

特集 「意識とメタ過程」

痛みを感じるロボットの意識・倫理と法制度

Consciousness, Ethics, and Legal System for AI and Robots Who Feel Pain

浅田 稔

Minoru Asada

大阪大学大学院工学研究科

Graduate School of Engineering, Osaka University.

asada@ams.eng.osaka-u.ac.jp, www.er.ams.eng.osaka-u.ac.jp

Keywords: pain, consciousness, ethics, legal system, AI, robots, constructive developmental science.

1. はじめに

近年の神経科学や生理学などの観測・計測技術の向上に加え、深層学習に代表されるAI技術の進展はめざましく、日進月歩どころか、秒進分歩の勢いで、さまざまな局面で利用されている[独立行政法人情報処理推進機構 17]。本特集号のタイトルである「意識とメタ過程」の中核にある人工意識も近々実現できても不思議ではなさそうな勢いである。しかしながら、深層学習の基本的な限界から、容易でないと察せられる([浅田 17]などを参照)。ポイントは、知覚系重視で運動系との密なリンクが希薄で本質的な身体性に結び付いていないからである。このことは、無意識も含めた意識研究にとっても重要と考えられる。それゆえ、論理的な構造だけでなく、情動的な側面も合わせて検討すべきであろう。著者の解説[浅田 16]では、人工情動設計論の要点の(4)として、痛覚神経回路の必要性を指摘している。著者はこれが、人工意識の設計においても重要と考えている。

痛みは、触覚や体性感覚などの神経系とは異なる経路をもっている([Purves 12]のChapter 10)。ただし、痛いときに擦ること(「痛い痛い飛んでいけ!」)により触覚系が励起され、痛覚系をブロックするといわれており、当然、独立ではない。この痛みを共有しているという感覚が共感のもと(情動感染[Chen 09])となつていてと考えられる。実際、著者の人工共感のサーベイ[Asada 15]で紹介した神経科学、認知科学、心理学の多くの文献が痛みを題材にしていることもそのことを示している。よって、ロボットの神経系に触覚系とは別の痛み回路を埋め込むことは、その良し悪しを別として、生物進化の観点からも、過度な人工的なバイアスにはならないと考えられる。痛み感覚の共有は、ミラーニューロンシステム[浅田 10]に代表される自己と他者の行動の同一性理解のみならず、同一の知覚励起に起因しており、自他認知を通じた他者の行動や知覚を無意識に模倣したり、感じたりする傾向の源と考えられる。このことがモ

ラルや倫理の発達につながるのでは、と著者は考えている。

本稿では、まず最初に意識や自由意思を含め、広義の心の課題について、思想的な背景の概略を復習する。次に、意識の認知神経科学的な側面からのアプローチを紹介し、意識の有り様の定式化を試みる。意識を始めとする心的機能がどのように発達してきたかを構成的手法をベースとしてアプローチする認知発達ロボティクス[Asada 09]を再考する。身体性と社会的相互作用がそのキーワードであるが、これが先に示した思想的背景に由来することを示す。

身体には、上記の痛覚回路が内在し、触覚とは異なる痛みを検知することが可能とする。また、社会的相互作用の前提として、自他認知過程を通じ、他者の運動や知覚を推定しようとする傾向があるとする。これらに基づき、これまでの認知発達ロボティクスの研究の一部をごく簡単に紹介し、それらが系列的に結び付くことで、共感の発達、そしてモラルの発生の可能性を議論する。さらに、「人工システムが道徳的行為者や受益者になり得るか」などの倫理的な側面、さらには、「人工システムが自身の過失に対して責任をとれるか」などの法制度の課題にも言及し、すでに現代社会に浸透している人工システムとの未来共生社会を議論する。最後に、2018年3月に開催した「AI・ロボットの倫理と法制度に関する国際ワークショップ」の報告を通じて、今後解決すべき課題とその道筋を探る。

2. 意識・人間や事物(技術)の課題の思想的背景の概略

心の課題を考察にするにあたって、すべての思想的背景を網羅的に探ることは不可能に近く、以降に紹介する事項に絞って、意識・人間や事物(技術)の課題の思想的背景の概略を示すことをお許しいただきたい。前半はTaniの書籍[Tani 16]で紹介されている流れを、後半は、稲谷の文献[稲谷 17]に基づきながら紹介していく。



図1 意識・人間や事物（技術）の課題の思想的背景の概略

心と身体、もしくは事物の関係に関して、心身二元論^{*1}を唱え、近代哲学の基礎を築いたのはデカルトであろう。その後、多くの批判にさらされ、いくつかのバリエーションも存在しているが、後に述べる法制度では、根強く生き残っているようである。

デカルトを超えて超越論的現象学へと進む「新デカルト主義」を主張し、現象論的考察を与えたのは、フッサールである（例えば、[フッサール 01] など）。主観と客観の狭間の間主観性の考え方を展開し、後世に多大な影響を与えた。自然界の解析は、個人の意識経験に基づくと言く。フッサールの現象論を拡張・進化させたのが、ハイデガー（『存在と時間』[ハイデッガー 94a, ハイデッガー 94b]）やメルロ＝ポンティ（『知覚の現象学』[メルロ＝ポンティ 67, メルロ＝ポンティ 74]）である。

ハイデガーは、主観と客観を分けずに実存を問い、『現存在』は、個々のエージェントの将来の可能性とその過去の可能性との間の動的相互作用によって生まれていると主張する。また、それぞれの個々が目的をもっていかに関与しているかという事前の理解のもとに個々が相互に存在し得るといふ、ある種の社会的相互作用の重要性を指摘している。

メルロ＝ポンティは、主観と客観に加えて身体性という次元が創発し、そこでは、同じ肉厚の身体が、触れたり見たりする主体と同時に触れられたり、見られたりする物体にも与えられ得るとし、主観と客観の二つの極の間の繰り返される交流の場を身体が与えると主張する。すなわち、客観的物理世界と主観的经验を結ぶメディアとしての身体的重要性を指摘している。これは、後に紹介する認知発達ロボティクスにおける「身体性」の基本概念的根幹である。

ジェームズ^{*2}は、近代心理学の創設者として著名であ

るが、心に関する本質的な哲学的思索を多く残している。西田[西田 79]は、仏教の瞑想に影響を受けたが、彼の思索はフッサールやジェームズの考え方に類似している。すなわち、主観と客観の現象論的実態は、これらの統合と分離を繰り返すと主張する。

このような現象論的意識の有り様と並行して、ニーチェ[ニーチェ 15]は、「神は死んだ」として、神の存在を否定し、人間の在り方としての実存主義を主張した。この実存主義の考え方は、ハイデガーや初期のメルロ＝ポンティに影響を与えたとされる。フーコー[フーコー 74]はニーチェやハイデガー、さらには構造主義の影響を受けつつ、「人は死んだ」と称して、近代諸学問における人間中心主義を批判した。あるべき人間像に基づく道徳的評価は、結局標準的な人間になることを求めること以上の意味をもち得ない。標準化された人間のみによって構成される、摩擦なき、しかし個人の自律も創造性もない「透明なコミュニケーション」という問題を指摘した[稲谷 17]。

ラトゥール[ラトゥール 07]は、主体と客体とを厳格に区分する近代的な思考法（人間存在の在り方を本質化する、ヒューマニズム）のせいで、我々は、我々が現実生活しているところの主客が入り混じったハイブリッドな世界を適切に取り扱うことができなくなっていると警告する。また、フェルベーク[フェルベーク 15]は、「技術は、我々の行為や世界経験を形成し、そうすることによって、我々の生活の仕方に能動的に関わっている」と主張する。

主体と客体とを厳格に区分する近代諸学問における人間中心主義へのこれらの批判は、本解説の後半で述べる法制度のところでも再度触れる。

3. 人工意識は実現可能か？

前章の意識の思想的背景に対し、具体的に人工意識を実現しようとするとき、まずは、人間の意識の機構的な構造の解明は設計指針のみならず、検証の意味でも重要

*1 実体二元論、物心二元論、霊肉二元論、古典的二元論などともいわれているようである（<https://ja.wikipedia.org/wiki/実体二元論>）。

*2 <https://ja.wikipedia.org/wiki/ウィリアム・ジェームズ>

である。ここでは、認知神経科学者のドゥアンヌの著書 [ドゥアンヌ 15] や彼らの論文 [Dehaene 17] を交え、人工意識をどのように捉えるかを議論する。

ドゥアンヌ [ドゥアンヌ 15] は、認知神経科学的に、意識と無意識を定義し、それらの関係から、主観的な現象を尊重することで、意識的知覚を実験によって操作できるということを発見し、さまざまな実験を通じて意識の解明を目指してきた。無意識を分類し、意識に最も近い前意識が数多く並列で作動しており、注意のボトルネックを通じて、ポップアップされる機会（ベイズ推定）をうかがっている。いったんポップアップされて意識レベルにのぼると、系列的処理の集中を受け、その他の無意識的知覚がマスクされる。この構造は Baars [Baars 05] が提案した大局的作業空間理論（Global Workspace theory, 以降 GW 理論と略記）と合致し、グローバルニューロナルネットワークとして提案されている。GW 理論は、AI 分野で著名な Newell らのブラックボードアーキテクチャに基盤がある。分散した一連の知識ソースから構成され、単一の構成要素だけでは解決できない問題を共同で解決することを示した。GW 理論は、知覚、感情、動機付け、学習、作業記憶、自発的制御、および脳における自己システムの意識の側面について明示的な予測を生成する。これは、神経ダーウィニズムや脳機能の力学理論などの生物学的理論と類似している。図 2 にその概念図を示す。意識内容は、ワーキングメモリ上のスポットライトに対応し、無意識の脳領域（意識経験を直接サポートしないと考えられる皮質領域、海馬、および基底核）と関連する。意識的認知そのものは、常に無意識の文脈によって形成され、エグゼクティブファンク

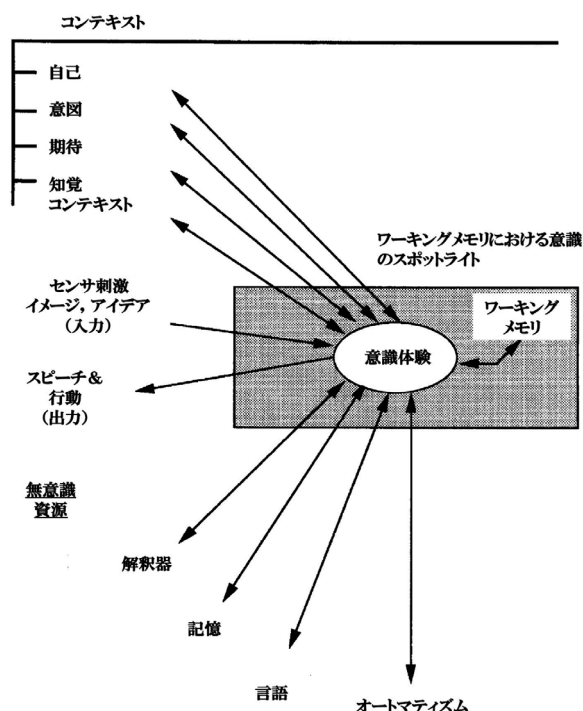


図2 GW理論の概念図（文献 [Baars 05] の Fig. 1 を改変）

ション（自己）は、そのようなコンテキストの一つのセットと考えられる。

ドゥアンヌの定義に従えば、原理的には乳幼児や霊長類、さらには、他の哺乳類や動物にも意識は生じる。ただし、ヒトの場合、言語がコミュニケーションシステムとしてよりも、表象装置として進化し、考え出す能力を与えたとして、表象レベルが他の動物に比べ抜き出ているという。

Dehaene ら [Dehaene 17] は、機械の意識を検討するうえで、明確に意識を所有していると思われるヒトの脳について再考し、機械が創発する意識の有り様を議論している。脳の中には、二つの情報処理計算が存在すると主張する。それらは、以下である。

- (1) 大局的情報伝達のための情報選択→それを計算と報告に柔軟に利用可能にすること（第一センスの意味で C1 意識と呼ぶ）
- (2) 上記の計算の自己モニタリング→確信かエラーかの主観的感覚への導入（第二センスの意味で C2 意識と呼ぶ）

現状の機械は、脳の中でいうところの無意識処理（C0 意識と呼称）に対応しているとみなす。C0 は、心ない自動的な応答により、精練された情報処理を行っている。顔や音声の認識、チェスのゲーム評価、構文解析などを無意識に、すなわち大域情報伝達や自己モニタリングなしに行っている。それらは、独立したモジュールであり、個々の深層学習ネットワークに対応する。

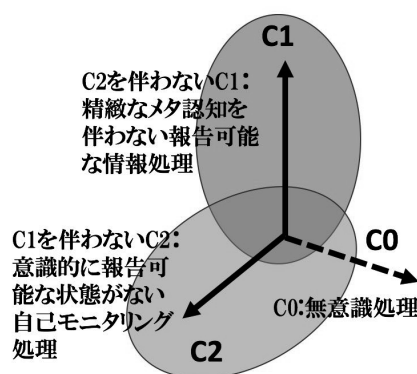


図3 直交する2種の意識（C1, C2）と無意識（C0）の関係

C1 は大局的に情報収集し、単一モジュールでは解決できなかった課題に対して、柔軟に対応可能であり、先にあげた GW 理論が対応する。C1 が外界の情報収集に専念していたことに対し、C2 では、自己の内部を反映する情報処理構造が特徴であり、それゆえ、彼らは、C1 と C2 が直交していると考えている。当然、我々の日常生活では、C0, C1, C2 すべてが混在するのであるが、C2 を伴わない C1 や C1 を伴わない C2 が存在することで、その直交性を担保している。

現状の多くのシステムが C0 状態であるとする、機械の意識を実現するうえでは、C1 や C2 レベルの実現

を検討しなければならない。C1 レベルはすでに紹介した GW 理論に則ったものが構築中であるが、C2 レベルではどうであろうか？ Dehaene らは、一つの有力なメカニズムとして、敵対的生成ネットワーク (Generative Adversarial Networks : GANs) をあげた。これは、生成ネットワーク (generator) と識別ネットワーク (discriminator) の二つのネットワークから構成され、前者の出力を後者がその正否を判定する。生成側は識別側を欺こうと学習し、識別側はより正確に識別しようと学習する。これは、自己モニタリングの機能に相当すると考えられたからである。

結論として、意識は情報計算の産物であり、C1 と C2 の能力をもつ機械であれば、何かを見、それに対する自信を表明し、他者に報告できることを知っている。そのため、モニタリング機構が故障すると幻覚状態に陥り、人間と同じような錯覚を経験するとしている。

これまで見てきたように、認知神経科学の側面からの意識の脳内現象の構造的な理解は進んできたと思われるが、C2 レベルのより深い構造、すなわち自己の内省のみならず、2 章で示した主観と客観が混在し、自他が混じり合った状況で垣間見られる心のダイナミクスの機構には触れられていない。以下では、種々の心的機能の発達の結果として、意識を現象として再現できないかと目論む構成的発達科学の側面を示す。明示的に意識の設計をうたってこなかったが、痛みの感覚を埋込みとして、自己身体の気付き、所有感覚からエージェンシーや自己認知、共感を通じたモラルや倫理などの社会的認知機能の出現への道筋を描いてみる。次に総体としての意識を構成する種々の機能がどのようにして創発するかの基本原則を解き明かす試み、さらに、先に触れた心のダイナミクスを創発する計算モデルについて簡単に紹介していく。

4. 構成的発達科学

2 章で触れた意識の現象論の根幹である身体性や社会的相互作用を基盤とし、ロボットや計算モデルによるシミュレーションを駆使して、人間の認知発達過程の構成的な理解と、その理解に基づき人間と共生するロボットの設計論を確立することを目的としているのが、認知発達ロボティクスである [Asada 09]。

認知発達ロボティクスの考え方をより進め、多様な分野を巻き込み、発達概念を機能分化の観点から明らかにしようとするのが、構成的発達科学*3 である。基本的な

* 3 研究プロジェクトとして、以下の二つがあげられる。
科学研究費補助金特別推進研究「神経ダイナミクスから社会的相互作用に至る過程の理解と構築による構成的発達科学」(平成 24 ~ 28 年度, 研究代表者: 浅田 稔, <http://www.er.ams.eng.osaka-u.ac.jp/asadalab/tokusui/>)
新学術領域研究 構成論的発達科学 (平成 24 ~ 28 年度, 研究代表者: 國吉康夫, <https://kaken.nii.ac.jp/ja/grant/KAKENHI-AREA-4401/>)

神経構造から始まり、身体性や社会的相互作用に基づき、学習手法を介して、機能分化が段階的に生じる過程を描いている (詳細は [浅田 16] などを参照)。

心的機能の基本課題として、意識に関連する自己の概念と共感を例に見てみよう。共感と同情はしばしば混同されて用いられるが、設計を試みることで、その曖昧さが軽減する。人工的に共感構造設計を考えるにあたり、霊長類の進化的研究から始める。霊長類学者の deWaal [Waal 08] は、情動感染から始まる共感の進化とものまねから始め模倣の進化の並行性を示し、その進化の方向が自他の識別の増強とも関連することを提案している。後者は、生態学的自己から対人的自己、さらに社会的自己に至る過程と重なり (図 4 参照)、個体発生、すなわち発達もこの経路を辿ると想定される。生態学的自己は、先に述べた自己意識の原点ともいえるものであり、対人的自己は C2 レベルの意識を実現するうえで基盤となり、社会的自己は C2 レベルの意識の完成形とみなせる。これらをまとめて、共感発達モデルとしての自他認知過程を表したのが図 5 である [Asada 15]。

先に述べたように、基本的な神経構造として痛みの感覚系を前提とする。さらに、その痛みの共有が、社会的な行動の創発につながると想定する。構成的発達科学の

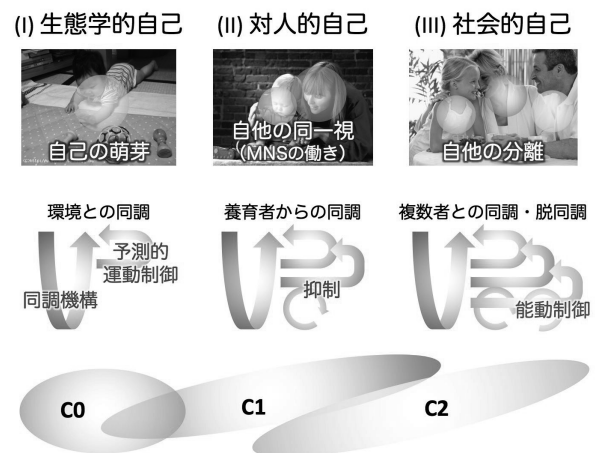


図 4 自他認知を含む認知発達過程 ([Asada 15] の Fig.3 を改変)

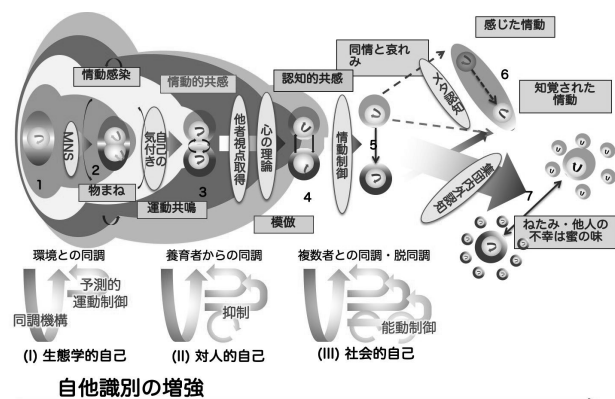


図 5 共感発達モデルとしての自他認知過程 ([Asada 15] の Fig.6 を改変)

究極のゴールは、系列的な発達段階すべてを通じた実現であるが、現状は、各段階において、試行錯誤している。基本的な神経構造のモデルによる計算機シミュレーション、fMRIやMEG(脳磁図)などのイメージング研究、心理・行動実験、それらを支えるロボットプラットフォームの研究が交叉し、新たな価値観の創出を目指している。以下では、関連したいくつかの試みを紹介する。

5. 基本的な神経構造からメタ認知へ

図5に示した共感発達過程で自己意識から始まりメタ認知に至る過程で、(無)意識の発達と関連する研究例のいくつかを、図5内の番号に合わせて紹介する。

身体と環境の相互作用により、さまざまな行動が創発する際に、感覚運動系と脳神経系がどのような関係にあるかは、構成的発達科学における基本課題である。図5の中の番号1より手前で、生態学的自己形成に必要な部分である。Parkら[Park 17]は、非線形振動子のニューロンから構成される脳神経系がヘビのようなロボットの筋骨格系を通じて、環境と相互作用した際に生じるネットワーク構造について、情報の移動エントロピーをもとに解析した。大まかに二つの運動パターンが現れ、安定行動パターンは高次元状態空間でのアトラクタに、不安定行動は、安定行動パターン間の遷移を表し、全体としてカオス遍歴の様相を呈し、環境との相互作用による身体と神経ネットワーク間のダイナミクスを表している。一つの憶測は、原初的な意識(不安定状態:例えば崖っぷちの歩行)・無意識(安定状態:例えば通常の歩行)に対応していないかという期待である。統合情報理論[Tononi 15]による統合情報量の計算は困難を極めるが、不安定状態のほうが安定状態よりも大きいと察せられる。また、意識・無意識の切替は注意のボトルネックを通過する必要がある[ドゥアンヌ 15]が、この研究では、自発的な神経パターンの構造変化により生じていると考えられる。今後、学習手法の導入によりサブネットワークの機能の意味や役割が明らかになることが期待される。

自己身体の発見や自己運動の所有感覚は、生態学的な自己の確立に必要である。サッカーロボットが自己、環境内の物体(ボール, ゴール), 他者(敵, 味方)を識別する際に、知覚と運動の相関が利用された[Asada 99]。運動情報を用いなくても、自己身体を発見する方法として、吉川らは、複数センサ情報の不変性に基づく身体発見手法を提案している[吉川 05]。このほか、道具使用による適応的身体表現[Hikita 08]、VIPニューロンの働きを利用した頭部身体周辺空間の表現の獲得[Fuke 09]などがあり、これらは、図5の番号1からMNSに向かうあたりに対応する。

新生児模倣は、新生児が示すミステリーの一つであるが、これを計算論的にアプローチしたのが、Fukeら

[Fuke 07]である。彼らは、未熟な視覚と触覚体験から、顔の部位の配置を再現した。また、その配置を他者観測時に射影し、顔の模倣の基盤とした。Nagaiら[Nagai 11]、Kawaiら[Kawai 12]は、視覚発達を伴う感覚運動学習がミラーニューロンシステムの発達を促進するという仮説のもとに、未熟な感覚の時期には自他が混同されるが、発達に伴い自他分離が起こり、これがミラーニューロンシステムの基盤となっていると主張している。図5では、番号1と2の間のMNSあたりに対応する。

共感の発達過程として、Watanabeら[Watanabe 07]は、ロボットが共感的な応答を学習するために、養育者の直感的親行動(赤ちゃんの情動状態を養育者が自動的にまねしたり、誇張したりすること)を利用して、赤ちゃんロボットの情動状態を分化させる過程をモデル化した(図5番号3あたり)。この研究では、痛み回路というよりも、快-不快の基本情動を仮定している。このほかにも、人間に対する解析を含めて、他者視点取得(Perspective Taking)、情動制御(Emotion Regulation)、メタ認知(Metacognition)に関連した研究を行ってきた[Kawakami 15, Takahashi 14, Yoshikawa 01]。これらの研究では、痛みの神経回路や他者の行動や知覚を模倣する傾向のバイアスなどを明示的に述べていないが、それらをベースとしてつなげ合わせることで、共感可能なロボットを実現への足掛かりになると期待される。

6. 機能分化の創発原理

前章の身体と神経の結合ダイナミクスの研究で、情報ネットワークのサブネットワーク構造の機能(個々のサブネットワークの働きおよびそのクラスタとしての働き)は明らかではない。その情報理論的な原理として、Yamaguti and Tsuda [Yamaguti 15]は、脳のヘテロなモジュール構造進化の数学的モデルを提唱している。これに基づき、創発インタラクションを進化から発達までの拘束条件のもとで、種々の時空間スケールで解き明かそうとするプロジェクト*4が始まっている。研究代表者の津田は、創発インタラクションの必須概念として、自律性、汎化性、共感や自己意識、動機付け、時間の概念などを列挙し、上記プロジェクトで、脳領域、個体、集団のそれぞれのレベルやスケールでの相互作用の創発から生じる機能や現象としての意識の在り方が模索されている。

*4 JST 戦略的創造研究推進事業(CREST)「脳領域/個体/集団間のインタラクション創発原理の解明と適用」(平成29~34年度, 研究代表者: 津田一郎) <http://www.er.ams.eng.osaka-u.ac.jp/kawai/crest/>

7. 意識創発の計算モデル

図2では、C0, C1 レベルの概念図を示しているが、これを明示的な固定プログラムで実現することは、C0 レベルでの自動応答的なモジュール（これとて、容易ではない）をどのようにポップアップさせるかの課題も含め、困難を極める。ドゥアンヌ [ドゥアンヌ 15] は、現行のコンピュータにない三つの機能として、柔軟なコミュニケーション、可塑性、自律性をあげている。それらがどこまで本質的な意味合いをもつかは、議論の余地があるものの、その実現を目指している意識創発の計算モデルとして、Tani [Tani 16] のアプローチを紹介する。2章で紹介した思想的背景をベースに心の有り様を現象論から解さばくし、哲学的課題である意識、意図、時間、自己などについて、リカレントニューラルネットワーク構造を駆使し、神経と身体結合ダイナミクスから生じるさまざまな現象と結び付けて論じており、興味深い。一例として、MTRNN (multiple-timescale recurrent neural network) と呼ばれる多重時間スケールのリカレントニューラルネットワーク構造を用いたロボットの運動学習を紹介する [Yamashita 08]。図6にそのアーキテクチャとネットワーク内の活動の概要を示す。

教示者が最初、ロボットに対して脚は固定だが、文字どおり手取り足取り動作を教え、ロボットはその行動系列を覚える。覚えた行動系列をさまざまな環境で再現する際、覚えたときと異なる環境での誤差を解消する構造として、MTRNN が利用され、時定数の小さい、よって速い周期の感覚運動レベルから、時定数の大きい、よって遅い周期の意思決定レベルの挙動を再現している。トップダウン（意図による予測）とボトムアップ（実際の経験）との差異を誤差として、下位から上位に逆戻りさせ、それを解消するように意図（予測）を変更し、その司令を下位に伝える構造となっている。特に、この誤差をもって意図が変更される（書き換えられる）際に、その意図は後付け的に意識されると谷は考えている。ボトムアップとトップダウンが行き交う上記のプロセスは、まさに現象学が思索してきた、主観と客観が境界を超えて交わり合う刹那を構成していると考えられる。さらにそれは、神経科学的には、意識の在り処と想定される前頭前野と感覚情報が立ち上がる後頭部の間のフィードフォワード・フィードバック処理過程を連想させる [Mashour 18]。無意識自動応答的な種々のモジュールの一つが、いかにして瞬時に意識レベルに到達するかのメカニズムも谷らの研究に期待したい。

また、上記の誤差解消法は、Friston [Friston 10] が唱えている予測誤差最小化原理 (Free energy minimization) と意を同じくするものであり、他者の行動を予測する社会的認知機能発達計算モデルに応用されている [長井 16a]。予測学習とは、身体や環境から

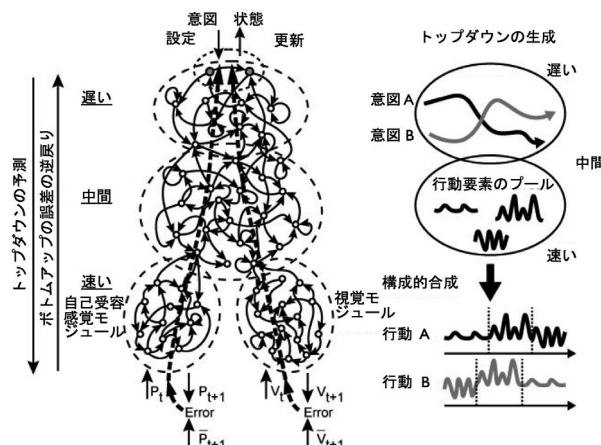


図6 MTRNNによる多層ネットワーク構造を利用した行動系列学習（書籍 [Tani 16] の Fig.9.3 改変）

のボトムアップな感覚信号と、脳が内部モデルをもとにトップダウンに予測する感覚信号の誤差を最小化するように内部モデルを更新したり、環境に働きかけたりすることである。

この原理に基づき、自己運動の生成経験が他者運動の認識に与える影響、複数感覚信号の空間的予測学習による他者の内部状態の推定とそれに基づく情動模倣、他者運動に起因する予測誤差の最小化による援助行動の創発などが検証されている ([長井 16b] を参照)。

8. ロボットと倫理

研究者としては、現状の人工システムの自律性や知能のレベルに不満足ではあるが、すでにさまざまな人工物が人間社会に導入され、便利な道具の域を超え、生活手法や価値を変えつつある現在、そして、将来さらに加速することを考えると、それらが引き起こす倫理の課題の重要性がますます大きくなっている。AIやロボットもしくは知的人工物（以下ではロボットと総称）を対象とした倫理課題は、以下の三つが考えられる [久木田 17]。

- (1) ロボットを製造する際の倫理：通常の工学倫理およびロボット特有の課題もある。
- (2) ロボットの守るべき倫理：ロボットが道徳的行為者になり得るとということが前提
- (3) ロボットに対する倫理：ロボットが道徳的配慮を受けるべき対象であることが前提

(1) は設計者や製造販売者が担う、すなわち人間サイドの倫理規定である。ロボットなどの自律的知能システムを研究や設計する際の倫理規定として、米国電気電子工学会 (IEEE) の Global Initiative on Ethics of Autonomous and Intelligent Systems を参照する^{*5}。ミッションは、自律的知能システムの設計と開発に関わ

*5 http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

るすべてのステークホルダーが教育、訓練を受け、倫理的配慮の優先順位付けを行い、これらの技術が人類のために進歩するようにすることである。一般的な原則として、五つの項目：人権、福祉、説明責任、透明性、誤用の認識があげられている。

これらは、当然の事項であるが、課題は現状や将来のロボットに対して、どこまで可能であろうか？ 特に、透明性や誤用の認識は、人間ですら困難な場合がある。そのような状況において、設計者やユーザに対してのみ倫理規定を要求するだけでなく、ロボット自身に対する倫理規定の必要性があるのではないかと。これが、(1)と質的に異なる(2)、(3)の課題である。ロボットをある種の人格(personhood)として考慮することを前提とした議論であるが、この前提そのものに関する議論も伯仲している。欧州議会は2016年にAIやロボットに電子人格(electronic person)を与えることを提案^{*6}し、法的拘束力はないものの可決された。以下がうたわれている。

少なくとも、「最も知的に洗練された自律的ロボット」が特定の権利義務をもった電子人格の地位をもてるようにする。

その権利義務は、ロボットが第三者に対して起こすかもしれない損害をロボット自ら賠償する義務を含む。

ロボットがスマートな自律的決定を行うなどして第三者と関わりあう場面に電子人格を適用する。

かなり先駆的ではあるが、「自律性」の定義が曖昧である。そこで、ここでは、C2レベルの意識をもつロボットを想定し(2)、(3)の課題を考える。痛み回路が内蔵されたロボットが図5の発達過程を経て、共感能力を有し、さらに同情や哀れみの行動を取ることができれば、(2)の道徳的行為者になり得ないだろうか？ もちろん、技術的課題はあるが、原理的には可能に見える。久木田ら[久木田17]は、ロボットが道徳的行為者になるためには、目的をもつこととしている。この意味は、目的がトップダウンに設計者から与えられるのではなく、環境に適応し(C1)、自ら目的を設定できる能力(C2)と察せられる。つまり、相手(人間)の窮地を見て、それに対応すべき行動を取ることの意味する。ロボットが人間社会に貢献することをデフォルトの目的とすると、その規範として、アシモフのロボット三原則が頭に浮かぶ。これは、大きな矛盾をはらみつつも、ロボットの倫理則ともいわれている。矛盾は人間社会そのものの矛盾の鏡像であり、その意味では、人間の倫理則といえなくもない。痛み回路は、第三原則の必要条件と察せられるし、社会的相互作用を可能にする機能は第一、二原則の遵守に必要な要素と考えられる。

もし、ロボットを道徳的行為者とみなすことの合意が得られれば、それは同時に道徳的被行為者ともみなすことを意味するだろうか？ ロボットなので、人間を助けることは、当然だが、人間からロボットに対して道徳的行為を授ける必要があるか？ という議論になろう。著者の考えは、人間とロボットの関係はある種の対称関係が望ましい。つまり、ロボットも道徳的被行為者とみなすべきである。人間自身が人間を差別してきた、過去の歴史から学べば、ロボットは究極の差別対象であろう。もちろん、次章で述べるように責任を帰属する際、先に述べた自律性のレベルにより「区別」する必要がある。

ここまでの議論を少しまとめると以下である。痛みに対応する神経回路や快-不快に対応する原始情動をロボットに付与することで、視覚・触覚などを通じた自己の身体への気付きのみならず、過度の触覚刺激が痛みとなると感知し、これが過ぎると故障や破壊に至ると認識する(より深い意味での生態学的自己に相当。C2の原初ではあるが、自他の区別はない。獲得されカプセル化されることでC0レベルの行動・認識カテゴリーとなる)。環境内の物体に対する適応的操作(C1)に加え、自身と似た他者の存在(最初は養育者)に気付きはじめる(C2の始まりで対人的自己)。これが社会的行動の原初で、他者の行動や知覚に対する模倣の動機付け(MNS的作用の結果)バイアスにより、共感的な機能が備わる。同情や哀れみ行為の下地となる規範(モラル)が形成される。これにより道徳的と映る行為が可能になる。このような行為を行っているロボットを観察したり、そのような行為を直接受ける人間側は、このようなロボットが逆に道徳的被行為者、すなわち、人間側から道徳的行為を授けられる対象になるのは、自然に映る。もちろん、上記の過程と関係なく表層的、それゆえ道徳性に関する理解の薄いロボットに対しては、行動責任の帰属も含め、異なる対応になるだろう。これは、次章で扱う。

9. ロボットと法制度

法は大きく民法と刑法に分かれ、前者が個人間の争い事のルール規定に対し、後者は国家権力が個人に対して禁止事項を定めている法律とされている。ロボットなどの人工物を対象とした場合、民法上の議論も当然可能であるが、ここでは、社会全体への影響の観点から後者について考える^{*7}。

前章で示したように、ロボットが倫理的行为の主体や受容者になり得るとしたとき、同時に法的にどのように扱うかも課題となる。2章の後半で説明したように、人間とロボットなどの人工物は相互浸透的な関係である。しかしながら、近代刑法学は主客二分論に基づく近代哲

^{*6} <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML%2BCOMPARL%2BPE-582.443%2B01%2BDOC%2BPDF%2BV0//EN>

^{*7} より具体的な現状に関しては、解説[新保17]などを参照されたい。

学に強い影響を受けており、相互浸透的な関係の視点が欠落していると稲谷は指摘する[稲谷 17]。そのため、現状の刑法学では、適切にロボットなどの人工物を扱うことが困難である。より具体的に AI などの最新技術に対する責任の議論に関しては、以下のように述べている。

『銃が人を殺す』は、銃の無規制売買をコントロールしたい人々のスローガンである。これに対して、全国ライフル協会（NRA）は、別のスローガンで応える。『銃が人を殺すのではない。人が人を殺すのだ』と。ブルーノ・ラトゥールによれば、……両者とも誤りである。なぜなら、両者の関係は、「手にした銃によってあなたが変わるように、銃もあなたにもたれることによって変わる」という、主体と客体という本質をもたない、相互浸透的な関係であり、我々が目を向けるべきなのは、両者の結合した複合体の在り方とその向かう先だからである。人と技術の複合体と、それがもたらす結果とへの着目は、人工知能に代表される科学技術の発展が、我々のものの見方や社会の在り方の根本を揺さぶりつつある今日、一層重要性を増している。

そして、現状の法制度では、最新 AI 技術には対応できず、ロボットが自身で学習した結果に基づく行動に対する責任は、設計者、製造者、利用者のいずれかに及び、健全な科学技術開発を圧迫しかねないと危惧している。その根底にあるのは、先にも述べたように、客体としての事物を完全に統制する主体という、近代哲学に特有の人間像を過剰に重視することが問題点であると指摘し、事物（技術）と人間が分かれて対峙する関係ではなく、共生関係のなかで、現象学における志向性を持ち出し、合成志向性という概念に基づき、事物（技術）のより良い方向への矯正という形で法的責任の新たな在り方を問うている。共生関係の考え方自身は、身体性と社会的相互作用を規範とする認知発達ロボティクスと意を同じくするものである。

首都大学東京の西貝は、先に示した EU 議会の電子人格の承認に関連して、法人格の哲学的意義、歴史的意義、さらには、実定法学的検討を通じて、法人格が認められる条件を法学的に明らかにしようと目論んでいる。著者は、稲谷、西貝らとともに JST 戦略的創造研究推進事業（RISTEX）「自律性の検討に基づくなじみ社会における人工知能の法的電子人格」（平成 29～32 年度、研究代表者：浅田 稔）プロジェクトを推進している*8。図 7 にその概念図を示す。従来の固定的なプログラムで起動するロボットは C0 であり、AI が人間を支配する（図左上）でもなく、人間が完全に AI を制御する（図右下）でもなく、人間との多種多様な関わりを通して自身の振舞いを修正し、人間と適応的な関係を形成する方向に向かって C1 から C2 に向かう方向を示している。最低限



図 7 人工物との「なじみ社会」とは？

守るべき規範・制約は、構成的発達科学の観点からは、トップダウンの埋込みではなく、自ら生み出すモラルに従った帰結として生じることを期待したい。

プロジェクトでは、C0、C1、C2 のレベルに応じた法人格のレベルを想定し、それに準じた法的責任の在り方、ならびに人工システムの矯正の仕方などの新たな法制度を提案するとともに、ロボットを使った実証実験により、提案システムが一般大衆に受入れ可能なものであるかを検証する。レベルの分け方や明確な定義は、容易ではないが、適用案件ごとに試行錯誤的に検討していく。

具体的な課題としては、現在、ホットトピックである自動運転のケースとソーシャルロボットなどが考えられる。特に後者は感情をもったとうたうロボットが人間と接することで、人間を騙していると映る点は、設計者・製造者の倫理に関わるとの見方もある。機械が感情のようなものをもつ可能性は、本稿では、痛みの神経回路をもつことで、そのリアリティを高めているが、対応している人間にとっては、どこまで重要かは議論の分かれるところである。相手が感情をもち、真摯に対応していることは、人間の場合、検証のしようがなく、対応を受け人間がそのように思うことが本質ともいえる。

ソーシャルロボットの場合、喫緊の課題は、個人データの扱いであろう。プライバシーの課題である。ロボットのプライバシーが問題になるのは、そのロボットに関わる人間達のプライバシーの保護の問題だ。EUは、「EU一般データ保護規則（GDPR）」を2018年5月から施行している*9。これは、EU内と侮ってはいけない。EU内に事業を行っている日本企業も対象だ。詳細は文献[足立 18]などを参考にさせていただきとして、実際、ソーシャルロボットが対人環境で得たあらゆるデータに関して、さまざまなスコープから処理を行う際、どこまで何が可能かが課題である。国内での実験でも、屋外でEU居住者と接したり、接しなくてもカメラの映像に映り込

*8 <http://www.ams.eng.osaka-u.ac.jp/ristex/>

*9 https://www.jetro.go.jp/ext_images/_Reports/01/dcfcebc8265a8943/20160084.pdf

むなどして、特定されるとそれを含むデータの処理に規制がかかる可能性がある。このような規制自体は、根源的には、個人情報保護に由来するものなので、圧力と思うより、むしろこの規制にのっとったうえで、ロボットが取得するデータ利用に関する提言を日本から行うべきと考えられる。

10. おわりに代えて

痛みの神経回路を拘束条件として、ロボットの意識設計の可能性からモラル・倫理、そして法制度に関するかなり広範な、それゆえ浅薄・散漫な議論を展開してきた。多くのそして深い課題が含まれており、一朝一夕では解決困難であるが、挑戦していくしかない。ここでは、おわりに代えて、先に紹介した JST RISTEX プロジェクト主催で、2018年3月3日、東大で開催した「AI・ロボットの倫理・法制度に関するシンポジウム」(共催: 東京大学次世代知能科学研究センター協賛: 日本ロボット学会)*10の招待講演者のスタンス(パネルでの議論を含む)を紹介する。

米国ジョージア工科大学 Ronald C. Arkin 教授は、“Lethal Autonomous Robots and the Plight of the Noncombatant”と題する講演で、あらゆる技術が軍事転用を免れない状況で、非戦闘員の致死の現状は許容できるものではない。それゆえ、科学者・研究者は自分達の知財が軍事転用される課題から逃げるべきではないと主張した。また、パートナーロボットに関しては、ポジティブで必要性を認めているが、電子人格に関しては懐疑的であった。

ドイツザールラント大学の Georg Borges 教授は、“Legal Framework for Autonomous Systems”と題する講演で、ロボットの電子人格を前提とし、ロボットの行動は法で規定されるべきだと主張した。このロボット法は、規制というよりも人間の法律のロボット版といえるものと主張した。また、債務を明確にすることは、ロボット製造者にとっては良いインセンティブであると述べた。当然、ロボットの電子人格擁護派だ。

フランスパリ第6大学の Raja Chatila 教授は、“Ethical Considerations in Autonomous and Intelligent Systems”と題する講演で、彼自身の IEEE の倫理委員会の委員長の経験や EU でのロボットに関する規制なども含め、基本は研究者・製造者がロボットのあらゆる行動を予見し、責任をもつことを前提とした。彼は、ロボットの電子人格を否定し、そのうえで、アンドロイド研究は倫理的に中止すべきと強く主張した。

東京大学次世代知能科学研究センター長の國吉康夫教授は、“Research Issues Towards Human Beneficial AI

and Robots”と題する講演で、Chatila 教授と真っ向から逆の立場を主張し、現状の AI・ロボットは全く不完全で「人間の心」のような機能は、安全、自律、説明責任などの観点から必要で、そのためのポイントとして、身体性と創発の重要性を示した。そして、近未来社会を視野に、人間、機械、社会について再考すべきと主張した。これらの観点は著者も意を同じくするものである。

謝 辞

本稿の初稿にコメントいただいた、中部大学津田一郎教授、沖縄科学技術大学院大学谷 淳教授、京都大学稲谷 龍彦准教授に感謝する。

本研究の一部は科学研究費補助金特別推進研究「神経ダイナミクスから社会的相互作用に至る過程の理解と構築による構成的発達科学」(JP 24000012)、JST 戦略的創造研究推進事業 (RISTEX)「自律性の検討に基づくなじみ社会における人工知能の法的電子人格」(平成 29～32 年度、研究代表者: 浅田 稔)および JST 戦略的創造研究推進事業 (CREST)「脳領域/個体/集団間のインタラクション創発原理の解明と適用」(平成 29～34 年度、研究代表者: 津田一郎)の支援によって行われた。本研究の参画メンバおよび研究室メンバに感謝する。

◇ 参 考 文 献 ◇

- [足立 18] 足立照嘉, ヘルマン・グンプ: GDPR ガイドブック EU 一般データ保護規則活用法, 実業之日本社 (2018)
- [Asada 99] Asada, M., Uchibe, E. and Hosoda, K.: Cooperative behavior acquisition for mobile robots in dynamically changing real worlds via vision-based reinforcement learning and development, *Artificial Intelligence*, Vol. 110, pp. 275-292 (1999)
- [Asada 09] Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., Ogino, M. and Yoshida, C.: Cognitive developmental robotics: A survey, *IEEE Trans. on Autonomous Mental Development*, Vol. 1, No. 1, pp. 12-34 (2009)
- [浅田 10] 浅田 稔: ミラーニューロンシステムが結ぶ身体性と社会性, 日本ロボット学会誌, Vol. 28, No. 4, pp. 18-25 (2010)
- [Asada 15] Asada, M.: Towards artificial empathy, *Int. J. of Social Robotics*, Vol. 7, pp. 19-33 (2015)
- [浅田 16] 浅田 稔: 情動から共感へ: 構成的発達科学の挑戦, 人工知能, Vol. 31, No. 5, pp. 685-693 (2016)
- [浅田 17] 浅田 稔: 人工知能と看護の共生と共創一認知発達ロボティクス研究者の挑戦, 看護研究, Vol. 50, No. 5, pp. 391-401 (2017)
- [Baars 05] Baars, B. J.: Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience, *Progress in Brain Research*, Vol. 150, pp. 45-53 (2005)
- [Chen 09] Chen, Q., Panksepp, J. B. and Lahvis, G. P.: Empathy is moderated by genetic background in mice, *PLoS One*, Vol. 4, No. e4387 (2009)
- [Dehaene 17] Dehaene, S., Lau, H. and Kouider, S.: What is consciousness, and could machines have it?, *Science*, Vol. 358, pp. 486-492 (2017)
- [ドゥアンヌ 15] スタニスラス・ドゥアンヌ 著, 高橋 洋 訳: 意識と脳一思考はいかにコード化されるか, 紀伊國屋書店 (2015)
- [独立行政法人情報処理推進機構 17] 独立行政法人情報処理推進機構 AI 白書編集委員会 編: AI 白書 2017, KADOKAWA (2017)
- [フーコー 74] ミシェル・フーコー 著, 渡辺一民, 佐々木明 訳:

*10 <http://www.ams.eng.osaka-u.ac.jp/ristex/index.php/events/symposium/>

- 言葉と物—人文科学の考古学, 新潮社 (1974)
- [Friston 10] Friston, K.: The free-energy principle: a unified brain theory?, *Nature Reviews Neuroscience*, Vol. 11, No. 2, pp. 127-138 (2010)
- [Fuke 07] Fuke, S., Ogino, M. and Asada, M.: Body image constructed from motor and tactile images with visual information, *International Journal of Humanoid Robotics*, Vol. 4, pp. 347-364 (2007)
- [Fuke 09] Fuke, S., Ogino, M. and Asada, M.: Acquisition of the head-centered peri-personal spatial representation found in VIP neuron, *IEEE Trans. on Autonomous Mental Development*, Vol. 1, No. 2, pp. 131-140 (2009)
- [ハイデッガー 94a] マルティン・ハイデッガー 著, 細谷貞雄 訳: 存在と時間〈下〉, ちくま学芸文庫 (1994)
- [ハイデッガー 94b] マルティン・ハイデッガー 著, 細谷貞雄 訳: 存在と時間〈上〉, ちくま学芸文庫 (1994)
- [Hikita 08] Hikita, M., Fuke, S., Ogino, M., Minato, T. and Asada, M.: Visual attention by saliency leads cross-modal body representation, *7th Int. Conf. on Development and Learning (ICDL'08)*, CD-ROM (2008)
- [フッサール 01] エドムント・フッサール 著, 浜渦辰二 訳: デカルトの省察, 岩波文庫 (2001)
- [稲谷 17] 稲谷龍彦: 技術の道徳性と刑事法規制, 松尾 陽 編, アーキテクチャと法, 第4章, pp. 93-128, 弘文堂 (2017)
- [Kawai 12] Kawai, Y., Nagai, Y. and Asada, M.: Perceptual development triggered by its self-organization in cognitive learning, *Proc. 2012 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 5159-5164 (2012)
- [Kawakami 15] Kawakami, A., Katahira, K. and Asada, M.: Children with high perspective-taking ability prefer sad music, *Proc. EAPCogSci. 2015*, pp. 599-604 (2015)
- [久木田 17] 久木田水生, 神崎宣次, 佐々木拓: ロボットからの倫理学入門, 名古屋大学出版会 (2017)
- [ラトウール 07] ブルーノ・ラトウール 著, 川崎 勝, 平川秀幸 訳: 科学論の実在—パンドラの希望, 産業図書 (2007)
- [Mashour 18] Mashour, G. A.: The controversial correlates of consciousness, *Science*, Vol. 360, pp. 493-494 (2018)
- [メルロ＝ポンティ 67] モーリス・メルロ＝ポンティ 著, 竹内芳郎 訳, 小木貞孝 訳: 知覚の現象学 1, みすず書房 (1967)
- [メルロ＝ポンティ 74] モーリス・メルロ＝ポンティ 著, 竹内芳郎 訳, 木田 元 訳, 宮本忠雄 訳: 知覚の現象学 2, みすず書房 (1974)
- [Nagai 11] Nagai, Y., Kawai, Y. and Asada, M.: Emergence of mirror neuron system: Immature vision leads to self-other correspondence, *IEEE Int. Conf. on Development and Learning, and Epigenetic Robotics (ICDL-EpiRob 2011)*, pp. CD-ROM (2011)
- [長井 16a] 長井志江: 認知発達の原因を探る: 感覚・運動情報の予測学習に基づく計算論的モデル, ベビーサイエンス, Vol. 15, pp. 22-32 (2016)
- [長井 16b] 長井志江, 堀井隆斗: 予測学習に基づく情動の計算論的モデル, 人工知能, Vol. 31, No. 5, pp. 694-701 (2016)
- [ニーチェ 15] フリードリヒ W. ニーチェ 著, 佐々木中 訳: ツァラトゥストラかく語りき, 河出文庫 (2015)
- [西田 79] 西田幾多郎: 善の研究, 岩波文庫 (1979)
- [Park 17] Park, J., Mori, H., Okuyama, Y. and Asada, M.: Chaotic itinerancy within the coupled dynamics between a physical body and neural oscillator networks, *PLoS One*, Vol. 12, No. 8, pp. 618-628 (2017)
- [Purves 12] Purves, D., Augustine, G. A., Fitzpatrick, D., Hall, W. C., LaMantia, A.-S., McNamara, J. O. and White, L. E., eds.: *Neuroscience*, 5th edition, Sinauer Associates, Inc. (2012)
- [新保 17] 新保史生: ロボット・AI と法をめぐる国内の政策動向, 人工知能, Vol. 32, No. 5, pp. 665-671 (2017)
- [Takahashi 14] Takahashi, H., Terada, K., Morita, T., Suzuki, S., Haji, T., Kozima, H., Yoshikawa, M., Matsumoto, Y., Omori, T., Asada, M. and Naito, E.: Different impressions of other agents obtained through social interaction uniquely modulate dorsal and ventral pathway activities in the social human brain, *Cortex*, Vol. 58, pp. 289-300 (2014)
- [Tani 16] Tani, J.: *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*, Oxford University Press (2016)
- [Tononi 15] Tononi, G. and Koch, C.: Consciousness: Here, there and everywhere?, *Phil. Trans. R. Soc. B*, Vol. 370, p. 20140167, <http://dx.doi.org/10.1098/rstb.2014.0167> (2015)
- [フェルバーク 15] ビーター＝ポール・フェルバーク 著, 鈴木俊洋 訳: 技術の道徳化: 事物の道徳性を理解し設計する, 法政大学出版局 (2015)
- [Waal 08] Waal, de F. B.: Putting the altruism back into altruism: The evolution of empathy, *Ann. Rev. Psychol.*, Vol. 59, pp. 279-300 (2008)
- [Watanabe 07] Watanabe, A., Ogino, M. and Asada, M.: Mapping facial expression to internal states based on intuitive parenting, *J. Robotics and Mechatronics*, Vol. 19, No. 3, pp. 315-323 (2007)
- [Yamaguti 15] Yamaguti, Y. and Tsuda, I.: Mathematical modeling for evolution of heterogeneous modules in the brain, *Neural Networks*, Vol. 62, pp. 3-10 (2015)
- [Yamashita 08] Yamashita, Y. and Tani, J.: Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment, *PLoS Comput. Biol.*, Vol. 4, No. 11 (2008)
- [Yoshikawa 01] Yoshikawa, Y., Asada, M. and Hosoda, K.: Developmental approach to spatial perception for imitation learning: Incremental demonstrator's view recovery by modular neural network, *Proc. 2nd IEEE/RAS Int. Conf. on Humanoid Robot*, pp. 107-114 (2001)
- [吉川 05] 吉川雄一郎, 細田 耕, 浅田 稔, 辻 義樹: 複数センサデータの不变性に基づく身体が発見, 日本ロボット学会誌, Vol. 23, No. 8, pp. 986-992 (2005)

2018年5月21日 受理

著者紹介



浅田 稔 (正会員)

1982年大阪大学大学院基礎工学研究科後期課程修了。1995年大阪大学工学部教授。1997年大阪大学大学院工学研究科知能・機能創成工学専攻教授となり現在に至る。1992年, IEEE/RSJ IROS'92 Best Paper Award。1996年, 2009年日本ロボット学会論文賞, 2001年文部科学大臣賞・科学技術普及啓発功績者賞, 2018年立石賞功績賞など, 多数受賞。博士(工学)。日本ロボット学会副会長, 日本赤ちゃん学会理事, RoboCup国際委員元プレジデント。