3I3-10

自己観測原理: 他者認知の数理的枠組

Self-observation Principle: Mathematical Framework of Recognizing Others

牧野 貴樹[*1] 合原 一幸[*2]
Takaki Makino Kazuyuki Aihara

[*1]東京大学 総括プロジェクト機構 [*2]東京大学 生産技術研究所
Division of Project Coordination, Tokyo University Institute of Industrial Science, Tokyo University

We briefly describe *mentalizing*, or understanding others' mental states, in an active research area of cognitive science. To overcome inaccessiblity of others' mental states, theoretical studies assume some mentalizing mechanism in the brain, including simulating others' behavior within knowledge of behavior of the self. We also present our computer-simulation study that tackles the role of mentalizing in a social environment, which examines behavior of agents based on reinforcement learning in Iterated Prisoners' Dilemma games. The results show that agents that choose actions using the estimated policy (corresponding to the mental state) of the co-player, achieve higher cooperation rates than control agents do, which choose actions using only the expected action of the co-player, or only the recent history of game plays.

人間が、日常の中で認知するものの中で、もっとも高度で複雑なものは、ほかの人、すなわち、他者である。ただたんに、顔をみて名前がわかるというようなことではなく、相手の行動・振る舞い・表情などから、いまどのような状態で、何を考えているのか、ということが、ふつうの人であれば、ごく自然にわかる。人間がもつ、他者の心を認知する能力は、Mentalizing [Frith 03] と呼ばれる ("Theory of Mind (心の理論) [Premack 78]"、あるいは "Folk Psychology (素朴心理学) [Stich 94]" とも呼ばれることもある)。Mentalizing がとりたてて高度で複雑な認知であると思っていない人も多いかもしれないが、数理モデルとして考えると、その複雑さが浮き彫りになる。他者がなぜ高度で複雑な認知対象なのか、そして、その他者を認知するために、人間の脳内でどのようなことが行われているのか、を知ることは、認知症を考えるための新たな出発点となるのではないだろうか。

本稿では、Mentalizing を概説したうえで、Mentalizing を計算機によってシミュレーションする研究について紹介する。

1. Mentalizing とは

ある子供が、ケーキ屋の前で大声を出しながら親に八つ当たりしている場面を見たとしよう。普通の人であれば、たちどころに、子供がケーキを食べたいという状態にあり、また、その欲求を親がかなえてくれないことに怒っている、という気持ちであることがわかるであろう。このようなことがわかることは、Mentalizing の一例である。人間は、他者と接するとき、日常的に Mentalizing を行い、その結果を利用して、他者とかかわりあっている。

しかし、改めて考えてみよう。「食べたい」「怒っている」という心の状態そのものは、子供の内部の状態である。そのような内部の状態は、直接に観測することはできない。外から観察できるのは、そのような状態の結果引き起こされる行動 (イヤイヤする) や、表情 (歯を食いしばっている) だけであるが、さまざまに異なる心の状態から、同じような行動や表情が出てくるため、内部の状態についてすぐにわかるわけではない。それでもわかることは多いが、それは、我々人間が、さまざまな行動や表情などを、心の状態を基準として分類して学習し、認知しているからである。実際に目に入っている行動 (首を左右に大きく振る) や、顔の表情 (口を大きく横に開いて、合わせた歯を見せている) から、そのような分類をして、心の状態を推定することは、簡単ではないことがわかるであろう。さらに、その心の状態は、子供が受けるさまざまな刺激や、子ども自身の心の動きのため、時々刻々と変化してゆく。Mentalizing とは、このような非常に複雑な「他者」というダイナミカルシステムの内部状態を推定することに他ならない。先の例でいえば、「子供が『ケーキを食べたい』という状態にあるとわかること」は、まさに、子供というシステムの内部変数を推定したことに相当する。加えて、人間は、この推定した内部変数を利用して、システムの将来の挙動を予測したり (たとえば、そのうち泣き出すだろう、など)、システムに入力を与えて挙動を制御したり (チョコを渡して子供を落ち着かせる、など)、といった能力までも持っている。

このことが、どれだけ驚異的な能力であるかは、反応炉や神経細胞といったより少数の内部変数で表せるはずのシステムを相手に、研究者や技術者が悪戦苦闘していることを考えれば、よくわかるだろう。第一、他者の内部にどのような状態変数があるのか、ということすら、そもそも知る方法がないのである。そのようなブラックボックスのシステムの挙動予測や制御は簡単ではない。入力なしのダイナミカルシステムの場合では、元のシステムの次元の 2 倍よりも大きくなるよう、遅れ時間系からの再構成空間を作り、そこでの近似関数を学習することで、そのシステムのふるまいを再現できることがわかっている [Takens 81] が、しかし、この近似関数のパラメータ数は膨大なものになる。そもそも、人間ひとりの内部状態が何次元の空間で表現できるのかは、大きすぎて見当もつかない。上の例を考えても、目撃者は、「子供が親の内部状態を推定した内容」まで一瞬で推定できるのであるから、他者の内部における他者の推定という再帰的構造 (高次の意的スタンス [Dennett 04]) が表現できるような空間でなければならないと思われる。

また、Mentalizing の難しさと重要性を示すものとして、自閉症の存在がある [Baron-Cohen 95]。自閉症とは、コミュニケーションや社会的能力の障害であり、Mentalizing の能力が損なわれることが原因であると言われている。自閉症の有名な症例

連絡先: 牧野 貴樹: 277-8568 千葉県柏市柏の葉 5-1-5
Tel: 04-7136-3973
e-mail: mak@scint.dpc.u-tokyo.ac.jp

[Sacks 95] では、計算・図形・推論といった他の知的能力に関しては人並み以上である (実際、博士号を持ち活躍している) のに、他者の心に共感することができない。人間同士の感情の交流、心のふれあいやだましあいといったことが理解できないので、他者の行動に関して厖大な経験のライブラリーを作り、それをもとに論理的に他人とのやり取りを構築しているが、それでも表面的にしか社会に参加することができない。そんな彼女が自らを「火星の人類学者」と例えていることは、Mentalizing なしで他者と付き合うには、未知の反応炉やなにかを相手にするように、知的能力をフル活用して推論しつづけなければならない (そして、それでも難しい) ことを端的に表しているといえよう。

このように、普段は意識していないけれども非常に高度な能力である Mentalizing を、人間はどのように獲得しているのか、そのメカニズムを探ることは、認知科学研究の大きなテーマのひとつである。

この問題に関しては、20 年以上にもわたり、Simulation theory と Theory theory の 2 つの学説が真っ向から対立している。Simulation theory とは、「人間は、脳内で自己をシミュレーションすることを通して Mentalizing する」という考え方 [Gordon 86] である。ミラーニューロンの発見 [Pellegrino 92] (自分が把持などの動作をする時にも、他者が同じ動作をするときにも活性化する神経細胞。サルの脳で発見された) が大きな後ろ盾となり、支持を集めている。一方、Theory theory とは、「人間は生まれつき Mentalizing (他者の理解) に必要な推論の理論を脳に備えている」という考え方である [Carruthers 96]。新生児模倣 (生後数日〜数週間の新生児が、顔の表情を模倣する現象) [Meltzoff 77] という、Simulation Theory では説明できない現象があることから、こちらの仮説もまた有力である。

最近は、両者の仲立ちとなるような説が提唱されるようになってきた。たとえば、筆者らの提案する自己観測原理 [Makino 03] は、単なる自己の行動のシミュレーションではなく、自己に関する観測の変化の予測を他者に適用することで、他者の予測が可能になるという考え方である。また、Shared Circuit 仮説 [Hurley 07] は、自己のなかの、知覚と行動で共有される予測回路が、他者を認知する時にも活動することで、共通の推論が行われる、というものであり、自己観測原理の発生段階的説明とも解釈できる。こうした新しい仮説がどの程度正しいかについては、今後の研究の発展を待つ必要があるだろう。

2. 計算機によるシミュレーション研究

このような研究をするためには、必ずしも人間の被験者を相手にしなければいけないわけではない。どのような他者理解のメカニズムをもっていれば、どのような行動につながるのかを、仮説ごとに計算機シミュレーションを実施する、という方法論でも可能である。しかし、問題の性質上、現実に迫る精巧なシミュレーションモデルを構築することはまだ難しく、高度に抽象化した条件の下で、振る舞いを観察し、理論を検証する、という形にならざるを得ない。ここでは、そのような研究のひとつとして、我々が取り組んでいる強化学習による相互理解の計算機シミュレーションの研究を紹介する [Makino 06]。

2.1 繰り返し囚人のジレンマ

囚人のジレンマゲーム [Tucker 50] は、人間の社会的行動を説明するために考えられた枠組であり、経済学や進化生物学でよく利用されている。X と Y、2 人の参加者が、同時に「C (協力)」か「D (裏切り)」のどちらかを選び、その結果によって、表 3. のような利得 (報酬) を得るとする。ここで、両参加者が各々自分の利得を最大にする行動を選ぶと考えよう。X は、Y の行動を事前に知ることはできないが、Y が C を選ぶと仮定すると、X は C を選ぶとき (利得 3) よりも D を選ぶとき (利得 5) ほうが X の利得が増える。また、Y が C を選ぶと仮定しても、やはり自分は C を選ぶとき (利得 0) よりも D を選ぶとき (利得 1) のほうが利得が増える。よって、X の報酬を最大化する行動は D (裏切り) となる。Y も同様の推論をするので、結果的に双方が D を選び、お互いに 1 ずつの利得しか得ることができない。もし両者が C (協力) を選んでいれば、双方に 3 の利得が入ったはずであるが、各々が自己の利益だけを最大化しようとすると全体の利益を損なってしまうのである。このような、個人の利益と全体の利益が相反する状況が、共犯の疑いで逮捕された 2 人の容疑者が別々に自白を迫られるという例 (黙秘=協力、自白=裏切り) で最初に定式化されたため、囚人のジレンマと呼ばれている。

このようなジレンマは、囚人に限らず、様々な現実の例でも起こる。特に、軍拡競争、量販店の値引き合戦などでは、同じ相手に対して、終わりなくこのジレンマが繰り返される。これをモデル化したものが、「繰り返し囚人のジレンマ」(IPD) と呼ばれるゲームである [Axelrod 84]。面白いことに、繰り返しが入る場合には、先ほどの例とは異なり、参加者が個人の利益だけを追求しても、相互の協力行動に至る場合がありうる。有名な例として Tit-for-tat (しっぺ返し、TFT) と呼ばれる戦略がある。TFT では、最初の一回は C を選び、それ以後は前回の相手の行動をそのまま繰り返す。これは実質的に、相手が協力を選んだ後は C を選んで「返礼」し、相手が D を選んだ後には D を選んで「罰する」ことに相当する。相手にとって C を選ぶほうが長期的に得になる環境を作り出すことで、TFT エージェントは相互の協力を作り出し、結果的により大きな利得を得られるようにできるのである。

しかし、TFT は学習しないエージェントにおける固定的な戦略の例である。行動を学習するエージェントの場合には、裏切りによる目先の利得増加にとらわれず、相互の協力に至ることは簡単ではない。この研究のゴールは、人間のように、お互いに他者を理解できる学習エージェントが、自発的に協力行動に至るかどうかを調べることである。

2.2 強化学習

強化学習 [Sutton 98] は、正解 (正しい行動選択) についての情報が明示されず、行動の結果の報酬情報だけが与えられる場合に、将来の報酬和を最大化するような行動を学習する枠組である。ここでは、そのなかでも最もシンプルな Q-Learning [Watkins 92] を利用する。

エージェントは、各時間ステップ t において、環境の現在の状態 $s_t \in S$ を観測し、ポリシー $\pi : S \rightarrow A$ に基づいて行動 $a_t \in A$ を選択する (図 1)。この s_t と a_t に応じて、(ときには確率的に) 環境から報酬 r_t と次の状態 s_{t+1} が与えられる。このとき、将来の r_t の和が大きくなるような行動を考えたいのだが、未来の報酬を単純に合計すると発散することがあり扱いにくいので、価値として割引報酬和 $V = \sum_{t=0}^{\infty} \gamma^t r_t$ を考える。ここで、$\gamma (0 \leq \gamma < 1)$ は割引率と呼ばれる、学習アルゴリズムのパラメータである。このとき、価値 V を最大にするような π を選ぶことが、強化学習の目的である。

Q-Learning では、行動価値関数 $Q(s, a)$ を利用して学習する。$Q(s, a)$ は、状態 s において行動 a をとり、その後は最適な行動を選択した場合に期待される価値である。関数 Q が正しいとすれば、最適な π は、それぞれの状態 s において $Q(s, a)$ を最大にする a を選ぶことである。実際には、適当な値の関数 Q からスタートして、次のような学習則を使い、そのとき

の Q に基づく π をもとに行動を選びながら関数 Q の値を更新していくことで、関数 Q が正しい値に収束し、最適な π が得られることが証明されている [Littman 94]。

$$Q(s(t),a) \leftarrow Q(s(t),a)+\alpha\left[r(t)+\gamma \max_{a'} Q(s(t+1),a')-Q(s(t),a)\right] \quad .$$

しかし、その収束が保証される条件として、環境のマルコフ性 (つまり、状態 s の中に将来の環境変化を予測できる情報がすべて含まれること) が満たされる必要がある。実際上、IPD のようなマルチエージェント問題の場合には、これは不可能である。なぜなら、相手のエージェントも環境の一部であると考えられるため、マルコフ性を満たすには、相手のエージェントの学習状態すべてが s の中に表現できるような状態空間 S が必要になるが、状態空間 S_X, S_Y をもつ 2 体のエージェント X、Y がお互いに相手の学習状態すべてを表現しようとすると、S_X の中に Y のポリシー (S_Y から A へのマップ) が表現されなければならない。逆も同様であるので、状態空間が無限の入れ子構造となってしまい、必要な S の次元が発散してしまう。また、そもそも、他者の内部の学習状態を、外部から完全に把握できるということもありえない。一般には、マルチエージェント環境では収束の保証はあきらめ、観測できる情報から他者の限定されたモデルをなんらかの形で推定し、そのモデルの情報を S に含めた形で学習することになる。

従来のマルチエージェント強化学習研究の多くは、他者が次に選ぶと予測された行動の情報だけを S に含むものであった [Panait 05]。しかし、その場合、同じエージェント同士で協力行動を学習するためには、何らかの仕掛けが必要であった。我々は、S により詳細な他者の内部状態に関する情報があれば、何の仕掛けもなくても協力行動に至るのではないかと考え、実証のためのエージェントを設計した。以下では、その設計と実験について説明する。

2.3 他者の状態を推定するモデル

まず、他者の状態を推定するエージェントの前提として、他者の状態を利用しない、単純なエージェントを考えよう (Level 0 と呼ぶことにする)。Level 0 にとって、他者はあくまで環境のノイズ源としか捉えられていない。このようなエージェント同士で IPD ゲームを対戦させても、相互協力に至る確率が低いことが知られている [Sandholm 95]。

ここで、対戦相手が Level 0 であると仮定して状態を推定する Level 1 エージェントを考えてみよう。外部から観測できるのは過去の行動だけであるが、そこから、前回の行動 4 通りのそれぞれに対してどう行動を選ぶかという 4 ビットの戦略ベクトル (SV) を得ることができる。学習による変化を反映するため、以前と異なる行動が観測された場合には、その観測に合うよう逐次的に SV を更新していくものとする。この SV は、Level 0 が持つ関数 Q の、各状態において協力行動をとる場合の評価値と、裏切り行動をとる場合の評価値の大小関係だけを抜き出したものと捉えることができ、その意味でも他者の内部情報を観測できる情報だけから再構成したものと考えられる。そこで、Level 1 エージェントは、この SV と前回の行動の直積を状態として学習し、行動決定する。直感的には、相手の取りうる戦略のひとつひとつに対応して、関数 Q を用意して、ベストな応答戦略を学習するものと思ってよい。

我々は、これらに加えて、他者の内部にある自己のモデルまで含めて再帰的にモデル化する、Level 2、Level 3 などのエージェントモデルも同時に構築したが、詳細については、論文を参照されたい [Makino 06, Makino 07, 牧野 07]。

表 1: 囚人のジレンマにおける、参加者 X と Y の行動に応じた利得の表

X \ Y	C (協力)	D (裏切り)
C (協力)	3 \ 3	0 \ 5
D (裏切り)	5 \ 0	1 \ 1

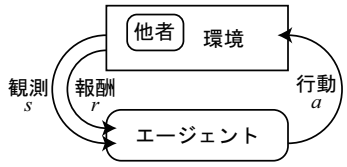

図 1: 強化学習におけるエージェントと環境の関係。他者は環境の一部として扱われる。

2.4 シミュレーション結果

次に予測される他者の行動だけではなく、他者の内部状態の推定を利用して行動を決定するという方法のメリットを確かめるために、同レベルのエージェント同士での IPD 対戦を 1000 試行ずつ実施し、得られる平均利得を比較した。ただし、同じ行動系列の繰り返しから脱出できるよう、弱いノイズを導入した (行動が確率 1×10^{-4} でランダムに置き換わる)。

図 2 は、前節で説明した Level 1 エージェント (ここでは policy-based と記す) と、統制条件として、Level 1 と基本構造が同一ながら、SV そのものではなく SV から予測される次の行動を利用して学習したもの (action-based)、Level 0 エージェント が利用する過去の履歴の長さを変化させたもので、学習開始からの平均利得の時間変化を比較したグラフである。どのエージェントも、最初の間は得られる平均報酬が低いが、ある時点を境に平均利得が上昇し安定する。この安定に至るまでの期間は、S のサイズによって決まっているようである。しかし、安定したあとに得られる報酬は、エージェントの設計によって大きく異なることがわかる。policy-based エージェントが、ほぼ 3 近く (表 3. において、相互に C=協力を選んだ場合に得られる利得) に達しているのに対し、ほかのエージェントは、3 よりも幾分低い値にしか至らない。これは、試行のうちのいくつかで、相互協力に至らず、裏切りの行動を含む形で安定した行動系列に至るケースが見られたためである。確実に相互協力に至るということは、他者に関する情報をもとに、適切な行動の学習ができたことを示唆している。

この実験から、他者の内部状態に関する推定能力があることで、社会的な利他的行動が促進されることが示唆された。今後、IPD 以外の条件などで、さまざまな実験をすすめる必要がある。また、ヒトやサルなどの社会的行動がみられる動物のふるまいとの比較も検討していきたい。

3. おわりに

本稿では、Mentalizing、すなわち他者の理解という人間の認知の仕組みに関する認知科学研究と、その理論を検証するための計算機シミュレーション研究を紹介した。他者を認知するという日常的な活動の中に、脳のもっとも複雑な機能がフル活用されていることを知ると、認知という、われわれがふだんなにげなく行っていることへの見方も広がるのではないだろうか。

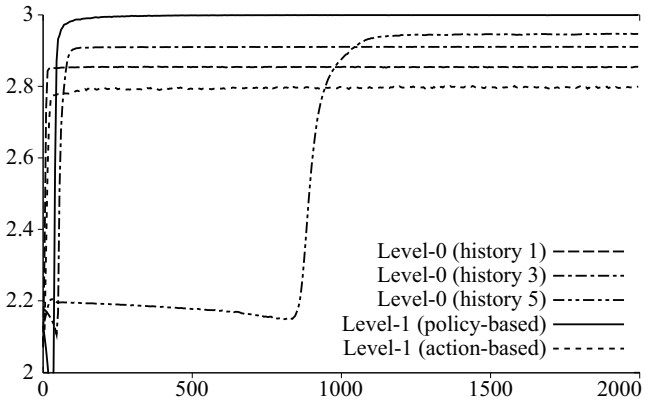

図 2: IPD ゲームにおける平均利得 (縦軸) の時間変化。パラメータは $\gamma = 0.99$ である。横軸は時間ステップ ($\times 10^4$) 。

参考文献

[Axelrod 84] Axelrod, R.: *The evolution of cooperation*, Basic Books, New York (1984)

[Baron-Cohen 95] Baron-Cohen, S.: *Mindblindness: An Essay on Autism and Theory of Mind*, MIT Press (1995)

[Carruthers 96] Carruthers, P.: Simulation and self-knowledge: a defence of the theory-theory, in *Theories of theories of mind*, Cambridge University Press, Cambridge (1996)

[Dennett 04] Dennett, D.: Three kinds of intentional psychology, in Heil, J. ed., *Philosophy of mind: A guide and anthology*, Oxford University Press, UK (2004)

[Frith 03] Frith, U. and Frith, C. D.: Development and neurophysiology of mentalizing, *Phylosophical Transactions of the Royal Society B*, Vol. 358, No. 1431, pp. 459–473 (2003)

[Gordon 86] Gordon, R.: Folk Psychology as Simulation, *Mind and Language*, Vol. 1, pp. 158–171 (1986)

[Hurley 07] Hurley, S.: The Shared Circuits Model: How Control, Mirroring and Simulation Can Enable Imitation, Deliberation, and Mindreading (2007), To be published in *Behavioral and Brain Sciences* (in press)

[Littman 94] Littman, M. L.: Markov games as a framework for multi-agent reinforcement learning, in *Proc. of ICML 1994*, pp. 157–163 (1994)

[Makino 03] Makino, T. and Aihara, K.: Self-observation Principle for Estimating the Other's Internal State, Technical Report METR 2003–36, the University of Tokyo (2003)

[Makino 06] Makino, T. and Aihara, K.: Multi-agent Reinforcement Learning Theory to Handle Beliefs of Other Agents' Policies and Embedded Beliefs, in *Proc. of AAMAS-2006* (2006)

[Makino 07] Makino, T. and Aihara, K.: Policy-based Beliefs and Recurrent Embedding: Algorithms for Multi-agent Reinforcement Learning with Embedded Beliefs (2007), In submission to *Autonomous Agents and Multi-Agent Systems*

[Meltzoff 77] Meltzoff, A. N. and Moore, M. K.: Imitation of facial and manual gestures by human neonates, *Science*, Vol. 198, pp. 75–78 (1977)

[Panait 05] Panait, L. and Luke, S.: Cooperative Multi-Agent Learning: The State of the Art, *Autonomous Agents and Multi-Agent Systems*, Vol. 11, pp. 387–434 (2005)

[Pellegrino 92] Pellegrino, di G., Fadiga, L., Fogassi, L., Gallese, V., and Rizzolatti, G.: Understanding motor events - a neurophysiological study, *Experimental Brain Research*, Vol. 91, pp. 176–180 (1992)

[Premack 78] Premack, D. G. and Woodruff, G.: Does the chimpanzee have a theory of mind?, *Behavioral and Brain Sciences*, Vol. 1, pp. 515–526 (1978)

[Sacks 95] Sacks, O. W.: *An Anthropologist on Mars*, Knopf, New York (1995), 邦訳『火星の人類学者』(訳:吉田利子、早川書房)

[Sandholm 95] Sandholm, T. and Crites, R.: Multiagent Reinforcement Learning in the Iterated Prisoner's Dilemma, *Biosystems, Special Issue on the Prisoner's Dilemma* (1995)

[Stich 94] Stich, S. and Ravenscroft, I.: What is Folk Psychology?, *Cognition*, Vol. 50, pp. 447–468 (1994)

[Sutton 98] Sutton, R. S. and Barto, A. G.: *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA (1998)

[Takens 81] Takens, F.: Detecting strange attractors in turbulence, in Rand, D. A. and Young, B. S. eds., *Dynamical systems and turbulence, Vol. 898 of Lecture notes in mathematics*, pp. 366–381, Springer-Verlag, Berlin (1981)

[Tucker 50] Tucker, A. W.: A Two-Person Dilemma, in Poundstone, W. ed., *Prisoner's Dilemma*, Doubleday, New York (1950)

[Watkins 92] Watkins, C. J. C. H. and Dayan, P.: Technical Note: Q-learning, *Machine Learning*, Vol. 8, No. 3/4, pp. 279–292 (1992)

[牧野 07] 牧野 貴樹, 合原 一幸：他者理解をシミュレーションする, シミュレーション, Vol. 26, No. 3, pp. 171–175 (2007)