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)

本日の話題.

- 心の理論研究関連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.
「心の理論機構」のメカニズムに踏み込んだ学習モデルの構築に向け，本日はモデルの要請を列挙し，要請に適合したモデルの枠組みを提唱する.

  - 他者理解における「心の理論機構」
  - 生体情報処理モデルとしての要請
  - 他者心的状態の推定方法と知識
  - 多元的様相推論の学習モデル
  - まとめ

1.他者理解における「心の理論機構」

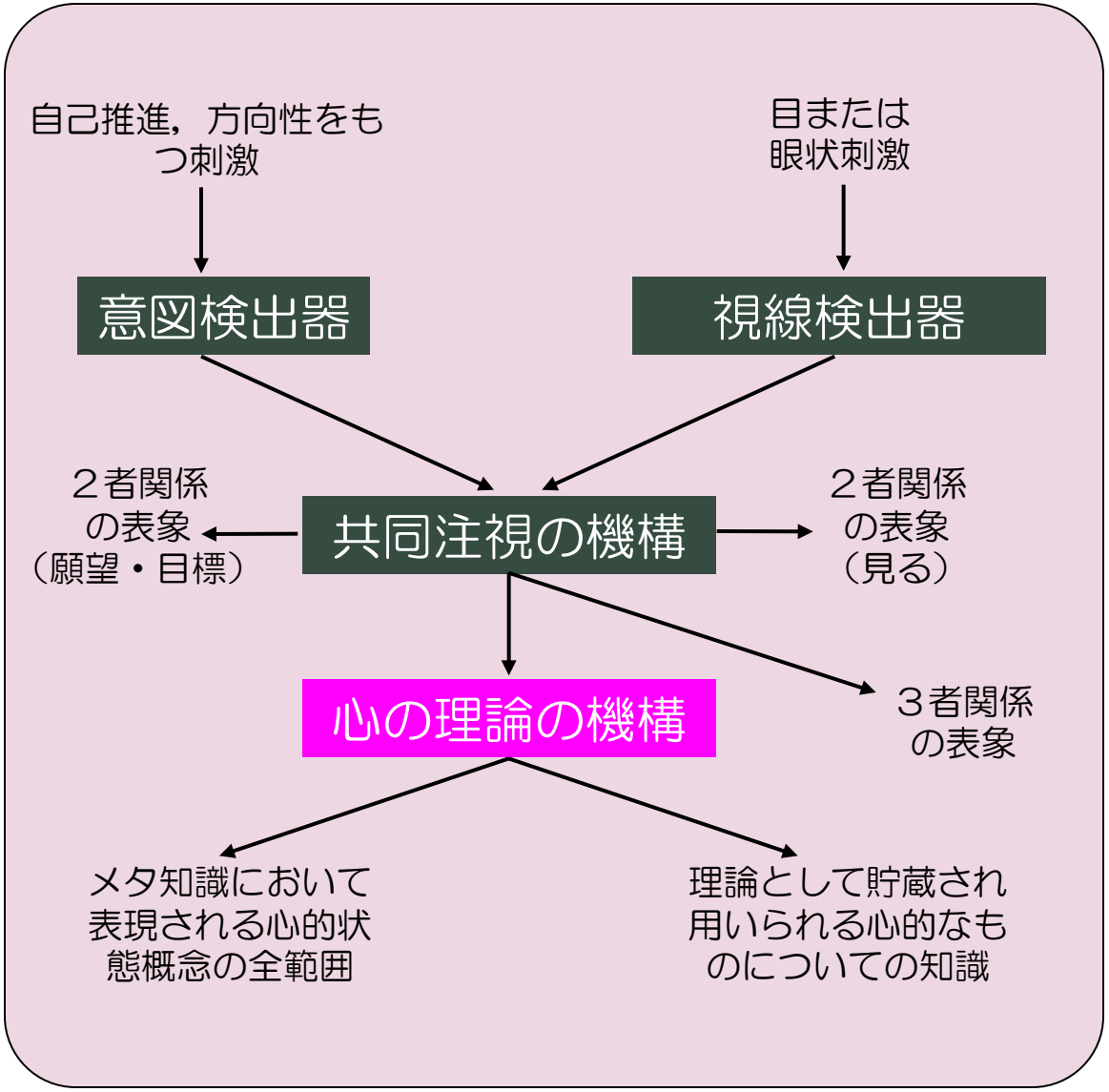
.

心の理論

D. Premack and G. Woodruff, Does the chimpanzee have a theory of mind?

- **心の理論**（Theory of Mind）
  - ヒトが、他者の心の動きを類推したり、他者が自分とは違う信念を持っているということを理解したりする機能のことである
  - 自己および他者に「思う（think）」や「知っている（know）」のような心的状態を帰属させること

他者理解における「心の理論機構」



Baron-Cohenのモデル

- 心の理論には複数の機構が必要だと考えられている.

- 心の理論の機構の学習モデル
  - 他者の心的状態を推論する機構

2.生体情報処理モデルとしての要請

.

生体情報処理モデルとしての要請

- モデル化における大局的な課題を探りたい.

- 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.

- ヒトの脳で実現可能なモデルの要請
  - 【要請1】利用可能な記憶機能(学習)
  - 【要請2】統計学習表象の流動性と唯一性
  - 【要請3】多元的心的状態(様相)の統一的処理
  - 【要請4】推定他者表象の解釈能力

【要請1】利用可能な記憶機能(学習)

脳の神経回路では以下2種の記憶機能で実現.

- (A)統計的記憶:
  多くの学習データから統計的な性質を抽出する，関連する情報に関する予測性を持つ点で優れる.
  - 長期的な意味記憶など.

- (B)即時的記憶:
  一度限りの経験(情報間の関係)についての記憶なので，統計的性質を反映せず，予測性を持たない.
  - エピソード記憶や，語彙獲得の即時マッピングなど

【要請2】統計学習表象の流動性と唯一性

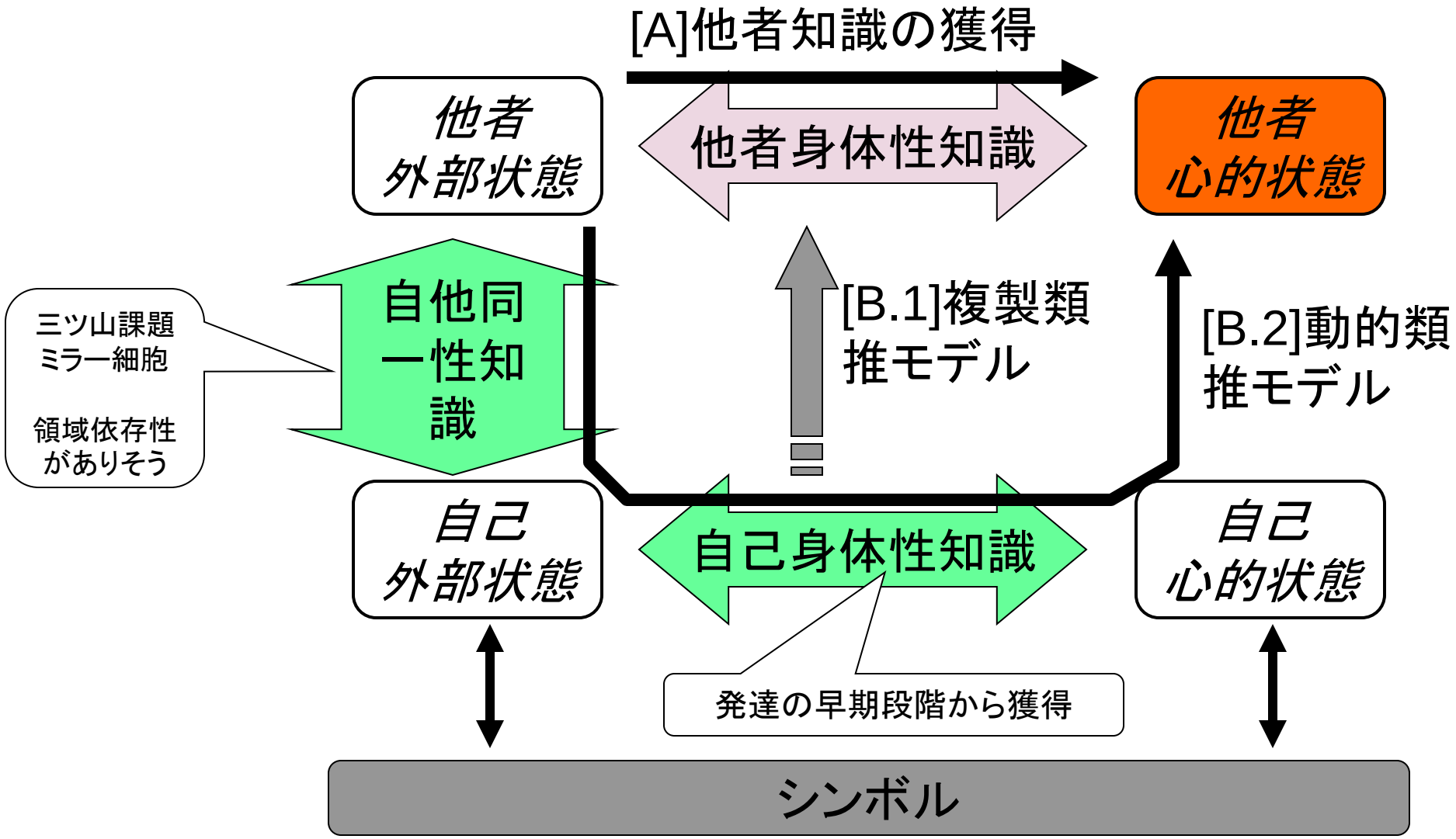
- 統計的記憶の表象は，常に学習が行なわれる性質のため，流動性を持つ．また，生体の神経回路は，コンピュータと異なり表象の複製はできない．
- 複製不能性と流動性のために，統計学習表象は唯一の存在で，他部位の表象と直接比較できない．

【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

他者心的状態の推定と知識



[A]他者知識の獲得によるモデル化

他者身体性知識を獲得するモデル.

《問題点》

- 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他者身体性知識の取得は，学習データも少なく困難【要請1】.
- 他者心的状態からシンボルへの写像を決定できず，解釈に問題が発生する【要請4】.

※工学系研究者からの学習モデルではしばしば見かける.

[B]自己知識による類推モデル

[B.1] 複製類推モデル

自己身体性知識を，他者身体性知識として複製し，
他者外部状態から他者心的状態を類推する.

《問題点》

- 知識の複製を含む問題がある【要請2】
- 複製後の学習により自他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】.

[B.2] 動的類推モデル

自他同一性知識を自己身体性知識を組み合わせて，
他者外部状態から他者心的状態を類推する.

- 前記4つの基本要請に抵触しない．⇒この枠組みを採用

４．多元様相推論の学習モデル

.

多元的様相(心的状態)の統一的処理

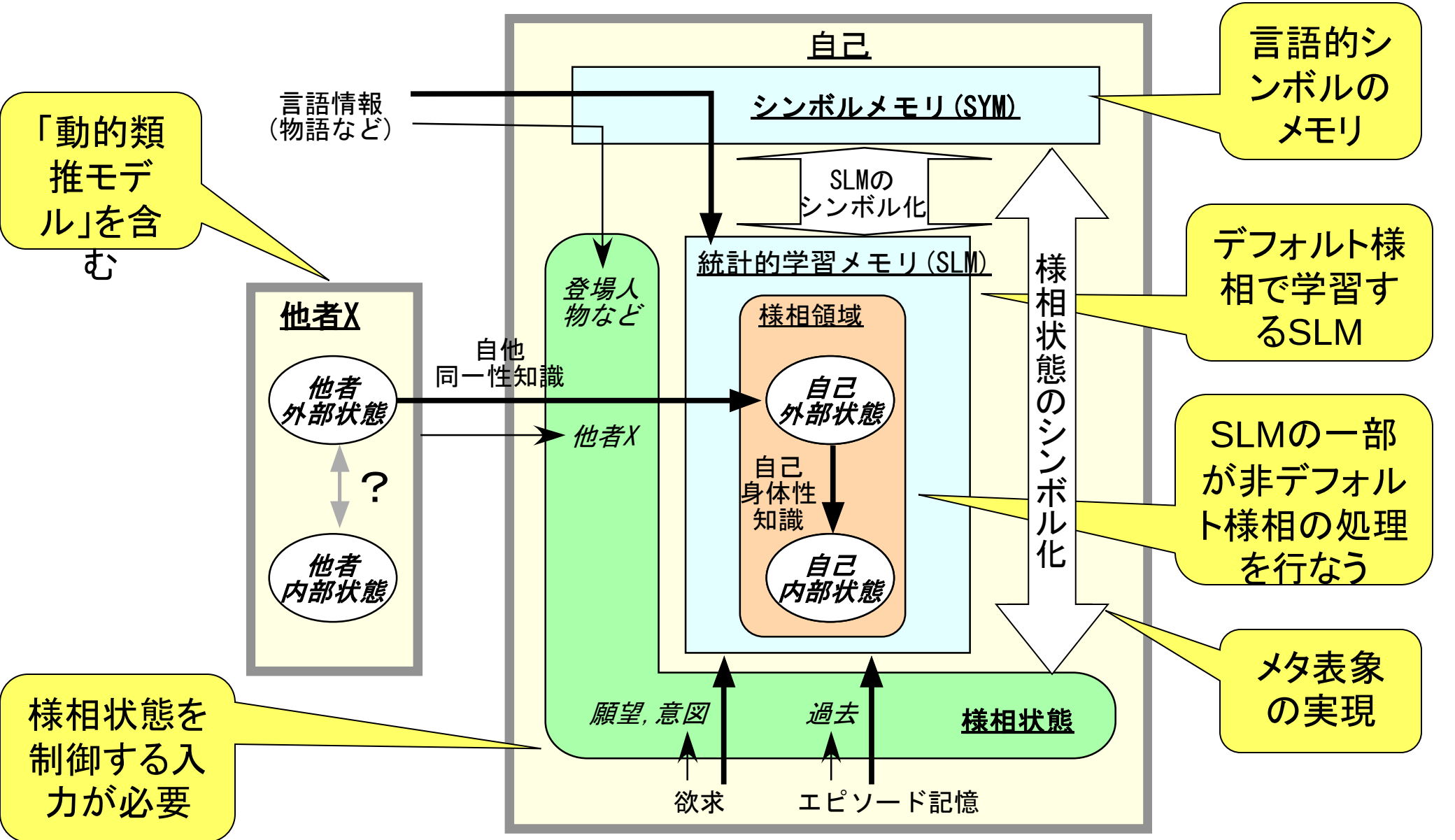
- 【要請3】多元的な様相を統一的に扱えるべき
  - 自己／他者
  - 過去／現在／未来
  - 信念／願望／意図

他者心的状態の推定だけが様相ではない.

- 一時に，意識に昇るのは上記の組み合わせ内の一つ
- 統計的学習には大量のデータが必要【要請1】
  - デフォルト様相＝自己の現在の信念
    - 自己身体性知識などの事象間の関係の知識
  - 非デフォルト様相：デフォルト様相以外

単一の統計的学習メモリ(SLM)がデフォルト様相で学習

多元的様相推論の学習モデル

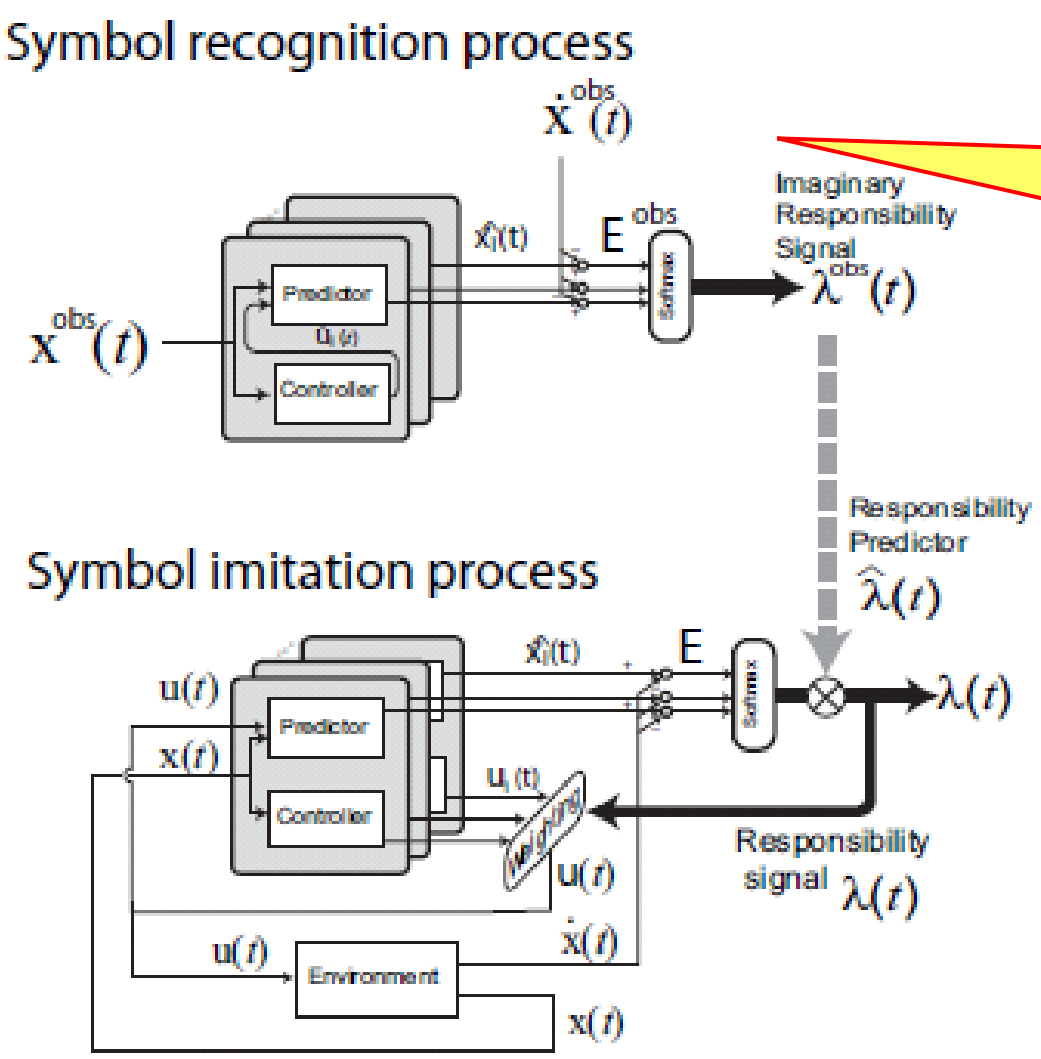


モデル化における課題

- SLM上で非デフォルト様相の推論領域の限定手段
  - 非デフォルト様相に関わる推論処理中も，SLMの多くの領域はデフォルト様相に関わる処理を行なう.
- 非デフォルト様相処理結果の記銘(学習)
  - 他者や意図など様相に関わる処理結果が，SLM上に上書きされるのは，記憶の混乱を招く可能性がある.
- 様相状態のシンボル化
  - 過去や意図などの様相状態は，(言語的表現を除けば)外界に現れないので，これらを操作可能な対象とするには，シンボル化が必要となるかもしれない.

関連研究1: MMRLによる見まね

鮫島和行(ATR)
2002



自己身体性知識の学習による獲得

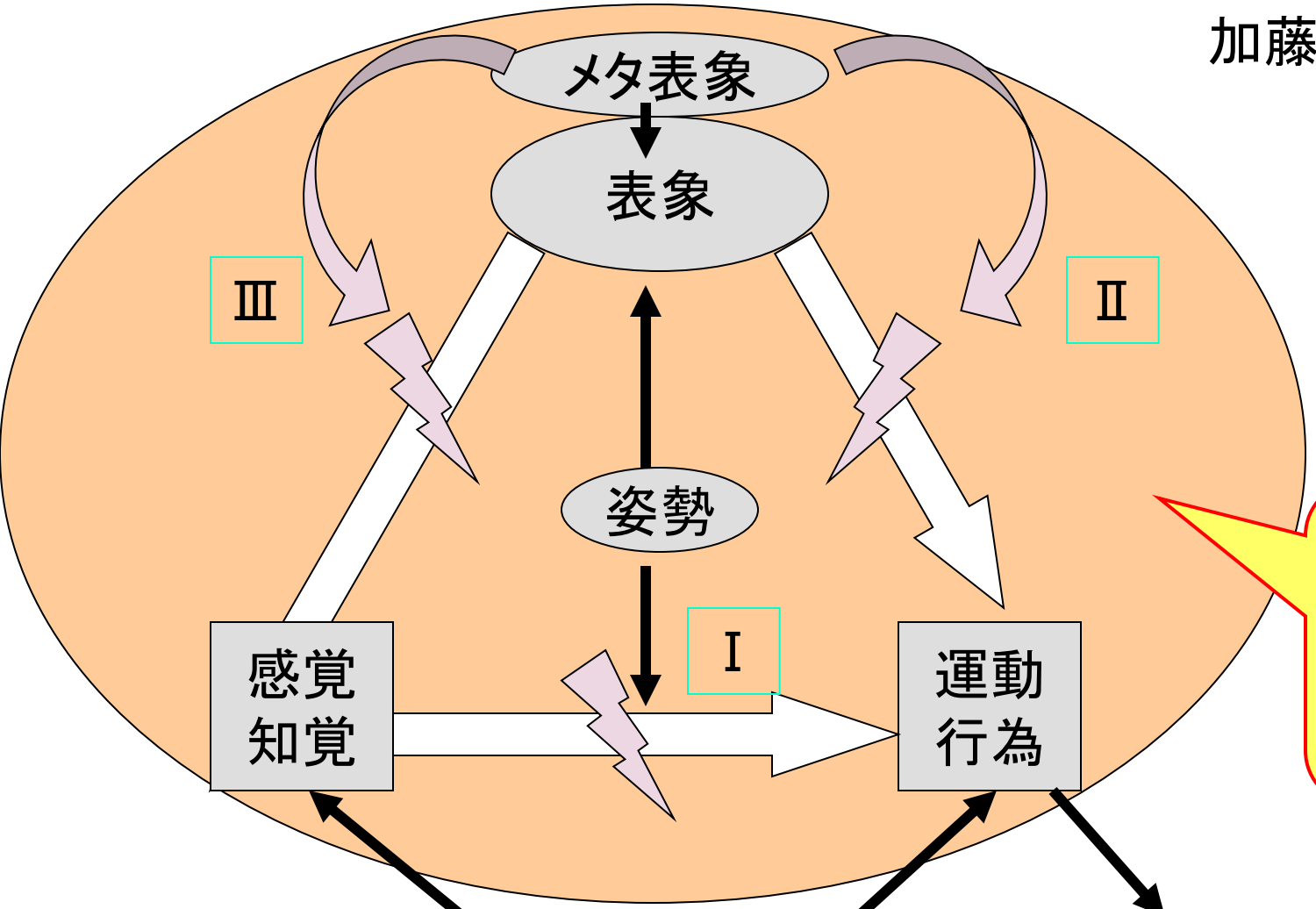
マルチモジュール強化学習(MMRL)による見まね

【上段】他者の軌道$x^{obs}(t)$を観測し、対となっている予測器と制御器を使うことで他者が使っているであろう責任信号$\lambda^{obs}(t)$を推定する。

【下段】推定した他者責任信号を使って，自らの行動を生成.

関連研究2:知覚，行為，表象の切り離しモデル

加藤義信(愛知県立大学)
2004



非デフォルト様相の処理機構.

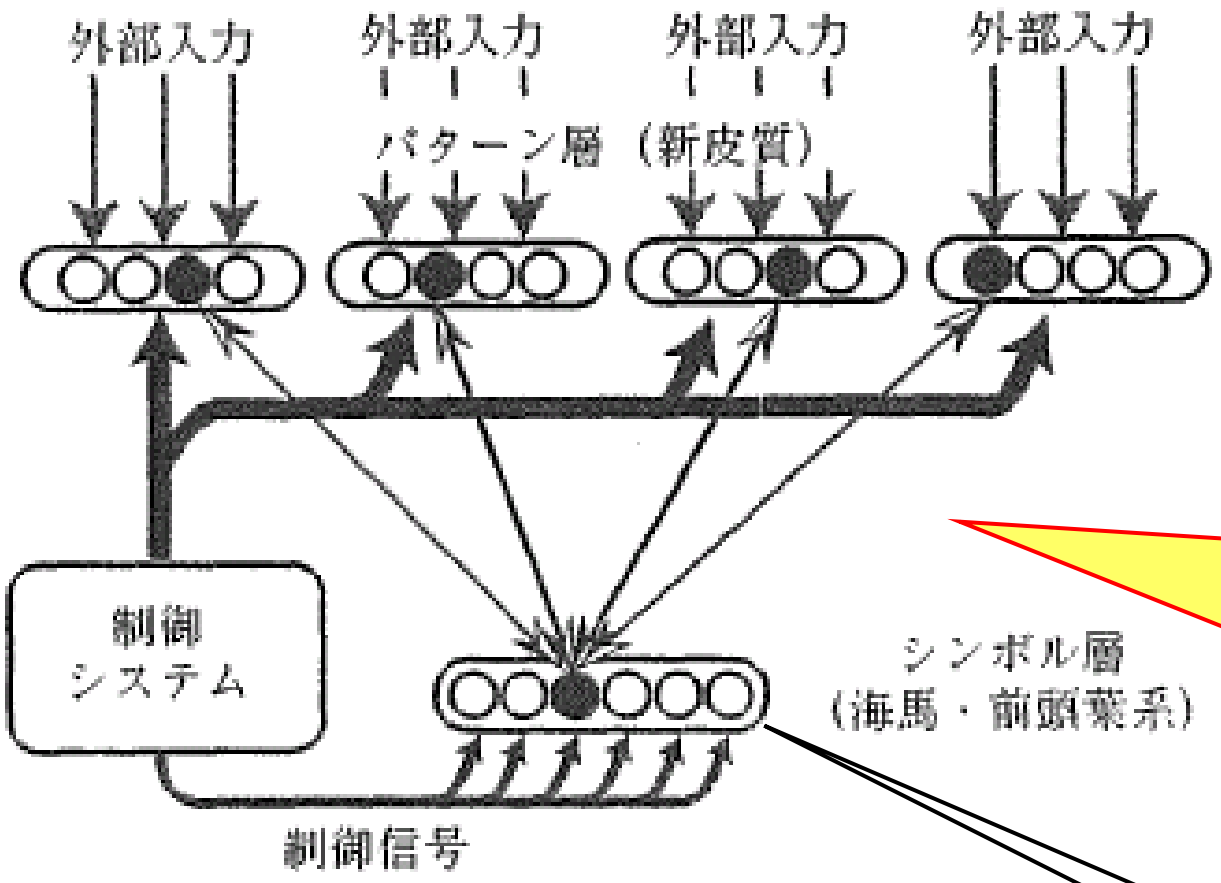
デフォルト様相

「いま，ここ」の世界

「これまで，これから，あそこ」の世界

関連研究3:思考する記憶機械 PATON

大森隆司(北大)



SLM(統計的学習メモリ)

SLMとSYMの存在.

制御システムは，時として様相状態の制御.

シンボル層とパターン層からなり，それに制御システムが注意信号を送る

SYM(シンボルメモリ)

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，構成が類似する他のモデルとの比較検討したい.

- 提案モデルを具体化し，検証する実験系について検討したい.