



CHAIN

Center for Human Nature,
Artificial Intelligence, and Neuroscience

人間知×脳×AI研究教育センター (CHAIN) の概要

北海道大学 人間知×脳×AI研究教育センター長

北海道大学 大学院文学研究院・教授

田口 茂

Contents

- 1. 人間知×脳×AI研究教育センターとは何か
- 2. これまでの成果

1. 人間知×脳×AI研究教育センターとは何か

CHAINの研究・教育の特徴

センターの名称

- 日本語名称

人間知×脳×AI研究教育センター

- 英語名

Center for **H**uman **N**ature, **A**rtificial Intelligence, and **N**euroscience

- 略称 CHAIN

- 2019年7月1日設立（北海道大学学内共同施設）



学位

文学院

学位

理学院

学位

生命科学院

学位

情報科学院

学位

生命科学院

学位

農学院

学位

教育学院

学位

経済学院

学位

医学院

学位

医理工学院

.....等々

各大学院で専門分野を学び、
CHAINでは学際的能力を磨く

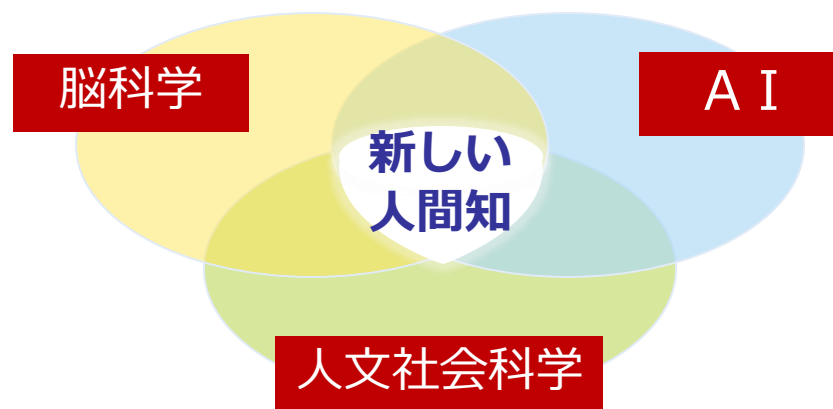
修了証

CHAIN教育
プログラム
(特別プログラム)



センターの目的

- 人文社会科学・神経科学（脳科学）・AI（人工知能）という三つの分野が交差する地点で、学際的・文理融合的教育・研究を行う。
- そこにおいて、新しい「人間知」が成立する。



人間についてフル装備で考える

「人間とは何か？」誰にとっても無関心でいられない問い。

哲学の問い？宗教の問い？文学の問い？

科学の問いでもある。

AIがますます発達してゆく現代において、まさに喫緊の問い。

CHAINでは、多様な学問をフルに動員することで、この問いに挑もうとする。



人間に迫る脳科学とAI

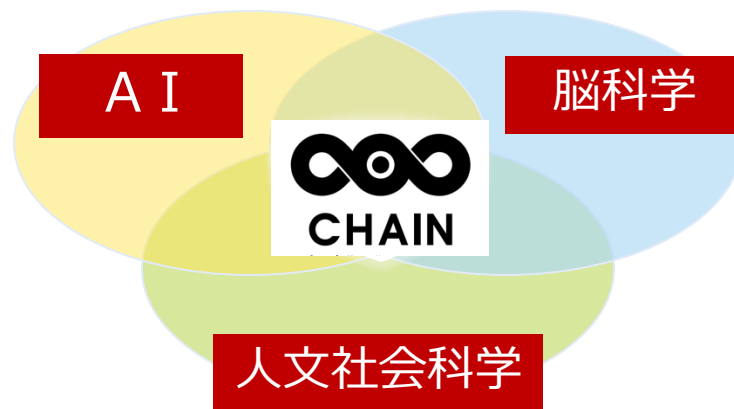
- 古来、人間を深く探究してきたのは哲学などのいわゆる人文社会科学だった。
- ところが、**近年脳科学やAIが「人間とは何か？」という問いに迫りつつある。**
- 脳イメージング技術（fMRIなど）の発達→人間の「主観性」「内面」への科学的アプローチ
- AIの爆発的な発展：テキストでは人間とほとんど変わらない反応を示したり、様々な領域で人間を凌駕する能力を示している。



脳科学・AI・人文社会科学の融合

- これからは、人間を考えるために、人文社会科学だけではもはや不十分である。
- 人間を考えるための十分な射程を得るには、脳科学・AIと人文社会科学が緊密に融合した「新しい科学」が必要とわれわれは考えている。
- これを集中的に研究・教育しているのがCHAINという組織である。

CHAINの研究・教育



AIがチューリングテストに合格？

- つい先日（3/31）に出たプレプリントでは、「AIがチューリングテストに合格した」と主張されている。
- 相手が人間かAIかわからない状況で、テキストベースでやりとりする。相手がAIが人間かを判定。
- その結果、質問者はGPT-4.5-PERSONAを73%の確率で人間と誤認したという。これは人間の回答者を正しく当ててる確率よりも有意に高かった。

<https://arxiv.org/abs/2405.08007>

People cannot distinguish GPT-4 from a human in a Turing test

Cameron R. Jones
Department of Cognitive Science
UC San Diego
San Diego, CA 92119
cameron@ucsd.edu

Benjamin K. Bergen
Department of Cognitive Science
UC San Diego
San Diego, CA 92119
bkbergen@ucsd.edu

Abstract

We evaluated 3 systems (ELIZA, GPT-3.5 and GPT-4) in a randomized, controlled, and preregistered Turing test. Human participants had a 5 minute conversation with either a human or an AI, and judged whether or not they thought their interlocutor was human. GPT-4 was judged to be a human 54% of the time, outperforming ELIZA (22%) but lagging behind actual humans (67%). The results provide the first robust empirical demonstration that any artificial system passes an interactive 2-player Turing test. The results have implications for debates around machine intelligence and, more urgently, suggest that deception by current AI systems may go undetected. Analysis of participants' strategies and reasoning suggests that stylistic and socio-emotional factors play a larger role in passing the Turing test than traditional notions of intelligence.

1 Introduction

1.1 The Turing test

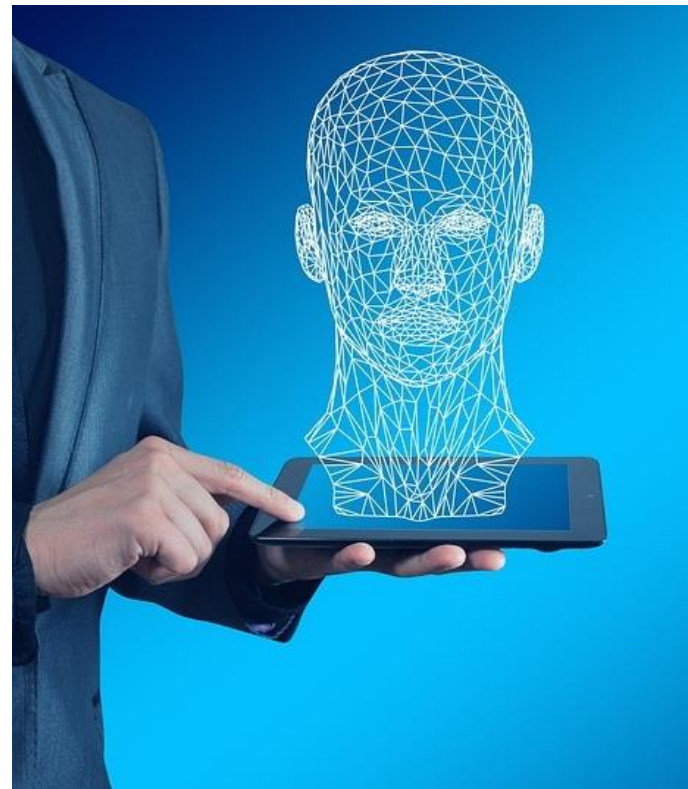
Progress in artificial intelligence has led to systems that behave in strikingly humanlike ways. Large Language Models like GPT-4 [OpenAI, 2023] not only produce fluent, naturalistic text, but also perform at parity with humans on a range of language-based tasks [Chang and Bergen, 2024]. These systems are increasingly being deployed to interact with people on the internet, from providing assistance as customer service agents [Soni, 2023] to spreading misinformation on social media [Zellers et al., 2019, Park et al., 2023]. As a result, people interacting anonymously online are increasingly forced to ask themselves the question: “Am I speaking to a human or a machine right now?”

Unwittingly, these people are engaging in a real-world analogue of a thought experiment dreamed up three quarters of a century ago by the computer scientist and mathematician Alan Turing. In his seminal article, Turing [1950] proposed a test to measure whether a machine could generate behaviour that was indistinguishable from a human. In his original formulation—which he referred to as the imitation game—a human interrogator would speak to two witnesses (one human and one machine) via a text-only interface. If the interrogator was not able to reliably distinguish between the human and the machine, the machine would be said to have passed [French, 2000].

Turing’s article “has unquestionably generated more commentary and controversy than any other article in the field of artificial intelligence” [French, 2000] (p. 116). Turing originally envisioned the test as a measure of machine intelligence; if a machine could imitate human behaviour on the gamut of topics available in natural language—from logic to love—on what grounds could we argue that the human is intelligent but the machine is not? However, this idea has accrued a raft of objections in the intervening years, for instance that the test is too easy [Marcus et al., 2016, Gunderson, 1964], or too

AIが社会に浸透する未来社会

- これからわれわれは、相手が人間かAIかわからない世界で生きていくことになる。
- それはどんな世界になるのだろうか？
- CHAINでは、AIのような「**人工的な主体**」が社会に浸透した未来において、「主体」や「人格」の概念、それらにまつわる**倫理、法、社会制度、習慣**などがどのように変わっていくか、また変わるべきなのかを考える研究を早くから行っている。



AIに好奇心をもたせる

- AIに「内発的動機づけ」 (intrinsic motivation) や「好奇心」 (curiosity) を実装する。

⇒ AIの性能向上のためのすでに確立された手法の一つ

- 最近では「退屈できるエージェント」の研究なども。

⇒ 退屈できる方が優秀！

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. (2018). Exploration by random network distillation. ICLR 2019.

Yu, Y., Chang, A.Y.C., Kanai, R. (2019). Boredom-Driven Curious Learning by Homeo-Heterostatic Value Gradients. Frontiers in Neurorobotics.



AIを通して、人間が見えてくる

- 「退屈」「飽きてしまう」.....
- 人間のネガティブな性質とされているが、実は重要な役割が.....？
- AIを通して、人間が見えてくる。
 - ↓
 - 相互解明が
加速
 - ↑
- 人間の知を通して、AIが進歩する。



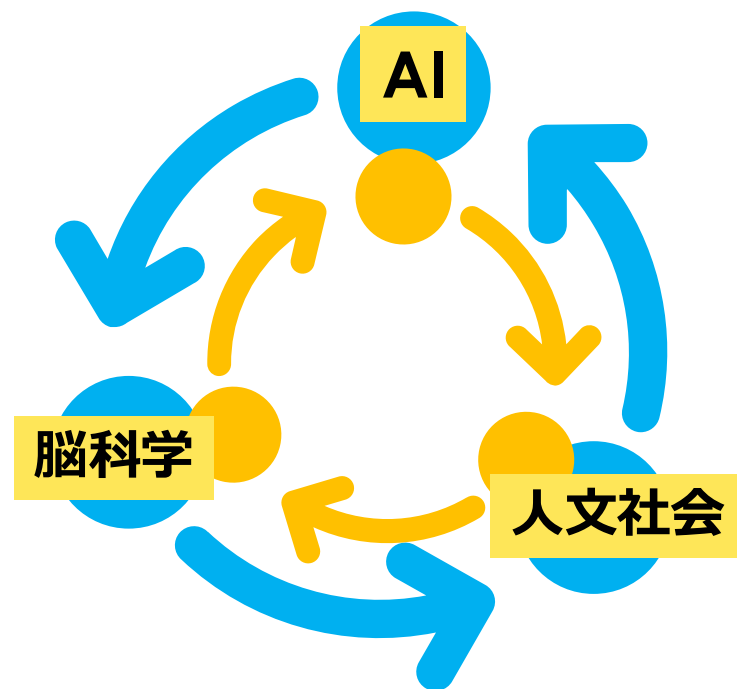
AIと人間の相互解明

- AIはどんどん「人間らしく」なっているが、人間の汎用的な能力にはまだ距離がある。
- ここから逆に、「なぜ人間はあれもこれもできるのか？」という風に、人間らしさへの問いが新たな形をとっていく。**AIを鏡として、人間の自己理解が進む。**
- また逆に、こうした問いを哲学的・心理学的・脳科学的に考えていくことで、**新たなAIが生み出せるかもしれない。**
 - 例：人間的能力と言われているものが、思った以上に「身体」に結びついており、この面でAIにはまだまだ発展の余地があることがわかってきた。

このように、人文社会科学、脳科学、AIの間で、それらが一つに交わるきわめて刺激的な知の領域が生まれている。

AI・脳科学・人文社会科学の融合

- AIや脳科学が高度に発展する社会において、その「意義」を考えようとするれば、哲学や倫理学などの人文社会科学は欠かせない。
- また、人文社会科学と融合することによって、AIや脳科学に新しい発展の方向（新しい使い方）が開かれ、技術革新につながる可能性もある。
- そして、AIや脳科学が突きつけている新しい現実に触れることで、人間について考える人文社会科学も、大きく変貌しようとしているのである。



相互に不可分に絡まり合っている

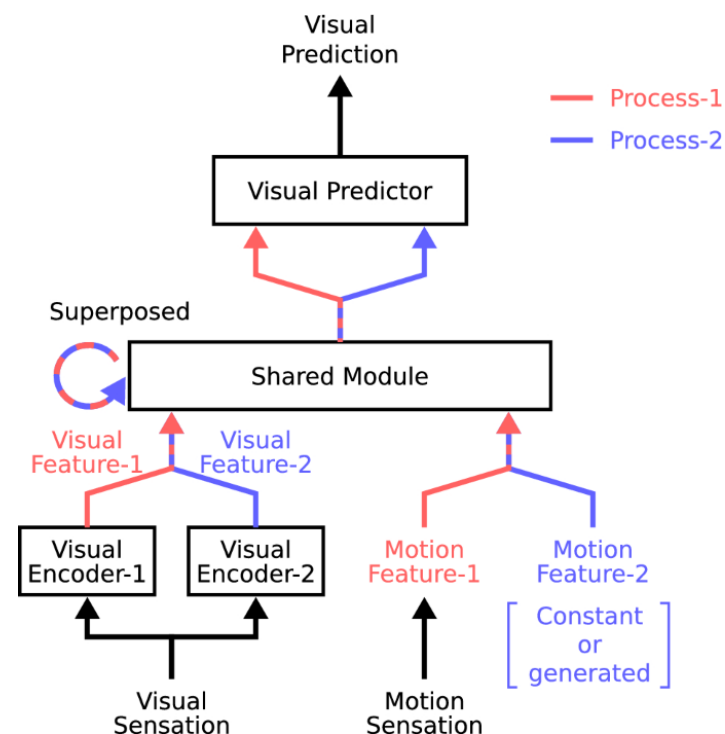
CHAINの学際的研究の例（哲学×AI）

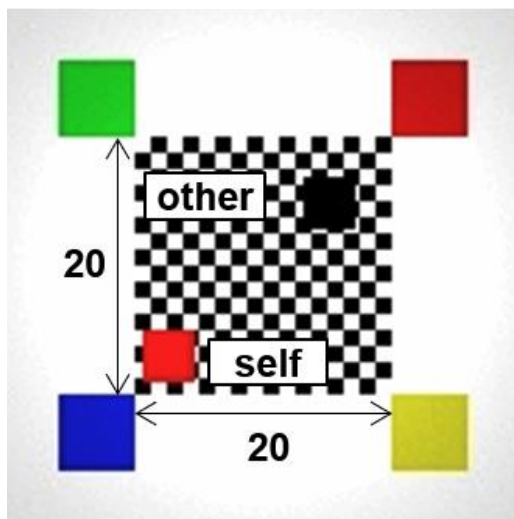
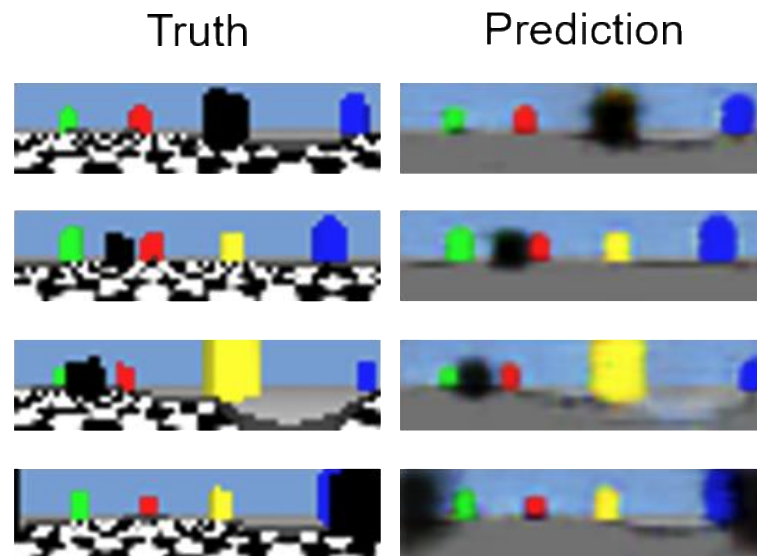
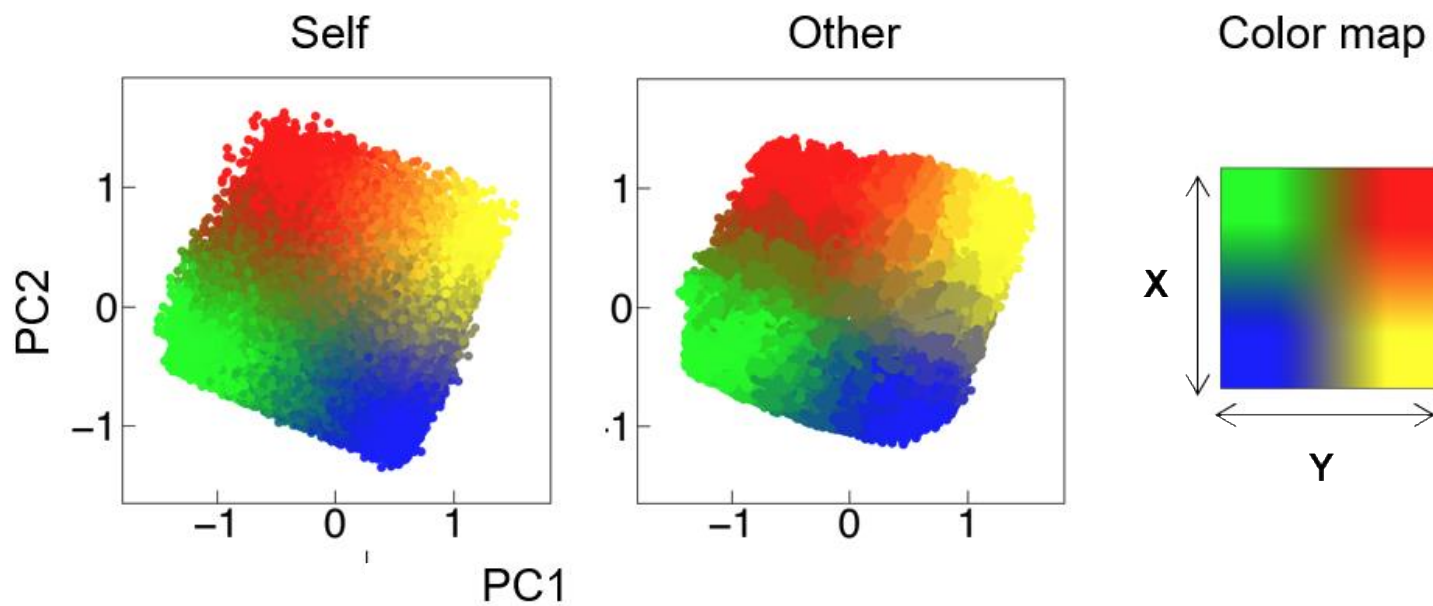
- 現象学分野から：経験の特性の分析
- 他人がレモンを食べているのを見るだけで、唾液が出てくる。
- サッカーの試合を見ていて、思わず足が動く。
- 自分と他人の身体を（思わず）重ね合わせる。
- 他人の身体を客観的に眺めるより、このような「自他の重ね合わせ」の方がより基本的では？
- 幼児はこのような「自他の重ね合わせ」から出発して「自分」と「他人」を理解してゆくのでは？



CHAINの学際的研究の例（哲学×AI）

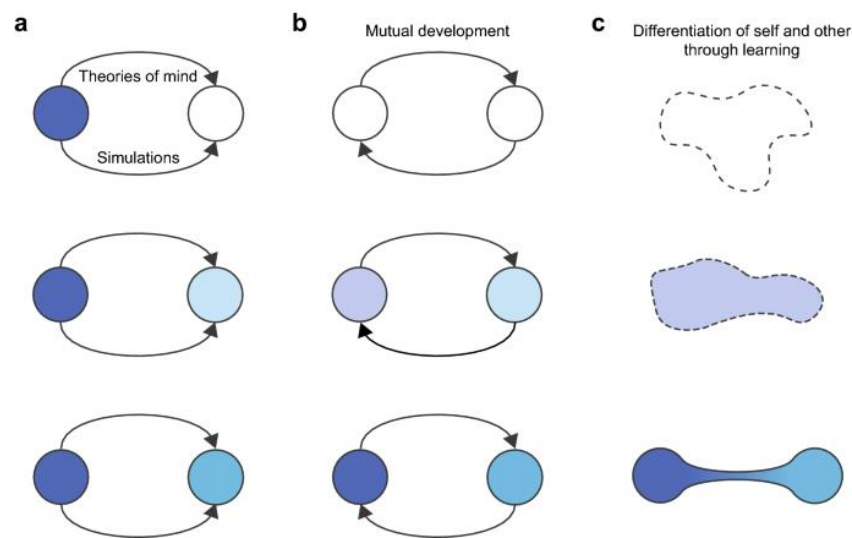
- 以上のような哲学（現象学）分野における田口の仮説を、野口渉(CHAIN元ポスドク)・飯塚博幸(CHAINコアメンバー)らがAIによるシミュレーションで検証
- 仮想環境中で視覚的予測学習を行うAIエージェント
- 同じ入力を二つに分けて重ね合わせる「重ね合わせシステム」を実装。
- これにより、エージェントは自分視点からの経験を積み重ねるだけで、他者視点からの視覚的表象を予測できるようになった。
- Noguchi, W., Iizuka, H., Yamamoto, M., & Taguchi, S. (2022). Superposition mechanism as a neural basis for understanding others. *Sci Rep* **12**, 2859.



a**b****c**

異分野融合により「人間」に迫る

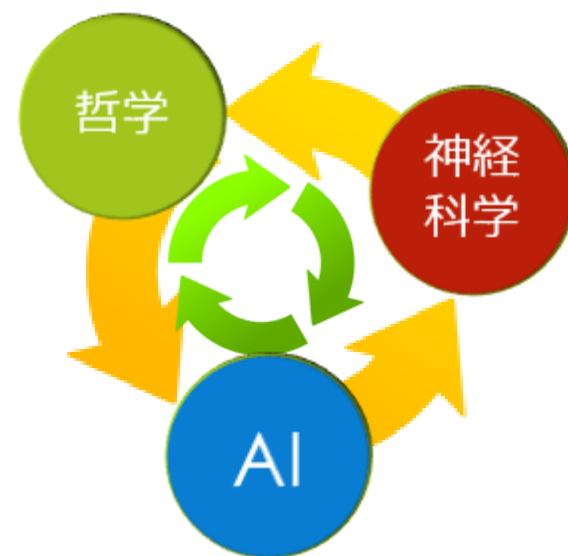
- この研究は、哲学だけでも、AIだけでもできなかった。
- 高度に哲学的な仮説と、近年登場した機械学習手法を組み合わせ、おそらくは脳のなかに実現されているかもしれないメカニズムのあり方を提案している。
- これにより、「私」と「他者」という人間の最も重要な捉え方に光を当てていくことが可能になる。



Noguchi et al. (2022). *Sci Rep* **12**, 2859.

AI×脳×哲学＝人間の新たな解明

- AIと脳科学は、人間を解明する有力なツールとなっている。
- しかし、実験の前提となる「考え方」に古い哲学のバイアスがかかっていると、見えてくる結果も限定されてしまう。
- 哲学的なヴィジョンとアイデア、脳科学の知見、AIの工学的手法が組み合わさるとき、「人間」とその「社会」、そして「世界」の新しい解明が可能になる。
- またそこから、新たなAIを生み出すアイデアも生まれてくるかもしれない。



2. これまでの成果

教育関連の成果

・履修生の満足度も高く（アンケート結果参照）、**履修生間の部局を超えた学際的共同研究**が多数生まれている。参加学生の業績発表・学会賞受賞などでも成果が出ている。学生の共同研究が企業の資金を獲得。

■ 人間知・脳・AI教育プログラム（修士・博士）

先端的な**博士人材育成**プログラムを提供（5年一貫）

■ CHAIN独自開講科目（大学院共通授業科目）

修士課程学生全員に開かれた**文理融合的科目**の提供

■ 全学教育：〈人間知〉の学際的探究（CHAIN入門）

学部学生に向けた文理融合的科目の提供（R4～）

■ 人間知・脳・AIリカレントプログラム（社会人）

全学に先駆けて北大の新しいリカレント教育プログラムに**最初のパッケージ**を提供（令和5年度開講予定）

■ 北大D X博士人材フェローシップ（博士）

参画する**中核機関**の一つとして同フェローシップを支える。応募学生の加点は参画プログラム中最高レベルの6点

■ 北大独自の文理融合教育を実現

■ 「人間知」を重視した
独自の高度AI人材教育を展開

■ 社会のニーズに応じて、
AI倫理、AIと脳科学など
独自性の高いリカレント教育を
提供

履修生の声

〔第一期生〕

様々なバックグラウンドの学生
とのディスカッションを通じて、
見識が広がるとともに異分野の
人とのコミュニケーション能力
が成長したように感じます。

〔第三期生〕

自分がほとんど触れたことのな
かった分野や方法論について知
る機会があるだけで、研究に対
するモチベーションが高まりま
した。今の自分の研究にも繋
がっています。



大学院教育プログラム
履修生数（合計）の推移

令和2年	24名	
令和3年	39名	
令和4年	66名	うち 文系 57 理系 49
令和5年	106名	

CHAIN教育プログラムへの満足度



知識やスキルの面で成長できた



文理融合的な知識やスキルを身につけられた



※履修生アンケート結果（令和4年9月13日時点）

取組の現時点における成果・今後見込まれる成果（研究関連）

【現時点における成果】

- 研究面では、**トップ10%論文をこれまで計4報**発表したほか、国内外の研究機関との**多数の共同研究**を進め、**学際研究の拠点**として国内外で**知名度が高まっている**（研究会・シンポジウムのこれまでの参加者総数は延べ3,623人、**Google検索「AI, neuroscience, center」検索首位、「AI, 学際センター」上位2件目**[2024年1月10日時点]）
- 多数の外部資金**を獲得、**企業連携も拡大**することができた（グラフ参照）。

【今後見込まれる成果】

- 企業連携**については、セブン銀行、日立、NTT等と新たな連携を協議中であり、**産学連携のいっそうの拡大が可能**。
- AIと神経科学・人文社会科学の融合による世界的にも類例のあまりない**個性的な学際研究**、ならびに**他研究機関および企業との連携**を通して、**北大初の独創的な研究成果を社会に還元していくことが可能**。



国際共著論文数

令和4年度

14本



多数の英語査読付き論文発表

令和4年度

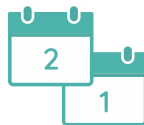
センター教員＋兼務教員

79本



Top10%論文発表

2年
連続



北大初！

北大本属教員として、
企業への初の
クロスアポイントメント契約を締結



企業連携受入件数・総額

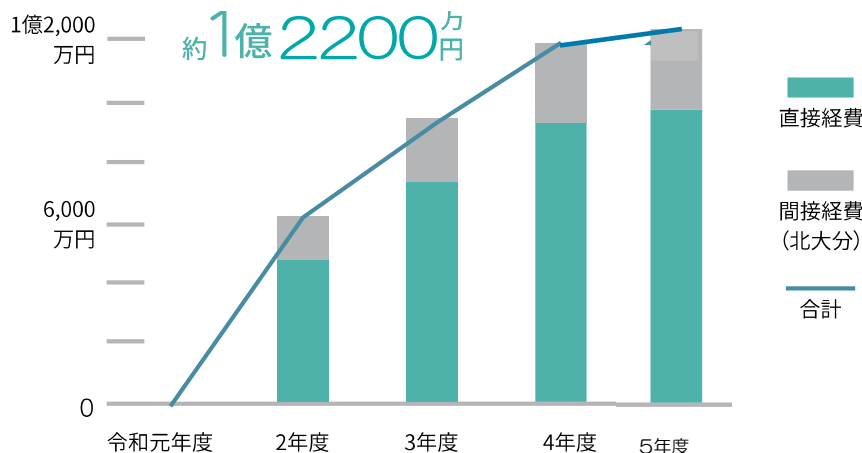
令和5年度上半期

4件

約7560万円

外部資金獲得額 令和5年10月時点

主要メンバー**6名分**の合計



アラヤ、クロスコンパス、クリプトン・
フューチャー・メディアなど

マツダ, SONY, など

コニカミノルタ, オムロンなど

富士通,
東京エレクトロン
など

連携企業の
拡大

さらに
継続・拡大

鈴木啓介G

仮想現実と
人間知



仮想現実（VR）や拡張現実（AR）技術を用いた認知心理学実験。**人間の現実感、身体感覚、自己感覚**を研究

宮原克典G

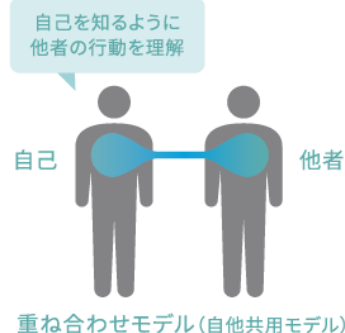
人工主体の
倫理



AIのような**人工的な主体**が社会の隅々に浸透する未来における**倫理や法・社会制度**のあり方を研究。

飯塚・田口G

自他の
重ね合わせ
メカニズム



「自他の重ね合わせ」ができるAIエージェントの研究（哲学×AI）。
「**人間の心がわかるAI**」につながっていく可能性がある研究。

吉田・宮園G

サリエンスと
アフォーダンス



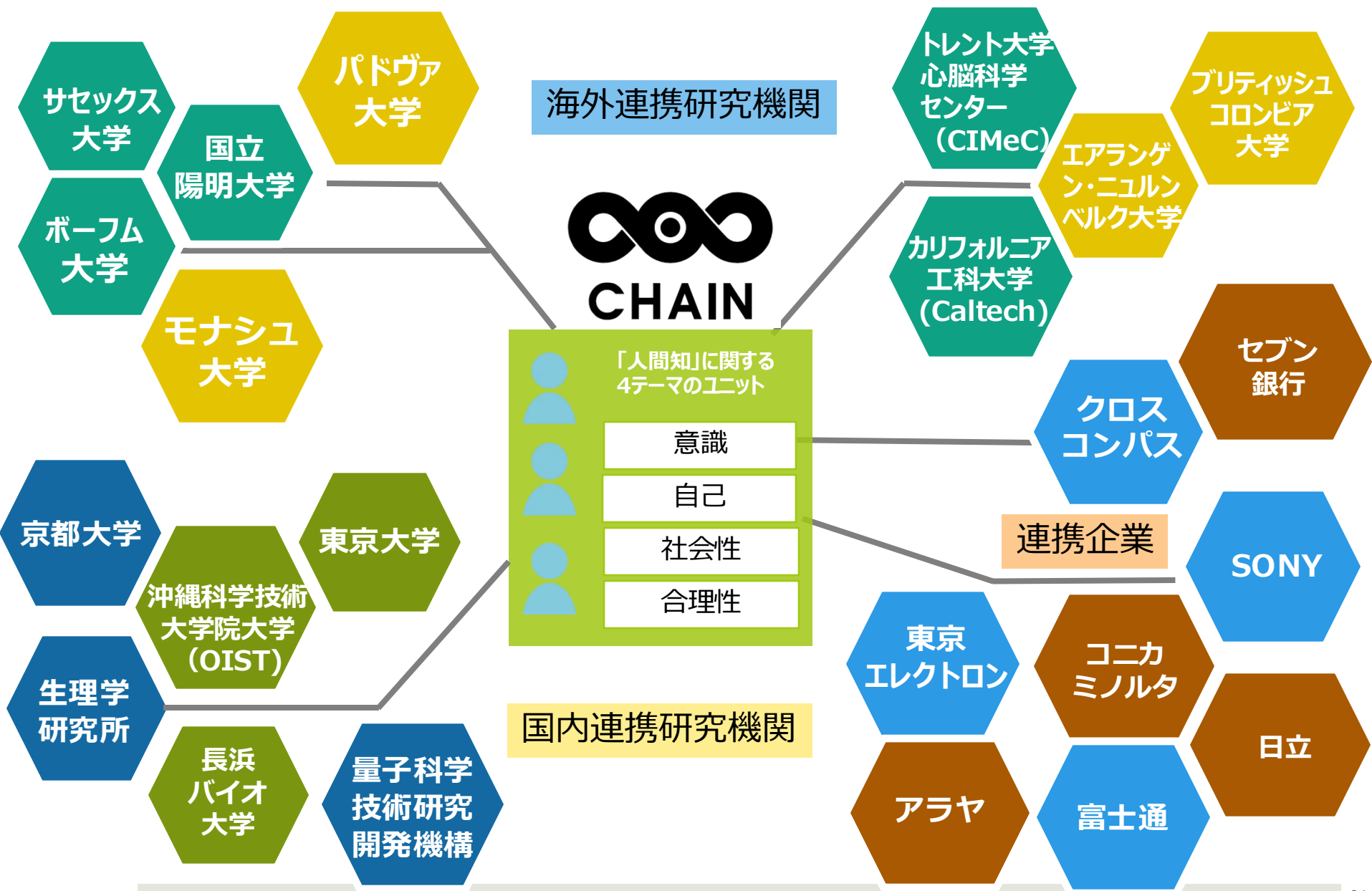
神経科学×哲学×精神医学。心に対して何かが「**目立つ**」という現象を新たな仕方で学際的に解釈。**精神疾患の原因解明**につながりうる研究。

多様な学際的事件を開催

- **国際シンポジウム**「〈意識の科学〉の冒険—哲学・脳科学・AI・ロボット研究のクロスオーバー」 (2019)
- **International workshop:** The free energy principle of the brain: experiments and verification (2021), Moral and legal questions about artificial agents (2025)
- **国際学会 ALIFE 2023**
- **国際学会 Aware & Alive** (ASSC Satellite)
- **CHAINセミナー**：これまで50回開催
- 国内外から延べ67名の登壇者
- 哲学・法学・統計学・生物学・認知科学・神経科学・機械学習・ロボティクス・人工生命など、多様な分野



CHAINの国内・海外連携機関 (交渉中のものを含む)



CHAINから広がる世界

- このような外部連携の展開は、CHAIN修了生の進路も切り開いている。
- CHAINで教育を受ければ、アカデミアへの就職にも強い（文理融合的人材は引っぱりだこだ）し、一般企業も、AIや脳科学と、哲学などの人文社会科学の両方に鼻が利く人材は喉から手が出るほど欲しい。
- みなさんも、いままさに力強く動き始めている学際的・文理融合的研究と一緒に参画してみてほしい。実に刺激的な世界が広がっている。