記号と意味とが相互作用する場のモデル
A model of field where signs and meanings interact with

犬童 健良†
Kenyro Indo
†関東学園大学
Kanto Gakuen University
kindo@kanto-gakuen.ac.jp

概要
記号表現は言葉や記号で考えを表明するが，その表された意味と相互作用する場として捉えられる．記号表現により思考は文脈から切り離され，自律的なエージェントとして分散する．そのため心内エージェンシーは免疫ネットワークとしての自己決定性と自己言及性を有する．一方記号表現の自壊はその非凸性を意味する．本研究は認知的領域の非凸性を阻止する耐戦略的メカニズムについて論じる．また阻止条件を用いて認知的領域を生成する実験について報告する．

キーワード:記号と意味の相互作用，記号的分離可能性，免疫ネットワーク，認知的阻止，クジ比較，社会選択

1. はじめに

本研究では言葉や記号を用いて考えを表明すること，またその意味的な相互作用を抽象化して記述するための理論的なモデルを提案する．まず意味作用ないし意識というものを認知空間のオペレーター（作用）が認知空間の対象を変化させるプロセスと考えてみよう．記号論ではシニフィアン（Sa），つまり記号として作用する側とシニフィエ（Sé）との対応関係，あるいはより短く記号部と意味部のペアを意味作用の基本ユニットとする．これらの用語はソシュール以来の記号論の伝統であるが，一方，記号と意味の対応関係は恣意的に発生するものではなく，一定の安定性をもった構造として定まるという観点をその背景にもつ．

記号をモノのように考えれば，モノ同士のときと同様，記号と記号が衝突して動く方向が変わったり，分解して中のものが飛び出たり，化学反応を起こして，全体として滑らかに変化するが，質的に不連続な変化（カタストロフ）を引き起こすこともある．もちろん書いた文字や話された音声自体が物理的に相互作用するのではなく，それに対応して認知表象が生成したり消滅したりする．直観的には，ある記号（シンボル）を認知してその意味を知ること，あるいはその逆に思ったことを記号で表現したりすることが，記号を用いた思考である．つまりその記号とは別の記号や音声・映像・匂いなどのイメージ，行動選択の機会・必要性・危険性など何らかの知覚や解釈が導出される．こうした一連の変化が起こるのが意味作用の場である．それは表象（記号）が配置される認知空間の別の，しかし等価な表現であり，また意識と私たちがふつう考える対象に相当する．しかしながら，量子場理論の解釈のしかたにならえば，記号はあらかじめ外的に与えられたものではなく，意味作用の場において生じた（エネルギー状態の）励起として検出される(Stone, 2000)．記号は意味作用の場に生じた励起であり，有徴化され，気になるものである．

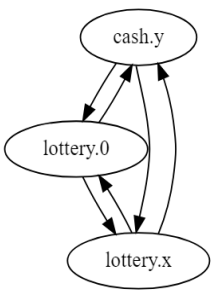
以下に示すネットワークモデルはより直観的に理解できるだろう．記号とその意味は，カギとカギ穴の関係が連鎖するように時間発展する．

(1) Sa(1) → Sé(1) →Sa (2) →Sé (2) →…

各時点 t において記号として作用する Sa (t)とその効果として生じる意味 Sé (t)との間には自由度の非対称性がある．記号部 Sa (t)が意味部 Sé (t)より自由度が高く，部屋ごとの暗号が次々と解かれていくイメージである．また個々の Sa→Sé, あるいはそのつながり・鎖(chain)は，条件文のモデルとしても解釈しうる．

数理的には Sé(t)は抽象的な多様体上のある点ないしそのファイバーであり，直観的に，図 1 に示すようなネットワークのノードとそのポテンシャルとして理解できる．その解釈は認知的コネクターで結合された認知空間の領域であり，メンタルスペース理論(Fauconnier,1997)における操作の非文法的側面である．

図 1 認知的ネットワーク



式(1)中の矢印はネットワーク内では（有向）エッジに対応し，ポテンシャルの勾配に沿って動く粒子，あるいは全体解いて波のように動くものと解釈される．図 1 ではクジ選択の可能な結果がノードとして示され，また可能な結果のペアを比べることがエッジに対応する．つまり図 1 は「X 円が確率 P で当たるクジ」と「現金 Y 円がもらえる」という 2 つのオプションを選択する問題を理解する意味作用の場を表現する．

異なるクジの選択結果を比較するときの認知空間は特殊相対性理論と類似する．つまり選んだ行為を一つの動く座標系として，選ばなかった行為に対応する別の座標系の動きがどう見えるかということである．クジ比較の例での基底状態は，その不変量としての気になる度合いを保存する．自明なネットワークの均衡として，同じエッジは同じように気になる．しかしより一般的には，それは場から影響を受けて変化していく．

もしネットワークが安定（均衡）していれば，その動きはポテンシャルの高い方から低い方への流れとして容易に理解できる．クジ選択者は「X 円が当たる」ときの「選ばなかった現金 Y 円」よりも「クジが外れて 0 円」だったと考えたときの同じ「選ばなかった現金 Y 円」をより気にするかもしれない．

ネットワークのエッジやポテンシャル自体が変化すると，もちろんその動きはより複雑になる（ニューラルネット学習の進歩を考えてみよう）．ノード間のポテンシャル差はエッジの気になる度合いに対応し，エッジに沿った移動の費用とみなせる．実際，(lottery.X→cash.Y) と(lottery.0→cash.Y)から含意される(lottery.X→lottery.0)という潜在的な注目の移動が含意されるが，これは最短路アルゴリズムにおけるポテンシャルの更新のアナロジーである．犬童（2013, 2014a）の行った実験では気になる度の差の競合パタンが検出される．これは全回答者に共通のラングやサブグループのパロールと考えうる．6 節ではその生成実験を論じる．

2. 心の中の他者

生きた認知システムは認知領域の滑らかな表面，凸性(convexity)を生み出そうとする潜在的な傾向をもっていると思われる(Gärdenfors, 2004)．それは知ろうとする対象を制限するバイアスとして異なる分野の間に散ばって現れることからも示唆される．認知心理学における確証バイアス，統計学的推論における選択バイアス，科学哲学にけるラムジーテスト，経済学・経営管理における満足化原理はその例である．

直観的には，凸性とは仲間同士を混ぜても他者（あるいは異物や敵など）にはならないという性質である．トポロジー的には，もしそのような組み合わせがつかれば，その認知領域が穴，くぼみ，あるいは凹みをもっていることを意味する．自己に属するカテゴリーと属さない他者とを区別する手続きは，抽象化したメカニズムとしては自己－他者の識別，あるいは認証を行う免疫と同じである．またそれはカテゴリー判断や命題の真偽決定の基礎となる．また一定の安定した意味作用（意味作用の均衡）を保証できなければ，言語的なコミュニケーションは不可能である．

ところで生成 AI で製作されたチャットボットはヒトのように言語（書かれた言葉）で会話し，感情や人格すらもっているかのようにふるまうことができる．しかし事実とその解釈は区別しなければならない．半世紀前にワイゼンバウムの Eliza がすでにそうであったように，ヒトは感情をもたない他者にも心を見出すことがある．また AI は感情をもたぬゆえ平気で嘘をつくことができる（これは論理としては正しい）．この場合，騙す知能が騙される知能に勝っている．もっとも Eliza の仮想人格 DOCTOR のように，そのスクリプトがヒトに偽の記憶をもつ動機がある状況（カウンセリング）を作り出し，ヒトがそれを無自覚に受容するのだとも考えうる．

AI がヒトのプロンプトに応答して言葉を生成しているということ自体は一つの事実だが，「ヒトと区別できない（ゆえに知能をもっている）と思った」ということは，それとは異なる別の事実である．後者の事実は通常，その元となった事実についての解釈であると考えられる．同様に，記号表現された自分自身の思考はその都度，その意味作用を再生産するためにエージェント化されると考えることができる．事実は認知空間に送り出されたエージェントである．式(1)が表す思考活動によって成立する認知領域は，したがってエージェンシーとしての性質を帯び，心の社会を形成する．それらのインタラクションは時間的に遠く離れた記号表現の間ですら可能である．

3. 記号的分離と心のエージェンシー

分散して長距離依存する心のエージェンシーは免疫ネットワーク(Vaz & Varela, 1978)と同じく，自己決定性と自己言及性を有するフィードバックループとなる．

それはいわば嘘つきパラドックスの動学化である．記号を用いて考えを表明する行為は，エージェンシーのもつ投機的性質によって，表現される意志や思考そのものを破壊する可能性がある．またそれを防ぐため免疫と似たしくみを備えていると推察できる．

記号（シンボル）は広義の文脈としてのエージェントの環境全体から切り離され，一定の境界をもった範囲としてその意味を理解される．本研究ではこれを記号的分離可能性（symbolic separability）と呼ぶことにする．記号的分離は凸解析における分離定理のアナロジーであるが，認知領域におけるその成立は自明ではなく，探求するための確固とした枠組みはまだないと思われる．先駆的には免疫ネットワークやカタストロフ理論が参考になる．ルネ・トム(Thom, 1977)は言語的な意味をモデル化するためトムは力学系の特異点に至る安定軌道，つまりカタストロフに着目した．カタストロフを含む安定な閉軌道パタンを生物の動作の語彙（動詞）に対応させている．

記号表現における自壊的な意向は，より一般的に言うと，文脈における認知領域の囲い込みの失敗，その結果としての認知領域の非凸性である．実際，限定された合理性(bounded rationality)とは人が複雑な問題をそのまま解こうとすると失敗することであり，またそれを人は避けようとする思考と行動の傾向としての満足化(satisficing)を意味する（Simon, 1976）．

記号と意味との相互作用は記号の使用される文脈に依存する．記号論者は意味作用を固定された対応関係（符号化規約）のように扱うのを止めて，物理や生物のメタファーから構造生成や文脈埋込のモデルを探求したが，認知科学でもニューウェルとサイモンの物理記号系やフォーダーの思考の言語仮説といった心的表象の計算論から離れて，生態学的知能や埋込まれた学習が研究されるようになった．フォーダーは記号と意味の関係を信頼できる因果関係として論じた(Fordor, 1998)．これは脳は統計学的推論のエラーとしての他者性を抑えるしくみをもつことを示唆する．

記号使用が埋め込まれる文脈には，記号だけでなく，記号の使用者，伝達を意図する相手，彼らの記憶や理解や反応，またそれを第 3 者が知ることで生じる潜在的な影響が含まれる．その拡がりは明示的に伝達行為が行われている場すら超える（たとえば外為市場や債券市場は金融機関のトレーダーが電話などの通信手段を用いて直接やりとりするネットワークのことだが，経済学者の言う市場はそれとは異なる）．しかし一定の安定した意味作用（意味作用の均衡）を保証できなければ，言語的なコミュニケーションは高々無意味な連想ゲームになり，意味理論として物足らない．

以降の節では免疫的エージェンシーとその記号分離，認知的阻止のメカニズムがその抽象化された認知領域上のオペレーターから必然的に生じるということを，クジ比較における可能な結果間の注目ネットワークモデルを用いて具体的に説明する．

4. 認知空間における阻止メカニズム

自己に属するカテゴリーと属さない意味作用を確定するには，文脈のどこまでが記号の意味に影響するかを，ある境界で囲い込む必要がある（おそらくフォーダーが意味の全体論に反対する観点とも一致する）．またその結果，エージェントとその環境との相対的な定義が生じる．実際，満足化は当面の問題解決・目標達成に影響する情報を選び，無関係な情報を捨てることを繰り返しながら問題表現の均衡に達するような認知プロセスである．思考の表明は不確実な結果と不明瞭な境界を伴う意思決定，あるいは他者や自分自身に対してコミットするための（戦略的な）ゲームプレイとして捉えられる．そこで，生きた知能システムはとうぜん記号表現された自己の意志であるエージェント，つまりその認知領域に不可欠な参加者と，そのネットワークにとっての脅威となる異物・他者を区分し，監視したり排除したりする防御的あるいは予防的なメカニズムを備えると考えられる．本研究ではこれを認知的な阻止（cognitive blocking）と呼ぶことにする．またこのメカニズムは，認知的領域の滑らかな変換を要求すると考えるのが自然だろう．それが認知空間の凸性と記号的分離によって支持され，自己という不動点（あるいは定常状態）をもつ特別な認知プロセスの基盤となっているはずだ．

認知的ネットワークはときとしてより複雑な，必ずしも凸ではない領域をなす．自律化したエージェンシーは自己のメンバを他者から区別し異物を攻撃・排除するが，他者を取り込むことで免疫的寛容性を示すこともあり，それによってネットワーク自体を拡張する．いわば敵から身を守る（損失回避する）とともに捕食して生きていく（利得を追求する）必要がある．

5. 二者択一の定理

環境は，エージェントの認知活動が行われる場(field)

として抽象化される．意味作用の場ではエージェントが環境と相互作用し，基底状態からの励起としての記号を発生させる，あるいはその記号自体がエージェントである．簡単化のため，記号と意味の相互作用が発生する場を点の集まりN={1,2, …, n }と考える．各点には場の量（多様体のファイバー）と局所座標系が対応する．各点xは局所座標系と同一視される．計量テンソルのアナロジーとして，ある$x \in N$におけるエージェントの認知活動はその局所座標系における場の量の観測であり，局所座標系同士の変換である．またこの活動の結果エージェントは$x \in N$に移動すると仮定して力学系を考える．そこで認知活動を$x \to y$，$x, y \in N$あるいは$y = F(x)$と書くことができる．

直観的には意識状態・認知表象が場の量に依存した注意配分によってベクトルv^xとして生成される．ただしこの定式化では認知は行為と形式上区別できない．意識的活動では意味作用Fに伴い，その結果として，知った・分かった，知らない・分からないといった自己意識（情報）が伴う．ヒトの場合，その行為は自己意識を伴うのが通常であるが，例外もある．実際，もし生成される認知表象が生成前と同じなら情報に気づかない．

二者択一の定理(Minkowski-Farkas lemma)から，n次元ベクトルvとn行m列の行列Fに対して$Fu = v$となるベクトルuが存在し，かつあらゆるm次元ベクトルpに対し$F^T p \geq 0 \to v^T p \geq 0$（$T$は転置）が成立するか，あるいはどちらも成り立たない．つまりvがFの列ベクトルがなす錐の中にある（つまりその非負成分の一次結合として特定される）．あるいはその代わりにpをいくつかサンプリングして共単調性をチェックしながら支持超平面を形成する認証手続きが考えられる．

図 1 のネットワークのエッジの気になる度はクジ選択問題がエンコードされた認知領域をおおまかに捉えるものと考えうる．犬童（2014a）は実際に実験データから正の差が生じるエッジ同士の競合を検出した．そのペアは多くが[i→j]と[j→k]の形（[ ]内は複数ありうる）で，その同時正差は二者択一定理における共単調性の違反である．すなわちＦの錐に入らない他者，部分経路[ijk]を，反事実として弱く排除する．ただし等号で成り立つ境界例は免疫的寛容を受ける．例えば X=400, Y=300（金額単位：万円），P=0.8 とすると，i=(lottery.0→cash.Y)，j= (lottery.X→cash.Y)，k = (cash.Y→lottery.0)である．i→j，j→k は 23 件と 34 件，等号も含めると回答全体の 46%と 77%にあたる．両方成り立つのは 0 件であった（期待件数 8）．

6. 認知的可能領域を生成する

Gärdenfors(1993)は意味の創発に社会選択の公理系を転用した．仮想選択されるクジに対応する 2 人のエージェントのエッジのペアごとに気になる度を組にして，もし合意できるならそれを尊重するように合成する手続きはこのモデルと同型である．Arrow の不可能性定理から組全域をカバーしようとすると独裁になる（α 独裁）．また認知的阻止を最も注目されにくいエッジを選ぶ関数とする．Gibbard-Satterthwaite 定理から認知的阻止は全域で耐戦略的だと独裁になるか強制・禁止を伴うかのいずれかである（γ 独裁）．独裁でない領域は，α 可能なら捕食，γ 可能なら免疫と解釈される．これらは複雑に相互作用し，とくにα 可能かつγ 独裁であるなら自壊的な記号表現が予測される．ポスターでは Indo（2014b）の方法を用いた生成実験も紹介する．

文献

Fauconnier, G. (1997). Mappings in Thought and Language. CUP（『思考と言語におけるマッピング』坂原茂・田窪行則・三藤博(訳) (2000) 岩波）

Gärdenfors, P. (1993). The emergence of emergence of meaning. *Linguistics and Philosophy*, 16(3), 285-309.

Gärdenfors, P. (2004). *Conceptual Spaces: The Geometry of Thought*. MIT Press.

Fordor, J.A. (1998). *Psychosematics*. MIT Press.

犬童健良(2013). アレの背理における注目と注目の流れ. 行動経済学会誌, 6, 70-73. http://dx.doi.org/10.11167/jbef.6.70

犬童健良(2014a). アレの背理における反事実的注目とリスク選好の認識的安定性. 関東学園大学経済学紀要, 39, 53-80. https://doi.org/10.20589/kantogakueneconomics.39.0_53

Indo, K. (2014b). Parallel possibility results of preference aggregation and strategy-proofness by using Prolog. In *Proc. the 6th International Conference on Agents and Artificial Intelligence – Vol. 2,* 243-248. https://doi.org/10.5220/0004913302430248

Simon, H. A. (1976). *Administrative Behavior 3rd edition*. Free Press（サイモン, H.A. 松田武彦・高柳暁・二村敏子(訳) (1989) 経営行動 ダイヤモンド社）

Stone, M. (2000). *The Physics of Quantum Fields*. Springer（ストーン, M. 樺沢宇紀(訳) (2012) 量子場の物理 丸善）

Thom, R. (1977). *Stabilité Structurelle et Morphogénèse, Deuxième Édition*. InterEditions（ルネ・トム, ジーマン, E.C. 彌永昌吉・宇敷重弘(訳) (1980) 構造安定性と形態形成 岩波）

Vaz, N. M., & Varela, F. J. (1978). Self and non-sense: an organism-centered approach to immunology. *Medical Hypotheses*, 4(3), 231-267 (小泉俊三(訳) 自己と無意味：免疫学への生体中心の一アプローチ」 現代思想 1984 12 月 166-88）