックが停車している．
＊観察者は，Y の位置のトラックは確認できるが，S の位置にいる A には見えない．
観察者は、知覚の部分観測性を考慮して A に信念を帰属させなければならない．
＊A はスタート地点 S から行動を開始し，X の位置の寿司のフードトラックの横を素通りし，P の位置まで移動して Y の位置にあるハンバーガーフードトラックを確認した後に，最終的に X の位置まで移動する．

・ハンバーガーに向かわなかったので，ハンバーガーは寿司よりも低い選好とがわかる．
Aは、カレーが一番好きで，カレーのトラックがYの位置に停まっていると信じて行ったが停まっていなかったために，2番目に好きな寿司に向かった，という解釈ができる．

→★この推測はわかりやすい。寿司の店を通り過ぎたので、寿司が一番好きではないとわかり、ハンバーガーの店を見て戻り、寿司を購入したので、ハンバーガーより寿司が好きとわかる。

3･2 GPT-4 による信念と選好の同時推論

・大規模言語モデルのChatGPT（GPT-4）に、この推論ができるかを検証した（2023.11.25）．
図3にChatGPTに与えたプロンプト，図4にChatGPTの回答を示す．

・プロンプトの問 1は行動の生成モデルを明示的に問う問題で，回答の三つの行動パターンは，仮説推論のための仮説となる命題群を構成している．ChatGPTが正解を出力できるまで，いくつかのプロンプトを試したが，問 1を入れないプロンプトでは正解を出力できなかった．

【図3の問1】
Aさんが寿司、ハンバーガー、カレーのそれぞれが最も好きだった場合に
どのような行動をとるか、それぞれの場合について推論してください。

・また，一般的にGPT-4などのLLMは自分自身の独自の選好はもたないが，
特定の選好をもつ人がどのような行動を出力するかの一般的知識を有している．
すなわち，合理的行動の生成原理に関する一般的知識を有していることを示唆する．

→★この課題について、人間には推論が可能だが、ChatGPTには解けなかった。
そして、課題の前に、上記問1の誘導尋問を与えると解けたとのこと。
ただ、その理由を説明するのは、大規模言語モデルでは難しいと思われる。
いわゆるニューラルネットワークの出力の説明が難しいのと同じ。

4. インタラクションにおける他者理解

・サリー・アン課題やフードトラック課題で，回答者に求められるのは一方的な観察であるが、現実の社会的状況は，インタラクション中に他者を理解することが求められる．

4･1 価値観の違いを考慮したインタラクション

・ゲーム理論，社会心理学，経済学などの分野では，競争と協力という単純化された枠組みで人々のインタラクションをモデル化する．

・よく使われるゲーム，囚人のジレンマでは，二人の犯罪容疑者が互いに相手の行動を予測して意思決定を行う状況を表している．**表1**に囚人の利得表を示す．容疑者A，Bの二人は，コミュニケーションが取れない状態で，検事は自白させるために次の司法取引をもちかける．
●もし両者が黙秘（協力）すれば，証拠不足で両者ともに軽い刑になる（懲役 1年）．
●一方が自白し（裏切り），もう一方が黙秘すれば，自白したほうは釈放され，
黙秘したほうは重い刑（懲役 3年）に処される．

●両者が自白すれば，二人とも中程度の刑（懲役 2 年）に処される．

・自分が黙秘して相手が裏切る（自分は懲役 3 年）可能性と自分が裏切って相手が
黙秘する可能性（自分は釈放）を考慮すると，自白を選択するのが妥当である．

・相手が協力する意図をもっていることがわかった場合、自己利益だけ考えれば，
裏切るのが妥当（経済合理的）であるような状況でも，一定数の人は協力を選択する．

・搾取できる状況であっても搾取せずに協力することは合理的な選択になり得る．
直接互恵的関係における協力は、将来の返報が期待できるために合理的である．
間接互恵的関係における利他は、自分に利益が戻って来る可能性があるために合理的である．

・協力性や利他性などの性格傾向は，そのような行動を促進する内在的な動機となる．
行動ゲーム理論，心理学的ゲーム理論，相互依存理論は個人の価値観を導入している．

・個人が競争的であるか協力的であるかは、**社会的価値志向性（SVO）**としてモデル化される．
SVO は，お金などの資源を分配する際に自他にどのように配分するかの好みであり，
一見非合理であるが、長期的に合理的な利他行動を熟考なく生成する動機となる．

・利他と協力を**向社会**，利己と競争を**向自己**と呼ぶ．協力は共同利益を向上させる一方，
相手が協力的という思い込みや無防備な協力的態度は容易に搾取されるので，
向社会者ほど相手の意図に敏感でなければならない．

→★一般的で当たり前の話と思う。

・SVO と利得表を与えると各選択における主観的な効用が計算され，合理的な選択が決定する．
エージェントがある選択をした場合の期待効用 u：

$$u = [W_{self}, W_{other}] \cdot [R_{self}, R_{other}] \quad (2)$$

＊Wself，Wother：自他の価値を重視する程度を表す重み
＊Rself，Rother：当該選択が行われた場合の自他それぞれの報酬

・SVO によって**表 1** のプレーヤ A の利得がどのように変換されるかを**表 2** に示す．
プレーヤ B の行動が予測できない場合，プレーヤ A の SVO が自虐，自己犠牲，利他，協力では
（協力、黙秘）を選択し，個人主義では（裏切り、自白）を選択することが合理的である．

→★相手の気持ちが不明で、協力してくれるか、裏切られるか、等確率の場合、
自分の社会的価値志向性 SVO が「自分が損してもよい」なら［協力］を選択し、
逆に「自分が損するのが嫌」なら［裏切り］の選択が合理的とのこと。
金融商品購入のときに、ハイリスク・ハイリターン／ローリスク・ローリターンの
どちらにするかの選択と同じかな。

・他者とインタラクションする**エージェント**は，明示的な情報交換と，非言語的な手掛かり，

特に感情表現を観察することで，相手の心的状態を知る．感情表現は，目標や好みなどの心的状態を相手に伝えるなど，重要な社会的機能を果たす．感情表現は曖昧で，観察者は意味を同定するために文脈に注意して推論しなければならない．

・著者は，囚人のジレンマを用いた実験で，人がインタラクション中の相手の表情から相手の協力的か競争的かの意図を推論し，意思決定に反映できることを示した．

・実験参加者はゲーム状況に対する相手の行動，表情を観察し，心的状態を推論した．その結果，相手の戦略は全く同じであったにもかかわらず，協力的表情パターンの相手には協力を選択する傾向が高く，競争的表情パターンの相手には裏切りを選択する傾向が高いことが示された．

→★人は相手の表情から気持ちを読み取るという一般的な結論と思う。

・表情から心的状態を推論する過程は逆評価と呼ばれ，ベイズ推論によってモデル化される．

4･2 より詳細な価値観を考慮したインタラクション

・人は、物事に対してより詳細な偏りをもち、物事は、数多くの**属性（特徴）**から構成される．物事の最終的な価値は，これらの特徴に割り当てられる価値の総和として決定される．

・ここで，属性は物事に関する主観的な理解と知識であるために，**信念**として扱う．ある物事が **N 次元**の特徴ベクトルで構成され，各次元が**命題 A→B** を表すとすると、ある物事に関する個人の**信念**は次のように表現できる．

$$Bel=\{x\in RN \mid -1\leq x\leq 1\} \quad (3)$$

＊ x は，それぞれの**命題［A→B］**を信じる程度
＊ 正の値のとき A→B が真であると信じる程度
＊ 負の値のとき A→¬B が真であると信じる程度
＊ ゼロは，A と B の無関係性の認識を意味する

・多次元の特徴をもつ事物は単一の表現に抽象化され，この表象は効用 u をもつ．
u は、**命題に対する価値**（重み）

$$Val=\{w\in RN \mid -\infty\leq w\leq +\infty\} \quad (4)$$

と Bel の**内積**で次のように表す．

$$u=w\cdot x \quad (5)$$

・例えば、飲酒が［飲酒→肝臓病，飲酒→快楽］の二次元の特徴で構成されるとし、飲酒が肝臓病を引き起こすと信じてネガティブに評価し，

＜（x［飲酒→肝臓病］＝1)、(w 飲酒→肝臓病＝－1）＞

飲酒による快楽を肯定してポジティブに評価する人の場合，

＜（x［飲酒→快楽］＝1)、(w［飲酒→快楽］＝1）＞

飲酒の効用は次のように計算できる．

$$u=[-1,\ 1]\cdot\ [1,\ 1]=0$$

・一方，飲酒が肝臓病を防ぎ、快楽を発生させると信じ，ポジティブに評価する人の場合，
飲酒の効用は、u＝[1，1]･[1，1]＝1、となる．

・著者は，ある人にとって負の価値かつ「真」の信念をもつ命題（懸念）に対して，
インタラクティブな AI エージェントが，その懸念を払拭するような命題を提示することで，
人を説得できるかを実験で検証した結果，効用そのものを変化させることは難しいが
「○○するべきである」という義務感については変化させられることがわかった．

→★実験結果がわかりにくい。「飲酒が肝臓病を引き起こす」という信念は変えられないが、
飲酒の程度と肝臓病の発病との統計情報から少量の飲酒は問題ないと示すことで、
「飲酒は避けるべきである」という義務感は変えられるということかな？
ＡＩエージェントは意図を持たないので、「懸念を払拭する命題を提示する」ように
利用者が意図的にプロンプトを工夫する必要があるのでは？

4･3 価値観の多様性を考慮した Win-Win 関係

・先に述べた囚人のジレンマでは，懲役の価値が両者にとって等しいことを前提としているが，
懲役の価値は個人によって異なる．より具体的かつ詳細な個人の価値を考慮して、
非ゼロ和ゲーム的状況で Win-Win 関係を追求可能な課題として，**複数論点交渉ゲーム**がある．

・囚人のジレンマでは，相手が協力的か競争的かの意図を見極めることが重要であったが、
複数論点交渉ゲームにおいては，相手の選好，限界，最良代替案の真偽といった，
より複雑な相手の心的状態を見極めることが求められる．

・熟練した交渉者は，双方に利益をもたらす Win-Win 解の可能性を理解しているが，
非熟練者は一方が利益を得れば他方が損をするゼロ和として認識することが多い．

・固定パイバイアス，アンカリング，過信，フレーミングなど
Win-Win 解を見つけるのを妨げる多くの認知バイアスが特定されている。

・相手の好みやゲームのクラスの特定に影響を与えるのは**固定パイバイアス**で，
相手の利益が自分の利益と完全に対立しているという誤った信念である．
固定パイバイアスの低減には感情表現を通じた情報交換が有効であるが、
これは，相手の感情表出から相手の心的状態を逆推論することで，
交渉における相手の選好や限界を推論できるからだと考えられる．

→★わかりにくいが、熟練した交渉者は、相手の感情表出から心的状態を推論して、
Win-Win 解（妥協案）を導いているということかな。

・著者は、交渉における感情の生成モデルの学習により Win-Win の交渉実現の可能性を示した．
エージェントの効用 u＝w･x の x＝[x1 ，…，xn] は交渉対象となる n 個の論点の水準，
w は論点に割り当てた価値である．**エージェント**は効用に従って次のように表情を表出する．

expression＝
　anger　　if u ＜ limit,
　neutral　if u ＝ limit,
　joy[k]　　if u ＞ limit.　(6)

limit は，提案がそれ以下であれば拒否するという限界である．
k∈{1, …, K} は喜びの程度であり，効用値が大きくなるほど大きな喜びを表出する．

・実験参加者のタスクは，**エージェント**に拒否されないようにし，スコアを上げることである．
拒否を回避するには表情が怒りに変化する限界を見極めなければならない．
しかし，参加者のスコアを上げるには**エージェント**の選好 w を正確に推測する必要がある．
数学的には，u と x についての n 個の連立方程式を解くことで求められるが，
実際には図 8 に示すインタフェースを用いてインタラクティブに推定することができた．

・実験の結果，生成モデルを学習した参加者は相手の心を読む能力が向上したので、
より複雑な社会的状況で，人が生成モデルを利用した読心を行っている可能性を示唆する．

→★結論として、生成モデルを利用すれば「熟練した交渉者」になれるということかな。
　『人が生成モデルを利用した読心を行っている』という文の意味が理解できないが…
　「行っている」は「行うことができる」の書き間違いかな？

5. 心を読み人の振舞いを予測する AI

・本稿では、他者の行動を予測するための認知モデル（他者モデル）について，
心の理論の計算モデルであるベイジアン心の理論を中心にこれまでの研究を振り返った．
ベイジアン心の理論では，観察者が，行為者のどのような心的状態が，どのような状況で，
どのような行動を出力するかについての**合理的行動の生成モデル**を用い，
観測した**振舞いから逆に心的状態を推論**する．これは逆推論と呼ばれる．

・**ベイジアン心の理論**では，合理的行為者の生成モデルがあれば，逆に心的状態が推論できる．
図 4 のように，LLM は合理的行為者の生成モデルをもっている．さらに，図 7 のように，
LLM は人の多次元効用関数を構成するもとになる概念特徴についても詳細なモデルをもつ．

・パーソナライズされた LLM が、個人と長期的にインタラクションすれば，
AI による個人の詳細な効用関数のモデル化は可能であるが，
人の「歪ゆがんだ」認識および「不合理な」意思決定を発生させる
さまざまな認知バイアスを含んだものになるだろう．
詳細かつ**完全な**個人の効用関数をニューラルネットワークの中に構築し，
個人の完全な行動予測が可能な AI が実現されると考えられる．

→★個人の行動予測が可能な AI は実現しても、「**完全な**」予測は簡単でないと思う。

・心理学，認知科学は，予測できない人を理解し、予測可能にするモデルの探求だった．

今後，他者の行為予測において，人工ニューラルネットワークを用いたAIが
人を凌駕する性能を発揮するのはほぼ間違いない．

・個人の行為を予測するための生成モデルの重要な要素である効用関数も
人工ニューラルネットワークの中に複雑さを保ったまま表現されるだろう．
そして，それらはもはや，説明や理解が不可能かもしれない．工学的にはそれで問題ない．

→★ニューラルネットワークは出力について説明できないが、すでに実用にはなっている。

・相手が詐欺師かどうかを正確に見分けてくれるパーソナル**エージェント**，
友達が最も喜ぶ誕生日プレゼントを提案するパーソナル**エージェント**がつくられるだろうが、
科学的にはどうだろう．これまでのように，**エージェント**の入出力を単純化したり，
人を有限のカテゴリカルな属性に基づいて理解することはまだ意味をもつだろうか．

→★この解説の「はじめに」で述べた『数理モデルの価値が薄れる』という指摘のことかな？

・多次元の交互作用の理解と的確な説明は難しい．
超多次元の入出力関係をモデル化できる人工ニューラルネットワークの中には，
超多次元の交互作用が正確に表現されているだろう．
もちろん、人が理解できるようにやさしい説明を生成してくれるだろう．
その説明を甘受するのかそれとも理解の限界の敗北を認めるのか，
それを迫られる時が来るのではないかと思う．

→★危機感を具体的に説明してほしい。
著者は、近い将来、ＡＩ技術が完全に人間の能力を凌駕すると信じたうえで、
人間がＡＩに従属する時代の到来を危惧しているようにみえるが・・・

→★ＡＩはあくまでも人間の道具なので、使い方を誤らないことが重要と思う。

→★私が第２次ＡＩブームの時に検討した、自分がやりたいことを私に代わってやってくれる
ソフトウェアクローンは、本解説のパーソナルエージェントの概念に似ているかな。
この発想の原点は、日常的業務（ルーチンワーク）をソフトウェアクローンにまかせて、
人間は創造的業務に専念できるようにすることで、システム名は知的秘書システムとした。

（参考URL） https://1968start.com/M/rd/rd/IC.html
（引用元）拙著「ソフトウエア危機とプログラミングパラダイム」(啓学出版、1992.8)
https://1968start.com/M/keigaku/index.html

以上