題目：社会性の脳計算の解明、そして人工知能

中原裕之
Hiroyuki Nakahara

理化学研究所 脳科学総合研究センター 理論統合脳科学研究チーム〔〒351-0198 埼玉県和光市広沢 2-1〕
Laboratory for Integrated Theoretical Neuroscience, Riken Brain Research Institute, 2-1 Hirosawa, Wako City, Saitama 351-0198, Japan

1. はじめに

脳科学と人工知能（AI）が交錯するフロンティアはいかに形成されるのか？そして，そのフロンティアで, 社会知性を実現する脳機能はどこまで解明されているのか？本稿ではこれらの疑問に答えるべく，計算脳科学（計算論的神経科学：computational neuroscience）分野で行われている，社会知性を形成する脳計算の研究について，AI 研究をされている読者を念頭に概説する．また，この機会に，計算脳科学により何が明らかになるのかについて触れ，さらに，深層学習と脳科学の関連についても説明する．社会知性に関わる脳計算の研究は脳科学のフロンティアの一つである．この研究を理解するには，脳科学の発展の流れを簡潔にでも押さえておく必要がある．また，この視点は AI 研究で脳の知見を活かそうとするときにも肝要である．なぜなら，脳の知見をうまく活かすには，脳とは「なに」かを知ることが重要だが，それが「いかに」成立しているかを知ることも同時に重要だからである．このような議論に AI 研究者が触れる機会はそれほど多くないのではないかと推察し，その点も意識して書いてみた．それでは始めよう．

2. 脳という情報処理システム

脳は情報処理をする．この情報処理が私たち人間のすべての行動の源泉である．ここで，以下の議論の整理のために 2 点確認しておきたい．まず 1 点目は，以下の記述は人間の脳に語るように書いているが、それは便法である。実際，脳の働きを学ぶのにマウスやラット，サルなどを用いた動物実験による脳科学研究は必須であり，それらの研究から人間の脳の働きについて多くを学ぶことができる．以下の、人間を対象とした語り口は、わかりやすさを重視しているだけである。二点目は，すべての行動の源泉という意味を確認しておきたい. 本稿では「行動」という言葉を最大限広い意味で使っている．たとえば，視覚・認知・運動を含んでおり，言語や推論，さらには情動・感情・社会相

互作用なども含めて使っている．脳というシステムが実現するあらゆるアウトプット・ふるまいを指す言葉として使っていることを了解しておいていただきたい．

脳のことをわかろうとして脳科学研究について、また AI 研究が脳の知見を取り込んだ研究を進めようとするとき，この「わかる」という点に注意が必要である．脳機能には物質原理と情報原理の両方が働く．脳は，遺伝子・分子レベルから，多様な神経細胞や神経伝達物質，さらに局所回路と大局的回路に至るまで，脳という生物的物質として実現されている．それらは物質原理のもとで，脳神経細胞の集団活動のパターンと相互作用により情報処理を実現しているのである．「脳のことをわかる」というとき，「わかる」という言葉を物質原理面で主に使っている脳科学者は多い．一方で，もっと深い意味で脳のことをわかるには，あるいは AI 研究が脳科学に向ける関心に応えるには，この物質原理と情報原理の交錯を踏まえた脳の情報処理についての理解が必要である．それを理解してはじめて，その脳機能がどうして実現されたのかがわかるのである．計算脳科学ではそこに重点が置かれている．

2.1 情報処理システムを理解する 3 水準

脳という情報処理システムを理解するには，あるいはもっと一般的に，ある情報処理システムを理解するには，いわゆるマーの 3 つの水準を併存させた理解を進めることが重要である(Marr 1982, デビット・マー 1987)．三つの水準とは，（A）問題（計算論），（B）表現とアルゴリズム，（C）ハードウェア，である．（A）問題（計算論）とは端的には作動目的である．何の処理を，なぜ，どのような制約条件のもとで実現するのか，その実行可能な方略は何か，である．（B）表現とアルゴリズムは，その入力と出力の表現は何か，その入出力変換のためのアルゴリズムは何か，である．そして，（C）ハードウェアは，その処理を実現するための物理的実装である．

この 3 水準が重要であることは，全く未知の物体の情報処理を理解しようとする仮想的な試みを想像してみれば理解できる．たとえば，宇宙からの未知の物体の情報処理を理解することを想定するとよい．あなたは当然そのハードウェアを調べるだろう（「(C) ハードウェア」にあたる）．しかし，そのハードウェアの配線図が仮にわかったとしても，その回路が何を実現しているかはわからない．ハードウェア上のアルゴリズムと入出力表現がわかれば，どのような計算をしているかはおぼろげに見えてくる（「(B) 表現とアルゴリズム」にあたる）．一方で，たとえばパソコンの CPU のアルゴリズムがわかったとしても，そのアルゴリズムが現在実行しているソフトウェアの働きが何かはわからない．つまりアルゴリズムが何を実現しているのか，それが一体，何のために，なぜ，何を実現しているのか，その実行方略を理解する必要がある（「(A) 問題 (計算論)」にあたる）．

これら 3 つの水準にわたって，脳の情報処理についての理解が進められる．ここで注

記しておきたいのは，それぞれの水準で，ある程度別個に理解を発展させることが可能だということである．実際，脳科学でも，それぞれの水準の研究が並行して進められている．それは，この 3 つの水準に不定性と相補性があるからである．たとえば，足し算が電卓でも算盤でも実現できるように，ある問題(計算論)を 実現するアルゴリズムあるいはハードウェアは複数存在しうる(不定性)．一方で，たとえば，ハードウェアがわかってもそれが実現するアルゴリズムは必ずしもわからないが，ハードウェアは確かに可能なアルゴリズムの実現を制約する(相補性)．そして，この不定性と相補性を踏まえたうえで，問題(計算論)を押さえていくことが，脳の機能理解を進めるときには重要である．これについては次節で詳しく述べる．脳という複雑な働きをするハードウェアを目の前にしたとき，つまり，その物質原理に圧倒されているとき，しばしば私たちは(C)ハードウェアに心を奪われがちだが，脳の情報処理を理解するには、(B)表現とアルゴリズムと(A)問題（計算論）を押さえていくことが重要である．

2.2 脳計算の視点：脳情報処理と行動／認知をつなぐ

脳機能理解において，また一般的な知能の理解や実現においては，(A) 問題（計算論）のスコープを意識することが重要である．AI 研究における汎用型 AI と専門型 AI の議論も，このスコープの観点から理解できる．

問題（計算論）のスコープは狭くすることも広くすることもできる．それは対象とする情報処理の課題によって変わってくる．たとえば，視覚脳科学で「立体視」という情報処理課題がある(Marr & Poggio 1976)．これはコンピュータ・ビジョンの研究課題でもある．人間で言えば，それぞれの目に入ってきた 2 つの画像から，外界の立体構造を再構成する課題である．これは人間の視覚認識の基本機能として重要な課題ではあるが，広義の「行動」のもとでは，物体認識のなかのさらに立体視という課題に絞っているという点で，比較的狭めのスコープで問題（計算論）を明確化している例だといえる．一方で，たとえば「注意」や「作業記憶」（または短期記憶）は，それよりも比較的広めの認知レベルのスコープで問題（計算論）を明確化する．なお，スコープが狭い／広いことと，その問題を解くのが簡単／困難であることは無関係である，ただ，一般には広い方が問題の定義自体が難しくなることが多い．言い換えれば，スコープが広いほど，計算論的に問題を定義すること自体が，研究の重要な一部になる傾向が強い．

計算論的問題を定義する際には，対象となる「行動（や認知)」をどのように計算論として定式化するかが問題となる．スコープが広くなればなるほど，行動（すなわち情報処理として実現されるふるまい）をどのような情報処理と対置させるかが難しくなる．その意味では，物体認識よりは，強化学習で扱われる「行動」――すなわち獲得する報酬を最大化するための意思決定と学習の行動――のほうが，行動の計算論的な定式化が難しい．同様に，強化学習よりも社会知性のほうが定式化に難しさがある．なお，繰り返すが，これらの難しさの程度の差とその情報処理の難しさは別の話で，物体認識の情

報処理のほうが簡単だと言っているのではない. ここでは定式化の難しさを議論しているのである．さらに，定式化の難しさの違いには，その階層性や入れ子構造によるものもある．たとえば，物体認識を単なる対象物の認識（上ではその意味で使った）から，その重要性の認識あるいは社会的状況での認識まで含めて考えれば，それは強化学習や社会知性とも関係してくる．実際，あとで述べる深層強化学習では，そのような研究が主流になりつつある．

脳機能を脳計算の視点から理解するとは，この行動の定式化とそれを実現している脳の情報処理の対応を理解すること，つまり，問題（計算論）を明確にしながら，そのアルゴリズムとハードウェアの関係を見ようとすることである．

2.3 ヒト fMRI と脳計算モデル化解析

脳科学の発展は目覚ましく，現在も急速に発展しつつある．たとえば，カルシウム・イメージングを利用した多数の神経細胞の活動の同時記録や，光遺伝子学の利用により特定の神経細群の活動を励起または抑制する操作，遺伝子操作を利用することによる詳細な神経回路構造の同定などの研究が進んでいる. これらはマウスなどの動物脳による研究である. そこでは 3 つの水準の中で特にハードウェアの理解が現在爆発的に進みつつある．また，これらはアルゴリズムの理解にも影響を与えつつある．

一方で，もう少し長期的な視点に立つと，脳科学の発展はすなわち，問題（計算論）が対象とする認知と行動の範囲の拡大・発展であることに気づく．心理学や認知科学が対象としてきたような心理的・認知的概念が，脳回路・脳活動と結びつけられて脳科学的概念に発展し，さらには，それが脳計算の概念へと成熟する．一例を挙げると，物体認識あるいは視覚神経科学の基礎を築いたともいえる，ヒューベル＆ウイーセルの研究がある(Hubel & Wiesel 1959, Hubel & Wiesel 1962). これらの研究で，彼らは 1981 年にノーベル生理学賞・医学賞を受賞した（大脳半球の研究を進めたロジャー・スペリーと共同で受賞）．彼らは主に麻酔下の動物を用いて，初期視覚野の神経細胞における発火特性の研究をはじめとする，初期視覚から中期視覚に関わる研究を主導し，のちの後記視覚（物体認識）の研究の基礎を築いた．これらの知見はじつはふんだんに深層学習で用いられている．一方で，同じく 1960 年代末に，ヒトに十分近いサルを用いて、サルが実際の行動をしているときの、その行動と神経活動を結びつける研究，いわば現代的な認知神経科学の創始とも言える研究が始まった(Wurtz 1968, Wurtz & Goldberg 1971). この研究は，眼球運動，つまり対象に対して目を向ける神経機構の解明を促進すると同時に，“外界の知覚”から“外界への働きかけ”（運動）の途中に存在する過程，いわば認知・心理の過程を知覚―認知―運動の一連の流れの中に位置づけ，それを脳活動と対応させて研究する端緒となった．これにより，認知神経科学は大きくその研究領域を広げることになった．その後，脳計算へと成熟していった例としては，たとえば短期記憶や注意

に関する研究を挙げることができるが，これらについては後に深層学習に関連して触れる(Wang 2002).

心理的・認知的概念，特に人間の高次機能に関わる心理的・認知的概念は，この 20 年ほどの間に急速な勢いで脳科学的概念へと変貌を遂げようとしている．それを促進してきたのは，1990 年代に始まった磁気共鳴機能画像法（functional magnetic resonance imaging: fMRI）(Ogawa et al 1990)を用いて人間の脳活動を非侵襲でイメージングする，ヒト fMRI 実験である．このヒト fMRI 研究が，ヒト脳機能の理解を大いに進めることとなった．しかし，ヒト fMRI にもその時間・空間解像度にはまだまだ限界があるのは確かで，これがあればすべて解決するというものでもない．一方で，ヒトの行動と脳活動を同時に観察できることは大変な強みである．古来からの文学書や歴史書，哲学書など書物に見られる広汎な人間の心理や行動についての論述，あるいはウイリアム・ジェームズから数えれば 100 年以上の長きにわたる心理学，そして人工知能の命名がなされたダートマス会議（1956）がほぼ開始点と考えられる認知科学，これらが扱おうとした「行動」（概念・心理・認知）を「脳活動」と同時計測しながら検証できるのである．後で述べる社会知性を司る脳機能の解明でも，ヒト fMRI が決定的な役割を果たしている．

ヒト fMRI による行動と脳活動の同時計測だけで，すぐに脳計算のしくみがわかるわけではない．脳科学的概念を脳計算の概念にしていくことが必要である．その強力な手段となるのがモデル化解析（model-based analysis）である(O'Doherty et al 2001, O'Doherty et al 2007, Suzuki et al 2012, 中原裕之 & 鈴木真介 2013)．ほかにニューラル・ディコーディングも有力な手法であるが(Kamitani & Tong 2005, Nishimoto et al 2011)，ここではモデル化解析について概説しておく．モデル化解析では，脳の情報処理モデルを行動と脳活動の両方の側面から検証する．モデルはその行動データを十分に説明でき，さらに，その情報処理は脳活動に存在するはずである．この考え方に基づき，モデル化解析では，まずは，複数の異なる仮説について，それぞれの脳計算モデルがどれだけ行動データを説明できるかを，たとえばモデル選択などを用いて調べる．この行動データへのフィッティングの際に，モデルの中にある自由パラメータを同時に推定する．この推定は重要な役割を果たしていて，その情報処理の際，脳の情報処理の内在変数が各試行でどのような値をとるのかも推定できる．この内在変数は，行動データから直接観測することは不可能であり，モデルがあることで初めて推定可能となる脳内の情報処理である．その後，この推定された内的変数を用いて，脳活動データを調べることで、脳のどこでその情報処理が行われているかを検証する．このように，行動と脳情報処理の両側面をつなげることで脳計算を明らかにしていくのがモデル化解析である.

2.4 深層学習と脳科学

以上，脳科学における脳計算の理解の発展について述べてきたが，ここでは，その発展の影響を受けた深層学習による物体認識の例を挙げ，深層学習と脳科学の関係についてごく簡単に触れておく．深層学習による物体認識モデルには随所に脳科学の知見が織り込まれている(Hassabis et al 2017)。その発展の傾向が 3 つの水準にわたって見られる．それを以下で簡単に見ておこう．まず，言い古されたことではあるが，深層学習（deep learning）はその命名がなされる以前，脳の働きにヒントを得たニューラルネットワークと勾配降下法の研究を端緒として開発され始め(Amari 1967, Rosenblatt 1958, Rumelhart et al 1986)，物体認識への応用はネオコグニトロンがその原型である(Fukushima & Miyake 1982)．深層学習にすでに用いられているさまざまなノウハウにも，脳科学と対応する部分が多い．脳科学用語で言うと，シナプス可塑性はもとより，先に挙げたヒューベル＆ウィーセルの研究(Hubel & Wiesel 1959, Hubel & Wiesel 1962)を起源とする，受容野，コントラスト正規化，正規化線形関数，単純細胞や複雑細胞の反応特性などがそれである．これらは前述の 3 つの水準で言うと，主にハードウェアとアルゴリズム（問題(計算論)としてはスコープが狭い局所的なアルゴリズム）に該当する．さらに，深層学習による物体認識は，たとえば強化学習を組み合わせることで，ビデオゲームや囲碁などでも驚異的なパフォーマンスを示した(Mnih et al 2015, Silver et al 2016)．じつは，そこでも脳の知見が巧みに利用されていて，たとえば「注意」(Xu et al 2015)や「エピソード記憶」がそうである（他の例については(Hassabis et al 2017)参照）．LSTM を用いた深層学習に関する議論も「作業記憶」と関連が深い．

面白いのは，深層学習における脳科学の利用は，上にあげたハードウェア／アルゴリズムから、徐々に問題としてスコープが広めに、たとえば上に挙げたような注意・エピソード記憶・作業記憶などの問題（計算論）とアルゴリズムの両方にまたがる展開に徐々に広がってきていることである．この傾向は今後さらに深まっていくだろう．また，興味深いのは，深層学習の情報処理を理解するために，脳科学の手法を取り入れようという提唱で(Ritter et al 2017), これはとりもなおさず深層学習という人工脳の脳計算論研究の提唱であると言えよう．なお，深層学習を脳機能の理解に直接適用しようとする研究も始まっている(Hong et al 2016, Yamins & DiCarlo 2016).

3. 脳の強化学習: neural reinforcement learning

脳の強化学習理論は，脳科学における脳計算理解の発展例として顕著である(Rangel et al 2008, Rushworth et al 2011)．これはまた，社会知性を実現する脳計算の土台としても重要である．さらに，近年の AI 研究では深層強化学習の研究の発展が目覚ましい．これらに触れながら，ここで簡単に脳の強化学習理論を振り返ることで次節の準備としたい．なお，強化学習の数理としての基本的枠組は，ここでは既知の事柄として扱う。ま

た，脳機能の記述も具体的な記述は極力省いた．より詳しく知りたい方は、引用記事を参照されたい．

そもそも脳の強化学習には，行動のもっとも土台となる要素が多く含まれる．報酬は人間や動物がそれを求めて行動を起こす機序になる．報酬はもっとも原初的な行動の動機である．そのことが，脳強化学習が動機・情動・感情などの脳機能の解明に深く関わる所以である．また，報酬を得るための行動の選択には意思決定が介在し，報酬を得るには適応的に行動を学習することが必要である．つまり，脳の強化学習では，意思決定と学習という人間／動物の基本的な脳機能にアプローチしている．その行動では，予測（報酬予測）が本質的な土台となっている．そのため，脳の強化学習理論は，注意や作業記憶などの高次認知機能を解き明かす基盤ともなる．

脳強化学習がこの約 20 年で急速な発展を遂げた背景には，ドーパミン神経細胞の報酬予測誤差仮説の発展がある．これは，ドーパミン神経細胞は時間差強化学習（temporal difference reinforcement learning）の学習信号に対応する，とする仮説である(Schultz et al 1997)（図 1）．この仮説が脳強化学習の理解の土台になり，大脳皮質―大脳基底核回路が関わる意思決定と学習の脳機能理解が進んだ（本特集の坂上論文の記事 参照）(中原裕之 2005, 中原裕之 2009)．脳強化学習の分野は，3 つの水準の研究が並行して進み，互いに刺激を与え合う成功例として傑出している．それは，上述した脳回路（ハードウェア）の詳細な研究と大脳皮質―大脳基底核回路における情報処理（アルゴリズム）の研究において，問題（計算論）を，たとえば古くは条件づけ（conditioning）や採餌行動（foraging）など，また近年では神経経済学（neuroecnomics）や計算精神医学（computational psychiatry）に関連して議論される意思決定と学習行動について，その脳計算と対比して議論することが可能になりつつあるからである(Glimcher & Fehr 2014, Montague et al 2012).

近年，脳強化学習の分野では，その土台となるモデルフリー強化学習と，それに対して，外界環境の内部モデルを用いて行うモデルベースド強化学習，この 2 つの強化学習を実現する脳機能の分化に関する研究が大いに進みつつある(Daw & O’Doherty 2014). そして，これら 2 つの強化学習の中間的な形態も提唱されつつある(Momennejad et al 2016)．また，私たちの研究でも，報酬予測誤差仮説を発展させたドーパミン報酬構造学習仮説を提唱している（図 2）(Nakahara 2014, Nakahara & Hikosaka 2012, Nakahara et al 2004, 中原裕之 2013)．これらの研究は計算脳科学として強化学習と表現学習の融合を議論しており，その意味で，AI における深層強化学習の議論と関連する点が多い．

脳強化学習の研究は今後の AI 研究，特に深層強化学習の研究と交流を深めながら進展していくと考えられる．まず興味深いのは，脳強化学習と AI 強化学習はその起源から交流がある．AI の強化学習の研究では忘れられがちなようだが，強化学習という人工知能／機械学習の分野の確立に中心的役割を果たしたサットン＆バルトの研究(Sutton & Barto 1998)は，動物心理学による条件づけ行動のモデルを構築する試み(Sutton

& Barto 1981)に起源をもつ．そして逆に，AI の研究で発展した時間差強化学習(Barto et al 1983)が，脳強化学習の理解の土台となった報酬予測誤差仮説に結びついたのである．最近の研究に目を移すと，深層強化学習の研究では，最近の成功をベースに(Mnih et al 2015, Silver et al 2016)，さらなる発展として物体認識を外界モデルの利用に組み合わせることが盛んに試みられつつある(LeCun et al 2015, Neverova et al 2017)．これは，脳強化学習で議論されているモデルベースドの脳機能と対応しており，今後の研究交流が注目される．また，深層学習による物体認識に脳視覚系のアルゴリズムまたはハードウェアの知見がうまく取り込まれているのに対して，深層強化学習では，まだ脳強化学習で議論されているようなアルゴリズム・ハードウェアの知見はそれほど取り込まれていない．たとえば，Alpha GO の学習で用いられた段階的学習（教師あり学習・ポリシー学習・バリュエーション学習の 3 段階）と、大脳皮質－大脳基底核回路の並行回路での逐次系列の段階的学習は関連が深いと思われる(Hikosaka et al 1999, Nakahara et al 2001)．今後の交流が待たれる．

4. 社会知性を実現する脳計算

社会知性は人間の知能の中でどのような位置を占めるのか．「人間」＝「ヒト」であることの多くは社会性・社会行動に現れる．私たちの日常生活を思い起こしてみよう．あるいは，個人における重要なライフイベントが何であるかを考えてみるとよい．日常生活の大半は何らかの社会行動である．重要なライフイベント，たとえば結婚，家族との別れ，それらの多くが社会的である．深夜にベッドに横たわって一人で寝ようとしているときでさえ，社会知性を働かせることが多い．寝る前に小説を読むこともあるだろうし，音楽に耳を傾け，歌詞を聞いているかもしれず，そこで歌われていることが恋愛であったりする．一日に起きたことを思い浮かべるときも，たとえば，学校あるいは職場での出来事，パートナーとのこと，それらはすべて社会知性に関わる事柄である．人間を本当に理解したければ，その社会知性と行動がいかに生み出されるかを理解しなければならない．人間の脳機能を理解したと言うためには，その社会知性を実現する脳機能を理解しなければならないのである．社会知性を司る脳機能は，人間を人間としている本質的な脳機能である．これが脳機能研究の 1 つの出発点となる．

ただし，1 つだけ誤解のないようにしておくべきことがある．それは，人間の知能はさまざまな機能が重層的に調和して働くことで実現されているということである．たとえば，社会知性を実現する脳機能は人間が人間としての本質的な脳機能であるが，一方で，その脳機能が単独で日常の社会行動を実現できるわけではない．社会知性を実現するには，たとえば視覚などの知覚，そして運動の機能が前提となる．相手を識別する必要があるし，相手に働きかけることができる必要がある．知覚機能・運動機能だけでなく，学習や意思決定などの認知機能，動機づけのもとになる情動機能，さらには感情や

言語などの諸機能が重層的に働くことで，日常生活の社会知性は実現されている．社会知性の実現は人間の本質的な脳機能ではあるが，同時にそれだけで人間の社会知性が実現されるわけではない．そのことを踏まえたうえで，以下に社会知性を実現する脳機能について考えていく．

4.1 社会脳科学の著しい発展

社会知性の脳科学は，この 15 年ほどで著しく発展してきた．これには，上述したヒト fMRI 実験が重要な役割を果たした．この発展には大きく分けて 2 つの流れがある．一つは，社会心理学などで重要とされる社会的認知などの心理学的概念を脳科学と結びつけ，ヒト fMRI 実験を用いて行動と脳活動を対比させるタイプの研究である．もう一つは，行動経済学などから経済ゲームの手法を取り入れることで，社会的インタラクションの重要な特徴づけ（「社会的選好：social preference」と呼ばれる）を行い，行動と脳活動を対比させる研究である．これらの研究により，社会知性の重要な要素や諸々の心理的概念と脳活動・脳部位との対応が明らかになってきた(Fehr & Krajbich 2014, Rilling & Sanfey 2011)．また，これに呼応するかのように，サルやネズミを用いた社会知性の脳研究も発展してきている(Chang et al 2013, Tremblay et al 2017, Yoshida et al 2012).

それでは，重要と考えられているいくつかのトピックを挙げてみよう．たとえば，中心的な心理的概念として「心の理論」「共感」「利他性」などが挙げられる．他者の考え・気持ち・心を推断する能力を問うのが「心の理論」（Theory of Mind）である．この研究の大きな源流とされるのは，1 つはプレマックの研究(Premack & Woodruff 1978)である．そこでは，チンパンジーが行動するときに，他のチンパンジーの“心の中にある”もの，たとえば，なぜそれをするのか，その意図は何かなどを想定して行動しているかどうかが問われた．もう一つの大きな源流は，自閉症に関わる研究として発展したバロン―コーヘンの研究(Baron-Cohen et al 1985)である．彼はのちに，子供が「心の理論」を発達のなかで形成するメカニズムについて論じている．「共感」（empathy）も他者を推断する能力に関わる．「共感」とは，他者の気持ち・感情をより直接的あるいは自動的に感じとる能力，それがあたかも自らのことのように推断できる能力のことを指す．なお，共感について，たとえば感情的共感（emotional empathy）と認知的共感（cognitive empathy）の区別がされることもある．この区別を用いるときには，前者が他者の感情を察すること，後者はより認知的側面で他者の感情や考えを感じとることが強調される．「利他性」（altruism）とは，自らの利益のみを考えるのではなく，あるいは自らの損失を省みずに，他者の利益を図るように行動する性向を指す．この性向は，社会行動が円滑に行われるもとになっていると考えられている．自己の利益のみを追求せず他者についての考慮を行動選択に反映する性向は，より細かく見るとさまざまなタイプに分けられる．そ

れらは社会的選好（social preference）の総称でくくられ，それには自らと他者の利益の分配が等しくなることをよしとする「公平性」（fairness）や，お互いへの何らかの見返りを考慮する「互恵性」（reciprocity）などがある．

これらの概念に共通しているのが自己と他者の区別である．自己表象と他者表象が脳内で形成されるにあたっては，有名ないわゆる「ミラーニューロン」，つまり，自己の行動と他者の行動観察を同じように表象する脳機能も社会知性に関わってくる．さらには，「主体感」そして自己と他者の境界も，社会知性の脳研究の課題となる．また，心の理論および利他性と関連が深い課題として，社会が成立していくには，相手の将来のふるまいを信じることができる，つまり「信頼」の脳機能も重要である．社会的同調（social conformity）や社会的規範（social norm），また社会的正義いわゆるモラルが，脳でいかに表象されるかなども重要な研究課題である．

これらの社会知性の概念に関わる脳活動は多岐にわたる．本稿では紙面の都合上，詳細を説明するゆとりがないため，いわゆる社会脳（social brain）と呼ばれる脳部位を中心にざっと列挙しておく．大脳皮質の脳部位としては，内側前頭葉（medial prefrontal cortex），背外側前頭葉（dorsolateral prefrontal cortex），側頭頭頂移行部／後側上側頭溝（temporoparietal junction/posterior superior temporal sulcus），島皮質（insular cotex）など，そして，皮質下の脳部位としては，たとえば扁桃体（amygdala）や大脳基底核回路の線条体（striatum），側坐核（nucleus accumbens）を挙げることができる．大変粗い言い方になるが，上に挙げた社会知性の諸処の心理的概念を調べるヒト fMRI 実験で，これらの部位の脳活動が報告されている．

4.2 社会知性を実現する脳計算：自己システムと他者システムの視点から

このように，社会脳科学の発展により，社会知性の心理的概念が脳活動に対応づけられたことで，社会知性は脳科学的概念になりつつある．これに呼応するように，脳計算理解の展開も勢いよく進みつつある．一方，まだまだ未開拓な領域も多く，社会知性の脳計算の解明は非常に魅力的な研究分野となっている(Behrens et al 2009, Ruff & Fehr 2014)．ここで重要な役割を果たしているのが先に紹介したモデル化解析である．これを用いることで，先に挙げた脳部位で，どのような社会知性の情報処理が行われているかが明らかになりつつある．また社会知性一般，そして，特に社会的意思決定と学習に関して，先に挙げた脳強化学習のフレームワークが，社会知性の脳計算モデルの堅固な土台となっている．

以下に，一例として，他者のシミュレーション学習についての筆者らの研究(Suzuki et al 2012)を，以前のテクニカルレポート(中原裕之 2014a) に沿って紹介する．当研究の出発点を 2 点，まずは確認しよう．一つは，脳の情報処理の中に他者が存在するという視点である．そもそも社会知性とは何だろうか．わざわざ「社会」がつくことで何が大

きく変わるのか．それは，そこに「他者」が存在することである．物理的にその場に他者がいるという場面はもとより，仮に一人でいるとしても，そこで他者のことが念頭にあれば，そこには社会知性の働く場が生まれる．すなわち，他者に関する情報処理を行うことが，広い意味で脳による社会知性の実現ということになる．もう 1 つは，脳計算の基本要素の徹底解明という姿勢である. 対象とする情報処理が複雑になればなるほど，基本要素をまず明らかにすることが大切になる. 社会知性のように多様かつ複雑な行動を対象に脳計算を解明するときには，この基本姿勢が必須である．

筆者は，この基本要素の解明の手がかりとして，脳における「自己システム＋他者システム」による社会的な意思決定と学習を考えている（図 3）．脳の中にある他者（のモデル）を他者システムと捉えるのである．脳はもともと（非社会的状況でも）“自己システム”を持っている．この自己システムと他者システムが協働して，他者の心や行動を推測し，それを生かす社会知性の脳計算を可能にすると考える．この「自己システム＋他者システム」の観点から社会知性の脳計算の研究を加速するには，まず自己システムの脳計算が強固に理解できていることが望ましい．そこに，脳強化学習の知見を使うのである. 脳強化学習では（非社会的状況での）意思決定と学習の脳計算の基本要素，たとえば，各選択肢の価値，さらにそれらの価値の比較に基づく行動選択確率，また，学習に使われる報酬予測誤差などが明確化されている. これらの基本要素と脳活動の対応も進んでいる．これらを自己システムの脳計算の土台として、社会的意思決定が「自己システム＋他者システム」から実現されるとし、その脳計算の解明を進めるのである．

4.3 シミュレーション学習

先に述べた「心の理論」に関わる諸説の中で「心のシミュレーション」説は有力な考え方である．これは，他者の心を理解あるいは予測するには，他者の心的過程（つまり他者の脳で起きている脳計算）を，同様の脳計算をする自己の脳部位を使ってシミュレーションすればよいというものである．一方，シミュレーション説とは違う考え方もある．それは，たとえば「セオリー・セオリー」と呼ばれる行動パターン説に代表されるように，わざわざシミュレーションを用いなくても，他者が対応している状況（他者の知覚の入力）と他者の行動を観察し，その他者の入出力を直接学んでしまえば，シミュレーションをしなくても他者の行動を予測できるようになるとするものである．

筆者らは強化学習を題材として取り上げ，ヒト fMRI 実験にモデル化解析を適用することで，この「心のシミュレーション」説の検証に真っ向から挑んだ．その結果，シミュレーションは存在すること，しかしシミュレーション理論が正しいわけではないことを示すことができた. 他者の心を理解あるいは予測するにはシミュレーションだけでは不十分で，入出力の直接学習も利用される．ただし，この直接学習の土台にはシミュレーションがあることを示したのである（図 4）．つまり，シミュレーションと入出力の

直接学習のハイブリッドな学習を通じて, 人間は他者の報酬予測と行動選択の学習を行っていることが明らかになった.

筆者らはこの研究を進めるときの方略として, これらの脳計算の基本構成要素を重点的に抽出するための実験パラダイムを開発し, 脳数理モデルを構築し, モデル化解析によりシミュレーション学習（他者の報酬にもとづく学習のシミュレーション学習）を調べた. この実験のコントロール課題は, 強化学習での標準的な価値意思決定課題（自己の報酬予測による選択）であり, これを用いて非社会的状況における自己システムの脳計算（報酬予測誤差に基づく学習信号）について調べた. 一方, メイン課題は他者の価値意思決定予測課題であった（他者がコントロール課題をしていて, 被験者はその選択を当てる課題）. この 2 つの実験課題の学習信号に対応する脳計算モデルと脳活動を比べることで, 2 つのシミュレーション学習信号が発見された. それが「他者報酬予測誤差」（他者が実際に得た報酬量と, シミュレーションによる他者の予測報酬量の差）と「他者行動予測誤差」(他者の実際の行動とシミュレーションから予測された行動の差)である. この他者報酬予測誤差は, 全脳の中でほぼ前頭葉腹内側部のみで見いだされた. この脳部位ではもともとのコントロール課題での自己報酬の報酬予測誤差に関連する活動も見られた. これは他者の強化学習をシミュレーションで実現するのに, この前頭葉腹内側部が中心的役割を果たしていることを示している. 一方で, 他者行動予測誤差の場合は, 前頭葉背外側部や背内側部, 側頭頭頂接合部, 後部上側頭溝で脳活動が見られた. 同じ社会脳に関わる部位のうち, これらはむしろシミュレーションを利用したうえで, 他者の具体的な行動に対する学習信号を形成していることが示されたのである. 端的に述べれば, 実際の行動に一致する脳活動の実体（脳計算）を示すとともに, 他者の心あるいはその価値意思決定のシミュレーション学習の新たなモデルを提示できたと考えている. 興味を持たれた読者諸氏にはぜひ原著論文にあたられたい(Suzuki et al 2012, 中原裕之 & 鈴木真介 2013).

5. まとめ

AI と脳の研究が相互に刺激を与え合うフロンティアは豊かに広がっている. 社会的インタラクションを実現する AI を作る研究は, 人間の社会知性を表現する行動やその心理・認知, 脳回路・脳活動の研究から刺激を受けることだろう. そして, もっとも重要なのは, AI 研究と, これらの心理・認知・行動を脳の情報処理と対応づける脳計算の研究が, 互いに学び合うことだろう. 上に見たように, AI のさまざまな学習理論はこれまで脳計算研究に大きく貢献してきており, 未来を見据えても, 現在発展中の深層強化学習と脳強化学習の協奏が大きく期待できる.

社会知性の脳計算研究はこれから大いに発展するにちがいない.「自己システム＋他者システム」のアプローチは, 多くの課題解明の土台になる. たとえば, 上の例に挙げ

たシミュレーション学習の研究では，じつは「自己システム＋他者システム」の限局した場合を扱っている．たとえば，そのメイン課題では被験者は他者システムの出力を直接回答している．しかし，実際の日常では他者の行動予測を利用して，自己の意思決定を調整することがある．こういった現実をさらに考慮に入れた研究を私の研究室ですでに進めつつある．また，私たち人間は自分の報酬だけではなく，他の人を気にかけて行動を変える．これは他者システムを用いた他者への配慮が自己システムの意思決定の脳計算を変容するプロセスと見なすことができる．こちらの研究も進展中である．こういった人間の反応は，社会的インタラクションの基本要素である．社会知性を実現する脳計算の基本要素を徹底的に解明する研究と，社会的インタラクションを実現するためのAI 研究の交流は実り多いものとなるだろう．これらの研究例は比較的「心の理論」と関連が深く，心理的，認知的，脳科学的な概念を，脳計算へと展開させる．AI 研究が社会知性の脳研究に目を向けるとき，脳計算に注目することは有意義であると思われる．これはすなわち計算論（問題）の研究との交流である．たとえば，社会知性の他の重要な概念として共感があるが，そもそも共感には感情が必要であり，共感・感情の脳科学における計算論の展開はまだこれからの段階にあるので，AI 研究と脳計算論の交流は意義深いものとなるはずである．筆者もいずれ感情と共感の脳計算論を展開したいと考えている．

今後さらに，脳と AI のフロンティアに両方の側から参画する研究者が増えることが期待される．AI 研究の側から，そして脳研究の側からも，その交流のツボを押さえた有意義な刺激を与え合う研究を大いに促進したいものである．脳と AI の境界領域の研究が発展し，脳と AI の両領域の研究で成果が上がることを期待したい．

Figures

図 1

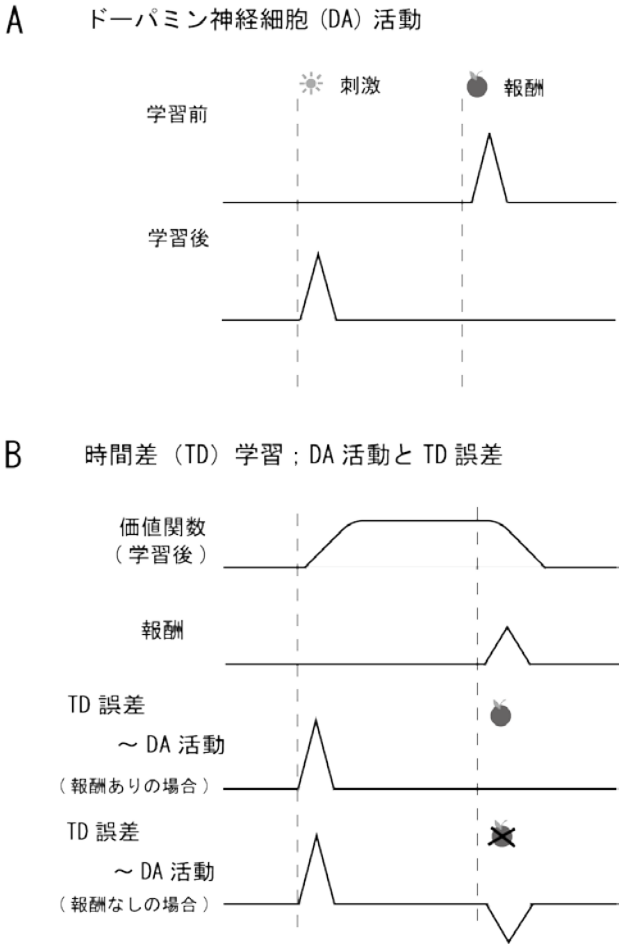

図 1：刺激―報酬課題（古典的条件付課題）における，ドーパミン神経細胞活動（DA 活動）（A）と，それに対応する時間差学習（TD 学習）および誤差（TD 誤差）（B）：B の下の 2 段は，ドーパミン神経細胞活動と TD 誤差が一致することを，報酬がもらえる場合ともらえない場合の両方について示している．(中原裕之 2014b)から転載.

図 2

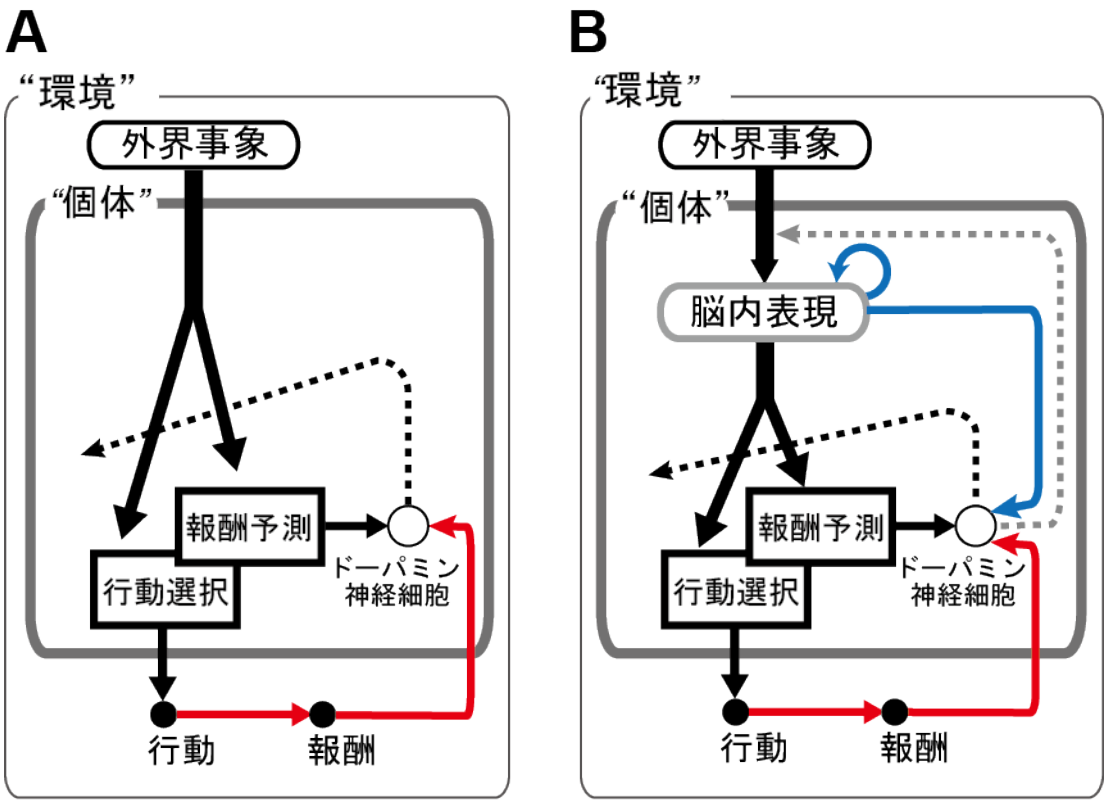

図 2：報酬予測誤差仮説（A）と報酬構造学習仮説（B）

（A） ドーパミン神経細胞活動は報酬予測誤差を表す（黒色点線矢印）．誤差は実際に得た報酬（ドーパミン神経細胞に入っていく赤色矢印）と予測された報酬との差である．ここで用いられている報酬予測は，基本的にはその各時点での外界事象（ドーパミン神経細胞に入っていく黒色矢印）で表現される予測に限られる．ドーパミン神経細胞活動で表される報酬予測誤差は，学習信号として報酬予測と行動選択の学習に寄与する（黒色点線矢印：本文中の学習で変化する「重み」が交差している黒色実線矢印に対応）．報酬予測誤差仮説では，ドーパミンの活動は特異的な報酬予測誤差シグナルを符号化すると考えられている．

（B） ドーパミン神経細胞活動は報酬構造を学習するための信号を表す（黒色点線矢印および灰色点線矢印）．ドーパミン神経細胞は，報酬予測のみならず，学習された報酬構造に関する入力（青色矢印）を受ける．さらに，ここでの報酬予測は，各時点での外界事象と学習された報酬構造の両者を反映した脳内入力から生成される．この予測は，原理的に（A）で用いられた報酬予測より優れている．この予測を利用した報酬予測誤差信号は，より優れた学習信号として報酬構造学習の信号の一部となる（黒色点線矢印）．報酬構造成分をより多く含むドーパミン神経細胞活動は，報酬構造を反映した内的表現の学習にも用いられる（灰色点線矢印）．文献(Nakahara & Hikosaka 2012)の図を元に改変．(中原裕之 2013)から転載．

図 3

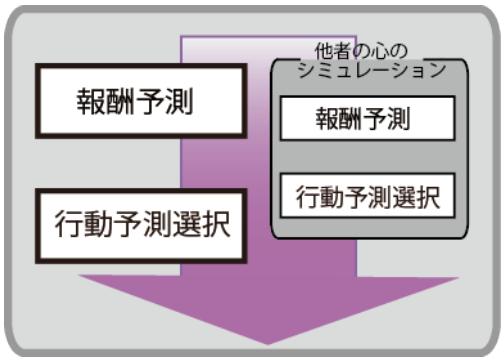

図 3：自己システム＋他者システム
(中原裕之 2014a)から転載.

図 4

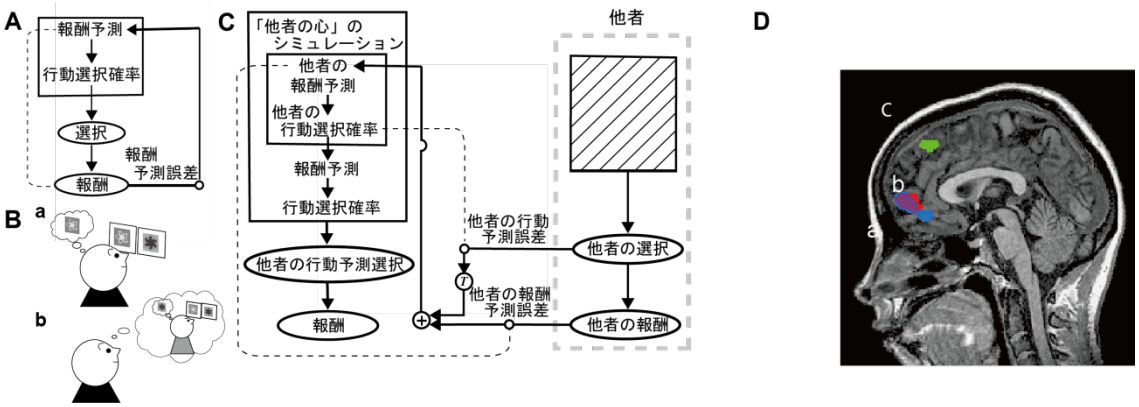

図 4：他者の心のシミュレーション学習

A. 価値意思決定の脳内過程：四角の枠は個体の内部プロセスを表す．B.（a）価値意思決定の実験課題．（b）他者の価値意思決定予測課題．C. B の（b）の課題で他者の意思決定をシミュレーション学習する脳内過程：2 つの学習信号である“他者の報酬予測誤差”（他者が実際に得た報酬量とシミュレーションによる他者の予測報酬量の差）と“他者の行動予測誤差”（他者の実際の行動と予測された行動の差）の存在と脳機能局在を示した．D. 主な脳活動：価値意思決定課題を遂行中の脳活動部位（a）と他者予測課題を遂行中の脳活動部位（b，c）．a：被験者自身の“報酬予測誤差”（前頭葉腹内側部：画像の青色部分）．b：“他者報酬予測誤差”（前頭葉腹内側部：画像の赤色部分）．c：“他者行動予測誤差”（前頭葉背内側部：画像の緑色部分）．（中原裕之 2014a）から転載.

参考文献

Amari S. 1967. A theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers*: 299-307

Baron-Cohen S, Leslie AM, Frith U. 1985. Does the autistic child have a "theory of mind"? *Cognition* 21: 37-46

Barto AG, Sutton RS, Anderson CW. 1983. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*: 834-46

Behrens TE, Hunt LT, Rushworth MF. 2009. The Computation of Social Behavior. *Science* 324: 1160-4

Chang SW, Gariepy JF, Platt ML. 2013. Neuronal reference frames for social decisions in primate frontal cortex. *Nat. Neurosci.* 16: 243-50

Daw ND, O'Doherty JP. 2014. Multiple Systems for Value Learning In *Neuroeconomics: Decision Making and the Brain*, ed. PW Glimcher, E Fehr, pp. 393-410: Academic Press

Fehr E, Krajbich I. 2014. Social Preferences and the Brain In *Neuroeconomics*, ed. Elsevier, pp. 193-218

Fukushima K, Miyake S. 1982. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition In *Competition and cooperation in neural nets*, pp. 267-85: Springer

Glimcher PW, Fehr E. 2014. *Neuroeconomics: Decision making and the brain*. Academic Press.

Hassabis D, Kumaran D, Summerfield C, Botvinick M. 2017. Neuroscience-Inspired Artificial Intelligence. *Neuron* 95: 245-58

Hikosaka O, Nakahara H, Rand MK, Sakai K, Lu X, et al. 1999. Parallel neural networks for learning sequential procedures. *Trends Neurosci.* 22: 464-71

Hong H, Yamins DL, Majaj NJ, DiCarlo JJ. 2016. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* 19: 613-22

Hubel DH, Wiesel TN. 1959. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology* 148: 574-91

Hubel DH, Wiesel TN. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology* 160: 106-54

Kamitani Y, Tong F. 2005. Decoding the visual and subjective contents of the human

brain. *Nature neuroscience* 8: 679-85

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521: 436-44

Marr D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Company. 397 pp.

Marr D, Poggio T. 1976. From understanding computation to understanding neural circuitry.

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518: 529-33

Momennejad I, Russek E, Gershman S. 2016. The successor representation in human reinforcement learning. *bioBxiv*: 1-27

Montague PR, Dolan RJ, Friston KJ, Dayan P. 2012. Computational psychiatry. *Trends in Cognitive Sciences* 16: 72-80

Nakahara H. 2014. Multiplexing signals in reinforcement learning with internal models and dopamine. *Curr. Opin. Neurobiol.* 25: 123-29

Nakahara H, Doya K, Hikosaka O. 2001. Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences - a computational approach. *J. Cognit. Neurosci.* 13: 626-47

Nakahara H, Hikosaka O. 2012. Learning to represent reward structure: A key to adapting to complex environments. *Neuroscience Research* 74: 177-83

Nakahara H, Itoh H, Kawagoe R, Takikawa Y, Hikosaka O. 2004. Dopamine Neurons Can Represent Context-Dependent Prediction Error. *Neuron* 41: 269-80

Neverova N, Luc P, Couprie C, Verbeek J, LeCun Y. 2017. Predicting Deeper into the Future of Semantic Segmentation. *arXiv preprint arXiv:1703.07684*

Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL. 2011. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology* 21: 1641-46

O'Doherty J, Kringelbach ML, Rolls ET, Hornak J, Andrews C. 2001. Abstract reward and punishment representations in the human orbitofrontal cortex. *Nat. Neurosci.* 4: 95-102

O'Doherty JP, Hampton A, Kim H. 2007. Model-Based fMRI and Its Application to Reward Learning and Decision Making. *Ann. N. Y. Acad. Sci.* 1104: 35-53

Ogawa S, Lee T-M, Kay AR, Tank DW. 1990. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences* 87: 9868-72

Premack D, Woodruff G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1: 515-26

Rangel A, Camerer C, Montague PR. 2008. A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience* 9: 545-56

Rilling JK, Sanfey AG. 2011. The neuroscience of social decision-making. *Annual Review of Psychology* 62: 23-48

Ritter S, Barrett GTD, Santoro A, Botvinick MM. 2017. Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study. *arXiv* (In Press)

Rosenblatt F. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review* 65: 386

Ruff CC, Fehr E. 2014. The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience* 15: 549-62

Rumelhart DE, McClelland JL, Forschungsgruppe P. 1986. Parallel Distributed Processing: Extrapolations in the Microstructure of Cognition, vol. 1, 2: Cambridge: MIT Press

Rushworth MF, Noonan MP, Boorman ED, Walton ME, Behrens TE. 2011. Frontal cortex and reward-guided learning and decision-making. *Neuron* 70: 1054-69

Schultz W, Dayan P, Montague PR. 1997. A neural substrate of prediction and reward. *Science* 275: 1593-9

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529: 484-89

Sutton RA, Barto AG. 1998. *Reinforcement Learning : An Introduction*. pp. 1. Cambridge, MA,: MIT Press. 1-551 pp.

Sutton RS, Barto AG. 1981. Toward a modern theory of adaptive networks: expectation and prediction. *Psychological review* 88: 135

Suzuki S, Harasawa N, Ueno K, Gardner JL, Ichinohe N, et al. 2012. Learning to Simulate Others' Decisions. *Neuron* 74: 1125-37

Tremblay S, Sharika KM, Platt ML. 2017. Social Decision-Making and the Brain: A Comparative Perspective. *Trends in Cognitive Sciences* 21: 265-76

Wang X-J. 2002. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* 36: 955-68

Wurtz RH. 1968. Visual cortex neurons: response to stimuli during rapid eye movements. *Science* 162: 1148-50

Wurtz RH, Goldberg ME. 1971. Superior colliculus cell responses related to eye movements in awake monkeys. *Science* 171: 82-84

Xu K, Ba J, Kiros R, Cho K, Courville A, et al. 2015. Show, attend and tell: Neural image caption generation with visual attention. *International Conference on Machine Learning*: 2048-57

Yamins DL, DiCarlo JJ. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19: 356-65

Yoshida K, Saito N, Iriki A, Isoda M. 2012. Social error monitoring in macaque frontal cortex. *Nat. Neurosci.*

デビット・マー. 1987. *ビジョン-視覚の計算理論と脳内表現*. 産業図書. 448 pp.

中原裕之. 2005. 大脳基底核の計算モデル In *脳の計算機構*, ed. 銅谷賢治, 五味裕章, 阪口豊, 川人光男, pp. 140-61: 朝倉書店

中原裕之. 2009. 意思決定とその学習理論 In *脳の計算論*, ed. 甘利俊一監, 深井朋樹, pp. 159-221: 東大出版会

中原裕之. 2013. 報酬構造学習―ドーパミン神経細胞をめぐる新仮説―. *BSI ITN Technical Report* 13

中原裕之. 2014a. ヒトの心を汲みとる社会知性――脳神経基盤と脳計算の解明―― *BSI ITN Technical Report* 14

中原裕之. 2014b. 脳の計算理論：強化学習と価値に基づく意思決定 *BSI ITN Technical Report* 14

中原裕之. 2017. 社会性の脳計算の解明、そして人工知能. *BSI ITN Technical Report* 17

中原裕之, 鈴木真介. 2013. 意思決定と脳理論：人間総合科学と計算論的精神医学への展開. *Brain&Nerve* 65: 973-82