自己言及を基盤とした類推による他者理解の学習モデル
― 心の理論の学習モデル構築に向けて ―

山川宏，岡田浩之(東海大学理学部)

１．はじめに

心の理論研究を中心とした多くの実験事実が蓄積され，他者理解のモデル化に挑む条件が整いつつあるが[1]，現状ではメカニズムに踏み込んだ学習モデルはまだない．しかし今回は，実験知見との詳細な整合性を議論以前に，生体で実現可能な学習モデルの基本的要請を列挙し，それに適合するモデルの枠組みを提案する．

メカニズムを含むモデル研究では，個別の技術的課題に拘泥し，大局的な課題を見失う危険性を避けるため，このような検討が重要である．

２．生体情報処理モデルとしての要請

ヒトの脳で実現可能な他者理解モデルとしては少なくとも，以下の要請を満たす必要がある．

【要請 1】利用可能な記憶機能(学習)

脳の神経回路で実現可能な以下の２種類の記憶機能により実現されるべきである．

(A)統計的記憶： 取得情報の統計的な性質を反映した，長期的な意味記憶などに対応する．多くの学習データが必要だが，関連する情報に関する予測性を持つ点で優れる．

(B)即時的記憶： エピソード記憶や，自己の心的状態の記憶，語彙獲得の即時マッピングのように一度限りの経験(情報間の関係)についての記憶．統計的性質を反映しないために，予測性を持たない．

【要請２】統計学習表象の流動性と唯一性

統計的記憶の表象は，常に学習が行なわれる性質のため，流動性を持つ．また，生体の神経回路は，コンピュータと異なり表象の複製はできない．

複製不能性と流動性のために，統計学習表象は唯一の存在で，他部位の表象と直接比較できない．

【要請３】多元的心的状態(様相)の統一的処理

心的状態として自己／他者の区別する必要があるが，ヒトはこの他に，過去／未来や，信念／願望／意図などの心的状態を区別して用いる様相処理能力を持つ．様相論理を用いることで，これらの多元的世界を統一的に扱える．

「実現し得るモデルは単純である可能性が高い」というモデル選択原理に従えば，様相処理も，統一的な処理機構により実現されるべきである．

【要請４】推定他者表象の解釈能力

自身がシンボルで表現できる心的表象を，他者心的状態として推定したら，直ちに対応付けしてシンボルとして表現できるべきである．

３．他者心的状態の推定方法と知識

他者身体性知識，自己身体性知識，自他同一性知識という３種類の知識の観点から，他者心的状態の推定モデルとして妥当な枠組みを検討する．そして，動的な類推であることを述べる．

3.1 他者身体性知識を用いたモデル化

他者の外部状態と心的状態の関係である他者身体性知識を利用する「他者身体性知識獲得モデル」では，他者心的状態の推定は容易である．

この知識の取得は，学習データが少ない他者の，直接観測できない心的状態を隠れ変数とみなして推定するために非常に難しい【要請１】．

さらに，得られた他者心的状態から自己心的状態への写像を決定できず，他者心的状態の解釈問題が発生する問題がある【要請４】．

3.2 自己言及を基盤とした類推によるモデル化

次に，自己を他者の一つであると客体化する自己言及に基づいて類推を行なうモデルを考える．

自己の外部状態と心的状態の関係である自己身体性知識は，自身が経験する多くのデータを用いて，発達の早期段階から獲得される．

一方，ミラー細胞の存在からも示唆されるよう，ヒトも自己外部状態と他者外部状態の対象レベルでの同一性を反映する自他統一性知識を持つ．

事前獲得した自己身体性知識を，他者身体性知識として複製する「複製類推モデル」は，生体の神経回路網では困難な知識の複製を含む問題がある【要請２】．そして異なる他者や，部分的な知識領域に応じて何れの時期に複製を行なうかも問題である．さらに，複製後の学習により自己と他者の対応付けが次第に失われるため解釈に支障をきたす問題もある【要請４】．

必要に応じて，自他同一性知識と自己身体性知識を組み合わせて類推を実行する「動的類推モデル」は，前記４つの基本要請に抵触しない．そこでの我々は，他者理解の学習モデルとして，この枠組みを採用する．

4．多元様相推論の学習モデル

「動的類推モデル」を，【要請３】を満たすよう拡張した多元様相推論の学習モデルを提案する． これは，統計的学習メモリ(SLM)，シンボルメモリ(SYM), 様々な様相状態により構成される．

4.1 デフォルト様相と統計的学習メモリ(SLM)

心的状態に関わる様相としては，過去／未来，信念／願望／意図，自己／他者などの組み合わせがある．特権的な位置を占める，自己の現在の信念に関わる様相をデフォルト様相と呼び，それ以外を非デフォルト様相と呼ぶことにする．

SLM は，大量のデータが得られるデフォルト様相を利用して，統計的学習を行なう【要請１】．ここには自己身体性知識などの事象間の関係についての知識が獲得・保持される．

4.2 非デフォルト様相に対する推論処理

非デフォルト諸様相における推論にも，SLM 上の知識を用いるため，その心的状態を SLM 上に表現する必要がある．以下で【要請３】を満たす，諸様相に対する統一的な推論処理を説明する．

まず，他者心的状態の推定においては，既に述べた「動的類推モデル」に従い，(1)自他同一性知識による他者外部状態から自己外部状態への変換，(2) 様相状態が*他者 X* であることの保持，(3)SLM 内の自己身体性知識による自己外部状態から自己心的状態の推論が行なわれる．

物語など言語情報として他者外部情報が与えられた場合には SYM 上のシンボルから自己外部状態への変換が行なわれる．シンボルグラウンディングに関わる SYM 内表象と SLM 内表象の関連付けは【要請１】から即時的学習で獲得すると考えるが，本稿ではこれ以上立ち入らない．

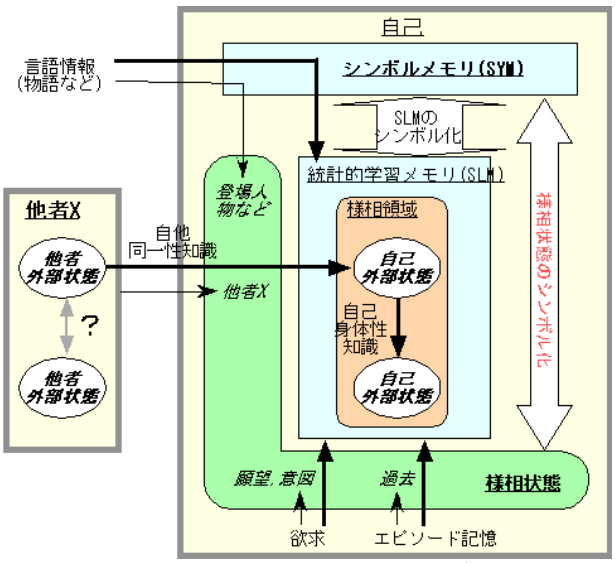

図１ 多元様相推論の学習モデル

他者心的状態推定以外の処理としては，自己内部の欲求によって，*願望／意図*など態度が様相状態に登録され，その内容について SLM では実践的推論が行なわれる．また，エピソード記憶からは*過去*という様相状態が登録され，その内容について SLM では記憶の再現が行なわれる．

4.3 非デフォルト様相推論の性質

ヒトの様相状態は，意図的に制御されるばかりでなく，景色から過去を思い出したり，お腹が空いて冷蔵庫を開けたりと不意に切り替わるが，同時に複数の様相は扱えないだろう． SLM を時分割して用いる提案モデルは，上記事実と整合する．

何れの非デフォルト様相も，SLM，様相状態への入力情報を生成するために事前の推論過程(SYM など)を含み得るが図１では表現しきれない．

またヒトは，何れの非デフォルト様相に関わる推論処理の最中も，外界に近い領域はデフォルト様相に関わる処理を行なうため，SLM 上で非デフォルト様相の推論領域を限定するはずである．この実現には SLM をサブモジュールに分割するなどの方法を考え得るが，今後の検討課題である．

4.4 処理結果の出来事としての記銘

非デフォルト様相における SLM 上の推論結果の SLM 自体への影響は，過去様相の場合は記憶の強化として，創造的な要素を含む意図様相の場合は関係性発見の促進として作用するだろう．

一方，推論処理結果(意図した計画など)を一つの出来事として記銘するには，様相状態と SLM 上の推論結果を関連付けて記銘する必要がある．

欲求に基づく意図などの様相状態を直接に記銘可能であるかどうかについては議論を要する．

仮に，直接記銘が不能な場合は， SYM 上において一種のメタ表象として**様相状態のシンボル化**を行なえば，出来事としての記銘が可能となる．

5. おわりに

モデル研究の立場から，他者理解モデルの要請としては，多元的な様相の統一的扱いが重要と考え，新たな学習モデルの枠組みを提案した．

今後，様々な実験課題における様相制御を具体化し，既存実験知見との整合性を検証したい．また，構成が類似する PATON[2]モデルとの比較検討したい．考察にあたり[3]の諸記事を参考とした．

参考文献

[1] 子安増生, "幼児期の他者理解の発達",1999.

[2] T. Omori et al., *Neural Networks*, Vol.12, No.7-8, p1157-1172, 1999.

[3] 山川宏ら編, "特集:意図研究のスペクトル", 人工知能学会誌,Vol.20, No.4, 2005.