2025/3/10　言語処理学会第31回大会（NLP2025）
チュートリアル

人工知能の哲学入門

鈴木貴之

（東京大学大学院総合文化研究科）

tkykszk@g.ecc.u-tokyo.ac.jp

最終版のスライドは下記URLからダウンロードできます。

https://tkykszk.net/nlp2025/

自己紹介

- 東京大学大学院総合文化研究科（科学史・科学哲学研究室）の教員です。
- 専門は心の哲学（とくに意識のハードプロブレム）です。

- そのほかの研究テーマ（心の哲学の関連領域）：
  - 心理学の哲学・認知科学の哲学
  - 精神医学の哲学
  - **人工知能の哲学**

- そのほかの研究テーマ（メタ哲学）：
  - メタ哲学（哲学についての哲学研究）
  - 実験哲学（質問紙調査などを用いた哲学研究）

人工知能研究と哲学

- 第2次人工知能ブーム期まで、人工知能研究者と哲学者のあいだでは活発な論争が行われていた。
- 真の人工知能は実現不可能だと主張する哲学者も多かった。
- 1990年代以降、議論は下火に。
- 人工知能研究が飛躍的な進展を遂げた現在、再検討が必要なのでは？

JST/RISTEXの研究開発プロジェクト（2018年度-2021年度）

ウェブサイト：https://updatingphilosophyofai.net

尾形哲也先生インタビュー（その1）

検索...

尾形哲也先生は、早稲田大学基幹理工学部教授で、ロボット技術と人工知能、とくに深層学習を融合させた構成論的アプローチに基づいて、学習や人間機械協調などの研究をされています。このインタビューでは、そのようなユニークな研究に至った経緯や、現在の深層学習の課題などについてお話をうかがいました。

RISTEX

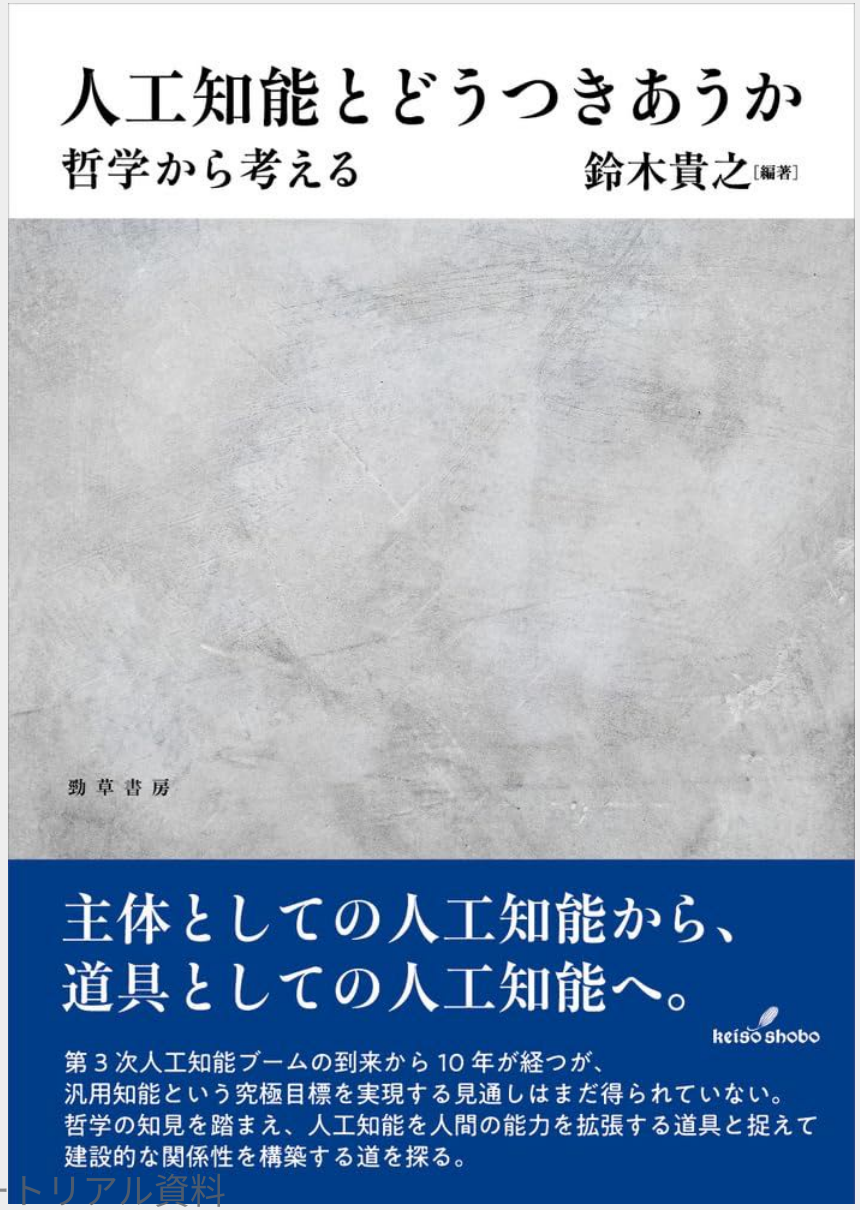
HITE Human Information Technology Ecosystem

ロボット研究と人工知能研究

【一一】　今日は尾形哲也先生にお話を伺いたいと思います。尾形先生は、深層学習とロボット工学の中間の、ユニークな位置で人工知能の研究をされています。まずは、人工知能あるいはロボットに関するいままでの研究の経歴を伺いたいと思います。

【尾形】　はい。もともと、加藤一郎先生という、世界で最初の人型ロボットWABOT-1を1973年に開発された「日本のロボットの父」と呼ばれた先生ですが、その加藤先生の研究室に入りたいというモチベーションで1989年に早稲田に入りました。その加藤先生の研究室に配属されるのと同じぐらいの時期に甘利俊一先生の『神経回路網の数理』を読んで、加藤先生に神経回路とロボットをやりたいですとお願いしたのです。

当時、「学習」でロボットを動かすという手法はなかったわけではないのですが、人工神経回路モデルを用いるのは非常に珍しい発想でした。しかしそのお願いをした直後に、加藤先生から、それなら君は「ロボットの心」をやれ、と言われたのです。単なる知能ロボットを考えていたので、「心」という言葉に驚いたのを覚えています。私の修士の１年の時の最初の学会発表（ロボット学会）のタイトルは「ロボットにおける心の発生」です。


人工知能とどうつきあうか
哲学から考える
鈴木貴之[編著]
勁草書房
主体としての人工知能から、
道具としての人工知能へ。
keisoshobo
第３次人工知能ブームの到来から10年が経つが、
汎用知能という究極目標を実現する見通しはまだ得られていない。
哲学の知見を踏まえ、人工知能を人間の能力を拡張する道具と捉えて
建設的な関係性を構築する道を探る。


人工知能の哲学入門
鈴木貴之
勁草書房
人工知能の可能性と限界をめぐる
哲学的議論をアップデート！
第２次人工知能ブーム期まで盛んに行われた人工知能をめぐる
哲学的批判は、すでに乗り越えられたのだろうか？
第３次人工知能ブームの現在の活況を捉え直し、
今こそ求められる理論的検討の道を探る。
keiso shobo

1　古典的人工知能：基本的発想と問題

2　現在の人工知能：基本的発想と成果

3　人工知能の哲学：主な問い

4　人工知能の哲学：言語をめぐる問い

古典的人工知能：基本的発想

- 知能の本質は計算（形式的な規則にしたがった記号操作）。
  - 知的過程はアルゴリズムとして表現可能。コンピュータはアルゴリズムを実行可能。それゆえ、コンピュータは知能をもちうる。

- 例：手書き数字認識
  - 手書き数字をデジタルデータに変換し、データにさまざまな規則を適用して、書かれた数字を特定する。
  - 数字認識のための規則：
    - データが曲線を含まなければ、数字は1か4か7である。
    - データが交点を含むならば、数字は4か8である。

- 古典的人工知能研究の特徴：
  - アルゴリズムで用いられる規則や、規則で用いられる特徴（直線、交点など）を、エンジニアがすべて考える必要がある。
  - それゆえ、メカニズムの透明性は高い。

古典的人工知能：問題

- **問題1**：現実世界の課題は複雑。そこで必要な知識や規則を特定することは困難。
  - 例：イヌやネコの画像認識、チェスや将棋における手の選択（写真：Adobe Fireflyによる生成画像）

- **問題2**：現実世界の課題の多くは明確な境界をもたない (open-ended) 。そこで必要な知識や規則をすべて特定することは困難。
  - 例：雑談、コンビニ店員の仕事（写真：Adobe Fireflyによる生成画像）
  - 常識の明示化

- **問題3**：必要な知識や規則を特定できたとしても、ある状況で必要な知識や規則を素早く特定することは困難。
  - 課題が複雑になると必要な知識や規則の数が増加し、計算量が指数関数的に増大する（組み合わせ爆発）。
  - フレーム問題 (cf. Dennett, 1984)

- **問題4**：多くの規則には例外がある。どうしたらすべての例外的状況に対処できるか？
  - 「例外的な状況を除けば、赤信号では停車すべき」

- **古典的人工知能研究の本質的な困難**：単純な課題には有効な手法が、現実世界の複雑な課題には利用できない。

ヒューバート・ドレイファス『コンピュータには何ができないか』

「第一部でわたしが示す一般理論は、人工知能分野があるお決まりのパタンをもっているということである。すなわち、初期のめざましい成功と、それに続く突然の思いもよらない困難というパタンである。」（Dreyfus, 1992, p. 85; 邦訳 p. 153）

「知能は理解を要求し、理解はコンピュータに常識という背景を与えることを要求するが、その常識を成人した人間がもっているのは、彼が身体をもち、技能を通じて物質世界と相互作用し、ある文化へと教育されるからだ」（Dreyfus, 1992, p. 3; 邦訳 p. 5）

第2次人工知能ブーム期までの人工知能の哲学

- 古典的人工知能研究の問題点を正しく指摘していた。
- しかし、なぜ人間は同様の問題に直面しないのかを具体的に説明することはできなかった。

1　古典的人工知能：基本的発想と問題

2　現在の人工知能：基本的発想と成果

3　人工知能の哲学：主な問い

4　人工知能の哲学：言語をめぐる問い

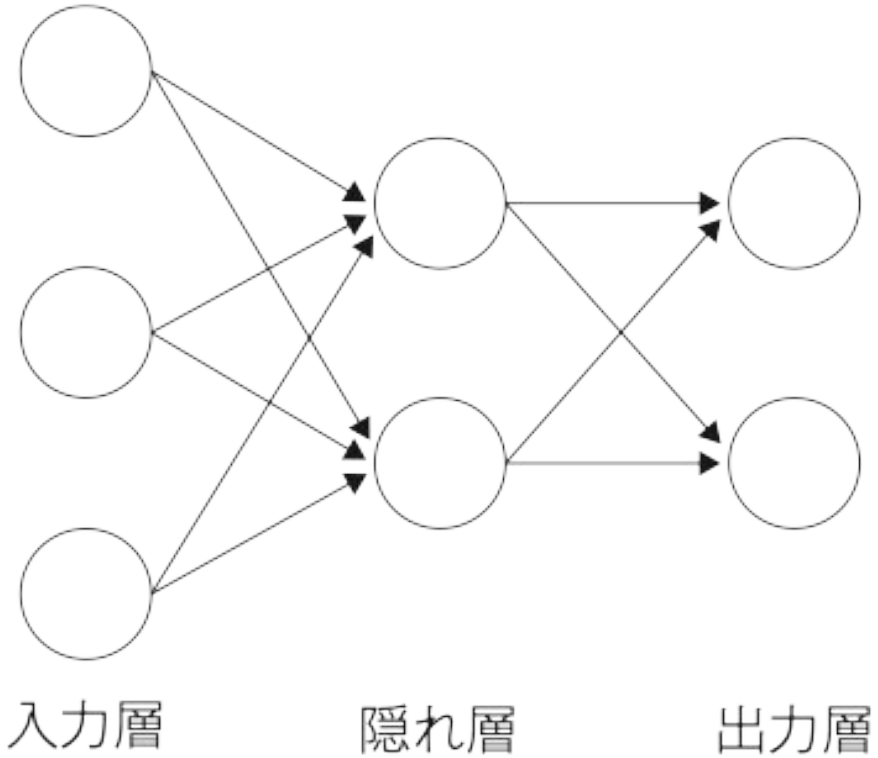
現在の人工知能：基本的発想

- データからの学習（機械学習）
- ニューラルネットワークの利用
  - 深層ニューラルネットワークを用いた機械学習＝深層学習

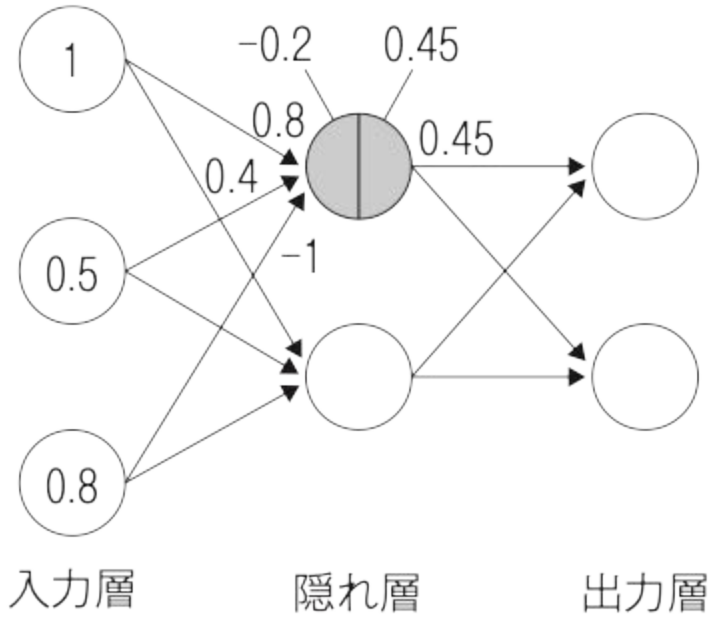
ニューラルネットワーク

- ユニットと、ユニット同士の結合からなる。
- 各ユニットは実数値を入出力とする。
- 上流のユニットからの入力に基づいてユニットの出力値が決定される。
- ネットワーク全体は、入力ベクトルを出力ベクトルに変換する。

ニューラルネットワークの基本的な構造



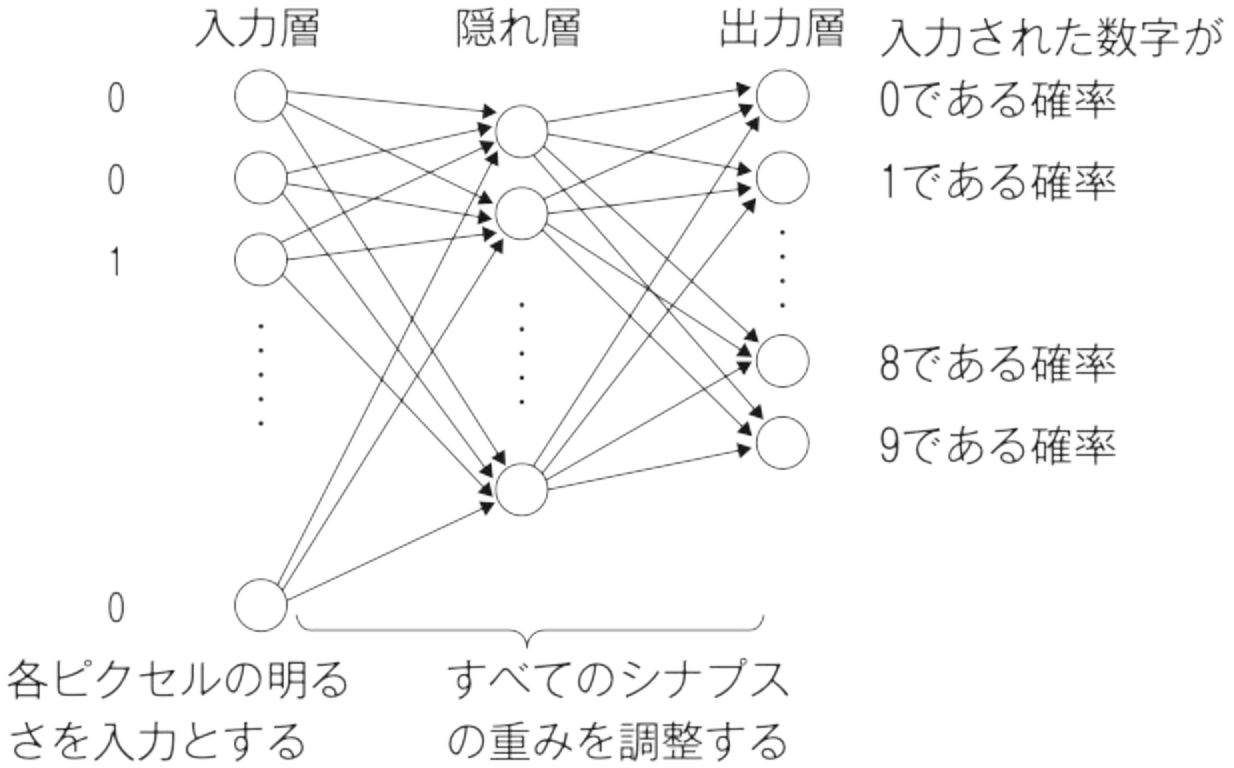
（図：鈴木, 2024, p. 80）



灰色のユニットへの入力は、上流のユニットの出力に対応する重みを掛けたものの総和に閾値（ここでは-0.4）を加えたもの（1 x 0.8 + 0.4 x 0.5 + (-1) x 0.8 + (-0.4)= -0.2）。この値を活性化関数に入力した出力値（0.45）が、このユニットの出力となる。（図：鈴木, 2024, p. 81）

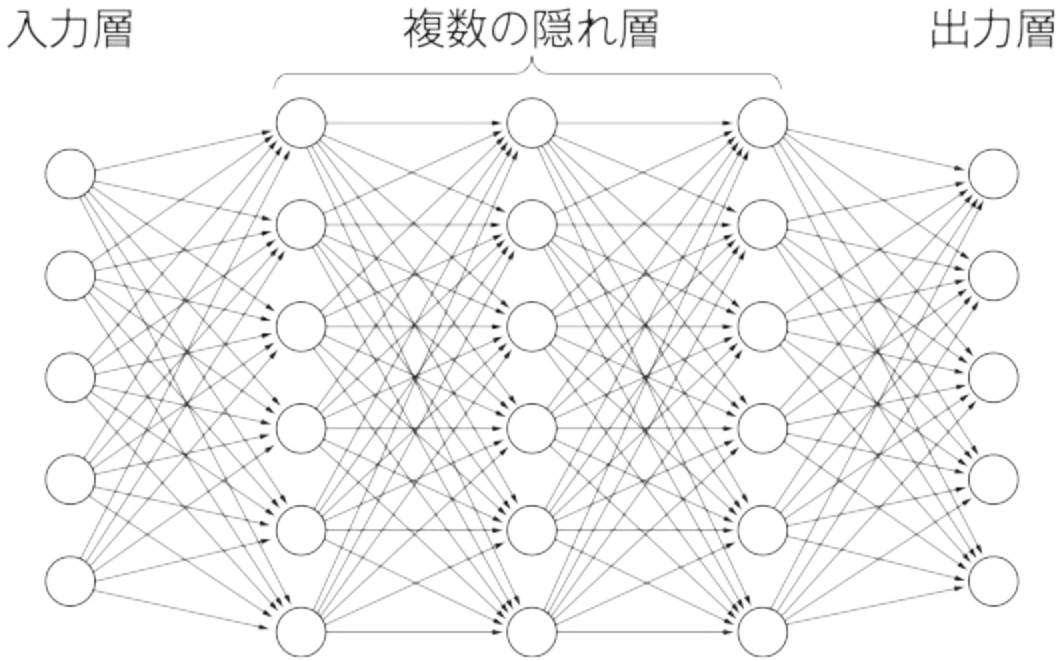
- 例：手書き数字認識
  - 入力は、ピクセルごとの明るさを数値化したベクトル。
  - 出力は、入力が0から9の数字である確率。
  - 訓練データを用いて、できるだけ正解に近い出力が得られるように重みの値を調整する。

手書き数字認識ネットワークの基本的な構造



（図：鈴木, 2024, p. 86）

深層ニューラルネットワーク



（図：鈴木, 2024, p. 97）

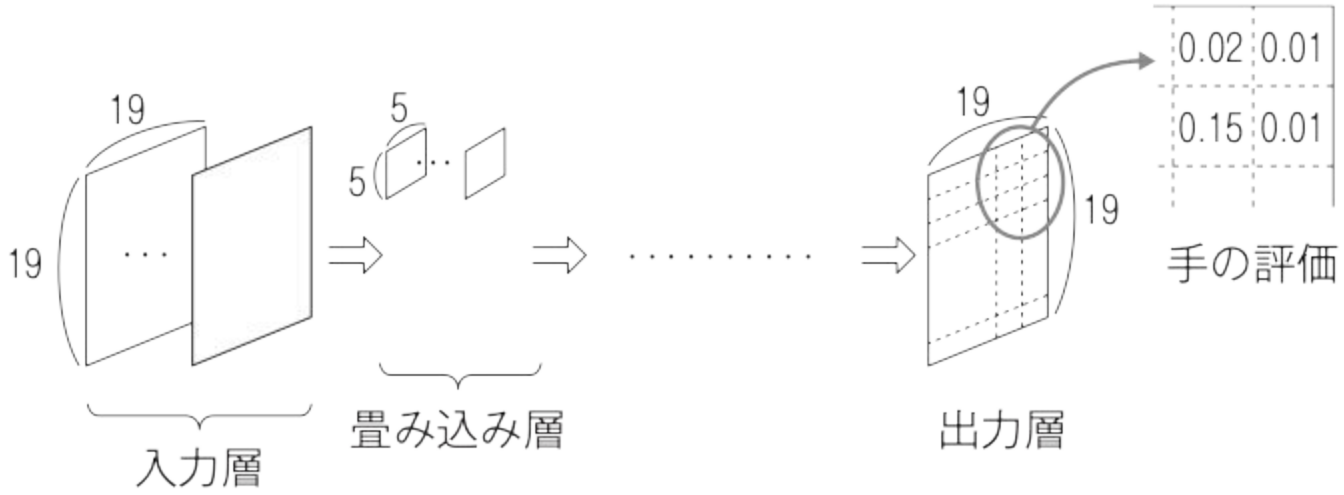
現在の人工知能：特徴

- **特徴1**：入力と出力の関係を、ニューラルネットワークが訓練データからみずから学習する。
  - 中間でどのような情報処理が行われるかをエンジニアが考える必要はない（end-to-endの学習）。
  - 特徴量設計が不要。

- **特徴2**：大規模な深層ニューラルネットワークは複雑な関数を表現している。
  - イヌの画像とネコの画像の識別のような複雑な課題も実行可能。

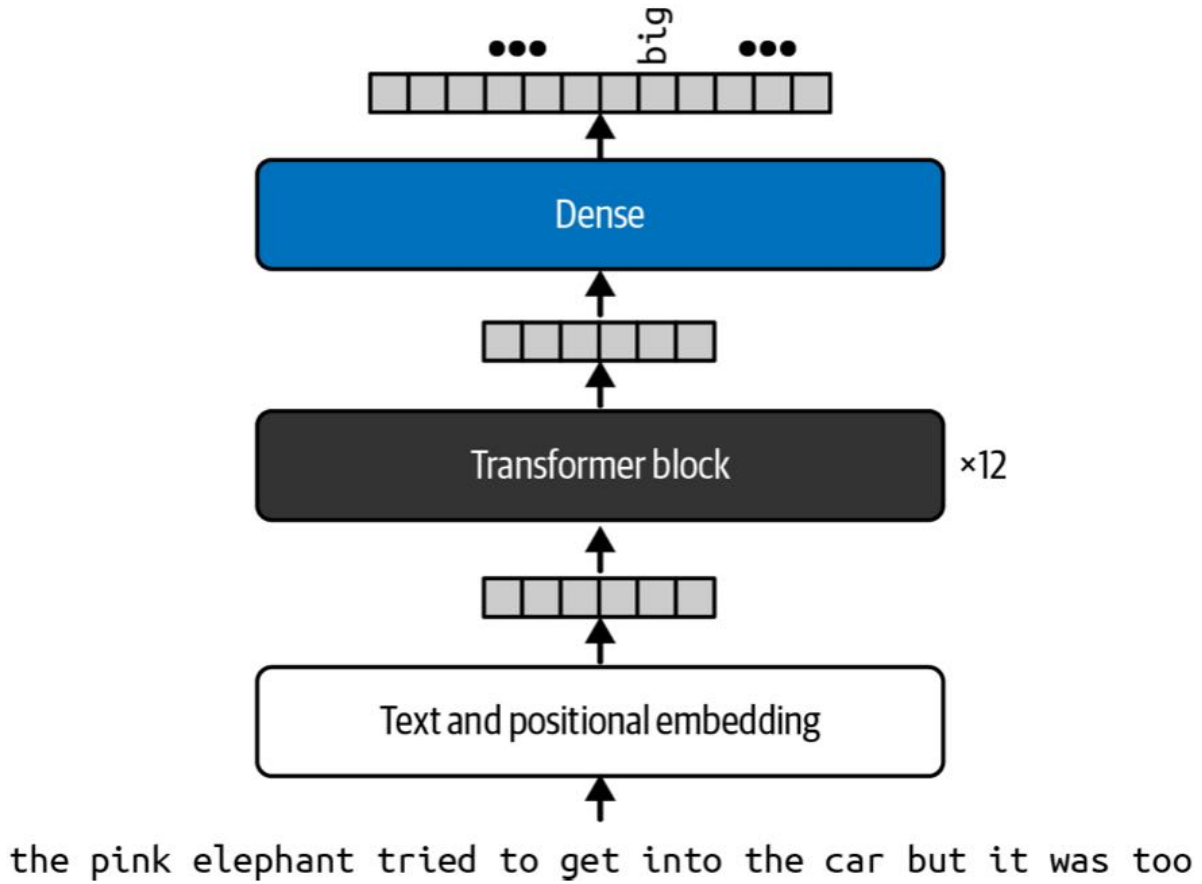
- **特徴3**：入出力の設定を工夫すれば、この手法はさまざまな課題に利用可能。
  - 囲碁：現在の盤面を入力、次の手の評価を出力とする。
  - 文章生成：単語をベクトル化したものを入力、入力に続く単語の確率を出力とする。

囲碁AIによる手の選択



（図：鈴木, 2024, p. 157）

大規模言語モデルによる文章の生成



（図：Foster, 2023, p. 251）

1　古典的人工知能：基本的発想と問題

2　現在の人工知能：基本的発想と成果

3　人工知能の哲学：主な問い

4　人工知能の哲学：言語をめぐる問い

現在の人工知能をめぐる理論的な問い

- 人工知能そのものに関して：
  - 現在の人工知能の限界とは？
  - 深層ニューラルネットワークは何をしているのか？
- 認知科学との関連で：
  - 深層学習は人間の認知の原理か？

問い①：現在の人工知能の限界とは？

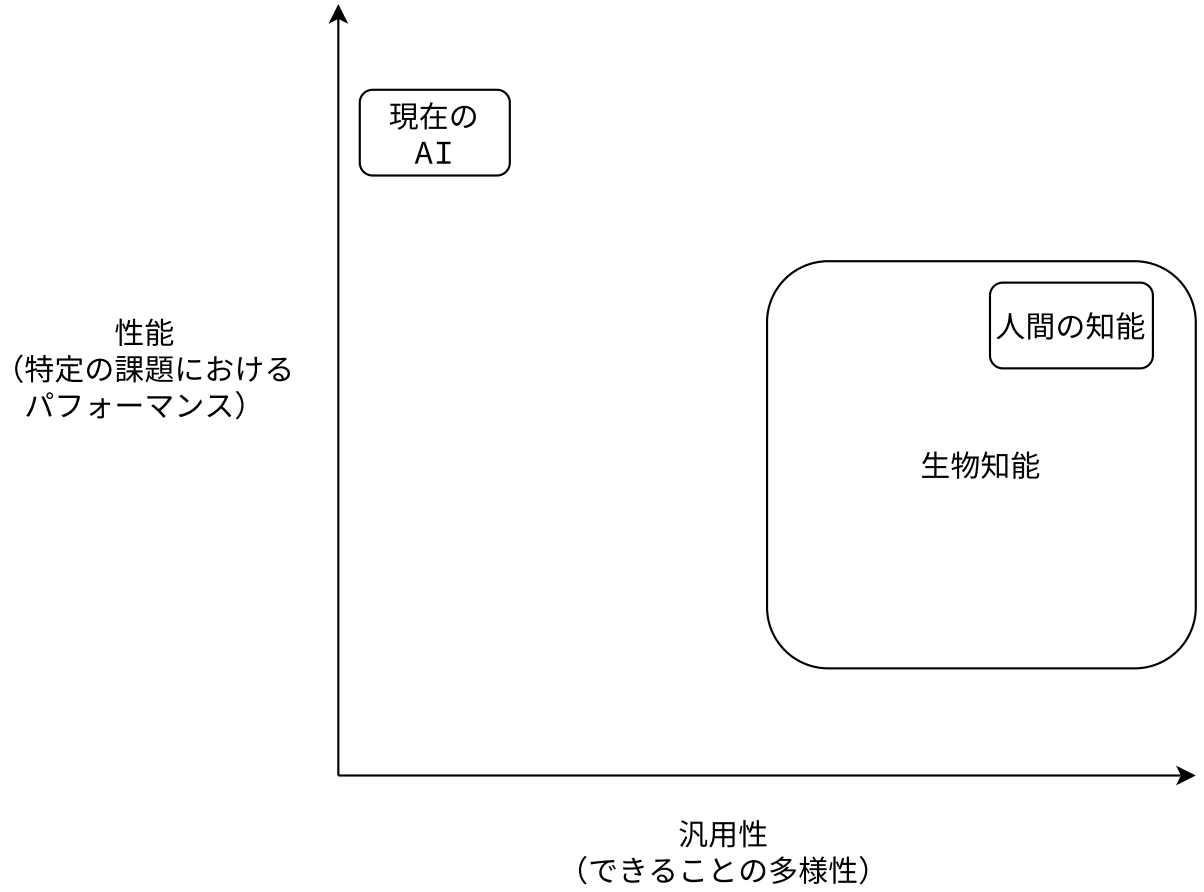
- 問題の明確化：
  - 人工知能一般でできないこと／特定の手法でできないこと
  - 原理的にできないこと／相対的に難しいこと

  →生産的な問いは、特定の手法（たとえば深層学習）では実現困難なことは何か、というもの。

- 課題①：大量の訓練データが存在しない課題への対応
  - 対比：画像診断／政策決定
  - 画像認識やゲームは特殊な課題かもしれない。
  - 記号計算的な手法との融合が必要？ (Marcus, 2020)
    - 人間の介入を排したことこそが成功の秘訣？ (Sutton, 2019)

- 課題②：汎用人工知能の実現
  - 現在の人工知能の多くは課題特化型。
  - 汎用かつ高性能な人工知能は実現可能か？
    - 問題①：汎用性と性能のトレードオフ
    - 問題②：汎用学習メカニズムの必要性
    - 問題③：生物知能との違い (cf. Brooks, 1990)

汎用性と性能の関係



（図：鈴木, 2024, p. 187）

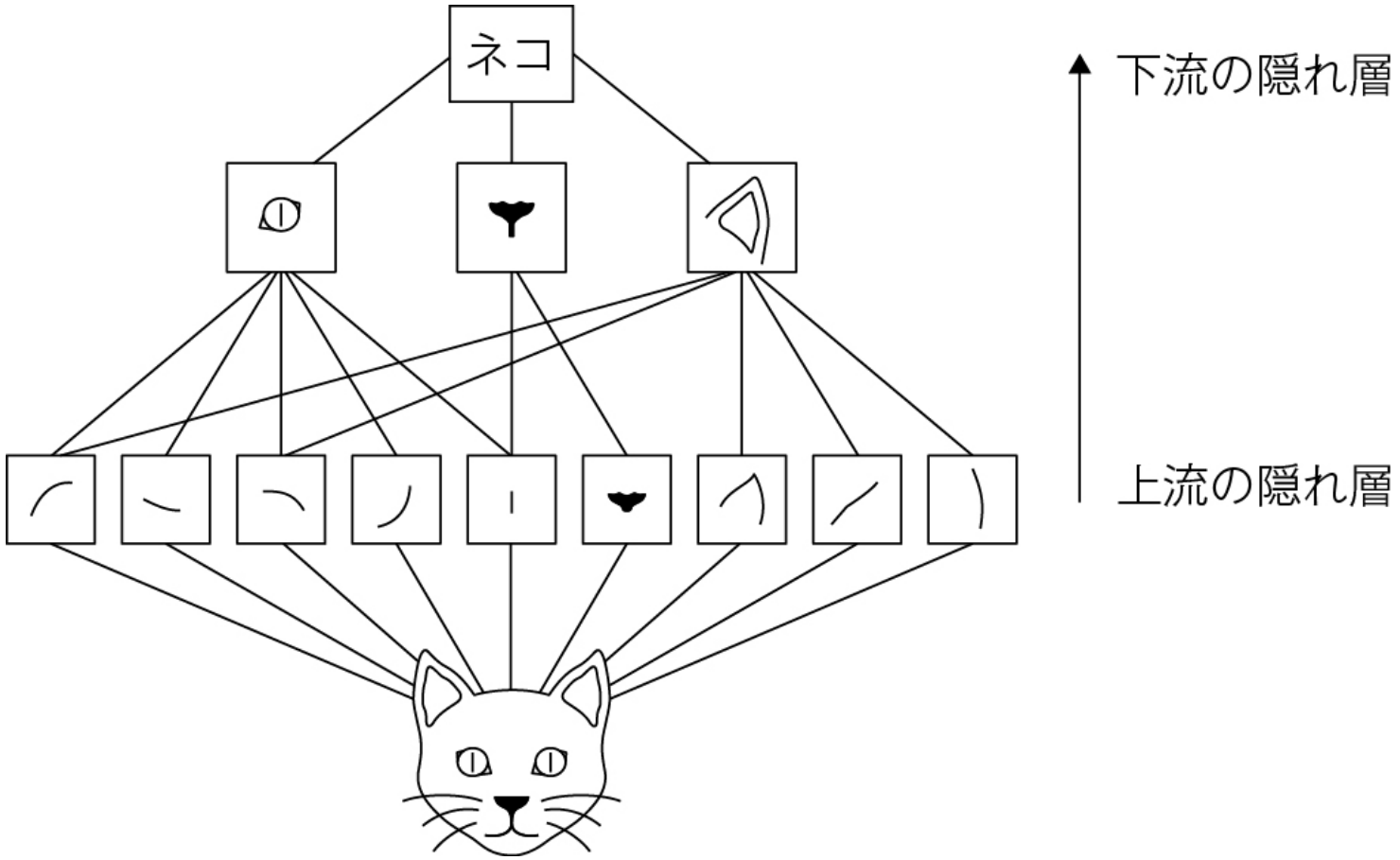
問い②：深層ニューラルネットワークは何をしているのか？

- 問題の明確化：
  - 深層ニューラルネットワークの局所的な仕組みは明らか。
  - 全体として情報処理をしていること、ベクトルの変換をしていることも明らか。

  →両者の中間レベルにおける有用な特徴づけとは？

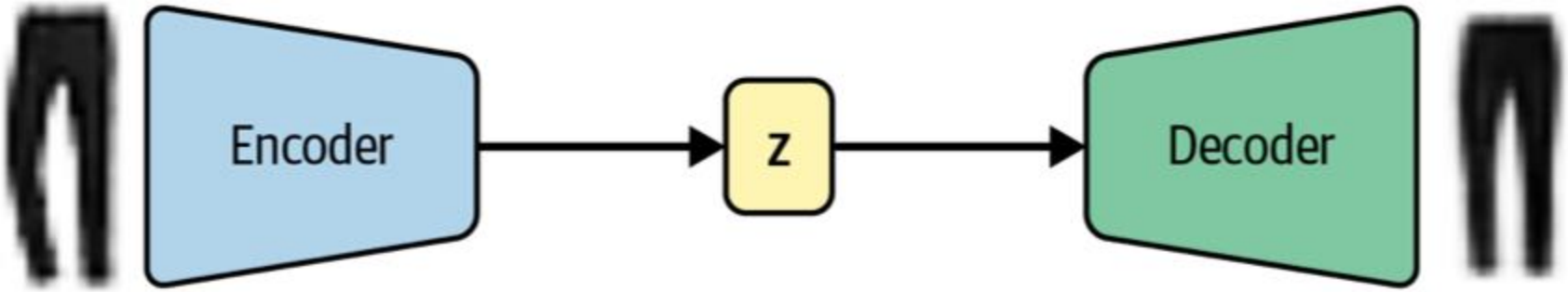
- いくつかのアイデア：
  - 局所的で単純な特徴にもとづいて、より大域的で複雑な特徴を検出している（畳み込みニューラルネットワーク）
  - 次元削減によって本質的な情報を抽出している（自己符号化器）
  - 局所的な情報に文脈情報を追加している（自己アテンション）

畳み込みニューラルネットワーク



（図：鈴木, 2024, p. 116）

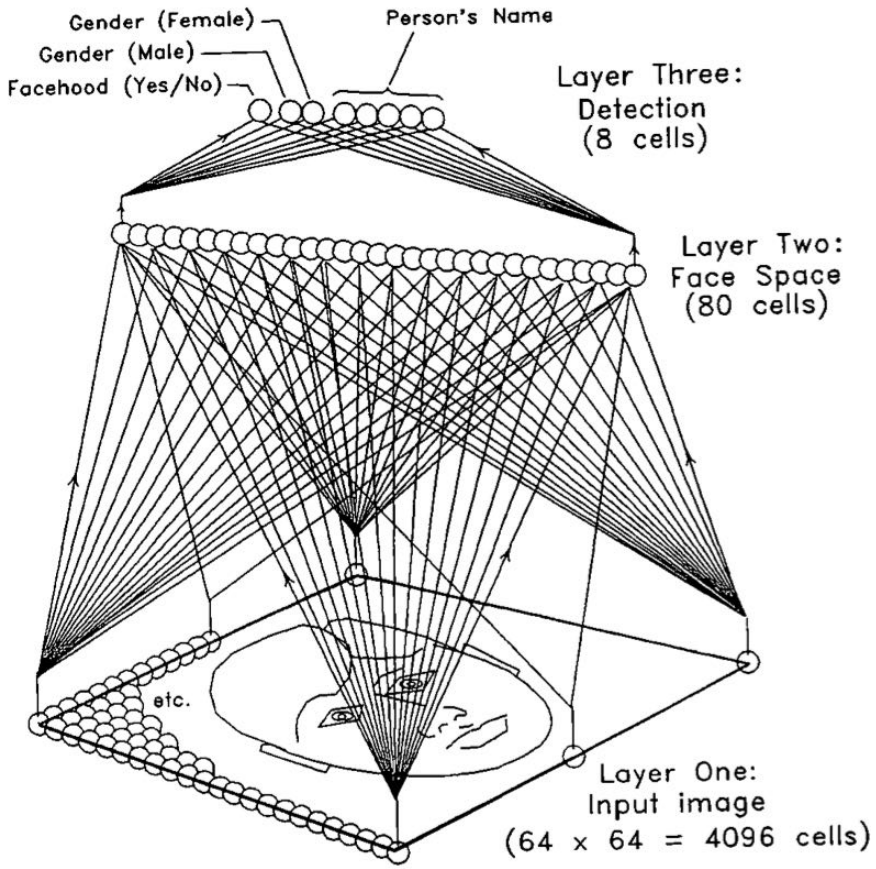
自己符号化器



（図：Foster, 2023, p. 63）

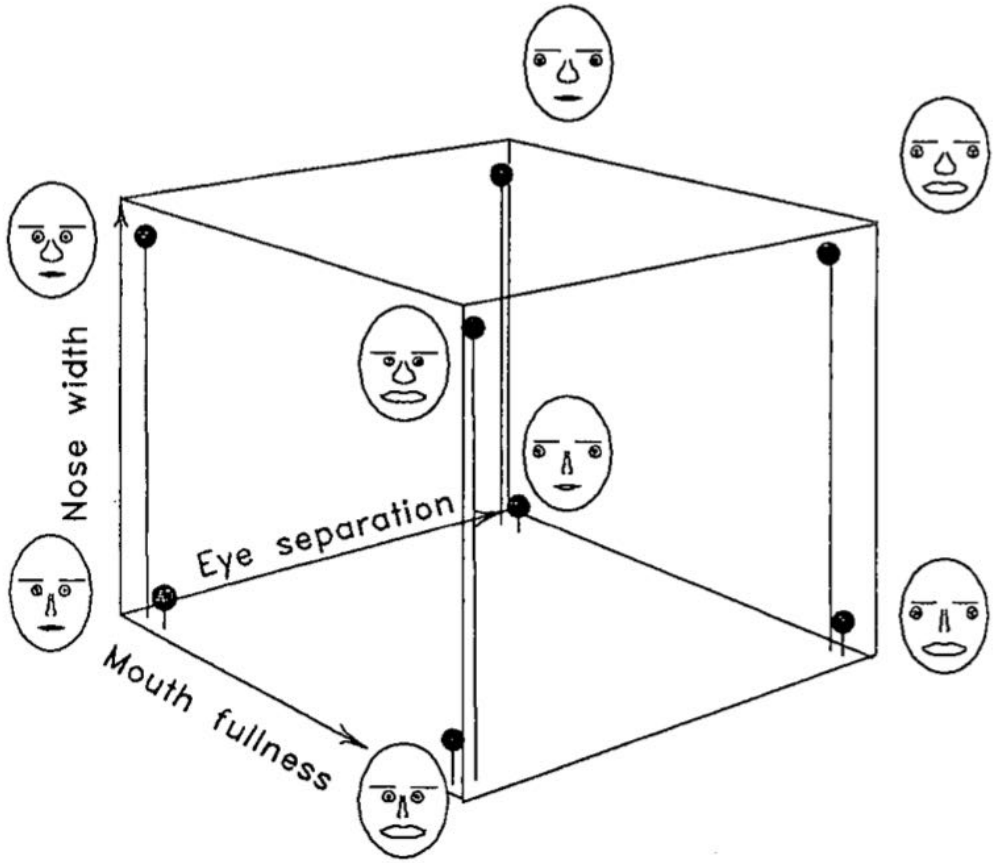
- チャーチランド (Churchland, 2012) の状態空間意味論 (state space semantics)：
  - 隠れ層の活性化パターン＝隠れ層が表現する状態空間内のある点
  - 状態空間内における点同士の類似性関係は、世界内の何らかの特徴同士の類似性関係の写像となっている。

顔認識ネットワーク



（図；Chuchland, 1995, p. 40）

顔認識ネットワークの隠れ層が表現する特徴空間



（図：Churchland, 1995, p. 28）

- これらの特徴づけには共通点もあるが、相違点もあるように思われる。

  →多様な深層ニューラルネットワークに適用可能な（かつ抽象度が高すぎて無内容になることのない）特徴づけとは？

- 考えるべき問題：
  - ユニット数＝次元数が異なるネットワークも同じ情報を表現できるのか？
  - 個々のユニット＝次元には、つねに有意味な解釈が可能か？
  - 隠れ層が表現する状態空間には、つねに何らかの解釈が可能なのか？

隠れ層のユニットが強く反応する刺激の例

（図：Churchland, 2012, p. 64）

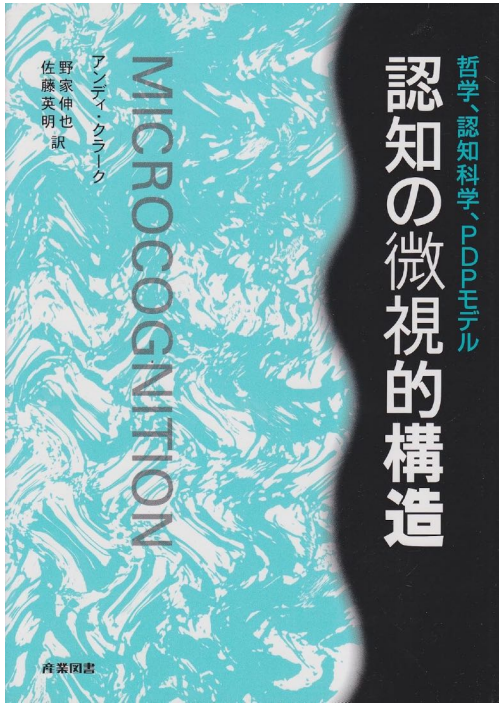
現在の人工知能：倫理的・社会的問題

- バイアス
- 透明性
- AIアラインメント（価値整合問題）
  - いずれも現在の人工知能の仕組みに由来する問題であるという点が重要。

問い③：深層学習は人間の認知の原理か？

- 人工知能研究と認知科学：
  - 人間の認知の原理がわかれば、それを利用して人工知能が実現できるはず。
  - ある原理で人間と同様の人工知能が実現できるならば、人間の認知の原理もそれと同じである可能性が高い。

- 認知科学における論争 (cf. Clark, 1989; 2014)：
  - 計算主義：人間の認知の基本原理は、形式的な規則に従った記号操作。
  - コネクショニズム：人間の認知の基本原理は、ニューロンの活性化パターンの変換。

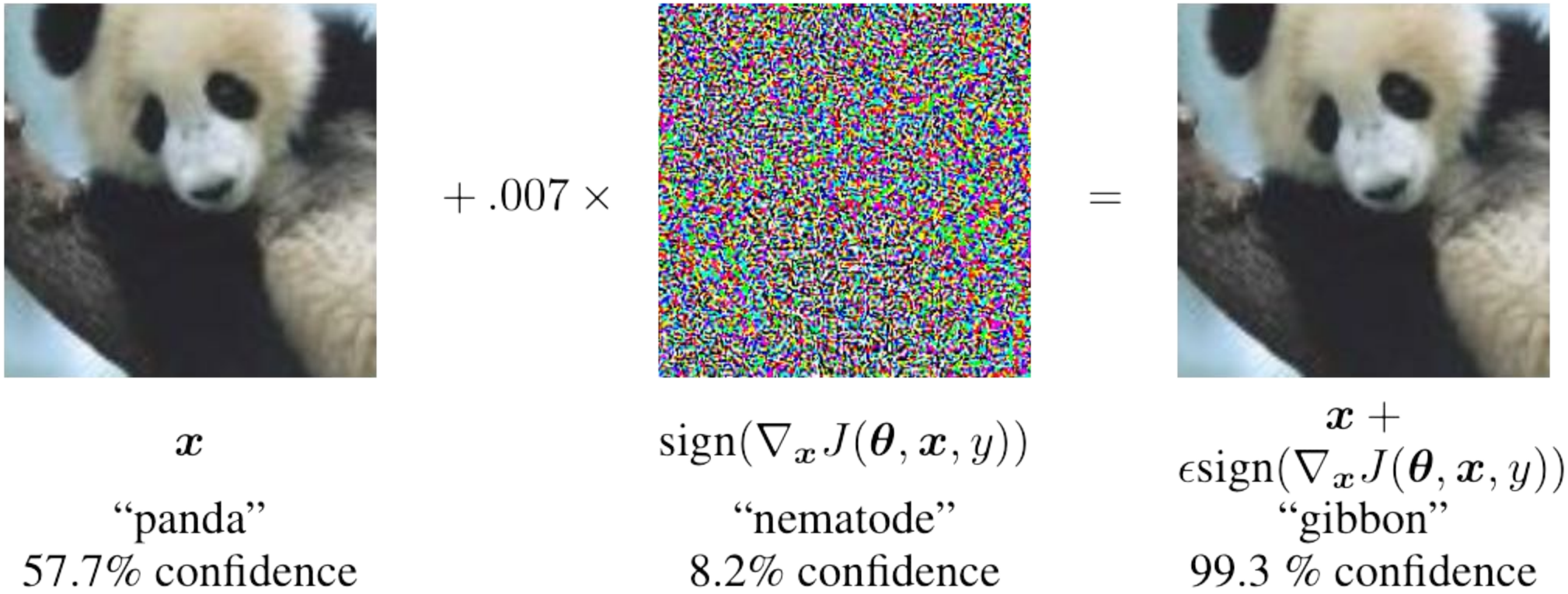


- コネクショニズムの論拠：
  - 生物学的妥当性
  - 認知の重要な特徴の説明：
    - 汎化能力
    - 曖昧な入力への耐性
    - ダメージに対する耐性

- 計算主義者からの批判：
  - 思考の体系性・生産性が説明できない。
  - コネクショニズムは実装レベルの記述にすぎない。

- 現在の人工知能は、コネクショニズムが正しいことを示しているのか？
  - 一方で：深層ニューラルネットワークと人間の脳の相違点
    - ネットワーク構造の違い
    - 学習メカニズムの違い
    - 学習に必要なサンプル数
    - 誤りの種類

敵対的事例の例



（図：Goodfellow et al., 2015, p. 3）

- 現在の人工知能は、コネクショニズムが正しいことを示しているのか？（続き）
  - 他方で：深層学習からの示唆
    - 高次元の情報を低次元に圧縮するという自己符号化器の原理は、脳の基本原理でもあるかもしれない。
    - 大規模言語モデルが表現している事象同士の確率的関係は、脳における知識の基本形式かもしれない。

1　古典的人工知能：基本的発想と問題

2　現在の人工知能：基本的発想と成果

3　人工知能の哲学：主な問い

4　人工知能の哲学：言語をめぐる問い

問い①’：大規模言語モデルの限界とは？

- 形式的な規則に従った言語使用？
- ハルシネーションの克服？
- 十分な訓練データが存在しない内容に関する文章生成？

→これらは人間にとっても実行が困難な課題なのでは？

問い②’：大規模言語モデルは何をしているのか？

- トランスフォーマの出力層が構成する状態空間は、何を表現しているのか？
- この状態空間において、語と文や文章はどのような関係にあるか？

- LLMは語と語の確率的関係以上の情報を表現しているのか？ (cf. Jawahar et al., 2019)
  - 語と語の確率的関係だけでほとんどの言語処理課題は実行可能？
  - 計算主義とコネクショニズムの論争との関連性

問い③’：大規模言語モデルの原理は人間の言語使用の原理か？

- LLMの言語使用に対する3つの見方：
  - LLMは意味理解を欠いており、それゆえ十全な言語使用は不可能。
  - LLMと人間の言語使用のメカニズムは同じ。
  - LLMと人間の言語使用のメカニズムは異なるが、LLMは意味理解なしに十全な言語使用が可能。

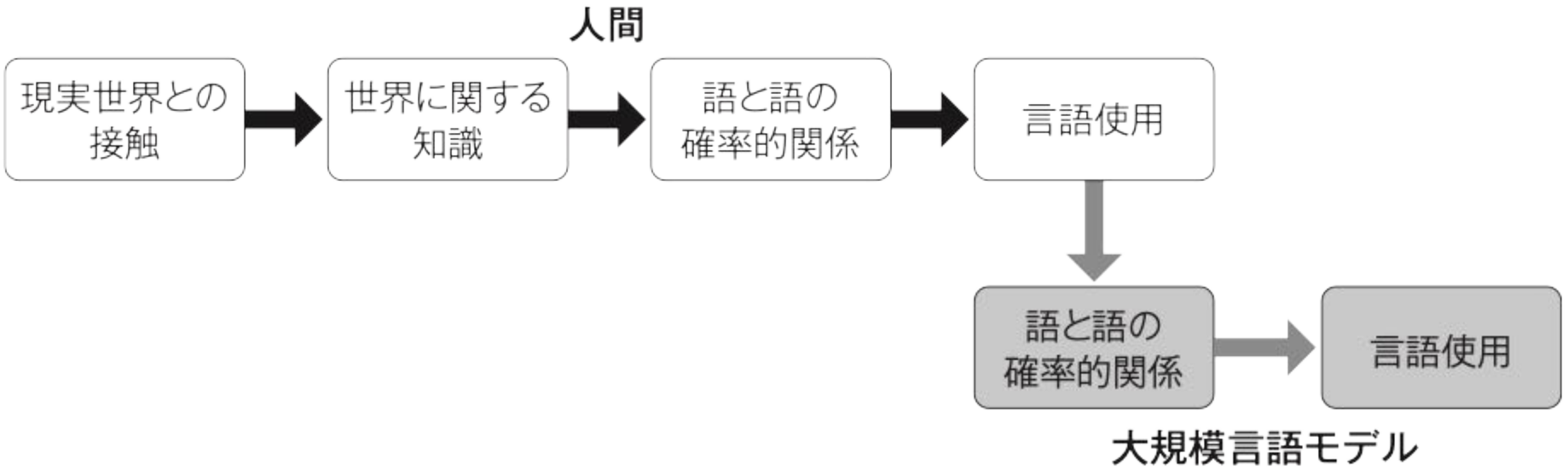
問い③’：大規模言語モデルの原理は人間の言語使用の原理か？

- LLMの言語使用に対する3つの見方：
  - **LLMは意味理解を欠いており、それゆえ十全な言語使用は不可能。**
  - LLMと人間の言語使用のメカニズムは同じ。
  - LLMと人間の言語使用のメカニズムは異なるが、LLMは意味理解なしに十全な言語使用が可能。

- 人間との相違点：
  - アテンションや自己回帰のメカニズムはLLMに特有では？
  - LLMは人間の言語実践に寄生的。
  - LLMは語と現実世界のつながり（記号接地）を欠く (cf. Bender & Koller, 2020) 。

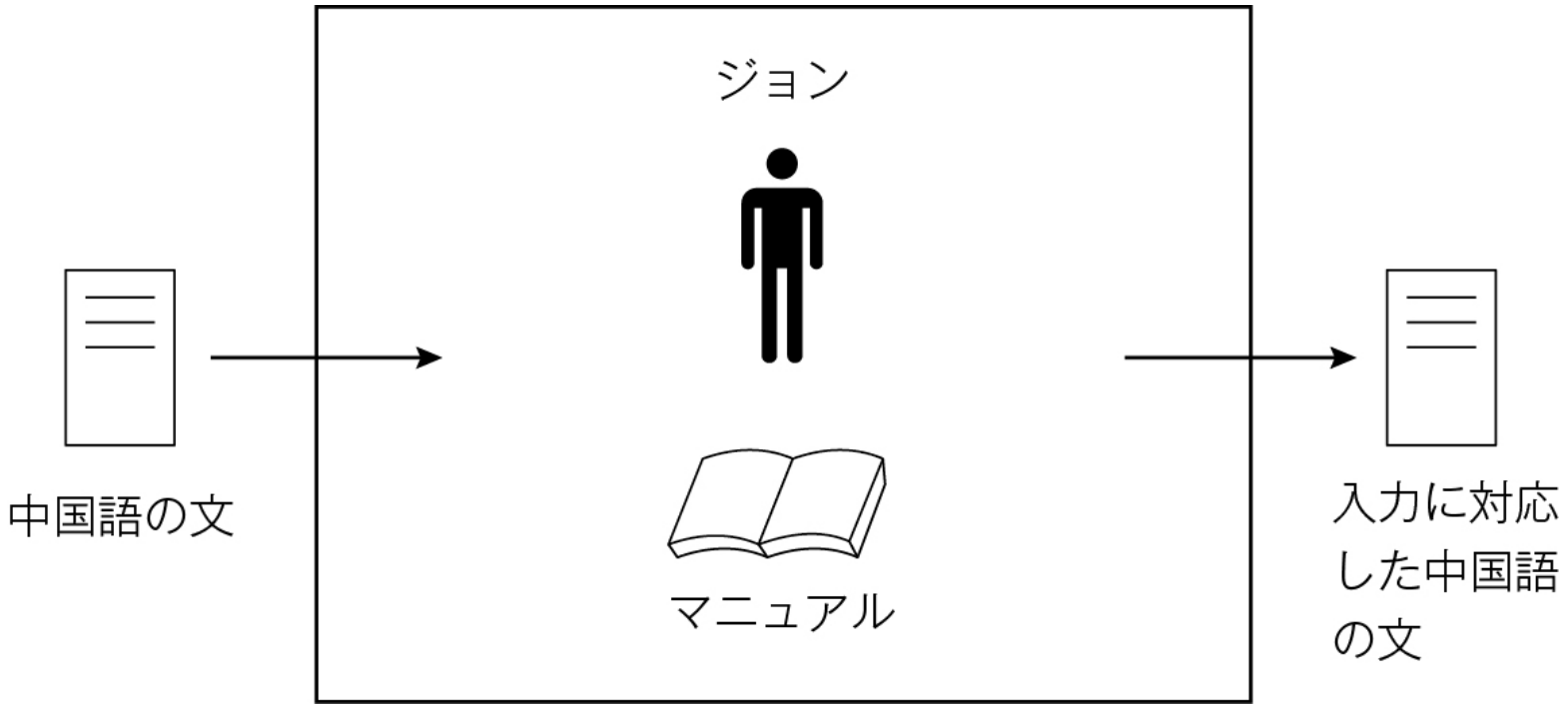
- しかし…
  - 記号接地は語と語の確率的関係を知るために必要なだけかもしれない。
  - LLMをロボットと組み合わせれば記号接地は可能かもしれない。
    - このようなロボットも意味理解を欠いているのか？

人間と大規模言語モデルの関係



（図：鈴木, 2024, p. 171）

中国語の部屋の思考実験



（図：鈴木, 2024, p. 44）

問い③’：大規模言語モデルの原理は人間の言語使用の原理か？

- LLMの言語使用に対する3つの見方：
  - LLMは意味理解を欠いており、それゆえ十全な言語使用は不可能。
  - **LLMと人間の言語使用のメカニズムは同じ。**
  - LLMと人間の言語使用のメカニズムは異なるが、LLMは意味理解なしに十全な言語使用が可能。

- LLMと人間の言語使用が本質的に類似しているとしたら、それは何を意味しているのか？
  - 従来の言語哲学・言語学の言語観には根本的な修正が必要？
    - 言葉の意味とは何か？意味理解とは何か？
    - 言語使用に意味理解は必要か？

- LLMと人間の言語使用が本質的に類似しているとしたら、それは何を意味しているのか？
  - 従来の言語哲学・言語学の言語観には根本的な修正が必要？
    - **言葉の意味とは何か？意味理解とは何か？**
    - 言語使用に意味理解は必要か？

- 人工知能研究と言語哲学・言語学：
  - 古典的AI研究は、現在主流の言語哲学・言語学の考え方（真理条件意味論、生成文法…）に対応する。
  - LLMはどのような言語哲学・言語学の考え方と対応するのか？
    - 意味の使用説（次田, 2023）？

- 言語使用の二側面：
  - 言語内使用：言語→言語
  - 言語外使用：知覚→言語、言語→行動

  →（現在の）LLMに可能なのは言語内使用だけ。両者の関係は？

- 言葉の意味に関する2つの見方：
  - 外在主義：言葉の意味は言語と言語以外のものの関係によって決まる（真理条件意味論など）。
  - 内在主義：言葉の意味は言語内部の要因によって決まる（概念役割意味論など）。

  →LLMは、従来とは異なるタイプの内在主義の可能性を示唆している?

- LLMと人間の言語使用が本質的に類似しているとしたら、それは何を意味しているのか？
  - 従来の言語学・言語哲学の言語観には根本的な修正が必要？
    - 言葉の意味とは何か？意味理解とは何か？
    - **言語使用に意味理解は必要か？**

問い③’：大規模言語モデルの原理は人間の言語使用の原理か？

- LLMの言語使用に対する3つの見方：
  - LLMは意味理解を欠いており、それゆえ十全な言語使用は不可能。
  - LLMと人間の言語使用のメカニズムは同じ。
  - **LLMと人間の言語使用のメカニズムは異なるが、LLMは意味理解なしに十全な言語使用が可能。**

- LLMと言語に関する根本的な問い：
  - 意味理解とはいかなる現象なのか？
  - 言語使用に意味理解は必要か？
  - 言語使用の原理は、現在われわれが有している概念で説明可能か？

人工知能の哲学の意義

- **哲学にとって**：心の本質、知能の本質を理解する手がかり
- **人工知能研究にとって**：新たな課題の提示、新しい手法への手がかり

文献

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47(1), 139–159.（公刊前バージョンの邦訳：ロッドニィ・A・ブルックス「表象なしの知能」柴田正良訳『現代思想』第18巻第3号、1990年）

Churchland, P. M. (1995). *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain.* MIT Press.（ポール・M・チャーチランド『認知哲学ー脳科学から心の哲学へ』信原幸弘・宮島昭二訳、産業図書、1997年）

Churchland, P. M. (2012). *Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals.* MIT Press.

Clark, A. (1989). *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing.* MIT Press.（アンディ・クラーク『認知の微視的構造ー哲学、認知科学、PDPモデル』野家伸也・佐藤英明訳、産業図書、1997年）

Clark, A. (2014). *Mindware: An Introduction to the Philosophy of Cognitive Science (Second Edition).* Oxford University Press.

Dennett, D. C. (1984). Cognitive Wheels: The Frame Problem of AI. In C. Hookway (Ed.), *Minds, Machines and Evolution: Philosophical Studies* (pp. 129-151). Cambridge University Press.（ダニエル・デネット「コグニティヴ・ホイールー人工知能におけるフレーム問題」信原幸弘訳『現代思想』第15巻第5号、1987年）

Dreyfus, H. L. (1992). *What Computers Still Can’t Do: A Critique of Artificial Reason*. MIT Press.（1979年版の邦訳：ヒューバート・L・ドレイファス『コンピュータには何ができないか ー哲学的人工知能批判』黒崎政男・村若修訳、産業図書、1992年）

Foster, D. (2023) *Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play (Second Edition).* O’Reilly.（デイヴィッド・フォスター『生成Deep Learning ー絵を描き、物語や音楽を作り、ゲームをプレイする（第2版）』オライリー・ジャパン、2024年）

Jawahar, G., Sagot, B., & Seddah, D. (2019). What Does BERT Learn about the Structure of Language? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3651-3657). Association for Computational Linguistics.

Marcus, G. (2020). The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence (arXiv:2002.06177). *arXiv*.

Sutton, R. (2019). The Bitter Lesson. (http://www.incompleteideas.net/IncIdeas/BitterLesson.html)

鈴木貴之. (2024). 人工知能の哲学入門. 勁草書房.

次田瞬. (2023). 意味がわかるAI入門ー自然言語処理をめぐる哲学の挑戦. 筑摩書房.