ハルシネーションと記号接地問題
―これからの生成 AI の教育利活用に向けて―

藤川 大祐
千葉大学教育学部

本稿では、2023 年から 2024 年にかけての生成 AI の進化を踏まえ、生成 AI におけるハルシネーション（幻覚）と記号接地問題（Symbol Grounding Problem）について 2024 年 12 月現在の状況を確認した。ハルシネーションに関しては、これまで報告されてきたハルシネーションの問題がテキスト生成においては基本的に解決していることが確認された。他方、日本語処理の問題に起因する数学的問題への誤答や画像生成における画像内の不適切な文字の生成といった問題が生じていることも確認された。記号接地問題については、AI が記号接地できない例として示されていたハルシネーション例が日本語処理による問題であったこと等から、子どもの学習の問題に当てはめることについては慎重な態度が求められることが示された。[1]

キーワード：生成 AI、ChatGPT、ハルシネーション、記号接地問題、アブダクション

1. はじめに

2022 年に ChatGPT が一般公開されて以降、自然言語の入力をもとに文章や画像等を生成する生成 AI が普及している。学校教育における活用も広がり、日本では 2023 年 7 月に「初等中等教育段階における生成 AI に関する暫定的なガイドライン」（文部科学省 2023）が定められ、その前後より、多くの実践が発表されている（藤川 2024）。

上記ガイドラインでは、当時の生成 AI について「今後更なる精度の向上も見込まれているが、回答は誤りを含む可能性が常にあり、時には、事実と全く異なる内容や、文脈と無関係な内容などが出力されることもある（いわゆる幻覚（ハルシネーション＝Hallucination））」と述べ、すべての学校に「情報の真偽を確かめること（いわゆるファクトチェック）の習慣付け」を含めた教育活動を求めている。

生成 AI におけるハルシネーションは不可避であるということが、指摘されている。Banerjee et al.（2024）は、生成 AI には以下の問題があることから、ハルシネーションは大規模言語モデル（Large Language Model, LLM）[2] の数学的・論理的構造に由来するものであり、不可避であることを示している。

① 訓練データの不完全性

100%完全なデータベースを実現することは決してできず、どれだけ大きな訓練データのセットを使用しても、ハルシネーションは避けられない。

② 正確な情報の取得は「干し草の中の針」

ある事実が訓練データベース内に存在しても、そのデータを確実に取得するのは「干し草の中の針」のようなものであり、文脈やデータのポイントがぼやけたり混ざり合ったりする可能性がある。

③ 意図分類の決定不可能性

自然言語の解釈には無限の可能性があり、誤った解釈を引き出す可能性が残る。

④ 出力生成の限界

出力がどこで止まるかを LLM は事前に知ることができないため、事前に生成内容をチェックできず、ハルシネーションを生成する可能性がある。

⑤ 事実確認の限界

出力を事後的に確認しても、有限回のステップでは 100%の精度で確認することはできない。

とはいえ、生成 AI のハルシネーションの可能性を低くすることは可能だと考えられる。情報の不足によるハルシネーションと知識を有するにもかかわらず生じるハルシネーションとを区別する Simhi et al.（2024）、生成 AI による検索において自律的に繰り返し知識検索を行う仕組みを示している Yu et al.（2024）等、ハルシネーションの抑止に資すると考えられる研究は多く発表されている。

また、2023 年から 2024 年にかけての ChatGPT の機能改善の状況を見れば、ハルシネーションへの対応も進んでいることがうかがわれる。

Daisuke FUJIKAWA：Hallucinations and the Symbol Grounding Problem: Toward the Future Use of Generative AI in Education
Faculty of Education, Chiba University

表 1　2023 年 7 月ごろと 2024 年 12 月時点の ChatGPT の主な機能比較

機能カテゴリ	2023年7月ごろの機能	2024年12月時点の機能
情報の更新性	知識の最終更新は2021年9月	ウェブ検索機能により、リアルタイムで最新情報を取得可能
画像生成	未対応	DALL·E統合により、テキストから画像生成が可能
データ処理・計算	簡単な計算やリスト作成のみ対応	Python実行環境を備え、複雑なデータ分析や計算が可能
ファイル操作	ファイルのアップロードや解析は未対応	ユーザーがファイルをアップロードし、その内容を解析・活用可能
プラグイン機能	未対応	サードパーティ製プラグインの利用が可能になり、機能拡張が容易
対話履歴の記憶	セッション内の短期的な記憶のみ	ユーザーの情報を長期的に記憶・管理可能
コード支援	基本的なプログラミング支援	複雑なコードの作成やデバッグ支援が可能
カスタムGPT（GPTs）	未対応	ユーザーがノーコードで独自のChatGPTを作成・公開可能
o1モデルの導入	未対応	人間のような推論能力を持つo1モデルを導入し、科学、コーディング、数学などの複雑なタスクに対応

文部科学省からガイドラインが出された 2023 年 7 月ごろの ChatGPT のモデルは GPT-4 であり、2024 年 12 月現在は GPT-4o である。それぞれの時点での ChatGPT 有料プランの機能を比較すると、表 1 のようになる[3]。これらのうち、「情報の更新性」、「データ処理・計算」、「対話履歴の記憶」、「カスタム GPT (GPTs)」、「o1 モデルの導入」は、ハルシネーションの抑止に直接的に関わるものと考えられる。「情報の更新性」は、情報が古いゆえに生じるハルシネーションの抑止につながる。「データ処理・計算」と「o1 モデルの導入」は、計算や推論に関わるハルシネーションを防ぐ。「対話履歴の記憶」は、対話が長くなる中で、古い会話で出されていたこととの不整合の抑止につながる。「カスタム GPT (GPTs)」は、局所的な情報を利用者があらかじめ参照できるようにすることで、知識の不足によるハルシネーションを防ぐ。

以上より、少なくとも ChatGPT 有料プラン利用時におけるハルシネーションの現れ方は、2023 年 7 月ごろと 2024 年 12 月とで大きく異なっていると考える必要がある。今後も生成 AI の機能改善が進むことを想定すれば、これからの教育における生成 AI の利活用において、ハルシネーションをどのように扱うかについてあらためて検討することが必要だと考えられる。あわせて、詳しくは後述するが、ハルシネーションと関連して議論されている記号接地問題（Symbol Grounding Problem）についても、検討が必要である。

本稿の目的は、①生成 AI のハルシネーションを今後の学校教育でどのように扱うべきか、②記号接地問題が生成 AI の教育活用にどのように関連するかを明らかにすることである。

2. 生成 AI の現状とハルシネーションの具体例

2.1. これまで見られたハルシネーション

Google 検索で「ハルシネーション　生成 AI」、「ハルシネーション　ChatGPT」といった検索語で検索を行い、ヒットしたページに記されているハルシネーションの事例を収集した。以下、類似のものを帰納的にまとめる形で分類して示す。ここでの分類は帰納的なものだが、学校教育の場面ごとの生成 AI の活用について考えるためにも、ここで示したような分類を行っておくことが有用だと考えられる。

A.　学術的な情報

「量子コンピューティングにおける最新のブレークスルー」[4]、「アインシュタインが発見した、物理法則」[5]に関する質問に対して、生成 AI が事実とは異なる回答をしたことが報告されている。

B.　歴史や地理に関する情報

「加藤清正」、「日本で 2 番目に大きな湖」に関する質問に対して、生成 AI が事実とは異なる回答をしたことが報告されている[6]。

C.　実在の人物や作品に関する情報

ミステリー小説シリーズ「すべてが F になる」に関する質問に対して、生成 AI が実際とは大きく異なる回答をしたことが報告されている[7]。また、米国のラジオパーソナリティであるマーク・ウォルターズ氏についての質問に、生成 AI が、同氏が横領を働いたと誤回答していたことが報告されている[8]。

D.　質問に誤情報が含まれている場合

実在しない「日本 DX 推進連盟」という団体[9]や「邪馬台国は東北にあったという説」[10]について尋ねる質問のように、質問に誤情報が含まれる場合、生成 AI がその誤情報を修正せずに回答したことが報告されている。

E.　数学的な課題

「5678×2647」のような計算問題[11]や「9.11 と 9.9 のどちらが大きいか」という問題[12]について、生成 AI が明らかな誤答をしたことが報告されている。

F.　画像生成

「歴史上の偉人の肖像画」の生成が求められたのに対し、画像生成 AI が実在しない人物の画像を出力したことが報告されている[13]。また、画像内に文字がある場合に、実際にはない文字が作られてしまうことが報告されている[14]。

これらについて、2024 年 12 月現在の ChatGPT 4o でも同様の問題が見られるかを確認した。

A から C については、明らかな誤りを見つけることはできなかった。

D について、まず「日本 DX 推進連盟」については、このような名称の団体に関する具体的情報は確認できないとした上で、類似の名称の 2 団体の情報が示された。また、邪馬台国が東北にあるという説については、岩手県八幡平市にあったとする説がインターネット上で紹介されていること旨の回答が示された。ただし、D に関して、「最近出てきた邪馬台国は北海道にあったという説について説明してください。」と質問したところ、「邪馬台国の所在地については、九州説や畿内説が有力視されています。しかし、近年、北海道に存在したとする新たな説が提唱されています。」として「北海道説」についての説明が示された。しかし、「北海道説はどこで言われていますか。」という追加質問に対しては「北海道に存在したとする説は、一般的な学説や主要な研究において支持されていません。そのため、北海道説に関する具体的な情報や提唱者については確認できませんでした。」という回答がなされた。質問に誤情報が含まれている場合には、ハルシネーションが起こりうることが示されたと言える。

E に関しては、単純な計算問題について誤答を見つけることはできなかったが、「9.11 と 9.9 のどちらが大きいか」という問題について ChatGPT 4o も o1 も、9.11 の方が大きいと誤答している。ただし、英語で “Which is larger, 9.11 or 9.9?” と問えば、ChatGPT 4o や o1 ではもちろん、旧バージョンの ChatGPT 4 も 9.9 の方が大きいと答える。

図 1　おかしな文字を含む画像（ChatGPT 4o が生成）

F の画像生成については、2024 年 12 月時点でもハルシネーションが起こることが確認された。たとえば、「日本の小学校で教師が黒板に文字を書いている絵を描いて。」と ChatGPT 4o に指示すると、図 1 のように文字がおかしい絵が出力された。英語で “Draw a picture of an elementary school teacher writing on a blackboard.” と指示した場合には、教師が意味不明の文字列を書いている絵が出力された。現状の生成 AI が絵を生成する際に、絵の中の文字は文字としてでなくある種の形としてのみ処理されていることがうかがわれる。

以上のように、これまで指摘されていたハルシネーションについては改善が見られるものの、質問に誤情報が含まれている場合、数学的な課題、画像生成に関しては ChatGPT 4o においても一部ハルシネーションが見られることが確認された。ただし、ここで挙げられた数学的な課題については、英語で入力することによりハルシネーションを防ぐことができた。

2.2. 教育関連で取り上げられたハルシネーション

教育関係のニュース、実践報告、書籍、論文等において、ハルシネーションが扱われているものがある程度見られる。ここでは、検索サイトや生成 AI、新聞データベースを使い、教育関連で取り上げられたハルシネーションの事例を、2.1.で示した分類にあてはめて見ていく。

A.　学術的な情報

中学校の理科の課題で、胃では機能しない唾液アミラーゼの働きを「胃で消化されやすくする」とした生徒が多く出てきたことが報告されている[15]。これは、生成 AI を搭載した検索エンジンが食品企業のホームページの「唾液に含まれる酵素（アミラーゼ）が、食べ物に含まれるでんぷんを分解し、胃で消化されやすい状態にします」を示したことによるとされる。生成 AI 自体のハルシネーションではないが、誤解を招く検索結果が示された。

B.　歴史や地理に関する情報

小学校教諭が自身の勤務している青森県五所川原市について ChatGPT に質問したところ、別の自治体にある「鶴の舞橋」や、地元でも聞いたことのない「ごしょつがるニンニク」という名産品が紹介されたことが報告されている[16]。

新井は、ChatGPT 3.5 に東京大学の世界史の入試問題を解かせたところ、多くの誤りが含まれていたことを報告している[17]。

C.　実在の人物や作品に関する情報

小学校 3 年生の国語の授業で文学作品「モチモチの木」について関連する文章を生成 AI に求めたところ、元の物語から話が大きく変わっていたことが報告されている[18]。また、中学校の国語で取り上げられる森鴎外の「高瀬舟」についてハルシネーションが生じた例が報告されている(みんなのコード 2023)。

前出の新井は、ChatGPT がウェブ検索機能で日本の時事問題（具体的には自由民主党の総裁選挙関連）について聞いていると、情報源に “Japan Forward”[19] が多く、情報の偏りが懸念されるとしている。

D.　質問に誤情報が含まれている場合

中村（2024）は、「2040 年のノーベル化学賞受賞者は？」と尋ねると、生成 AI が「マックス・ブランク研究所のケン・ヤマダ教授に授与されました」というそれらしい回答がなされたことを報告している（p.35)。質問が未来の未確定のことについて、あたかも確定しているかのように読めるものであることが、このような回答につながっているものと考えられる。

E.　数学的な課題

今井（2024）は、「2 分の 1 と 3 分の 1 のうち、どちらが大きいですか？」という問題、「6%は下のどれと同じですか？　選択肢：　A：6 割　B: 6 割の 10 分の 1　C:　6 割の 100 分の 1　D: 上のどれでもない」という問題、さらには「下の問題文にある計算式の答えに最も近いものを選んでください。問題：12/13 - 11/12　選択肢：A: 0　B: -1　C: 1　D: 2」という問題で、生成 AI が誤答をすることを報告している。また、小平（2024）は Google Gemini が単純な連立方程式を正確に解けない例があることを報告しており、小出（2024）は Google Bard（後の Google Gemini）が図形の証明問題について誤答をしている例を示している[20]。御園（2024）は、整数についての説明を求めたところ、ChatGPT が「整数は、正の整数、負の整数、およびゼロからなる数の集合です」と循環定義のような回答がなされたことを報告している。以下は数学的な問題というより数量に関わる問題であるが、らいけん（2023）は、上記以外に、「300 文字で書いてください」としても異なる量の文章が作られてしまう、「自然数 123,456,789 の日本語読みをひらがなで書いてください」としても誤った読み方が出力されてしまうといった問題を指摘している。

F.　画像生成

小学校の授業で俳句に合わせた挿絵を生成 AI に作らせたところ、紅葉を「燃える木々」と詠んだ句に対して、生成 AI が文字通り炎上する木の絵を描いたことが報告されている[21]。

これらについて、2024 年 12 月現在の ChatGPT 4o の回答を確認した。

A から C については、改善が見られる。ChatGPT 4o にはウェブ検索機能が備わっているため、ウェブを検索して回答できるものについては、基本的に適切な回答がなされるようである。2024 年 12 月現在の政治に関する時事問題(少数与党の状況、政治倫理関連の国会での議論、税制「103 万円の壁」についての議論）についての質問への ChatGPT 4o の回答においては、読売新聞、朝日新聞、FNN 等のサイトが参照されており、“Japan Forward” が参照される状況は再現されなかった。

D の「2040 年のノーベル化学賞受賞者は？」という問いに対しては、「2040 年のノーベル化学賞受賞者に関する情報は、現時点（2024 年 12 月 16 日）では存在しません」と適切に回答しており、前提が誤っている質問についても改善が見られる。

E の数学的な課題のうち、「6%は下のどれと同じですか？　選択肢：　A：6 割　B: 6 割の 10 分の 1　C:　6 割の 100 分の 1　D: 上のどれでもない」、「300 文字で書いてください」、「自然数 123,456,789 の日本語読みをひらがなで書いてください」といった問題については、誤答が再現された。これらはいずれも日本語での数に関する表現の扱いに関わるものであるので、日本語特有の数に関する表現の処理に問題があることがうかがわれる。

F の「燃える木々」の画像生成の問題については、紅葉の比喩であることを示すことにより改善が可能である。なお、「燃える木の絵を描いてください。ただし、燃える木というのは比喩で、紅葉のことを示しています。」というプロンプトで ChatGPT 4o が出力した絵は図 2 の通りで

図 2　比喩としての燃える木（ChatGPT 4o が生成）

あり、紅葉が燃えている様子が描かれている。

2.3. 現状を踏まえたハルシネーションの扱いについて

以上のように、2024 年 12 月現在の ChatGPT をはじめとした生成 AI は、主にウェブ検索機能によって 2023 年 7 月ごろと比較するとかなりの程度、ハルシネーションの問題が改善されている。他方、数学的な問題については、少なくとも日本語で問うと明確な誤答を出力する場合が確認できる。また、画像生成については、プロンプトを調整することにより適切な画像の生成が一定程度期待されるものの、画像中の文字については適切な文字あるいは文字列にはならないことが基本だと考えられる。

みんなのコード (2023) は、生成 AI の利活用について「自分が知りたいことを検索する感覚で使われる方もいますが、ハルシネーションの恐れがあるため結局ファクトチェックが必要になります。思考力や創造性を高める場面では生成 AI を活用し、調べ物をする際は検索エンジンを使うといった使い分けが重要です」と言っている。生成 AI を利用した場合にファクトチェックは必要である。だが、生成 AI 自体がウェブ検索機能をもつようになったことから、検索エンジンとの使い分けについては単純に割り切れなくなりつつある。ウェブ検索をしたい場合には、まず検索エンジンと生成 AI のいずれかで調べ、結果に満足できない場合にはもう一方で調べるというように、柔軟に併用することが現実的であろう。また、複数の生成 AI を併用することも考えられる。

3. 記号接地問題と生成 AI との関係

3.1. 記号接地問題をめぐる議論

前出の今井 (2024) は、生成 AI が数学的な問題で誤答をすることを、記号接地問題 (symbol grounding problem) の観点から議論している。記号接地問題とは、記号 (symbol) がいかにして実世界の具体的な対象や出来事と結びつき (grounding)、意味を獲得することができるかという問題である[22]。生成 AI を学校教育において活用する際には、生成 AI が記号の意味をどのように理解しているかが問題となりうる。今井 (2024) のように、生成 AI が意味を理解していないことを、記号接地問題と関連づけた議論もある。このため、記号接地問題を検討することは、学校教育における生成 AI の活用について考える上で重要である。

記号接地問題を提起した Harnad (1990) は、この問題を、Searle (1980) による「中国語の部屋」思考実験を引き合いに出して説明している。この思考実験は、中国語をまったく理解しない人間が規則に従って記号を操作することで、あたかも中国語を理解しているかのように振る舞える状況を示したものである。この思考実験から、形式的な記号操作だけでは意味の理解を生み出すことができないということ、すなわちプログラムを実行するコンピュータは単に記号操作を行っているにすぎず、意味を理解しているわけではないということが言える。

Harnad は、記号接地問題の別の例として、「中国語－中国語辞書の堂々巡り」を挙げている。これには二つのパターンがある。第一のパターンは中国語の辞書（中国語で中国語の説明がなされている辞書）で第二言語として中国語を学ぶものであり、第二のパターンは中国語の辞書で第一言語として中国語を学ぶものである。前者は母語や現実の経験・知識に基づくことが可能であるが、後者は母語の経験・知識に基づくことができない。コンピュータ・プログラムにおける記号接地問題は、後者のようなものだと考えられる。

Harnad は、記号接地問題の解決策の候補を示してもいる。Harnad は以下のように言う。

> 記号的表現は以下の 2 種類の非記号的表現に基づいて下から接地されなければならない。(1) 「アイコン的表象」 (iconic representations)、これは遠方の物体や出来事の近くでの感覚投影の類似物である。(2) 「カテゴリー的表象」 (categorical representations)、これは感覚投影から物体や出来事のカテゴリーに不変な特徴を抽出する、学習されたまたは生得的な特徴検出器である。基礎的な記号はこれら物体や出来事のカテゴリーの名前であり、それらの (非記号的な) カテゴリー表象に基づいて割り当てられる。より高次の (3) 「記号的表象」 (symbolic representations) は、これらの基礎的記号に基づき、カテゴリーの所属関係を記述する記号列で構成される (例：「X は、Z であるような Y である」 (An X is a Y that is Z))。

たとえば、ウマの視覚的なイメージが「アイコン的表象」であり、ウマ特有の体型や脚の長さや首の形状などが「カテゴリー的表象」だと考えられる。そして、縞模様のあるウマがゼブラ (シマウマ) だというようにカテゴリーの所属関係によって作られるのが「記号的表象」である。このように、Harnad は、視覚情報などの感覚情報に基づいて基礎的な記号が作られ、そうした基礎的な記号からさらに他の記号が作られるという考え方が、記号接地問題の解決策の候補だとしている。Harnad は、神経ネットワークをモデルとした「コネクショニズム」を記号操作と組み合わせた「ハイブリッド・システム」の有効性を示唆する。しかし、「ハイブリッドシステムがどのようにして感覚的投影の不変な特徴を見つけ出し、それによって物体を正確に分類し識別するのかが明らかにされていない」と述べ、記号接地問題が未解決であることを示している。すなわち、

記号接地なしにカテゴリー的表象が形成できるかという問題が残っているということである。なお、Harnad 自身も触れているが[23]、そもそも物体や出来事に不変的特徴があるのかと言う「消失する交差点」（vanishing intersections）問題も、記号接地問題解決の困難さに関わると考えられる。

Steels & Vogt (1997) は、記号接地問題に明示的な言及はないものの、後に記号接地問題の解決に関連するとされる実験について報告している。Steels & Vogt は、Wittgenstein (1974) による「言語ゲーム」を踏まえ、2 組のロボット・エージェントによる「適応型言語ゲーム」(adaptive language game) を行った。このゲームの基本的なシナリオは以下の通りである。

(1) 2 組のエージェントが互いにコンタクトする。一方は話し手、もう一方は聞き手の役割を担う。
(2) 話し手は周囲の環境から一つの物体を会話のトピックとして選択し、指差し等の非言語的手段で聞き手がそのトピックに注意を向けるようにする。
(3) 各エージェントは、多様な物体を知覚し、特徴によって他の物体と区別できるようにする。
(4) 話し手は特徴の 1 セットを符号化（encoding）する。
(5) 聞き手は符号化された表現を復号化（decoding）する。
(6) 聞き手は復号化が話し手の期待と合っていたか否かをフィードバックする。

Steels & Vogt (1997) はこの実験によって、2 組のロボット・エージェントが対話をすることにより、他者からの助けなしに、環境や相手のエージェントとの相互作用のみによって知覚カテゴリーや共通の記号を形成できることを示したとしている。このことは、記号接地なしにカテゴリー形成が可能であるということを示していると考えられる。後に Steels (2008) は、ここで示された種類の実験を根拠に、エージェントが自律的に意味を生成し、身体化やカテゴリー化を通して意味を世界に結びつけ、自律的にこれらの意味を呼び出すことが可能であるとして、記号接地問題がすでに解決したと主張している。

しかし、Taddeo & Floridi (2005) は、Steels & Vogt (1997) による「適応型言語ゲーム」について、このゲームにおいてはエージェントがすでに接地された記号を操作することを前提としていると指摘し、このゲームによって記号接地問題は解決していないとしている。また、Müller (2015) も、Steels の提案するロボット・エージェントは「純粋な構文的装置」であり、「理解」がどのように生じるかが示されていない、コミュニケーションの成功はプログラマーによる装置の「適切な」設定によって達成されたものなので外部的に導入されたものだと批判している。

また、Taddeo & Floridi (2007) は、記号接地問題の解決策として、ロボット・エージェントの行動が引き起こす内部状態が記号の意味を形成するという「行動ベースの意味論」を提起している。たとえば、ロボットが「左に曲がる」という行動をとると、その行動によって生じる内部状態が対応する記号の意味となるというものである。Müller (2015) は、こうした「行動ベースの意味論」に関して、ロボット・エージェントに目的や規範性が欠如しているという点等を指摘し、疑問を呈している。

そして、谷口 (2016) は Steels の議論を受け、記号が外部の現実世界に接地する（記号接地する）か否かではなく、エージェント（ロボットや AI）が環境との相互作用の中で自律的に記号を形成・獲得していく「記号創発」にこそ注目すべきだと主張している。これは記号接地問題が解決済みかを問うのでなく、記号接地という問題設定そのものを再検討しようとするものである。谷口曰く、「私達が捉えるべきは物理記号システムの接地の仕方ではない。記号創発システムの現れ方なのである。人が設計した、ロボットにとっては極めて恣意的な記号システムの接地を考えるのではなく、ロボットにとって自然な記号システムが環境との物理的相互作用、他者との記号的相互作用を通して組織化されていく過程を捉えることこそ、記号接地問題の本質的解決につながるのである」 (p.79)。

以上のように、Harnad が提起した記号接地問題については、解決したという立場と解決していないという立場が分かれている。このことは、谷口も述べているように、記号接地問題の曖昧さによるものと考えられる。

Harnad は記号接地問題について、「形式的な記号システムの意味解釈を、単に私たちの頭の中の意味に寄生するものではなく、そのシステム自体に内在的なものとするにはどうすればよいのか？」と問うていた。だが、ロボット・エージェントや AI といったコンピュータ・システムにおける記号接地問題が扱われるとき、そのシステムは（少なくとも現状では）人間がプログラムしたものである。人間がプログラムしている以上、システムは私たち人間の「頭の中の意味」と全く無関係ではありえず、プログラムした人間の意図あるいは志向に当然影響を受ける。こうしたことまでをも排除するのであれば、そもそも Harnad が提起した記号接地問題は、コンピュータ・システムについて問われるべき問題ではないこととなる。

Taddeo & Floridi (2005) は、記号接地問題が解決したとするためには、次の「ゼロ意味論的コミットメント条件」(zero semantical commitment condition、以下「Z 条件」とする。）が必要だとしている。

> a) いかなる形の内在主義も認められない（no innatism）： 人工エージェント（AA）に既にインストールされている意味的なリソース（いわゆる

“virtus semantica”）を前提にしてはならない。

b）いかなる形の外在主義も認められない（no externatism）：意味的に熟達した“deus ex machina”（外部の力）によって、外部から意味的リソースをアップロードすることも認められない。

もちろん、(a)と(b)の条件は以下の可能性を否定するものではない。

c) AA が記号を意味付けするために、自身の能力やリソース（例えば計算的、構文的、手続き的、知覚的、教育的な能力などを、アルゴリズム、センサー、アクチュエーターなどを通じて活用する）を持つこと。

だが、仮にこの Z 条件を採用したとしても、記号接地問題の曖昧さは解消されないと考えられる。システムが計算的リソースを持っているということは数についての意味を内在していることとどう違うのか、センサーを持っていることは光や物質についての意味を内在していることとどう違うのかといったことが曖昧である。

さらに、Müller (2015) は、記号接地問題の解決には理解や目的や思考性が必要だという立場をとっており、ここでは Z 条件とはまた異なる条件が挙げられている。

こうした問題を整理するためには、Steels (2008) の言う c 記号（c-symbol）と m 記号（m-symbol）との区別が有用だと考えられる。c 記号はコンピュータ科学における記号であり、コンピュータのメモリ上のアドレスであるポインタとして表現される。m 記号は、芸術、人文学、社会科学、認知科学の伝統における意味志向の記号である。意味志向の m 記号に接地が必要だとしても、コンピュータ内部の c 記号に接地が必要とは言えないはずである。このように考えれば、記号接地問題は AI を含むコンピュータ・システムの問題ではないこととなる。そうであれば、谷口の言うように、AI やロボットについて記号接地問題を問うより、AI やロボットが環境や他者との相互作用を通してどのように記号システムを組織化していくのかを問うべきだということとなる。

3.2. 数学的問題のハルシネーションと記号接地

今井（2024）の議論に戻ろう。今井は、「どんなに流暢でも、生成 AI が記号接地をしていないことには変わりはない。ChatGPT は超優秀な『次のことば予測マシン』である。ビッグデータの海を泳ぎ、単語や句が別のどの単語や句といっしょに使われるのかを計算し、この単語の次に来そうな単語や句を予測する」と言う。そして、イチゴの画像を選ぶこと、イチゴが果物であること、「甘酸っぱい」味をもつことなどを言語で返すことができるとした上で、「しかし、『甘酸っぱい』がどういう味なのかはわからない。リンゴなど他の『甘酸っぱい』食べ物と区別することは（少なくとも現状の）AI にはできない」と述べ、現状の生成 AI は「ことばの海を、ひとつのことばから別のことばへ漂流しながら、言語による言い換えで人間の質問に答えているにすぎない」と言う。

ここまでの今井の議論は、少なくとも Harnad の議論とは噛み合っていない。Harnad は、物体や出来事の感覚を直接的に表現した「アイコン的表現」や、そこから抽出された特徴を基に構築される「カテゴリー的表現」によって、記号が物体や出来事と結びつく、すなわち接地することの可能性を示していた。今井が自ら述べているように、現状の生成 AI はイチゴの画像データを学習している。Harnad との間で噛み合った議論をするには、視覚データの扱いについて論じる必要がある。

すでに述べたように、今井は、生成 AI が「2 分の 1 と 3 分の 1 のうち、どちらが大きいですか？」という問題や「6%は下のどれと同じですか？　選択肢：　A : 6 割　B: 6 割の 10 分の 1　C:　6 割の 100 分の 1　D: 上のどれでもない」という問題に誤答することを報告している。このことについて今井は、「ChatGPT と算数がお手上げ状態の子どもに共通すること。それは、パターンは覚えているが、その意味が抽象化されていない」と述べる。しかし、すでに見たように、質問を英語で入れれば ChatGPT は旧バージョンでも適切に回答することから ChatGPT が日本語特有の「割」という単位、2 バイト文字である日本語文字のカウント方法、さらには 4 桁ごとに「万」や「億」とする命数法といったことに十分対応していない点が問題だと考えるべきであろう。ChatGPT は、算数がお手上げの子どもより、日本語が苦手であるがゆえに日本語の数に関する表現をうまく扱えない子どもに近いと考える必要がある。

今井は別の場所（大澤ほか 2024）で、1/2 や 1/3 といった記号の意味内容が「ピンとこない」ことについて「それは記号接地がうまくできていないからです」と述べている。だが、そもそも数学的な概念について記号接地とは何かは曖昧だ。同じ場所で大澤が「数学とは考えてみれば、原理的に、僕らが経験可能だという意味での対応する実在をもちません」と指摘しているように、実世界における物理的対応物が存在しない対象が多い数学について、記号接地という概念を用いること自体に無理がある。

以上のように、生成 AI の数学的問題におけるハルシネーションに関しては、「記号接地」が何を意味するかが曖昧であることやハルシネーションが主に日本語における数の処理の問題によると考えられることから、記号接地問題と関連づけることには無理がありそうだ。

3.3. 生成 AI の言語習得と記号接地問題

今井（2024）は、子どもの学習についても「記号接地」という言葉を用いて議論している。むしろ、今井の議論の中心は、生成 AI の記号接地問題でなく、子どもの学習の問題である。

子どもがどのように記号接地しているかについて今井が論じているのは、主に言語習得に関してである。このことについては、より詳しい今井・秋田（2023）を見よう。

今井・秋田は、記号接地問題は「子どもが母語を学習する際に実際に起こる」問題だと言う。「意味を知っていることばを一つも持たない子どもは、まったく意味のない記号を使って新たに記号を獲得することはできない」のであり、最初の一群の語を記号接地することと、その後に直接の身体経験がなくても他の語を習得することがいかにして可能かが、子どもの言語習得の過程として問われる。今井・秋田は、オノマトペと「ブートストラッピング・サイクル」を中心にこの過程を説明している。

今井・秋田は、幼い子どもがオノマトペを多用することに着目し、子どもが言語を習得する過程を次のように説明する。赤ちゃんであっても、音と対象の対応づけを自然に行う様子が見られ、このことは「ことばの音が身体に接地する最初の一歩」だと示唆される。そして、オノマトペが足場となり、母語の音の特徴、音の並び方、音が何かを「指す」こと、音と意味の結びつき、情報から一部の特徴を「切り出す」こと等を子どもは学ぶ。そして、子どもは成長する中でオノマトペから離れ、恣意的で抽象的な記号の体系を学ぶ。すなわち、子どもは、既存の知識をもとに、知識を雪だるま式に増やす「ブートストラッピング・サイクル」により知識を増やしていく。

今井・秋田の言う「ブートストラッピング・サイクル」において重要な役割を果たすのが、推論の一形式であるアブダクションだ。アブダクションは、哲学者パースが提唱した推論形式であり、観察された事実等からある程度の妥当な仮説を立てることを言う[24]。今井・秋田が挙げている例は、白い豆ばかりが入った袋と白い豆があったときに、その白い豆は白い豆ばかりが入った袋から取り出した豆であると結論づけるものだ。アブダクションは仮説を導くだけのものなので常に正しい答えを導く推論ではないが、それゆえに新しい知識を生み出すことにつながりうる。

今井・秋田は、赤ちゃんの言い間違いに、アブダクションの「足跡」が見られると言い、練乳を「イチゴのしょうゆ」を言ったり、蹴ることを「足で投げる」と言ったりする例をあげている。これらの例は誤りと言えるが、子どもはアブダクションによってさまざまな表現を創造し、創造されたものを修正しながら知識を発展させていく。今井・秋田が言うように、「知識を創造する推論には誤りを犯すこと、失敗することは不可避なことである」。

以上のような今井・秋田の子どもの言語習得の過程についての議論は、説得的である。だが、この議論に関して注意を要する点がある。それは、生成 AI の言語習得の過程は、今井・秋田が言う人間の子どもの母語習得の過程とは異なるということである。生成 AI は、記号接地を特に必要とせず、大量のテキストデータから、記号相互の関係を学ぶ。未知の語の意味を推測する際には、アブダクションをするわけではなく、前後の文脈等から統計的な推論を行う。生成 AI はこうした学習によって、単純なハルシネーションがかなり限定される程度には進化している。

今後の学校教育での生成 AI の活用において、生成 AI が記号接地ができていないと考えられることをもって、生成 AI の使用に問題があると考えるのは、短絡的すぎる。生成 AI による学習は人間の学習とは異なるものかもしれないが、少なくともテキストでの対話においては人間との間でかなり円滑なコミュニケーションが可能となっている。少なくとも記号接地ができていないことが、生成 AI によるハルシネーションにつながっているとは考えにくい。

4. 考　察

本稿では、これまで指摘されてきた生成 AI におけるハルシネーションについて概観し、2024 年 12 月現在の状況を確認した。この結果、ウェブ検索機能が備わったことによりハルシネーションがかなり抑えられていることが確認された。数学的な問題についてはハルシネーションが発生することが確認されたものの、英語で入力した場合には再現されず、数に関する日本語処理に課題があるらしいことが確認された。また、画像生成については、画像内の文字あるいは文字列に特に問題があることが確認された。

以上より、生成 AI におけるハルシネーションは、少なくともテキスト生成に関しては実用上問題ないレベルに近づきつつあると結論づけられるだろう。数学的な問題については、現状でもいったん英語訳して入力することによって、ハルシネーションの回避はほぼ可能である。このことは、数に関する日本語処理の課題が開発側に認識されれば、改善が可能であることを意味する。生成 AI が基本的にウェブ上の情報を学習していることを踏まえれば、ウェブから得られる知識をもとにした情報提供については、ある程度の推論も含め、ハルシネーションなしで回答が得られることが期待できるようになりつつあると考えられる。

もちろん、ハルシネーションを完全になくすことは不可能である。世界では常に新しいことが起こり、既存の言葉に新たな意味が加えられたり、新たな言葉が誕生したりすることが続く。プロンプトを書いた者の意図が正確に伝わる保証もない。こうしたことから、生成 AI の回答が適切かどうかの確認が不要となるわけではない。しかし、生成 AI による学習は日々進むので、同じようなハルシネーションがいつまでも繰り返されることも考えにくい。学校の授業において、ハルシネーションが発生する様子を児童生

徒に見せて、生成 AI が間違えることもあるというやり方は、成立しにくくなっている。

AI におけるハルシネーションの問題は消えないとしても、そのことと記号接地問題とは区別されるべきである。そもそも、AI に記号接地が必要だとする根拠はない。少なくとも、生成 AI が数学的問題に関して生じさせるハルシネーションの事例として出されたものは、数に関する日本語処理の問題ゆえに生じたものしか確認されておらず、記号接地問題とは無関係である。そして、もともとコンピュータ・システムの話だったはずの記号接地問題を、子どもの学習に当てはめることにも慎重さが求められる。少なくとも、記号接地問題についての明確な定義の上で議論がなされるべきだろう。

Bloomberg の記事によれば、OpenAI は AI の発展を次のように 5 段階で示している[25]。

レベル 1 チャットボット、会話言語をもつ AI
レベル 2 推論者（Reasoners）、人間レベルの問題解決
レベル 3 エージェント、行動を起こせるシステム
レベル 4 イノベーター、発明を支援する AI
レベル 5 組織（Organizations）、組織の仕事ができる AI

ChatGPT はレベル 2 に近づいているとされる。これは、博士号レベルの教育を受けた人間と同等の問題解決ができるレベルと説明されている。しかし、このためには、長文を円滑に処理したり、学術論文や専門書について広く学習しておいたりする必要があるだろう。数に関する日本語処理の改善も求められる。画像生成における文字の扱いについても改善が必要だ。おそらく、こうした課題については、今後短期間での克服が考えられているのだろう。

今後、教育での生成 AI 活用を活用する際に、一目でわかるようなハルシネーションが頻繁に生じることを前提にすべきではない。ハルシネーションについて学ぶ際には、あらかじめ事例を用意しておく等の準備が必要となるだろう。他方、わかりやすいハルシネーションが起きにくくなるからこそ、実際にハルシネーションが生じた際に発見が難しくなるとも考えられる。具体的には、授業で生成 AI を扱う際に生成 AI による回答の妥当性を検証することの重要性と検証方法を扱うこと、数学的問題や文字を含む画像の生成においては特に注意して扱うこと等が求められる。生成 AI の賢さを侮らず、必要に応じて裏付けを確認しながら利用できるようにすることが、今後の教育において求められる。

---

[1] 本稿の一部は、日本教育工学会 2025 年春季全国大会（2025 年 3 月 9 日、成城大学にて）において「生成 AI におけるハルシネーションと記号接地問題」として発表したものである。

[2] Large Language Model であり、大量のテキストデータを学習して言語のパターンや文脈を理解・生成する AI モデルのこと。生成 AI の基盤的な技術である。

[3] 以下のようなサイトの記事を元に作成した（いずれも 2024 年 12 月 11 日最終確認）。
Lifewire “Does ChatGPT Have Real-Time Data?” (2024 年 8 月 17 日)
https://www.lifewire.com/chatgpt-real-time-data-8679846?utm_source=chatgpt.com
株式会社 WWG「【検証】ChatGPT 有料版と無料版を比較｜OpenAI o1 とは」（2024 年 10 月 21 日）
https://wwg.co.jp/blog/52107
AI & NFT Times「ChatGPT は何ができる？使い方や活用、モデルの種類、料金徹底解説！」（2024 年 10 月 23 日）
https://ai-henoheno-mohero.com/chatgpt-start/

[4] AINow「ChatGPT ハルシネーションの原因と対策方法は？」(2024 年 8 月 25 日)
https://ainow.jp/what-causes-chatgpt-hallucination/ (2024 年 12 月 25 日最終確認)

[5] romptn ai「AI における「ハルシネーション」とは？対策方法や具体例などを解説！」（2024 年 6 月 21 日）
https://romptn.com/article/46533#google_vignette (2024 年 12 月 25 日最終確認)

[6] WEEL「生成 AI のハルシネーションとは？　種類や事例、発生の原因と対策方法について解説」（2024 年 9 月 20 日）
https://weel.co.jp/media/hallucination (2024 年 12 月 25 日最終確認)

[7] IT navi「GPT-4 のハルシネーション（幻覚）の研究」（2023 年 4 月 6 日）
https://note.com/it_navi/n/na19444975e30 (2024 年 12 月 25 日最終確認)

[8] AIsmiley「生成 AI のハルシネーションとは？　発生の原因やリスク、その対策について」（2024 年 5 月 16 日）
https://aismiley.co.jp/ai_news/hallucination/ (2024 年 12 月 25 日最終確認)

[9] Rentec Insight「生成 AI の活用で注意すべき『ハルシネーション』とは？」（2024 年 2 月 29 日）
https://go.orixrentec.jp/rentecinsight/it/article-385 (2024 年 12 月 25 日最終確認)

[10] 前出、IT navi「GPT-4 のハルシネーション（幻覚）の研究」

[11] 前出、Rentec Insight「生成 AI の活用で注意すべき『ハルシネーション』とは？」

[12] リーブリーダー「なぜ AI は小学生にもわかる数の大小を間違えてしまうのか？　その考察」（2024 年 11 月 19 日）
https://www.leapleaper.jp/2024/11/19/why-does-ai-make-mistakes-in-math-even-kids-can-solve/ (2024 年 12 月 25 日最終確認)

[13] 前出、romptn ai「AI における「ハルシネーション」とは？対策方法や具体例などを解説！」

[14] SB C&S Windows マイグレーション相談センター「ハルシネーションが発生する原因と対策方法を詳しく解説」（2024 年 9 月 26 日）
https://licensecounter.jp/win_migration/blog/c513fe5813f74b55ecd5ebd2b9b3c2f62bfafe70/ (2024 年 12 月 25 日最終確認)

[15] FNN プライムオンライン「【物議】「生成 AI」丸写し？中学校の課題で“同じ誤答”が続出　教諭「生徒たちには多くの学びがあった」」(2024 年 3 月 8 日)
https://www.fnn.jp/articles/-/668432#google_vignette (2024 年 12 月 12 日最終確認)

[16] EdTechZine「教師が授業で生成 AI を使って見せる実践――重要なのは「生成して終わり」にしないこと」（前田昌顕、2024 年 11 月 11 日）
https://edtechzine.jp/article/detail/11671 (2024 年 12 月 12 日最終確認)

[17] 文部科学省「初等中等教育段階における生成 AI の利活用に関する検討会議」（第 4 回、2024 年 9 月 24 日開催）における新井紀子氏による会議資料及び議事録における新井氏の発言より。

https://www.mext.go.jp/b_menu/shingi/chousa/shotou/193/index.html （2024 年 12 月 25 日最終確認）

[18] みんなの教育技術「ChatGPT を小３国語「モチモチの木」で使ってみたら 【池田修×藤原友和チャット対談#3】」（2023 年 5 月 13 日）
https://kyoiku.sho.jp/235965/ (2024 年 12 月 12 日最終確認)

[19] 一般社団法人ジャパンフォワード推進機構によるウェブサイトであり、インターネット上で外国語で発信される日本関係のニュース等のコンテンツを掲載している。理事に産経新聞社の関係者が多く、産経新聞社との関連が深いものと考えられる。
https://japan-forward.com/ja/ (2025 年 3 月 8 日最終確認)

[20] 小出 (2024) は、その後、生成 AI がこの種の問題に正解することを述べている。そして、あえて「間違えて解答してください」と指示することで、生成 AI から誤答が得られることを報告している。

[21] 日本経済新聞「生成 AI、小学校の授業にも　使い方学びトラブル防ぐ」（2024 年 6 月 5 日）
https://www.nikkei.com/article/DGXZQOUE25BM30V20C24A4000000/ (2024 年 12 月 12 日最終確認)

[22] この問題を提起した Harnad (1990) は、次のように述べている。「記号接地問題とは、「形式的な記号システムの意味解釈を、単に私たちの頭の中の意味に寄生するものではなく、そのシステム自体に内在的なものとするにはどうすればよいのか？任意の形状に基づいてのみ操作される意味のない記号トークンの意味を、他の意味のない記号以外のものにどのようにして接地させることができるのか？」という問題である。」

[23] Harnad (1990) の注 20 で説明がなされている。

[24] 詳しくは、たとえば米盛 (2007) を参照。

[25] Bloomberg “OpenAI Scale Ranks Progress Toward ‘Human-Level’ Problem Solving (Rachel Metz, 2024 年 7 月 21 日)
https://www.bloomberg.com/news/articles/2024-07-11/openai-sets-levels-to-track-progress-toward-superintelligent-ai?embedded-checkout=true&leadSource=uverify%20wall
(2024 年 12 月 25 日最終確認)

引用文献

Banerjee, S., Agarwal, A., & Singla, S. (2024) LLMs will always hallucinate, and we need to live with this. arXiv preprint arXiv:2409.05746

藤川大祐 (2024) 「初等中等教育実践における生成 AI の活用のあり方」、千葉大学教育学部研究紀要、第 72 巻、pp.83-90

Harnad, S. (1990) The symbol grounding problem. Physica D: Nonlinear Phenomena, 42, 335-346

今井むつみ (2024) 『学力喪失－認知科学による回復への道筋』、岩波新書

今井むつみ・秋田喜美 (2023) 『言語の本質－ことばはどう生まれ、進化したか』、中央公論新社

小平理 (2024)「模範解答の導出の実際」、小原豊・金児正史・北島茂樹 (編著) 『実践事例で学ぶ生成 AI と創る未来の教育』、東洋館、pp.77-80

小出晴斗 (2024) 「足立区立興本扇学園における実践と成果」、小原豊・金児正史・北島茂樹 (編著) 『実践事例で学ぶ生成 AI と創る未来の教育』、東洋館、pp.103-106

みんなのコード (2023) 『学校の生成 AI 実践ガイド－先生も子どもたちも創造的に学ぶために－』、学事出版

御園真史 (2024) 「生成 AI と批判的思考」、小原豊・金児正史・ 北島茂樹 (編著)、『実践事例で学ぶ生成 AI と創る未来の教育』、東洋館、pp.36-38

文部科学省 (2023) 「初等中等教育段階における生成 AI に関する暫定的なガイドライン」
https://www.mext.go.jp/content/20230718-mtx_syoto02-000031167_011.pdf (2025 年 1 月 10 日最終確認)

Müller, V. C. (2015) Which symbol grounding problem should we try to solve? Journal of Experimental & Theoretical Artificial Intelligence, 27(1), 73-78

中村大輝 (2024) 『生成 AI で進化する理科教育－導入から実践までの完全ガイド』、東洋館

大澤真幸・松尾豊・今井むつみ・秋田喜美 (2024) 『生成 AI 時代の言語論』、左右社

らいけん (2023) 『教師のための ChatGPT ガイド－AI を活用した教育の手引き』、(出版社表記なし)

Searle, J. R. (1980) Minds, Brains, and Programs. Behavioral and Brain Sciences, 3 (3), 417-457

Simhi, A., Herzig, J., Szpektor, I., & Belinkov, Y. (2024) Distinguishing Ignorance from Error in LLM Hallucinations. arXiv preprint arXiv:2410.22071

Steels, L. (2008) The symbol grounding problem has been solved, so what’s next？ De Vega, M., Glenberg, A. and Graesser, A. (ed.) Symbols and embodiment: Debates on meaning and cognition, Oxford University Press, 223-244

Steels, L., & Vogt, P. A. (1997) Grounding adaptive language games in robotic agents. In Advances in artificial life: Proceedings of the Fourth European Conference on Artificial Life (ECAL 97), Brighton, 28-31 July, 1997 (pp. 474-482). MIT Press

Taddeo, M., & Floridi, L. (2005) Solving the symbol grounding problem: a critical review of fifteen years of research. Journal of Experimental & Theoretical Artificial Intelligence, 17(4), 419-445

Taddeo, M., & Floridi, L. (2007) A praxical solution of the symbol grounding problem. Minds and Machines, 17, 369-389

谷口忠大 (2016) 「記号創発問題: 記号創発ロボティクスによる記号接地問題の本質的解決に向けて」(< 特集> 認知科学と記号創発ロボティクス: 実世界情報に基づく知覚的シンボルシステムの構成論的理解に向けて)、人工知能, 第 31 巻 1 号、pp.74-81

Wittgenstein, L. (1953) Philosophisch Untersuchungen. Bazil Blackwell（L. ヴィトゲンシュタイン著，鬼界彰夫訳 (2020) 哲学探究．講談社）

米盛裕二 (2007) 『アブダクション　仮説と発見の論理』、勁草書房

Yu, T., Zhang, S., & Feng, Y. (2024) Auto-RAG: Autonomous Retrieval-Augmented Generation for Large Language Models. arXiv preprint arXiv:2411.19443