公開

研 究 終 了 報 告 書

「脳情報に基づいた AI の信頼性評価技術の開発」

研究期間：2020 年 12 月～2024 年 3 月
研 究 者：西田 知史

1. 研究のねらい

人間が AI の提示情報に対して主観的に信頼性を感じる脳内メカニズムを解明し、それに基づいて主観的信頼性を脳活動から解読して AI を評価するための技術を開発する。信頼できる AI を開発するうえで、AI が扱う情報の安全性や透明性といった客観的信頼性の向上が課題である一方、人間が AI の提示情報に対して主観的にどれくらい信頼できるかという主観的信頼性の向上も同等に重要な課題である。主観的信頼性は客観的信頼性と必ずしも一致せず、人間が持つ AI への負の感情（例：AI による支配への恐怖）などにも強く影響を受ける。本研究では、主観的信頼性を生み出す脳内メカニズムに着目するという斬新な発想と、研究代表者が有する世界最先端の脳科学の知識と手法に基づいて、AI の主観的信頼性の向上という課題に対し、脳活動の解読に基づいた主観的信頼性の評価技術を開発する形で解決を目指す。まず、脳科学の実験手法を用いて、AI に対する主観的信頼性を感じる脳内メカニズムを世界で初めて解明する。続いて、計測した脳活動から主観的信頼性を解読して評価する新規技術を開発する。さらに、この技術における脳計測コストの問題を解決するため、研究代表者が最近開発した、シミュレートした脳活動から認知内容を解読する革新的手法を取り入れる。本研究の成果は、信頼できる AI の開発において汎用的に利用可能な主観的信頼性評価のための画期的技術を生み出し、信頼できる AI の開発のブレイクスルーとなりうる。

2. 研究成果

（1）概要

本研究計画では、信頼される AI の基盤技術開発に資することを目的として、認知神経科学の方法論と知見を活用した 2 つのアプローチで AI の信頼性に関する研究を進めてきた。1 つ目のアプローチは、人間が AI を信頼するとは何かという根本的な問いに対する認識論的説明を得ることを目的とした、AI への信頼性を生み出す脳内基盤の探究である。2 つ目のアプローチは、信頼される AI に不可欠な要因として人間らしい判断を行う AI を実現することを目的とした、脳情報のモデルを AI へ融合する技術の開発と検証である。1 つ目のアプローチにおける成果として、AI 生成情報に対する人間の価値判断において潜在的にはたらく負のバイアスと、それが脳の応答性にもたらす影響を明らかにした。また、人間が AI に抱く不安感の個人差を生み出す脳内基盤について理解を得た。2 つ目のアプローチにおける成果として、AI と脳情報の融合によって、AI に脳情報の個人差を反映させることに成功した。また同じく脳融合が、AI の内部情報を脳の内部情報へと近づける効果を持つことも判明した。

（2）詳細

【AI 生成情報に対する人間の負の価値判断バイアス】

AI に抱く負のイメージが AI 生成画像に対する人間の魅力度判断にもたらす影響とその脳内機序についての探究を行った。この目的のため、被験者 60 名（男性 30 名、女性 30 名、平均年齢 23.3 歳）を対象に、図 1 に示す認知課題（魅力度評定課題）を実施し、その際の脳活動を fMRI により計測した。

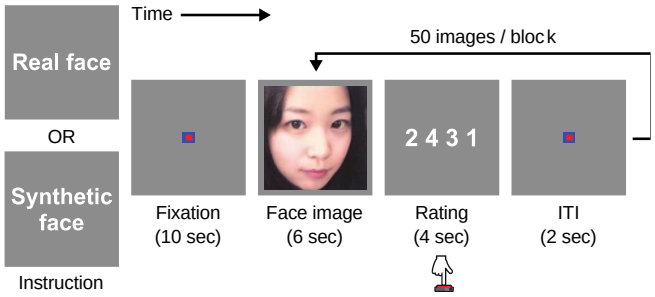

図 1　魅力度評定課題

この課題で被験者は、提示される顔画像が実在の人物のもの（実在顔）であるか、AI が作り出した架空の人物のもの（AI 顔）であるかについて事前に教示を受け、連続して提示される顔画像に対して被験者が感じる魅力度を 4 段階で評定した。評定はボタン押しで行うが、特定の指と特定の評定値が結びつかないように、試行ごとに数字とボタン位置の対応をランダムに変化させた。50 枚の顔画像から構成されるブロックごとに、実在顔と AI 顔に関する教示が交互に切り替わり、全部で 6 ブロック（300 枚）の評定を実施した。ただし、実際には実験統制のため、全ての顔画像は AI（StyleGAN [Karas+20]）が生成したものであり、被験者は教示を信じて一方を実在顔、他方を AI 顔という事前知識のもと、魅力度を評定した。300 枚の顔画像は全て異なるものであり、半分を実在顔、残り半分を AI 顔に割り当てて、その割り当てを被験者ごとに反転して、被験者全体でバランスをとった（図 2）。

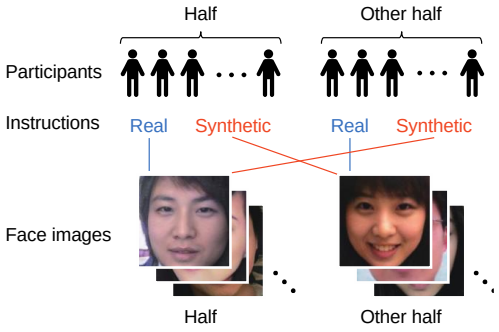

図 2　被験者による割り当ての反転

この偽の教示を利用した実験デザインを用いると、全く同じ顔画像の魅力度評定に対して、実在顔と AI 顔という思い込みだけを操作することが可能になり、画像に依存しない負のイメージの効果そのものを検証することができる。

この課題中に被験者から得られた評定値を基に、「AI に対する負のイメージは、AI 生成顔画像の魅力度を低く見積もらせる」という仮説を検証した。この仮説が正しければ、被験者平均の評定値において、実在顔に比べて AI 顔の方が低くなると予想される。実在顔条件の評定値から AI 顔条件の評定値を差し引いた値を underestimation score として、図 3 に underestimation score の分布を示す。図 3A は顔画像ごとに計算した underestimation score の分布、図 3B は被験者ごとに計算した underestimation score の分布に相当する。図中の赤い垂直線は平均値で、赤い領域は 95%信頼区間である。統計検定の結果、顔画像および被験者のいずれの underestimation score においても、0 より有意に大きくなった（対応あり t 検定、$P < 0.05$）。これは、AI 顔条件で魅力度が低く見積もられることを示しており、仮説の正しさが確認できた。

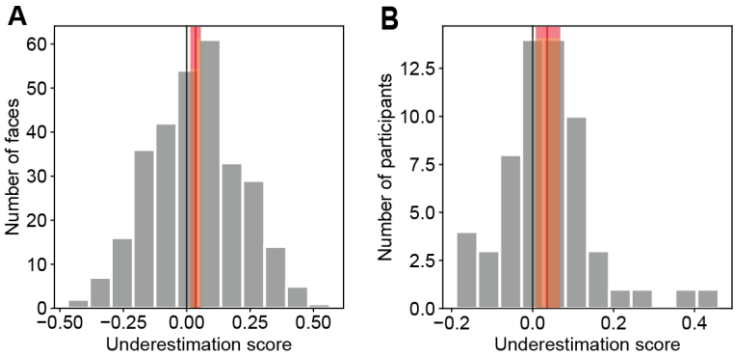

図 3　実在顔条件から AI 顔条件への魅力度評定の低下量（underestimation score）の分布

さらに、この課題中に脳活動を fMRI で計測し、顔の魅力度に依存して活動に変化が生じる脳領域を同定した。そして、その領域の中から、教示によって魅力度への応答性に変化が生じる脳領域を特定した。その結果、図 4 に示す領域（右紡錘状皮質）において、条件の違いによる魅力度評定の低下量（underestimation score）と、魅力度に対する応答性の変化量に有意な相関関係が確認できた（ピアソン相関、$r = 0.45$、$P < 0.005$）。

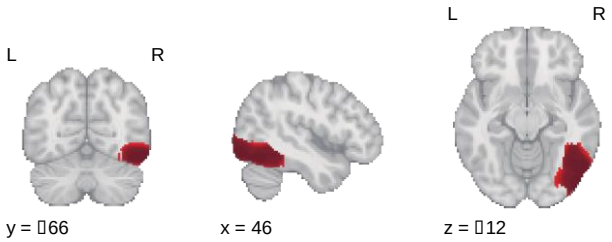

図 4　条件による魅力度評定の低下と相関して応答性が低下した領域（右紡錘状皮質）

以上のように、提示された顔画像が AI 生成物であるという思い込みだけで、顔の魅力度判断が変わると同時に、脳の特定領域において魅力度への応答性の変化が生じることが分かった。特に、紡錘状皮質は顔知覚に重要なはたらきを行う脳領域として知られているため [Kanwisher+97]、AI 生成物であるという思い込みは顔の知覚情報処理を変容し、顔に対する

魅力度の主観的経験に変化を生じさせている可能性が示唆された。

これまで、人工顔に対する魅力度の低下は、例えば不気味の谷現象[Mori+12]などの文脈で調べられてきた。しかし、不気味の谷現象は、自然物と人工物の見た目に差異があるときに生じるものであり、人工物が自然物と識別できないくらいに近づけば現象は解消されるとされてきた。しかし、本研究の成果は、例えば見た目で両者が見分けられなくなっても、見ている対象が人工物（AI 生成物）であるという思い込みだけで魅力度の低下が生じることを示唆している。したがって、AI 生成物に対する負のバイアスを生じさせないためには、AI 生成物の見た目を自然物に近づけるだけでなく、AI に対して人間が潜在的に持つ負のイメージの払拭も重要であるという示唆が得られた。

【AI に抱く不安感の個人差を生み出す脳内基盤】

近年の AI 技術の発展は目覚ましく、我々の社会への普及も順調に進んでいるように見える。しかし、依然として AI 技術に対して不安感（例：AI による支配、AI の暴走）を抱く人々は多い。そのような不安感は人間と AI が調和のうちに共存する社会の実現を阻む大きな要因の一つになりうる[Siau18]。したがって、そのような不安感を払拭するために、不安感を生み出す認知機序の理解が喫緊の課題だといえる。

本研究の目的は、AI に対する不安感の個人差に着目し、そのような個人差に伴って変化する感覚応答の局在を調べることにより、不安感に関わる脳内基盤を同定することである。不安感のような AI に対する潜在的イメージは、イメージと関連する感覚情報の価値判断を変容するとともに、上述の研究のように脳の感覚応答にも変化をもたらす。したがって、AI に対する不安感と関わる脳領域では、AI 関連情報に対する感覚応答の個人差が、不安感の個人差と共変することが予想される。

これを検証するため、被験者 59 名を対象に、AI・ロボットに関わる映像を視聴している際の脳応答を fMRI により計測し、被験者間相関 [Hasson04] に基づく指標を用いて、各脳領域における感覚応答の個人差を評価した。さらに、同一被験者において、質問票を用いて AI に対する不安感を定量化したうえで、不安感の個人差を評価した。そして、感覚応答の個人差が不安感の個人差と相関関係を持つ脳領域を分析した。

その結果、いくつかの脳領域における活動の非類似度が AI に抱く不安感の非類似度と相関することが分かり、特に上側頭皮質における相関が顕著であった（図 5）。

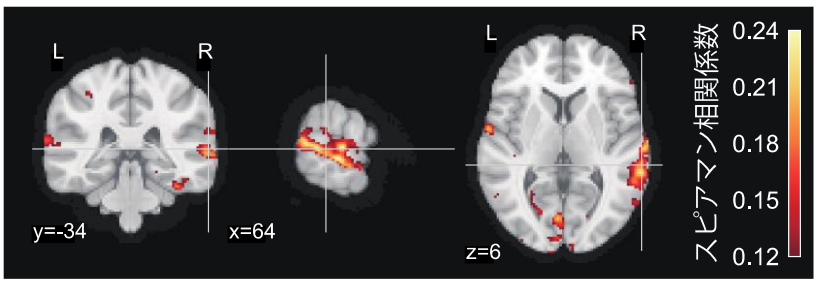

図 5 AI に対する不安感の個人差に相関する個人差を示した脳領域

上側頭皮質は社会性認知に関与すると考えられており、特に心の理論[Leroy 15]、表情認識[Direito 19]、生物運動知覚[Herrington 11]などに関わることが知られている。本研究で用いた映像視聴実験にて、被験者は AI・ロボットが登場する映像を視聴した。そのため、上側頭皮質にて相関が見られた一つの解釈として、AI に不安感を感じる人ほど AI・ロボットを無機的な物体として捉え、不安感を感じない人ほど AI・ロボットを社会的なエージェントとして捉えるという傾向が、上側頭皮質における個人差相関として表出したと考えられる。

【脳情報の個性を反映する脳融合 DNN】

図 6 に技術の概要を示す。この技術では、まず映像を入力した畳み込みニューラルネット（CNN）の中間層活性化パターンから個人の脳活動を予測する回帰モデル（予測モデル）を学習する。続いて、学習済みの予測モデルを用いて算出した予測脳活動から、映像に対応する多様なラベルを推定する回帰モデル（解読モデル）を学習する。これらの学習が完了すると、新たな脳計測を要さずに任意の映像入力からラベルを推定するシステムが構築できる。

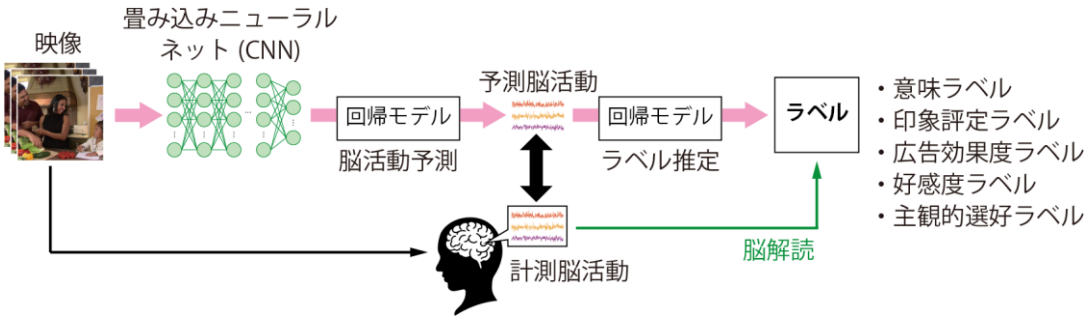

図 6　DNN へ脳情報を融合する技術の概要

ここでは、個人の脳活動から作成した脳融合 AI の推定ラベルが、脳情報処理の個人差を反映することを確かめるため、同じ映像を視聴時の個人の計測脳活動から解読したラベルとの関係を調べた。具体的には、脳融合 AI の推定ラベルの個人間被類似度と、計測脳活動から解読したラベルの個人間非類似度をそれぞれ計算し、両者の個人間非類似度の相関関係を分析した。もし脳融合 AI が個人差を反映するのであれば、統計学的に有意な相関関係が認められるはずである。

図 7A に、有意な相関関係が見られたラベルの例を示す。このラベルは映像シーンに対して付与された文章記述であり、映像の意味認知内容の個人差を反映するかを検証した結果である。図中の各点は脳融合 AI と脳解読における各個人ペア間の非類似度を示す。スピアマン相関係数で 0.74 という高い値が観測され（$P < 0.0001$）、脳融合 AI システムが高い精度で個人差を反映することを示唆した。また、全 87 個のラベルの推定結果にて検証したところ、図 7B に示す全ラベルの相関係数の分布から見て取れるように、実に 81 個のラベルで有意な相関関係が認められ（色付きの棒グラフ、多重比較補正後 $P < 0.05$）、脳融合 AI システムは脳情報処理の個人差を大部分のラベル（93.1%）で反映できていることが確認された。

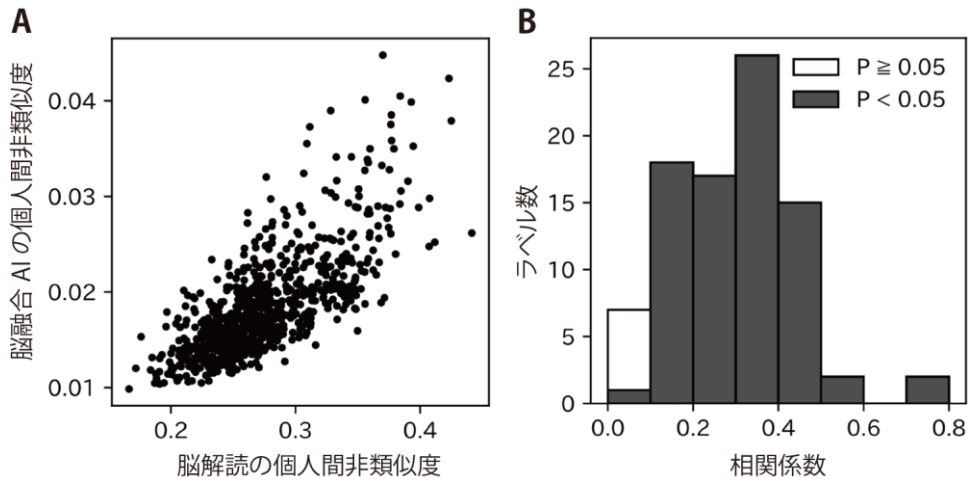

図 7 ラベル推定結果における脳融合 AI と脳解読の相関関係

以上の結果は、研究代表者が開発した脳融合 AI 技術が、個人の脳情報を利用することによって、人間らしい個性を反映しうることを示唆している。前年度の成果で、同技術は AI の判断を人間の判断に近づけることも分かっているため、人間らしい AI を実現するうえで、同技術が大きく貢献しうるといえるだろう。また、同技術は AI が提示する情報に対する個々人の信頼性評価にも利用することを計画している。脳融合技術が AI に人間らしい判断と個性を持たせることは、その用途においても重要なファクターとなるといえ、今年度までの研究でその基礎固めに成功したといえる。

【DNN の内部情報を脳らしくする脳融合】

同様の脳融合技術を用いることで、CNN の内部情報が脳の内部情報に近づくことも確認できた（図 8）。ここでは、脳融合 CNN および標準 CNN から解読モデルで推定したラベルの時系列と、計測脳活動から脳解読で読み取ったラベルの時系列の類似度を、ラベルごとに比較した。脳融合で脳の内部情報に近づくなら、脳解読の結果に近づくはずである。

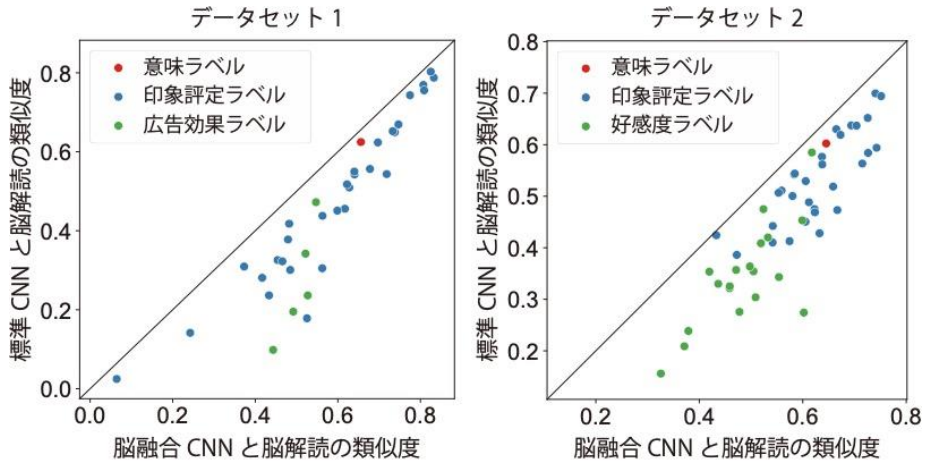

図 8 脳融合 CNN および標準 CNN と脳解読との類似度

図 8 の左図と右図は異なる映像と脳活動のペアデータで検証した結果を示す。図中の各点は、映像に付与された異なる推定ラベルに対応し、赤色の点は意味ラベル、青色の点は印象評定ラベルのグループ、緑色の点は映像の広告効果度または好感度のグループを表す。X 軸が脳融合 CNN と脳の推定ラベルの類似度、Y 軸が標準 CNN と脳の推定ラベルの類似度に対応する。いずれの映像・脳活動ペアデータにおいても、全ての推定で脳融合 CNN と脳の類似度が、標準 CNN と脳の類似度を有意に上回った（対応あり t 検定、$P < 0.00001$）。したがって、CNN へ脳情報を融合することにより、推定結果が脳の解読結果に近づくこと、すなわち脳の内部情報に近づくことが示唆された。

このように AI の内部情報を脳の内部情報へ近づける技術は、人間らしい振る舞いを示す AI の実現に寄与すると考えられる。また、本技術を介して人間と感じ方や考え方が似ることで、AI アラインメントを効果的に実現する技術として、人間中心の未来情報社会の実現に多大な貢献を示すと期待される。

3. 今後の展開

本研究の最終目標は、AI の信頼性を評価するために、脳情報を取り入れた評価技術を創出することであった。しかし、認知課題において AI の信頼性を評価するパラダイムを考案することが難しく、また AI に対する信頼性とは何かという根本的な疑問に直面しているため、信頼性自体を評価するにはまだ至っていない。ただし、この課題についても二方向から解決に取り組んでおり、近日中の進展が期待できる。まず、一つの方向性として、AI の信頼性とは何かについて、質問票調査などを取り入れて探究するとともに、哲学者との共同研究を介して AI の信頼性の概念分析を進めている。もう一つの方向性として、AI に対する信頼性の一側面を、人工物らしさや外見の信頼性などに単純化して捉え、それらを評価する技術の実装と実証を進めている。

本研究で進めてきた 2 つのアプローチのうち、AI に対する信頼性の認識論的研究では、AI に対する潜在的イメージがいとも簡単に AI 自体や AI の提示情報への感じ方に変容をもたらすことを示した。不気味の谷現象に挙げられるように、従来の研究では AI の外見が重視されてきたが、本研究の成果は外見と無関係な潜在的イメージも同様に、信頼性を形成する重要な要因であることを示し、関連研究分野にパラダイムシフトを促す重要な貢献を果たした。今後の研究でも、そ

のような人間の内在的な性質の観点から、人間と AI の関係性を捉える研究を進めていく予定である。さらに、内在的な性質も考慮して、人間の AI に対する認知・神経機序を包括的に捉えた計算論モデルの構築も進めている。そのようなモデルが創出できれば、人間と AI のインタラクション研究に大きな進歩をもたらすと考えている。

一方、2 つのアプローチのうち、人間らしい AI の実現を見据えた脳融合技術の開発については、信頼性の評価という方向からだけではなく、別の方向から信頼される AI の実現に貢献していくビジョンも描いている。特に、AI に対して未だに人間が優れている、汎用性、柔軟性、ロバスト性などを、脳融合によって AI に付与する可能性を検証していくつもりである。具体的には、脳融合によって AI の敵対的攻撃に対する耐性を向上させたり、フェイク映像の検出性能を上げたりするような利用方法も視野に入れ、実際に研究を進めている。

また、本研究において脳融合技術の成果として得られた、脳情報の個人差の模倣に関する知見は、脳融合技術の応用可能性を格段に向上させたといえる。つまり、脳情報の個人差の模倣は、脳情報処理の個性を計算機上で人工的に再現する脳デジタルツインの実現に直接つながる大きな発見である。今後は、個々人の脳情報を模倣する性能を向上させていき、脳デジタルツインを世界に先駆けて実現することを目標に見据えている。そのような技術は、人間を空間的・時間的制約から解き放ち、社会に素晴らしい変革をもたらすと期待している。

4. 自己評価

上述のように、主観的信頼性の評価技術を確立するには至っていないが、当初の研究目標から派生して得られた、AI に対する潜在的イメージの脳内メカニズムについての知見や、脳情報の個人差を反映する脳融合技術の知見は、学術および社会に変革をもたらしうる重要性と発展性を有している。また、成果は順調に蓄積されており、論文や学会等での発表件数は十分に多く、近日中にもさらなる成果の更新が望めることから、研究は極めて順調に遂行されているといえ、本研究は高い評価に値すると自負している。

5. 主な研究成果リスト

(1)代表的な論文(原著論文)発表

研究期間累積件数: 件

1. Nishida S, Blanc A, Maeda N, Kado M, Nishimoto S. Behavioral correlates of cortical semantic representations modeled by word vectors. PLOS Computational Biology, 17(6): e1009138, 2021.	1.
2. Kawahata K, Wang J, Blanc A, Maeda N, Nishimoto S, Nishida S. Decoding Individual Differences in Mental Information from Human Brain Response Predicted by Convolutional Neural Networks. bioRxiv:2022.05.16.492029, 2022.	
3. Shinkuma R, Nishida S, Maeda N, Kado M, Nishimoto S. Reduction of information collection cost for inferring brain model relations from profile. IEEE Transactions on Systems, Man and Cybernetics: Systems, 52(7):4057–4068, 2022.	2.
4. Kawasaki H, Nishida S, Kobayashi I. Hierarchical Processing of Visual and Language	

Information in the Brain. Proceedings of AACL-IJCNLP 2022, 405–410, 2022.	
5. Matsumoto Y, Nishida S, Hayashi R, Son S, Murakami A, Yoshikawa N, Ito H, Oishi N, Masuda N, Murai T, Friston K, Nishimoto S, Takahashi H. Disorganization of semantic brain networks in schizophrenia revealed by fMRI. Schizophrenia Bulletin, 49(2):498–506, 2023.	3.
6. Nishida S. Behavioral and neural evidence for the underestimated attractiveness of faces synthesized using an artificial neural network. bioRxiv:2023.02.07.527403, 2023.	
7. Kawasaki H, Nishida S, Kobayashi I. Exploring Hierarchical Changes in Functional Brain Network Hubs through Brain-Activity Prediction with Convolutional Neural Networks. Proceedings of 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 4740–4745, 2023.	
8. Wang J, Kawahata K, Blanc A, Maeda N, Nishimoto S, Nishida S. Asymmetric representation of symmetric semantic information in the human brain. bioRxiv:2024.02.09.579613, 2024.	

(2)特許出願

研究期間全出願件数:0 件(特許公開前のものは件数にのみ含む)

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

1. Kuroda E, Nishimoto S, Nishida S, Kobayashi I. A Deep Generative Model imitating Predictive Coding in the Human Brain. The 22nd International Symposium on Advanced Intelligent Systems, G04-4. Online. Dec 16, 2021. (Best Session Award)
2. Kawahata K, Blanc A, Nishimoto S, Nishida S. Individual differences of natural perceptual content in the human brain can be estimated via brain response prediction using deep neural networks. The 7th CiNet Conference: New horizons in brain mapping, 1-3. Online. Feb 1, 2022.
3. Wang J, Blanc A, Nishimoto S, Nishida S. Symmetric pairs of semantic information are represented with little overlap in the human brain. 2022 Organization for Human Brain Mapping Annual Meeting, WTh084. Glasgow, UK. Jun 8, 2022.
4. Kawahata K, Wang J, Blanc A, Nishimoto S, Nishida S. Making information representations of deep neural networks more brain-like via models for brain-response prediction. International Symposium on Artificial Intelligence and Brain Science 2022, A08. Okinawa, Japan. Jul 4, 2022.
5. Wang J, Kawahata K, Blanc A, Nishimoto S, Nishida S. Asymmetry in Representations of Semantic Symmetry in the Human Brain. International Symposium on Artificial Intelligence and Brain Science 2022, B14. Okinawa, Japan. Jul 4, 2022.
6. Taguchi H, Nishida S, Nishimoto S, Kobayashi I. Validation of the Role of Attention Mechanism in Predicting Brain Activity. SCIS&ISIS2022, T-3-G-123. Mie, Japan. Dec 1, 2022.
7. Maeda C, Nishida S. Diversity of music preferences is reflected in intersubject synchronization of fMRI signals. IBRO 2023, AS13-A13. Granada, Spain. Sep 9, 2023.
8. Wang J, Nishida S. Individual variability in AI anxiety reflected in intersubject

synchronization of movie-evoked fMRI signals. IBRO 2023, AS13-A19. Granada, Spain. Sep 11, 2023.
9. 人工知能学会 2022 年度全国大会: 優秀賞, Jul 2022
10. 日本神経回路学会: 優秀研究賞, Sep 2023
11. 西田知史. 日常的な認知に関わる脳情報処理のモデル化と人工脳への応用. 情報通信研究機構研究報告 68(1): 11‐19, 2022.
他 49 件