3E1-05

他者意図理解のBDI論理による解釈

In explanation of understanding the intention of others by BDI logic

岡田浩之*1　　山川宏*2
Hiroyuki OKADA　　Hiroshi YAMAKAWA

*1東海大学理学部
Tokai University

In this paper, we show the possibility of modeling and explanation of the mental situation thought to be an important element of the theory of mind like Belief, Desire, Intention by BID logic. And also, we suggest the possibility of united discussion about theory of mind by BID logic. And we try to describe mental process in a theory of mind problem that is a certain kind of social problem solving in chimpanzees by BDI logic.

1. 心的過程の記述と心の理論

自己や他者の心の状態をどのように理解しているのか、その理解のための枠組みが「心の理論」であり、人間はヒトは他者とのコミュニケーションを実現する手段として,「心の理論」を有効に利用していることが知られている。「心の理論」とは,自己や他者の行動をその人がもつと予測される内的表象（知識や信念, 願望、意図、欲求などの心的状況）に帰属させて考える仕組みであり, この理論をもつことで他者の行動予測が可能になると考えられている。

Premack[Premack 78] はチンパンジーが人間の行動をある程度的確に予測できることを非言語的な実験を用いて示し、チンパンジーも「心の理論」を有することを示唆した。Wimmer[Wimmer 83] によるとヒトの 3 才児は誤信念課題に答えるのは難しいが、4, 5 才児になると、自分の目で見ている状態と実際に生じている事実との関係を理解できるようになり誤信念課題に正答する。

このように、Premack 以降、哲学、心理学、脳科学など様々な分野の研究者によって「心の理論」の様々な解釈が我々に与えられてきたが、では「心」の本質は何なのか、その存在がヒトやチンパンジーの脳にあるとすればどのようなアルゴリズムとして動作しているのかについては未だに見解の一致を見ていない。

本稿では信念、願望、意図といった心の理論の重要な要素と考えられる心的状況を BID 論理でのモデル化および説明が可能であることを確認し、心の理論の統一的な議論の可能性を示唆する。最近になって, ヒトや動物を対象にした「心の理論」に関する実験結果が数多く蓄積し、それらを適切に比較する必要が要請されているが、BDI 論理による記述により様々な実験結果の見通しよい比較が可能になると思われる。また、Jadex などの BDI エージェントシステムの利用でモデルの動作を計算機上で確認することが可能になりつつあることを述べる。

2. BDI 論理と BDI アーキテクチャ

信念 (B:Belief), 願望 (D:Desire), 意図 (I:Intention) の様相とその時間的変化を記述する論理体系の BDI 論理がある [Hagiya 94]。BDI 論理で扱う様相演算子には表 1 に示すような信念、願望、意図を表す BEL,DESIRE,INTEND がある。たとえば、BEL(p) は「p を信じている」という解釈をする。同様に表 1 に示したような、未来方向の時間分岐を表す時相演算子がある。たとえば、AX($p \vee q$) は「全ての未来において、その次の時刻に p または q が成り立つ」を意味する。

表 1: BDI 論理の演算子

	演算子	意味
様相	BEL	信じている
	DESIRE	望んでいる
	INTEND	意図している
時相	A	全ての未来において
	E	ある未来において
	X	次の時刻に
	G	現在を含み永遠に
	F	現在を含む未来のいつか
	U	条件が成立するまで

Bratman は著書である「意図と行為」で人間のような合理的な主体の心的状況の中での意図の重要性を指摘した [Bratman 94]。Bratman の意図の理論を BID 論理により実現するものに Rao の BDI アーキテクチャがある [Rao 91]。BID アーキテクチャでは願望世界は信念世界の一部、意図世界は願望世界の一部であるような分岐時間可能世界モデルを定式化に用いている。Rao はこの世界で, (1) 願望を達成したという信念を持つまで, (2) 願望を達成できる手段があるという信念を持っている限り, (3) 願望を達成する状況でなくなるまで、などの, 条件を組み合わせて, 幾つかの意図持続のタイプを提案している。

3. チンパンジーの他者意図理解

3.1 餌を奪い合う 2 匹のチンパンジー

B.Hare [Hare 2000] は実験室で 2 匹のチンパンジー (Pantroglodytes) が餌を奪い合う場面において、チンパンジーは同種の他個体の意図を推測していることを示唆する結果を得た。この実験では社会的に優位な個体と劣位な個体（例えばボスザルと子分ザル）が実験室環境において 2 個の餌をめぐって競争する。双方から餌がよく見える（図 1 Door-Door 条件)、あるいは優位個体からしか餌が見えない簡単に餌の場所に到達できるような設定（図 1 Dominant-Door 条件) において、すべての実験で優位な個体がすべての餌を独占した。

しかし、実験室に置かれた衝立によって餌が劣位個体にしか見えない場合は餌を獲得できる回数が増加した（図 1

連絡先: okada@ss.u-tokai.ac.jp

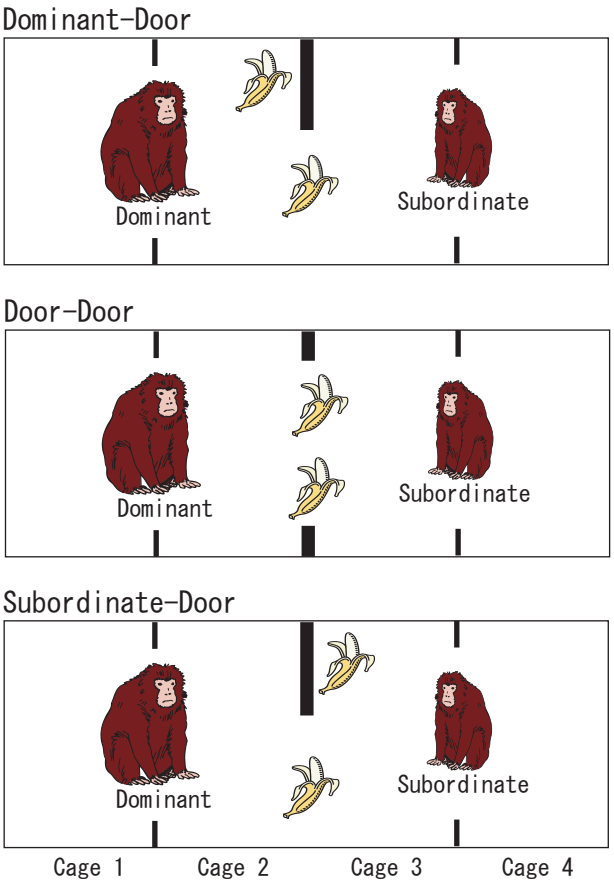

図 1: 2 匹のチンパンジーが餌を奪い合う

Subordinate-Door 条件)。この結果は競争相手である優位個体から見て衝立が障害になり餌が見えていないということを劣位個体が理解し、優位個体の意図を推測し、優位個体が狙っていないであろう餌を取りに行くことで、餌の獲得確率が向上したと解釈できる。

3.2 チンパンジーの心的過程の記述

ここでは、前述したチンパンジーの心的過程を BID 論理で記述することを試みる。

まず、それぞれのチンパンジーは空腹を満たすという願望を達成するためのプランを選び、未来志向的意図として形成する。たとえば、空腹を満たしたいと欲求したなら、バナナを食べる意図を生成するは次のように書ける。

$$\mathbf{DESIRE}(\mathbf{AF}\ \text{空腹が満たされる}) \supset (\mathbf{BEL}(\mathit{like}(\text{バナナ}))) \supset \mathbf{INTEND}(\mathbf{AF}\ \text{バナナを食べる}))$$

また、「バナナを食べる」ことを意図して、それが成功すれば「空腹が満たされる」ことを信じているということは以下のように書ける。

$$\mathbf{INTEND}(\text{バナナを食べる}) \supset \mathbf{A}(\mathbf{DESIRE}(\text{バナナを所有}) \wedge \mathbf{G}(\mathbf{BEL}(\text{バナナを所有}) \supset \mathbf{INTEND}(\mathit{does}(\text{食べる})) \wedge \mathbf{X}(\mathbf{BEL}(\mathit{succeeded}(\text{食べる})) \supset \mathbf{BEL}(\text{空腹が満たされる}))))$$

このチンパンジーの実験における餌の奪い合いにおいて重要なのは子分ザルが自分の経験から得た次のような知識をもとに

- 衝立があるとバナナが見えなくなる
- 自分だけが見えているバナナは奪える可能性が高い

「バナナの向こうに衝立があるということはボスザルから見てバナナが見えていない」と推論し、そのバナナを奪いに行く行動を生成することである。つまり、子分ザルがボスザルの心的状態をシミュレートとしている。これは「自らの観測から相手の状態を予測し、相手の意図を推測する」という「心の理論」の重要な要素の一つであり、その間のチンパンジーの心的過程を以下のように記述することができる。

「見えているバナナしか奪えない」

$$\mathbf{BEL}(\mathbf{AG}\neg((\text{バナナが見えない}) \wedge (\text{バナナを所有}))$$

「衝立があるとバナナが見えない」

$$\mathbf{BEL}(\mathbf{AG}((\text{衝立がある}) \supset \neg(\text{バナナが見える}))$$

これを相手の意図に投影して、次のような意図を生成する。

「自分だけに見えているバナナは楽に奪い取れる」

$$\mathbf{INTEND}(\mathbf{AF}\ \text{バナナを奪いたい}) \wedge \mathbf{BEL}(\text{ボスザルから見えていない}) \supset \mathbf{INTEND}(\text{そのバナナを奪う})$$

4. 終わりに

本稿では、信念、願望、意図といった「心の理論」の重要な要素と考えられる心的状況を BID 論理でのモデル化および説明が可能であることを確認した。現在のところ、チンパンジーの他者意図理解の実験結果を Jadex シミュレータに実装中であり、詳細なシミュレーション結果については講演会で報告する。

奈良女子大の新出氏には日頃から議論いただき貴重なコメントを頂いています、ここに感謝いたします。

参考文献

[Premack 78] D.Premack and G.Woodruff: Does the Chimpanzee Have a Theory of Mind?, The Behavioral and Brain Sciences,4,pp.515-526 (1978).

[Wimmer 83] H.Wimmer and J.Perner: Beliefs about beliefs :Representation and constraining function of wrong beliefs in young children 's understanding of deception. Cognition, 13, pp.103-128(1983).

[Hagiya 94] 萩谷昌己: ソフトウェア科学のための論理学, 岩波書店 (1994).

[Bratman 94] M.E.Bratman:意図と行為—合理性、計画、実践的推論、、門脇 俊介, 高橋 久一郎 (訳)、産業図書 (1994).

[Rao 91] A.S.Rao and M.P.Georgeff: Modeling Rational Agents within a BDI-Architecture, Int. Conf. on Principles of Knowledge Representation and Reasoning ,pp.473-484(1991).

[Hare 2000] B.Hare, J.Call, B.Agnetta and M.Tomasello: Chimpanzees know what conspecifics do and do not see, ANIMAL BEHAVIOUR, 59, pp.771-785 (2000).