HISTORY OF AI

A HISTORICAL SURVEY OF IDEAS, SYSTEMS, AND INSTITUTIONS

AIの歴史

知能の探求、その軌跡と展望

記号主義、ニューラルネットワーク、機械学習、深層学習、大規模言語モデル、AIガバナンスまでを通史としてたどる。

構成

全26章 / 7部構成

版

2026年3月版

発行

haya株式会社

CONTENTS

目次

第I部

基盤と黎明期（1950年代〜1960年代）

第1章〜第4章

第1章　知能の機械化への道 —— 思想的源流と制度的基盤

1.1　マッカロック＝ピッツの形式ニューロンモデル（1943）

人工知能の歴史を語るにあたって、その出発点をどこに置くかは一つの知的判断を要する。古代ギリシアの自動機械（アウトマトン）の伝説や、ライプニッツの普遍的計算の夢にまで遡ることも可能だが、本書では1943年を一つの起点として採用する。この年、神経生理学者ウォーレン・マッカロックと数学者ウォルター・ピッツが発表した論文「神経活動に内在する観念の論理的計算（A Logical Calculus of the Ideas Immanent in Nervous Activity）」が、知能の機械的実現に向けた最初の数理的基盤を提供したからである。

マッカロック＝ピッツモデルの核心は、神経細胞の活動を命題論理の演算として形式化した点にある。彼らのモデルでは、各ニューロンは二値的（発火するか、しないか）な状態をとり、十分な数の興奮性入力が同時に入った場合にのみ発火する。さらに、抑制性シナプスには「拒否権（veto）」が与えられ、単一の抑制性入力が活性化するだけでニューロンの発火を阻止できるとされた。

この単純なモデルがもつ理論的含意は深遠であった。マッカロックとピッツは、これらの形式ニューロンを適切に接続することで、多様な論理関数を実現できることを示した。重要なのは、ここで後世の意味での「万能な学習機械」が与えられたことではなく、神経活動を形式的な記号操作として扱いうることが初めて明瞭に提示された点である。この発想は二つの方向に射程を広げた。一つは神経科学の方向であり、脳の機能を計算理論として理解する道（後の計算論的神経科学）を開いた。もう一つは工学の方向であり、論理回路やオートマトン研究に接続される人工的ネットワークの構築可能性を示唆した。後者がニューラルネットワーク研究の系譜の最初の一歩となる。

ただし、マッカロック＝ピッツモデルには重大な限界があった。学習の仕組みが含まれていなかったのである。結合の重みと閾値は設計者が事前に決定するものであり、経験から自律的にパラメータを調整する機構は存在しなかった。この欠落は、後にローゼンブラットのパーセプトロン（第3章）やヘッブの学習則によって補われることになるが、「構造を与えれば計算ができる」という静的な証明と、「構造を経験から獲得する」という動的な問題との間には、AI研究が数十年にわたって格闘する本質的な溝が横たわっていた。

1.2　ウィーナーのサイバネティクスとフィードバック制御（1948）

マッカロック＝ピッツが神経活動の論理構造を明らかにしたのに対し、ノーバート・ウィーナーは知能のもう一つの本質的側面——目標指向的な行動の制御——に注目した。1948年に出版された『サイバネティクス——動物と機械における制御と通信（Cybernetics: Or Control and Communication in the Animal and the Machine）』は、生物と機械の双方に共通する制御原理としてフィードバックの概念を体系化し、後の人工知能研究に根本的な影響を与えた。

ウィーナーの着想の源泉には、第二次世界大戦中の対空射撃管制システムの研究があった。敵機の将来位置を予測するためには、レーダーから得られる情報を用いて逐次的に軌道の推定を修正する必要がある。この経験から、ウィーナーは「目的のある行動（purposeful behavior）」を、環境からのフィードバック情報に基づいて自己修正する過程として一般化した。1943年にはローゼンブルース、ウィーナー、ビゲローの共著で「行動、目的、目的論（Behavior, Purpose, and Teleology）」と題する論文が発表され、目的論的な行動を機械にも帰属しうるという大胆な主張がなされている。

サイバネティクスの意義は、学問分野の垣根を越えた統合的視座を提供した点にある。ウィーナーは、生物の恒常性維持（ホメオスタシス）、神経系の信号伝達、社会組織の自己調整を、すべてフィードバックループという共通の枠組みで理解しようとした。この学際性は1940年代後半から1950年代にかけてメイシー会議（Macy Conferences on Cybernetics, 1946–1953）という一連の学際的会合を生み、そこには神経科学者、数学者、人類学者、心理学者が集った。

サイバネティクスとAIの関係は複雑である。一方では、ウィーナーの思想はAI研究の知的土壌を準備した。フィードバックによる適応的行動という概念は、後の強化学習（第9章）の理論的先駆であり、制御理論はロボティクス（第12章）の基盤となった。他方、サイバネティクスは「AI」という名称が定着した後、次第に独立した運動として後景に退いていった。ダートマス会議（1956年）の主催者たちは、自分たちの新しい分野をサイバネティクスとは意識的に区別しようとした節がある。マッカーシーは後年、人工知能という語を選んだ理由の一つは「サイバネティクス」と距離を置くためだったと明言している。この分離は、初期AI研究における制度的・知的ポリティクスの一側面として記憶されるべきである。

1.3　シャノンの情報理論 —— 通信・計算・知能の数理的統合

1948年はAI前史における画期的な年であった。ウィーナーの『サイバネティクス』と同じ年に、クロード・シャノンはベル研究所の技術誌に「通信の数学的理論（A Mathematical Theory of Communication）」を発表し、情報という概念そのものを数学的に定義した。

シャノンの理論の核心は、情報を確率論的に定量化した点にある。情報量（エントロピー）は、メッセージの「驚き」の度合い——すなわち、そのメッセージがどれだけ予測困難であったか——として定義される。この定義は通信工学に革命をもたらしただけでなく、「知る」「学ぶ」「予測する」といった知的活動の本質を数理的に捉える枠組みを提供した。

シャノンのAIへの貢献は理論的基盤の提供にとどまらない。彼自身が直接的にAI研究に関与していた。1950年には「チェスをプレイするコンピュータのプログラミング（Programming a Computer for Playing Chess）」と題する論文を発表し、ゲーム木の探索戦略を体系的に論じた。この論文で提示された「タイプA戦略」（網羅的探索）と「タイプB戦略」（選択的探索）の区別は、後のゲームAI研究（第12章）の基本的枠組みとなる。また、ベル研究所ではシャノンは迷路を解くロボットマウス「テセウス（Theseus）」を構築している。これは電磁リレーを用いて試行錯誤的に迷路の解を記憶する装置であり、機械学習の最初期のデモンストレーションの一つとみなしうる。

さらに重要なのは、シャノンがダートマス会議の共同提案者であったことである。1955年の提案書には、マッカーシー、ミンスキー、ロチェスターとともにシャノンの名前が記されている。通信理論の創始者が人工知能の創設に関与したという事実は、AI研究が当初から「情報」という概念を中心に据えていたことを示している。シャノンの情報理論は、後にベイズ推論、最大エントロピー法、情報量基準によるモデル選択など、AI・機械学習の方法論に繰り返し現れることになる。

ただし、シャノンの情報理論には「意味」が含まれていないという根本的な制約がある。シャノン自身が明確に述べたように、彼の理論における「情報」は意味内容とは無関係であり、純粋に統計的な量である。この「意味の不在」は、後にAI研究が記号処理（意味をもつ表象の操作）と統計的学習（意味を持たないパターンの抽出）のどちらに重点を置くべきかという根本的論争の伏線となる。

1.4　チューリングテストと「計算する機械」（1950）

1950年、アラン・チューリングは哲学雑誌『マインド（Mind）』に「計算する機械と知能（Computing Machinery and Intelligence）」を発表した。この論文は「機械は考えることができるか？（Can machines think?）」という問いから始まるが、チューリングはこの問いそのものを回避する。「考える」という概念が曖昧で定義困難であるため、彼はこの問題を「模倣ゲーム（imitation game）」と呼ぶテストに置き換えることを提案した。

模倣ゲームの原形は三人のプレイヤーからなるパーティーゲームである。質問者Cが、別室にいるプレイヤーAとBに筆談で質問し、どちらが男性でどちらが女性かを判定しようとする。チューリングの革新は、プレイヤーAを機械に置き換えた場合にこのゲームの結果がどう変わるかを問うたことにある。もし質問者が機械と人間を安定的に区別できなくなるならば、その機械は「考えている」と見なしてよいのではないか——これがチューリングテストの骨子である。

チューリングはこの論文で、機械の思考に対する九つの反論を検討し、それぞれに対して再反論を展開している。神学的反論（魂をもたない機械は思考できない）、「頭を砂に突っ込む」反論（機械が考えるなどという結論は恐ろしいので考えたくない）、数学的反論（ゲーデルの不完全性定理による限界）、意識に基づく反論（機械は何も感じない）など、多岐にわたる反論を系統的に退けていく。

とりわけ注目すべきは、チューリングが「学習する機械」の可能性に言及した箇所である。彼は、あらかじめすべての知識をプログラムするのではなく、「子供の心」のような初期状態から教育によって知能を発達させる方法を提案した。この発想は、後の機械学習研究の根本理念を先取りするものであった。また、チューリングは「50年以内に、平均的な質問者が5分間の質問で正しく判定できる確率が70%を超えないほどうまく模倣ゲームをプレイするようコンピュータをプログラムすることが可能になるだろう」と予測している。

チューリングテストは、AI研究において最も頻繁に引用され、また最も激しく批判されてきた概念の一つである。ジョン・サールの「中国語の部屋」論証（1980年）はその代表的な批判であり、テストを通過する能力と真の理解は同一ではないと主張した。この論争は第24章で再び取り上げる。ここで重要なのは、チューリングの論文が「知能とは何か」という哲学的問いを、「知能をどうテストするか」という操作的問いに変換したことであり、この転換がAI研究に実証的・工学的性格を与える上で決定的な役割を果たしたことである。

1.5　ダートマス会議と「人工知能」の命名（1956）

1955年8月31日、ジョン・マッカーシー、マービン・ミンスキー、ナサニエル・ロチェスター、クロード・シャノンの四名は、ロックフェラー財団に対して「人工知能に関するダートマス夏季研究プロジェクト（Dartmouth Summer Research Project on Artificial Intelligence）」の提案書を提出した。この提案書が、「人工知能（Artificial Intelligence）」という用語の公式な初出とされる。

提案書には、次のような宣言が含まれていた。「学習のあらゆる側面、あるいは知能のあらゆる特徴は、原理的に、機械がそれをシミュレートできるほど正確に記述できるという推測に基づいて研究を進める」。この一文は、AI研究の基本的前提——知能は形式的に記述可能であり、したがって機械的に再現可能である——を簡潔に表現している。

1956年の夏、ダートマス大学で開かれたこの集まりは、提案書では「2か月・10名」の研究計画として構想されていたが、実際には参加者の出入りがあるローリングなワークショップとなった。したがって、後世にしばしば描かれるような、少数の創設者が一堂に会して統一綱領を打ち立てた場ではなかった。むしろ、マッカーシー、ミンスキー、シャノン、ロチェスターに加えて、サイモン、ニューウェル、ソロモノフ、サミュエルらがそれぞれ異なる研究アプローチを持ち寄り、並行的に議論した場として理解する方が実態に近い。

会議で特に強い印象を与えた成果としてしばしば言及されるのが、ハーバート・サイモンとアレン・ニューウェルによるLogic Theorist（論理理論家）である。これは、ホワイトヘッドとラッセルの『プリンキピア・マテマティカ』の定理を自動的に証明するプログラムであり、のちに初期AIの象徴的成果として記憶されることになる。Logic Theoristについては第2章で詳述する。

ダートマス会議の歴史的意義は、特定の科学的発見にあるのではなく、学問分野としてのAIを「創設」したことにある。会議後、AI研究の歴史的中心地が形成された。MITではミンスキーが、スタンフォードではマッカーシーが、カーネギーメロン大学ではサイモンとニューウェルが、それぞれ独自の研究プログラムを展開していくことになる。

「人工知能」という命名自体にも注意を払う必要がある。マッカーシーがこの用語を選んだ理由について、彼は後に複数の動機を述べている。一つはサイバネティクスやオートマトン理論といった既存の用語との差別化であり、もう一つはこの新しい分野の独立性を強調するためであった。しかし、「人工的な知能」という言葉は、当初から過大な期待を喚起する両刃の剣でもあった。この命名が後のAIブームとその反動（AIの冬、第5章・第8章）にどの程度寄与したかは、興味深い問いである。

1.6　初期の楽観 —— サイモン、ニューウェル、ミンスキーの展望

ダートマス会議の後、AI研究の創設者たちは驚くほど楽観的な予測を公にした。これらの予測を記録しておくことは、AI史を理解する上で重要である。なぜなら、楽観と幻滅の振幅こそが、AI研究の制度的・社会的ダイナミクスを規定してきたからである。

ハーバート・サイモンとアレン・ニューウェルは1957年から1958年にかけて、十年ほどでデジタルコンピュータがチェスの頂点に達し、重要な数学的定理の発見と証明にも到達するという趣旨の予測を示した。実際にチェスで人間世界王者級に到達するのは1997年のDeep Blueまで待たねばならず、当初の見通しは大幅に早すぎたことが分かる（第13章）。

ミンスキーも1960年代後半には、一世代のうちに知的能力の主要部分が機械の領域に入るという強い見通しを語っていた。ニューウェルとサイモンは1958年以降、彼らの「物理記号システム仮説」の原型となる主張を展開し始めていた。この仮説は、汎用的な知的行動に必要かつ十分な条件は物理的記号システムであるというものであり、1976年に正式に定式化される。

これらの楽観的予測は、単に研究者の個人的な見通しの誤りとして片付けるべきではない。背景には構造的な理由があった。第一に、初期のAIプログラム（Logic Theorist、General Problem Solver、第2章）が狭い領域で目覚ましい成果を示しており、これが領域の拡張によって汎用的知能に到達しうるという線形的な外挿を誘発した。第二に、コンピュータの処理速度が指数関数的に向上しつつあり、性能の限界は原理的なものではなく計算資源の問題に過ぎないという見方があった。第三に、知能の本質的な困難さ——常識推論、身体性、社会的文脈の理解——がどれほど深いかが、まだ十分に認識されていなかった。

この楽観は、研究資金の確保には有利に働いたが、長期的にはAI研究の信頼性を損なうことになる。もっとも、後に「AIの冬」（第5章・第8章）が到来した理由を、この楽観だけに還元することはできない。実際には、ARPAの優先順位の変化、機械翻訳やロボティクスへの失望、英国でのライトヒル報告（1973年）などが重なり、過大な約束と不十分な成果のギャップが制度的に可視化されたのである。AI史における楽観の構造を理解することは、2020年代の生成AIをめぐる期待を冷静に評価するためにも有益である。

1.7　AI研究の制度的基盤 —— MIT、スタンフォード、CMU

ダートマス会議の後、AI研究は三つの主要な拠点で制度化された。各拠点の知的特色と制度的条件を理解することは、その後のAI研究の分岐を理解する鍵となる。

MITでは、マービン・ミンスキーが1959年に人工知能プロジェクト（AI Project）を共同創設し、1963年にはARPA資金を背景にProject　MACが発足した。AI研究はその内部で大きく成長し、1970年にはMIT人工知能研究所（AI　　Lab）として独立する。MITのAI研究は、ミンスキーの知的多様性を反映して、知識表現、ロボティクス、自然言語理解、コンピュータビジョンなど幅広い領域にまたがっていた。また、マッカーシーがMIT在籍中（1958年）に開発したプログラミング言語LISPは、その後数十年にわたりAI研究の標準言語となる。

スタンフォードでは、マッカーシーが1962年に復帰したのちAI研究の拠点形成を進め、1963年頃から研究体制が整い、1965年までにスタンフォード人工知能研究所（SAIL）が制度的に定着した。SAILは当初、キャンパスから離れたスタンフォードの丘陵地に位置し、独自の研究文化を育んだ。マッカーシーの関心は形式論理に基づく推論システムにあり、「状況計算（situation calculus）」をはじめとする知識表現の形式的手法を発展させた。また、SAILは1960年代後半からロボティクスとコンピュータビジョンの研究でも先駆的な成果を出している。

カーネギーメロン大学（CMU）では、サイモンとニューウェルが情報処理心理学と人工知能を架橋する研究プログラムを推進した。彼らのアプローチの特色は、人間の問題解決過程の心理学的研究とコンピュータによるシミュレーションを密接に結びつけた点にある。General Problem Solver（第2章）はその代表的成果であった。CMUのアプローチは、AIを純粋に工学的な問題として捉えるのではなく、人間の認知のモデルとして位置づけるという知的伝統を確立した。

これら三つの拠点に共通するのは、DARPA（および前身のARPA）からの潤沢な研究資金に支えられていたという点である。冷戦期の米国において、AI研究は国防上の潜在的重要性を認められ、比較的自由度の高い資金が提供された。この資金構造は1960年代のAI研究の急速な発展を可能にしたが、同時に、研究成果が軍事的期待に応えられなかった場合に資金が急減するという脆弱性をも内包していた。第5章で論じるライトヒル報告（1973年）に端を発する第一次AIの冬は、まさにこの構造的脆弱性が顕在化した事例である。

また、この時期のAI研究が米国の三大学に集中していたことは、初期AI研究の地理的・文化的偏りを反映している。ただし、米国だけが唯一の舞台だったわけではない。英国ではエディンバラ大学で1963年からドナルド・ミッキーらがAIグループを形成し、ソビエト連邦でも1950年代末からサイバネティクスと計算機科学の制度化が進んでいた。したがって、より正確には、1950年代末から1960年代前半にかけて研究資金・制度・国際的可視性の三点で米国が主導権を握っていた、と述べるべきである。

> 本章では、人工知能という学問分野が成立するまでの思想的・制度的基盤を概観した。マッカロック＝ピッツの形式ニューロン（1943）、ウィーナーのサイバネティクス（1948）、シャノンの情報理論（1948）、チューリングテスト（1950）、そしてダートマス会議（1956）という一連の知的事件は、知能の機械的実現という構想を学術的に正当化し、制度的に確立するための必要条件を整えた。次章では、この基盤の上に最初に構築されたAIプログラム群——Logic　Theorist、General Problem Solver、そしてLISP——を検討する。

参考資料（本章）

本文中の主要な記述を追跡するため、最小限の典拠を挙げる。

- Warren S. McCulloch and Walter Pitts, “A Logical Calculus of the Ideas Immanent in Nervous Activity” (1943).
- Arturo Rosenblueth, Norbert Wiener, Julian Bigelow, “Behavior, Purpose and Teleology” (1943); Norbert Wiener, Cybernetics (1948).
- Claude E. Shannon, “A Mathematical Theory of Communication” (1948); “Programming a Computer for Playing Chess” (1950); Nokia Bell Labs History.
- Alan M. Turing, “Computing Machinery and Intelligence” (1950); Stanford Encyclopedia of Philosophy, “The Turing Test”.
- John McCarthy, Marvin Minsky, Nathaniel Rochester, Claude Shannon, “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence” (1955).
- John McCarthy, “The Question of Artificial Intelligence” 所収書評ページ。人工知能という語をサイバネティクスから区別した理由を回顧している。
- Nils J. Nilsson, The Quest for Artificial Intelligence (2010). ダートマス会議の参加実態と初期AIの制度史の整理に有用。
- MIT News, “MIT’s computer science, AI labs merge” (2003). MIT AI Project, Project MAC, AI Lab の年表確認に使用。

- Stanford AI Lab 公式サイト、Stanford University History、John McCarthy memorial page. SAIL の成立時期と制度化の経緯の確認に使用。
- University of Edinburgh, “History of Artificial Intelligence at Edinburgh”. 英国における初期AI研究の制度化の確認に使用。
- Edward A. Feigenbaum, “Soviet cybernetics and computer sciences, 1960” (Communications of the ACM, 1961). ソ連圏における同時期の研究動向の確認に使用。

第2章　初期のAIプログラムと探索的アプローチ

2.1　Logic Theorist と General Problem Solver

ダートマス会議の前後、AI研究の最初の実質的な成果がハーバート・サイモン、アレン・ニューウェル、そしてジョン・クリフォード・ショーによってもたらされた。1955年から1956年にかけて、彼らがランド・コーポレーション（RAND Corporation）で開発したLogic Theoristは、人間の問題解決能力を意図的に模倣しようとした最初期の代表的プログラムとなった。

Logic Theoristの本質は、シンボリック論理における定理証明を機械で実行可能な探索問題として定式化したところにある。このプログラムは、ホワイトヘッドとラッセルの『プリンキピア・マテマティカ』第2章の最初の52個の定理のうち38個を証明することに成功したと報告された。注目すべきは、定理2.85として広く言及される証明例で、プログラムが開発者の予期を超え、より簡潔な証明を発見したという点である。この出来事は、機械が高次の記号操作に入りうるという印象をAI研究初期に強く与えた。

けれども、Logic Theoristの設計には根本的な制約があった。それは「扱える知識」と「それを処理する方法」が堅く結合していた点である。証明に必要な論理規則は事前に組み込まれており、新しい領域へ適用するたびにプログラムの内部構造を修正する必要があった。言い換えれば、「論理」という一般的な知識領域であってさえ、具体的な問題領域ごとに異なる実装が求められたのである。この観察がもたらしたのが、一般的な問題解決メカニズムの模索である。

1957年から1959年にかけて、ニューウェルとサイモンは**General Problem Solver（GPS）**を開発した。GPSの革新性は、「問題領域の知識」と「問題解決の戦略」を分離したところにある。これはプログラミングの実装パターンとしても、AI設計の哲学としても大きな進歩であった。

GPSが採用した戦略は「手段―目的分析（means-ends analysis）」と呼ばれるものである。このアルゴリズムは、現在の状態と目標状態との差分を認識し、その差分を段階的に縮小していく過程として問題解決を定式化した。たとえば、チェスから幾何学定理証明まで、異なる領域の問題に対しても同じ基本戦略を適用することが原理的に可能であった。ニューウェルとサイモンはこの成功から、「汎用的な知的行動は物理的記号システム（physical symbol system）」で十分に説明可能であると主張する「物理記号システム仮説（Physical Symbol System Hypothesis）」を段階的に発展させていった。

しかし、GPSもまた試験的な範囲を超えることはなかった。このプログラムが取り扱える問題は「よく定義された問題（well-defined problem）」に限定されていた。すなわち、初期状態と目標状態が明確であり、許可される操作と状態遷移がはっきり規定されている領域である。実世界の問題のほとんどは、このような形式化に抵抗する。現実の医療診断、自然言語理解、知覚や運動といった領域では、問題その

ものが曖昧であり、目標も複数の競合する価値の間で平衡を取る必要があった。GPSの手段―目的分析は、こうした「曖昧な領域」には適用困難であった。これは単なる技術的な欠陥ではなく、AI研究が数十年にわたって格闘することになる根本的な制約を示唆していた。

2.2　サミュエルのチェッカープログラムと機械学習の萌芽

Logic Theoristと同時代、まったく異なるアプローチからAI研究に接近しようとする者がいた。アーサー・サミュエルは、IBM 701上でチェッカー（英語では checkers、英国英語では draughts）を指すプログラムを1952年に開発し、1950年代半ばには学習機構を備えた版へ発展させた。1956年のテレビ実演は広く注目を集め、ゲームを通じた機械学習の可能性を可視化した。

サミュエルのアプローチの根本的な差異は、「プログラマが明示的に知識をコード化する」という初期AIの原則を相対化したところにある。代わりに、彼は自己対局と「rote　learning」を組み合わせる方法を採用した。今日の語彙で言えば、これは自己対局型の強化学習の先駆とみなすことができる。プログラムは何千回もの自己対局を通じて、ゲーム局面を評価するための関数を段階的に改善していった。この関数は、盤上の駒の数、キング昇格の数、駒の配置など複数の要素を組み合わせて、その局面での優劣を推定するものであった。

サミュエルは、ゲーム木探索の効率を高めるため、後に**アルファ・ベータ枝刈り（alpha-beta pruning）**と呼ばれることになる探索削減技術を実装した。すべての手を末端まで評価するのではなく、盤面評価関数に基づいて有望な手のみを深く追求する戦略である。この工学的洗練により、チェッカーという比較的閉じたゲーム領域において、プログラムは人間の有力アマチュア級に近い実力に到達した。1962年、プログラムはロバート・ニーリーとの対戦で勝利を収めている。

サミュエルの仕事の理論的意義は、機械学習という観念を具体化し、それが実際に動作することを実証した点にある。彼は1959年に「機械学習（machine learning）」という用語を定着させ、後の機械学習研究の知的伝統を確立した。しかし同時に、彼の業績は初期AIコミュニティにおいて相対的に軽視される傾向があった。なぜか。それは、Logic TheoristやGPSのような「記号推論」のアプローチと比べ、サミュエルの方法論は、より「経験的」で「統計的」に見えたからである。1950年代のAI哲学は、人間の知能を形式的推論の能力に求める傾向が強かった。学習や経験への依存は、むしろ不完全で粗野な近似として見なされることすらあった。

その結果、初期AI研究の主流は記号処理とシンボリックな推論へと偏向していく。サミュエルの先見性――統計的学習こそが知能の本質ではないか、という問い――が本当に評価されるのは、1990年代の統計的機械学習革命（第9章）を待たねばならなかった。初期AI研究においては、複数のパラダイムが競合していたが、その中で「どのパラダイムが正当性を持つか」を決めるのは、単なる科学的論証ではなく、研究資金、制度的サポート、そして各分野の知的指導者たちの好みであった。

2.3　LISP言語の誕生とAI向けプログラミング環境

初期AIが成功するには、単に良いアルゴリズムがあるだけでは不十分であった。記号的知識を表現し、それを操作するためのプログラミング言語が必要であった。この必要性に応答したのが、ジョン・マッカーシーが1958年にMITで構想し、1960年の論文で公表した**LISP（List Processing）**である。

LISPの発想は、数学の関数的形式から生まれた。マッカーシーは、部分再帰関数と記号式の操作を組み合わせることで、「シンボリック表現」を計算対象とすることができると考えた。ラムダ計算に由来する関数の抽象化と適用という概念を導入することで、高度な抽象化を行いながらもAI問題を記述できる言語が実現された。とりわけ重要だったのは、**リスト構造（list　structure）を言語の基本データ型とした点である。プログラムもデータも同じリスト形式で表現されるというこの特性が、後のメタプログラミング**、すなわちプログラムがプログラムを生成・操作する自己言及的な計算を可能にした。

LISPは迅速にAIコミュニティの標準言語となった。1960年の発表以降、1980年代にかけてLISPなくしてはAI研究が成立しないほどの地位を確立した。とりわけDECのPDP-10系計算機上でのLISP実装は、AI研究者たちに相対的に扱いやすい計算環境を提供した。MITやスタンフォードのAI研究所では、LISPを基盤として、より高レベルの表現言語や推論システムが次々と構築されていった。

しかし、LISPの言語設計には、後年の視点からすれば問題もあった。シンボル処理に最適化されたLISPは、数値計算に関しては非効率であり、また言語そのものが柔軟すぎるがゆえに、大規模なプログラムの保守が困難であった。1980年代のLISPマシン産業の衰退（第7章）は、単にハードウェアの経済的な競争敗北だけではなく、ソフトウェア開発の管理可能性の問題も含んでいたのである。

2.4　幾何学定理証明機と記号推論

Logic　Theoristが一般的な論理定理の証明を目指したのに対し、より特定の領域に焦点を絞った記号推論システムも同時期に発展していた。その代表が、**幾何学定理証明機（Geometry　Theorem　Prover）**である。

ハーバート・ゲレンター（Herbert　Gelernter）がIBM　704上で開発したこのシステムは、1959年の春に初めて初等ユークリッド幾何学における定理を完全に自動で証明することに成功した。約20,000個の命令から構成されるこのプログラムは、Logic　Theoristよりも高い成功率を示し、複数の定理について段階的な証明を生成できた。

ゲレンターのアプローチで特筆すべきは、**数値的図式（numerical　diagram）を証明プロセスに統合したことである。幾何学的推論は本質的に、「点Aがこの直線の左にある」といった空間的関係の理解に依存している。ゲレンターは、各定理について数値的な座標を付与した図式を構築し、それを「フィル**

タ」として機能させた。すなわち、論理的に導出された中間目標が、この数値図式と矛盾していないかを検証するのである。矛盾する目標は、証明木の早期に刈り込まれ、計算資源が浪費されるのを防ぐ。この手法は、現代的な語彙ではセマンティック・フィルタリングと呼びうる概念の先駆であった。

幾何学定理証明機のもう一つの創意は、**補助点の自動生成（automatic construction of auxiliary points）**である。多くの非自明な幾何学的証明は、元の定理には明示的には現れない追加的な点や線を導入することで成り立つ。人間の数学者は直感的にこのような補助線を引くが、機械がこれを自動で行うことは自明ではない。ゲレンターのシステムは、証明に詰まったとき、系統的に補助点を生成し、それが新しい定理を導くかどうかを探索した。

これらの工学的洗練にもかかわらず、幾何学定理証明機はやはり限定的な領域に留まった。ユークリッド幾何学は、数学の中でもきわめて形式化しやすい領域である。述語論理による完全な記号化が可能であり、証明規則も明確に定義されている。しかし、この限られた成功ですら、より広い数学分野や現実世界の知識推論に直結することはなかった。この観察は、AI研究において繰り返し現れる教訓を示唆する。すなわち、「狭い領域での形式的完全性」と「広い領域での有用性」の間には、単なる量的な拡張では埋められない本質的な溝が存在するということである。

2.5　初期の自然言語処理 —— 機械翻訳の夢と挫折

同じ1950年代から1960年代にかけて、全く異なる知的野心がAIコミュニティを駆り立てていた。それが**機械翻訳（machine translation）**である。人類が望遠鏡や顕微鏡を発明して物理世界を拡張したように、コンピュータは言語の壁を消去する翻訳機械として機能しないだろうか。この問いは、特に冷戦期の米国にとって戦略的重要性を持っていた。ソビエト連邦の科学・技術文献を迅速に英語に翻訳することは、科学的競争力に直結していたのである。

初期の機械翻訳の象徴的事件は、**ジョージタウン・IBM実験（Georgetown-IBM experiment）**である。1954年1月7日、ジョージタウン大学とIBMが共同で、ロシア語を英語へ自動翻訳するシステムを公開デモンストレーションした。60以上のロシア語文が完全に自動で英語に翻訳されたのである。このシステムは、わずか250語彙と6個の文法規則しか持たなかったにもかかわらず、新聞記者やコンピュータ科学者の想像力を強く刺激した。実験の報告者たちは「3年から5年以内に機械翻訳は解決される」と楽観的に述べた。

この楽観は、単なる技術者の過信ではなく、当時のAI理解の反映であった。もし知能が本質的に「シンボルの操作」に帰着するのなら、言語も結局は一種のシンボルシステムであり、規則に基づいた変換で十分なはずだと考えられたのである。文法規則を詳細に記述し、辞書を充実させれば、翻訳品質も線形に向上するであろうと予想されたのである。

1960年代を通じて、米国とソビエト連邦の両国で大規模な機械翻訳プロジェクトが展開された。しかし、現実はこの楽観を裏切った。翻訳品質は期待ほど向上せず、新しい文法規則や語彙を追加しても、改善幅は次第に縮小していった。より深刻な問題は、機械翻訳の本質的な困難が、単なる「コンピュータパワー」や「データの量」では解決できないということが徐々に認識されたことである。

たとえば、“The pen is in the box“という単純な英文を考える。これを「ペンは箱の中にある」と訳すか、「箱の中にペンがある」と訳すか、あるいは「ペンが箱に入れられている」と訳すか――言語の選択肢は文法だけでは決定できない。それは、ペンが何か、箱が何か、そしてこの文が何を伝達しようとしているのかについての、より広い知識や文脈に依存している。**常識知識（commonsense　knowledge）**の欠落こそが、機械翻訳の根本的な障壁であることが、次第に認識されていった。

この認識は、1966年の**ALPAC報告書（Automatic Language Processing Advisory Committee Report）**という形で、制度的結論に至った。ジョン・R・ピアスを委員長とするこの委員会は、米国政府から機械翻訳研究の進捗を評価するよう求められ、十年間の大規模投資の後、厳しい結論を下した。報告書は、当時の機械翻訳は実用的有用性に乏しく、「近い将来に有用な機械翻訳が実現する予見可能な見通しはない」と結論づけたのである。

ALPAC報告書の影響は計り知れない。これは単なる学術的評価ではなく、米国政府による資金配分の再判断を正当化する文書となり、機械翻訳研究への連邦支援を大幅に縮小させた。以後かなり長い期間、米国国内で機械翻訳研究は主流の座から退き、この分野に関心を示すこと自体が研究者のキャリア上のリスクと見なされる局面すら生まれた。

しかし、この挫折の中にも、AI研究の深刻な自己反省が隠されていた。機械翻訳の失敗は、単なる「言語処理の問題」ではなく、知能そのものの本質についての理解の不足を明らかにしたのである。人間の言語理解は、テキストの表面的な構造だけでは成り立たない。それは、物理世界についての常識的知識、社会的文脈、発話者の意図、そして言語外のパラダイムまで含む、広大な知識ネットワークに根ざしている。この認識は、第5章で論じる「第一次AIの冬」を導くことになる出発点であると同時に、後年のより謙虚で精密なAI研究へ向かう転換点でもあった。

実は、機械翻訳への関心が完全に消えたわけではなかった。ソビエト連邦ではプロジェクトが低調ながら継続されたし、米国内でもウェイン州立大学やテキサス大学での研究が1970年代まで細々と続いていた。しかし、主流の学術コミュニティからの後退は明確であり、その結果、機械翻訳研究は「失敗した分野」という烙印を押されることになった。この事実は、AI史における資金と期待の相互作用の重要性を教える。すなわち、初期の過剰な楽観が、後年の過度な悲観を招き、その結果として有望な領域が不当に放棄されるという悪循環である。

並行して、別のアプローチから自然言語処理へ接近する者もいた。1966年、MITのジョセフ・ワイゼンバウムは、**ELIZA**と呼ぶプログラムを発表した。これは、テキスト入力に対してパターンマッチングと置換規則を適用することで、ロージャン派の心理療法士を模倣するものであった。ELIZAの回答は表面的には巧妙であり、多くの利用者が感情的な深さがあると感じた。ワイゼンバウムの秘書すら、ワイゼンバウムに部屋を出るよう求め、ELIZAと「本当の会話」をしたいと述べたという逸話が残されている。

しかし、ワイゼンバウム自身は、このプログラムの成功に深刻な哲学的懸念を抱いた。人間がこれほど簡単に機械的なテキスト処理に騙されるのであれば、人間の理解とは何か？機械が人間との対話を「模倣する」ことと、真に「理解する」ことの間には、何が横たわっているのか？ワイゼンバウムはその後、人工知能批判の最も説得力ある声の一人となる。彼の1976年の著作『コンピュータと人間の知性（Computer Power and Human Reason）』は、AI研究に対する知的批判の古典として記憶されることになる。ELIZAの一見の成功と、それが露呈した人間理解の深刻な限界は、初期AI研究の矛盾を象徴していたのである。

まとめ：初期AIの遺産と限界の自覚

本章で検討したLogic Theorist、General Problem Solver、サミュエルのチェッカープログラム、LISP言語、幾何学定理証明機、そして機械翻訳の試みは、1950年代から1960年代初頭のAI研究の知的景観を描き出している。

これらの初期プログラムに共通するのは、「形式化可能な領域では、機械は人間に匹敵する、あるいは超越する知識処理を行える」という信念である。Logic Theoristの優雅な定理証明、GPSの汎用的な問題解決戦略、サミュエルの学習プログラムは、その信念が空疎なものではなく、実際に計算可能な成果を生むことを示した。

しかし同時に、これらの取り組みの相対的な狭隘さもまた明らかになった。Logic Theoristは論理学にしか適用できず、GPSは「よく定義された問題」に限定され、チェッカープログラムはボードゲームという単一領域に留まった。機械翻訳の挫折は、言語のような形式化困難な領域では、単なるシンボル処理では不十分であることを示唆していた。ELIZAの「成功」は、逆説的に、人間がいかに容易に機械的な反応に応答するかを露呈し、真の「理解」とは何かの問いを引き出した。

第I部の次の章（第3章）では、これらの記号主義的アプローチと対立する別のパラダイムが台頭してくることになる。ローゼンブラットのパーセプトロン（1958）はニューラルネットワークという全く異なる計算原理に基づき、知能を「学習される構造」として捉え直した。初期AI研究の黄金期は、実は複数の知的方向が競合し、その中での主導権争いの時期でもあったのである。この競合の結果として、1960年代後半から1970年代にかけて、AI研究全体が深刻な危機に直面することになる。その危機は単なる技術的な困難ではなく、知能とは何かについての根本的な理解の再考を強いるものになるだろう。

参考資料（本章）

本文中の主要な記述を追跡するため、最小限の典拠を挙げる。

- Allen Newell and Herbert A. Simon, “The Logic Theory Machine” (1956).
- Nils J. Nilsson, The Quest for Artificial Intelligence (2009). Logic Theorist、GPS、機械翻訳初期史の整理に有用。
- Arthur L. Samuel, “Some Studies in Machine Learning Using the Game of Checkers” (IBM Journal of Research and Development, 1959).
- IBM History, “The games that helped AI evolve”. サミュエルのチェッカープログラムと1962年の対戦史の確認に使用。
- John McCarthy, “Recursive Functions of Symbolic Expressions and Their Computation by Machine, Part I” (1960); John McCarthy, “History of Lisp”.
- Herbert Gelernter, J.R. Hansen, D.W. Loveland, “Empirical Explorations of the Geometry Theorem Machine” (1960).
- Georgetown-IBM実験の記録および National Research Council, Language and Machines: Computers in Translation and Linguistics (1966). ALPAC報告の確認に使用。

第3章　パーセプトロンとニューラルネットワークの最初の波

3.1　ローゼンブラットのパーセプトロン（1958）

1950年代半ばまで、知能の機械化をめぐる議論は二つの相互補完的な流れを示していた。一つは第1章で述べたマッカロック＝ピッツの静的な形式ニューロンモデルであり、もう一つはウィーナーのサイバネティクスやシャノンの情報理論といった動的な制御・学習の枠組みであった。しかし、これらのアプローチには共通の欠落があった。形式ニューロンモデルはネットワークの構造を所与として、その計算能力を論じるものであり、経験からパラメータを自律的に調整する学習メカニズムを欠いていたのである。ローゼンブラットのパーセプトロンは、この欠落を埋めるための最初の本格的な試みであった。

フランク・ローゼンブラットは、コーネル航空研究所（Cornell Aeronautical Laboratory）に属する心理学者・工学者であった。1957年、彼はIBM 704上でパーセプトロンをシミュレートし、翌1958年に「脳における情報の保存と組織化の確率モデル（The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain）」と題する論文を『心理学評論』に発表した。この論文は、単なる技術的提案ではなく、学習する機械の理論的基盤を示そうとするものであった。

パーセプトロンの構造は一見単純である。入力層（sensory units）と呼ばれる多数のニューロンからの信号を受け取り、これらの信号の重み付き和が一定の閾値を超えた場合に出力ニューロンが発火するというものである。しかし、その真の革新は学習ルールにあった。ローゼンブラットはドナルド・ヘッブの神経可塑性に関する理論に触発され、「神経が協働して発火すれば、その結合は強化される」というヘッブの原則を機械的に実装した。パーセプトロンの学習アルゴリズムは、誤分類された事例に対して、その方向に重みを調整するという単純な規則であった。ここで重要なのは、このアルゴリズムが「教師信号（teacher　signal）」に基づいており、正しい分類と実際の出力との誤差が学習を駆動するという点である。

1960年6月23日、ローゼンブラットはニューヨークのコーネル航空研究所で、自分が構築した物理的機械「Mark Ⅰ パーセプトロン」の公開デモンストレーションを行った。この機械は米国海軍研究局と空軍ローム開発センターからの資金を得て製作されたもので、400個の光電池を20×20格子状に配置した感覚ユニット、中間の連合ユニット群、そして調整可能な結合荷重を持つ出力部から構成されていた。Mark Ⅰの印象的な性質は、その学習能力にあった。当時の報道では、この機械は試行錯誤を通じて新しい技能を獲得する最初期の装置として大きく宣伝された。Mark I パーセプトロンは現在、スミソニアン米国歴史博物館に所蔵されている。

しかし、ローゼンブラットの理論的な業績は、この物理的な実装よりも深刻な含意を持っていた。彼は「パーセプトロン収束定理」と呼ばれるものを証明した（正確には、後に数学者アルバート・ノビコフが1962年に厳密な証明を与えたが、ローゼンブラットの直感的な理解が先行していた）。この定理が述べるところは以下の通りである。データが線形分離可能（すなわち、超平面によって完全に分離可能）であるならば、パーセプトロンの学習アルゴリズムは有限のステップで収束し、すべての訓練例を正しく分類する解に到達するということである。

この定理の意義は見かけ以上に深い。それは、学習という営みが原理的に可能であること、そして単純な局所的な規則（誤差に基づいた重み調整）の繰り返しが、全体的な問題解決へと導くことを証明していた。これは哲学的には重要であった。なぜなら、学習が計算可能な過程として形式化されたからである。また実務的には、この定理は単層パーセプトロンに対する理論的な信頼を与え、より複雑なネットワーク構造への拡張への道を開いた。

ローゼンブラットは自らの発見に対して楽観的であった。1958年の新聞報道では、数年以内にパーセプトロン型の機械が人間並みの認識能力へ近づくかのような見通しが語られた。この予測は、技術的な限界と理論的な限界の双方について、十分な認識を欠いていた。パーセプトロンが扱える問題は、実は厳格に制限されていたのである。線形分離可能な問題に限定されるというこの制約は、当初は十分に共有されていなかった。その理由は、線形分離不可能な問題がいかに多く実世界に存在するかが、まだよく理解されていなかったからである。

3.2　適応線形素子（ADALINE）とウィドロウ＝ホフ学習則

ローゼンブラットのパーセプトロンとほぼ同時期に、スタンフォード大学ではバーナード・ウィドロウと彼の博士課程の学生マルシアン・ホフが、異なるアプローチによるニューラルネットワーク学習の問題に取り組んでいた。1960年、彼らは適応線形素子（Adaptive Linear Element, ADALINE）と呼ぶシステムを開発し、これに伴う「ウィドロウ＝ホフ学習則」を提案した。

ADALINE は外観上はパーセプトロンに似ているが、その学習メカニズムは根本的に異なっていた。パーセプトロンでは重みの調整は出力層での誤分類（離散的なエラー）に基づいていた。一方、ADALINE では重みの調整は「最小平均二乗誤差（Least Mean Squares, LMS）」に基づいていた。すなわち、活性化関数を適用する前の線形出力と教師信号との間の二乗誤差を最小化する方向に、連続的に重みを調整するのである。

この違いは技術的な詳細に見えるかもしれないが、実は深い理論的含意を持っていた。LMS アルゴリズムは勾配降下法の特殊な場合であり、凸最適化の理論に基づいている。ウィドロウとホフが採用したアプローチは、学習を「最小化可能な目的関数の勾配に沿った移動」として形式化したのである。この枠組みは後に機械学習の主流となる統計的最適化の論理の先駆けであった。

ADALINE　の実用的な意義は、その理論的な優雅さ以上に、実装可能性にあった。ウィドロウとホフは、ウィドロウ＝ホフ学習則を使用して、適応フィルタリング、特に電話線のノイズキャンセレーション問題に取り組み、実際の成功を収めた。この応用は現在まで続いており、デジタル信号処理との結合において ADALINE の理論は今日でも使用されている。これは、単なる学術的興味の対象ではなく、実際の工学的問題を解決できるツールとしてニューラルネットワークが機能することを示した最初の事例の一つであった。

ウィドロウとホフのアプローチはまた、別の重要な含意を持っていた。LMS　ルールは「オンライン学習」に適していた。すなわち、データが一つずつ到着する環境でも、その都度重みを更新できるのである。このオンライン学習の特性は、後に強化学習（第9章）の理論的基盤となる概念であり、適応制御の枠組みとも親和的であった。

しかし、ADALINE が達成した実用的な成功にもかかわらず、その理論的な影響力はパーセプトロンほどではなかった。理由は複数ある。第一に、ADALINE もまた単層の線形モデルであり、表現力の限界をパーセプトロンと共有していたこと。第二に、その時代の計算機環境では、ADALINE の導入とシミュレーションにかかるコストが高かったこと。第三に、ミンスキーとパパートが神経ネットワークの限界を論じる際、代表例としてパーセプトロンを選んだため、批判の象徴性がそちらに集中したことである。

それでもなお、ウィドロウ＝ホフ学習則の理論的重要性は過小評価されるべきではない。それは、ニューラルネットワークの学習を最適化問題として形式化した最初の体系的な試みの一つであり、後の機械学習理論に与えた影響は深刻である。特に、確率的勾配降下法（Stochastic　Gradient　Descent）という、現在のディープラーニングを支える基本的なアルゴリズムは、本質的にウィドロウ＝ホフ学習則の拡張に他ならない。この歴史的な系譜は、1950年代後半から1960年代にかけてのニューラルネットワーク研究が、統計的最適化の論理を早期から統合していたことを示している。

3.3　ミンスキー＝パパート『Perceptrons』の衝撃（1969）

1969年5月、MIT　の人工知能研究所から一冊の書籍が出版された。『Perceptrons:　An　Introduction　to　Computational　Geometry』（『パーセプトロン――計算幾何学への導入』）と題するこの書は、マービン・ミンスキーとシーモア・パパートによる共著であった。この著作は、ニューラルネットワーク研究に歴史的な転機をもたらすことになる。

ミンスキーとパパートがこの著作に取り組み始めたのは1963年頃のことであった。当初は簡潔な論文になるはずであったが、彼らが書きながら予期しない数学的困難に直面したため、執筆は6年にわたり、書籍という形で完成することになった。著作の中心的な主張は、次のようなものであった。パーセプトロン――特に単層パーセプトロン――は、本質的に限定された計算能力を持つものであり、比較的単純な述語関数さえも計算できない場合があるということである。

ミンスキーとパパートの批判を今日もっとも直感的に説明する例として広く使われるのが「排他的論理和（Exclusive OR, XOR）問題」である。XOR は入力が二つの二値変数（0 または 1）を取る場合、一方が 1 で他方が 0 のとき出力が 1、両方とも 1 または両方とも 0 のとき出力が 0 となる論理関数である。データを座標平面にプロットすれば、点（0, 0）と（1, 1）が一つのクラスを形成し、点（0, 1）と（1, 0）が別のクラスを形成する。単層パーセプトロンが学習できる関数は、データを直線で分離する線形判別関数に限定される。XOR に対しては、どのような直線を引いても二つのクラスを完全に分離できない。

ミンスキーとパパートの著作でより中心的に扱われたのは、XOR　　だけでなく、より一般的な「パリティ」や「連結性（connectedness）」のような性質であった。彼らは、単層パーセプトロンがこうした述語を計算するうえで本質的な制約を持つことを示した。彼らの論証の数学的厳密性は高く、その時点での多くの数学者にとって説得力があった。著作の後半では、単層パーセプトロンの理論的限界を超えるには多層構造が必要であることも示唆されていた。しかし、多層パーセプトロンを現実に学習させる効率的手法は、まだ確立していなかった。

『Perceptrons』が与えた社会的・制度的な影響は、その理論的内容の重要性よりも、むしろ大きく受け取られたと考えられる。当時、パーセプトロン研究に対する資金は主に米国防高等研究計画局（ARPA/DARPA）などから供給されていた。1960年代末から1970年代にかけて、ニューラルネットワーク研究への支援は縮小していくが、その因果を『Perceptrons』一冊だけに還元するのは不正確である。計算資源の不足、記号主義的AIへの制度的傾斜、そして多層学習法の未成熟が重なっていたからである。その後の10年前後、ニューラルネットワーク研究は主流から外れ、多くの研究者がこの領域から離れた。

ここで重要な歴史的な再評価が必要である。『Perceptrons』がニューラルネットワーク研究の凋落の「原因」であったのか、それとも「象徴」であったのか、という問いである。実際のところ、1960年代後半までに、パーセプトロン研究は内発的な限界に直面していた。ローゼンブラットは1971年に亡くなったが、彼の存命中でさえ、パーセプトロン研究に対する熱意は減少していた。また、記号主義的な AI　　アプローチ（ニューウェル・サイモンの一般問題解決機、マッカーシーの形式論理に基づくシステム）が、1960年代前半には既に DARPA の関心を集めており、これがニューラルネットワーク研究への関心を競合的に圧迫していた。

その上、パーセプトロン研究そのものの内部にも、解決困難な問題が累積していた。線形分離可能性の制約は原理的なものであり、より複雑な問題に適用するには多層ネットワークが必要であることは既に認識されていた。しかし、多層ネットワークの効率的な学習アルゴリズムは存在しなかった。誤差逆伝播法（backpropagation）は、パウル・ウェルボスによって1974年に提案されたが、その重要性が広く認識されるまでには、さらに12年を要した。つまり、『Perceptrons』が出版された1969年の時点では、多層パーセプトロンの学習問題を解決する道は理論的に開かれていたが、実践的には実装されていなかったのである。

しかも、ミンスキーとパパート自身も、『Perceptrons』が与える印象ほど、ニューラルネットワークの将来に対して絶望的ではなかった。ミンスキーは後年、自分たちは単層パーセプトロンの限界を明確にすることが目的であり、多層ネットワークの研究を進めることを排除しようとしたのではないと述べている。しかし、組織的・制度的なレベルでは、『Perceptrons』は「ニューラルネットワークは終わった」というシグナルとして機能した。DARPA　の資金配分者、大学の管理者、そして多くの研究者自身が、このシグナルを受け取った。これは、学術的な議論と制度的現実の間に存在する非対称性を典型的に示す事例である。

さらに注目すべきは、ミンスキーとパパートが1969年に提示した多層パーセプトロンに対する批判が、必ずしも現代の深層学習には適用されないという点である。『Perceptrons』での彼らの議論は、パーセプトロンの重みが「局所的な」ランダムな接続を持つ場合を前提としていた。つまり、各ニューロンが受け取る入力が限定的であり、入力間の相互作用が弱い場合を想定していたのである。しかし、現代のニューラルネットワークでは、重みは勾配降下法によって大域的に最適化されており、局所的な構造の制約はない。この点をミンスキーとパパート自身が十分に認識していたかどうかは、歴史的に議論の余地がある。1988年に『Perceptrons』の拡張版が出版された際、ミンスキーとパパートは自分たちの議論が依然として妥当性を持つと主張したが、その主張の説得力は、多くの神経網研究者の間で限定的であった。

3.4　コネクショニズムの一時的後退

『Perceptrons』の出版は、より広大な知的・制度的転換の中で生じた。1960年代から1970年代への転換期に、人工知能研究は根本的なパラダイムシフトを経験した。その中心は、学習可能なニューラルネットワークから、明示的な知識表現と論理的推論に基づくシステムへの移行であった。この転換は、単に一冊の本の影響の結果ではなく、より深い構造的理由を持っていた。

第一に、計算資源の問題がある。1950年代から1960年代にかけてのコンピュータの性能向上は指数関数的であったが、それでもなお、大規模なニューラルネットワークの訓練は極めて困難であった。パーセプトロンやADALINEは単層ネットワークに限定されており、より複雑な問題への適用は想像しがたい状況であった。一方、記号主義的な　AI　アプローチ――論理定理証明機（Logic　Theorist）、一般問題解決機（GPS）――は、計算の透明性と追跡可能性の点で優れていた。それらは人間の推論プロセスをより直接的にシミュレートするかに見えたのである。

第二に、制度的・社会的な要因がある。ダートマス会議の主催者であったマッカーシーとミンスキーは、「人工知能」という概念を確立する過程で、サイバネティクスやニューラルネットワーク研究からの「独立」を強調した（第1章で既に述べた）。この独立は、単なる命名の選択ではなく、より広い知的戦略の一部であった。1960年代を通じて、MIT　のAI研究所はニューラルネットワーク研究に対して距離を置くようになった。研究資源の配分でも、明示的な知識表現と記号推論の路線が相対的に優位であった。

第三に、理論的な限界がある。単層パーセプトロンやADALINEの線形性の制約は、本質的であった。より複雑な非線形関数を計算するには多層ネットワークが必須であるが、その訓練法が存在しなかった。1960年代を通じて、何人かの研究者がこの問題に取り組んでいたが――例えば、アレクセイ・イヴァハネンコとヴァレンチン・ラパが階層的ネットワーク（Group Method of Data Handling, GMDH）の研究を進めていた――それらの成果は限定的であり、広く認識されることはなかった。

1970年代は、ニューラルネットワーク研究にとって「冬」であった。英国のライトヒル報告（1973年）は、人工知能研究全般への資金削減を勧告し、その中にニューラルネットワークも含まれていた。米国では、DARPA　　の資金削減は1970年代を通じてより徐々に進行したが、その効果は同様に深刻であった。多くの研究者がこの領域から去り、ニューラルネットワーク関連の学会発表や論文掲載数は著しく減少した。

しかし、完全な沈黙ではなかった。いくつかの研究グループは、その間にも地道にコネクショニズムの研究を続けていた。スティーブン・グロスバーグは、生物学的に着想を得たニューラルネットワークのモデルを1970年代に展開した。テウヴォ・コホネンは1970年代初頭から連想記憶研究を進め、自己組織化マップ（Self-Organizing　Map,　SOM）として結実するのは1982年である。ジェームス・アンダーソンも1970年代初頭から記憶と連想に関する神経モデルを発表していた。これらの研究者たちは、主流から外れた地位にありながらも、ニューラルネットワークの理論的・実装的な進展に貢献していた。

その上、ADALINE　の実用的な成功は、ニューラルネットワークアプローチの完全な放棄を招きはしなかった。適応信号処理の領域では、ウィドロウとホフの遺産が継続され、新たな応用が開拓されていた。しかし、これらの応用は「ニューラルネットワーク」という標識の下で語られることはなく、むしろ「適応フィルタリング」や「適応制御」といった別の名称で参照されていた。

1980年代中盤になると、コネクショニズムの再興の兆候が現れ始めた。1986年には、デービッド・ルメルハート、ジェフリー・ヒントン、ロナルド・ウィリアムズが『Nature』誌に「学習の表現」と題する論文を発表し、誤差逆伝播法の有効性を実証的に示した。このペーパーは、多層ニューラルネットワークの訓練が原理的に可能であること、そして実装的にも実行可能であることを実証した。この「PDP（Parallel Distributed Processing）革命」と呼ばれる運動は、コネクショニズムに第二の生命を与えることになる（第8章）。

第3章のまとめとして述べるならば、1950年代後半から1960年代にかけてのパーセプトロンとニューラルネットワークの最初の波は、機械学習の理論と実装に対して根本的な貢献をした。ローゼンブラットの収束定理は、学習が計算可能な過程であることを示した。ウィドロウとホフの最小平均二乗誤差法は、最適化を通じた学習の論理を確立した。これらの知見は、1970年代の沈黙の期間を経て、1980年代以降の統計的機械学習の台頭に再び影響を与えることになる。一方、『Perceptrons』が示した単層パーセプトロンの理論的限界は、実は二層以上のネットワークの必要性を論証的に確立したのであり、その意味では深層学習への理論的な道を開いたとも言える。

ただし、この時期の歴史を正当に評価するには、通説の再検証が必要である。『Perceptrons』がニューラルネットワーク研究を「殺した」という標準的な物語は、部分的には事実であるが、複雑な因果関係を単純化しすぎている。実際には、パーセプトロン研究は内発的な理論的限界に直面しており、その限界を突破するための技術（誤差逆伝播法）は、ミンスキーとパパートの批判と並行して開発されていたのである。また、コネクショニズムの衰退は、ニューラルネットワークという方法論そのものの無価値性ではなく、むしろ当時の計算機資源と理論的ツールの不十分さを反映していた。同時に、この沈黙の時代が完全ではなかったこと――グロスバーグ、コホネン、アンダーソンらの地道な研究の継続――が、後の再興を可能にしたのである。

1969年の『Perceptrons』から1986年のPDP革命までの17年間は、単なる空白期ではなく、新しい理論的・技術的基盤が静かに構築される時期であった。次の第4章では、同じ時期に記号主義的な AI アプローチが何を成し遂げたのか、そしてなぜそのアプローチもまた限界に直面することになったのかを検討する。これにより、知能の機械化の歴史における、異なるパラダイム間の相互的な脆弱性と補完性が明らかになるであろう。

> 本章では、パーセプトロンとニューラルネットワークの最初の波を、その理論的基礎、実装的展開、そして歴史的な挫折まで、多角的に検討した。ローゼンブラットのパーセプトロンは、機械学習の理論に学習の収束可能性をもたらした。ウィドロウ＝ホフ学習則は、最適化を通じた学習の枠組みを確立した。『Perceptrons』は、単層ネットワークの根本的な限界を証明し、多層構造の必要性を論証的に示した。その結果として訪れたコネクショニズムの一時的後退は、決して無意味な空白期ではなく、新たな理論的基盤が構築される時期であった。これらの知見は、次の時代の記号主義的 AI の隆盛、そしてその後の統計的機械学習への転回を理解する上で不可欠である。

参考資料（本章）

本文中の主要な記述を追跡するため、最小限の典拠を挙げる。

- Frank Rosenblatt, “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain” (1958).
- Smithsonian, “Electronic Neural Network, Mark I Perceptron”. Mark I パーセプトロンの現物情報の確認に使用。
- Cornell Chronicle, “Professor’s perceptron paved the way for AI – 60 years too soon” (2019). ローゼンブラット周辺の制度史整理に使用。
- Bernard Widrow and Marcian E. Hoff, 初期 ADALINE / LMS 関連論文群; Stanford Engineering / Stanford Report による Widrow の業績解説。

- Marvin Minsky and Seymour Papert, Perceptrons (1969, expanded edition 1988).
- Paul J. Werbos, The Roots of Backpropagation. 誤差逆伝播法の系譜確認に使用。

第4章　対話・理解・ロボティクスの初期探求 —— 模倣から意味へ、限定された世界から汎用知能へ

初期AIの最初の勢いは、1960年代に新しい領域への試行的な拡張を促した。第1章から第3章で述べた数理的基盤、記号処理パラダイム、そして学習能力の理論は、いま実際の問題に適用されはじめた。人間と機械の対話は可能か。言語理解とはどのような計算過程なのか。物理的環境で行動する知能あるロボットは実現可能か。こうした問いが、1966年から1970年代初頭にかけて、AI研究の新しい前線を開いた。しかし同時に、これらの試みが暴露した困難さは、その後のAI史における根本的な危機の予兆でもあった。本章では、対話、言語理解、ロボティクスという三つの領域における初期の重要な成果と、それらが明らかにした本質的な限界を検討する。

4.1　ELIZA —— 自然言語対話の先駆（1966）

驚くべき単純性と予期せぬ説得力

1966年、MITの認知科学者ジョセフ・ヴァイツェンバウムは、計算機によるテキストベースの対話を実験するためのプログラムを発表した。ELIZAと名付けられたこのシステムは、記号的な洗練さには欠けるものの、AI思想史上最も重要な実験のひとつとなった。なぜなら、それが明らかにしたのは、知能や理解に関する人間の根深い錯誤であったからである。

ELIZAの動作原理は極めて単純である。ヴァイツェンバウムが1966年の論文「ELIZA–A Computer Program for the Study of Natural Language Communication Between Man and Machine」で説明したように、プログラムは「パターンマッチングと置換（pattern matching and substitution）」を用いた。最も有名なバリエーション「DOCTOR」は、ロジャーズ派の人間中心的心理療法を模倣していた。ユーザーが「I am feeling depressed」と入力すれば、ELIZAは「Why are you feeling depressed?」と応答する。「My mother hates me」と述べると、「Tell me more about your family」と返す。このように、ユーザーの発話を解析し、事前に設定された規則に従って質問文に変換するだけである。その背後にあるのは、意味の処理ではなく、文字列のパターン操作に過ぎない。

しかし、ここで歴史的に重要な現象が発生した。ユーザー、とりわけヴァイツェンバウムの秘書をも含む多くの人が、ELIZAが真に自分を理解し、共感していると感じたのである。心理的な悩みを打ち明け、深い対話が進行していると信じたのである。ヴァイツェンバウムはその後、「I had not realized … that extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people」と述懐している。ヴァイツェンバウム自身の意図は、AIの研究者たちが人間の言語理解の複雑さを過小評価していることを示す批判的デモンストレーションであったはずだった。ところが、その成果は逆に、「機械でも言語対話ができる」という楽観主義を増幅させた。

「ELIZA効果」と理解の幻想

後に「ELIZA効果（ELIZA effect）」と呼ばれるこの現象は、「人々が、コンピュータが生み出した記号列に対して、正当化されるよりもはるかに大きな理解度を帰属させる傾向」を指す。この効果がもつ理論的含意は深刻である。ナイーブに解釈すれば、人間は記号の表面的な並列パターンだけで「理解」を感じる生物である可能性がある。しかし、より重要な洞察は逆である。すなわち、人間は社会的・心理的文脈において対話者に深い知的作用を帰属させる傾向が強く、その帰属がしばしば表面的な応答テクニックによって欺かれうるということである。

ヴァイツェンバウムはこの発見に衝撃を受け、1976年の著作『Computer Power and Human Reason』で、機械にはけして代替できない人間的判断の領域を強調するようになる。彼は、計算論的思考（computational thinking）に還元できない人間固有の領域が存在すると主張した。この批判は、同時代のAI楽観主義に対する最も誠実で深刻な警告となった。

言語理解の困難さへの隠れた示唆

しかし、ELIZAが意図せずに明かした最も本質的な問題は、言語理解とは何かということである。もしパターン置換によって人間が「理解」を感じるならば、「理解」とは何か。もし単純な規則操作によって対話が成立するならば、対話成立に必要な条件は何か。ヴァイツェンバウムは、この問いを逆説的に提示した。実は、ELIZAはユーザーの心理的期待に依存していたのである。ユーザーが「共感的な聞き手としての心理療法士」という社会的・文化的枠組みを投影したからこそ、単純な置換が深い対話として経験されたのだ。言い換えれば、「理解」は計算機の内部過程だけでは定義できず、ユーザーとシステムの相互作用の文脈に本質的に依存している。この洞察は、1980年代以降の自然言語処理理論に深い問題を提起し続けることになる。

4.2　STUDENT・SIR ―― 意味理解への挑戦

制限された言語の代数問題解法

ELIZAが言語対話の表面性を露呈させたのに対し、ほぼ同時期にMITで進められていた別の研究は、言語の意味理解をより直接的に問題化していた。1964年、ダニエル・ボブロウが博士論文「Natural Language Input for a Computer Problem Solving System」で発表したSTUDENTは、限定された自然言語表現を理解して代数の文章問題を解くプログラムであった。

STUDENTの設計思想は、ELIZAとは異なる方向を示していた。ELIZAが構文を機械的に操作するのに対し、STUDENTは意味表現（semantic representation）を通じて問題を解こうとした。例えば、「Tom is twice as old as Mary. If Tom is 24, how old is Mary?」という英文に対して、STUDENTは以下のようなプロ

セスを実行する。（1）言語をパースし、概念的な関係を抽出する。（2）その関係を代数方程式に変換する。（3）方程式を解く。（4）自然言語で答えを返す。このアプローチは、言語理解の本質を「意味への変換」と見なしていた。

STUDENTが扱える言語の範囲は厳密に限定されていた。利用可能な述語は「is」「equals」「contains」など十数個に限られ、許可された構文パターンも有限であった。しかし、この制限の中では、プログラムは確実に意味を処理していた。システムは複数の表現「Tom's age is 24」「Tom is 24 years old」「Tom, who is 24」を同じ意味表現に統合できた。こうした柔軟性は、純粋なパターン置換では実現不可能である。

意味ネットワークと世界知識の問題

STUDENTに続いて、1960年代末から1970年代には「意味ネットワーク（semantic network）」と呼ばれる知識表現法が注目を集めた。これは、概念をノードとし、概念間の関係をエッジで表現するグラフ構造である。例えば「鳥は動物である」「動物は生命体である」といった知識を、階層的なネットワークとして表現する。この手法により、STUDENTのような限定的なシステムを超えて、より豊かな世界知識を機械的に処理できるのではないかと期待された。

しかし、意味ネットワークのアプローチにも根本的な限界が存在した。第一に、知識を有限のネットワークとして明示的に表現しようとすると、「知識の爆発（knowledge explosion）」という困難に直面する。世界についての知識は事実上無限であり、必要な知識を事前に予測することは困難である。第二に、単語の意味そのものが文脈に依存するという深刻な問題がある。「銀行」という言葉が「river bank」なのか「financial institution」なのかは、文脈によってのみ決定される。この文脈依存性は、事前に構築されたネットワークでは対応できない。第三に、推論のプロセスそのものが組合せ的爆発を引き起こす。利用可能な知識が増えるほど、可能な推論径路も指数関数的に増加し、計算不可能になるのである。

STUDENTとその後継者たちの研究は、言語理解が単純な規則適用ではなく、広大で複雑な世界知識と推論能力に依存していることを明らかにした。しかし同時に、その依存性の深さゆえに、限定された領域を超えた汎用的な言語理解は、当時の記号処理的アプローチでは根本的に困難であることも示唆していた。

4.3　Shakey ── 初の汎用移動ロボット（1966–1972）

知覚・推論・行動の統合

1960年代を通じてAI研究が自然言語処理に集中している傾向が強かったのに対し、スタンフォード研究所（Stanford Research Institute, SRI）では、より統合的なアプローチが進められていた。1966年から1972年にかけてのShakey　プロジェクトは、単なる思考実験のプログラムではなく、実世界に接続された知能ある移動ロボットを構築する試みであった。

Shakeyは、カメラや距離センサーから得られる情報によって環境を知覚し、その知覚情報を記号的な表現に変換し、論理的な計画を立案し、その計画に基づいて物理的に行動するロボットであった。ロボットの基本的な構成は、視覚的知覚システム（computer　vision）、知識表現システム（knowledge representation）、計画システム（planning　system）、そして運動制御システム（motor　control）からなっていた。

Shakeyが実現した最初の成果は、「環境のモデル化」と「計画の自動生成」であった。与えられたタスク（例えば「部屋Dに行き、ブロック9をドアウェイ4の近くに押す」）に対して、Shakeyは以下のプロセスを実行した。（1）カメラから得た画像をプロセッシングし、物体の位置、障害物の配置を認識する。（2）その認識情報から、遂行可能な行動（移動、ドアの開閉、スイッチのオンオフ、物体の移動）のシーケンスを計画する。（3）計画された行動を実行する。（4）行動の結果を知覚し、必要に応じて計画を修正する。

この統合的なループの実現は、当時のテクノロジーでは極めて困難であった。ロボットビジョンは、今日のような洗練されたアルゴリズムを持たず、基本的なエッジ検出と図形認識に頼っていた。移動も遅く、不安定であった。しかし、この困難さにもかかわらず、Shakeyプロジェクトは、知能あるロボットシステムの可能性を実証する上で決定的に重要であった。

STRIPS計画システムと記号的推論

Shakeyの中核をなす技術的革新は、STRIPS（Stanford Research Institute Problem Solver）と呼ばれる計画システムであった。STRIPSは、与えられた初期状態、目標状態、利用可能なアクションの集合から、目標を達成するための行動シーケンスを論理的に導出するシステムであった。

STRIPSの動作原理は、状態空間探索（state　space　search）に基づいていた。初期状態から出発して、各ステップで実行可能なアクション（前提条件を満たすアクション）を試行し、その結果として新しい状態に遷移する。この探索は、目標状態に到達するまで続けられる。各アクションは、前提条件（precondition）、追加効果（add　list）、削除効果（delete　list）として定義される。例えば、「ドアを開ける」というアクションは、「ドアが閉まっている」という前提条件を持ち、実行されると「ドアが開いている」という効果を追加し、「ドアが閉まっている」という事実を削除する。

STRIPSの理論的意義は、記号的に表現された目標を、機械が実行可能な行動シーケンスへ自動変換できることを示したことである。第1章で述べたように、物理記号システム仮説は、適切に定義された記号の操作によって汎用的知能が実現可能であると主張していた。STRIPSは、その仮説をロボット計画の文脈で具体化したものであった。

限界と隠れた仮定

しかし、STRIPSの成功は、同時に深刻な隠れた仮定を露呈させていた。第一に、世界は完全に既知であると仮定されている。つまり、システムが知る必要のあるすべての関連事実は、事前に明示的に与えられていると見なされている。しかし現実は、環境は不確定性を含み、予期しない事象が発生し、知覚は不完全であり、新しい事実が常に発見される。第二に、アクションの効果が確定的（deterministic）であると仮定されている。実世界では、同じアクションが環境の状態によって異なる結果を招く。第三に、変化は明示的に記述されたアクションの結果によってのみ起こると想定されている。この「フレーム問題」は、後に第5章で詳述するように、AIの冬の一因となった。

Shakeyプロジェクトにおいても、これらの制約は顕著であった。ロボットが行動できるのは、極めて単純化された環境（整然とした室内、限定された物体、予測可能な照明条件）に限定されていた。複雑な環境では、知覚に誤りが生じ、計画が破綻し、ロボットは対応できなくなった。にもかかわらず、Shakeyの試みは、自律的知能ロボットの構想を最初に実現したという歴史的意義を持つ。後のロボティクスの発展（第12章）は、Shakeyの遺産の上に、確率的手法や適応的制御を付加することによってなされることになる。

4.4　マイクロワールドとBLOCKS WORLD ―― 限定世界での知識表現

簡潔性と制御可能性への希求

1960年代の後半から1970年代初頭にかけて、AI研究において新しい戦略が台頭した。それは、複雑な現実世界ではなく、人為的に簡潔化された「マイクロワールド（microworld）」と呼ばれる限定された領域に焦点を絞るアプローチである。マービン・ミンスキーとシーモア・パパートらがその可能性を論じ、多くのAI研究者がこのアプローチに従った。その理由は実用的であった。現実世界の複雑さはAIシステムにとって圧倒的であり、限定された世界であれば、システムの能力と世界の複雑さのバランスが取れる可能性があるのだ。

最も有名なマイクロワールドは「BLOCKS　　WORLD」である。これは、限定された数の立方体、ピラミッド、その他の幾何学的物体が平坦な表面（テーブル）の上に配置された、極めて単純化された環境である。BLOCKS　WORLDの特徴は以下の通りである。（1）有限で既知の物体セット。（2）色や形といっ

た明示的な属性。（3）「上に」「左に」「接触している」といった明確な空間関係。（4）曖昧性の完全な排除。例えば「赤いブロック」は、世界に赤いブロックが一つしかなければ、完全に明確に指示対象が定まる。

このマイクロワールドのアプローチが有効であった理由は、それが知識表現と推論の困難さを根本的に削減するからである。現実世界では、システムが理解する必要のある関連知識は無限に近く、また新しい知識が常に発見される。しかし、BLOCKS　WORLDでは、関連知識がすべて有限で事前に決定可能である。かつ、知識の曖昧性や文脈依存性がほぼ完全に排除されている。このため、AI研究者は、知識表現の形式（述語論理、フレーム、セマンティックネット）と推論アルゴリズムに集中できるのである。

マイクロワールドの学術的生産性

実際に、マイクロワールドのアプローチは、短期的には生産的であった。BLOCKS　WORLDを舞台にした様々なプログラムが開発され、問題解決、自然言語理解、視覚認識、計画立案の各領域で具体的な進展があった。制約された世界でも、複数の複雑な課題を同時に解く必要があり、システムは相互参照の解決（anaphora resolution）、常識推論の模擬、階層的な目標分解といった高度な処理を行わねばならなかった。この困難さは、現実的な問題のスケール削減版である。

ただし、学問的な意味での生産性と、現実的応用可能性との間には、深刻なギャップが存在した。マイクロワールド内での成功は、しばしば、システムが世界について非常に多くのことを「知っている」ために達成されるのであり、それは一般化可能性の低さを意味していたのだ。世界を記述する知識が多いほど、別の世界への適用は難しくなる。

4.5　SHRDLU —— 限定世界での言語理解（1970）

統合的な言語理解システムの実現

1968年から1970年にかけて、テリー・ウィノグラドがMITで開発したSHRDLU は、BLOCKS WORLDのマイクロワールドに徹底的に特化した自然言語理解システムである。SHRDLUという奇妙な名前は、活字植字機（Linotype machine）のキーボード配列「ETAOIN SHRDLU」に由来する。これは英語で頻度の高い文字列を機械的に並べたもので、当時の計算文化を反映した命名であった。

SHRDLUの革新的側面は、自然言語理解を単なる言語処理ではなく、言語・世界モデル・推論の統合として捉えたことである。ユーザーが「Put the green cone on the red block」と指示すると、SHRDLUは以下のプロセスを実行する。（1）英文を構文解析し、意味表現（semantic representation）に変換する。（2）その意味表現の参照対象（reference resolution）をBLOCKS WORLDの世界知識に基づいて解決する。つまり、「green cone」がどのオブジェクトを指しているかを現在の世界状態から判定する。（3）

その指示対象への到達可能性、物理的実行可能性を評価する。（4）必要な移動と操作の計画を立てる。（5）その計画をマイクロワールド内で実行する。（6）実行結果を世界モデルに反映し、ユーザーに報告する。

SHRDLUの言語能力は、当時の基準でも後の基準でも、驚くべき水準にあった。システムは長範囲の照応参照を解決できた。例えば、「Pick up the big red block」の後に「Put the cone on it」と述べると、「it」が「the cone」を指すのではなく、直前に取り上げられた赤いブロックを指すことを理解した。さらに複雑な場合も処理できた。「I had you set aside two red blocks. Why did you do that?」といった質問に対して、システムはその行動の根拠（過去の会話コンテキスト）を参照して応答できた。

統合的理解の達成と同時の限界の露呈

SHRDLUの最大の成果は、限定世界であれば、言語理解をかなり深い水準で実現しうるシステムが構築可能であることを示した点にある。ELIZAのように表面的なパターン置換ではなく、実際に世界のモデルを保持し、言語を意味表現に変換し、その意味表現を基礎として推論し、行動するシステムが構築可能であることを示した。この意味では、SHRDLUは認知科学とAIの境界を画定する重要な実験でもあった。

しかし、同時に、SHRDLUの設計上の制約は、その成功の本質的な限界をも明らかにしていた。第一に、BLOCKS WORLDへの徹底的な特化である。プログラムは、ブロック、円錐、ピラミッド、玉といった数種類の物体の性質について、詳細な知識を組み込まれていた。色、形、サイズ、位置といった属性の組み合わせ方、物理的制約（「物体は宙に浮かない」「積み重ねの高さには限界がある」）、動作の前提条件（「掴むには物体が掴める状態になければならない」）といったすべてが、ウィノグラドによって明示的にプログラムされていた。

第二に、意味の基礎の限定性である。SHRDLUが理解できる意味は、BLOCKS WORLDの物理的性質に直接結びついた意味だけである。その外の領域への拡張は、ほぼ不可能である。例えば、同じシステムで「Put the economic policy into effect」といった抽象的な指示を理解させるには、根本的な再設計が必要になるだろう。

第三に、言語現象への対応の限定性である。SHRDLUは、幾何学的・物理的領域の言語表現については精密に処理できるが、メタファー、曖昧性、修辞、言語遊戯といった人間の言語使用の豊かさには対応できない。システムが拒否する不正な入力も多い。例えば、若干の倒置やレジスター（話し言葉のスタイル）の変化にも敏感に反応してしまう。

ウィノグラド自身は、SHRDLUの成功に慢心することなく、その限界を認識していた。1970年代末から1980年代にかけて、彼はAI的な記号処理アプローチそのものの根本的な問題に直面することになり、やがてそのパラダイムから距離を置くようになる。

第4章のまとめ —— 対話・理解・ロボティクスの困難な現実

本章で検討した四つのシステム（ELIZA、STUDENT、Shakey、SHRDLU）は、初期AIの知的野心の頂点を示している。それぞれが、人間的な知的活動——対話、理解、計画、行動——を機械的に実現しうることを示唆していた。マッカーシーとミンスキーが1950年代に描いた楽観的な展望は、1960年代から1970年代初頭にかけて、実在するプログラムによって具体化されつつあるように見えた。

しかし、この成功の内実は複雑である。ELIZAの成功は、実は人間の心理的投影に依存していた。STUDENTとその後継者たちは、意味理解には膨大な世界知識が必要であることを示唆した。Shakeyは、実世界との相互作用の複雑さを露呈させた。SHRDLUは、言語理解が成立するためには、言語対象領域の徹底的な形式化が必要であることを示した。

これらの成果が共通して示していることは、知能とは記号の操作だけでは成立しないということである。知能には、（1）広大で複雑な世界知識、（2）不確定な現実との相互作用に対応する適応性、（3）文脈に依存した推論、（4）言語と実世界の接続、といった複数の層が必要なのだ。

同時に、この時期の研究成果は、別の深刻な問題をも露呈させていた。記号主義的なAIパラダイムの下では、複雑性は「組合せ爆発」によって制御不能になる傾向がある。世界をより精密に表現するほど、利用可能な知識が増えるほど、推論の計算量は指数関数的に増加する。マイクロワールドは、この組合せ爆発を人為的に回避するための戦略であったが、その代償は深刻であった。限定された世界での成功は、そのまま一般化可能ではないのだ。

ウィノグラドとシャンク、そしてミンスキーら初期AIの指導者たちが1970年代後半から直面することになった課題は、AI研究の根本的な方向転換を促すことになる。言語理解、知識表現、推論システムの「本質的困難さ」が、単なる計算資源の限界ではなく、アプローチそのものの問題であることが次第に明らかになっていったのである。

第5章で述べるように、1973年のライトヒル報告と1970年代から1980年代の「AIの冬」は、この知的限界を制度的・経済的現実として結晶化させたものである。しかし、その危機を生み出した根本的な思想的張力は、本章で検討したシステムの内部にすでに存在していたのだ。ELIZAの幻想、STUDENTの知識問題、Shakeyのフレーム問題、SHRDLUの領域特化性——これらはすべて、記号主義的AIパラダイムの限界を先取りする徴候でもあった。

AIの歴史は、しばしば「冬」と「春」の循環として描かれる。しかし、より精密には、各時代の成功の内に隠された困難が、次の時代の制度的危機を準備しているのだ。初期AIの対話・理解・ロボティクスへの挑戦は、その後のAI研究が何度も直面することになる根本的な課題——知識の表現と獲得、推論の計算複雑性、言語と意味の関係、身体性と知能の関係——を最初に可視化した歴史的営為であった。

本章では、1966年から1970年代初頭にかけてのAI研究における対話・理解・ロボティクスの初期的成果と限界を検討した。ELIZAは自然言語対話の可能性と幻想を示し、STUDENTは意味理解が膨大な世界知識を必要とすることを示唆し、Shakeyは実世界との相互作用の複雑さを露呈させ、SHRDLUは言語理解の成立条件と同時にそのマイクロワールド依存性を明かした。これらの成果は、ダートマス会議以来の記号主義的AIパラダイムの最初の本格的な実験結果であり、同時にその根本的限界を示唆していた。次章で述べる第一次AIの冬は、この知的張力が制度的危機として爆発した事象であり、その後のAI研究は、ここで露呈した困難に対応する新しいパラダイムと手法を模索することになるのである。

参考資料（本章）

本文中の主要な記述を追跡するため、最小限の典拠を挙げる。

- Joseph Weizenbaum, “ELIZA–A Computer Program for the Study of Natural Language Communication Between Man and Machine” (1966).
- Daniel G. Bobrow, Natural Language Input for a Computer Problem Solving System (1964).
- SRI, “Shakey the Robot”. Shakey の時期と基本構成の確認に使用。
- Nils J. Nilsson, The Quest for Artificial Intelligence. Shakey と STRIPS の制度史整理に有用。
- Terry Winograd, Procedures as a Representation for Data in a Computer Program for Understanding Natural Language (1971) および Computer History Museum の SHRDLU 記録。

第II部

知識の時代と冬の到来（1970年代〜1980年代）

第5章〜第8章

第5章　第一次AIの冬 —— 期待と現実の乖離

概観

1973年から1980年代初頭にかけて、AI研究は資金的・制度的・知的な危機に見舞われた。この時期は後に「第一次AIの冬」と呼ばれるようになる。しかし、この表現は一定の注意を要する。「冬」という用語は、研究が完全に停止したことを示唆するが、実際には理論的・技術的な営みは継続していた。むしろ重要なのは、初期AIの楽観的なパラダイムが根本的に問い直され、研究コミュニティが自らの前提を再検討するようになったことである。本章では、第一次AIの冬を単なる「失敗の時代」としてではなく、AI研究が直面した構造的問題と、その問題への対応の過程として分析する。

5.1　ライトヒル報告と英国における研究資金の凍結（1973）

ライトヒル報告の背景と構成

1973年、英国科学研究会議（Science Research Council, SRC）は応用数学者ジェームズ・ライトヒル卿による包括的な報告書を発表した。公式には「人工知能：総括的調査（Artificial Intelligence: A General Survey）」と題されたこの報告書は、英国におけるAI研究の到達点を評価し、その制度的価値を問い直す検証となった。

ライトヒル報告は、英国AI研究の主要拠点——とりわけエディンバラ大学を中心とする研究群——を念頭に置きつつ、その成果と限界を冷徹に検討した。報告書は以下のような複合的な批判を構成していた：

1. **約束と成果のギャップ**：初期のAI研究者（サイモン、ミンスキー）が掲げた野心的な予測——10年以内にチェスの世界チャンピオンを倒す、常識推論を可能にするなど——が実現されなかった。
2. **スケーラビリティの欠如**：ミクロワールドや限定領域では成功を収めたAIプログラムも、現実の複雑さに直面すると失敗した。
3. **根本的な理論的困難**：組合せ爆発、フレーム問題、常識推論の困難など、単なる計算資源の増加では解決不可能な理論的問題の存在。
4. **知識表現の限界**：機械が形式的に記述可能な知識の範囲が、想像されたより遥かに限定的であった。

ライトヒルは、汎用問題解決プログラムも定理証明もロボットの常識的判断も、結局は組合せ爆発によって深く制限されると論じた。この診断は、初期AIの基本的仮定の妥当性そのものに疑問を投げかけるものであった。

報告書の政治的・制度的影響

ライトヒル報告の影響は、学術的な批判を超えた。1973年5月9日、ロンドンの Royal Institution（王立研究所）でライトヒルと主要なAI研究者（ドナルド・ミッチー、ジョン・マッカーシー、リチャード・グレゴリー）の間で公開討論会が開催された。この討論はBBCの「Controversy」シリーズで放映され、「汎用ロボットは蚤気楼か」という趣旨の挑発的な論点をめぐって激しい議論が交わされた。

マッカーシーを始めとするAI研究者たちは、報告書の批判に対して強く異議を唱えた。彼らは、基礎研究が常に短期的な応用を生み出すとは限らず、長期的視野から根本的な問題に取り組むことの重要性を主張した。しかし、英国科学研究会議はこれらの反論にもかかわらず、ライトヒルの評価を受け入れることを決定した。

その結果は劇的であった。1973年以後、英国の大学におけるAI研究に対する資金提供は大幅に絞られた。すべての拠点が一律に閉じられたわけではないが、エディンバラを含む主要拠点は研究計画の縮小や再編を余儀なくされた。多くの研究者は他分野への転向、あるいは国外移動を選ばざるを得なかった。この一撃によって、英国が20世紀中盤に保有していたAI研究の国際的存在感は著しく損なわれることになった。

ライトヒル報告の再評価

後世の史家による再検討は、ライトヒル報告に対する評価の複雑性を明らかにしてきた。一方では、報告書がAI研究に向けた批判の多くは本質的に正当であったこと——組合せ爆発、常識推論の困難、知識獲得ボトルネックなどは、その後の数十年間を通じてAI研究を悩ませ続けた——は否定できない。

しかし他方で、報告書の論理的構造に対する疑問も提起されている。例えば、ライトヒルが「AI研究が失敗した」という結論に至る際、彼は初期の研究者の過度な楽観性を主な原因として指摘した。しかし、科学的進歩は必然的に乗り越えるべき楽観と失敗の両者を含むプロセスである。また、報告書は英国AI研究の特定の弱点に焦点を当てるあまり、同時期に米国で進行していた知識表現や自然言語処理の地味だが着実な進歩を見落とした可能性がある。

5.2　DARPA資金の削減と米国AI研究への影響

米国における資金配分体制とその変化

英国の場合とは異なり、米国でのAI冬の到来はより漸進的で複雑であった。しかし、その影響は同じく決定的であった。

1960年代から初期1970年代にかけて、DARPA（国防高等研究計画局）はAI研究の最大の資金提供者であった。冷戦期の軍事的競争の文脈において、AI研究は自動翻訳、音声理解、画像解析、推論システムなど、潜在的に軍事的価値を有するものとして位置づけられていた。MIT、スタンフォード、カーネギーメロン大学の三つの主要拠点は、その恩恵を大きく受けていた。

しかし1970年代に入ると、状況は急速に変わり始めた。第一に、ベトナム戦争の経験が軍事予算に対する議会の監視を厳しくした。1969年に可決されたマンスフィールド修正案は、軍事的な特定の機能と明確に関連しない研究に対する軍事資金の使用を制限した。初期AIプログラムの多くは、直接的な軍事応用との結びつきを示しがたく、この修正案の対象となりやすかった。

第二に、DARPA内部でのAI研究への評価が低下していた。既存のプロジェクトが公約した成果を上げないばかりか、要求される予算が増嵩する傾向にあったからである。自動翻訳、自動推論、ロボット制御などの領域では、期待される進捗が得られていなかった。

1973年から1974年にかけて、DARPAはAI研究への資金配分を大きく見直した。基礎的なAI研究への包括的支援は後退し、より明示的な応用目標を持つプロジェクトへ重点が移された。ここで重要なのは、「AIへの支援が突然ゼロになった」のではなく、「ミッション志向の強い資金制度へ再編された」という点である。

米国における反応と適応

しかし、ライトヒル報告がもたらした英国の劇的な市場撤退とは異なり、米国のAI研究は完全には消滅しなかった。重要なのは、資金削減の方向性と規模であった。

MIT、スタンフォード、CMUなどの主要拠点は、DARPA資金の削減に直面した際、異なる戦略をとった。その一つは、研究の焦点を「知識ベース」システムへと移行させることであった。これは、AIの普遍的な汎用性を前提とした初期のアプローチを放棄し、特定の領域（医療診断、化学分析など）における「専門知識」の形式化と応用に的を絞るものであった。このアプローチが後の「エキスパートシステム」の時代（第7章）を準備することになったことは注目される。

また、一部の大学では産業界との結びつきを強化する動きが見られた。IBM、ゼロックス、テキサス・インスツルメンツなどの企業がAI研究に投資を始めた。これらの企業の関心は、汎用的な知能システムの開発より、実用的な商用応用の可能性にあった。この産学協力の展開は、AI研究のアカデミック・セクターから産業セクターへの重心移動を示唆するものであった。

5.3　組合せ爆発問題とフレーム問題

組合せ爆発の論理的構造

第一次AIの冬の根底にあった技術的困難のうち、最も根本的なものの一つが「組合せ爆発」である。この問題は新しいものではなく、むしろ初期のAI研究が構造的に抱えていた問題であった。

組合せ爆発とは、問題空間における選択肢の数が入力サイズに対して指数関数的に増加する現象を指す。例えば、チェスの場合、平均的な局面において1手進むごとに約35の指し手選択肢が存在する。したがって、n手先を見通す場合、評価すべき局面数は（35^n）となり、この数は急速に天文学的規模に達する。この問題は、シャノンがチェス探索における「タイプB戦略」（選択的探索）を論じた1950年の段階で、すでに予感されていた。

しかし、1960年代後半から1970年代にかけて、この問題の深刻性が徐々に認識されるようになった。それは、初期のAIプログラムが達成した成功が、実は極めて限定的な問題領域での成功であったことが明らかになったからである。ELIZAが精神科医の相談者として機能したのは、その応答が実際には高度に定型化されており、真の意味での言語理解を必要としなかったからであった。SHRDLUがブロックスワールドで完璧に機能したのは、その世界の複雑性が意図的に制限されていたからであった。

現実世界の問題——自然言語翻訳、ロボット制御、医療診断——に対して同じ論理ベースの推論を適用しようとした場合、組合せ爆発は致命的な障害となった。一般的問題解決機（General Problem Solver）が、より大きな問題に直面すると完全に失速するのを観察することで、研究者たちは漸進的な認識に到達した：単に計算能力を増すだけでは十分ではなく、より根本的な推論アーキテクチャの再設計が必要であるということ。

フレーム問題の哲学的含意

しかし、組合せ爆発以上に深刻だったのが、「フレーム問題」（frame problem）である。この問題は、1969年にジョン・マッカーシーとパトリック・ヘイズが論文「人工知能の観点からのいくつかの哲学的問題」の中で初めて形式的に定式化した。

フレーム問題の技術的な形態は、行動の効果を論理で表現する際に直面する表現の問題として現れた。例えば、「ロボットがAを動かす」というアクションを表現する場合、形式論理では以下のようなことを記述する必要がある：

```
Action(MoveA)
Precondition: At(A, LocationX)
Effect: At(A, LocationY)
```

しかし、問題はここから始まる。Aを動かすというアクションが実行された場合、Aの位置は変わるが、「Bはまだ元の場所にある」「机の色は変わっていない」「重力は働いている」など、無数の「変わらない」事実をすべて明示的に記述する必要があるのか？あるいは、変わらないと想定する事実の集合をどのように定義するのか？

この問題の中核は、人間の推論が本質的に「何が関連的か」を直感的に判断する能力を持っていることにある。私たちは、「机を動かした」という情報から、自動的に「空気の分子の配置は変わっていない」「太陽までの距離は変わっていない」などという無数の非言及的な真実を無視する。しかし、機械はこの「何が重要か」というフィルタリング能力を持たない。したがって、形式論理によるアクション表現は、人間にとって自明な常識を明示化する必要に直面するのである。

マッカーシーは、その後この問題に対する解決策として、状況計算の精密化や「周囲化」（circumscription）のような形式的手段を提案した。しかし、1980年代にイェール大学の研究者が提示した「イェール銃問題」（Yale shooting problem）は、この種のアプローチがなお困難を抱えていることを露わにした。この問題はおおむね以下のように設定される：

1. Fred は最初は生きている。
2. 銃は最初は空であり、その後装填される。
3. 少し待ってから Fred を撃つ。通常なら Fred は死ぬはずである。

このシンプルな状況設定から、妥当な推論システムは「Fred　は死んでいる」と結論すべきである。しかし、変化しない事実をどう最小化するかという規則を素朴に与えると、銃がなぜか空に戻る、といった直感に反する解釈が競合してしまう。イェール銃問題は、フレーム問題が1970年代だけの論点ではなく、その後も長く知識表現研究を悩ませたことを示す象徴的事例である。

フレーム問題の意義は、単なる形式論理の技術的困難を示すだけではない。それは、以下の根本的な認識をもたらした：「知識」を形式化する過程で、人間が当たり前と考える常識的推論は、実は極めて複雑で、多層的な暗黙的知識に支えられているということである。

5.4　常識推論の困難 —— なぜAIは「当たり前」を理解できないのか

常識知識の量と複雑性

AI研究が直面した第三の根本的困難は、「常識推論」（common sense reasoning）の必要性である。人間の知的活動のかなりの部分は、明示化されない背景知識に支えられている。例えば、「太郎は病院に行った。医者を見た」というテキストを理解するためには、以下のような無数の背景知識が必要である：

- 病院は医師がいる場所である。

- 「医者を見た」は、医師と対面したことを意味する可能性が高い。
- 医師は通常、患者に対して医療提供行為を行う。
- これらの行為は特定の物理的・社会的文脈に依存している。

1970年代のAI研究者たちは、この「常識」を機械に教え込むことの難しさに直面した。ジョン・マッカーシーの「Advice Taker」（1959年）以降、常識推論を形式化しようとする努力が続けられていたが、その困難さは予想をはるかに上回っていた。

1970年代を通じて、研究者たちは常識知識の規模に関する推定値を段階的に上方修正せざるを得なかった。初期には、「数千から数万の事実を入力すれば、機械が常識的推論を行えるようになるだろう」と考えられていた。しかし、より詳細な分析により、人間の常識知識は極めて膨大であることが明らかになった。一般的な大人が保有する常識知識は、単純な推定でも数百万から数十億の事実・ルール・ヒューリスティックスから構成されている可能性がある。

常識推論とデフォルト論理

1970年代末から1980年前後にかけて、常識推論の困難に対する理論的応答が現れ始めた。その一つが「デフォルト論理」（default logic）である。この論理形式は、確実な知識（例：「すべての人間は死ぬ」）と、デフォルト的な規則（例：「通常、鳥は飛ぶ——ただしペンギンを除く」）を区別することで、常識推論をより柔軟に扱おうとしたものであった。

マービン・ミンスキーの「フレーム理論」（1974年）もまた、このような常識推論の問題に対する一つの応答であった。ミンスキーは、知識を統合的なフレーム（「ステレオタイプ的な状況」を表現するデータ構造）として組織することを提案した。フレームは、デフォルト値（通常の場合に成り立つ値）、例外処理（デフォルトが成り立たない場合）、関連するサブフレームへの参照を含むことができる。例えば、「レストラン」というフレームには、「テーブルがある」「メニューがある」「ウェイターがいる」などのデフォルト要素が含まれる。新たに登場人物がレストランに入った場合、システムは自動的に「この人はメニューを見るだろう」「食事をするだろう」といった推論を行える。

しかし、フレーム理論もまた、その限界を持つことが次第に明らかになった。フレームを構築し、例外を記述し、デフォルト値を設定するプロセスは、本質的に人間の手作業を要求した。また、複数のフレームが相互に関連する複雑な状況では、どのフレームを適用するかの判断そのものが非自明な問題として現れた。

5.5　AI研究コミュニティの自己反省と方向転換

「AIの冬」という共有された認識

1970年代後半にかけて、AI研究コミュニティは徐々に危機的認識を共有し始めた。会議報告や回顧的論考では、AI研究の「行き詰まり」について率直な議論が交わされるようになった。主要な研究機関でさえ、プロジェクトの縮小や方向転換を余儀なくされた。

この時期における重要な現象は、研究者たちが過去の過度な楽観性を批判的に再検討し始めたことである。ハーバート・サイモンやアレン・ニューウェルは、1960年代後半の自らの予測に対する反省的発言を公表した。彼らは、知識表現の困難さと組合せ爆発の深刻性を過小評価していたことを認めた。この知的謙虚性は、後のAI研究の新しい方向性を準備するために重要な心理的基盤となった。

パラダイムの多元化と共存

興味深いことに、第一次AIの冬は、AI研究を「完全に」停止させなかった。むしろ、初期の記号主義的パラダイムの一元的支配が緩和され、複数の研究パラダイムが併存する状況が生まれた。

第一に、「論理プログラミング」の展開である。1970年代前半、アラン・コルメローとロベール・コワルスキーらの仕事により、Prolog言語が開発された。Prologは、形式論理（特に述語論理）に基づくプログラミング言語として、記号的推論を形式化する新しい枠組みを提供した。これは、初期のAIが直面した表現と推論の問題に対する一つの技術的応答であり、後に「知識表現」の体系化（第6章）に繋がるものであった。

第二に、「心の社会」（Society of Mind）理論の発展である。マービン・ミンスキーは、1970年代から1980年代にかけて、従来の統一的な知能モデルに替わって、複数の専門的エージェント（agents）が相互作用する分散的システムとして知能を構想した。この理論は、知識表現と推論の統一的解決を求める努力を放棄し、局所的・並列的な計算メカニズムの集合として知能を理解する新しい視点を示唆するものであった。

第三に、研究の応用志向化である。1970年代後半から1980年初頭にかけて、一部の研究者たちは、高度な汎用知能の実現よりも、限定された領域における「専門知識」の形式化に目を向け始めた。このアプローチが、後にエキスパートシステムのブームをもたらすことになる（第7章）。

研究が完全には止まらなかった事実

ここで重要な歴史的修正が必要である。「AIの冬」という隠喩は、AI研究が完全に消滅したことを示唆するが、実際には研究活動は継続していた。例えば以下の点に注意すべきである：

1. **ニューラルネットワーク研究の地下水流**：ミンスキー＝パパートの『Perceptrons』（1969年）がニューラルネットワーク研究に打撃を与えたことは確かであるが、この分野が完全に消滅したわけではなかった。1970年代を通じて、小規模なグループが継続的にニューラルネットワークの理論的・実験的研究を進めていた。特に、適応線形素子（ADALINE）の研究を続けていたバーナード・ウィドロー（スタンフォード大学）や、隠れ層を持つネットワークの学習可能性に関する理論的研究が進められていた。

2. **ロボティクス研究の継続**：ハンス・モラヴェックらは、1970年代を通じて、ロボット制御と視覚認識に関する地道な研究を継続していた。ロドニー・ブルックスの身体性重視のロボティクスが大きな影響力を持つのは、やや後の1980年代後半からである。これらの研究は、主流のAI研究から外れたものとして扱われることが多かったが、身体性を備えた知能の理解に向けた重要な知的資産を蓄積していた。

3. **知識表現研究の多様化**：メディカルAI（特に診断システム）、データベースシステム、知識工学など、特定の応用領域に焦点を当てた研究が展開されていた。これらは、普遍的知能の実現という当初の野心的目標とは異なるものであったが、実用的価値を示し、後の技術発展の基礎となった。

「冬」の複数性に関する考察

現在の歴史研究から見直される点として、「第一次AIの冬」がモノリシック（一枚岩的）な現象ではなかったことが指摘されている。例えば、英国と米国では状況が異なっていた。また、時間軸に沿っても、1973年のライトヒル報告による即座の影響と、1980年代後半のLISPマシン市場崩壊を軸とする第二次AIの冬とは、メカニズムが異なっていた。

さらに、当事者たちの認識も一様ではなかった。フロンティアの研究者の中には、この時期を「冬」とは見做さず、むしろ「興奮に満ちた時間」と回想する者もいた。AI研究の規模が縮小したことは事実だが、その結果として行われた理論的・哲学的検討が深まったこともまた事実なのである。

第5章のまとめと次章への橋渡し

第一次AIの冬は、AI研究の発展史における分水嶺である。しかし、それは単なる「失敗」や「衰退」ではなく、初期AI研究の根本的な前提が問い直された時期として理解すべきである。

初期AI研究の基本的仮説は、以下のようなものであった：

1. 知能は原理的に形式記号で表現可能である。
2. この形式記号の操作により、汎用的な推論を実現できる。
3. 計算能力の増加により、より複雑な問題が解決可能になる。

第一次AIの冬を通じて、これらの仮説の妥当性が根本的に問い直された。特に、常識推論、フレーム問題、組合せ爆発といった障害は、単なる技術的課題ではなく、知識表現と推論のアーキテクチャに関わる本質的な問題であることが明らかになった。

この認識の深化は、次の時代への知的準備となった。1970年代後半から1980年初頭にかけて、AI研究は「普遍的知能」から「知識の体系化」へと焦点をシフトさせ始めた。フレーム理論、スクリプト、意味ネットワーク、述語論理とPrologなどの知識表現の新しい形式は、この時期に精力的に発展させられた。これらが第6章で詳述される「知識表現と推論の体系化」の時代への道を開くことになるのである。

また、第一次AIの冬の経験は、後の研究者たちに重要な教訓をもたらした。それは、技術的限界を謙虚に認識し、長期的視野から根本的な問題に取り組むことの重要性である。この教訓は、エキスパートシステムのブームを経て、さらに統計的機械学習への転回（第9章）を経た後も、AI研究の知的文化の一部として継承されていくことになる。

> 本章では、第一次AIの冬を、単なる資金の枯渇や技術的失敗ではなく、AI研究が直面した構造的課題と、その課題に対する理論的応答のプロセスとして分析した。ライトヒル報告（1973）による英国の急速な撤退、DARPA資金の戦略的再編、そして組合せ爆発・フレーム問題・常識推論といった根本的な理論的困難は、初期AI研究のパラダイムの限界を露呈させた。しかし同時に、Prolog言語、フレーム理論、非単調論理といった新しい知識表現と推論の手段が開発される知的環境も醸成した。次章では、このような危機的認識から生まれた「知識表現と推論の体系化」の時代を検討し、AI研究がどのように新しい理論的基盤を構築していったかを論じる。

参考資料（本章）

本文中の主要な記述を追跡するため、最小限の典拠を挙げる。

- James Lighthill, “Artificial Intelligence: A General Survey” (1973).
- Jon Agar, “What is science for? The Lighthill report on artificial intelligence reinterpreted” (British Journal for the History of Science, 2020).
- Bodleian Archives, Lighthill / BBC Controversy 関連記録。1973年討論会の開催形態確認に使用。

- IEEE Spectrum, “Freddy the Robot and the Great Debate over AI’s Future” (2025). 1973年討論会の歴史的位置づけ整理に使用。
- NSF, “The Mansfield Amendment”. 米国における研究資金制度の転換の背景確認に使用。
- John McCarthy and Patrick J. Hayes, “Some Philosophical Problems from the Standpoint of Artificial Intelligence” (1969).
- Steve Hanks and Drew McDermott, “Nonmonotonic Logic and Temporal Projection” (1987). Yale shooting problem の原典。
- AAAI Digital Library, AI Magazine archives. 創刊年が1980年であることの確認に使用。

第6章　知識表現と推論の体系化

6.1　フレーム理論（ミンスキー、1974）とスクリプト（シャンク）

1970年代初頭、AI研究コミュニティは深刻な危機に直面していた。第5章で述べたように、初期の汎用問題解決プログラムや探索的アプローチは、限定された領域を超えて拡張することができず、組合せ爆発と常識推論の困難に翻弄されていた。この窮状を打開するために、研究者たちは新しい視点に転じた。知識の獲得と表現をより体系的に、そして現実的に捉え直す必要があったのである。

この転機を象徴する重要な出来事が、1974年のマービン・ミンスキーによるフレーム理論の提唱であった。ミンスキーはまず1974年のAI Lab memoとして『知識表現のための枠組み（A Framework for Representing Knowledge）』を提示し、翌1975年に公刊された論集で広く流通させた。彼は、従来の論理的記号処理から一歩進んだ知識表現の枠組みを提案した。その核心は、人間が新しい状況に遭遇するとき、記憶の中から既存の典型的な「フレーム（frame）」――すなわち定型的な状況の構造化されたイメージ――を呼び出し、それを現実に適合させるというメカニズムにある。

フレーム理論の特徴は、知識を静的な論理式ではなく、動的にして階層的な「スロット（slot）」と「デフォルト値（default value）」を持つ構造として表現する点にある。たとえば「部屋」というフレームであれば、その上位レベルには「部屋には床と壁と天井がある」といった不変の特性が固定されている。一方、下位レベルには「床の色」や「家具の配置」といった変動可能な詳細が多くのスロットを持つ形で配置される。さらに注目すべきは、各スロットには「これが満たされない場合どうするか」という手続き的な情報も付属するということであった。これは単なるデータ構造ではなく、推論の指針を組み込んだ知識表現の形式だったのである。

ミンスキーの枠組みはまた、記憶と想像の関係について新しい光を当てた。複数の関連フレームが階層的に連結された「フレーム・システム（frame systems）」を通じて、推論は「期待に合致するか例外か」という二項的な過程を辿る。人間が視覚的イメージを操作するときのように、AIもフレーム構造を変形させることで推論を効率化できるというこのモデルは、認知心理学とAIの接点を新たに開いた。

フレーム理論の影響はきわめて大きかった。次節で述べるスクリプト理論をはじめ、後のオントロジーやセマンティックウェブの構想にも直接的な思想的継承が見られる。しかし同時に、フレーム理論は技術的な限界を抱えていた。フレームそのものを自動的に生成・修正する仕組みが明確ではなく、知識エンジニアが人手でフレームを構築するしか方法がなかったのである。これは本章の終盤で論じる「知識獲得ボトルネック」問題の最初の現れであった。

フレーム理論とほぼ同じ時期に、スタンフォード大学のロジャー・シャンク（Roger Schank）は異なるアプローチから知識表現の問題に取り組んでいた。シャンクが1972年に発表した「概念的依存関係（Conceptual Dependency, CD）」理論は、自然言語の意味を言語の形式に依存しない根本的な「原始行

為（primitive　acts）」の集合で表現しようとするものであった。たとえば「与える」「取る」「移動する」といった少数の原始行為に、「物体転移（ATRANS）」や「位置転移（PTRANS）」といった抽象化されたカテゴリを与え、複雑な文の意味をこれらの組み合わせで表現できると主張したのである。

シャンクの理論は、さらに発展してスクリプト理論へと結晶化する。1977年にシャンク自身とロバート・アベルソンによる『スクリプト、計画、目標、理解（Scripts, Plans, Goals, and Understanding）』が出版されると、AI　における知識の可視化は新たな段階に進んだ。スクリプトとは、「レストランで食事をする」というように、定型的な状況下で生じる一連の行為の流れを時間的に構造化した知識である。レストランスクリプトには、「入店する」「席に座る」「メニューを見る」「注文する」「食事する」「会計する」「退店する」という標準的な段階が含まれ、各段階には通常期待される行為や登場人物が暗黙的に紐付けられる。

フレーム理論とスクリプト理論のいずれもが、AI が直面していた根本的な問題――常識知識をどのように表現し、それを効率的に利用するか――に対する異なるアプローチを提供していた。フレームは空間的・構造的な知識（「部屋とは何か」）の表現に、スクリプトは時間的・手続的な知識（「レストランで食事をするとはどのようなプロセスか」）の表現に、それぞれ適していた。この双方の枠組みは、次の十年のエキスパートシステム（第7章）の知識ベース構築において重要な影響を及ぼすことになる。

6.2　意味ネットワークとオントロジーの起源

フレーム理論やスクリプト理論と並行して、別の知識表現の伝統が発展していた。それが意味ネットワーク（semantic　network）である。意味ネットワークの起源は1960年代初頭に遡る。ロス・クイリアン（Ross Quillian）は、意味記憶研究の一環として、言語理解のための知識を「概念ノード」と「ラベル付き関係弧」から構成される有向グラフとして表現する方法を開発した。

意味ネットワークの基本的な構造は極めてシンプルである。各ノードは概念や実体を表し、ノード間の弧は概念間の関係を表現する。たとえば「犬」というノードから「動物」というノードへ向かう「is-a 関係（is-a　relation）」の弧が引かれれば、「犬は動物である」という知識が表現される。さらに「犬」から「四本の足」へ向かう「持つ（has）」の弧があれば、より詳細な属性情報が統合される。この単純さゆえに、意味ネットワークは直感的でプログラム化しやすく、多くの初期NLPシステムに採用された。

意味ネットワークの認知的妥当性についても、当時の研究者たちは強い確信を持っていた。人間の意味記憶がネットワーク構造で構成されているという心理学的な仮説は、同時代の認知科学者アラン・コリンズ（Allan　Collins）とマーク・クイリアン自身による実験的支持を得ていた。彼らの「認知的経済（cognitive　economy）」仮説によれば、人間は属性情報を最小限のノードに格納することで、脳の記憶負荷を軽減しているというのである。たとえば「犬には毛がある」という事実を個々の犬についてそれぞれ記憶するのではなく、「犬は動物である」「動物には毛がある」という二つの関係から推論的に導く方法によって、記憶容量を節約しているとの主張である。

しかし、意味ネットワークには重要な限界があった。推論の計算的効率性の問題もさることながら、より根本的には「どのような関係をノード間に定義するか」「どのレベルの粒度で概念を分割するか」という設計上の決定が、ネットワークの有効性を大きく左右するということであった。異なるネットワーク設計は異なる推論結果をもたらし、そのいずれが「正しい」かを判断する原則が欠けていたのである。

意味ネットワークの理論的な拡張は、1980年代を通じて進行する。特に重要なのが「オントロジー（ontology）」という概念の浮上である。オントロジーは、単なる知識表現の形式ではなく、「特定の領域について何が存在するか」「それらの存在物がいかなる関係にあるか」を明示的・形式的に定義する知識体系である。この用語が計算機科学の脈絡で本格的に使用されるようになるのは1990年代のことであり、トム・グルーバー（Tom　Gruber）の1993年の著作『知識共有のためのオントロジー設計原則に向けて（Toward Principles for the Design of Ontologies Used for Knowledge Sharing）』によってその重要性が確立された。しかし、その知的根源は確実に1970年代の意味ネットワーク研究に遡ることができるのである。

意味ネットワークとオントロジーの系譜は、知識表現の「構造化」と「形式化」の漸進的な努力の歴史として理解できる。自由形式の記号処理から出発した初期AI が、次第により厳密で相互運用可能な知識表現を求めるようになったのは、実践的な必要性に駆られてのことであった。単一のAIシステム内での知識利用から、複数のシステム間での知識共有へと問題領域が拡張するにつれて、より明示的で共有可能な形式が不可欠となったのである。

6.3　述語論理とProlog ―― 論理プログラミングの台頭

知識表現の多様化と並行して、別の強力なパラダイムが台頭していた。それは述語論理（predicate logic、一階論理）に基づく論理プログラミングである。述語論理そのものは数学基礎論の古典的な道具だったが、それを計算可能な形式に変換し、プログラミング言語として具現化したのは、1970年代の象徴的な成果であった。

述語論理に基づくAI推論の理論的基盤は、1960年代後半の幾つかの重要な発展に遡る。1969年、ジョン・マッカーシーとパトリック・ヘイズは『人工知能の立場からの哲学的問題（Some　Philosophical Problems from the Standpoint of Artificial Intelligence）』を発表し、「状況計算（situation calculus）」と呼ぶ形式体系を導入した。状況計算では、世界の各時点を「状況（situation）」として表現し、アクションの実行によって一つの状況から別の状況へ遷移する過程を記述する。この形式的な枠組みは、ロボット計画生成（robot planning）における推論を数学的に厳密に扱う道を開いた。

しかし、述語論理の AI への適用にはより根本的な問題があった。一般的な一階述語論理は計算上無制限の表現力を持つ一方で、その推論は決定不可能である。つまり、任意の公理系と問い合わせが与えられたとき、その問い合わせが真であるか偽であるかを有限時間で判定するアルゴリズムは存在しないのである。この理論的な桎梏を打破する道を切り開いたのが、1965年のジョン・ロビンソン（John Robinson）による「解法原理（resolution principle）」であった。

解法原理は、与えられた論理式群から矛盾を導き出すことで（背理法の形で）定理を証明する方法である。この方法の決定的な利点は、与えられた前提が一定の制約を満たす場合、特にホーン節（Horn clause）と呼ばれる特殊な形式に限定される場合に、完全な計算手続きを設計しやすくなる点である。ホーン節とは、高々一つの正のリテラル（肯定的な原子式）を持つ選言式のことであり、「P ← Q ∧ R」という形式で表現可能である。この制限的だが十分に表現力のある論理形式こそが、計算可能な論理プログラミングへの道を拓いたのである。

1972年、フランスのエクス・マルセイユ第二大学の AI グループに属していたアラン・コルメロー（Alain Colmerauer）とフィリップ・ルッセル（Philippe Roussel）は、ホーン節の手続き的解釈に基づくプログラミング言語「Prolog」を開発した。Prolog という名称は、ルッセルの妻の提案で、「論理プログラミング（programmation en logique）」の頭文字を取った略語として選ばれたものである。Prolog の開発にあたって、コルメローは当初、フランス語で記述された概念を自動的に処理し、それに対して照会できるツール（自然言語処理）を作成することを目指していた。しかし、その過程で彼は、論理そのものがプログラミングの媒体として機能しうることに気付いたのである。

Prolog の特異性は、命令型プログラミング（「何をせよ」という指示）と宣言型プログラミング（「何が真であるか」という陳述）の境界を曖昧にしたことにある。Prolog プログラムは一連の「事実（fact）」と「規則（rule）」から構成される。事実は「father(tom, bob).」のような原子式であり、規則は「grandfather(X, Z) :- father(X, Y), father(Y, Z).」のような含意式である。問い合わせ「?- grandfather(tom, X).」に対して、Prolog システムは、バックトラッキングと呼ばれる探索戦略を用いて、この含意を満たす X の値を探索する。プログラムの「実行」と「証明」の同一視は、論理に基づくプログラミングの本質を象徴するものであった。

1970年代から1980年代にかけて、Prolog は急速に普及した。特に日本の第五世代コンピュータプロジェクト（第7章）では、Prolog を知識表現と推論の標準言語として採用することが決定され、大規模な投資がなされた。Prolog の利点は、知識を形式的かつ直感的に表現でき、その形式的意味論に基づいて自動的な推論が可能という点にあった。しかし、同時に Prolog の限界も明らかになっていった。第一に、確実性を欠く知識（不確実性推論）の扱いが困難であった。Prolog では「true」と「false」の二値論理が基本であり、確実性が部分的な推論（「おそらく真である」「強く疑わしい」といった推論）をネイティブにサポートしていなかった。第二に、計算効率の問題である。バックトラッキングによる探索は、問題領域によっては指数的な時間を要する可能性があった。

これらの限界を補うために、1970年代後半から1980年代にかけて、不確実性下での推論（uncertain reasoning）を扱うための新しい形式主義が登場することになる。

6.4　不確実性推論 ―― ベイジアンネットワークとデンプスター＝シェーファー理論

AI　システムが現実世界で動作するためには、完全に確実な知識のみに依存することはできない。医学診断から物体認識に至るまで、実装可能なほぼすべての　AI　システムは、不完全で矛盾を含みうる情報に基づいて推論を行わねばならない。この現実的要請に応えるために、1970年代から1980年代にかけて、不確実性を形式的に扱う理論的枠組みが次々と提案された。

不確実性推論の最も初期の実装は、確信度因子（certainty　factor,　CF）と呼ばれるアドホックな数値スキームであった。これは医学診断専門家システム MYCIN（第7章）で 1976 年に導入されたもので、各推論規則に「この規則が適用される場合、結論は何パーセント確実か」を数値（通常 -1 から +1 の範囲）で付与する方法である。MYCIN　　の実装的成功は確信度因子の有用性を示したが、理論的には不十分であった。複数の規則から得られた確信度をどのように組み合わせるか、その組み合わせ規則に論理的な正当性があるかについて、明確な基礎が欠けていたのである。

不確実性推論の理論的な転機は、1980年代初頭の「ベイジアンネットワーク」の登場によってもたらされた。UCLA　　のユダ・パール（Judea　　Pearl）は、1988年に出版された『確率的推論の知的システム（Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference）』において、グラフ理論と確率論を統合した強力な推論形式を提示した。ベイジアンネットワークは、ノードが確率変数を表し、有向辺が条件付き依存関係を表す有向非環グラフである。

ベイジアンネットワークの革新性は、複雑な結合確率分布を、小規模な条件付き確率テーブル（conditional probability table, CPT）の乗積として分解できることを示した点にある。これは計算論的には大きな利点をもたらした。従来の確率論的推論では、n 個の変数を扱う場合、指数的に増加する状態空間に対応する 2^n 個のパラメータが必要であった。しかし、ベイジアンネットワークの構造が局所的な依存関係を正確に反映していれば、必要なパラメータの数は大幅に削減される。さらに、パールが開発した「信念伝播（belief　propagation）」と呼ばれる推論アルゴリズムは、木構造やそれに近いグラフでは効率的に周辺確率を計算することができた。

パールの貢献の重要性を過大評価することはできない。彼は確率的推論を単なる数値計算の問題ではなく、グラフ構造による知識表現の問題として再定義したのである。このパースペクティブの転換により、「独立性構造」「d-分離」「有向非環グラフの性質」といった形式的・グラフ的な概念が推論の中心に据えられることになった。これは、従来の記号主義的知識表現（フレーム、スクリプト、セマンティッ

クネットワーク）と統計的推論との統合を理論的に可能にした。後の統計的機械学習時代（第9章）において、グラフィカルモデル、条件付きランダムフィールド、構造学習といった発展の基礎が、パールのベイジアンネットワークに遡ることができるのである。

一方、ベイジアンネットワークと競合する不確実性推論の形式として、「デンプスター＝シェーファー理論（Dempster-Shafer theory, DST）」もまた注目を集めていた。アーサー・デンプスター（Arthur P. Dempster）による統計的推論の文脈での理論的提案を、グレン・シェーファー（Glenn Shafer）が証拠の一般的な数学理論へと発展させたこの理論は、ベイジアンアプローチとは異なる不確実性の扱い方を提供する。

デンプスター＝シェーファー理論の特徴は、「どの仮説が真であるか確定できない無知」を「信念度（belief）」と「妥当性度（plausibility）」という二つの尺度で表現する点にある。ベイズ確率では、知識が不足している場合でも各仮説に確率を割り当てる必要がある。これは、一度も投げていないコインの表裏に対してもなお 0.5 の確率を仮定することに相当し、この強い仮定が合理的でない場合がある。これに対して、デンプスター＝シェーファー理論は、複数の証拠源から得られた不完全な情報を組み合わせるとき、確実に述べられない信念を「未割り当て（unassigned）」のまま保持することを許容する。この柔軟性は、複数のセンサーからの矛盾した情報や、専門家の意見が分かれている場合など、現実的な多くの局面において理論的には有利である。

しかし、デンプスター＝シェーファー理論にも実装上の課題がある。仮説の空間が大きくなるにつれて、Dempster の合成規則の計算複雑度が指数的に増加するという問題である。このスケーラビリティの課題が、当時のコンピュータの性能制約と相まって、ベイジアンネットワークほどの広範な採用には至らなかったのである。結果として、1990年代の統計的機械学習への転換の中では、ベイジアンアプローチ（および後の graphical models）が主流となることになる。

不確実性推論の諸理論は、AI研究における根本的なパラダイムシフトを象徴している。初期AI の決定論的な記号操作から、統計的・確率的な手法への移行は、単なる形式的な選択肢の問題ではなく、「知能とは何か」という根本的な問い自体の再設定であった。確実性のない世界での有効な推論こそが、真の知能の本質であるという認識が、1980年代後半から次第に支配的になっていくのである。

6.5　知識獲得ボトルネック問題

本章で述べてきた知識表現の様々な形式――フレーム、スクリプト、セマンティックネットワーク、述語論理、確率的グラフモデル――は、すべてある共通の問題に直面していた。それは「知識をどのように機械に入力するか」という極めて実践的だが、同時に根本的な困難である。この問題は、1980年代以降「知識獲得ボトルネック（knowledge acquisition bottleneck）」と呼ばれるようになり、AI研究における最大の制約として認識されるようになった。

知識獲得の困難さは、ダイナミックに機械の性能を制限する構造的な要因に由来していた。第一に、人間の専門知識は多くの場合「暗黙的（tacit）」であり、明示的に言語化することが困難である。医学診断の専門家が「この患者の症状の組み合わせはがんの可能性が高い」と判断する過程には、医学教科書には書かれていない経験的な「勘」や「直感」が含まれている。この暗黙的知識を「規則」や「フレーム」といった形式的な表現に変換するプロセスは、専門家と知識エンジニアの間での綿密で時間的な対話を必要とした。

第二に、知識の領域的複雑性である。たとえば医学診断システム MYCIN では、感染症診断のための知識を取得するために、医学の専門家と人工知能の研究者が何年にもわたって共同作業を行った。採取された規則の数は数百に及び、各規則の確信度を決定するプロセスだけで多大な時間を要した。この労力の賦課が、たとえ一時的には診断性能において優れた結果をもたらしたとしても、その知識ベースを保守・更新することは極めて困難であった。

第三に、獲得した知識の「脆弱性」と「文脈依存性」である。ある領域で慎重に構築された知識ベースは、その領域内では有効に機能しても、わずかな変動や新しい状況に対しては極めて脆弱であった。フレーム理論やスクリプト理論では、「標準的な」状況を想定した構造を構築することはできたが、想定外の変動や例外的な状況に対応する柔軟性が不足していた。この脆弱性は、後にエキスパートシステムの限界（第7章）として顕在化することになる。

知識獲得ボトルネックを打開しようとするアプローチは、大別して二つあった。一つは、知識獲得そのものを自動化・半自動化しようとする「機械学習」的なアプローチであり、もう一つは、表現形式や推論メカニズムそのものをより形式的・一般的に設計しようとする「知識工学」的なアプローチである。

前者の方向は、主に1980年代後半から1990年代の機械学習の発展（第9章）につながっていく。データから自動的に規則やパターンを抽出できれば、人間の専門家への依存度を低減できるという発想である。後者の方向は、オントロジー設計や知識表現形式の形式化の推進へと向かった。より厳密で相互運用可能な知識表現の枠組みを構築できれば、異なるシステム間での知識の転用や、知識ベースの自動生成・検証が可能になるという期待に基づいていた。

しかし、1980年代の時点では、この二つのアプローチのいずれも、知識獲得ボトルネックを根本的に解決するまでには至っていなかった。むしろ、この問題は、1980年代後半のエキスパートシステム産業の衰退（第8章）と第二次AIの冬を招く主要因の一つとなる。知識獲得ボトルネックは、AI研究が「知識」という概念に直面する限り、常に存在し続ける根本的な課題であり、2020年代のLLM 時代においても、依然として「学習データの質と量」「学習後の知識の更新」という形で繰り返し現れ続ける問題なのである。

本章では、1970年代から1980年代にかけての知識表現と推論の体系化の過程を検討した。フレーム理論（ミンスキー、1974）とスクリプト（シャンク、1977）は、構造化された知識表現の初期の成功事例であり、その後のオントロジー研究の思想的源泉となった。意味ネットワークは、グラフ的な知識表現の可能性を示し、後のセマンティックウェブの基礎となった。述語論理と Prolog は、形式的・宣言的な知識表現と自動推論を可能にしたが、不確実性の扱いに課題を残した。ベイジアンネットワーク（パール、1988年）とデンプスター＝シェーファー理論は、不確実な世界での推論を理論的に基礎付けた。しかし、これら多様なアプローチに共通する問題として、知識獲得ボトルネックが立ちはだかっていた。この問題の深刻さは、次章で論じるエキスパートシステムの実装とその後の衰退の過程で明らかになっていく。

参考資料（本章）

本文中の主要な記述を追跡するため、最小限の典拠を挙げる。

- Marvin Minsky, “A Framework for Representing Knowledge” (MIT AI Memo, 1974; book chapter, 1975).
- Roger C. Schank, “Conceptual Dependency: A Theory of Natural Language Understanding” (1972); Roger C. Schank and Robert P. Abelson, Scripts, Plans, Goals and Understanding (1977).
- M. Ross Quillian, “Semantic Memory” (1968); Allan M. Collins and M. Ross Quillian, “Retrieval Time from Semantic Memory” (1969).
- John A. Robinson, “A Machine-Oriented Logic Based on the Resolution Principle” (1965).
- Prolog Heritage project. Prolog の起源と 1972年マルセイユ版の確認に使用。
- Judea Pearl, Probabilistic Reasoning in Intelligent Systems (1988).
- Glenn Shafer, A Mathematical Theory of Evidence (1976).

第7章　エキスパートシステムの隆盛と限界

導入

1970年代から1980年代にかけて、AI研究は新しい方向性を見出した。第5章で述べたように、初期の汎用問題解決プログラムは期待に応えられず、研究資金は削減され、AI研究は冬の時代を迎えていた。しかし同時に、一つの新しいパラダイムが急速に台頭していた。それが「エキスパートシステム（Expert System）」である。

知識の形式化と表現を中心とするこのアプローチは、第6章で述べた知識表現と推論の体系化を実装の形へと変え、実際のビジネス領域で成果を挙げ始めた。1970年代から1980年代前半は、エキスパートシステムが学術的に注目され、商業的に投資されていた黄金期であった。しかし同時に、この時期の成功は、より深刻な構造的問題を隠蔽していた。知識獲得の困難、保守の非現実性、領域外での脆弱性――これらの問題が表面化するのは1980年代後半のことである。

本章では、エキスパートシステムの誕生から隆盛、そして限界の露呈までの軌跡を追う。また、この時期の日本の第五世代コンピュータプロジェクトは、エキスパートシステムへの国家的投資を象徴するものとして詳述する。そして、この章はやがて来る第二次AIの冬への橋渡しとなる。

7.1　DENDRAL ―― 最初のエキスパートシステム（1965–）

起源と動機

エキスパートシステムという用語の厳密な定義については議論があるが、通常、最初期の成熟したエキスパートシステムとして挙げられるのが、スタンフォード大学で1965年頃から開発された**DENDRAL**である。

DENDRALが誕生した背景には、具体的な科学的問題があった。スタンフォード大学の化学者カール・ジェラシ（Carl Djerassi）、遺伝学者ジョシュア・レダーバーグ（Joshua Lederberg）、そしてAI研究者エドワード・ファイゲンバウム（Edward Feigenbaum）の三者の共同研究である。1965年、ファイゲンバウムはレダーバーグに対し、「科学的仮説の自動形成に適した問題を探している」と相談した。レダーバーグが提案したのが、**有機化合物の構造を質量分析スペクトロスコピーのデータから推定するという問題**であった。

レダーバーグ自身は、NASAの火星探査計画に参加し、火星表面での生命の痕跡を検出するための自動質量分析装置の設計に関わっていた。この経験から、分析データから化学構造を自動推定する必要性が生まれたのである。

システムの構造と特徴

DENDRALの革新性は、三つの層からなる推論構造にあった。

第一段階は**「生成」（generation）**である。与えられた分子式（例えば、$C_6H_{12}O$）に基づいて、化学結合の規則に従い、可能な構造異性体を網羅的に生成する。この段階は、化学的制約を知識として組み込むことで、可能性の空間を大幅に削減した。

第二段階は**「テスト」（testing）**である。**生成された各候補構造について、その構造が与えられた質量分析スペクトロスコピーデータとどの程度整合するかを判定する。スペクトロスコピーでは、分子が衝突によって特定のパターンで断裂するが、この断裂パターン（フラグメンテーション）は分子の構造に依存する。DENDRALには、化学専門家が記述した**多数のフラグメンテーション規則**が組み込まれていた。

第三段階は**「ランク付け」（ranking）**である。スペクトロスコピーデータとの整合度が高い構造候補を上位から順に提示する。

この三段階のアプローチにより、DENDRALは**特定の条件下では熟練した有機化学者に匹敵する水準**で未知の有機化合物の構造候補を絞り込めるようになった。1970年代までには、質量分析データの解釈支援という限定領域で高い有効性を示していた。

エキスパートシステムとしての意義

DENDRALが「最初のエキスパートシステム」と称される理由は、以下の特徴にある。

第一に、知識の明示的な形式化である。化学の専門知識——化学結合の規則、スペクトロスコピーの解釈規則、化学者の推論戦略——を、コンピュータが処理できるIF-THENルールの形で明示的に記述した。これは初期のAIプログラム（Logic Theorist、General Problem Solver）とは根本的に異なっていた。初期プログラムはあらかじめ定められた探索アルゴリズムを用いていたのに対し、DENDRALは**ドメイン固有の専門知識**をシステムの核とした。

第二に、実際の科学的成果である。DENDRALは学術論文の発表、および産業界での利用に至った。1970年代には、実際の研究者がDENDRALを利用して未知物質の構造解析を行い、学会で報告する例もあった。これは、AIシステムが学術的な専門知識領域で実際に使用されたことを意味する。

第三に、後続するエキスパートシステム開発の原型を示したことである。DENDRALの開発過程で得られた知見——知識ベースの構築方法、ドメイン専門家との協働、推論エンジンの設計——は、その後のMYCIN、XCON、R1といったシステムに直接受け継がれた。

初期の限界

しかしながら、DENDRALの成功は限定的でもあった。システムが扱える化合物の種類は、ルール群に組み込まれた知識に依存していた。新しい種類の化合物が現れるたびに、新しいフラグメンテーション規則を手作業で追加する必要があった。この「知識獲得ボトルネック」は、DENDRALのみならず、その後のすべてのエキスパートシステムが共有する根本的な課題となるのである。

7.2　MYCIN ―― 医療診断と確信度因子

開発の背景

1970年代初頭、スタンフォード大学でエドワード・ショートリフ（Edward Shortliffe）による博士論文プロジェクトとして開発されたのが**MYCIN**である。対象領域は医療診断、特に**細菌感染症の同定と抗生物質の推奨**という極めて実践的で重要な領域であった。

血液感染症（bacteremia）と髄膜炎（meningitis）に起因する細菌を特定し、患者の体重に応じた用量調整を加えた抗生物質を推奨する。この領域が選定された背景には、強い実務的ニーズと同時に、診断知識の明示化が可能だと考えられたことがある。医学の中でも感染症診断は、一定のルール性を持つ領域として認識されていた。

確信度因子（Certainty Factors）の導入

MYCINが初期のエキスパートシステムと決定的に異なる点は、**確信度因子（Certainty Factor, CF）**というメカニズムの導入である。

医学診断には本質的な不確実性がある。患者の症状、検査値、背景情報など、個々の証拠はすべて確率的であり、どの証拠も決定的ではない。また、複数の証拠を組み合わせるとき、それらが独立しているのか相関しているのかも不明確である。DENDRALのような確定的な推論エンジンでは、このような不確実性を扱うことができない。

MYCINでは、各ルールに**確信度**が付与された。例えば：

```
IF [患者の血液培養が陽性] AND [細菌がグラム陽性球菌]
THEN [Staphylococcus aureus を支持する CF = 0.8]
```

ここで0.8という値が確信度因子である。確信度因子は-1から+1の値をとり、+1は強い支持、0は無関連、-1は強い反証を意味する。これは確率そのものではなく、ルールが結論をどれだけ支持または反駁するかを表す**経験的な重み付け**であった。0.2未満の確信度を持つ仮説は、計算を継続するに値しないものとして探索を打ち切るという**カットオフ（cutoff）**機構も設けられた。

複数のルール結論から同じ仮説に到達した場合、複数の確信度を組み合わせる計算式が定義された。このアプローチは、厳密なベイズ確率論というより、領域専門家の直感的な確実性評価を計算可能な形にした**実務的ヒューリスティクス**であった。

臨床性能と実装上の問題

MYCINの診断性能は著しく高かった。およそ450〜500個の生産ルール（production rule）を用いて、**限定された感染症診断課題では、専門医と同等水準の推奨を示す**ことが報告された。1970年代末から1980年代初頭にかけて、MYCINは複数の医学雑誌で報告され、医学AI研究の代表的成功例として認識された。

しかし、MYCINが実際に臨床現場で使用されることはなかった。その理由は、単なる技術的限界ではなく、医療業界とAIシステムの間の根本的な不一致にあった。医師は自動化されたシステムを信頼せず、また医療責任の問題も存在した。実験室での診断正確性と、臨床現場での受容可能性とは、別の問題なのである。

構造的意義と後続研究

MYCINの重要性は、実装の成功だけでなく、それが後続のエキスパートシステム研究に提供した**メソドロジーの枠組み**にある。

第一に、不確実性推論の形式化である。確信度因子そのものは数学的には粗いものであったが（ベイズ確率論の厳密性を欠く）、専門家の判断を定量化し、システムに統合する一つの実装的方法を示した。このアプローチは、後のベイズネットワークやデンプスター＝シェーファー理論（第6章）への理論的足がかりとなった。

第二に、説明可能性（explainability）である。MYCINはなぜある診断に到達したのかを、使用されたルール群と確信度の積み重ねとして説明できた。この「説明可能性」は、ブラックボックス的な深層学習とは対照的に、エキスパートシステムの重要な特性となる。その後のNEOMYCINというプロジェクトでは、MYCINの推論過程をより詳細に説明し、教育的に活用することが試みられた。

第三に、知識獲得プロセスの可視化である。MYCINの開発では、感染症の専門医との対話を通じてルール群を構築していった。この過程で、医学的知識がいかに複雑で、暗黙的であるか、そしていかに難しく形式化されるべきかが明らかになった。これが後に「知識獲得ボトルネック」という概念の源泉となる。

7.3　R1/XCON ―― 商用化の成功とDEC社での実績

問題設定と開発

カーネギーメロン大学のジョン・P・マクダーモット（John P. McDermott）によって1978年に開発された**R1**（後にXCON、eXpert CONfigurer）は、エキスパートシステムの商用化における最初の大規模な成功例であった。その対象領域は、**Digital Equipment Corporation（DEC）のVAXコンピュータシステムの構成（configuration）**という、一見地味ながら経済的には極めて重要な問題である。

DECの営業プロセスでは、顧客の要求仕様に基づいて、数千の部品の中から適切なコンポーネントを選択し、システムを構成する必要があった。メモリサイズ、ディスク容量、I/Oインターフェース、周辺機器など、選択肢の数は組み合わせとしてほぼ無限に近い。さらに、コンポーネント間の互換性制約が複雑に絡み合っている。人間の技術者による手作業での構成には、数日を要し、エラーが頻繁に発生していた。顧客にはエラーで不要な部品が届き、最悪の場合、システムが起動しないという事態も発生していた。

システムアーキテクチャと知識表現

R1/XCONは、**OPS5（Official Production System 5）**というプログラミング言語で実装された。OPS5は生産ルール（production rule）の効率的な処理のために設計された言語であり、エキスパートシステムの実装に最適化されていた。

システムは、運用の拡大とともに**数千規模の生産ルール**を持つ知識ベースへと成長した。各ルールは、部品構成の制約、互換性、最適化原理などを表現していた。例えば：

```
IF [顧客がディスク容量10GB以上を要求] AND [予算が限定的]
THEN [リスト価格が最も低いディスク製品を選択]
```

R1の重要な特徴は、**前向き連鎖（forward chaining）**と呼ばれる推論戦略を採用していたことである。既知の事実（顧客の要求仕様）から出発し、マッチするルールを次々と適用して新しい事実（部品選択）を導出していく。これは、医学診断のような後向き連鎖（ゴール駆動型推論）とは異なり、入力データから確実に解に到達する方式を採用したのである。

商用上の成果

R1/XCONが実運用に入った**1980年前後**から、その成果は広く注目を集めた。

- **処理能力**：人間の技術者が数日要する構成タスクを、システムはわずか数時間で完了した
- **精度**：95〜98%の精度でシステム構成を完成させ、人間のエラー率を大幅に低減

- **スケール**：1980年代半ばまでに数万件規模の注文処理に用いられた
- **経済効果**：DECは年間数千万ドル規模の節約効果を見込んだ。これは、不要部品の削減、組立時間の短縮、顧客満足度の向上によるものである

R1/XCONの成功は、企業内でのAIシステム導入を劇的に推進した。他の企業もこれを見習い、同様のルールベースシステムの開発投資を開始した。エキスパートシステムは、もはや学術的な研究対象ではなく、**ビジネス上の価値を実証した商用技術**として認識されるようになったのである。

保守の現実

しかし、R1/XCONの成功の背後には、ほぼ見えない苦労があった。それが**保守（maintenance）**である。

VAXコンポーネントのラインナップは常に変化していた。新型コンポーネントが導入され、旧型は廃止され、互換性も常に変動した。このたびに、R1の知識ベース（数千規模のルール）の更新が必要になった。更新には、DECの技術者への頻繁なインタビューが必須であり、ルール間の相互作用を検証し、新しいルールが既存のルールと矛盾しないことを確認する必要があった。

実際には、システムの保守は、DEC社内の専任チームが継続的に担当していた。この保守コストは、システムの経済効果を相殺しない程度には抑えられていたが、やはり**知識獲得と保守が継続的な負担**であることを示していた。

7.4　LISPマシンと専用ハードウェアの産業

エキスパートシステムとハードウェアの共進化

1980年代のエキスパートシステムの急速な成長は、専用のハードウェア産業をも生みだした。その中心が**LISPマシン**である。

LISP言語は、1958年にジョン・マッカーシーによってAI研究のために開発された言語である（第2章）。1970年代から1980年代にかけて、LISP言語で記述されたエキスパートシステムの処理が急速に増加する。しかし、当時の汎用コンピュータ（IBM System/360系など）上でLISPプログラムを実行することは、計算効率の点で極めて悪かった。汎用コンピュータは、主に数値計算や商用処理を想定して設計されていたからである。

この問題を解決するため、LISPの効率的な実行を目的として専用に設計された計算機、つまり**LISPマシン**が開発された。

LISPマシンの技術的特徴

LISPマシンは、以下の特徴を持つ単一ユーザーワークステーションであった。

第一に、LISP固有の演算をハードウェアレベルでサポートした。LISPでは、リスト（list）の操作、型チェック（type checking）、ガベージコレクション（自動メモリ管理）が頻繁に行われる。LISPマシンは、これらの演算を専用のハードウェアユニットで実行できるよう設計された。例えば、「リストの先頭要素を取得する（CAR演算）」「リストの末尾を取得する（CDR演算）」といった基本的操作は、汎用CPUでは複数のマシン命令を要するが、LISPマシンではハードウェアで直接実装された。

第二に、タグ付きアーキテクチャを採用した。メモリの各ワードに、そのワードが数値なのか、シンボルなのか、リストへのポインタなのかを示す**タグ**が付加されていた。これにより、実行時の型チェックが高速化された。

第三に、インクリメンタルガベージコレクションの実装である。通常、ガベージコレクションは実行を一時停止して行われるため、応答性が損なわれる。LISPマシンでは、ハードウェアサポートにより、ガベージコレクションをバックグラウンドで行いながら計算を継続できるようにした。

Symbolicsと商用展開

LISPマシンの最も成功した商業化は、**Symbolics**という企業による展開であった。SymbolicsはMIT AI Labの研究者・技術者らによって1980年に創立され、LISPマシンの高級ラインを製造していた。

Symbolicsの主力製品は、1983年に発表された**3600シリーズ**である。これは高性能な単一ユーザーワークステーションであり、LISPマシンの中では最も洗練され、最も高性能であった。動作クロック周波数は数MHz程度（当時の汎用ワークステーションと同程度）であったが、LISPの実行に特化した設計により、汎用マシンよりもはるかに高速なLISP実行性能を実現していた。

他にも、Lisp Machines Inc.（LMI）の**Lambda**、Texas Instrumentsの**Explorer**、Xeroxの**Interlisp-D**ワークステーションなど、複数の企業がLISPマシンを製造していた。

LISPマシン産業の展望と限界

1985年頃までのLISPマシン産業は、急速に成長していた。エキスパートシステムの開発ブーム、特にR1/XCONの成功により、企業における AI開発への投資が増加していたからである。LISPマシンは、AIシステム開発の最適なプラットフォームとして認識され、多くの企業や研究機関で採用された。

しかし、この産業の長期的な展望は、当初の楽観的予測とは異なっていた。より詳しくは第8章で述べるが、1980年代後半のマイクロプロセッサの革命的進歩により、汎用ワークステーション（特にSun Workstations）がLISPの実行速度において追いつき、やがて追い越すようになった。また、マイクロプロセッサの価格性能比の改善は、LISPマシンの高価格を正当化しなくなった。結果として、LISPマシン産業は1990年代初頭までに衰退していくのである。

7.5　日本の第五世代コンピュータプロジェクト（1982–1992）

国家的野心と制度的背景

1982年、日本の通商産業省（MITI）は、一つの野心的なプロジェクトを開始した。**第五世代コンピュータシステム（Fifth Generation Computer Systems, FGCS）**と呼ばれるこのプロジェクトは、エキスパートシステムと知識処理を基盤とした、次世代コンピュータを開発することを目的としていた。

背景には、以下の戦略的考慮があった。当時、米国ではエキスパートシステムが急速に商用化され、経済的価値を生み出しつつあった。IBM、DARPA、シリコンバレーの企業群が、AI技術の開発競争に熱中していた。日本が「第四世代」（汎用コンピュータ：IBM互換機）の世界でIBMに後れを取ったという認識のもとで、AI技術という「第五世代」において日本が主導権を握ることを目指した戦略的決定であった。

プロジェクトの中心機関として**新世代コンピュータ技術開発機構（Institute for New Generation Computer Technology, ICOT）が設立された。予算規模は約540億円**に上り、10年間のプロジェクト期間が設定された。なお、FGCS本体は1992年に終了し、その後もICOT自体は1994年まで存続した。

技術的目標と研究方針

FGCSプロジェクトの技術的目標は「知識ベース情報処理に向かった新しいコンピュータ技術に関する研究開発」と定義された。より具体的には、以下の三つの要素を含んでいた。

第一に、ロジックプログラミングの並列化である。Prolog（述語論理を基盤とする宣言的プログラミング言語、第6章）を極めて高速に実行し、かつスケーラビリティを備えた並列推論マシンを開発することが目標であった。

第二に、大規模並列ハードウェアの設計・製造である。FGCSで開発された並列推論マシン（PIM, Parallel Inference Machine）は、数百から数千のプロセッサを搭載し、推論タスクを並列処理できる設計を目指していた。

第三に、知識表現と推論の基盤技術の確立である。Prologの拡張、制約論理プログラミング（constraint logic programming）、非単調推論、知識データベース管理システムなど、知識処理の理論的・実装的基盤を構築することであった。

成果物と技術的貢献

FGCSプロジェクトが生み出した具体的成果は、以下の通りである。

ハードウェア：複数世代のPIMが開発された。最も先進的なモデルは、512個のプロセッサを搭載する**PIM/p**、および256個のプロセッサを搭載する**PIM/m**である。これらは、当時としては極めて大規模な並列計算機であった。ただし、外部との接続性やソフトウェア生態系の整備では、後発的な限界があった。

プログラミング言語と処理系：並行論理プログラミングの基盤として**GHC**（Guarded Horn Clauses）が提案され、その実装言語として**KL1**（Kernel Language 1）が整備された。KL1は、ロジックプログラミングの宣言的特性を保ちながら、並列実行を可能にする言語として設計された。

理論的成果：並列ロジックプログラミングの理論的基礎が確立された。特に、**ガード（guard）**の概念や、非決定性の制御に関する理論的研究は、その後の並列計算理論に貢献した。

商用化の困難と評価

しかし、FGCSプロジェクトが目指した「商用レベルの第五世代コンピュータ」は、実現しなかった。以下の複合的な要因がこれに寄与した。

第一に、ソフトウェア生態系の不在である。PIMは高性能な推論マシンであったが、それを活用するアプリケーションやツールの開発が不十分であった。欧米の企業が既に構築していたエキスパートシステム開発環境（例えば、KEEやARTなど）に相当するものが、FGCSプロジェクト内では十分に広がらなかった。

第二に、Prolog自体の限界が明らかになっていたことである。Prologは理論的には強力であるが、実務的なエキスパートシステム開発では、OPS5などの生産ルールベースの言語の方が、より自然で効率的であることが判明していた。FGCSが過度に理論的なアプローチに依存していたのに対し、実務界ではより実用的なパラダイムが求められていた。

第三に、米国でのLISPマシン産業が急速に衰退し始めたことである。並行してFGCSプロジェクトが進行していた1980年代後半から1990年初頭は、ちょうどAIハードウェア産業全体が衰退の危機に直面していた時期であった。このタイミングの悪さは、プロジェクトの商用化を大きく阻害した。

第四に、汎用ワークステーション技術の急速な進化である。Intelの80386、i486、Pentium、そしてSun Microelectronicsなどの汎用ワークステーションの性能が指数関数的に向上していた。結果として、専用のAIハードウェアの価格性能比の優位性が失われていった。

しかし、FGCSプロジェクトの評価は、商用成功の有無だけで判断されるべきではない。**並列ロジックプログラミングの理論的基礎、大規模並列アーキテクチャの設計経験、知識処理システムの研究**などにおいて、プロジェクトは重要な知的遺産を残した。特に、KL1やGHCは、その後の並列プログラミング言語の発展に影響を与えた。また、日本の研究機関での研究人材育成という点でも、プロジェクトは一定の役割を果たした。

1992年のプロジェクト終了時、日本はこれまでのAI推進戦略の見直しを迫られることになった。しかし、この見直しの中で、FGCSが培った研究人材と理論的基盤が、その後の日本のコンピュータサイエンス研究を支えることになるのである。

7.6　エキスパートシステムの脆弱性と保守問題

知識獲得ボトルネックの実態

1980年代中盤から後半にかけて、エキスパートシステム業界で一つの深刻な問題が顕在化していた。それが**知識獲得ボトルネック（knowledge acquisition bottleneck）**である。

エキスパートシステムの開発には、以下のプロセスが必要である：

1. **領域分析**：対象領域の専門知識を概観し、対応すべき問題を明確化する
2. **知識抽出（知識獲得）**：領域の専門家にインタビューし、彼らの推論過程・判断基準・ノウハウを言語化させる
3. **形式化**：抽出した知識をIF-THENルールや知識グラフの形に変換する
4. **実装**：形式化された知識をプログラミング言語（例：OPS5）で実装する
5. **検証**：実装されたシステムが、専門家の判断と一致するかをテストする
6. **調整と反復**：テスト結果に基づいてルール群を修正し、再度テストする

このプロセス全体を通じて、**知識獲得（第2から第3ステップ）が最も困難で時間を消費する段階**であることが明らかになった。

知識獲得の困難さの原因

知識獲得が困難である理由は、単に「知識が複雑だから」ではなく、より深い認識論的問題にあった。

第一に、暗黙的知識の言語化の困難である。熟練した専門家の知識の大部分は、彼自身が明確に意識していない暗黙的な形をしている。医師が患者を診察するときに、その医師自身が意識しない複数のシグナルを同時に処理している。化学者が分子構造を推定するとき、スペクトロスコピーのデータをどのような重み付けで解釈しているのか、本人も完全には説明できない。この暗黙的知識を、明示的な規則として言語化することは、極めて困難である。

第二に、領域固有の文化と専門用語である。各領域には独自の概念体系と言語がある。医学用語、化学用語、工業用語などは、その領域内部では自明であるが、領域外の知識工学者には理解しにくい。また、専門家が「常識」と見なす背景知識を、外部者が完全に習得することは困難である。

第三に、知識の動的性質である。実務的な領域の知識は、決して静的ではない。新しい研究成果が発表される、業界の標準が変わる、顧客のニーズが変化する。このたびに、知識ベースの更新が必要になる。R1/XCONの場合、VAXコンポーネントの変更のたびに、知識工学者がシステムを再度調整する必要があった。

知識エンジニアの役割と限界

このような困難に対応するため、「知識エンジニア（knowledge engineer）」という新しい職業が誕生した。知識エンジニアは、領域の専門家と協働し、専門知識をコンピュータが処理可能な形に変換する役割を担った。

しかし、知識エンジニアの仕事は極めて困難であった。理想的には、知識エンジニアは、領域の深い理解と、知識表現の形式的理解の双方を持つべきであり、同時に人的コミュニケーション能力にも優れていなければならない。このような人材は極めて稀であった。

また、知識エンジニアがシステムを開発する過程で、新しい知識が発見されることもある。つまり、専門家自身も認識していなかった知識が、インタビューと形式化のプロセスの中で浮き彫りになることがあるのである。これは知識工学の価値の一つであるが、同時に、プロジェクトの予測可能性を低下させた。

保守と更新の非現実性

知識獲得ボトルネックは、システムの初期開発段階だけの問題ではなかった。むしろ、システムが稼働を開始した後の**保守と更新**においてより深刻な問題となった。

R1/XCONの事例で見たように、ビジネス環境や技術環境が急速に変化する領域では、エキスパートシステムの知識ベースも常に更新する必要がある。しかし、既存の知識ベース（特に数千規模のルール）を修正することは、極めて危険である。一つのルールを変更すると、他の多くのルールとの相互作用が変わり、予期しない副作用が生じることがある。

この問題はしばしば**「スパゲッティ効果（spaghetti　effect）」**と形容された。複雑に絡み合ったスパゲッティのようなルール相互作用の中で、一つのルール変更が全体にどのような影響を与えるかを予測することは、事実上不可能になるのである。

ブリットルネス（脆弱性）

エキスパートシステムのもう一つの根本的な限界は、その**ブリットルネス（brittleness）**である。これは、システムが学習した領域内では高い性能を発揮するが、領域外の問題に対しては、時に劇的な失敗をすることを意味する。

医学診断システムが感染症診断では優秀でも、外科手術の判断には適用できない。化学構造解析システムが有機化合物には有効でも、無機化合物や新しい有機化合物には失敗することがある。汎用性の欠如、新規性への対応能力の欠如が、エキスパートシステムの構造的限界なのである。

この限界は、人間の知能の特徴の一つである「汎化（generalization）」能力と対照的である。人間は、特定領域での経験を、他の領域に活かすことができる。しかし、ルールベースシステムには、このような抽象的な概念レベルでの汎化能力がない。

1980年代後半から1990年初頭のエキスパートシステム市場の崩壊

これらの構造的問題が徐々に認識されるにつれ、エキスパートシステム市場は急速に縮小していった。以下が市場崩壊の主要な要因である。

第一に、期待と現実のギャップである。1980年代初期には、エキスパートシステムが、経営意思決定から医療診断、科学研究まで、あらゆる領域で汎用的に応用されるという楽観的な見方が広がっていた。しかし、実際には、特定の狭い領域でのみ有効性を発揮し、汎化可能性が著しく限定されていたのである。

第二に、保守コストの上昇である。システムが老朽化し、知識ベースが肥大化するにつれ、保守に要する費用が増加した。特に、ビジネス環境が急速に変化する企業では、保守コストがシステムから得られる利益を上回るようになった。

第三に、新しいパラダイムの出現である。1980年代後半から1990年代初頭にかけて、ニューラルネットワークやコネクショニズムの研究が復活し始めた（第8章）。データから自動的に知識を獲得できるニューラルネットワークアプローチは、知識エンジニアによる手作業を避ける可能性を提供しているように見えた。

第四に、ハードウェア環境の劇的な変化である。汎用ワークステーションと汎用プロセッサの急速な性能向上により、専用のAIハードウェア（LISPマシンなど）の必要性が失われた。これにより、エキスパートシステム開発環境そのものの経済的基盤が失われた。

構造的教訓

エキスパートシステムの隆盛と衰退は、AI研究に対する重要な教訓をもたらした。

第一に、「知識は獲得可能である」という仮定の限界である。エキスパートシステムは、領域専門家の知識をシステマティックに抽出し、形式化できるという信念に基づいていた。しかし、実際には、知識の暗黙性、領域の動的性質、知識の相互依存性などの問題が、この信念の実現を困難にした。

第二に、「知識があれば問題は解ける」というアプローチの限界である。問題解決には、知識だけでなく、学習能力、環境への適応能力、類推能力、常識的推論能力など、複数の認知能力が必要である。これらの能力をすべて記号ベースの知識表現で表現することは、実質的に不可能であることが明らかになった。

第三に、「ドメイン特化的システムは汎用システムよりも優れている」という前提の限界である。確かに、狭い領域での高い性能は、エキスパートシステムの強みであった。しかし、経済的には、多くの企業が必要としているのは、「すべての領域で一定水準の性能を発揮できるシステム」であり、「医学診断では高性能だが、工業設計では使えないシステム」ではないのである。

これらの教訓は、1990年代以降のAI研究を方向づけることになる。統計的学習、機械学習、データドリブンなアプローチへのシフトは、エキスパートシステムの限界を乗り越えるための、必然的な応答だったのである。

第8章への橋渡し

エキスパートシステムの時代は、1980年代中盤から後半にかけて、その終焉を迎えつつあった。R1/XCONの商用成功、LISPマシン産業の成長、そして日本の第五世代プロジェクトという国家的投資は、一見すれば、AIが新しい黄金期を迎えたかのように見えた。しかし、その成功の底流には、構造的な限界が潜んでいた。知識獲得ボトルネック、システムの脆弱性、保守の困難さ。そして、急速に進化する汎用ハードウェアと、新しい学習パラダイムの出現。

1987年から1989年にかけて、AI産業は崩壊へと向かう。LISPマシン企業は経営危機に直面し、エキスパートシステムプロジェクトは次々と放棄されていく。一見、AI研究全体が失敗に向かっているかのように見えた。これが「第二次AIの冬」である。

しかし、この冬は、単なる産業的失敗ではなく、認識論的な転換点でもあった。記号主義的な知識表現から、統計的機械学習へ。手作業による知識獲得から、データからの自動学習へ。そして、汎用コンピュータ上でのスケーラブルなアプローチへ。

次章では、この劇的な転換を語ることになる。

まとめ

本章では、1965年から1980年代後半にかけてのエキスパートシステムの時代を概観した。DENDRALから始まる領域固有の知識表現、MYCINによる確信度因子の導入、R1/XCONによる商用化の成功、LISPマシン産業の興隆、そして日本の第五世代プロジェクトの野心的な取り組み。これらは、AI研究が特定の領域で実質的な成果を上げた時代を記録している。

同時に、この時代は、記号主義的アプローチの構造的限界が明らかになっていった時代でもあった。知識獲得ボトルネック、システムのブリットルネス、保守の非現実性。これらの問題は、エキスパートシステムパラダイムの本質的な弱点であり、単なる実装上の問題ではなかった。

AI史の中で見れば、エキスパートシステムの時代は、記号主義的AI（1956–1980年代）の最後の繁栄であり、同時に、その限界を最も明確に露呈させた時期でもあったのである。この露呈が、その後のパラダイムシフトを招来することになる。

参考資料（本章）

- Edward A. Feigenbaum, Bruce G. Buchanan, and Joshua Lederberg, DENDRAL 関連論文群。DENDRAL の起源と知識工学的構成の確認に使用。
- Stanford Medicine / Edward H. Shortliffe 関連資料および Edward H. Shortliffe, Computer-Based Medical Consultations: MYCIN (1976). MYCIN の対象領域、確信度因子、評価実験の確認に使用。
- John P. McDermott, “R1: A Rule-Based Configurer of Computer Systems” (1982). R1/XCON の問題設定とアーキテクチャの確認に使用。
- Randall Davis and Douglas B. Lenat, Knowledge-Based Systems in Artificial Intelligence (1982). 知識獲得ボトルネックと保守問題の整理に有用。
- Computer History Museum および MIT 関連史料。Symbolics と LISP マシン産業の成立過程の確認に使用。
- 情報処理学会（IPSJ）等の FGCS / ICOT 回顧資料。FGCS の期間、KL1 と GHC の関係、PIM の位置づけの確認に使用。

第8章　第二次AIの冬 —— バブル崩壊と再出発

8.1　LISPマシン市場の崩壊（1987–）

1980年代のAI産業は、エキスパートシステムの商業化に牽引される好況にあった。その中心を占めたのが、LISP言語向けに特化した専用ハードウェア、すなわちLISPマシンである。Symbolics、LISP Machines Inc.、Xerox、Texas Instrumentsといった企業は、AIプログラムの実行を最適化する専用ハードウェアを開発し、これを研究機関や企業に販売した。1986年時点では、Symbolicsの売上は1億1500万ドルに達し、AI産業全体がテクノロジー投資の一大領域となっていた。

しかし、1987年前後から、この市場は急速に収縮する。その引き金は、半導体テクノロジーと汎用ワークステーションの急速な発展であった。Sun Microsystems などのRISCワークステーションや汎用マイクロプロセッサは、ムーアの法則に従って性能を向上させ、LISPマシンの価格性能比上の優位を侵食していった。こうして、高価で専門的なLISPマシンの必要性は数年のうちに急速に薄れていった。

Symbolicsは1986年の高収益から数年で経営が悪化し、Lisp Machines Inc.も1980年代末には行き詰まった。Texas InstrumentsとXeroxもこの市場から撤退し、LISPマシン産業は1990年代初頭までにほぼ終息した。長年の投資と期待は、数年のうちに急速にしぼんでいったのである。

LISPマシン市場の崩壊は、単なる経済的出来事ではなく、AI研究の知的・技術的基盤に対する一つの信号であった。ハードウェア最適化に依存したアーキテクチャは、汎用性能の前に無効化されたのだ。この教訓は、後のAI研究がハードウェア依存からの独立を考慮する際に重要となる。

8.2　第五世代プロジェクトの終焉と評価

LISPマシン市場の危機よりも早く、日本の第五世代コンピュータプロジェクト（Fifth Generation Computer Systems, FGCS）の限界が明らかになっていた。1982年から1992年にかけて、日本の通商産業省（MITI）は総額約540億円を投じ、知識基盤型コンピュータの開発を推進した。その目的は、大規模並列処理とロジックプログラミングを基盤とした「時代を先導するコンピュータ」を実現し、AI時代の覇権を確立することにあった。

第五世代プロジェクトの技術的標的は、従来のフォン・ノイマン型アーキテクチャからの根本的な転換であった。プロジェクトは、Prolog系の並行論理プログラミングを活用した並列推論マシンの開発に注力し、PIM系の試作機を複数世代にわたって実装した。研究そのものは、並列論理プログラミングの理論的発展に貢献するなど、学術的には意義があった。しかし、商業化可能な製品への転化という点では、プロジェクトは成功しなかった。1992年にFGCS本体が終了した時点で、その成果は主として理論的知見と研究基盤の形成にとどまり、世界市場を変えるまでには至らなかった。

第五世代プロジェクトの失敗が示した教訓は複層的である。第一に、ハードウェア最適化による性能向上の追求は、汎用プロセッサの指数関数的な進化に対抗できないということであった。第二に、知識ベース型システムの本質的な困難さ——知識の表現、獲得、保守の複雑性——は、ハードウェア改善のみでは解決不可能であった。第三に、国家主導の大規模プロジェクトであっても、科学的・技術的なパラダイムの根本的な過誤は補正できないということである。

第五世代プロジェクトはまた、国際的波及効果をもたらした。このプロジェクトの発表は、世界中のコンピュータ産業と政府に対し、「並列処理が将来の性能向上の源泉である」という認識を強く印象づけた。米国のStrategic Computing InitiativeやMCC、英国のAlveyプロジェクト、欧州のESPRITなどは、それぞれ独自の背景を持っていたが、FGCSを重要な外部刺激の一つとして受け止めたと広く論じられている。皮肉なことに、日本の野心的なプロジェクト自体は商業的成功に至らなかったが、それが喚起した国際的な競争は、各地域のAI・コンピュータ科学研究に複雑な遺産を残したのである。

8.3　記号AIの限界に対する認識の深化

LISPマシンと第五世代プロジェクトの失敗の根底にあったのは、記号主義的アプローチ（symbolic AI）そのものの構造的限界であった。第一次の冬（第5章）では、組合せ爆発問題とフレーム問題により記号推論システムの無力さが露呈していたが、エキスパートシステムの一時的な成功によって、この問題は表面上は「解決可能」なものとして扱われていた。しかし、1980年代後半から1990年代初頭にかけて、この楽観主義の虚構が明らかになる。

エキスパートシステムの実装と運用において、最大の障害となったのが「知識獲得ボトルネック（knowledge acquisition bottleneck）」である。理論上は、ドメイン専門家のルールを記号形式で表現すればシステムが構築できるはずであった。しかし現実には、MYCINが感染症診断の専門家知識を格納するのに複数年を要したように、ルール化可能な知識の抽出は労力集約的であり、知識エンジニアと専門家の継続的協力を必要とした。

さらに深刻だったのは、メンテナンスの問題である。DECのXCON（構成管理システム）のような成功事例でも、知識ベースが数千のルールに成長するにつれて、各ルール間の整合性の維持は困難を増していった。新しい知識を追加しても、既存のルールとの相互作用が予測不可能な結果を生じさせることが頻繁に起こった。その結果、保守コストは次第に運用上の利益を上回るようになった。1980年代後半には、多くの企業がかつて多額の投資をした既存のエキスパートシステムの維持を放棄するか、急速に縮小させるようになった。

記号AIのもう一つの根本的欠陥は、その「脆弱性（brittleness）」にあった。ニューラルネットワークは部分的な損傷や入力の変動に対してある程度のロバスト性を示すが、記号システムは仕様の範囲外の入力に対して極度に脆弱である。学習用に限定された領域では高い性能を示しても、わずかに異なる問題が提示されると、システムは完全に機能不全に陥る。この欠陥は、人間の知能の柔軟性と適応性との隔たりを如実に示していた。

これらの実務的な困難さは、同時に理論的な認識の転換をもたらした。1980年代後半のAI研究コミュニティ内では、「常識知識（common sense knowledge）」の必要性と、そのような知識をいかにして獲得・表現するかという問題について、より深い悲観的認識が広がった。マービン・ミンスキーやロジャー・シャンクといった初期AIの有力研究者たちでさえ、「AIの挫折の根本的原因は、シンボルの意味と現実世界の対応を定義することの困難さにある」という趣旨の発言をするようになった。記号主義は、知能の全体を説明する理論的枠組みではなく、むしろ特定領域での最適化に過ぎないという認識が、徐々に主流派内にも浸透していったのである。

8.4　コネクショニズムの復活 —— PDP研究グループ（1986）

1986年は、AI史における転換点である。記号AIの限界が深刻化する同じ時期に、ニューラルネットワーク研究に対する関心が劇的に復活する。その象徴が、デイビッド・ラメルハート、ジェームス・マクレランド、およびPDP（Parallel Distributed Processing）研究グループによる『Parallel Distributed Processing: Explorations in the Microstructure of Cognition』の出版である。

PDP研究グループは、1980年代初頭からカーネギーメロン大学とカリフォルニア大学サンディエゴ校を拠点として、ニューラルネットワークとコネクショニズムの理論的・実験的研究を進めていた。ラメルハートとマクレランドを中心とするグループは、認知科学と神経科学の文献から着想を得て、並列分散処理という統一的枠組みで複数の認知機能をモデル化する試みを推進していた。

1986年に刊行された二巻本『PDP』は、単なる技術書ではなく、AIとコグニティブサイエンスの基本的パラダイムの再構成を提示する著作であった。その核心は、知識と認知機能を「分散された相互接続のパターン」として理解する観点にあった。記号主義がシステムの計算を記号の操作と見なすのに対し、コネクショニズムは基本単位をニューロン風の単純な計算ユニットとし、これらの大量の並列相互作用によって知能的挙動が創発されるものと見なした。この枠組みは、脳の神経生物学的現実に一層接近したものと考えられ、また計算可能性の面でも有利であると期待された。

PDP研究グループの登場は、1969年のミンスキー＝パパート『Perceptrons』以降停滞していたニューラルネットワーク研究に対して、学術的な正当性を回復させた。当時のニューラルネットワーク研究は、周辺的で非主流の領域と見なされていた。しかし、PDP研究グループの発表と並行して、複数の研究機関でニューラルネットワーク研究が復活し始める。米国ではMIT、トロント大学などが、ニューラルネットワーク研究を本格化させた。英国ではジョフリー・ヒントンが活動の中心の一人となった。

PDP研究グループの歴史的意義は、記号主義と統計的・接続主義的アプローチとの「科学的和解」を学術的に正当化したことにある。1970年代から1980年代中盤までは、これら二つのアプローチはほぼ敵対的な関係にあり、同一の研究機関内でも支持者が分裂することがしばしばであった。しかしPDP研究グループは、むしろ異なるレベルの分析（シンボル的記述と実装的な神経回路網との関係）が共存可能であり、相補的であることを示唆した。この視点は、後の統計的機械学習への転換（第III部）を知的に準備する上で重要であった。

8.5　誤差逆伝播法の再発見とニューラルネットの再燃

PDP研究グループの復活をもたらした最大の技術的基盤が、「誤差逆伝播法（backpropagation）」の再発見である。この歴史は、科学史における独立同時発見と優先権の問題を如実に示す例である。

誤差逆伝播法の最初の導出は、1974年にポール・J・ウェルボスが博士論文「Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences」の中で提示されていた。この論文は当時、学術出版の中心的流通経路に乗らず、認識されないまま忘却の彼方に沈んでいた。1985年には、パーカー（David Parker）が独立に同様のアルゴリズムを導出し、「Learning Logic」と題する技術報告書で発表している。しかし、これもまた広く知られるには至らなかった。

転機は1986年に訪れた。ラメルハート、ヒントン、ウィリアムス（Ronald J. Williams）は、1986年に誤差逆伝播法を独立に再発見し、それをマルチレイヤーニューラルネットワークの学習に適用する系統的方法論を確立した。彼らは同年、Nature誌に「Learning Representations by Back-Propagating Errors」と題する論文を発表した。この論文は、従来のパーセプトロンが克服できなかった限界——多層ネットワークにおける隠れ層（hidden units）の役割と最適化の方法——を解決するものであった。

重要なのは、このアルゴリズムが何をもたらしたかである。単一層のパーセプトロンは線形分離可能な問題のみを解くことができ、XOR問題のような非線形分離可能な問題には対応できなかった（これがミンスキー＝パパートの批判の核心であった）。しかし誤差逆伝播法により、複数の隠れ層を持つニューラルネットワークが、非線形な関数を近似できることが実証された。さらに、隠れ層のユニットが学習過程で自動的に重要な特徴表現（feature representations）を発見することが観測されたのである。

1987年には、ラメルハート、マクレランド、およびPDP研究グループ全体の二巻本『Parallel Distributed Processing』が刊行された。この著作には、誤差逆伝播法を用いた言語認識、視覚処理、象徴的推論など多様な認知タスクの実装例が含まれていた。この出版は、誤差逆伝播法の認知度を劇的に高め、1980年代後半から1990年代初頭にかけてのニューラルネットワーク研究ブームの引き火となった。

誤差逆伝播法の「再発見」という表現が適切である理由は、Werbosの1974年の論文がすでに存在していたにもかかわらず、ラメルハート、ヒントン、ウィリアムスの1986年の論文がAI研究コミュニティに与えた影響が、その百倍以上であったからである。これは学術出版システムと知識の流通における偶発的・

構造的な非効率性を示している。ラメルハート自身が後年に認めたところによれば、彼らが先行研究を引用しなかった理由は、それらの論文が「あまりに隔絶していた」からであり、再発見は実質的には必然であったというのである。

8.6　第二次AIの冬と第一次の冬との構造的比較

第一次の冬と第二次の冬は、しばしば「同じ現象の繰り返し」として語られるが、その構造は大きく異なる。この違いを理解することは、AI産業・研究の周期的ダイナミクスを把握する上で重要である。

第一次の冬（1973年前後からおよそ1980年中盤まで）の主因は、研究資金の急激な削減にあった。ライトヒル報告（1973年）は英国のAI研究を大幅に収縮させ、同時期のDARPA資金削減により米国の基礎研究も圧迫された。この外的な資金環境の変化が、研究活動そのものを制約したのである。その意味で、第一次の冬は「資金危機型」の冬であった。

これに対し、第二次の冬（1987年から1990年代中盤まで）の構造は異なる。この時期、AI関連の企業は依然として存在し、研究機関への資金流入も続いていた。むしろ問題は、市場と研究の期待値の乖離にあった。エキスパートシステムは約束された機能を十分に提供できず、LISPマシンはコスト・パフォーマンス面で汎用プロセッサに敗北した。投資家と政府が期待していた「AIによる産業革新」は現実化せず、失望が広がったのである。その意味で、第二次の冬は「期待崩壊型」の冬であった。

また、第一次の冬中でも記号推論やエキスパートシステムの理論的・実装的な改善は進められていたが、第二次の冬は同時に、新しいパラダイムへの転換の時期でもあった。1986年のPDP研究グループの登場と誤差逆伝播法の再発見は、単に既存の記号主義的アプローチの改良ではなく、根本的な変更を示唆していた。この意味で、第二次の冬は「パラダイム転換期」であった。つまり、個別の技術的改善と新しい基本理念の同時的出現が特徴となっている。

第三に、資金メカニズムの違いがある。第一次の冬では、DARPA資金の喪失が痛手であったが、民間企業のAI関心は低く、したがって資金喪失はほぼ公的セクターの研究を襲った。これに対し、第二次の冬では、LISPマシン企業の破綻に見られるように、民間セクターがAIバブルの直撃を受けた。公的研究機関への影響は相対的に限定的であり、むしろ大学のコンピュータサイエンス部門では新しい研究パラダイムへの投資が可能であった。この民間セクターの縮小と公的セクターの新パラダイムへのシフトが、第二次の冬が「クリエイティブな冬」になり得た理由の一つである。

8.7　知的継続性と新しい予感 —— 第III部への橋渡し

第二次AIの冬を通じて、AI研究は知的に大きく変貌した。しかし、この変化を「断絶」と見なすのは誤りである。むしろ、1980年代末から1990年代初頭にかけてのニューラルネットワーク研究の復活は、第3章のパーセプトロンの挫折から数十年の歳月を経た「長い弧の回復」と見なすべきである。

ミンスキー＝パパート『Perceptrons』（1969年）が単層パーセプトロンの根本的限界を指摘してから17年後の1986年に、その限界を超える方法（多層ネットワークと誤差逆伝播法）が実装可能になったのである。この歴史的弧は、人工知能の歩みが決して直線的ではなく、理論的批判、技術的制約、計算資源の進化が複雑に絡み合う過程であることを示している。

同時に、1980年代後半以降のニューラルネットワーク研究の復活は、新しい知識の時代への過渡期であった。誤差逆伝播法とPDP研究グループは、記号処理の限界を認識しながらも、まだ完全には統計的機械学習へのパラダイム転換を達成していない。むしろ、「表現学習（representation　learning）」という中間的な概念を通じて、記号的理解と統計的パターン抽出の間を橋渡しするものであった。隠れ層が学習過程で自動的に特徴を発見する仕組みは、数十年後のディープラーニング（第IV部）に直接つながる知的・技術的資産となるのである。

第二次の冬が明けるにあたって、AI研究コミュニティはもう一つの重要な認識に到達していた。それは、「知能とは何か」という哲学的問いに答える前に、「知能はいかにして学習するのか」という認識論的問いが根本的であるという洞察である。この転換は、単なる技術的シフトではなく、AI研究全体の理論的基礎の再構築を意味していた。記号主義の時代には、知識は与えられたものであり、問題はいかにしてそれを操作するかにあった。しかし、統計的・接続主義的視点からは、知識は経験から自動的に抽出されるべきものであり、「学習」こそが知能の本質的メカニズムであるとの認識が浮上していたのである。

本章で論じた第二次AIの冬と記号主義から接続主義への転換は、次の第III部「統計的転回と機械学習の勃興」（1990年代〜2000年代）へと直結する。LISPマシンの栄光と衰退、第五世代プロジェクトの野心と失敗、エキスパートシステムの脆弱性の認識を経て、AI研究は「知識」から「学習」へと関心の重心を移動させるのである。この転換が、データ駆動型のアプローチ、統計的アルゴリズム、そして最終的には大規模ニューラルネットワークの台頭につながっていく道程を、次の三章で検討することになる。

> 本章では、1980年代後半から1990年代初頭の第二次AIの冬を、複数のレベルで分析した。ハードウェア市場の劇的な収縮（LISP　　マシン、第五世代プロジェクト）、基本的なパラダイムの限界（記号AIの脆弱性と保守性の問題）、そして新しい知的運動の萌芽（PDP研究グループと誤差逆伝播法の再発見）が、同時的に発生した。この冬は、第一次の冬のように研究資金が消滅したわけではなく、むしろ市場の期待値の喪失と理論的パラダイムシフトの加速が特徴であった。ニューラルネットワークの再燃は、パーセプトロンの長い沈黙からの回復であり、第3章から数十年の間に蓄積された理論的・技術的知見が、初めて広く実装可能になった瞬間でもあった。

参考資料（本章）

- Computer History Museum および MIT 関連史料。LISP マシン産業と Symbolics / LMI の系譜の確認に使用。
- 情報処理学会（IPSJ）等の FGCS / ICOT 回顧資料。FGCS の予算規模、期間、PIM の位置づけの確認に使用。
- David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams, “Learning Representations by Back-Propagating Errors” (Nature, 1986).
- David E. Rumelhart and James L. McClelland (eds.), Parallel Distributed Processing (1986/1987).
- Paul J. Werbos, Beyond Regression (1974) および David Parker の技術報告。誤差逆伝播法の先行史の確認に使用。

第III部

統計的転回と機械学習の勃興（1990年代〜2000年代）

第9章〜第13章

第9章　統計的機械学習の時代

概観

1990年代から2000年代初頭にかけて、AI研究は根本的なパラダイムシフトを経験した。それまで知識表現と記号推論によって知能を実装しようとする「記号主義」が圧倒的に支配していたAI研究の世界に、統計的機械学習という新しい思想がしだいに浸透していった。この転換は、単なる手法の交代ではなく、知能とは何か、それをどのように機械で実現するかという根本的な問い立ての変化であった。

本章では、このパラダイムシフトの構造的背景を明らかにした上で、当時の主要な手法群――サポートベクターマシン、決定木とアンサンブル学習、ベイズ的アプローチ、強化学習――を検討する。各々の手法は表面的には独立した工学的な成果に見えるかもしれないが、実は深い統一的な志向を共有していたことが重要である。それは、「データから直接知識を抽出する」という原理、「確率と統計を通じて不確実性を扱う」という方法、そして「汎化性能を保証する理論的枠組み」の探求という三位一体である。

9.1　パラダイムシフト ―― 記号主義から統計主義へ

9.1.1　第二次AIの冬からの回復

第8章で述べたように、1987年から1990年代前半にかけての第二次AIの冬は、専用ハードウェア市場の崩壊とAI産業への期待の後退をもたらした。だがこの冬は、同時に大きな知的転換をも促した。LISPマシンの衰退とともに、記号処理を中心とするAIの研究戦略そのものへの反省が深まったのである。

エキスパートシステム（第7章）が直面していた根本的な問題は、いまや明確に認識されるようになっていた。第一に、知識獲得のボトルネック問題である。人間の専門家が保有する知識を記号的に表現し、システムに組み込むことは、想像以上に困難で時間を要した。エドワード・ファイゲンバウムらが繰り返し強調した通り、知識の抽出は「単に遅いだけでなく、本質的に困難」であった。第二に、スケーラビリティの限界である。知識ベースが拡大するにつれて、矛盾の解決、保守性の確保がますます難しくなった。第三に、新しい領域への転移の困難さである。あるドメインで構築されたエキスパートシステムは、異なるドメインではほぼ無用の長物であった。

これらの問題に直面した研究コミュニティは、別の道を模索し始めていた。それは、人間の明示的な知識提供に依存するのではなく、大量のデータから自動的にパターンを学習する機械学習へのシフトであった。

9.1.2　三位一体の収束

このシフトが可能になった背景には、三つの外部条件の同時的な成熟があった。

（1）データの可用性

1980年代後半から1990年代にかけて、デジタルデータの量と多様性が飛躍的に増加した。郵便局の手書き数字認識問題（USPS dataset）、医療記録、金融取引記録、テキストコーパスなど、次々と大規模で構造化されたデータセットが問題ドメインから生成されるようになった。カリフォルニア大学アーバイン校によるUCI機械学習リポジトリは、このような公開ベンチマークデータセットを一堂に集め、研究者たちが共通の課題に対して異なるアルゴリズムを比較検証する文化を確立した。データは、かつての記号AIにおいて想像もされていなかった、根本的に重要な資産へと変貌していったのである。

（2）計算資源の進化

ムーアの法則に従い、プロセッサの処理速度は継続的に向上していた。しかし、統計的機械学習の興隆において注目すべきは、単なる計算速度の向上だけではなく、「統計計算」を効率的に実行するアルゴリズムの発展である。特に、マルコフ連鎖モンテカルロ法（MCMC）と詳細な実装技術は、ベイズ推論における計算の実現可能性を劇的に拡大させた。ゲマン兄弟によって1984年に提案されたギブスサンプリングは、1990年代に統計学の主流へと台頭し、複雑な確率モデルの学習を可能にした。

（3）学習理論の成熟

1960年代から70年代にかけてソビエト連邦でウラジミール・ヴァプニク（Vladimir Vapnik）とアレクセイ・チェルヴォーネンキス（Alexey　Chervonenkis）により開発されてきた統計学習理論が、1990年代には英語圏の研究者たちに認識されるようになった。VC次元（Vapnik–Chervonenkis　dimension）によって定義される「仮説空間の複雑さ」と「汎化誤差の確率的上界」との間に、厳密な数学的関係があることが明らかになったのである。これは、過去のニューラルネットワークやパターン認識手法が直感的に行っていた「訓練誤差と汎化誤差のバランス」を、理論的に正当化する基盤を与えた。

この三つの要素の収束が、1990年代という時期を、AI研究における**第二の興隆期**へと押し上げた。記号主義から統計主義への転換は、偶然ではなく、外部条件と知的準備の同時的な成熟の産物であったのである。

9.1.3　パラダイムとしての意味

パラダイムシフトをいかに特徴づけるべきか。単に「記号的表現」を「統計的パターン」に置き換えるという技術的変化ではなく、もっと深い認識論的な転換である。

記号主義（symbolic AI）は、知識を「記号的な表現」として外部から構成することを前提としていた。「ドメイン知識」「ルール」「事実」は、設計者が言語的に記述し、体系的に整理するべきものであった。言い換えれば、知能とは「正しい知識をいかに正確に表現し、推論するか」という問題だと見なされていたのである。

これに対し、統計的機械学習は、知識を「データに潜在するパターン」として見なす。知識獲得の問題は「何が正しい知識であるか」を明示的に述べることではなく、「大量のデータからその確率的構造を学習すること」へと転換した。この転換は、知識を「外部から与えられるべき記述」から「経験から自動的に抽出されるべき資源」へと根本的に再定義するものであった。

言語的に述べれば、記号AIは「処方的（prescriptive）」知識を求めていたが、統計機械学習は「記述的（descriptive）」知識を求めるのである。前者は「何をするべきか」を問い、後者は「データは実際に何を示しているか」を問う。この違いは、工学的にも、また知識観としても、深い意味を持つ。

9.2　サポートベクターマシン（SVM）とカーネル法

9.2.1　SVMの誕生：VC理論から実装へ

ウラジミール・ヴァプニクとコンスタンティン・コルテスが1995年に発表した「Support-Vector Networks」は、統計的機械学習の時代における最初の大きな「勝利」となった。このアルゴリズムは、VC理論という抽象的な学習理論をはじめて実践的で強力な学習機械へと具現化したものであった。

SVMの基本的な考え方は、シンプルながら洗練されている。与えられた訓練データを、高い次元の空間へ**写像**し、その空間において、異なるクラスのデータを「最大マージン（maximum margin）」で分離する超平面を探す。ここで「マージン」とは、超平面から最も近いデータ点までの距離である。VC理論による厳密な分析の結果、このマージン最大化原理は、単なる訓練誤差の最小化だけでなく、未知のテストデータに対する高い汎化性能を保証することが示されたのである。

このことの意義は極めて大きい。それまでのニューラルネットワークやその他の機械学習手法は、「いかに訓練誤差を小さくするか」ということに専念する傾向があった。だが、複雑なモデルを訓練データに完全に適合させれば、テストデータでの誤差が増加する「過学習（overfitting）」という問題が生じることは、経験的に知られていた。SVMが革新的であったのは、「訓練誤差」と「汎化誤差」のバランスを理論的に分析し、そのバランスを最適化する学習原理を数学的に導き出したことにあった。

9.2.2　カーネル法の威力

しかし、SVMの真の威力は、カーネル法の組み込みによってはじめて十分に発揮された。元々、データが線形分離可能な場合、上述の最大マージン超平面を見つけることは、二次計画問題として効率的に解くことができる。しかし、現実の多くの問題では、元の特徴空間での線形分離は不可能である。

カーネル法は、この困難を優雅に解決する。その基本的な原理は、「データを高次元の空間へ非線形に写像すること」である。マーサー定理として知られる数学的事実は、適切なカーネル関数（kernel function）を定義することで、高次元空間での内積計算を、元の空間での単純な関数計算に帰着させられることを保証している。つまり、実際には高次元空間での計算を行わずに、その効果を得ることができるのである。このトリック（kernel trick）により、SVMは計算コストをほぼ変えることなく、複雑で非線形な決定境界を学習できるようになった。

カーネル法の自由度は高い。ガウスカーネル、多項式カーネル、シグモイドカーネルなど、様々なカーネル関数が提案された。各々のカーネルは異なる仮説空間に対応し、したがって異なる「学習能力」を持つ。この柔軟性により、SVMは自動顔認識、テキスト分類、バイオインフォマティクスなど、極めて多様な問題領域で高い成功を収めた。

9.2.3　実践的影響と限界

1995年から2000年代初頭にかけて、SVMは統計的機械学習の「王者」として扱われた。その理由は、理論的な正当性と実践的な有効性の両者を兼ね備えていたからである。特に、テキスト分類や手書き数字認識、生物医学的なパターン認識などの問題で、それ以前の手法よりも優れた性能を示したのである。

ただし、SVMが直ちに衰退したわけではないが、やがて新たな課題が浮上する。第一に、大規模データセットに対するSVMの訓練の計算コストは、二次以上に増加する傾向があった。クラウドコンピューティング時代の到来とともに、数百万件、数十億件のデータを扱う問題が一般的になるにつれて、この計算複雑性の問題は深刻化していく。第二に、SVMの意思決定プロセスは「ブラックボックス」的であり、なぜそのような決定に至ったかを説明することが困難であった。第三に、深層学習の台頭（第14章以降）により、2010年代には、より単純で並列化が容易なニューラルネットワークが、実践的には徐々にSVMを置き換えていくことになる。

9.3　決定木・ランダムフォレスト・アンサンブル学習

9.3.1　決定木の簡潔性と解釈可能性

決定木（decision tree）は、SVMほどは理論的な光彩を放たなかったが、実践的には極めて重要で、今日でも広く用いられている手法である。

決定木の基本的な構想は古く、1960年代まで遡ることができるが、1980年代にジョン・クインラン（J. Ross Quinlan）によってID3アルゴリズムが、そして1993年にはその改良版C4.5が開発された。これらのアルゴリズムは、与えられた訓練データを再帰的に分割することで、樹形状の判定構造を構築する。各々の分割は、その時点でのエントロピー（entropy）という情報理論的概念に基づき、「最も情報量を得られる特徴」を選択することで実行される。

決定木がSVMと大きく異なる点は、その解釈可能性にある。決定木が学習したルールは、「特徴aが値xより小さければ、特徴bが値yより大きいならクラスAに分類する」というように、人間が直接読み理解できる形で表現される。この特性により、医療診断、信用判定、故障診断など、「なぜそのような判定をしたのか」を説明する必要がある領域で、決定木は高い価値を持ってきた。

9.3.2　アンサンブル学習とランダムフォレスト

しかし、単一の決定木の予測精度は、SVMやニューラルネットワークに比べて劣ることが多い。この限界を克服するために、複数の決定木を組み合わせる「アンサンブル学習」の考え方が発展した。

1996年、レオ・ブレイマン（Leo Breiman）はバギング（bootstrap aggregating）と呼ばれる手法を導入した。このアプローチでは、元の訓練データからランダムに復元抽出により部分的なサンプルを複数個生成し、各々のサンプルに対して独立に決定木を構築する。その後、分類問題ではこれらの決定木の多数決、回帰問題では平均をとることで、最終的な予測を行う。バギングの威力は、複数の「弱い学習器」を組み合わせることで、個々の学習器よりも大幅に高い性能を達成できることにある。

2001年、ブレイマンは「ランダムフォレスト（Random Forests）」という、さらに洗練されたアンサンブル手法を発表した。ランダムフォレストは、バギングの考え方を進め、各々の決定木の構築時に、分割に用いる特徴を「ランダムに選択された部分集合」から選ぶという工夫を加えた。これにより、個々の決定木の多様性がより高まり、アンサンブルとしての予測精度が向上する。ランダムフォレストの成功は、統計的機械学習における「アンサンブル」という発想の強力さを示すものであった。

一方、別のアンサンブル手法として、1996年にはヨアヴ・フロイントとロバート・シャピレ（Yoav Freund and Robert E. Schapire）が「AdaBoost（Adaptive Boosting）」と呼ばれるアルゴリズムを発表した。AdaBoostの革新的な点は、構成する個々の学習器に対して「重み」を導入することで、過去に誤分類されたデータに対してより高い学習圧力を加えることができるという点である。ブースティングのアプローチは、理論的には、弱い学習器の集合から強い学習器を構成できることを示し、アンサンブル学習の正当性を強く後押しした。

9.3.3　なぜアンサンブルなのか

アンサンブル学習が強力である理由は、複数の異なる仮説の「多数決」という原理の有効性にある。個々の学習器が確率的に独立な誤りを犯すと仮定すれば、集団の判定の誤り確率は、個々の学習器の誤り確率よりも指数関数的に低下する。この原理は、統計的機械学習における最も単純で、かつ最も効果的な「知恵の集約（wisdom of crowds）」の実現形態である。

2010年代から2020年代にかけて、ランダムフォレストやGradient Boosting Machines（GBM）などのアンサンブル手法は、表形式（tabular）データを用いた実務的な予測問題において、深層学習を含む他のあらゆる手法より優れた性能を示すことが繰り返し報告されている。このことは、統計的機械学習の時代に確立された原理が、今日に至るまでその威力を失っていないことを示す重要な指標である。

9.4　ベイズ的アプローチとグラフィカルモデル

9.4.1　ベイズネットワークの理論的基盤

1980年代半ば、ジュディア・パール（Judea Pearl）は、ベイズネットワーク（Bayesian networks）という確率的グラフィカルモデルを体系化した。この手法は、複雑な確率分布を、変数間の条件付き独立性を明示的に表現する有向非環グラフの形で表現するものである。

パールの革新は、単に確率モデルを構築するツールを提供しただけではない。彼は、グラフの構造と確率的独立性の間に深い対応関係があることを示した。この対応関係により、複雑な結合確率分布を、より単純な条件付き確率のテーブルの積の形で因数分解できるようになったのである。結果として、変数の数が多い場合でも、確率推論の計算を実現可能な規模で実行できるようになった。

1988年に出版されたパールの著書『Probabilistic Reasoning in Intelligent Systems』は、AI研究者のバイブルとなった。それまで、記号的推論と確率的推論は、原理的に対立する二つの手法として捉えられていたが、パールはベイズネットワークを通じて、両者の統合を示唆したのである。具体的には、エキスパートシステムで用いられていたような「知識ベース」を、確率的な不確実性の枠組みに統合することが可能であることを示したのである。

9.4.2　推論アルゴリズムと計算の実現

ベイズネットワークが実用的な威力を発揮するためには、与えられた観測に対して、未知の変数の確率分布を効率的に計算する「推論」のアルゴリズムが必要である。パールは、条件付き独立性の構造を利用して、複雑な全体の推論問題を、グラフ構造に沿った局所的な計算に分解できることを示した。これを「信念伝播（belief propagation）」アルゴリズムと呼ぶ。

信念伝播は木構造のグラフに対して正確な推論を実現し、また環を含むグラフに対しても近似的な推論を与える。この分解可能性は、計算複雑性を指数関数的に削減するものであり、それまでのベイズ推論の計算不可能性という大きな障害を部分的に解決したのである。

同時期に、統計学の領域からは、マルコフ連鎖モンテカルロ法（MCMC）とりわけギブスサンプリングが、確率的推論の別の強力な手段として台頭してきた。ゲマン兄弟が1984年に提案した手法は、1990年代にはゲルファンドとスミスの1990年の論文に代表されるように、複雑な確率モデルの学習と推論を実現する標準的な計算手法となっていた。MCMCの利点は、高次元の確率空間での積分を、逐次的なサンプリングにより近似できることにあり、その応用の汎用性は高い。

9.4.3　知識獲得とモデル学習

ベイズネットワークは、最初は人間の知識として構築されていた。医療診断、故障診断など、ドメインの専門家がネットワークの構造（どの変数がどの変数に依存するか）を設計し、条件付き確率表を手作業で埋めるという手法である。

しかし、ほどなく研究者たちは、この構造と確率パラメータをデータから学習する方法の開発に着手した。1990年代から2000年代にかけて、観測データからベイズネットワークの構造を推定する手法（構造学習）と、与えられた構造の下でのパラメータを推定する手法（パラメータ学習）に関する研究が急速に進展した。特に、隠れ変数が含まれる場合の学習には、EM（Expectation-Maximization）アルゴリズムが強力な手段となった。

このように、ベイズネットワークは、記号的知識表現と統計的パラメータ学習を統一する枠組みを提供し、記号AIから統計機械学習への転換を「哲学的」レベルでも支持する理論的基盤となったのである。

9.5　強化学習の理論的基盤 —— サットン＝バルトの貢献

9.5.1　強化学習の古い起源

強化学習（reinforcement　learning）という概念そのものは、1990年代に初めて登場したわけではない。その根源は、動物心理学における「試行錯誤」の観念、そして制御理論におけるフィードバック原理に遡る。サイバネティクス（第1章）の思想、ノーバート・ウィーナーの「自己修正的システム」の概念も、広い意味で強化学習を予示していた。また、1950年代の初期AI研究において、シャノンが提案したゲーム木探索（第1章）の戦略や、サミュエルのチェッカープログラム（第2章）による自己改善も、強化学習的な発想を含んでいた。

しかし、強化学習が統計的機械学習の一分野として体系化されたのは、1980年代後半から1990年代にかけてのことである。その過程で決定的な役割を果たしたのが、リチャード・サットン（Richard　S. Sutton）とアンドリュー・バルト（Andrew G. Barto）である。

9.5.2　時間差分学習と価値関数

1988年から1998年の十年間に、サットンとバルトは一連の論文と著書を通じて、強化学習の理論的基盤を確立した。その核心となるのが、「時間差分学習（Temporal-Difference Learning, TD学習）」と呼ばれる方法である。

強化学習の基本的な設定を述べよう。エージェントが環境と相互作用する際、各々の時刻tにおいて、エージェントは現在の状態stを観測し、ある行動atを選択する。すると環境は報酬rtを与え、新しい状態st+1へ遷移する。エージェントの目標は、将来の累積報酬（割引した合計）を最大化する最適な行動選択戦略（方針、policy）を学習することである。

この問題を解くための古典的な手法は、動的計画法（dynamic　programming）であった。ベルマン方程式として知られる再帰的な関係式により、価値関数V(s)（状態sからの将来の期待報酬）を定義し、これを繰り返し計算することで最適方針を求めるというアプローチである。しかし、動的計画法は、環境の詳細なモデル（遷移確率、報酬関数）が既知であることを前提としており、実世界の多くの問題ではこの前提が成り立たない。

時間差分学習が革新的であったのは、環境モデルを必要とせず、また大域的な価値関数の更新ではなく、観測された遷移に基づいて逐次的に価値関数を改善できるという点にある。具体的には、各々の遷移（st, at, rt, st+1）を観測するたびに、実際に得られた報酬rtと次状態での価値予測V(st+1)の「時間差分」を用いて、現在の状態の価値V(st)を更新するのである。この更新則は、統計的には、観測による勾配法の一種として解釈される。

さらに、クリストファー・ワトキンスは1989年の博士論文で「Q学習（Q-learning）」を提案し、1992年のWatkins–Dayan論文でその性質が広く共有された。Q学習では、状態sにおける行動aの価値Q(s, a)を学習の対象とする。この手法の利点は、「オフポリシー」学習が可能であることである。つまり、現在の方針に従わないような行動を観測した場合でも、それから最適方針に関する情報を抽出できるのである。これは、コンピュータゲームの自動プレイやロボット制御など、多くの実際の応用で極めて有用である。

9.5.3　理論的保証と実装

サットンとバルトの重要な貢献は、単に新しいアルゴリズムを提案したことだけではない。彼らは、時間差分学習と価値関数学習の枠組みを明確化し、他の研究者による収束性解析や最適性証明と結びつけることで、強化学習を単なる工学的なヒューリスティクスから、統計的機械学習の正当な一分野へと引き上げたのである。

さらに、彼らは1998年に『Reinforcement Learning: An Introduction』という教科書を共著で出版した（初版；改訂版は2018年）。この教科書は、強化学習の理論と実装を体系的に解説し、複数の世代の研究者を育成してきた。その影響は、単に理論にとどまらず、実装上の知見（例えば、Q学習の近似による不安定性への対策）も含まれており、学術と実践をつなぐ標準的なテキストとして機能した。

9.5.4　強化学習から深層強化学習へ

統計的機械学習の時代における強化学習は、まだ比較的ニッチな研究領域であった。その理由は、いくつかの実装上の困難にあった。第一に、「探索と利用のトレードオフ（exploration-exploitation trade-off）」という根本的な問題がある。エージェントは、既知の好ましい行動を繰り返すべきか（利用）、

それとも未知の行動を試してみるべきか（探索）というジレンマに直面する。この問題の理論的最適化は、多腕バンディット問題として知られるが、実装においては常にトレードオフを伴う。第二に、「状態空間の大規模性」である。囲碁のような複雑なゲームでは、可能な状態の数が天文学的に多く、価値関数を明示的に表現することは不可能である。

これらの問題の解決は、深層ニューラルネットワークの台頭を待つことになる。2013年から2015年にかけて、デミス・ハサビスのDeepMind研究所による「深層強化学習（Deep Reinforcement Learning）」の成果――DQN、AlphaGo――が、強化学習を一躍脚光浴びせることになるのであるが、それは本書の第15章以降の物語である。

小括：パラダイムの統一性と展望

本章で概観した四つの主要なアプローチ――SVM、決定木とアンサンブル学習、ベイズネットワーク、強化学習――は、一見すると異なる問題を扱う異なる手法に見えるかもしれない。しかし、深い層においては、これらはすべて統計的機械学習というより大きなパラダイムの異なる側面を体現しているのである。

共通する原理は以下のようにまとめられる：

1. **データからの学習**：記号的に与えられた知識ではなく、観測されたデータから直接的にパターンを抽出する。
2. **不確実性の確率的表現**：真の世界についての不完全な知識を、確率分布や確信度として数学的に表現する。
3. **汎化性能の保証**：訓練データ上の精度だけでなく、未知のテストデータに対する予測精度を理論的に分析し最適化する。
4. **計算可能性の重視**：理想的な数学的解が存在しても、実装可能な計算量の内で近似解を求める工学的現実主義。

これら四つの原則の統一的な理解こそが、統計的機械学習の時代を特徴づける知的財産なのである。

1990年代から2000年代初頭のこの時期は、AI研究にとって「第二の興隆期」であったと言える。第一の興隆期は1950年代から1960年代のダートマス会議と初期のAI研究であり、新しい学問分野を創設することそのものが目標であった。第二の興隆期は、その基盤の上に、より洗練された、理論的に正当化されたアルゴリズムと手法を開発し、さらに重要なことに、それをスケーラブルに実装し、実世界の問題に適用することであった。

同時に、本章の記述の中には、やがて訪れる深層学習革命への伏線も多く含まれている。特に、計算統計（MCMC、ギブスサンプリング）の進化、大規模データセットの可用性、そして何より、複雑な非線形表現を学習する能力の追求は、ニューラルネットワークの再興へ向かう自然な知的流れであった。SVMが「最大マージン」という原理によって汎化性能を保証しようとしたのに対し、深層学習は「表現学習（representation　learning）」によって複雑なパターンを自動的に獲得しようとするアプローチである。この転換も、統計的機械学習の時代における問題の深い理解があってこそ可能であった。

次章では、自然言語処理の領域において、統計的パラダイムがいかに急速に、そして徹底的に、記号的手法を置き換えていったかを見ることになる。言語の機械的処理という伝統的な課題領域において、統計的機械学習の威力は最も劇的に証明されたのである。

参考資料（本章）

- Cortes, C., & Vapnik, V. (1995). “Support-Vector Networks.” Machine Learning, 20(3), 273-297.
- Breiman, L. (2001). “Random Forests.” Machine Learning, 45(1), 5-32.
- Freund, Y., & Schapire, R. E. (1997). “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.” Journal of Computer and System Sciences, 55(1), 119-139.
- Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann.
- Sutton, R. S., & Barto, A. G. (1998). Reinforcement Learning: An Introduction. MIT Press.
- Watkins, C. J. C. H., & Dayan, P. (1992). “Q-learning.” Machine Learning, 8, 279-292.
- Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.
- Gelfand, A. E., & Smith, A. F. M. (1990). “Sampling-Based Approaches to Calculating Marginal Densities.” Journal of the American Statistical Association, 85(410), 398-409.
- Geman, S., & Geman, D. (1984). “Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 6(6), 721-741.

第10章　自然言語処理の統計革命

10.1　コーパス言語学と統計的NLPの台頭

第2章で論じた初期の機械翻訳システムの挫折から、自然言語処理（NLP）は長く低迷の時代を迎えていた。ウィーノグラード、シャンク、ウィルクスらによる意味理解の試みは、限定された領域内では成果を挙げたものの、一般的な言語現象の複雑性に直面すると頓挫した。1980年代までのNLP研究は、言語学的な規則を手作業で記述し、知識ベースに格納する記号主義的アプローチに主に依存していた。その結果、システムは脆弱であり、新しい領域や表現に対応することが困難であった。

1980年代から1990年代にかけて、この状況は劇的に変化する。コンピュータの価格低下と性能向上に伴い、大規模なテキストコーパスがデジタル化され、利用可能になり始めた。同時に、計算言語学のコミュニティは、言語の統計的特性に着目する研究を開始した。この転換は単なる工学的な判断ではなく、言語学そのものの再考をも意味していた。

コーパス言語学（corpus linguistics）の系統的展開は、1980年代以降に加速する。特に英語コーパスの構築が進展し、1990年代初頭には「言語の一般的パターンは、統計的に十分規模の言語データから直接学習可能である」という仮説が、単なる可能性ではなく、実証可能な方法論として認識されるようになった。この認識の転換を可能にしたのは、計算能力の向上だけではなく、機械学習理論の発展でもあった。

統計的NLPの台頭は、1990年代を通じて急速に進行する。1990年から2000年代初頭にかけて、機械翻訳、音声認識、情報検索、形態素解析、構文解析といった主要領域で、記号主義的ルールベースのアプローチから統計的学習アプローチへの重心移動が進んだ。この転換の根本的な動因は、統計的手法がルールベース手法よりも本質的に優れていたのではなく、むしろ次の三つの実践的な要因にあった。第一に、統計的手法は新しい言語や領域への適応が相対的に容易であり、手作業によるルール記述のコストを大幅に削減できた。第二に、大規模コーパスの入手可能性の向上により、統計的学習に必要なトレーニングデータが確保されるようになった。第三に、ベイズ推論、最大尤度推定、EMアルゴリズムなどの統計的手法の理論的基盤が固まり、実装上の課題が解決されていった。

10.2　統計的機械翻訳 —— IBMモデルからフレーズベースへ

統計的機械翻訳（Statistical Machine Translation, SMT）は、統計的NLPの成功例として最初に注目を集めた領域である。その源泉は、IBM研究所（トーマス・J・ワトソン研究センター）の研究チームであった。1991年、ピーター・ブラウン、スティーブン・デラ・ピエトラ、ヴィンセント・デラ・ピエトラ、ロバート・マーサーらは、翻訳問題を統計的観点から根本的に再定式化した。

従来の規則ベース機械翻訳は、言語学者が言語対ごとに翻訳規則を手作業で記述し、辞書や文法を構築するというボトルネックに悩まされていた。IBMの研究チームは、翻訳を「対訳コーパス（並行コーパス）から学習可能な確率モデルの問題」として捉え直した。すなわち、原言語の単語シーケンスが与えられたとき、目標言語の単語シーケンスが観測される条件付き確率を推定するという枠組みである。

IBMが発表した五つの統計的翻訳モデル（公開された形で1993年に発表される）は、複雑さを段階的に増す設計となっていた。モデル1は純粋な語彙翻訳確率を学習し、単語ペアの対応関係を確立する。モデル2以降は、単語の並び替え（reordering）や単語削除（deletion）といった翻訳に必要な複雑な現象を段階的にモデル化する。この系統的な構築法は、統計的手法の有効性を実証するとともに、後続の研究者たちに方法論的な指針を与えた。

IBMモデルの重要な特徴は、「隠れた対応関係（alignment）」という概念の導入にある。対訳コーパスから観測されるのは、対訳文のペアだけであり、どの原言語の単語がどの目標言語の単語に対応するかは直接観測されない。IBMの研究チームは、この未観測の対応関係を隠れ変数として扱い、期待値最大化（Expectation-Maximization, EM）アルゴリズムによってこれを推定する方法を開発した。このアプローチは、後の自然言語処理におけるあらゆる隠れ変数モデルの理論的先例となる。

統計的機械翻訳の次の段階は、フレーズベースのモデルの登場である。1990年代後半から2000年代初頭にかけて、研究者たちは、単語単位の翻訳ではなく、「フレーズ」（統計的に抽出された単語の連続）を翻訳の基本単位とすることで、モデルの柔軟性と精度を向上させた。フレーズベース翻訳では、対訳文から統計的に有意なフレーズペアを自動抽出し、これを翻訳テーブルとして保持する。言語モデルが翻訳の流暢性を担当し、並び替えモデルがフレーズの再順序化を制御する。この三層構造（翻訳モデル、言語モデル、並び替えモデル）は、2000年代の統計的機械翻訳の標準的なパイプラインとなった。

統計的機械翻訳の学術的成功にもかかわらず、商用システムへの導入は段階的であった。Google Translateは2006年に公開され、2000年代後半には統計的機械翻訳を中核に据えることで、機械翻訳を広く実用レベルへ押し上げた。Googleのシステムは、インターネットから収集した大規模な対訳コーパスを活用し、統計的翻訳モデルを訓練した。この成功は、大規模データの入手可能性と計算資源の充実が、統計的手法の効果を引き出す上でいかに重要であるかを示した。ただし、統計的機械翻訳による翻訳品質の改善は、2010年代にニューラル機械翻訳の登場によって再び問い直されることになる（第15章）。

10.3　情報検索と検索エンジンの進化

統計的NLPと密接に関連するもう一つの領域は情報検索（information retrieval, IR）である。デジタル時代以前の情報検索は、ライブラリアンや専門家による人間が媒介する活動であり、ユーザーは厳密なBoolean問い合わせ言語を習得する必要があった。1980年代には、DialogやLexisNexisといった有料データベースシステムが専門家向けサービスを提供していた。

1990年代のインターネットの爆発的拡大は、情報検索に革命をもたらした。初期のウェブ検索エンジン（Altavista、Infoseekなど）は、ページ内のテキスト情報を単純に統計的にランク付けしていた。しかし、この方法は容易にスパム化可能であり、検索結果の品質は実用的とは言い難かった。

根本的な転換は、ウェブの構造的特性――すなわちハイパーリンク――に着目することから生まれた。1997年、スタンフォード大学の学部生セルゲイ・ブリンとラリー・ペイジは、ウェブページの「リンク人気度（link popularity）」に基づいてページをランク付けするPageRankアルゴリズムを開発した。PageRankの基本的な直感は単純である：あるページへのリンク数が多いほど、またそのリンク元のページが重要であるほど、そのページは重要である。数学的には、PageRankはマルコフ連鎖の定常分布に対応し、ランダムウォーカーがウェブ上を歩き続けたときにどのページにいる確率が最も高いかを計算する。

1998年に設立されたGoogle（初期の正式名は「BackRub」の後にGoogleに改名）は、PageRankアルゴリズムとテキスト関連性スコアを組み合わせることで、当時の他の検索エンジンをはるかに上回る検索品質を実現した。Googleの登場は、単なる技術的な改善ではなく、情報アクセスの民主化であった。ユーザーは複雑なクエリ言語を学ぶ必要がなく、自然言語に近い問い合わせで有用な結果を得られるようになった。

情報検索における統計的な進化は、検索エンジンに限定されない。自動分類、クラスタリング、質問応答（question answering）、テキスト要約など、あらゆるIR関連タスクが統計的学習アプローチを取り入れ始めた。特に、学習ランキング（learning to rank）という分野が発展し、機械学習モデルが直接的にランク付けの品質を最適化する方法が研究されるようになった。

10.4　Word2Vec以前の分散意味論 —— LSA、LDA

単語の統計的特性から意味的情報を抽出するという問題は、1990年代から活発に研究されていた。その最初の成功例は、潜在意味解析（Latent Semantic Analysis, LSA）である。

1990年、スコット・ディアウェスターらは、文書における単語の共起パターンから潜在的な意味構造を発見する手法を提案した。LSAの基本的な考え方は次の通りである。単語と文書の関係を行列形式で表現し（行が単語、列が文書、要素が単語の出現回数）、この行列に特異値分解（Singular Value Decomposition, SVD）を適用する。SVDにより、元の高次元の単語空間を低次元に圧縮することができ、この圧縮された空間では、同義語的な単語は近い位置に配置される。

LSAの重要な含意は、単語と文書の関係に潜在する意味的構造が統計的に抽出可能であるということを示した点にある。たとえ検索クエリが特定の単語を含まなくても、LSA空間では意味的に関連する文書が検出される。この「隠れた意味」へのアクセス可能性は、情報検索の精度向上に貢献し、LSAは1990年代の情報検索システムで広く採用された。

しかし、LSAは線形代数的手法であり、確率的な基礎を欠いていた。この限界を補うため、1999年にトーマス・ホフマンは確率的潜在意味解析（Probabilistic Latent Semantic Analysis, PLSA）を提案した。PLSAは、文書生成過程を確率的にモデル化する。具体的には、各文書は、複数の隠れた「トピック」の混合分布から単語をサンプリングして生成されると仮定する。トピックは単語上の確率分布として定義され、文書ごとにトピック混合比が異なるというモデルである。PLSAにより、統計的に厳密な基盤の上でセマンティック構造を学習することが可能になった。

2003年、デイビッド・ブライ、アンドリュー・ング、マイケル・ジョーダンは、PLSAの確率的基礎をさらに強化した潜在ディリクレ配分（Latent Dirichlet Allocation, LDA）を提案した。LDAは、PLSAの「トピック混合比が文書ごとに固定される」という制約を緩和し、ディリクレ分布を用いた事前分布（Dirichlet　prior）を導入している。この修正により、LDAはより数学的に望ましい性質を持つようになり、過学習に対してよりロバストになった。

LDAとPLSAの関係は、統計的学習における「事前分布の重要性」を示す典型例である。PLSAは最尤推定によってパラメータを学習し、これは観測データに完全に適合するモデルを学習する傾向がある。一方、LDAはベイズ的アプローチを採用し、パラメータの不確実性を明示的にモデル化することで、一般化性能を向上させている。

LSA、PLSA、LDAという一連の進化は、単なる手法の改良ではなく、統計的NLPの基本的な思考様式の発展を示している。「隠れた変数を用いた確率的生成モデル」というフレームワークは、後の深層学習時代まで自然言語処理の理論的基盤となる。

10.5　音声認識の進歩 —— 隠れマルコフモデルからDNNへ

音声認識は、統計的NLPが実用的成果を上げた領域の一つである。初期の音声認識システムは、テンプレートマッチングなどの単純な手法に依存していたが、1980年代以降、隠れマルコフモデル（Hidden Markov Model, HMM）が主流技術となった。

HMMの適用は、音声信号の時系列的性質を明確に数学的に定式化した。音声は、時間軸に沿って変化する音響特徴量の系列として表現される。各時点では、複数の隠れた音素状態の一つが活性化しており、その状態から特定の音響特徴が確率的に出力される。HMMはこのプロセスを、遷移確率と出力確率の二つの確率分布で記述する。1990年代を通じて、HMMベースの音声認識システムは、大規模音声コーパスの利用可能性の向上に伴い、着実に精度を改善していった。

HMMの効果的な運用には、音響モデル（HMM自体）と言語モデルの二つの成分が必要であった。音響モデルは、音声信号から隠れた音素列への確率的写像を学習し、言語モデルは、自然言語の確率的特性を捉える。言語モデルとしてN-gramモデルが広く採用された。N-gramは、テキストコーパスから単語

（あるいは文字）の連続パターンを統計的に抽出し、現在の単語の確率を直前のN-1個の単語に条件付けるモデルである。このシンプルな手法は驚くほど効果的であり、1990年代から2000年代初頭まで標準的な言語モデルとして使用され続けた。

音声認識システムの精度は、2010年代に新しい転換点を迎える。深層学習の普及に伴い、ガウス混合モデル（Gaussian Mixture Model, GMM）ベースの音響モデルが、深層ニューラルネットワーク（DNN）で置き換えられるようになった。DNN音響モデルは、複雑な非線形の特徴表現を学習することで、従来のHMM-GMM手法を大幅に上回る性能を実現した。このDNN-HMM統合システムは、HMMの時系列モデリング能力とDNNの表現学習能力を結合し、2010年代の標準的な音声認識プラットフォームとなった。

10.6　推薦システムの発展 —— 協調フィルタリングからNetflix Prizeへ

推薦システムも、統計的機械学習が実務的な成功を遂行した領域である。1990年代から2000年代初期、Amazonや映画販売サイトなどのオンラインプラットフォームでは、顧客の購買行動をモデル化し、個人ごとに異なる推薦を提供する需要が高まっていた。

推薦システムの基本的なアプローチは協調フィルタリング（collaborative filtering）である。その直感は単純である：自分と同じような好みを持つ他のユーザーが好んだアイテムは、自分にも好まれる可能性が高い。協調フィルタリングは、顧客-アイテム評価行列（各要素が特定の顧客による特定のアイテムの評価）から、観測されていない要素を予測する問題として定式化される。この問題は、行列補完（matrix completion）または低ランク行列近似の問題である。

1990年代の協調フィルタリング手法は、最近傍ユーザーを見つけてその評価を平均化するなど、比較的シンプルであった。しかし、大規模なデータセットに対しては計算上の課題が生じた。2000年代に入ると、行列因子分解（matrix factorization）技術がこの問題の主流ソリューションとなる。行列因子分解では、評価行列Rを二つの低ランク行列の積（ユーザー潜在要因行列とアイテム潜在要因行列）として分解する。これにより、スパースな観測行列からの有効な予測が可能になり、同時に計算量も削減される。

推薦システムの統計的手法の有効性が最も劇的に示された事例は、Netflix Prize（2006-2009）である。2006年、映画配信企業Netflixは、自社の推薦アルゴリズムを改善する目的で、公開競技コンペティションを開催した。訓練データとして、約48万人のユーザーが約1800万件の映画評価を行った大規模なレイティングデータセット（100,480,507件の評価）を提供した。優勝者には100万ドルの賞金が提供された。

このコンペティションは、複数の理由で歴史的意義を持つ。第一に、機械学習アルゴリズムの競争を通じた改善の可能性を、きわめて明確に実証した。第二に、大規模で現実的なデータセットがパブリックベンチマークとして提供されることで、研究コミュニティ全体で方法論の進展が加速されることを示した。第三に、単純な個別手法の組み合わせ（アンサンブル学習）が、個々の洗練された手法よりも優れた性能を発揮することを実証した。

優勝チーム「BellKor's Pragmatic Chaos」は、2009年9月21日、Netflixの内部アルゴリズムをテスト集合上で10.06%上回る精度を達成し、100万ドルの賞金を獲得した。彼らの方法論は、行列因子分解、勾配ブースティング、時間的効果のモデリングなど、複数の統計的手法を慎重に組み合わせたエンサンブル手法であった。Netflix　Prizeの競技期間全体を通じて、推薦システムの精度は継続的に改善され、研究者と実務家の間での知見共有が促進された。

Netflix　Prizeの意義は単なる技術的改善にとどまらない。このコンペティションは、機械学習による問題解決が、学術的な関心事ばかりでなく、商用的に重要な実務的価値を持つことを、産業界と一般社会に対して明確に示した。これは、その後の機械学習ブームの波及力を高める要因の一つとなった。

章末の展望

第10章では、統計的NLPの成立と展開を四つの軸で検討してきた。第一は、方法論の転換である――記号主義的ルールベースのアプローチから、データ駆動型の統計的学習へのパラダイム転換がいかにして成立し、正当化されたのかを検討した。第二は、このパラダイム転換を可能にした物質的・技術的基盤である――大規模コーパスの入手可能性、計算資源の充実、理論的手法の発展である。第三は、これらの抽象的な進展が、機械翻訳、情報検索、音声認識、推薦システムといった具体的な応用領域でいかに具現化されたかである。第四は、隠れた変数を用いた確率的生成モデルというフレームワークが、NLPの理論的基盤として機能し始めたことである。

1990年代から2000年代初期にかけての統計的NLPの勃興は、AI研究全体における方向転換の一部であった。同じ時期、コンピュータビジョン（第11章で詳述）、そして強化学習（第9章）の領域でも、統計的機械学習が主流となり始めていた。しかし、統計的手法の限界もまた明らかになり始めていた。統計的手法は、大規模な訓練データから規則性を抽出することに優れていたが、稀な現象、長い尾を引く語彙、訓練データに現れない複合的パターンに対しては脆弱であった。Word2Vec以前の分散意味論（LSA、PLSA、LDA）は、単語の静的な意味表現を学習したが、文脈依存的な意味、あるいは動的な表現学習には対応していなかった。

これらの制約を突破する次のパラダイムの転換は、深層学習の本格的な普及と、大規模データの一層の増大を待つことになる。次章では、コンピュータビジョンにおける統計的学習と深層学習の相互作用を検討する。その後、第14章以降は深層学習時代の到来と、その自然言語処理、ビジョン、そして科学応用への影響を追跡することになる。

統計的NLPの歴史は、単なる技術史ではない。それは、「知識とは何か」「学習とは何か」という根本的な問いに対する、AI研究の回答の変遷を示している。記号主義が「知識は明示的に記述可能な規則である」と想定したのに対し、統計的NLPは「知識は大規模データから統計的に抽出される暗黙的パターンである」という見方を提示した。この転換は、後の深層学習の台頭と、その現在の支配的地位を理解する上で、不可欠な知的遺産である。

参考資料（本章）

- Brown, P. F., et al., IBM statistical machine translation 関連論文群（1990年代前半）。
- Deerwester, S., et al. “Indexing by Latent Semantic Analysis” (1990).
- Hofmann, T. “Probabilistic Latent Semantic Analysis” (1999).
- Blei, D. M., Ng, A. Y., & Jordan, M. I. “Latent Dirichlet Allocation” (2003).
- Google 公式ブログ・研究ブログ。Google Translate が2000年代後半に統計的機械翻訳を中核化した経緯の確認に使用。
- Brin, S., & Page, L. “The Anatomy of a Large-Scale Hypertextual Web Search Engine” (1998).
- Netflix Prize 公式資料および BellKor’s Pragmatic Chaos 関連論文。推薦システム史の確認に使用。

第11章　コンピュータビジョンの発展

11.1　エッジ検出からオブジェクト認識へ

コンピュータビジョン研究は、画像から有意な情報を抽出するという最も基本的な問題から出発した。その最初の関心は、画像内の輪郭や境界の検出にあった。なぜなら、物体の認識とは根本的には、物体の形状的特徴をデジタル画像から抽出することであり、その最初の一歩は、画像内で急激に輝度が変化する領域――すなわちエッジを検出することだからである。

1960年代から1970年代初頭にかけて、Sobel作用素やLaplacian検出器といった単純な勾配ベースのエッジ検出手法が開発された。これらの手法は、隣接ピクセルの輝度差分を計算することで局所的な勾配を推定し、その大きさがしきい値を超える領域をエッジとして検出するというものであった。その後、1980年代初頭にはMarr-Hildreth流の理論化が進み、エッジ検出はより明確な計算論的基盤の上に置かれるようになった。

この時期のコンピュータビジョン研究の根本的な制約は、単なる計算資源の不足ではなく、視覚理解の問題構造そのものが十分に理解されていなかったことである。画像から三次元物体を認識するとはどのような計算問題なのか、その本質的な構造が明確でなかったため、研究者たちは本当に解くべき問題が何であるかさえ十分に把握していなかった。

11.2　Marrの計算理論（1982）とビジョン研究の枠組み

1982年、MITの認知科学者デイヴィッド・マーは『Vision: A Computational Investigation into the Human Representation and Processing of Visual Information』を出版した。この著作は、視覚に対する根本的に異なるアプローチを提示し、その後のコンピュータビジョン研究に深刻な影響を与えることになった。

マーの中心的な主張は、情報処理システム――特に視覚システム――を理解するには三つのレベルでの分析が必要だというものであった。第一が「計算論的レベル」であり、ここでは視覚の計算的目標が何であるか、そしてそれを達成するための戦略が何であるかを明確にする。第二が「表現とアルゴリズムのレベル」であり、計算論を具体的なデータ構造とそれを操作するアルゴリズムとして実装する。第三が「ハードウェア実装のレベル」であり、アルゴリズムを神経機構や電子回路として物理的に実現する。

マーが強調したのは、これら三つのレベルは互いに独立した分析対象であり、混同することは科学的誤謬につながるということであった。ハードウェア実装の詳細が理論的理解に必要とは限らず、逆に、計算論的な理解がアルゴリズムの選択を強く制約することもあるという発想は、AI研究とコンピュータ科学に根本的な視点の転換をもたらした。

マーのビジョン理論の具体的な構成は、画像から三次元世界の表現へと段階的に進む過程として描かれた。まず原画像から局所的な明度勾配を検出し「前原始スケッチ（primal sketch）」を生成する。次にこれを発展させて「2.5次元スケッチ」を作成する。これは視点依存的な三次元表現であり、表面の向きや奥行きを符号化する。最終段階では「三次元モデル表現」に到達し、オブジェクト中心座標系での物体の形状が表現される。この段階的な構成は、視覚処理の計算的本質を深く考えさせるものであり、後の深層学習研究においても、多層ニューラルネットワークが段階的に複雑な特徴を学習するという構造に響き渡ることになった。

マーの理論はまた、視覚研究に物理学的制約（physical constraints）の重要性を導入した。例えば、表面の光沢度は通常ゆるやかに変化すること、輝度の急激な変化はしばしば物体の輪郭であること、視差によって三次元構造を復元できることなど、物理的世界に成り立つ規則性を視覚アルゴリズムに組み込むべきだという主張であった。このアプローチは、単なるデータ駆動的なパターン認識ではなく、視覚の根本的な制約を考慮した問題設定の重要性を示唆していた。

マーの理論の限界もまた注目に値する。彼の枠組みは、画像解析の早期段階（エッジ検出、輪郭追跡）に関しては非常に説得力があったが、高度なセマンティック理解――つまり、検出された物体が「何であるか」を理解する段階――には十分に対応できなかった。この限界は、後に統計的機械学習と深層学習がコンピュータビジョンに進出する際の理論的基盤となるであろう。

11.3　SIFT・HOG ―― 手工的特徴量の時代

1990年代から2000年代初頭にかけて、コンピュータビジョン研究は「手工的特徴量（handcrafted features）」の時代を迎えた。これは、画像から人間が設計した特定の特徴を抽出し、それらを基にして物体認識や画像マッチングを行う方法論である。この時期の研究は、物体の本質的な視覚特性を数学的に形式化することに集中していた。

1999年、ブリティッシュコロンビア大学のデイヴィッド・ロウは「Scale-Invariant Feature Transform（SIFT）」の原型を発表し、2004年に詳しい記述を与えた。SIFTの本質的な革新は、スケール不変性と回転不変性を持つ局所的な特徴を検出できるという点にあった。従来の特徴検出手法は、画像の拡大・縮小や回転に対して不安定であり、実用的な応用場面ではほぼ利用不可能であった。ロウが提示した解決策は、ガウシアン差分（Difference of Gaussians, DoG）を用いて複数のスケールレベルで特徴点を検出し、各特徴点周辺の勾配方向の分布を記述することであった。

SIFTの手法の流れは以下のようである。まず、入力画像に対して複数のスケール段階でガウシアンフィルタを適用し、スケール空間を構築する。次に、隣接するスケール間のガウシアンフィルタ適用画像の差分を計算してDoGを生成する。このDoGの極大値と極小値がキーポイント（特徴点）の候補となる。低コ

ントラストの候補点やエッジ応答が強い点を除外した後、残存するキーポイントに対して支配的な方向を割り当てる。最後に、各キーポイント周辺の勾配方向のヒストグラムを計算し、128次元のディスクリプタ（特徴ベクトル）として表現する。

SIFTの実用的な意義は計り知れなかった。この特徴は画像のスケール変化、回転、照度変化に対してロバスト（robust）であり、また異なる視点からの同一物体の画像にも安定して対応できた。SIFTを用いると、大規模な画像データベースから特定の物体を検索できるようになり、画像ステッチング（複数の写真をつなぎ合わせてパノラマ画像を生成する）、3次元復元、物体追跡、さらには野生動物個体識別などの応用が可能となった。SIFTは2000年代のコンピュータビジョン研究における実質的なデ・ファクト・スタンダードとなり、この手法なしには多くの実践的問題は解けないと考えられていた。

SIFT成功の背景には、コンピュータビジョン問題が根本的には「パターンマッチング」の問題であるという認識があった。もし物体の本質的な視覚特性を正確に記述できれば、異なる画像間でそれらを確実にマッチングできるというのが基本的な想定である。

2005年、フランスの国立情報学自動化研究所（INRIA）の研究者ナヴィーン・ダラルとビル・トリグスは「Histograms of Oriented Gradients（HOG）」と呼ぶ特徴記述手法を発表した。HOGは、画像領域を細粒度の空間セルに分割し、各セル内でのエッジ方向（勾配方向）の分布をヒストグラムとして計算するというシンプルだが効果的な手法である。彼らの実験的発見は、9方向のヒストグラムビンを用い、4×4ピクセルのセルを使用し、16×16ピクセルのブロック内で局所的なコントラスト正規化を行うことが人物検出タスクにおいて有効であることを示していた。

HOGの特筆すべき点は、その単純性にもかかわらず、既存の手法を大幅に上回る性能を達成したことである。ダラルとトリグスの研究は、「局所的な物体形状と外観は、勾配方向の分布によって記述できる」という直感的ながら強力な原理を示唆していた。HOGは特にスケール変化に強く、照度変化にも比較的ロバストであり、その後の10年間、人物検出、車両検出、一般的なオブジェクト認識における標準的な特徴抽出手法として広く採用されることになった。

SIFTとHOGの成功は、統計的機械学習がコンピュータビジョンに本格的に進出する前夜の状況を象徴している。これら手工的特徴量に基づくアプローチは、物体認識問題を「適切な特徴抽出」と「効率的な分類器（通常はサポートベクターマシンなど）」の組み合わせに還元する。この二段階アプローチは、2010年代に深層学習が台頭するまで、コンピュータビジョンの実践的な方法論として君臨することになるのである。

11.4　ImageNetの構築と大規模ベンチマークの学術的意義

21世紀初頭のコンピュータビジョン研究は、パラダイムシフトの前夜にあった。SIFT、HOG、SVM（サポートベクターマシン）といった手法により、狭い領域での認識タスク――例えば顔検出や車両認識――では実用的な性能を達成するようになっていた。しかし、「任意の自然画像に写る任意の物体を認識する」という汎用的なオブジェクト認識問題は、依然として困難なままであった。

この状況を根本的に変えたのが、2007年から2009年にかけてフェイ=フェイ・リらによって構築された大規模画像データセット「ImageNet」であった。プリンストン大学に着任したリが直面したのは、次のような問題構造であった：機械学習の理論や計算アルゴリズムは急速に進歩しているのに対し、これらを訓練するためのデータは相対的に貧弱であり、また既存のデータセットは現実世界の多様性をほとんど反映していないという課題である。

認知心理学者アーヴィング・ビーダーマンの推定によれば、人間は約30,000個の物体カテゴリを認識できるとされていた。リはこの観察から着想を得て、人間の視覚認識能力の多様性に匹敵するような大規模で多様な画像データセットの構築を構想した。

ImageNetの構築は技術的・組織的な困難に満ちていた。合計1,400万枚規模の画像を22,000超のカテゴリに結びつけ、一部のサブセットには矩形領域でのバウンディングボックスのアノテーション（手作業による標注）も施す必要があった。リらが採用した戦略は、Amazon Mechanical Turkという群衆労働プラットフォームを活用し、世界中の労働者に小額の報酬の引き換えに画像のラベル付けを依頼するという大規模な衆知（crowdsourcing）アプローチであった。これにより、わずか数年で前例のない規模のデータセットを実現することが可能となった。

ImageNetの学術的意義は、単なるデータセットの提供にとどまらない。2010年より、ImageNetの部分集合（ILSVRC: ImageNet Large Scale Visual Recognition Challenge）の1,000カテゴリから成る画像分類タスクを用いた年次的な競技会が開催されるようになった。この競技会は、コンピュータビジョン研究の標準的なベンチマークとなり、新しいアルゴリズムの有効性を客観的に評価するための場を提供した。

ILSVRCの開催がもたらした影響の大きさは、今日のコンピュータビジョン研究から見返すと自明であるが、当時はそうではなかった。2012年のILSVRCで、トロント大学の研究グループがアレックス・クリジェフスキー、イリヤ・サツキヴァー、ジェフリー・ヒントンの指導の下で開発した「AlexNet」が、従来の手工的特徴量に基づく手法を圧倒的なマージンで上回る正解率を達成したとき、初めてこのベンチマークの歴史的重要性が広く認識されることになったのである。しかし、それは第14章の主題である。

ここで重要な点は、ImageNetという大規模で多様なデータセットの存在と、ILSVRCという標準化されたベンチマークが、コンピュータビジョン研究のパラダイム転換を可能にしたということである。「評価可能な問題設定」がなければ、深層学習の優位性を実証することすら困難であったであろう。この意味で、ImageNetはコンピュータビジョン研究における最も重要なインフラストラクチャの一つである。

11.5　顔認識技術の進化と社会的含意

顔認識は、コンピュータビジョンの応用として最も直接的に人間の生活に影響を与える領域の一つである。顔認識技術の発展は、純粋な学術的関心のみならず、セキュリティ、生体認証、法執行機関による監視など、社会的に重要な領域に関わっている。

2001年、ミツビシ電機の研究所（Mitsubishi Electric Research Laboratories）のポール・ヴィオラとマイケル・ジョーンズは、画像内の顔を高速かつ高精度に検出する手法を発表した。彼らは「Haar的特徴（Haar-like features）」と呼ばれる矩形フィルタを用いて、画像内の局所的なコントラストパターンを抽出し、これらを階層的に組み合わせたブースティング分類器（boosted classifiers）により高速に顔候補を絞り込んでいく方式を採用した。この「カスケード分類器」のアーキテクチャにより、彼らのアルゴリズムは当時のコンピュータで秒間15フレームもの速度で顔を検出することが可能になった。

Viola-Jonesアルゴリズムの実用的な威力は迅速に認識された。このアルゴリズムはシンプルで計算量が少なく、メモリ効率も良いため、組み込みシステムや携帯電話など、リソースが限定されたデバイスに実装することが容易であった。その結果、わずか数年のうちに、デジタルカメラやスマートフォンなどの標準的な撮影デバイスに組み込まれるようになったのである。

しかし、顔認識技術の急速な普及にともなって、深刻な社会的問題が顕在化し始めていた。第一に、プライバシーの問題である。大規模な顔認識システムが展開されるに伴い、個人の同意なしに、または個人が知らないうちに顔画像が収集・データベース化・検索される可能性が出現した。公共の場での監視カメラ映像、身分証や運転免許証のデータベース、逮捕者の顔写真を集めた警察の書類（mugshot）など、多様なデータベースが顔認識システムに接続されるようになったのである。

第二に、アルゴリズムバイアス（algorithmic bias）の問題である。複数の研究により、顔認識アルゴリズムは特定の人口集団に対して系統的に低い精度を示すことが明らかになった。特に、肌の色や性別によって誤識別率に差が出ることが文献で報告されている。2019年の米国標準技術研究所（NIST）による大規模評価では、多数のアルゴリズムにおいて人口集団間の誤識別率の差が観察された。

このバイアスの根本原因は、訓練データの統計的偏りにある。多くの顔認識システムが、特定の地域・肌色・性別分布に偏った画像群で訓練されていたため、その偏りが学習済みモデルに内在化されていたのである。さらに深刻な点は、このアルゴリズム的な不公正が、すでに歴史的に脆弱な立場にある人口集団に不均等な害をもたらしうるということである。

第三に、司法制度における濫用の可能性である。顔認識システムは、その性質上、完全な確実性を持たず、誤りの可能性が常に存在する。しかし一度顔認識システムが容疑者リストを生成すると、その結論は法執行機関の意思決定に強い影響を及ぼす傾向がある。結果として、無実の人物が取調べを受ける、航空機搭乗を拒否される、テロ容疑者リストに登録されるなど、深刻な害を被るケースが報告されている。

これらの懸念に対し、2020年代に入るにつれ、多くの民主主義国家は顔認識システムの公的利用に対する規制を強化し始めた。特に、生体認証データを含む個人情報のプライバシー保護、アルゴリズムの透明性と説明可能性の確保、バイアス除去のための技術的・制度的方策が重要な政策課題として認識されるようになったのである。この問題については、第22章で詳述する。

11.6　医療画像解析への応用の始まり

コンピュータビジョンの科学的・医学的応用は、2000年代から次第に本格化し始めていた。特に医療分野では、放射線医学（radiology）、病理学（pathology）、循環器医学（cardiology）などの領域で、高度な画像解析が診断の精度向上と効率化に不可欠になりつつあった。

医療画像解析における課題は、一般的なオブジェクト認識以上に複雑である。医療画像（CT、MRI、X線画像、超音波画像など）には、自然画像とは異なる統計的特性があり、またアノテーション（診断医師による標注）の品質が極めて重要である。さらに、医療診断の文脈では、偽陰性（実際には異常であるのに検出されない）の代償が偽陽性（正常なのに異常と判定される）の代償よりも遥かに大きいため、アルゴリズムの信頼性と説明可能性が人命に直結する。

2000年代初頭の医療画像解析は、依然として人間の医師による目視検査が中心であり、コンピュータは補助的な役割にとどまっていた。しかし、デジタル画像化の進展、電子カルテシステムの普及、そして統計的機械学習技術の成熟にともない、次第にコンピュータによる自動解析の可能性が認識されるようになった。SIFT、HOG、SVMといった技術は、腫瘍検出、血管抽出、肺結節分類など、特定の医療画像解析タスクに適用され始めたのである。

このころのアプローチは依然として領域特化的であり、汎用的なアルゴリズムはない。各医療応用領域は、その特有の画像特性と診断的目標に応じて、カスタマイズされた特徴抽出とパイプラインを必要とした。しかし、2010年代に深層学習が医療画像解析に本格的に進出すると、この状況は劇的に変わることになる。その発展については後の章で論じる。

本章ではコンピュータビジョンが、幾何学的エッジ検出（1970年代）から手工的特徴量（1990年代〜2000年代）への進化を追跡した。Marrの計算理論（1982）が視覚問題の問題構造そのものを根本的に問い直したのに対し、SIFT（1999）とHOG（2005）は実用的なパラダイムを確立した。ImageNetの構築（2009）とILSVRCベンチマーク（2010–）は、コンピュータビジョン研究に量的な飛躍をもたらす必須の基盤となったのである。

同時に、顔認識技術の急速な普及は、先制的な技術進歩と社会的正義の問題がいかに深く絡み合っているかを明らかにした。第11章を通じて、われわれが見いだした重要な洞察は、以下の通りである：コンピュータビジョンの進化は単なる技術的な問題設定の洗練ではなく、何を「見える」ことが重要であり、誰が見る権利を持つのか、そして誰が見られることを強制されるのかという、深刻な権力関係を含んでいるという点である。

第12章では、身体性を持つ知能――ロボティクス――へと視点を転ずる。視覚と運動制御の統合、そして環境との継続的な相互作用を通じた学習と適応という新しいテーマが現れることになるのである。

参考資料（本章）

- David Marr, Vision (1982).
- David G. Lowe, SIFT 関連論文（1999, 2004）。
- Navneet Dalal and Bill Triggs, “Histograms of Oriented Gradients for Human Detection” (2005).
- Fei-Fei Li, Jia Deng, et al. “ImageNet: A Large-Scale Hierarchical Image Database” (2009).
- NIST, Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects (2019). 顔認識における人口集団差の確認に使用。
- Paul Viola and Michael Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features” (2001).

第12章　ロボティクスと身体性の知能

12.1　産業用ロボットの発展 —— Unimate（1961）からFANUC・ABBへ

ロボティクスの歴史をたどるとき、その出発点は意外なほど古い。しかし、本章の冒頭では、産業用ロボットの実現という観点から、1961年を一つの起点として設定することにする。この年、ジョージ・デボルとジョセフ・エンゲルバーガーのコンビが開発した「ユニメート（Unimate）」が、ゼネラル・モーターズのイーウィング・タウンシップの鋳造工場に導入され、最初の実用的な産業用ロボットが現実のものとなったからである。

デボルは1954年に「自動化されたアーティクル転送（Programmed Article Transfer）」に関する特許を出願し、1961年に取得している。このユニメートは、油圧駆動の多関節マニピュレータであり、複数の位置を順序立てて記録し、その動作を再生する仕組みで動作した。オペレータが機械の先端を目的の位置に動かすと、位置センサがその座標を記録し、この一連の動作シーケンスをロボットが正確に繰り返すことができたのである。ゼネラル・モーターズでは、この初期のユニメートを用いて、溶融した金属を鋳型から取り出し、積み上げるという危険で反復的な作業を自動化した。

ユニメートの成功は、産業オートメーションの新たな可能性を示すものであった。それは単なる機械的な自動化ではなく、デジタルコンピュータと結合された「プログラム可能な」機械であり、作業内容の変更に応じて動作を柔軟に修正できるという柔軟性をもたらした。この特性は、個別・少量生産から多品種少量生産への移行を可能にし、製造業の産業構造そのものを変えていくことになる。

その後、産業用ロボットの市場は急速に成長した。日本のファナック（FANUC）は、1956年に富士通の数値制御装置部門として発足し、1970年代から本格的にロボット製造に乗り出した。ファナックは、ロボットアームの多軸制御、効率的なプログラミング、そして何より高い信頼性と耐久性を特徴としていた。ファナックの成功は、単に技術的優位だけではなく、製造業における日本の産業戦略——すなわち、電子制御技術とロボット化による競争力の確保——と密接に結びついていた。

スウェーデン発祥のASEAは、1974年に電動・全電子制御の産業用ロボットIRB 6を投入し、欧州ロボット産業の先駆となった。1988年にASEAとBBC Brown Boveriが合併して成立したABBは、その系譜を引き継ぎ、自動車産業における溶接・塗装・組立の領域で高い市場シェアを獲得した。後には協調ロボット（cobot）への進出によって、産業オートメーションの新たな段階へと移行していく。

2010年代から2020年代にかけて、ファナックとABBは共に、従来の産業用ロボット市場での支配的地位を保ちながら、同時にAIと機械学習の統合に向かい始めている。ここで重要なのは、産業用ロボットが単なる「自動機械」から、より柔軟で知覚的なシステムへと緩やかに移行しつつあるという歴史的転換点である。

12.2　行動ベースロボティクス ―― ブルックスのサブサンプション・アーキテクチャ

1980年代のAI研究は、一つの根本的な疑問に直面していた。それは、記号処理に基づいた伝統的なAIアプローチが、実世界のロボット制御にいかに不適切であるかという問題であった。従来のロボティクスは、「感知-計画-行動（sense-plan-act）」という基本サイクルに従っていた。すなわち、ロボットは環境からの入力を受け取り、内部モデルを構築し、その上で計画立案を行い、最後に行動を実行するというステップを踏んでいたのである。しかし、こうしたアプローチは、複雑な実世界環境では計算が爆発し、リアルタイムな応答が不可能になるという本質的な限界を抱えていた。

ロドニー・ブルックスは、マサチューセッツ工科大学（MIT）の人工知能研究所において、この課題に根本的な異議を唱えた。1986年に発表された彼のサブサンプション・アーキテクチャ（subsumption architecture）は、「知能は表象を必要としない（intelligence without representation）」という大胆な主張を体現していた。ブルックスの着想の源泉は、ゴキブリやロボットが実環境でどのように行動するかを観察することにあった。複雑な動物ですら、内部モデルなしに、環境との直接的な相互作用を通じて適応的に行動している。ならば、ロボットもまた、象徴的な世界表現を持たずに、感覚と運動の結合を通じて振る舞うことが可能ではないか――これがブルックスの問題設定であった。

サブサンプション・アーキテクチャの技術的特徴は、階層化された行動層の組織方式にある。各層は特定の行動能力（例えば、障害物回避、さまよい歩き、目標追求）を実装した有限状態機械として設計される。層間の制御は、抑制（inhibition）と抑圧（suppression）の機構を通じて行われる。下位層の出力が上位層に基づいて抑制され、より優先度の高い行動が低位層の信号を上書きすることで、複合的かつ適応的な行動が生じるのである。この設計哲学は、中枢神経系の実際の構造に着想を得ていると同時に、計算効率という工学的要求をも満たしていた。

ブルックスが1980年代末から1990年代初頭に開発した「アレン」（Allen）、「ハーバート」（Herbert）、「ジェンギス」（Genghis）といった自律型ロボットは、比較的単純な制御層の組み合わせだけで、複雑な地形移動や環境応答を実現することに成功した。これらのロボットは中央集権的な世界モデルに依存せず、局所的な制御回路の相互作用によって、全体としての協調的行動を生み出していた。このアプローチは、伝統的なAIの「頭脳集中型」の設計に対して、「分散型」あるいは「身体分散型」の知能概念を提示したのである。

さらに重要なのは、ブルックスの思想が引き出した哲学的含意である。もし知能が内部表象を必要としないのであれば、心身問題やクオーリアの問題をめぐる伝統的な問い立てそのものが再考を余儀なくされる。この問題は、第24章において再び論じられることになる。同時に、サブサンプション・アーキテク

チャは実践的な成功を収め、ブルックスが共同設立した「アイロボット（iRobot）」は、2000年代には掃除ロボット「ルンバ（Roomba）」の開発を通じて、家庭用ロボットの普及の先駆となった。これは、ブルックスの理論が単なる学術的好奇心ではなく、実世界的な価値をもつことを証明している。

12.3　確率的ロボティクス ―― SLAM、カルマンフィルタ、パーティクルフィルタ

ブルックスのサブサンプション・アーキテクチャが、知能の脱中央化と身体的即応性を強調したのに対し、1990年代から2000年代にかけて発展した「確率的ロボティクス」は、別の方向を指し示していた。これは、不確実な環境下での推論を確率論的に定式化し、ロボットが自己位置を推定し、同時に地図を構築するという複合的な問題を解くアプローチである。

その中核をなすのが、「同時位置特定と地図作成（Simultaneous Localization and Mapping, SLAM）」という問題の形式化である。移動ロボットが未知の環境に置かれたとき、それは二つの根本的な困難に直面する。一つは、自分の位置がわからないこと。もう一つは、その位置を正確に知らなくては地図を正確に構築できないという循環的な依存関係である。この鶏と卵の問題をいかに解くかが、1990年代のロボティクス研究の中心的課題となったのである。

カルマンフィルタは、この問題に対する最初の有力な解答であった。ルドルフ・カルマンが1960年に開発したこのフィルタリング手法は、もともと宇宙船の軌道推定のために設計されたものであるが、1990年代には拡張カルマンフィルタ（EKF）としてロボットのSLAMに応用されるようになった。カルマンフィルタの原理は、逐次的なベイズ推定に基づいている。ロボットの状態（位置・向き・速度）をベクトルで表現し、センサからの観測値と運動モデルの予測を組み合わせることで、事後確率分布を更新していくのである。この手法は、線形ガウシアン仮定の下では最適解を与え、計算効率もよいため、1990年代から2000年代初期まで産業用・研究用ロボットのナビゲーションにおける標準的な選択肢であった。

しかし、カルマンフィルタには本質的な制約がある。それは線形性とガウシアン性の仮定である。非線形な環境ダイナミクスやマルチモーダルな信念分布を扱う場合には、性能が劣化しやすい。こうした限界に対応するために、セバスチャン・スルンらは、パーティクルフィルタ（particle　filter）という手法をロボティクスへ広く導入した。パーティクルフィルタは、ベイズ推定を離散的な「パーティクル」（仮説的な状態）の集合により近似する手法である。各パーティクルは可能な状態を代表し、その重みは観測との合致度により決定される。観測が得られるたびに、パーティクルが重みに応じてリサンプリングされることで、非線形で複雑な信念分布を柔軟に追跡できるようになるのである。

パーティクルフィルタに基づいた「FastSLAM」アルゴリズムは、ロボットの自己位置推定と環境ランドマークの位置推定を効率的に融合させ、ランドマーク数に対して対数的な更新を実現した。この理論的進展は、2000年代の自律移動ロボットの商用化を可能にした重要な基盤となる。

12.4　自動運転技術の系譜 ―― DARPAグランドチャレンジからWaymo・Teslaへ

自動運転という社会的に重要な問題領域において、技術的ブレイクスルーが何によってもたらされるかを理解することは、AI史全体を通した学習の機会である。2004年から2007年にかけて実施された「DARPAグランドチャレンジ」は、まさにそうした転機となる出来事であった。

2004年3月13日、アメリカ国防高等研究計画局（DARPA）は、カリフォルニア州モハーヴェ砂漠において、初めてのグランドチャレンジを開催した。課題は、150マイル（240キロメートル）のオフロード経路を、無人の車両が自律的に走破することであった。しかし、この最初のチャレンジは完全な失敗に終わった。参加した15台のうち、どの車両も完走できず、最も遠くまで進んだカーネギーメロン大学の「サンドストーム」でさえ、11.78キロメートルで停止してしまったのである。参加者の多くは、未知の崖や落とし穴のある地形を自律的に走行させることの困難さを痛感することになった。

一年後の2005年10月8日、第二次チャレンジが開催された。ここで劇的な変化が生じた。スタンフォード大学のレーシングチームが開発した「スタンレー」と名付けられた青いフォルクスワーゲン・トゥアレグが、6時間53分58秒で完走し、200万ドルの賞金を獲得したのである。完走したのはスタンレーを含む5台のみであったが、他の参加車両も前年の最高記録を大幅に上回る距離を走破することに成功した。

スタンレーの技術的特徴は、後の自動運転研究の方向性を示すものであった。車両は複数のPentium Mプロセッサによる計算プラットフォームを搭載し、LiDAR、レーダー、カメラ、GPS/慣性計測を組み合わせて地形情報を取得していた。セバスチャン・スルンが指導したこのチームの核となったのは、機械学習と確率的推論の組み合わせであった。特に、不確実性の下での経路計画と障害物回避の問題を、統計的アプローチで解くという戦略が有効であることが実証されたのである。

2007年には、第三のチャレンジとして「アーバン・チャレンジ」が開催された。前二回が砂漠という単純な環境での走行に重点を置いていたのに対し、この都市チャレンジでは、交通信号の遵守、対向車線での追い越し、複数の他車両との相互作用といった、実際の交通環境での制約条件が導入された。カーネギーメロン大学の「ボス（Boss）」がシボレー・タホを用いて優勝したことで、自動運転の技術的困難が単なる物理的な地形走破から、複雑な社会的・法的ルール体系の理解へと進化していることが明らかになった。

DARPAグランドチャレンジが果たした歴史的役割は、技術的なものに留まらない。このコンペティションは、分散した学術・企業の研究者を一堂に集め、ベンチマークの力を通じた知識の可視化と共有をもたらした。失敗した参加者も成功した参加者も、共に「問題は何であるか」と「それをどう定式化するか」という根本的な問い直しを経験したのである。この経験は、後の自動運転技術の急速な発展――Google自動運転車プロジェクト（後のWaymo）、テスラのオートパイロット・Full Self-Driving（FSD）の開発――の知的基盤となる。

Waymoは、セバスチャン・スルン、アンソニー・レヴァンドウスキーらがGoogle内で立ち上げたプロジェクトとして2009年に開始された。Waymoのアプローチは、ハイブリッド・センシング（カメラ、LiDAR、レーダーの複合利用）と複数段階の冗長システムに基づいており、確率的ロボティクスの理論を大規模に展開したものであった。2025年5月時点で、Waymoは米国の複数都市で週25万件超の有償自動運転乗車サービスを運用しており、限定領域ではSAE Level 4相当の自律運行を実現している。

一方、テスラのアプローチは異なる戦略を採っている。イーロン・マスク率いるテスラは、複数の企業・学術機関のシステムとは異なり、カメラ中心のビジョンシステムに基づいた自動運転を指向している。テスラのFSD（Supervised）システムは、実走行データを機械学習の訓練に活用し、エンドツーエンド学習によって運動制御を獲得する戦略を採用している。もっとも、2026年3月時点で公道向け市販機能としてはなおSAE Level 2の監督付き運転支援に位置づけられる。その規模の経済と継続的学習能力は、Waymoとは異なる競争軸を示しているのである。

DARPAグランドチャレンジから20年以上経った現在、自動運転技術は、単なる「技術的問題」から「社会・倫理・法律的問題」へと軸足を移しつつある。これについては、第21章と第23章で改めて論じることになる。

12.5　ヒューマノイドとソフトロボティクス ―― 身体知の探求

ロボティクス研究において、人間型（ヒューマノイド）のロボットを構築することの意味は、単なる工学的目標以上のものである。二足歩行、両腕の並行処理、環境把握と身体操作の統合――こうした身体的能力は、人間にとっては自明であるが、機械の観点からは極めて複雑で非自明な問題を提起する。

ホンダが開発した「ASIMO」（Advanced Step in Innovative Mobility）は、このヒューマノイド研究における一つの象徴的な成果である。ホンダの人型ロボット研究は1986年に開始され、1990年代を通じて段階的に進化していった。P2（1996年）とP3（1997年）といった試作機を経由した後、2000年に正式なASIMOが発表された。ASIMOは身長130センチメートル、体重54キログラムという人間の子どもに近い寸法をもち、40度の斜面での歩行、階段の上り下り、そして複数のステップを用いた走行動作まで達成していた。さらに2011年には改良版ASIMOが発表され、後進走行、連続ジャンプ、片足ホップといった動作能力を獲得した。

ホンダの研究アプローチの特徴は、動的歩行メカニズムと制御アルゴリズムの複雑な統合にあった。人間が日常的に行っている二足歩行は、一つの統一的なシステムとしてではなく、脚部関節の協調、腰部の動的安定化、上半身とのバランス制御による「身体的知識」の集積である。ホンダの研究者たちは、計算機シミュレーションと物理的プロトタイプの反復的な相互作用を通じて、この身体的知識を段階的に獲得していったのである。

日本の別の研究グループとして、産業技術総合研究所（AIST）と川田工業による「人型ロボット（HRP）」プロジェクトも重要である。1998年に日本の経済産業省（METI）と新エネルギー・産業技術開発機構（NEDO）によって開始されたこのプロジェクトは、ホンダのP3を基盤として、2003年にはHRP-2 を完成させた。HRP-2 は、単なる歩行能力に止まらず、両腕を用いた作業能力――特に災害対応シナリオにおける瓦礫除去といった重労働――を目指していた。その後の HRP-3（2005年）、HRP-4（2010年）、HRP-5P（2018年）の系統は、身体性の段階的な拡張、特に力覚フィードバック（force feedback）を用いた物体操作能力の強化を示している。

これら日本のヒューマノイド研究は、単なる技術的成果ではなく、一つの哲学的立場を体現している。即ち、「身体を通じて知を獲得する」という発想である。統計的機械学習が高次元データから抽象的なパターンを抽出するのに対し、ヒューマノイド研究は、具体的な物理的相互作用を通じた「身体知」を重視する。2018年にホンダが ASIMO の開発終了を発表したことは、その営利性よりは、むしろ研究方向の転換を示唆していた。同時に、HRPプロジェクトは2020年代も継続され、より実用的な作業ロボット（特に建設・災害対応領域）への展開を指向していた。

ソフトロボティクスは、この身体性の探求をさらに急進化させた領域である。従来のロボットが硬い金属フレームと関節の組み合わせであったのに対し、ソフトロボティクスは、柔軟で適応的な素材と駆動メカニズムを用いて、生物的身体により近い機械的能力を実現しようとするものである。タコの触腕、ゾウの鼻、ハエトリグサのような生物から着想を得たソフトロボットは、従来のロボットが扱えない不規則な物体の把持、複雑な環境での変形と適応、そして最も重要なこととして、人間との物理的接触時における安全性を飛躍的に高めるものである。

これら硬いロボット、ヒューマノイド、ソフトロボットの系統は、一見すると異なる技術進化系統に見えるが、共通する根底的な発問を反映している。それは、「身体とは何か」そして「身体を通じて得られる知とは何か」という問いである。

12.6　計算論的神経科学とロボット設計 —— 脳に学ぶ運動制御

ロボットの行動能力と神経科学的知見の関係が深まった一つの重要な領域が、計算論的神経科学（computational neuroscience）による運動制御の理解である。脳がいかにして身体の複雑な運動を調整し、不確実な環境への適応的行動を生成するかについての研究は、同時に、その知見をロボットの制御アーキテクチャに逆流させる循環的なプロセスを生み出してきた。

人間の運動制御は、複数の階層的なシステムによって統制されている。脊髄レベルの局所反射回路から、脳幹の基本的な姿勢制御、小脳による運動の学習と調整、そして大脳皮質による高次の計画立案まで、この階層的な組織化は、コンピュータの設計者にとって重要な設計思想をもたらす。特に、内部運動

モデル（internal models）という概念が重要である。神経科学的証拠によれば、脳は身体の動作を予測する順モデルと、目標運動から必要な制御を推定する逆モデルを内部に保有していると考えられている。これは、機械学習における「予測誤差に基づく学習」と強く対応するものである。

シュテファン・シャールらによって追究された「運動プリミティブ（motor primitives）」は、この神経科学的知見を直接的にロボット制御に実装したものである。複雑な運動軌跡を、より単純で再利用可能な基本動作の組み合わせとして表現することで、学習の計算複雑性を削減しながら、同時により柔軟な適応を実現する方法論である。

また、確率的推論の脳内実装に関する研究（ベイズ脳仮説）は、ロボットが不確実な環境下でいかに合理的な判断を行うかについて、神経生物学的な正当性を与えた。脳が観測に対する信念の事前分布と尤度の統合を行い、事後分布を更新するという基本的なベイズ過程を実行しているとすれば、ロボットもまた同一の数理的原理に従うことで、脳と同等の適応性を獲得しうるという展望が生じる。

このように、神経科学とロボティクスの相互的な影響関係は、単なる一方向の「脳から機械へ」の知識移転ではなく、むしろ両分野が共通の計算原理を追究する過程として理解されるべきである。脳科学が神経回路の機能を理解するための計算的フレームワークを必要とするのと同様に、ロボティクスが生物的身体制御の秘密を解き明かすためには、神経科学との対話が不可欠なのである。

本章を通じて、ロボティクスと身体性の知能という領域が、AI研究全体の中でいかなる位置を占めているかが明らかになるはずである。産業用ロボットの実現は、記号的知能（第6章）の限界を示し、行動ベースロボティクスはその限界を理論的に自覚させた。確率的ロボティクスはベイズ的認識論の実装であり、自動運転という社会的に重要な問題へのアプローチである。ヒューマノイドとソフトロボティクスは、身体を通じた知の再発見を示唆している。そして計算論的神経科学は、生物と機械の間に深い構造的相似性があることを示唆している。

次章（第13章）では、ゲームという限定された領域でのAIの超越を見ることになる。Deep Blueから AlphaGoへと続く系統において、純粋な探索的知能がいかなる高さに到達しうるかが問われることになるのである。しかし、その過程で見落としてはならない点がある。身体をもたないゲームAIとは異なり、現実世界でのロボットは、本章で述べた身体性の制約と可能性の中で、真の知能とは何かについて、深い問いを投げかけ続けているのである。

参考資料（本章）

- George C. Devol, Unimate 関連特許および Unimation / History of Robotics 関連史料。
- ABB 公式史料。ASEA の IRB 6 と ABB への継承関係の確認に使用。

- Rodney A. Brooks, “A Robust Layered Control System for a Mobile Robot” (1986); “Intelligence without Representation” (1991).
- Sebastian Thrun, Wolfram Burgard, Dieter Fox, Probabilistic Robotics (2005). SLAM、粒子フィルタ、FastSLAM の整理に有用。
- Sebastian Thrun et al., “Stanley: The Robot that Won the DARPA Grand Challenge” (2006).
- DARPA 公式アーカイブ。Grand Challenge と Urban Challenge の基本事実確認に使用。
- Honda 公式 ASIMO 史料。ASIMO 系譜の確認に使用。
- Tesla 公式安全資料。FSD が監督付き Level 2 運転支援であることの確認に使用。
- Waymo 公式ブログ。2025年5月時点の週25万件超の有償乗車実績の確認に使用。

第13章　ゲームAIとベンチマークとしてのゲーム

問題設定

1970年代から1990年代にかけて、AI研究が第一次・第二次の「冬」を経験する中、一つの領域だけは着実に進歩し続けていた。それはゲームAIである。チェス、チェッカー、バックギャモンといった古典的ゲームから、テレビゲームに至るまで、ゲームはAIの能力を測定し、その限界を可視化する自然な舞台として機能してきた。

本章の中心的な問いは次のとおりである。なぜゲームがAI研究の主要なテストベッドとなったのか。そして、ゲームでの成功は、より広い汎用知能へのステップとなりうるのか。あるいは、ゲームという限定的な領域での勝利は、AI研究の進展を錯覚させる幻想に過ぎないのか。

この問題を理解するためには、単なる技術史ではなく、ゲームという課題が人工知能研究にもたらした哲学的・方法論的影響を追跡する必要がある。

13.1　Deep Blue vs. カスパロフ（1997）── 探索とドメイン知識

チェスと計算の歴史

チェスをコンピュータでプレイさせるという問題は、AI研究の黎明期から存在していた。シャノンは1950年の論文「チェスをプレイするコンピュータのプログラミング」において、ゲーム木の探索戦略を体系的に分類した。彼が示した「タイプA戦略」（網羅的探索）と「タイプB戦略」（選択的探索）の区別は、後のゲームAIの理論的基盤となる。ただし、1950年当時、チェスはコンピュータの計算能力をはるかに超えた複雑性を有していた。初期のコンピュータの処理速度では、数手先さえも十分に探索できなかったのである。

その後、アルファ・ベータ法（alpha-beta　pruning）という枝刈り技術が開発され、探索効率は飛躍的に向上した。アルファ・ベータ法は、ゲーム木の探索過程で、既に見つかった解よりも悪い解へ至る枝を早期に切り落とす。この単純にして優雅なアルゴリズムにより、実質的な探索深度を劇的に増加させることができた。

20世紀後半のチェスコンピュータの発展は、ハードウェア性能とアルゴリズム的洗練の相乗作用を示す典型的な例である。1980年代から1990年代にかけて、チェス専用ハードウェアが各社で開発された。その頂点が、IBMの Deep Blue である。

Deep Blue の設計思想

Deep Blue は、単なる汎用コンピュータではなく、チェス専用に設計された並列スーパーコンピュータであった。1997年の最終版は、IBM RS/6000 SP の上に 30 個の PowerPC 604e プロセッサと、480 個のカスタム VLSI チェスチップを搭載していた。このアーキテクチャにより、システムは毎秒約2億のチェスポジションを評価することが可能になった。

Deep Blue のアプローチは、徹底的な「探索」と「ドメイン知識」の融合であった。探索の側面では、従来のアルファ・ベータ法を高度に最適化し、通常は 6 手から 8 手先を探索し、特定の局面（典型的には終盤の戦術的危機的局面）ではより深く探索した。

探索と同じくらい重要だったのが、評価関数（evaluation function）の設計である。Deep Blue には、約70万局規模のグランドマスター対局データやオープニングブックからの知識が組み込まれていた。この知識抽出は、単なる統計的パターン認識ではなく、チェスの戦術的・戦略的原則をコンピュータ可能な形に翻訳する作業であった。評価関数は、単なるマテリアル（駒の価値）を計算するのではなく、駒の位置、ポーン構造、キング安全性、駒の中央集約度など、数十のチェス的特徴を組み込んでいた。

1996 年の第一次対局では、カスパロフは Deep Blue に勝利した。しかし IBM チームは、ハードウェアを高速化し、評価関数をさらに洗練させた。1997 年の第二次対局では、Deep Blue は 2勝1敗3引き分け、合計 3.5 対 2.5 でカスパロフを打ち負かした。

歴史的意義と制限

Deep Blue vs. カスパロフの勝利は、単なるコンピュータチェスの成功ではなく、AI史における一つのターニングポイントであった。この勝利は、AI研究が「AIの冬」（第8章）から確実に復興しつつあることを世界に示した。ただし同時に、Deep Blue の成功の本質を正しく理解することは、ゲームAIの性質と限界を認識する上で重要である。

Deep Blue は、チェスというドメイン固有の課題に対して、以下の三つの基本的戦略を駆使していた。

第一に、**網羅的探索**である。Deep Blue は、人間の直感的な選別ではなく、可能な限り多くの局面を計算で評価した。人間のグランドマスターは、数十の候補手を直感的に絞り込み、その中から最善手を探索する。Deep Blue は逆に、探索空間を徹底的に調べ尽くすことで、人間の直感を補った。

第二に、**領域特化的知識**である。チェスの戦術的・戦略的原則をコンピュータ化した評価関数は、一般的な学習メカニズムではなく、人間の専門家（チェスマスター）が事前に設計したものであった。この知識は、データベースから自動的に抽出されたというより、チェス理論の積み重ねが構成化されたものである。

第三に、**カスタムハードウェア**である。チェス専用チップの設計により、探索の並列化が最適化された。汎用プロセッサではなく、チェス評価に特化したハードウェアがあってこそ、この処理速度が達成されたのである。

これらの三つの要素は、Deep Blue の強さの源泉であると同時に、その限界をも定義していた。Deep Blue は、チェスという非常に構造化された、ルールが完全に定義された環境では圧倒的であった。しかし、確率分岐を含むゲームや、状態空間が非常に大きく、ドメイン知識の事前設計が困難なゲームには、この方法論をそのまま適用することはできなかった。

13.2　バックギャモンと TD-Gammon ―― 強化学習の実証

問題の難しさ：確率性と評価学習

チェスと異なり、バックギャモンは二つの重大な特徴を持つ。第一に、**確率的要素**がある。サイコロの出目によって、次の選択肢が大きく変わる。第二に、**評価関数の設計が難しい**。盤面は両者に完全に見えているので不完全情報ゲームではないが、出目の不確実性が絡むため、探索だけで最善手を安定に見積もることが難しい。

この二つの特徴により、チェスの Deep Blue で用いたアプローチ――網羅的探索と事前設計された評価関数――の適用は困難になる。バックギャモンのゲーム木は、チェスよりも大きい（チェスの典型的な分岐係数は約 35、バックギャモンは約 21 だが、確率的分岐の複雑さはより高い）。さらに重要な点として、人間の専門家でさえ、バックギャモンの最適戦略について完全な合意がなかった。チェスの場合、評価関数を設計するためのドメイン知識が豊富に存在したが、バックギャモンではそうではなかったのである。

Gerald Tesauro と TD-Gammon

この問題を解決したのが、IBM トーマス・J・ワトソン研究所の Gerald Tesauro である。Tesauro は 1992 年、**強化学習**の理論的枠組みを用いて、バックギャモンAI「TD-Gammon」を開発した。

TD-Gammon（Temporal Difference Gammon）の核心は、**評価関数を自動的に学習する**というアプローチにある。具体的には、以下のような仕組みであった。

まず、多層パーセプトロン（multilayer perceptron）という単純なニューラルネットワークが、盤面を入力として受け取り、その盤面から勝利する確率を出力する。初期状態では、このネットワークの重みはランダムに初期化されていた。

次に、TD-Gammon は **自己対局**によって学習した。プログラム自身と何百万回も対局を行い、各ゲームで得られた報酬信号（勝利または敗北）に基づいて、ニューラルネットワークの重みを調整する。この学習則が **Temporal Difference (TD) 学習**である。

TD 学習の創始者はリチャード・サットン（Richard Sutton）であり、1988 年に発表された論文が基礎となっていた。サットン＝バルト（Sutton-Barto）の強化学習理論は、報酬信号と価値関数推定を結びつけ、モデルのない学習（model-free learning）を可能にした。Tesauro はこの理論を、バックギャモンという複雑なゲームに初めて成功裏に適用したのである。

パフォーマンスと意義

1990年代前半、自己対局を通じて訓練された TD-Gammon（特にバージョン 2.1 以降）は、世界トップクラスのバックギャモン選手と同等かそれに近い実力を発揮した。これは、人間レベルのプレイを、明示的な手工知識の大量投入なしに実現した初めての印象的事例であった。

より重要なのは、TD-Gammon が人間が考えもしなかったプレイ戦略を発見したことである。その戦術は、当時のバックギャモン理論に新たな洞察をもたらし、人間のプレイヤー自身がそうした戦略を学んで採用するようになったほどである。これは、機械学習が単に人間を模倣するのではなく、人間の専門知識を超える戦略を発見しうることを示す初めての具体例となった。

Deep Blue との対比

Deep Blue と TD-Gammon を比較することは、AI 史の根本的な対立――**探索（search）と学習（learning）**の対立――を照らし出す。

Deep Blue は、与えられたタスク（チェス）に対して、人間が設計した知識（評価関数）と、高速な計算（大規模な探索）を組み合わせた。この方法は、ドメイン知識が豊富で、状態評価が明確なタスクでは極めて有効である。

TD-Gammon は対照的に、人間の専門知識に頼らず、大量の自己対局経験から評価関数を自動的に習得した。この方法は、知識工学（knowledge engineering）の負担を大きく軽減し、新しい領域への応用を容易にする。ただし、計算コストは Deep Blue 以上に膨大であり、数百万ゲームの自己対局が必要であった。

この二つのアプローチの違いは、単なる技術的選択ではなく、AI 研究における根本的なパラダイムの違いを反映していた。知識志向のアプローチ（Deep Blue）か、経験志向のアプローチ（TD-Gammon）か。あるいはその融合か。この問いは、後の深層強化学習（深層 DQN、AlphaGo など、第 15-16 章）へと引き継がれることになる。

13.3　Jeopardy! と IBM Watson（2011）── 質問応答の到達点

質問応答タスクの性質

Jeopardy!　は、一見するとゲームであるが、そのメカニクスは実は言語理解の困難さを象徴している。通常のゲーム AI とは異なり、Jeopardy! では：

- 対象者が「答え」（事実）を述べ、プレイヤーが「質問」を形成する（通常の質問応答と逆）
- 自然言語で提示される手がかりが、明示的なルール体系に従わない複雑な言語現象を含む
- 「トリビア」的知識、言葉遊び、文化的背景を要する

つまり、Jeopardy! は、**知識の深さ**と**自然言語理解の複雑さ**を同時に要求する課題である。

IBM Watson のシステム構成

2011 年 2 月、IBM の David Ferrucci 率いるチームが開発した Watson は、Jeopardy! の歴代最強プレイヤー（Brad Rutter、Ken Jennings）を破った。このシステムは、従来の「検索エンジン」的アプローチとは根本的に異なっていた。

Watson　の主要な特徴は、**複数の言語分析アルゴリズムの並列実行**にあった。システムは、以下のようなプロセスを同時多発的に実行していた：

1. **質問解析（Question Analysis）**：与えられた質問文から、その種類（人物、場所、事件など）を判定し、どのような情報を探すべきかを推定。
2. **情報検索（Passage Retrieval）**：広大な知識ベース（2 億ページ以上の構造化・非構造化データ）から、関連する文書・段落を検索。
3. **候補応答生成（Answer Candidate Generation）**：検索された文書群から、複数の候補応答を抽出。
4. **スコアリング（Scoring and Ranking）**：各候補応答に対して、数百の異なる特徴量（共起統計、言語的尤度、特定の信頼度メトリクス）に基づいてスコアを計算し、最も可能性の高い応答を特定。

この全体的なアーキテクチャで特に革新的だったのは、**新しいアルゴリズムの創造**ではなく、**既知のアルゴリズムの組み合わせと調整**にあった。自然言語処理、情報検索、機械学習のサブフィールドで既に開発されていた技術を、大規模かつ実時間的に統合したのである。

Watson は、10 個の出力ラックに 90 個のサーバー、合計 2880 個のプロセッサコアを含む、ルームサイズのコンピュータシステムであった。処理の並列化により、約 3 秒の制限時間内に数千の候補応答を評価することが可能になった。

パフォーマンスと限界

Jeopardy! での Watson の勝利は、AI 研究に重要な教訓をもたらした。第一に、**開放領域の知識課題（open-domain task）**においても、適切なシステム設計と大規模な計算資源があれば、人間レベルのパフォーマンスを達成可能なこと。第二に、言語理解は単一のアルゴリズムではなく、複数の異なる角度からのアプローチの統合によってこそ可能になることを、実証的に示したこと。

しかし同時に、Watson の限界も明白であった。

まず、Watson は **構造化知識への依存**が大きかった。Jeopardy! の質問は、最終的には既知の事実についての質問である。ウィキペディアや百科事典的な知識ベースが存在すれば対応可能な問題が多い。

次に、Watson は**複雑な推論**ができなかった。多段階の論理的推論、反事実的推論、価値判断を含む問題には対応できなかった。

第三に、Watson は**文脈（context）**の理解が浅かった。Jeopardy! では各質問は相対的に独立しているが、会話や文学的理解のような文脈依存的な課題では、Watson のアプローチは機能しにくかった。

IBM Watson その後

興味深くは、Jeopardy! 勝利後、IBM は Watson を医療診断、法律分析、企業助言といった領域に応用しようとした。ただし、これらの領域での成果は、Jeopardy! での成功ほど顕著ではなかった。その理由は、これらのドメインでは「正解が一つでない」「問題を完全に構造化できない」「領域固有の推論が複雑である」という点にあった。

Watson は、**境界のよく定義された、知識集約的な課題**では優れていたが、**複雑な推論と不確実性**を要する現実的な問題への汎化は難しかったのである。これは、エキスパートシステムが 1980 年代に経験した限界（第 7 章）と本質的に同じ制約であった。

13.4　ゲーム AI における評価手法と汎用知能の議論

ゲームが適切なベンチマークである理由

1990 年代から 2000 年代にかけて、AI 研究コミュニティは、ゲームを系統的な評価のためのテストベッドとして採用し始めた。この選択の背景には、複数の理由がある。

第一に、明確な勝敗基準。ゲームは、勝利・敗北といった客観的で定量化可能な結果を提供する。自然言語生成や画像認識と異なり、「正答率」がどうしても主観的になりようがない。

第二に、複雑性の段階的調整。チェス、チェッカー、Go といった古典的ゲームから、リアルタイムストラテジー、Atari 2600 といった複雑性の異なるゲームへと、段階的に難度を上げることができる。

第三に、人間のパフォーマンスとの直接比較。ゲームでは、AI のパフォーマンスを人間のプレイヤーと直接比較でき、その優劣が明白である。この比較可能性が、AI の社会的インパクトを可視化するのに重要であった。

第四に、汎用性の可能性。ゲームプレイに必要とされる能力――時間制約下での意思決定、不完全情報の処理、長期戦略立案、対抗者の行動予測――は、現実世界の多くの課題にも共通している。

しかし、ゲームの限界も顕著である

ゲームベンチマークが有用である一方で、その限界についても論じる必要がある。

第一に、単一目標の局所最適化。ゲームでの目標（勝つこと）は明確であり、単一である。しかし、現実的な多くの問題は、複数の（往々にして矛盾する）目標を同時に追求する必要がある。

第二に、完全なルール体系。ゲームはすべてのルールが明示的に定義され、ゲーム外のノイズや例外がない。しかし、現実の問題――交渉、医療診断、科学的発見――は、不完全で曖昧なルール体系の中で展開される。

第三に、身体性の欠如。ゲーム AI は、典型的には知覚と意思決定に焦点が当たり、行動の物理的実行やそのフィードバックは考慮されない。しかし、現実的なロボティクス（第 12 章）やヒューマン・エージェント協働では、身体性が本質的である。

第四に、静的環境。ゲームの世界は、プレイ中に根本的には変化しない。ルールセット自体、プレイ環境の構成も、固定されている。しかし、現実世界は開放的で動的である。AI システムが直面する環境は、常に新しい課題や異なるルール体系を提示する。

ゲームから学習の論争

1990 年代から 2000 年代初頭にかけて、AI 研究界では以下のような論争が繰り広げられていた。ゲーム AI の成功は、**本当に汎用知能への進歩を示しているのか、それとも、限定的なドメインでの最適化に過ぎないのか**。

楽観的な研究者たちは、ゲーム AI を、より広い知能的課題への基礎として位置づけた。彼らは、ゲームプレイに必要とされる能力――探索、評価、戦略的計画――が、他の領域でも本質的に必要とされていると論じた。

悲観的な研究者たちは、ゲーム AI を、領域特化的な工学的成功として見なした。彼らの見方では、Deep Blue はチェスの知識工学の産物であり、TD-Gammon は強化学習の証明概念に過ぎず、汎用知能とは関係がない、というものであった。

この論争は、単なる学術的な議論ではなく、AI 研究の資金配分、研究機関の戦略、研究者のキャリアパスに直接的な影響を与えていた。

13.5　ゲームからベンチマークへ ―― Atari・StarCraft の前史

Arcade Learning Environment と Atari 2600

2012 年、ゲーム AI のベンチマーク化における転換点が訪れた。Arizona State University と University of Michigan の研究者たちが、**Arcade Learning Environment（ALE）**を提案した。これは、Atari 2600 ゲーム機の 57 タイトルを、統一されたインターフェースを通じて AI エージェントに提供するプラットフォームであった。

Atari 2600 の選択は戦略的であった。これらのゲームは：

- **画像ベースの入力**：ゲーム画面のピクセル情報のみが与えられ、ゲームのルールは直接的には提供されない。したがって、画像認識と高レベルな戦略立案の両方が必要になる。
- **多様な複雑性**：Pong のような単純なゲームから Montezuma's Revenge のような複雑なゲームまで、難度の範囲が広い。
- **汎用性の証明の場**：異なるゲーム間での汎化能力が、AI システムの汎用性をテストする。

ALE は、単にゲームを提供するだけではなく、以下の点で従来のゲーム AI ベンチマークと異なっていた。

1. **学習可能性**：AI エージェントが、ゲームのルールを事前に与えられず、ゲームプレイから学習する。
2. **汎用エージェント設計**：単一ゲーム向けの特化設計ではなく、複数ゲーム間での汎用性を追求する。
3. **アーカイブ化**：ゲーム画像の処理（視覚認識）が必須となり、当時台頭していた深層ニューラルネットワークのテストベッドとして機能する。

Deep Q-Networks と深層強化学習の革新

ALE の登場から 1 年後の 2013 年、DeepMind（当時まだ Google に買収されていない独立した企業）の Volodymyr Mnih らは、**Deep Q-Networks（DQN）**を発表した。この研究は、強化学習と深層ニューラルネットワークを初めて成功裏に統合したものであり、AI 史における転換点を画した。

DQN は、以下のような革新を体現していた：

- **エンド・ツー・エンド学習**：raw pixels（画像のピクセル）を入力として、直接的に action values（各行動の価値）を出力する深層ニューラルネットワークを訓練。
- **経験再生（Experience Replay）**：過去の経験を蓄積し、ランダムに抽出して訓練。これにより、データの相関を低減し、学習を安定化。

- **Target Networks**：学習の安定性を向上させるため、異なるニューラルネットワーク複製を用いて TD ターゲットを計算。

DQN は、最初の論文で 7 つの Atari ゲームで訓練され、6 つのゲームでプロフェッショナルなヒューマンテスターを上回るパフォーマンスを達成した。その後の拡張版では、49　個のゲーム中で人間レベルのパフォーマンスに達したと報告されている。

この成功の意義は、単なるゲーム AI の技術的進展ではなかった。DQN は、**視覚認識と戦略的意思決定を統一的に扱う**深層強化学習モデルの実現可能性を、具体的に示したのである。これが後の　AlphaGo（第 16 章）への道を開いた。

StarCraft と複雑性の拡張

ALE が Atari ゲーム（典型的には単一エージェント対環境、あるいは単純な敵キャラクター）に焦点を当てていたのに対し、2010　年代後半には、より複雑なゲーム環境がベンチマークとして採用され始めた。その代表が **StarCraft II** である。

StarCraft II は、以下の点で Atari ゲームより本質的に複雑である：

- **実時間戦略**：ターンベースではなく、リアルタイムで意思決定が必要。
- **不完全情報**：対戦相手の全体像が見えず、戦術的な隠蔽（fog of war）が存在。
- **膨大なアクション空間**：各瞬間に利用可能な行動が非常に多い（推定 10^26 アクション）。
- **階層的計画**：低レベルの制御（個々のユニットの操作）と高レベルの戦略立案（経済管理、部隊編成）の両方を要する。

DeepMind は、こうした複雑性に対応するため、AlphaStar というシステムを開発し、2019 年にはプロフェッショナルプレイヤーを打ち負かしたと発表された。AlphaStar　は、**深層ニューラルネットワーク**、**人間プレイヤーのリプレイからの模倣学習**、そして**リーグ形式の自己対戦強化学習**を組み合わせた複層的なアプローチを採用していた（詳細は第 16 章で述べる）。

13.6　ゲーム AI の成果と盲点

この章の軌跡が示すもの

1950 年代のシャノンの理論的枠組みから、1997 年の Deep Blue、1993 年の TD-Gammon、2011 年の IBM Watson、2013 年の DQN、2019 年の AlphaStar に至る歴史を振り返ると、以下のパターンが浮かび上がる：

1. **技術的蓄積**：探索アルゴリズム（アルファ・ベータ法）から強化学習（TD 学習）、そして深層ニューラルネットワークへと、基本的な技術が段階的に進化してきた。

2. **ハードウェア依存性**：Deep Blue のチェスチップから、DQN の GPU 計算まで、計算能力の向上がブレイクスルーを可能にしてきた。

3. **知識獲得の進化**：事前に設計された評価関数（Deep Blue）から、自動的に学習された価値関数（TD-Gammon、DQN）へのシフト。これは、領域固有の知識工学への依存を減らす方向への進展である。

4. **複雑性の段階的上昇**：チェス（決定木の組み合わせ爆発）から、バックギャモン（確率性と評価学習）、Jeopardy!（自然言語理解）、Atari（高次元入力と学習）、StarCraft（リアルタイム、複雑なマルチエージェント環境）へと、段階的に複雑性が増している。

しかし同時に、以下の根本的な限界も見えてくる：

ゲーム AI の本質的限界

第一に、ゲーム・スコープの局所性。ゲーム AI の優れた成果は、必ずしも汎用知能への接近を意味しない。AlphaStar が StarCraft II で Grandmaster 達成したとしても、別のリアルタイムストラテジーゲームにそのアルゴリズムを適用できるかどうかは全く別問題である（実際、多くの場合、相応の再訓練が必要である）。

第二に、報酬信号の明示性。ゲーム AI は、明確で定量化可能な報酬信号（勝点、スコア）を利用できる。しかし、現実の多くの目標——医療診断の最適化、科学的発見、社会的福祉——では、何が最適化すべき「報酬」であるかが根本的に不確定である。

第三に、ルール外部性の欠如。ゲームのルールは所与であり、変更されない。しかし、現実的な問題では、しばしばルール自体を問い直す必要がある。

第四に、意図性と目的論の不在。ゲームAIは、勝つことを目指して設計され訓練されるが、「なぜゲームをするのか」という問いには無関心である。人間の目標指向的な行動は、局所的な報酬最大化だけでなく、より広い価値体系に埋め込まれている。

ゲームからの知見の汎化

しかし、完全に悲観的である必要もない。ゲーム AI から得られた知見の中には、より広い適用可能性を持つものがある。

まず、**強化学習の実現可能性**が実証された。TD-Gammon と DQN は、人間が事前に知識を設計することなく、経験から価値関数を学習できることを示した。これは、新しい領域への AI 適用における「知識工学の負担」を大幅に軽減する。

次に、**深層神経ネットワークの有効性**が確認された。DQN は、高次元の入力（ピクセル）から、直接的に意思決定信号を抽出できることを示した。これは、コンピュータビジョンと意思決定の統合を可能にし、自動運転（第 12 章）のような現実応用への基礎を提供した。

第三に、**複雑なマルチエージェント環境での学習**の基礎が確立された。AlphaStar　は、対抗的な環境で、自己対局や模倣学習を組み合わせることで、複雑な戦略を獲得できることを示した。

第IV部への橋渡し

ゲーム AI の軌跡は、AI 史における**統計的転回から深層学習革命への移行**を象徴している。

第 III 部では、統計的機械学習、強化学習の理論的基盤、そして自然言語処理・コンピュータビジョンの進展を論じた。これらは、ルールベース的なアプローチ（記号　AI）から、データ駆動的なアプローチへのパラダイムシフトを表現していた。

ゲーム AI は、このパラダイムシフトの**実験場**として機能した。Deep Blue は記号 AI 的アプローチの最後の大きな勝利であり、TD-Gammon は機械学習の可能性を示し、DQN は深層学習と強化学習の融合を示した。

第　IV　部で詳述する深層学習革命は、これらの洞察——探索による解法、価値関数学習、ニューラルネットワーク表現——を大規模化し、より複雑な問題領域へ適用していく過程である。AlphaGo は、ゲーム AI における深層学習と従来の探索手法（モンテカルロ木探索）の統合の頂点であり、同時に、次の段階への扉を開く。それは、科学的発見（AlphaFold）、言語理解（大規模言語モデル）といった、人工知能が本当の意味で「汎用的」な問題解決へと進むプロセスである。

> 本章では、ゲーム AI の進化を、技術的側面（探索、学習、評価関数）と方法論的側面（知識工学から自動学習へ）の両面から検証した。Deep Blue の専門知識と網羅的探索、TD-Gammon の強化学習的自動学習、IBM Watson の知識統合、そして DQN と AlphaStar の深層学習的アプローチは、AI 研究における異なるパラダイムを代表しており、その競合と融合の軌跡がそのまま、AI 史全体の主要な流れを構成している。ゲームの領域を超えた汎用知能への到達が可能か否か、それはこれからの章で展開される深層学習革命の成否にかかっている。

参考資料（本章）

- Claude E. Shannon, “Programming a Computer for Playing Chess” (1950).
- IBM, Deep Blue 関連公式史料。1997年版の構成と秒間評価局面数の確認に使用。
- Gerald Tesauro, TD-Gammon 関連論文群（1992, 1995）。
- IBM, Watson / Jeopardy! 関連公式資料。システム構成と 2011年対戦の確認に使用。
- Bellemare, M. G., et al. “The Arcade Learning Environment” (2012).

- Mnih, V., et al. “Playing Atari with Deep Reinforcement Learning” (2013); “Human-level control through deep reinforcement learning” (2015).
- Vinyals, O., et al. “Grandmaster level in StarCraft II using multi-agent reinforcement learning” (Nature, 2019).

第IV部

深層学習革命（2010年代前半〜中盤）

第14章〜第16章

第14章　深層学習の夜明け

導入

2006年から2016年にかけて、人工知能研究は第二次の冬からの目覚めを経験した。第8章で述べたように、1980年代のエキスパートシステムの衰退とLISPマシン市場の崩壊の後、ニューラルネットワーク研究は細々と続いていた。しかし21世紀初頭、三つの条件が揃うことで、この分野は劇的な復興を遂行することになる：第一に、ジェフリー・ヒントンらによる深層学習の訓練手法の革新である。第二に、画像認識ベンチマークとしてのImageNetの構築である。第三に、GPU計算の民主化とビッグデータの可用性である。これらの要素が相互に作用し、「深層学習革命」と呼ばれるパラダイムシフトをもたらした。

本章では、この変化の軌跡を追跡する。ヒントンの貪欲層別事前学習からAlexNetの衝撃、そしてCNNアーキテクチャの急速な進化まで、技術的イノベーションと計算基盤の変化を同時に描く必要がある。また、dropout、バッチ正規化といった規則化技法の発展、そしてGPU、TPU、AI専用チップへと至る計算基盤の系譜も記述する必要がある。この章の終わりには、深層学習がなぜそれほど急速に実用化され、産業への浸透が加速したのかが明らかになるはずである。

14.1　ヒントンとディープビリーフネットワーク（2006）

ニューラルネットワークの「第三波」の起点は、2006年7月に遡る。ジェフリー・ヒントン、サイモン・オシンデロ、イー・ウェイ・テーが『Neural Computation』に発表した論文「A Fast Learning Algorithm for Deep Belief Nets」である。この論文が提示した貪欲層別事前学習（greedy layer-wise pretraining）は、深層ニューラルネットワークの訓練における根本的な困難を部分的に解決するものであった。

その困難とは何か。第3章から第8章を通じて繰り返し現れた「消滅する勾配問題（vanishing gradient problem）」である。1991年、ザイプ・ホッホライター（Sepp Hochreiter）がその修士論文で形式化したこの問題は、4〜5層を超えるニューラルネットワークをシグモイド関数などの飽和活性化関数で訓練することを事実上不可能にしていた。逆伝播により計算された勾配が、層を深くするにつれて指数関数的に縮小し、最終的には深層の重みの更新信号が無視できるほど小さくなってしまう。この障壁が、1980年代のニューラルネット研究の停滞をもたらした重要な要因の一つであった。

ヒントンのアプローチは、一見して逆説的である。深い網を一度に訓練するのではなく、制限されたボルツマンマシン（Restricted Boltzmann Machine, RBM）と呼ばれる二部グラフ構造をなすニューラルネットワークを、層ごとに独立して、教師なしで事前訓練し、その後にラベル付きデータを用いた教師あり調整（fine-tuning）を行うというものであった。この「貪欲」戦略は、全層を同時に最適化する代わりに、局所的な最適化を順次実行することで、全体的な収束を促進するという発想に基づいていた。

RBMの学習は、ギブスサンプリングと対照度発散（contrastive divergence）という手法を用いて実行可能であり、各層における学習は無教師である。すなわち、入力データのみが必要であり、人間によるラベル付けは不要であった。データセットの大部分がラベルなしであるという現実において、この点の重要性を過小評価することはできない。

ヒントンらが提示した結果は、当時としては印象的であった。手書き数字認識のベンチマーク（MNIST）において、深層ビリーフネットワークは既存の教師あり学習アルゴリズムを上回る性能を示した。しかし、この方法にも限界があった。層別事前訓練は計算上の負担が大きく、また層数が増えるにつれて利得が減少することが後に判明する。さらに重要なことに、この手法は教師なし学習の有効性を仮定していたが、実際には、十分な教師ありデータが利用可能な場合には必ずしも必要ではないことが明らかになっていった。

それでもなお、ヒントンの2006年の論文は決定的な心理的・知的な転換点となった。「深い網は訓練可能である」という信念を研究コミュニティに再確立し、その直後の急速な発展の基礎を準備したのである。歴史的には、この論文はパラダイムの転換を告知するシグナルであり、次に続く技術的ブレークスルーの先駆けであった。

14.2　GPU計算とビッグデータの収束

深層学習が実用的な技術として急速に発展した背景には、ハードウェアの進化とデータの可用性の劇的な変化がある。

GPU（Graphics Processing Unit）は、当初はグラフィックス処理に特化した並列計算装置であり、AIの観点からは周辺的な技術であった。しかし2007年、NVIDIAがCUDA（Compute Unified Device Architecture）を公開した際、状況は変わった。CUDAは、NVIDIA製GPUを一般的な並列計算に利用することを可能にするプログラミングインターフェースであり、当初は科学計算向けであった。しかし2008年頃から、アンドリュー・ング（Andrew Ng）らの研究者がGPUを深層ニューラルネットワークの訓練に用いることの利点に気づき始める。GPUの高い並列性は、行列演算の大規模な並列実行、特に畳み込み演算の加速に極めて適合していたのである。

一方、データの側面でも同時並行的に革命が起きていた。2009年、スタンフォード大学の李飛飛（Fei-Fei Li）は「ImageNet」プロジェクトを立ち上げた。このプロジェクトは、WordNetの8万個の同義語集合のそれぞれについて、500〜1000枚の高解像度画像を収集し、アマゾン・メカニカル・ターク（Amazon Mechanical Turk）を用いてラベル付けするというものであった。2009年の時点では3.2百万枚の画像が組織化されていた。ImageNet Large Scale Visual Recognition Challenge（ILSVRC）は2010年に開始され、毎年のコンテストを通じて、視覚認識タスクにおける急速な進歩の源泉となることになる。

この二つの条件――計算能力とデータ――の出現は、深層学習の発展に必須であった。深いニューラルネットワークは、多くのパラメータ（AlexNetで6,000万以上）を持つ。過学習を避けながらこれを訓練するには、大規模で多様なデータセットが必要である。同時に、そうした大規模データセットで何千回もの逐次演算を行うには、CPUベースの計算では実用的な時間内に完了しない。GPUの並列性と、ImageNetのような大規模ベンチマークの存在は、相互的な前提条件として機能した。

14.3　AlexNet —— ImageNet 2012の衝撃

2012年9月30日、トロント大学のアレックス・クリズェフスキー、イリア・スツスケヴァー、ジェフリー・ヒントンからなるチーム「SuperVision」は、ImageNet Large Scale Visual Recognition Challenge 2012に、深い畳み込みニューラルネットワーク（CNN）を投入した。その結果は劇的であった。top-5エラー率（正答を5位以内に含まない確率）は、前年の最良手法である26.2%から15.3%へと急落したのである。これはコンピュータビジョンの歴史において、単一の方法による一度の改善としては記録的な進歩であり、当時のコミュニティに衝撃を与えた。

AlexNetの構成は、当時の標準的なCNNからいくつかの点で革新的であった。一つは深さである。5層の畳み込み層と2層の全結合隠れ層、そして1層の出力層――合計8層のネットワークは、当時としてはきわめて深かった。第二の革新は、活性化関数としてReLU（Rectified Linear Unit）を採用した点である。ReLU は $f(x) = \max(0, x)$ という単純な関数であり、シグモイド関数やタンハイパーボリック関数のような飽和関数ではない。この選択により、逆伝播による勾配計算の安定性が向上し、学習が著しく加速された。クリズェフスキーらの論文では、ReLUを用いたAlexNetが、同じアーキテクチャでシグモイド活性化関数を用いた場合に比べて約6倍高速に訓練されたことが報告されている。

アーキテクチャの詳細は以下の通りである。入力層は227×227×3のカラー画像を受け取る。第1層は11×11×3のカーネル96個を、ストライド4で適用し、応答正規化（response normalization）と3×3の最大値プーリングを行う。第2層は256個の5×5カーネルを適用し、再び正規化とプーリングを行う。第3層は384個の3×3カーネル、第4層は384個の3×3カーネル、第5層は256個の3×3カーネルを順次適用する。畳み込み層の後に、4096ニューロンの全結合隠れ層が2層続き、最後に1000クラスの出力層がある。全体で6,200万個のパラメータをもつこのネットワークは、NVIDIA GTX 580 GPU 2個を用いて、6日間かけて訓練されたと報告されている。

AlexNetの成功は、技術的な革新だけでなく、計算資源の民主化に支えられていた。クリズェフスキーは当時まだ学生であり、親の家の寝室でGPUによる訓練を行ったという逸話は、深層学習の成功がエリート的な大規模計算リソースに限定されるものではなく、意欲ある研究者による実験的なアプローチでも達成可能であったことを象徴している。また、CUDA技術の進化により、複数のGPUでの分散訓練も実用的になっていた。

AlexNetの成功は、多くの研究者にコンピュータビジョンの明確な転換点として受け止められた。この評価は単なる修辞ではなく、後続の研究の爆発的な増加によって裏づけられる。AlexNetの論文は、その後10年以上にわたって学術文献における最も引用される論文の一つとなり、深層学習の象徴的な到達点として扱われるようになった。

14.4　畳み込みニューラルネットワーク（CNN）の進化 ── VGG、GoogLeNet、ResNet

AlexNetの成功は、その直後に一連の建築的革新をもたらした。2013年から2015年にかけて、ImageNetチャレンジの各年の優勝者による新しいアーキテクチャが、急速に深層化・複雑化していった。この段階は、コンピュータビジョンにおける「アーキテクチャ設計の黄金期」と呼ぶことができる。

VGGネットワーク（Visual　Geometry　Group、オックスフォード大学）は2014年に提案された。カレン・シモニアン（Karen　Simonyan）とアンドリュー・ジッサーマン（Andrew　Zisserman）は、単純な3×3の小さなカーネルのみを用い、これを何層にも積み重ねることで深さを獲得するというシンプルかつ体系的なアプローチを採用した。VGGの最深版（VGG16/VGG19）は16〜19層の畳み込み層をもち、ImageNetベンチマークでtop-5エラー率7.3%を達成した。VGGは　AlexNetに比べて訓練時間が長く、パラメータ数が多い（500MB以上）という欠点を持つが、その単純さと一般性は、後の転移学習（transfer learning）の基盤として広く活用されることになる。

同じ2014年、Googleの研究チームはGoogLeNet（後にInception　v1と呼ばれる）を発表した。このアーキテクチャの革新は、「Inception　　モジュール」と呼ばれるサブユニットの導入である。各Inceptionモジュール内では、異なるサイズのカーネル（1×1、3×3、5×5）による畳み込みが並列に実行され、それらの出力が連結される。この並列構造により、ネットワークはより効率的に多スケールの特徴を抽出できた。GoogLeNetはImageNet 2014を制覇し、top-5エラー率6.7%を達成した。さらに特筆すべき点は、パラメータ数がAlexNetの1/12未満であり、計算効率が大幅に改善されたことである。

2015年、マイクロソフト・リサーチ・アジア（MSRA）のチームがResNet（Deep Residual Learning for Image　Recognition）を発表し、深層学習の理論的・実践的な理解に革命をもたらした。開発者はカイミング・へー（Kaiming　　He）、シャオキュン・ジャン（Xiangyu　　Zhang）、シャオティング・レン（Shaoqing Ren）、ジアン・スン（Jian Sun）である。

ResNetの革新は、「残差接続（residual connection）」或いは「スキップ接続（skip connection）」と呼ばれる仕組みの導入にある。従来のCNNでは、層の出力 y は y = F(x) のように、入力 x から目的の関数 F を直接学習することを目指していた。これに対してResNetは、y = F(x) + x という形で、本来学習すべき関数を F(x) = y - x という「残差」として定式化し、この残差をニューラルネットワークに学習させることを提案した。このアプローチにより、深度が増しても勾配の消滅が起きにくくなり、152層に達するきわめて深いネットワークの訓練が可能になった。

この革新がもたらした実践的な成果は劇的であった。ResNetはImageNet 2015で優勝し、アンサンブル構成のtop-5エラー率は3.57%に低下した。これはヒューマンレベルの性能（推定約5%）に初めて接近するものであり、コンピュータビジョン研究における象徴的なマイルストーンとなった。さらに理論的には、残差接続はなぜ有効なのかを理解する上での重要な洞察をもたらした。

残差接続が革新的である根本的理由は、以下のように分析される：第一に、恒等写像（identity mapping）を含むことにより、勾配が最悪の場合でも恒等写像を通じて流れることを保証し、深層への勾配伝播を確実にする。第二に、ネットワークが「何もしない」（すなわち F(x) = 0 に相当する状態）という局所的な最適解に容易に到達可能であり、その周辺での最適化を進めることが理論的に容易である。第三に、深い層では微細な調整に注力でき、浅い層では基本的な特徴抽出に注力できるというように、層ごとの役割分担が自然に生じるという仮説である。

ResNetの後、2016年には改良版のInception v3（バージョン3）が発表され、バッチ正規化と因数分解畳み込み（factorized convolution）により一層の効率化が図られた。しかし、CNNアーキテクチャの進化としては、ResNetによる残差接続の導入が、少なくとも2017年のTransformer登場まで、最も根本的な革新であった。

14.5　ドロップアウト、バッチ正規化 —— 深層学習を支える技法

AlexNetのアーキテクチャの成功と並行して、深層ニューラルネットワークの訓練を安定化し、正則化する技法の発展も、深層学習の実用化に不可欠であった。

ドロップアウト（Dropout）は、ジェフリー・ヒントンと同僚たちにより2012年に提案された。基本的な考え方は単純であるが、その効果は大きい。訓練時に、ニューロンをランダムに一定の確率（典型的には0.5）で無効化（ドロップアウト）し、その後の出力を調整する。このアプローチは、以下の直感的な理由に基づいている：ネットワークが訓練データへの過度な適合を避けるため、異なるニューロンの組み合わせによるアンサンブルの学習として機能する。すなわち、各訓練段階で異なる「部分ネットワーク」を訓練することで、相互に補完的な特徴表現の学習が促進される。AlexNetでもドロップアウトが用いられ、過学習の防止に重要な役割を果たしたと報告されている。

その理論的な正当化も後に与えられた。ドロップアウトの効果は、複数の部分ネットワークのアンサンブル平均と解釈できることが示された。また、訓練時のドロップアウトが、テスト時のネットワーク全体を用いた予測と（スケーリング調整を通じて）一貫性をもつことも理論的に確立された。ドロップアウトは現在、深層学習において最も広く用いられる正則化技法の一つであり、ReLUと並んで、AlexNetの成功を可能にした重要な技術的要素である。

バッチ正規化（Batch Normalization）は、セルゲイ・イオフェ（Sergey Ioffe）とクリスティアン・セゲディ（Christian Szegedy）により2015年に提案された。この手法は、各層への入力の分布が訓練中に変化する現象を「内部共変量シフト（internal covariate shift）」と名づけ、その是正を提案するものである。

具体的には、バッチ正規化は各訓練ミニバッチについて、各層の活性化を正規化（平均0、分散1に標準化）し、その後、学習可能なスケーリングとシフトパラメータを適用する。この操作により、以下の複数の利点が生じる：第一に、より高い学習率を用いても安定した訓練が可能になる。第二に、初期値の設定に対する依存性が低下し、ネットワークの訓練がより堅牢になる。第三に、バッチ正規化自体が正則化効果をもたらし、ドロップアウトなしでも過学習を制御可能になる。Ioffe と Szegedyの論文では、バッチ正規化を適用したネットワークが、同じアーキテクチャの元のネットワークに比べて14倍少ない訓練ステップで同一の精度に到達し、さらにアンサンブルではImageNetベンチマークで4.82%のtop-5エラー率を達成することを報告している。これはヒューマンレベルの性能を初めて超えるものであった。

バッチ正規化の重要性は、単なる性能向上だけではない。この手法により、より深いネットワークの訓練が実務的に可能になり、また学習率の選択という重要な超パラメータの影響が軽減された。Deep ResNet　の成功も、バッチ正規化なしには達成不可能であったと考えられる。バッチ正規化は、ドロップアウト、ReLU活性化関数と並んで、2010年代の深層学習の爆発的な発展を支える三本柱の一つである。

14.6　AI計算基盤の進化 —— GPU、TPU、AI専用チップの台頭

深層学習の実用化は、アルゴリズムとデータだけでは不十分である。計算基盤——すなわち、訓練と推論を実行するハードウェア——の継続的な進化が、その成功を支えた。この段階では、GPU から TPU へ、そしてAI専用チップへという三段階の進化を追跡する。

NVIDIA GPU と CUDA の登場は既に述べた。2007年のCUDA公開から2012年のAlexNetまでの期間、GPUの計算能力は指数関数的に向上していた。特にGTX 580（AlexNetで用いられた）は、浮動小数点演算において約1.5テラフロップスの性能を持ち、当時のハイエンドCPUの数十倍の並列計算能力を提供していた。しかし、ニューラルネットワークの訓練は、必ずしも最高精度の浮動小数点演算を必要としない。2014年、NVIDIAはCuDNN（CUDA Deep Neural Network library）をリリースし、ニューラルネットワークの基本的な演算（逆伝播、畳み込み、プーリング）に最適化されたルーチンを提供した。

一方、Google は異なる戦略を採用した。2010年代前半、Google は自社の機械学習ワークロードに最適化されたカスタムシリコンの開発に着手した。その結果が Tensor Processing Unit（TPU）である。TPU v1 は2016年5月のGoogle I/O カンファレンスで公開されたが、実は2015年から運用段階にあり、Google Search、Google Photos、Google Translate、YouTube などの推論ワークロードを支えていた。

TPU v1 の設計思想は、NVIDIA GPU と根本的に異なっていた。GPU は多目的の並列演算装置であり、様々な計算パターンに対応する柔軟性を備えている。これに対して、TPU は、テンソル演算（行列乗算とその関連演算）に特化した専用チップであり、不要な機能を削ぎ落とすことで極度の効率性を追求していた。TPU v1 は、28ナノメートルプロセスで製造され、わずか40ワットの電力消費で、毎秒92兆の8ビット演算を実行していた。

TPU の設計における革新的な点は、低精度演算（int8）の活用にあった。多くの推論タスクでは、完全な32ビット浮動小数点演算は過剰であり、8ビット整数演算で十分であることが実証的に示されていた。この洞察により、メモリ帯域幅と電力消費を大幅に削減しながら、高いスループットを達成することができた。

TPU の性能優位性は劇的であった。同時期のハイエンド GPU（Tesla K40 など）との比較では、一部の推論ベンチマークで大幅な速度向上と電力効率の改善が報告された。ただし、TPU v1 は主として推論向けであり、2016年の AlphaGo Lee の主要計算基盤は GPU であった。TPU が大規模学習の中核として存在感を強めるのは、後続世代以降である。

NVIDIA も対抗策を講じた。2017年に発表された Tesla V100 は、新しい Volta アーキテクチャを搭載し、Tensor Core と呼ばれるテンソル演算専用の演算ユニットを組み込んだ。これにより、NVIDIA GPU も、TPU に対抗する専用性能を獲得し始めた。

さらに、Intel も AI チップ開発の競争に参入した。2016年の Nervana Systems と Movidius の買収により、Intel は深層学習訓練用の Neural Network Processor（NNP）と、エッジコンピューティング向けの Vision Processing Unit（VPU）の開発を進めた。

この計算基盤の進化は、単なる性能向上ではなく、深層学習の質的な変化をもたらした。GPU の登場により、数週間を要していた訓練が数日から数時間で完了するようになった。TPU とその後継世代の登場により、Google のような大規模な機械学習プロジェクトは、前例のない規模の推論と学習を実行できるようになった。AI 専用チップの普及により、エッジコンピューティング（スマートフォン、IoTデバイス）での推論も現実的になった。

このハードウェア進化の背景には、経済的・戦略的な力学が存在していた。ムーアの法則が限界に近づきつつある中で、汎用プロセッサの性能向上は鈍化していた。これに対して、AI・機械学習はデータセンターの主要な負荷となりつつあり、カスタムシリコンの設計による性能向上と電力効率の改善は、企業の競争力に直結していた。NVIDIA、Google、Intel だけでなく、Apple、Qualcomm、その他多くの企業が AI チップ開発に投資し始める状況は、IT産業における構造的な転換を示唆していた。

結論 —— 深層学習の系譜と次への展望

本章では、2006年から2016年にかけての「深層学習の夜明け」を追跡した。ヒントンの貪欲層別事前学習（2006）が理論的な基礎を提供し、AlexNetの成功（2012）が実証的な証拠を示し、VGG、GoogLeNet、ResNet　の進化（2013〜2015）が建築的な原理を深化させ、dropout、バッチ正規化といった技法の発展（2012〜2015）が実務的な安定性をもたらし、そして GPU、TPU、AI 専用チップの進化がスケーラビリティを保証した。

この十年間の変化は、AI研究における「パラダイムシフト」と呼ぶに値する。第3章のパーセプトロン、第8章の PDP 研究グループによるニューラルネット復興から始まった系譜は、ここで初めて真の実践的威力を獲得した。第9章で論じた統計的機械学習の時代、特にサポートベクターマシンの全盛期は、この深層学習の登場により急速に背景に退いていくことになる。

同時に、深層学習の成功は、決定論的ではない。ヒントンの貪欲層別事前学習は、後に「実は層別事前訓練なしでも、十分なデータと計算資源があれば訓練可能である」ことが判明する。AlexNetの ReLU も、後年の研究で理論的な完全な説明が得られたわけではなく、むしろ経験的な有効性に基づいている。バッチ正規化の理論的メカニズムについても、「内部共変量シフト」の説明が全く正確であるかどうかは、現在でも議論の余地がある。すなわち、深層学習の成功は、革新的な直感と試行錯誤による経験的最適化の複合体なのである。

なお、2024年のノーベル物理学賞がジョン・ホップフィールドとジェフリー・ヒントンに授与されたことは、深層学習の科学的地位の象徴的な承認であった。受賞理由は「人工ニューラルネットワークによる機械学習を可能にした基礎的発見と発明」であり、1943年のマッカロック＝ピッツモデル（第1章）に端を発し、ホップフィールドネットワーク（1982年）、ボルツマンマシン、そして逆伝播法の再発見（第8章）を経て本章で論じた深層学習革命に至る80年の系譜が、物理学の最高栄誉をもって認められたのである。ヒントンは2019年にベンジオ、ルカンとともにチューリング賞を受賞しており、コンピュータ科学と物理学の双方の最高賞を受けた研究者となった。

次章（第15章）では、この深層学習の基盤の上に構築された、表現学習、生成モデル、強化学習といった発展的なテーマを扱う。特に、教師なし学習による「表現」の獲得、敵対的生成ネットワークによる生成的なパラダイム、そして深層強化学習による自律的な意思決定が、どのように統合されていくかが次の舞台となる。AlexNet 以後、わずか4年で AlphaGo がプロ棋士に勝利し（2016年、第16章）、生成 AI が大衆化する（2022年、第18章）という急速な展開へと至る道が、ここに準備されていたのである。

参考資料（本章）

本章の執筆にあたり、以下の一次資料および文献を参照した：

- Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). “A fast learning algorithm for deep belief nets.” Neural Computation, 18(7), 1527-1554.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). “ImageNet classification with deep convolutional neural networks.” In Advances in Neural Information Processing Systems (pp. 1097-1105).
- Simonyan, K., & Zisserman, A. (2014). “Very deep convolutional networks for large-scale image recognition.” arXiv preprint arXiv:1409.1556.
- Szegedy, C., Liu, W., Jia, Y., et al. (2015). “Going deeper with convolutions.” In IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-9).
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). “Deep residual learning for image recognition.” arXiv preprint arXiv:1512.03385.
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). “Improving neural networks by preventing co-adaptation of feature detectors.” arXiv preprint arXiv:1207.0580.
- Ioffe, S., & Szegedy, C. (2015). “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” International Conference on Machine Learning (pp. 448-456).
- Jouppi, N. P., et al. (2017). “In-Datacenter Performance Analysis of a Tensor Processing Unit.” arXiv preprint arXiv:1704.04760.
- Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). “ImageNet: A large-scale hierarchical image database.” In IEEE Conference on Computer Vision and Pattern Recognition (pp. 248-255).
- The Nobel Prize. (2024). “The Nobel Prize in Physics 2024.”

第15章　表現学習と生成モデルの登場

15.1　オートエンコーダと変分オートエンコーダ（VAE）

深層学習が画像認識において劇的な成功を収めた2010年代初頭、機械学習の研究コミュニティが直面していた課題は、「大規模かつ複雑な高次元データから、意味のある表現をいかに学習するか」という問題であった。ラベルが豊富に存在する教師あり学習の枠組みでは、畳み込みニューラルネットワークが優れた成果を上げていたが、一方でラベルなしデータを活用し、その内部構造を捉える教師なし学習の方法論については、理論的にも実装的にも未成熟な状態にあった。

オートエンコーダは、この問題に対する最も直接的なアプローチである。その基本的な原理は単純であり、入力データ x を低次元の潜在空間（ラテント空間）に圧縮するエンコーダ部分と、その圧縮された表現から元の入力を再構成するデコーダ部分から構成される。再構成誤差を最小化することによって、ネットワークは入力データの「本質的な特徴」を潜在空間に抽出することを強制される。2010年代初頭、複数の隠れ層を備えた「深いオートエンコーダ」の設計と訓練の方法が確立され、特にノイズ除去オートエンコーダ（denoising autoencoder）や縮約オートエンコーダ（contractive autoencoder）といった変種が、データの内部構造をより堅牢に学習する方法として提案された。

しかし、従来のオートエンコーダには理論的な限界があった。潜在空間が単なる「点群」として機能し、その確率的性質が十分に活用されていなかったのである。2013年、ディーデリク・キングマとマックス・ウェリングは「Auto-Encoding Variational Bayes」論文を発表し、この問題を根本的に解決する変分オートエンコーダ（Variational Autoencoder, VAE）を提案した。

VAEの革新性は、潜在変数を点ではなく、確率分布（典型的には多次元ガウス分布）として扱う点にある。エンコーダは、入力 x に対して潜在空間上の確率分布 $p(z|x)$ を出力し、その分布からサンプリングされた z を用いてデコーダが入力の再構成を試みる。この過程には根本的な課題が存在した。サンプリング操作は確率的で微分不可能であるため、標準的な勾配法による訓練が困難である。キングマとウェリングが導入した「再パラメータ化トリック（reparameterization trick）」は、サンプリング過程を微分可能な形式に変換し、この問題を優雅に解決した。

VAEの目的関数は、変分下限（Evidence Lower Bound, ELBO）として定式化される。これは再構成誤差と潜在変数の分布が標準ガウス分布から乖離する度合いを表す KL ダイバージェンスの二項から構成される。この二項の均衡は、訓練の進行に応じて動的に変化し、初期段階では再構成品質が優先され、後期には潜在空間の規則化が強調される。結果として、VAE は単なるデータの圧縮器ではなく、一貫性のある生成モデルとなり、潜在空間の任意の点からデータの新しいサンプルを生成できる能力を獲得する。

VAE が提供したのは、確率的推論と深層学習を統合する理論的枠組みである。この統合は、後の確率的深層学習モデルの設計に深刻な影響を与えることになる。

15.2　敵対的生成ネットワーク（GAN）の誕生（2014）

2014年、イアン・グッドフェロー（Ian Goodfellow）が発表した「Generative Adversarial Networks」論文は、生成モデルの設計に対してまったく異なるパラダイムを提示した。その基本的な構想は、単一の目的関数に基づいて最適化を行う従来の機械学習から、二つのニューラルネットワークが対抗的に学習する「ゲーム論的」なフレームワークへの転換であった。

GAN の構成は、二つのプレイヤーから成る。第一は「生成器（Generator, G）」であり、ノイズ z からデータを生成する。第二は「識別器（Discriminator,　D）」であり、与えられたサンプルが本物のデータセットから出自したものか、G が生成したものかを判別する。訓練の過程では、G は D を欺く能力を磨き、D は本物と偽物を区別する能力を洗練させるという相互作用が生じる。

数学的には、この対抗過程は二人零和ゲーム（two-player zero-sum game）として定式化される。グッドフェローの論文では、理想化された条件の下で、十分な容量を持つ生成器と識別器が適切に最適化されれば、訓練データ分布と生成分布が一致する均衡点に到達しうることが示された。そこでの識別器の判別精度は 1/2（ランダムと同等）に近づく。

GAN　が従来の生成モデル（例えば、最尤推定に基づく確率的モデル）と根本的に異なる点は、マルコフ鎖の展開を必要としない点、そして学習の過程で暗黙的なモデリング（implicit modeling）を行う点である。すなわち、G の出力分布の明示的な確率密度関数を定義する必要がなく、サンプルを生成する能力さえあれば十分である。この特性は、複雑で高次元の分布の近似において、柔軟性と計算効率をもたらした。

一方、GAN の訓練の実務的側面は、提案当初から困難が伴っていた。G と D の学習率、初期化、アーキテクチャのバランスが微妙であり、訓練が不安定化する「モード崩壊（mode collapse）」現象が頻繁に発生した。G が訓練データの多様性を学習せず、限定された種類のサンプルのみを生成し、D がそれらを区別する能力を失うという悪循環である。その後の研究では、Wasserstein　GAN、スペクトル正規化、プログレッシブ成長など、訓練の安定化を実現するための多数の技法が提案されることになる。

GAN　の理論的深さと実用的可能性が認識された当初、それは生成モデルの分野に革命をもたらすと予想された。実際、その後数年間に、GAN　は画像生成、スタイル変換、超解像度化、そしてテキスト・音声・動画生成へと次々と応用領域を拡張していくことになる。

15.3　Word2Vec・GloVe ―― 分散表現の革命

表現学習の革新は、視覚領域だけに限定されていない。自然言語処理（NLP）の領域では、2013年にトマス・ミコロフらが発表した「Efficient Estimation of Word Representations in Vector Space」が、言葉を高次元ベクトル空間上の「点」として表現する革命的な手法をもたらした。Word2Vec と総称されるこのアプローチは、深層学習時代の NLP 研究を規定する根本的なパラダイム転換であった。

Word2Vec の基本的発想は、極めて単純である。「言葉とは、その言葉と一緒に出現する周辺の言葉によって定義される」という分布意味論（distributional semantics）の古典的原則を、ニューラルネットワークの観点から実装した。ミコロフらが提案した二つのアーキテクチャ――CBOW（Continuous Bag of Words）と Skip-gram ――は、この原則を異なる方向から具体化している。Skip-gram では、与えられた単語から周辺の単語を予測し、その予測過程で学習された重み行列が単語埋め込み（word embedding）となる。

Word2Vec の革新性の核心は、計算の効率性と結果の質をともに達成した点にある。標準的なニューラルネットワークの訓練では、出力層が語彙サイズ分の次元を有し、その全体に対してソフトマックス正規化を計算する必要があり、数百万の単語を有する言語では計算が絶望的に遅い。これに対し、ミコロフらは「ネガティブサンプリング（negative sampling）」という技法を導入した。すべての負例（与えられた単語の文脈に出現しない単語）に対して勾配計算を行う代わりに、確率的に選ばれた小数の負例のみに対して更新を行う。この見た目では粗雑な近似が、驚くべき効果をもたらし、訓練の高速化と学習品質の向上の双方を実現した。

Word2Vec の結果がもたらす言語的意味性は、当初の発表以来、強い印象を与え続けている。学習された単語ベクトルの空間では、「king - man + woman ≈ queen」といった意味的な類推（analogies）が保存される。これは、単語埋め込み空間が、意味的および統語的な言語構造を線形な幾何学的構造として内在化していることを示唆する。この観察は、その後の転移学習やファインチューニングの理論的基礎となるとともに、「表現」の学習が人工知能における中心的課題であることを強調した。

2014年、スタンフォード大学のジェフリー・ペニントンらは、GloVe（Global Vectors for Word Representation）を発表した。Word2Vec が局所的な文脈ウィンドウに基づくのに対し、GloVe は全コーパスを通じた単語共起統計の「グローバル」な情報を活用する。具体的には、単語-文脈共起行列を重み付き最小二乗法で因数分解し、その際に共起頻度に基づいて異なる重みを付与する。単語の空間的表現は、この因数分解から直接得られるベクトルである。

GloVe と Word2Vec の比較は、深層学習における表現学習の多元的性質を明らかにする。局所的な予測タスク（Word2Vec）と全体的な統計構造（GloVe）は、まったく異なるアプローチでありながら、本質的に相補的な言語構造を学習する。この相補性は、後の BERT や GPT といった大規模事前学習モデルの設計に継承されることになる。

15.4　転移学習と自己教師学習 —— ラベルなしデータの活用

深層学習の時代において、ラベル付きデータの取得は依然として高コストの課題であった。ImageNet の出現（第14章）により、大規模なベンチマークデータセットが利用可能になったが、ほとんどの実世界の応用においては、特定タスクに対する十分なラベル付きデータが不足していた。この課題に対する解決策として、転移学習（transfer learning）と自己教師学習（self-supervised learning）という二つの学習パラダイムが重要性を増していった。

転移学習の原理は、大規模なソース領域で事前学習（pretraining）したニューラルネットワークのパラメータを、ターゲット領域の限定的なラベル付きデータを用いてファインチューニング（fine-tuning）することである。古典的な発想ではあるが、深層学習の文脈では、このアプローチが劇的な効力を発揮することが実証された。ImageNet で事前学習した CNN は、医療画像解析、物体認識、領域適応などの多様なタスクに対して、限定的なデータセットからでも汎化性を備えたモデルを構築することを可能にした。転移学習の成功は、「何を学ぶか」という知識の抽象性のレベルが、異なるタスク間で転移可能であることを示唆していた。

一方、自己教師学習（SSL）は、より根本的な学習原理の転換を示唆していた。ラベルなしデータから、入力の一部を予測する課題や対照的な比較課題など、学習目標そのものを自動生成するアプローチである。Word2Vec はその代表的な一例であり、「与えられた単語の周辺に現れる単語を予測する」というタスクが、外部から与えられたラベルではなく、データ自身から導出される教師信号として機能している。

2000年代後半から2010年代初頭にかけて、ヤン・ルカンら（LeCun et al.）は、自己教師学習の理論的枠組みとその潜在的な意義を繰り返し強調していた。彼らの主張は、監督信号は外部的にラベル付けされた目標に限定されず、データの内在的な構造（例えば、隠れた部分を可視部分から予測するなど）から導出可能であるというものである。この視座は、後の BERT（2018 年）における「マスク言語モデリング」や、コンピュータビジョンにおけるコントラスティブ学習の設計原理に継承されることになる。

転移学習と自己教師学習の相互補完的な発展は、ラベルなしデータという「眠れる巨人」をAI研究が活用し始めたことを意味していた。この転換は、深層学習が実世界の限定的なリソースに適応する能力を大幅に拡張し、その後の大規模言語モデルの出現を可能にする基盤となったのである。

15.5　計算論的神経科学からの知見 ―― 脳構造とアーキテクチャ設計

深層学習の飛躍的な成功は、純粋に統計的・最適化的な観点からのみ説明できるのではない。その背景には、脳の神経生物学的構造に着想を得たアーキテクチャ設計という側面が存在する。この相互作用は、計算論的神経科学（computational neuroscience）という領域における深層学習研究者と神経科学者の対話を通じて、より明示的に自覚されるようになった。

計算論的神経科学の観点から見た場合、従来の機械学習モデル（例えば、サポートベクターマシンやランダムフォレスト）は、脳の実装を説明するまったく異なるメカニズムに依拠していた。一方、ニューラルネットワーク、特に多層構造を備えた深層ネットワークは、脳の階層的情報処理構造をより忠実に反映するものとして見なされた。

畳み込みニューラルネットワーク（CNN）の設計は、視覚皮質の受容野（receptive field）という神経生物学的概念から直接着想を得ている。Hubel と Wiesel の古典的研究（1960 年代）では、視覚皮質の神経細胞が、特定の方向や周波数成分に選別的に応答することが示されていた。CNN のフィルタが、このような特異的な特徴検出器を実装することで、効率的な視覚処理を実現できるという理解は、神経生物学からの直接的な着想である。

さらに重要なのは、注意機構（attention mechanism）の設計である。脳における注意とは、膨大な感覚情報の中から、行動や認知課題に関連する情報を優先的に処理するメカニズムである。後に Transformer アーキテクチャで中核的な役割を果たすようになる自己注意（self-attention）メカニズムは、この神経生物学的注意の計算論的抽象化として理解することができる。

同時に、深層学習の研究者たちは注意深く警告を発していた。生物学的な一致は、必ずしも本質的であるわけではなく、むしろ帰納バイアス（inductive bias）という観点から理解すべきであるということである。CNN が視覚皮質の局所性や階層性を模倣するのは、「画像データは局所的な相関構造を持つ」という事前知識を符号化するためであり、その根底にあるのは脳そのものではなく、データの統計的性質である。

この認識は、計算論的神経科学から深層学習への一方向的な着想の流れではなく、むしろ双方向的な対話を生み出した。深層学習が高い予測精度を達成したモデルは、脳の情報処理メカニズムの新たな仮説を生成し、その仮説は神経生物学的実験によって検証される。Goal-driven hierarchical CNN が、霊長類の視覚皮質 V1、V2、V4、下側頭皮質（IT）のニューロン応答をよく予測することは、その相互作用の成果である。

15.6　深層強化学習 —— DQNからAlphaGoへ

強化学習（reinforcement learning）は、AI 研究の初期から存在してきた課題であり、その理論的基盤は第9章で述べた Sutton と Barto の研究にさかのぼる。しかし、従来の強化学習アルゴリズムが直面していた根本的な制約は、状態空間の維持であった。チェスやチェッカーのような離散的で有限の状態空間に対しては、価値関数（value function）や方策（policy）を明示的に表現することが可能である。しかし、画像から得られるような高次元の連続的状態空間に対しては、この古典的なアプローチは実用性を失う。

2013 年、DeepMind の Volodymyr Mnih らは「Playing Atari with Deep Reinforcement Learning」と題する論文を発表した。その核心は、深層ニューラルネットワークをQ値推定器として使用し、Atari 2600 のゲーム画面から直接学習することであった。Deep Q-Network（DQN）と呼ばれるこのアーキテクチャは、二つの重要な技術的創新を組み合わせていた。第一は「経験リプレイ（experience replay）」であり、エージェントが経験した状態遷移の履歴をバッファに保存し、訓練時にはそのバッファから無作為にサンプリングした過去の経験に対して勾配更新を行う。これにより、連続的な観測の相関性が軽減され、標本効率が向上した。第二は「ターゲットネットワーク」であり、Q値の目標値を計算する際に、現在のネットワークパラメータではなく、遅れたコピーを使用する。これにより、訓練の安定性が向上した。

2013 年の初期論文では、7 つの Atari ゲームに対して評価され、6 つのゲームでは従来のすべての手法を上回り、3 つのゲームでは人間の専門家を凌駕した。その後、2015 年には Nature 誌に掲載された改良版（Nature DQN）により、49 のゲーム全体でテストが行われ、より一層の一般化能力が実証された。この成功は、深層学習と強化学習の統合が高次元の知覚入力に対して機能することを初めて明確に示した。

DQN の成功は、ゲーム AI の領域にとどまらず、強化学習全体のパラダイム転換を促した。その後、Double DQN、Dueling DQN、Noisy Net DQN、Rainbow DQN などの亜種が提案され、各々がアルゴリズムの特定の側面（過大評価の軽減、アーキテクチャの効率化、探索と活用のバランス）を改善した。さらに重要なのは、DQN が、高次元の視覚入力から価値関数を近似し、勾配法によって直接学習できることを実証した点である。

しかし、DQN が直接的には対応しにくい問題も存在した。チェス、囲碁、将棋といった、行動空間が膨大な完全情報ゲームでは、純粋な価値関数ベースの学習だけでは探索効率に限界がある。これらのゲームでは、モンテカルロ木探索（Monte Carlo tree search, MCTS）という古典的な探索アルゴリズムが以前から使用されていた。深層ネットワークによる方策・価値の推定と、MCTS の探索を結びつける発想が、次の大きな飛躍をもたらすことになる。

本章では、深層学習の内部に「表現」という新しい問題領域が浮上し、それに対して複数のアプローチ――オートエンコーダ、GAN、Word2Vec――が競合・共存していた時期を描いた。同時に、ラベルなしデータの活用、脳構造からの着想、そして古典的な強化学習と深層学習の統合という三つの動向が、次章で論じられる　AlphaGo　という統合的な成果へと収斂していく過程を示した。第16章では、ゲーム　AI　という限定的な領域での勝利が、いかにして科学的発見という普遍的な課題へと接続されたかを検討する。

参考資料（本章）

- Kingma, D. P., & Welling, M. (2014). “Auto-Encoding Variational Bayes.” International Conference on Learning Representations.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. (2014). “Generative Adversarial Nets.” Advances in Neural Information Processing Systems, 27.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). “Efficient Estimation of Word Representations in Vector Space.” arXiv preprint arXiv:1301.3781.
- Pennington, J., Socher, R., & Manning, C. D. (2014). “GloVe: Global Vectors for Word Representation.” Empirical Methods in Natural Language Processing.
- Mnih, V., Kavukcuoglu, K., Silver, D., et al. (2013). “Playing Atari with Deep Reinforcement Learning.” arXiv preprint arXiv:1312.5602.
- Mnih, V., Kavukcuoglu, K., Silver, D., et al. (2015). “Human-level control through deep reinforcement learning.” Nature, 518, 529-533.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). “Deep learning.” Nature, 521, 436-444.
- Yamins, D. L. K., Hong, H., Cadieu, C. F., et al. (2014). “Performance-optimized hierarchical models predict neural responses in higher visual cortex.” Proceedings of the National Academy of Sciences, 111(23), 8619-8624.

第16章　AlphaGoと汎用AIへの問い

16.1　AlphaGoの設計思想 —— モンテカルロ木探索と深層学習の融合

2016年3月、デミス・ハッサビスらが率いるDeepMindのアルゴリズム群AlphaGoが、世界で最も強い囲碁棋士の一人である李世ドル九段との対局において、4勝1敗という圧倒的な勝利を収めた。このイベントは単なるゲーム AI の勝利ではなく、深層学習と古典的な AI 手法の統合がいかにして超人的な知能を実現しうるかを実証した、AI 研究史上の重要な転機であった。

AlphaGo の設計思想を理解するために、まず囲碁というゲームの計算論的性質を認識する必要がある。チェス（盤面8×8、初期局面後の平均的な着手可能数約35）と比較して、囲碁（盤面19×19、平均的な着手可能数約250）は行動空間が著しく大きい。さらに重要なのは、囲碁における局面の「良さ」を評価する方法論が存在しなかったことである。チェスでは、駒の物質的価値、キング安全度、ポーン構造といった、比較的に定量化しやすい評価関数の設計が可能であったが、囲碁では「領土」という概念の本質的な定義の困難さが存在し、評価関数の設計は經験的で恣意的にならざるを得なかった。

伝統的なゲーム探索では、Deep Blue（第13章）の場合のように、ミニマックス探索とアルファベータ枝刈りによって、何百万個の局面を評価することで、数手先を読み切ろうとしていた。しかし、囲碁の場合はこのアプローチが本質的に限界を有する。仮に、各局面で平均250の着手候補があり、深さ40手までを読むとするならば、250^40 という天文学的な数の局面評価が必要となり、物理的に計算不可能である。

モンテカルロ木探索（Monte Carlo Tree Search, MCTS）は、この問題に対する優れた解決策であった。MCTS の基本的原理は、確率的なシミュレーション（プレイアウト）によって局面の価値を推定することである。具体的には、現在の局面から、各プレイヤーが一定のランダムな方策に従って終局まで着手を続けた場合、どちらが勝つ確率が高いかを統計的に見積もるのである。何百万回のプレイアウトを行うことで、粗い確率推定値が得られ、それが局面評価の代替物として機能する。MCTS は 2006 年の Reuben Coulom による発見以降、コンピュータ囲碁の分野で標準的な手法となっていた。

AlphaGo の革新は、この古典的な MCTS に対して、深層学習から得られた二つの神経ネットワークを統合したことにある。第一は「方策ネットワーク（policy network）」であり、与えられた局面から、最も着手する可能性の高い着手を予測するディープ CNN である。第二は「価値ネットワーク（value network）」であり、局面がどの程度、その着手者にとって有利であるかを推定する。

AlphaGo の推論パイプラインは、以下のように構成されていた。（1）現在の局面から、方策ネットワークが着手確率分布を出力する。（2）各候補着手に対して、MCTS がシミュレーションを実行する。（3）ただし、完全にランダムなプレイアウトではなく、学習済みのプレイアウト方策が MCTS に組み込まれている。（4）各シミュレーション経路の終局値と、価値ネットワークの出力の両者から、局面値が推定される。（5）複数のシミュレーション結果に基づいて、最良の着手が選択される。

この統合の効果は劇的であった。MCTS のみを使用する従来の囲碁プログラムは、着手選択に膨大な計算量を要し、かつ終局判定に不完全なドメイン知識（ルールベースの優先度スコアリング）に依存していた。これに対し、AlphaGo は、深層ネットワークの効率的な推論と確率的探索を組み合わせることで、着手選択の計算量を大幅に削減しつつ、精度を向上させた。

AlphaGo の訓練も複層的であった。方策ネットワークは、まず人間の棋譜データセット（約16万局、約3000万手）に対して、教師あり学習によってプレイ予測を行った。その後、この初期化されたネットワークを初期値として、自己対局による強化学習（policy gradient 法）によって改善された。価値ネットワークの訓練も同様に、その後、自己対局データセットに対する回帰として実施された。

この多段階的で異質な訓練手法は、AI の設計における重要な一般的パターンを示していた。すなわち、外部データ（人間の棋譜）による監督信号から始まり、その後、自己対局を通じた強化学習へと移行するというアプローチである。この転換は、「ヒューマンノレッジからの脱却」を意味し、その後の AlphaGo Zero へと続く道を準備していたのである。

16.2　李世ドル戦（2016）——人類最強棋士との対決

2016年3月9日から3月15日にかけて、ソウルでの五番勝負は、世界的な注目を集めた。李世ドルは、2013年から2016年にかけて国際プロ棋戦で最高の成績を挙げた棋士であり、多くの観測者にとって「人類最強の棋士」の象徴であった。一方、AlphaGo は、その数ヶ月前に欧州チャンピオンの樊麾（Fan Hui）を 5 勝 0 敗で破っていたが、当時の評価はなお「有力な AI プログラムの一つ」という程度であり、人間の最高レベルの選手に対して確実な勝利を収めるとは広く予想されていなかった。

第一局から、AlphaGo の非人間的な着手が観察された。コンピュータ囲碁の従来的な着手と異なり、AlphaGo は「直感的」で「創造的」な着手を打つ。例えば、第一局において、AlphaGo は上辺への着手を打ったが、その着手の目的は一見して理解困難で、何十手も先の戦術的な複雑な計算に基づくものであった。李世ドルのような人間の棋士にとって、この計算の道筋を読むことは困難であり、それが AlphaGo の最大の強みであることが明白になっていった。

結果は、4勝1敗で AlphaGo の勝利に終わった。第四局では李世ドルが唯一の勝利を手にしたが、そこでは有名な78手目を含む非定型の攻防が AlphaGo の探索を乱したと広く受け止められた。この唯一の人間による勝利は、その後の AI 研究者たちに対して、機械的な最適化と人間の創造性の間には、依然として微妙な相互作用が存在することを示唆した。

李世ドル戦の社会的・文化的インパクトは、AI　技術史における先例のない大きさであった。それは単なる技術的達成ではなく、「知能」「創造性」「人間らしさ」といった哲学的な問いを、世界規模で提起する事件となったのである。チェスにおける Deep Blue の勝利（1997年、第13章）でさえ、比較的に限定的な関心にとどまったのに対し、AlphaGo　の勝利は、テレビ、新聞、学術メディア、ソーシャルメディアを通じて、広く報道された。この報道の波は、その後の AI への社会的な期待（および恐怖）の増幅に寄与し、AI 規制やアライメント研究への関心が高まる背景の一つとなったのである。

16.3　AlphaGo Zero ―― 自己対局による超越

李世ドル戦のわずか10ヶ月後、DeepMind は驚くべき報告をなした。方策ネットワークと価値ネットワークを訓練するために人間の棋譜データを使用せず、ゲームルールのみを与えて自己対局させることで、AlphaGo をはるかに上回る強度を有する AlphaGo Zero を開発したというのである。

AlphaGo　Zero　のアルゴリズムは、簡潔ながら根本的な転換を表現していた。すべての学習プロセスは、強化学習に統一される。初期状態では、方策ネットワークと価値ネットワークは、ランダムに重み初期化される。その後、このネットワークを使用して、MCTS による自己対局を行う。各対局の結果（誰が勝ったか）が報酬信号となり、その報酬に基づいて両方のネットワークが勾配法で更新される。この過程は、ネットワークがより強くなるにつれ、MCTS の探索がより目的志向的になり、より良いゲームプレイがもたらされ、さらに強いネットワークへと至るという正のフィードバックループを形成する。

その学習の速度は驚くべきものであった。3日間で、AlphaGo Zero は AlphaGo Lee（李世ドル戦で使用されたバージョン）を100勝0敗で圧倒した。21日間で、AlphaGo Master（中国のプロ棋士との60勝1敗の成績を挙げたバージョン）の水準に到達した。40日間で、それまでのあらゆるコンピュータ囲碁プログラムを凌駕し、人間による棋譜データなしに、純粋に強化学習のみで達成された。

AlphaGo Zero の意義は、幾つかの層で理解される必要がある。第一は、技術的な層である。人間の棋譜という外部的な教師信号なしに、ゲームルールという最小限の事前知識のみから、超人的な知能が生成されることが実証された。これは、AI　が人間の知識に依存することから脱却し、純粋な強化学習の論理的帰結に基づいて学習可能であることを示唆している。

第二は、認知科学的な層である。AlphaGo Zero が到達した棋力は、人間の直感とは大きく異なる「非人間的な」着手パターンを採用していた。これは、囲碁における最適な着手の集合が、人間の棋譜から統計的に推測される分布と根本的に異なる可能性を示唆する。言い換えれば、人間が「良い」と判断する着手が、実際には最適ではなく、人間の認知的限界によってフィルタリングされた部分集合に過ぎないということである。

第三は、AI 研究の方向性に対する層である。AlphaGo Zero の成功は、「ドメイン知識の手工的な設計」から「データと強化学習の論理」へのパラダイム転換を促した。その後の AI 開発では、人間の専門知識を明示的に実装するのではなく、より単純な機械学習の枠組みに埋め込み、大規模な計算能力と多数の自動的な試行錯誤によって、高性能を引き出すというアプローチが採用されるようになったのである。

16.4　AlphaFold ── タンパク質構造予測への展開

AlphaGo Zero の報告から2年後、DeepMind の同じチームは、ゲーム AI の領域から大きく異なる科学的問題へと視線を転じた。タンパク質の三次元構造を、アミノ酸配列から予測するという、生物学における50年来の「グランドチャレンジ」である。

この問題の重要性は、医学や薬学の根幹に関わっている。タンパク質は生命の主要な構成物質であり、その三次元構造は、機能、相互作用、疾患との関連を決定する。タンパク質構造の決定には、X 線結晶構造解析や NMR 分光などの実験的手法が必要であり、各タンパク質の構造決定には数ヶ月から数年を要する。配列データベースが急速に拡大する一方で、実験的に決定された構造はそれに比べてはるかに少なく、この「構造の空白」を埋めることが生命科学および医学の発展に不可欠であった。

Critical Assessment of Structure Prediction（CASP）という、タンパク質構造予測の国際的なコンペティションが、2年ごとに実施されている。このコンペティションでは、配列のみが与えられ、実験的に決定されていない新しいタンパク質に対して、参加チームが構造予測を行う。その後、実験的に決定された構造と比較することで、予測精度が評価される。

AlphaFold の第一版（AlphaFold1）は、2018年の CASP13 に参加し、全体ランキングで首位となった。特に、難易度が高いと分類されたターゲット（既存の類似構造がない）に対して、顕著な精度を示した。しかし、この時点での AlphaFold1 は、従来の物理的・統計的手法を拡張したものであり、革新的ではあるが段階的な改善の域を出ていなかった。

2020年の CASP14 では、状況は劇的に転換した。AlphaFold2 は、まったく新しいニューラルネットワークアーキテクチャ「Evoformer」を採用していた。Evoformer は、生物学的および物理的情報を処理するために特別に設計された革新的な層構造を備え、また「structure module」と呼ばれる新しい出力部分により、直接的に三次元構造を予測することが可能になった。

CASP14 での AlphaFold2 の成績は、科学コミュニティを震撼させた。多くの難問ターゲットで実験的手法に迫る精度を示し、全体の中央値 GDT_TS は 92.4 に達した。これは、少なくとも相当数の課題において、実験的構造決定に匹敵する水準の予測が可能になったことを意味していた。

その後、DeepMind と EMBL-EBI は AlphaFold Database を整備し、数億件規模の構造予測を無償で公開した。この決定は、生命科学の基盤データを広く共有することで、マラリアワクチン開発、酵素設計、疾患関連タンパク質の機能解析など、多様な研究を加速させた。

AlphaFold の成功が示唆する科学的インパクトは、単なる予測精度の向上ではない。それは、人工知能が「科学的発見のパートナー」となりうることを実証したのである。タンパク質構造予測は、従来は生物学の周辺的な技術的問題と見なされていたが、AlphaFold2　の成功により、それが生命科学全体の基盤的課題であることが明白になったのである。

16.5　ゲームAIから科学的発見へ —— 汎用知能の可能性と限界

AlphaGo から AlphaFold へいたる発展は、AI 研究におけるパラダイムシフトを象徴している。コンピュータチェスの時代には、ゲーム AI は AI 研究の周辺的な応用領域と見なされていた。しかし、深層学習とモンテカルロ法の統合、そして強化学習による自動的な知識獲得というアプローチは、ゲームという限定的な領域を超え、科学的発見という普遍的な課題へと拡張されたのである。

この転換の根底にある共通的な原理は何か。ゲーム AI と科学的発見の間には、一見して距離があるように見える。囲碁は、明確に定義されたルール、有限の状態空間、完全情報という特性を有する。一方、タンパク質構造予測は、不完全な情報（配列のみ）から、複雑で非決定論的な物理的系の構造を推測するという問題である。

しかし、両者に共通する原理がある。第一は、「目的関数の学習可能性」である。ゲームにおいては、勝利という明確な報酬信号が存在し、それが強化学習の基盤となる。タンパク質構造予測においては、実験的に決定された既知の構造が「教師信号」として機能し、それに基づいて教師あり学習が行われる。どちらの場合でも、外部的または内部的な目的関数が学習シグナルとして機能する。

第二は、「アーキテクチャと帰納バイアス」である。AlphaGo の MCTS と神経ネットワークの統合、AlphaFold2 の Evoformer と structure module の統合は、どちらも「問題特異的な知識を機械学習の内部に符号化する」という戦略である。純粋にエンドツーエンドの深層学習に頼るのではなく、問題の数学的・物理的構造に基づいて、ネットワークアーキテクチャをデザインすることで、サンプル効率と精度の双方が向上する。

第三は、「自動化による品質管理」である。AlphaGo Zero における自己対局、AlphaFold における大規模学習と反復的評価は、人間の介入を最小化しながら、高品質の学習プロセスを自動的に実行する。この自動化は、手工的な設計による誤りを軽減し、スケーリングによる性能向上を可能にする。

こうした共通原理にもかかわらず、AlphaGo と AlphaFold の間には根本的な限界も存在する。AlphaGo は、特定のゲーム（囲碁）における最適性を追求するシステムである。そのアルゴリズムは、囲碁という領域に高度に特化しており、チェスやシャンチー、さらには経済学や物理学といった全く異なる領域に直接応用することはできない。同様に、AlphaFold は、タンパク質構造予測という特定の生物学的問題に特化したアーキテクチャと訓練手法を採用している。

このような領域固有性は、「汎用人工知能（AGI）」というAI研究の長年の目標との間に深刻な緊張関係をもたらす。AlphaGo と AlphaFold の成功は、「特定の領域における超人的な知能」が達成可能であることを示した。しかし、それは必ずしも「多領域にわたる柔軟な知能」や「人間のような移譲可能な学習能力」へ一歩近づいたことを意味しない。むしろ、それぞれの問題に対して、精巧に設計されたシステムが必要であるという事実が浮き彫りになったのである。

この認識から生じた問いは、第V部（第17章以降）で扱われることになる。大規模言語モデルの出現により、「言語」という人間の認知の最も普遍的な側面を対象とした機械学習が可能になったことで、汎用性と特殊性の関係がより複雑に再構成されることになるのである。しかし、その道の準備はすでに、本章で論じた AlphaGo と AlphaFold の成功の中に埋め込まれていた。すなわち、異なる領域に対して統一的なパラダイムを適用するのではなく、各領域の構造に基づいてアーキテクチャを設計しながら、学習のプロセス自体を自動化するという、「構造化された自動機械学習」のパラダイムである。

> 第16章では、ゲーム AI という限定的な領域での深層学習と古典的探索の統合（AlphaGo）から、科学的発見という普遍的な課題への展開（AlphaFold）を描いた。この転換は、AI 研究が「知能とは何か」という抽象的な問いから、「特定の問題領域における超人的性能をいかに実現するか」という工学的・実証的な問いへと軸足を移していることを示している。次章では、この転換がいかにして「言語」という新しい領域において、根本的な再構成をもたらしたかを検討する。言語は、AlphaGo や AlphaFold のように限定的な特殊領域ではなく、人間の認知のあらゆる側面と深く接続した、普遍的な表現様式である。その意味で、Transformer アーキテクチャと大規模言語モデルの出現は、AI 研究における新たな章の開始を告げるのである。

参考資料（本章）

- Silver, D., Huang, A., Maddison, C. J., et al. (2016). “Mastering the game of Go with deep neural networks and tree search.” Nature, 529, 484-489.
- Silver, D., Schrittwieser, J., Simonyan, K., et al. (2017). “Mastering the game of Go without human knowledge.” Nature, 550, 354-359.

- DeepMind. (2016). “AlphaGo beats Lee Sedol in Seoul.”
- Jumper, J., Evans, R., Pritzel, A., et al. (2021). “Highly accurate protein structure prediction with AlphaFold.” Nature, 596, 583-589.
- DeepMind. (2020). “AlphaFold: a solution to a 50-year-old grand challenge in biology.”
- AlphaFold Protein Structure Database. EMBL-EBI / DeepMind.

第V部

大規模言語モデルと生成AIの時代（2017年〜現在）

第17章〜第20章

第17章　Transformer —— 注意機構が変えた世界

17.1　Attention Is All You Need（2017）—— アーキテクチャの革新

人工知能の歴史において、2017年6月に発表された一編の論文「Attention Is All You Need」（Vaswani et al.）ほど、後続の研究領域全体に構造的な変化をもたらしたものは稀である。この論文は、自然言語処理の中核的課題である機械翻訳を対象としながら、実は深層学習全般、さらには人工知能全体の技術基盤を再構成するアーキテクチャを提案していた。Transformerと呼ばれるこのアーキテクチャは、それまで支配的であった再帰型ニューラルネットワーク（RNN、LSTM、GRU）を一掃し、注意機構（attention mechanism）に基づく並列処理可能な新しいパラダイムをもたらした。

Transformerの登場以前、機械翻訳や言語生成のタスクは、シーケンス全体を逐次的に処理するRNNに依存していた。これは生物学的な脳の動作にも似た一つのメリット——時間的・因果的な順序の自然な扱い——を持つ一方で、致命的な欠点を抱えていた。並列化が困難であり、長いシーケンスの学習が不安定であり、計算時間が長いという、工学的な課題である。Vaswaniらは、これらの問題すべてに対して根本的な解決策を提案した。それが「自己注意機構（self-attention）」に基づく、RNNを完全に廃止した設計であった。

Transformerの基本的な考え方は、シーケンスの各トークンが「他のすべてのトークンとの関係性を直接計算する」という操作を並列化することで実現する。従来のRNNは時刻tでの状態が時刻t-1の状態に依存し、したがって逐次的な計算を余儀なくされていた。これに対してTransformerは、全トークン間の相互依存関係をグローバルに、かつ一度に計算する。これが何をもたらすか。第一に、シーケンスの全トークンが同時に処理でき、計算の並列化が可能になる。第二に、長距離依存関係（long-range dependency）の学習が容易になる。RNNの「勾配消失問題」によって遠い過去の情報が失われやすいのに対し、Transformerでは任意の距離にあるトークン間に「直結路」を確保できる。第三に、注意の重みを視覚化できるため、モデルの解釈可能性が向上する。

17.2　自己注意機構の数理的基盤

Transformerの核心は自己注意機構（self-attention）にある。この機構を理解することは、2017年以降のAI研究全体を理解するための必須の知識となる。

自己注意機構は、シーケンス内の各要素が「他のどの要素に注目すべきか」を学習する仕組みである。より正確には、入力シーケンス内の任意の位置における表現（representation）が、他のすべての位置の情報をどの程度に統合するかを学習する。この統合方法は固定的ではなく、入力に応じて動的に変化する。

自己注意機構の動作は、次のような概念的なステップで理解できる。各トークンに対して「クエリ（query）」「キー（key）」「値（value）」という三つのベクトル表現を生成する。クエリとキーの間の相似度を計算し、これを正規化（softmaxによって確率分布化）することで「注意の重み」を得る。その重みで値ベクトルを加重平均した結果が、その位置における「注意済み表現」となる。

この操作には深い論理的意味がある。クエリは「現在の位置で何を知りたいか」を表し、キーは「他の位置が提供できる情報の種類」を表し、値は「実際に提供される情報」を表している。自己注意の比喩的説明としては、「ある単語を理解しようとするとき、文中の他のすべての単語と照らし合わせて、どの単語が最も関連があるかを判定し、その情報を統合する」というものである。

Vaswaniらが提案した「多頭注意（multi-head attention）」は、さらにこの機構を拡張する。単一の自己注意ではなく、複数の独立した注意ヘッドを並列に実行し、それぞれが異なる「関連性の種類」をキャプチャすることを目指す。例えば、あるヘッドは文法的な関係（主語と動詞など）に注目し、別のヘッドは意味的な関連性に注目し、さらに別のヘッドは長距離の依存関係に注目する可能性がある。複数のヘッドの出力は連結され、線形変換を通じて最終的な表現へと統合される。

Transformerアーキテクチャ全体は、このような自己注意層を6層（或いはそれ以上）積み重ねた「エンコーダ」と、同様に積み重ねながら因果性を保つ「デコーダ」から構成される。各層内では自己注意の後に「位置別全結合層（position-wise feed-forward network）」が続く。層正規化（layer normalization）と残差接続（residual connection）が数値的安定性と勾配流を担保する。

自己注意機構は、単なる工学的な改善ではなく、知能そのものの本質に関わる問題提起を含んでいる。それは「コンテキスト依存的な表現学習」という課題に、初めて直接的で効率的な解決策を提供したのである。第15章で述べた「表現学習」という概念は、ここにおいて新たな段階に到達する。

17.3　BERT —— 双方向事前学習の威力（2018）

Attention Is All You Needが発表されたわずか一年後の2018年10月、Googleの研究チームは「BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding」を発表した。BERTは、Transformerの力を実装化し、自然言語処理の実践的課題に大規模に適用した最初の成功事例となる。

BERTの設計上の工夫は、Transformerのアーキテクチャをいかに有効活用するかという問題に対する一つの答えである。GPT（2018年、OpenAIにより提案）が「左から右へ一方向に」シーケンスを処理する「自動回帰型」の言語モデルであったのに対し、BERTは「両方向から」シーケンスを処理する「双方向」の表現学習モデルを採用した。この設計選択がもたらす理論的・実践的含意を理解することは、後続の大規模言語モデル研究を理解するための鍵となる。

BERTの学習方法は「マスク言語モデリング（masked language modeling）」と呼ばれる自己教師学習タスクに基づいている。入力テキストの約15%のトークンをランダムに選び、その一部を「[MASK]」トークンに置き換え、モデルに「元の単語が何であったか」を予測させる。重要な点は、マスクされたトークンを予測する際に、左右両方向の文脈が利用可能であるという点だ。一般的な言語モデリング（「次の単語を予測する」）では、未来の情報は利用できない。しかしマスク言語モデリングでは、隠された単語の左右両方の文脈が提供される。これにより、モデルは「文脈が何であるか」をより深く学習できる。

第二の学習タスク「次文予測（next sentence prediction）」は、二つの文が連続しているかどうかを判定させるものである。これにより、モデルは文間の関係、文書レベルの意味論を学習する。この二つの自己教師学習タスクを組み合わせることで、BERTは汎用的で強力な言語表現を学習する。

学習に用いられた規模（scale）も革新的であった。BERTは約1.1億パラメータの BERT-base と、約3.4億パラメータの BERT-large という二つのサイズで提案された。当時としては「大規模」と呼ぶに値する規模であり、大量のテキストデータ（Wikipedia、BookCorpusなど）で事前学習された。この学習済みモデルは、その後、特定のタスク（質問応答、テキスト分類、固有表現認識など）に対して「ファインチューニング」することで、素晴らしい性能を発揮した。

BERTが発表された際の実績は圧倒的であった。GLUE（General Language Understanding Evaluation）ベンチマークで80.5%のスコアを達成し、それまでの最高成績を7.7ポイント上回った。SQuADv1.1では93.2%のF1スコア、SQuADv2.0では83.1%のF1スコアを達成するなど、複数の標準的なベンチマークで同時に新記録を樹立した。

BERTの重要性は、単なる性能の向上にとどまらない。それは、「大規模な無監督データから有用な表現を獲得する」というアプローチが実現可能であり、しかも非常に効率的であることを証明した。これにより、言語処理タスクの工学的パラダイムが根本的に転換した。「タスク固有の手工的特徴抽出」から「タスク非依存の汎用的な事前学習表現」へのシフトは、後の大規模言語モデルへの道を開く。

17.4　スケーリング則の発見 —— パラメータ・データ・計算量の関係

2020年1月、OpenAIの研究チームはJared Kaplan、Chris McCandlishら執筆による論文「Scaling Laws for Neural Language Models」を発表した。この論文は、BERTやGPTといった個別の成功事例を越えて、より一般的で構造的な原理を明らかにした。それが「スケーリング則（scaling laws）」である。

スケーリング則とは、損失（loss）がモデルサイズ、データセットサイズ、計算予算の関数として、どのような数学的関係を持つかを記述するものである。Kaplan et al.の発見は、七桁以上の規模にわたって、これらの量の間に「べき則（power law）」関係が成立することを示した。具体的には、モデルサイズが大きくなれば、データセットもそれに応じて大きくする必要があり、また両者の最適な成長率が存在するというものであった。

この発見の含意は深刻である。第一に、それは「スケール（規模）こそが知能に至る道である」という仮説に定量的な根拠を与えた。単なる「大きければ良い」ではなく、パラメータ数、学習データ量、計算資源の「最適な配分」が存在し、その配分に従えば性能が予測可能に向上することが示された。

第二に、これはAI開発の方針に直接的な影響をもたらした。スケーリング則によれば、計算能力に制限がある場合、「短時間で非常に大きなモデルを学習する」戦略が最適であることが示唆される。すなわち、モデルが完全に収束する（損失が最小化される）まで学習するのではなく、「早期終了」して、その時点でのモデルを使用することが、与えられた計算予算の下では最適であるという、直感に反した結論である。

第三に、スケーリング則は「予測可能性」をもたらした。実験データに基づいて、より大規模なモデルの性能を事前に推定することが可能になった。これは、研究投資の効率化につながる。

Kaplan et al.のスケーリング則は、その後の大規模言語モデル開発の重要な羅針盤となった。OpenAI のGPT　系列をはじめ、多くの大規模モデル開発は、モデル・データ・計算量の関係を明示的に意識するようになった。ただし、この関係は固定的な「自然法則」というより経験的な設計原理として理解すべきであり、後続の Chinchilla 則などによって最適なデータ対パラメータ比は修正・洗練されている。

17.5　Transformer派生 ―― Vision Transformer、Decision Transformer

Transformerの真の汎用性は、自然言語処理以外の領域での成功によって実証された。Transformerアーキテクチャが本質的には「シーケンス処理」のための設計であり、「言語」に特有のものではないという理解から、様々な分野への適用が試みられた。

Vision Transformer（ViT）

2020年12月、Googleの研究チームはDosovitskii et al.により執筆された「An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale」を発表した。これは、Transformerを画像認識に適用した最初の大規模成功例である。

Vision　Transformerの基本的な工夫は簡潔である。画像を「パッチ（小さな正方形領域）」に分割し、各パッチを線形層を通してベクトル化し、これを「言語モデルにおけるトークン」と同様に扱う。一般的には16×16ピクセルのパッチが用いられ、224×224の画像であれば196個のパッチが生成される。各パッチは独立した「画像トークン」として、Transformerのエンコーダに入力される。

このアプローチの優雅さは、CNN（畳み込みニューラルネットワーク）の帰納的バイアス（局所性、平行移動不変性）に依存せず、純粋に「トークン間の注意関係」に基づいて画像の構造を学習することにある。驚くべきことに、Vision　Transformerは、数十年にわたり画像認識を支配してきたCNNと同等、あるいはそれ以上の性能を達成した。特に大規模な学習データセットと計算リソースが利用可能な場合、ViTはCNNを上回る性能を示すことが多い。

Vision　Transformerの成功は、Transformerの汎用性の証明であり、同時にAI研究全体に重要な示唆をもたらした。即ち、「帰納的バイアスに頼った専用設計」よりも、「大規模データと計算能力を活用した汎用的アーキテクチャ」の方が、多くの場合優れているということである。これは、後続の研究における「アーキテクチャ設計」の重要性を相対的に低下させ、「スケール」と「データ」の重要性を高める認識的シフトをもたらした。

Decision Transformer

Transformerのもう一つの興味深い応用分野は強化学習である。2021年6月、Berkeley/Google所属の研究者らは「Decision Transformer: Reinforcement Learning via Sequence Modeling」を発表した。

従来の強化学習は、「価値関数の推定」或いは「方策勾配の計算」という、あくまで古典的な制御理論的アプローチに基づいていた。Decision　Transformerは、これを完全に転換する。強化学習を「シーケンスモデリング問題」として再構成するのである。

具体的には、強化学習の軌跡（trajectory）を「状態→行動→報酬」の系列として表現し、これをTransformerに入力する。モデルが学習すべきは「期待リターン（期待される累積報酬）」を条件として、次の最適な行動を予測することである。これは、言語モデルが「文脈を条件として次の単語を予測する」という動作と、形式的には完全に同じである。

Decision　Transformerの革新性は、強化学習を言語モデリングの枠組みに統一することで、Transformerの強力な学習能力と、自己教師学習の効率性を強化学習に適用できたという点である。実験的には、Atari、OpenAI　Gym、その他の標準的なベンチマークで、既存の手法と同等か上回る性能を達成している。

これらの派生モデル（ViT、Decision　Transformer）の成功は、重要な認識をもたらした。即ち、Transformerは「言語処理のための」アーキテクチャではなく、「情報処理一般のための」基本的な構成要素であるということである。

17.6　Mixture of Experts —— スケーリング効率の追求

Transformerモデルのスケーリングにおける次の課題は、計算コストの増加である。パラメータ数がN倍になれば、推論時の計算量もおおよそN倍になる。大規模モデルの運用（特に推論フェーズ）は、膨大な計算資源を必要とするため、実用化における大きなボトルネックとなる。

この課題に対する一つの解決策が「Mixture of Experts（MoE）」アーキテクチャである。MoEの基本的な考え方は、全パラメータを「複数の専門化された部分ネットワーク（エキスパート）」に分割し、入力ごとに「ルーター」が「どのエキスパートを使用するか」を判定するというものである。重要な特徴は

「疎な活性化（sparse　activation）」である。全エキスパートのパラメータを毎回使用するのではなく、各入力に対して少数のエキスパートのみを活性化させる。これにより、パラメータ数（「容量」）を増やしながら、実際の計算量（「活性化計算」）は抑制できる。

Googleの研究チームは2021年1月に「Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity」を発表した。Switch Transformerは、MoEの設計を大幅に簡潔化した。従来のMoE手法では、「上位K個の専門性の高いエキスパート」を選択する複雑なルーティング機構を用いていた。これに対してSwitch　Transformerは、「最も適切な単一のエキスパートのみを選択する」というシンプルなルーティング戦略（k=1）を採用した。

この単純化がもたらす利点は劇的である。第一に、ルーティング機構の計算コストが削減される。第二に、通信オーバーヘッドが最小化される（複数エキスパート間の同期が不要）。第三に、学習の不安定性が低減される（複雑なルーティングに伴う勾配の流れの問題が軽減される）。実験結果は、同じ計算予算で学習速度が最大7倍向上することを示した。

Switch　Transformerによって、「兆単位のパラメータ」を持つモデルの学習・運用が初めて現実的になった。これは単なるパラメータ数の増加ではなく、Transformerのスケーリング特性に関する深い理解に基づいている。疎なアーキテクチャにより、計算効率を維持しながら「知識の容量」を指数関数的に増加させることが可能になったのである。

MoEアーキテクチャは、その後、Gemini や Mixtral など一部の大規模モデルで採用され、計算効率を高める有力な設計選択肢となっていく。ただし、すべての主要モデルが　MoE　を採用しているわけではない。

17.7　Transformerから大規模言語モデルへ

2017年から2021年にかけて、Transformerアーキテクチャの理解と応用は急速に深化・拡大した。自己注意機構の発見は、シーケンス処理における「長距離依存関係」の効率的な扱いをもたらした。スケーリング則の発見は、より大規模なモデルが本当に良い性能を持つことを定量的に立証した。Vision TransformerやDecision　Transformerは、Transformerが言語処理に限定されない汎用的な基盤であることを示した。MoEアーキテクチャは、スケーリングの計算効率を大幅に向上させた。

これらの要素の組み合わせにより、2020年代の初頭には、「大規模言語モデル（Large　Language Model, LLM）」という新しいクラスのAIシステムが出現する準備が整った。BERTやGPT-2の時点では、まだ「事前学習モデル」という位置づけであったが、次章で述べるGPT-3の登場により、パラダイムシフトが完成する。

Transformerアーキテクチャの重要性は、その技術的優秀性だけでは説明できない。それはまた、AI研究における「スケール中心的」なパラダイムへのシフトの象徴でもある。過去のAI研究では、「適切なアーキテクチャ設計」「手工的特徴エンジニアリング」「ドメイン知識の活用」といった、研究者の「知恵」と「工夫」が強調されていた。Transformerの成功は、これらの要因の相対的重要性を低下させ、「大規模データ」「大規模計算」「大規模パラメータ」という「規模（scale）」の力を強調する認識的転回をもたらしたのである。

この転回は、後の大規模言語モデルの爆発的な成功、ChatGPTの登場による社会的インパクト、そしてAI研究全体の産業化・商用化の加速をもたらす。しかし同時に、新しい課題——アライメント問題、解釈可能性、社会的影響——もまた急速に顕在化することになる。第18章では、Transformerを基盤とした大規模言語モデルの発展と、それに伴う変化を詳述する。

> 本章では、2017年のTransformer登場から2021年のMoEまで、わずか五年の間に起こった一連の技術的革新を概観した。自己注意機構の発見、双方向事前学習（BERT）、スケーリング則の発見、他分野への応用（ViT、Decision　Transformer）、そして疎なスケーリング（MoE）という各段階は、いずれも後続の大規模言語モデル時代を準備する必要不可欠な要素であった。次章では、これらの基盤技術の上に構築された、OpenAIのGPT系列から、その後の多様な大規模言語モデルの登場と発展を追跡し、いかにして2022年のChatGPT登場が、AI研究と社会の界面に構造的な転変をもたらしたかを明らかにする。

参考資料（本章）

- Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). “Attention Is All You Need.” Advances in Neural Information Processing Systems, 30.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” NAACL-HLT.
- Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). “Scaling Laws for Neural Language Models.” arXiv preprint arXiv:2001.08361.
- Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). “Training Compute-Optimal Large Language Models.” arXiv preprint arXiv:2203.15556.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” International Conference on Learning Representations.

- Chen, L., Lu, K., Rajeswaran, A., et al. (2021). “Decision Transformer: Reinforcement Learning via Sequence Modeling.” Advances in Neural Information Processing Systems, 34.
- Fedus, W., Zoph, B., & Shazeer, N. (2022). “Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity.” Journal of Machine Learning Research, 23(120), 1-39.

第18章　大規模言語モデル（LLM）の系譜

18.1　GPT-1からGPT-5へ ―― OpenAIの軌跡

第17章で述べたように、Transformerアーキテクチャの登場は自然言語処理に革命をもたらした。しかし、この革新的なアーキテクチャが言語処理の本質的な問題――意味理解、文脈的推論、知識の獲得――にどこまで接近しうるのかは、当初は明確ではなかった。2018年から2024年にかけて、OpenAIは一連の言語モデルを段階的に公開し、スケーリング戦略の可能性を実証した。この系譜はLLMの発展史そのものである。

GPT-1は2018年にOpenAIから発表された最初の生成事前学習済みTransformerであり、117百万パラメータを有していた。このモデルは、Transformerデコーダアーキテクチャを採用し、自己教師学習によって大規模テキストコーパスから言語パターンを学習した。その最大の特徴は、教師なし事前学習によって習得した表現が、わずかな教師あり微調整で多様な下流タスクに転移可能であることを実証した点にある。

2019年2月に発表されたGPT-2は、パラメータ数とデータセットサイズの両方を約10倍に拡大した。GPT-2は15億パラメータを備え、8百万個のウェブページから構成される40ギガバイトのWebTextコーパスで学習された。GPT-2の革新性は、**ゼロショット学習**――明示的な訓練を受けていないタスクを遂行する能力――の強力さを示したことにある。テキスト生成、質問応答、機械翻訳、要約など、多様なタスクで事前学習モデルが有効に機能することが実証された。OpenAIはGPT-2の初期バージョンを段階的にリリースしており、その過程で生成テキストの悪用リスクへの配慮を示している。これは、後のAI安全性研究への先駆的な取り組みとも解釈される。

GPT-3は2020年5月に発表され、AIの発展において最も象徴的なマイルストーンの一つとなった。1,750億パラメータ――GPT-2の約100倍――を有するGPT-3は、数ショット学習（few-shot learning）と一ショット学習（one-shot learning）の能力で際立っていた。わずか数例を示すだけで、多くのタスクに適応可能であることが実証され、これは人間の学習方式に近い柔軟性を示唆した。GPT-3は英語テキストにおいて、コード生成、詩作、論文作成、数学問題の解法など、前例のない多様な応用可能性を示した。このモデルの登場は、スケーリング則（scaling laws）の重要性を学術界に認識させた転機となった。

スケーリング則に関して特筆すべきは、2020年にOpenAIのJared Kaplanらによって発表された研究である。彼らは、神経言語モデルの損失関数がモデルサイズ、データセット規模、計算量に対して冪則に従うことを実証した。この発見は、より大きなモデルの構築がより小さなデータセットで効率的に訓練可能であることを含意していた。また2022年にはDeepMindのHoffmannらが「Chinchilla」スタディを発表し、与えられた計算予算に対して最適なモデルサイズとデータ量のバランスが存在することを示唆した。これらの理論的洞察は、GPT-3以降のモデル設計の指針となった。

GPT-3.5は、ChatGPTの基盤となったモデルであり、特に対話的なタスクに特化した微調整が施された。GPT-3.5はGPT-3の能力を保持しながら、より効率的な計算とより高速なレスポンス生成を実現し、APIを通じた大規模な商用展開を可能にした。

GPT-4は2023年3月に発表され、Transformerアーキテクチャの基本原則に忠実でありながら、複数の技術的改善を統合した。GPT-4の公式なパラメータ数は未公表であるため、本章では推定値には立ち入らない。重要なのは、GPT-4が**マルチモーダル能力**――テキストと画像の両方を処理する能力――を統合し、数学的推論、コード生成、複雑な質問応答においてGPT-3.5を大きく上回る性能を示したことである。GPT-4の登場は、LLMの能力が次第に単なる言語処理から知識統合的な推論へ移行していることを示唆している。

2024年5月にはGPT-4oが発表され、テキスト、画像、音声を統合的に扱う方向がさらに明確になった。ここで重要なのは、LLMが単独のテキスト生成器ではなく、複数の入出力をまたぐ基盤モデルとして位置づけ直されたことである。

2024年後半にOpenAIがo1を公開すると、開発の焦点は「より大きいモデル」だけでなく「より長く考えさせる推論手続き」へも移った。さらに2025年にはGPT-5が公開され、GPT系列と推論重視のo系列が併走する構図が明瞭になった。製品名や細かなSKUは短期間で変化するため歴史記述として重要なのは、LLM開発の最適化対象が、パラメータ数だけでなく**推論時計算**やツール使用の設計へ広がった点である。

一方、OpenAI以外の企業も推論能力の強化に注力し始めている。Anthropic、Google、DeepSeek　など各社は、長文脈処理、外部ツール使用、思考過程の拡張といった異なる方向から同じ課題に接近した。この傾向は、単なるパラメータの拡大から「推論時計算」や「推論手順の設計」へのシフトが業界全体の共通認識となったことを示している。

この進化系列――GPT-1の117百万パラメータからGPT-4o、o1、そして2025年のGPT-5へと至る道のり――は、スケーリング仮説の正当性を実証すると同時に、アーキテクチャと学習手法の継続的改善、そして推論能力の段階的深化の重要性を示している。

18.2　ChatGPTの衝撃（2022年11月）―― AIの大衆化

2022年11月30日、OpenAIはChatGPTを公開した。この出来事は、AIの歴史において最も社会的インパクトが大きい瞬間の一つとなった。ChatGPTはGPT-3.5をベースとしており、対話的な相互作用に特化した微調整が施されていた。

ChatGPTの成長曲線は歴史的に前例がない。**わずか5日間で100万ユーザーに到達**し、2カ月後には1億ユーザーを突破した。比較のために述べるならば、Instagramは250万ユーザーに達するまで2.5ヶ月を要し、Netflixは同じ規模に到達するまで3.5年を要した。ChatGPTの成長速度は既存のすべてのアプリケーションを圧倒していた。

なぜこのような爆発的な採用が起きたのか。第一に、ChatGPTは従来のAIシステムに比べてはるかに**ユーザーフレンドリー**であった。複雑なAPI呼び出しやコマンドラインインターフェースを必要とせず、自然言語での対話という人間にとって最も直感的な形式で機能した。第二に、その**汎用性**である。ChatGPTは、文章作成から問題解決、コード生成まで、多様な知的タスクに即座に応用可能であった。第三に、**社会的文脈**がある。2022年は、AI分野での技術的ブレイクスルーが次第に知識労働者のコミュニティに認識されつつあった時期であり、ChatGPTはその可能性を最も直感的に実証した。

ChatGPTの衝撃は技術的にも社会的にも多方面に及んだ。技術的には、大規模な言語モデルが単なる研究対象から実用的なツールへと転換したことを意味していた。これは、スケーリング仮説が実証的に正当化されたことと、言語モデルが**知識作業者向けのアシスタント**として機能しうることを示した。社会的には、AIの能力と限界についての公開的な議論がさらに活発化した。ChatGPTの利用者たちは、その多様な応用可能性とともに、時折の不正確さや幻覚（hallucination）現象を経験し、LLMの信頼性について深刻な問いを提起した。

また、ChatGPTは従来のエンタープライズAIシステムとは異なる展開戦略を採用した。OpenAIは月額20ドルのChatGPT Plusサブスクリプションモデルを導入し、**AI能力の民主化**と**商用化**の両立を図った。この戦略は、AI企業の事業モデルにおいて一つのテンプレートとなり、後続する企業の参入を促進した。

18.3　Claude、Gemini、o1 —— 推論能力と安全性の新たなフロンティア

ChatGPTの成功は、大規模言語モデルの商用的価値を明確化し、複数の企業による同様のシステム開発を刺激した。2023年から2024年にかけて、異なる技術的戦略と価値観に基づいた複数のLLMが相次いで登場した。

Anthropic社のClaudeは、OpenAIとは異なるAI安全性アプローチを標榜して登場した。AnthropicはOpenAIの前幹部であるDario　Amodei、Daniela　Amodeiらによって2021年に設立された企業であり、「Constitutional AI」という概念を中心に据えた安全で信頼できるAIモデルの開発を目指していた。

Constitutional　AIは、RLHF（強化学習による人間フィードバック）と並ぶアライメント手法として重要である。基本的なアイデアは、LLMに一連の原則（憲法）を明示し、その原則に照らして自己批判・自己修正させるという自己教師的なアプローチである。2020年代半ばにはAnthropicが憲法文書を更新し続けており、規則の列挙だけでなく、判断理由の明示を重視する方向が見られる。

Claudeの特徴的な能力の一つは、**長文脈処理**である。Claude 2 系列で 10 万トークン級、Claude 3 系列以降で 20 万トークン級の文脈長が実用化され、一部ではさらに大きな文脈窓も試験された。これは長大な文書分析やマルチターン対話において優位性をもたらした。

GoogleのGeminiは、同社の多年にわたる言語モデル研究（BERTから始まる）とスケール化への投資を結実させたモデルファミリーである。2023年末にGemini 1.0が公開され、2024年2月には Gemini 1.5 Pro が登場した。Gemini 1.5 は 100 万トークン級、のちには 200 万トークン級まで拡張された長文脈処理を掲げ、テキスト・画像・音声・動画を統合的に扱う Google のマルチモーダル戦略の中核を担った。

GoogleのGeminiファミリーの発展は、単なるパラメータの増加ではなく、長文脈処理、マルチモーダル統合、推論手順の最適化を通じて進行している。ここでは、OpenAIが牽引した対話型LLMの潮流に対して、Googleが「検索」「マルチモーダル」「長文脈」を強く結びつけた点が重要である。

OpenAIのo1は、2024年9月に発表され、スケーリング戦略において従来型の「パラメータ規模の拡大」から「推論時計算の拡大」へのシフトを象徴している。o1　は、より多くの内部計算時間を用いて複雑な問題に対して深い推論を行う方針を明示した。このアプローチは、単純な規模の拡大だけでは届きにくい能力領域が存在することを示唆し、AIの性能向上には複数の次元での最適化が必要であることを改めて認識させた。

2025年以降は、各社が「単一の万能モデル」を競うだけでなく、軽量モデル、長文脈モデル、推論重視モデルを使い分ける多層的戦略を採るようになった。歴史的に重要なのは、LLMの競争軸が単一ベンチマークではなく、安全性哲学、多言語能力、マルチモーダル性、推論能力、エージェント適性へと分岐したことである。

これら複数のモデルの登場は、LLMの発展が単一企業の独走ではなく**多元的競争**へ移行したことを意味する。特に注目すべきは、推論能力の重要性が業界全体で認識され、同時にAI安全性・AI倫理の問題がこれまで以上に本質的なものとして扱われるようになったという点である。

18.4　オープンソースLLM —— LLaMA、Mistral、DeepSeek

一方、大型テック企業による商用モデルの支配に対抗するように、オープンソース化されたLLMが次第に重要な位置を占めるようになった。オープンソースLLMは、学術研究の民主化、商用化への障壁の低下、および企業の研究成果の透明性向上をもたらしている。

Meta（旧Facebook）のLLaMAファミリーは、2023年に研究向けモデルとして発表された後、コミュニティによる改変と最適化を通じて急速に進化した。2023年7月には Llama 2 が公開され、研究利用にとどまらず商用利用にも開かれたライセンスが注目を集めた。Llama 2 は 7B、13B、70B の三つのサイズで提供され、**オープンウェイトのモデルが商用LLMに対して競争力を持ちうる**ことを実証した重要な事例である。

2024年7月には、Llama 3.1 の 405B パラメータ版が発表され、当時の公開系LLMとしては最大級の規模となった。Llama 3.1 は 15 兆トークンで学習され、文脈長も 128K に拡張された。ここで重要なのは、最先端級のモデルが一部でも公開されることにより、評価・蒸留・追加学習のエコシステム全体が加速したことである。

Mistral AIは、フランスを本拠とする新興企業であり、効率性重視のアプローチで知られている。2023年末に発表されたMixtral 8×7Bは、**Mixture of Experts（MoE）**アーキテクチャを採用し、スパースな計算を実現した。Mixtral 8×7Bは、8個の7Bパラメータ専門家（expert）から構成されるが、各トークンに対して2つの専門家だけがアクティブ化される。これにより、総パラメータ数は56Bであるにもかかわらず、アクティブパラメータ数は13Bに過ぎず、計算効率性を大幅に改善している。MoEアーキテクチャは、条件付き計算（conditional computation）という原理に基づいており、入力に応じて必要な計算リソースだけを動員する。このアプローチは、LLM開発における計算効率性の重要性が増す中で、重要な戦略的方向を示している。

DeepSeekは、2020年代半ばに効率的な言語モデル開発における新しいパラダイムを提示した企業である。DeepSeek は単なる既存モデルの追随ではなく、**推論重視**、**効率重視**のアプローチで知られ、その登場は2024年から2025年の言語モデル開発に大きな転換をもたらした。

2025年1月には、DeepSeek が **DeepSeek-R1** を公開し、業界に衝撃を与えた。R1 は推論に特化したモデルであり、強化学習を前面に出した訓練戦略と蒸留モデルの公開により、高度な推論能力の普及を加速させた。これにより、巨大な計算資源を持たない組織でも、高い推論能力を持つシステムを構築しやすくなった。

同時期の DeepSeek-V3 系列では、Mixture of Experts と蒸留を組み合わせた効率化が追求された。ここでの歴史的意義は、最先端性能の争点が「とにかく巨大なモデルを作ること」から、「限られた計算資源でどこまで推論能力を引き出せるか」へ移ったことにある。DeepSeek の急速な展開は、中国のAI企業が単なる追従者ではなく、効率的スケーリングと推論能力という次世代の問題においても先駆的なイノベータであることを示唆している。

オープンソースLLMの台頭は、AI開発の民主化をもたらしただけでなく、計算効率性をめぐる競争を加速した。LLaMA、Mistral、DeepSeekの登場により、大規模な計算リソースを有しない組織や研究機関でも、競争力のあるLLMを開発・配備することが可能になった。同時に、これらのモデルは学術研究の透明性と再現性を向上させ、AI開発全体の技術的水準を底上げしている。

18.5　中国のLLM開発と多言語・日本語LLMの発展

大規模言語モデルの発展は、必然的にグローバルな複数中心構造へと向かっている。特に中国のLLM開発は、2023年から2024年にかけて著しい進展を示し、単なる欧米モデルの追従ではなく、独自の技術的課題に取り組む地域的なパイオニアとしての地位を確立しつつある。

BaiduのERNIE（Enhanced Representation through Knowledge）シリーズは、中国語の理解と推論に特化したモデルファミリーである。ERNIE は、中国語の特有の言語特性――複合語、慣用表現、文化的文脈――を深く扱う方向で発展してきた。

AlibabaのQwen（通義千問）は、2023年に公開されたモデルファミリーである。Qwen は中国語と英語を中心にしながら多言語能力を重視し、公開ウェイトや派生モデルの豊富さによって、研究と実装の両方で大きな存在感を持つようになった。

中国におけるLLM開発の特徴の一つは、規制環境への適応である。中国の生成AI規制は、公開提供されるモデルに対して安全審査や届出を強く求めており、企業は性能競争と制度適合を同時に進める必要があった。この制約の中で、Baidu、Alibaba、DeepSeek などがそれぞれ異なる強みを形成したことは、中国のLLM史を理解するうえで重要である。

日本語LLMの発展は、オープンソースコミュニティと企業による協同によって進められている。Meta の Llama 系列や Qwen 系列を土台に、日本語継続事前学習や指示学習を施した派生モデルが複数の組織によって開発された。特に rinna や Stability AI による公開モデルは、日本語固有の語彙、文体、対話慣習への適応を進めた。

Stability AI は 2023 年に「Japanese Stable LM Beta」系列を発表し、Llama 2 を土台にした日本語継続学習モデルを公開した。これは、日本語向け公開LLMの整備が英語圏の周辺的模倣ではなく、独自の利用基盤を育てる試みであることを示していた。

rinna は、Llama 系列や Qwen 系列を基盤とした日本語継続学習モデルを公開し、2023 年末には Qwen ベースの「Nekomata」系列も発表した。これらのモデルは、日本語テキスト生成や質問応答などの実務的タスクでの利用可能性を高め、日本語LLMの公開エコシステム形成に貢献した。

多言語LLMの発展は、言語資源が限定的な言語コミュニティにおけるAIアクセスの民主化に大きく貢献している。同時に、言語の多様性を保存し、グローバルなAI能力の地域的な適応を促進する意義を持っている。

18.6　Constitutional AI・RLHF・DPO ―― アライメント手法の進化

LLMの能力向上と同様に重要なのが、その出力を人間の価値観や目標に整合させるための「アライメント」技術である。初期の大規模言語モデルは、統計的パターン学習に基づいているため、必ずしも社会的・倫理的に望ましい出力を生成するとは限らない。この問題に対処するために、複数の技術的アプローチが開発されてきた。

RLHF（強化学習による人間フィードバック）

Reinforcement Learning from Human Feedback（RLHF）は、2010年代後半の選好学習研究を背景に、2022年のInstructGPT論文によってLLMアライメントの中核手法として広く知られるようになった。RLHFの基本的な流れは以下の通りである：

1. **教師あり学習段階**：人間のラベラーが望ましい行動の例を示し、言語モデルを微調整する。この段階では、プロンプトに対する適切な応答を大量に生成させ、モデルを教師あり学習で初期化する。
2. **報酬モデル学習段階**：ラベラーが複数のモデル出力を比較し、どちらがより望ましいかをランク付けする。このランク付けデータを用いて、報酬モデル（reward model）を訓練する。報酬モデルは、与えられた入出力ペアに対して、人間の好みスコアを予測する。
3. **RL最適化段階**：報酬モデルをシグナルとして用いて、強化学習（PPO：Proximal Policy Optimizationなど）によって言語モデルを最適化する。この段階で、モデルは報酬を最大化するように更新される。

InstructGPT（13億パラメータ）がGPT-3（1,750億パラメータ）の出力を上回る品質を達成したことは、パラメータ数よりもアライメントの質が重要であることを示唆した。ChatGPTもこのRLHFアプローチに基づいており、LLMの商用化を可能にした重要な技術である。

RLHFの利点は、複雑な人間の価値観を直接的に言語モデルに組み込める柔軟性にある。欠点としては、報酬モデル自体の学習に必要なラベラーの努力が膨大であることと、報酬モデルの不完全性（reward hacking）がモデルを望ましくない方向へ導く可能性が挙げられる。

Constitutional AI（憲法的AI）

AnthropicによるConstitutional AIは、RLHFを超える安全性向上を目指したアプローチである。基本的なアイデアは、人間からのフィードバックだけに頼るのではなく、**一連の明示的な原則（憲法）**に基づいてモデル自身に自己批判・自己修正させることである。

Constitutional AIの手順は以下の通りである：

1. **赤チームの攻撃**：意図的に危害をもたらしうるプロンプトを生成し、モデルが問題のある出力を生成させる。
2. **自己批判段階**：モデル自身に、与えられた憲法の原則に照らしてその出力を批判させる。LLMは自然言語で、なぜその出力が原則に違反しているかを説明する。
3. **自己修正**：モデルに、批判に基づいて改善された応答を生成させる。
4. **教師あり学習**：改善された応答を正例として用いて、モデルを微調整する。

この手法の利点は、人間のラベリング作業量を削減できること、および複数の価値原則を同時に実装できることにある。

Constitutional AI をめぐっては、2020年代半ばに「単に禁止事項を列挙するだけでなく、判断理由をどのようにモデルへ埋め込むか」が論点となった。ここで重要なのは、アライメント研究が単なる有害出力の抑制から、モデルの振る舞いを説明可能な原則へ結びつける方向へ進んだことである。

DPO（直接選好最適化）

Direct Preference Optimization（DPO）は、2023年にスタンフォード大学の研究者により提案されたアプローチであり、RLHFの複雑性を大幅に簡化しながら同等かそれ以上の性能を実現する。

DPOの核心的な洞察は、報酬モデルを明示的に訓練せず、言語モデル自身が報酬モデルとして機能しうるという点である。基本的な流れは以下の通りである：

1. **選好データの準備**：同じプロンプトに対して、より良い応答（positive）とより悪い応答（negative）のペアを人間が評価する。
2. **直接最適化**：二値交差エントロピー（binary cross-entropy）目的関数を用いて、言語モデルを直接最適化する。この目的関数は、より良い応答の尤度を上げながら、より悪い応答の尤度を下げるように設計されている。

DPOは、RLHFが報酬モデルとRL最適化という二つの複雑なステップを必要とするのに対し、教師あり学習的な直接最適化で同じ結果を達成する。複数の実験により、DPOはPPOベースのRLHFと比較して、感情制御やダイアログ品質において同等またはそれ以上の性能を示すことが報告されている。

これら三つのアライメント手法——RLHF、Constitutional AI、DPO——は、AI安全性研究における段階的な進化を示唆している。RLHFは人間の判断を直接的に組み込むアプローチ、Constitutional AIはルール的原則を中核とするアプローチ、DPOは計算的効率性を重視するアプローチである。各手法は、異なる約束と制約のバランスを反映している。

18.7　パラメータ効率的学習 —— LoRA、QLoRA、アダプタ技術

LLMの能力の拡大に伴い、実運用における課題の一つが、これらの膨大なモデルの**微調整（fine-tuning）**にかかる計算コストである。GPT-4やLLaMA 70Bといったモデルを特定の用途に適応させるためには、数十ギガバイトのGPUメモリと膨大な計算時間が必要となる。この障壁を低減し、よりアクセス可能なカスタマイズを実現するために、複数の「パラメータ効率的学習」技術が開発された。

LoRA（Low-Rank Adaptation）

Low-Rank Adaptation（LoRA）は、Microsoft Researchが2021年に提案した技術である。LoRAの基本的なアイデアは、大規模言語モデルの全ウェイトを微調整するのではなく、**低ランク行列の積として表現される小規模な追加パラメータ**のみを訓練することである。

具体的には、元の重み行列Wが与えられたとき、LoRAは以下のように動作する：

- 元の重み行列Wと追加の低ランク行列の積（AとB）の合計：W'= W + AB^T
- ここで、Aは（d_in × r）、Bは（d_out × r）の行列であり、r（ランク）は元の次元d_inやd_outより遥かに小さい。

この手法により、LLaMA 65Bモデルでも、微調整に必要なパラメータ数を元の数%に削減できる。例えば、LoRA秩r=8を使用すれば、訓練パラメータ数は全パラメータの0.1%以下に削減される。これにより、消費者向けGPUメモリ（例：48GB）で大規模モデルを微調整することが可能になる。

LoRAのもう一つの利点は、複数のLoRA「アダプタ」を同じベースモデルに対して独立的に訓練でき、用途に応じて動的に切り替えられることである。例えば、医療用と法律用の二つのLoRAアダプタをLLaMA 70Bに適用し、入力に応じて適切なアダプタを選択できる。

QLoRA（Quantized LoRA）

QLoRAは、University of WashingtonとHugging Face の研究者により2023年に提案された、さらに激進的な効率化技術である。QLoRAは、LoRAにおける基盤モデルのウェイトを**4ビット量子化**することで、メモリ消費をさらに削減する。

QLoRAのプロセスは以下の通りである：

1. **基盤モデルの4ビット量子化**：事前学習済みLLMのウェイトを4ビット精度に圧縮する。これにより、例えばLLaMA 65BはGPUメモリで約48GBを要する（元は260GB以上）。
2. **LoRA微調整**：量子化されたモデルに対してLoRA技術を適用し、低ランク行列のみを訓練する。
3. **推論時の復号化**：推論時には、量子化された重みを必要に応じてより高精度で復号化する。

QLoRAの報告によれば、同じベースモデルに対してQLoRA微調整は、全ウェイト微調整に近い性能を保ちながらメモリ消費を大幅に削減する。これにより、研究者や小規模企業でも強力なLLMをカスタマイズできるようになった。

アダプタ技術との統合

LoRA/QLoRA以外にも、**アダプタネットワーク**と呼ばれる方法論がある。アダプタは、トランスフォーマーの層間に挿入される小規模なニューラルネットワークであり、限定的なパラメータのみを訓練する。複数の研究グループが異なるアダプタ設計（MADapterなど）を提案している。

これらのパラメータ効率的学習技術の登場は、LLM開発の民主化に大きく貢献している。かつて大規模企業のみが実現可能だった高度なカスタマイズが、限定的な計算資源しか持たない研究機関や企業でも実現可能になった。同時に、これらの技術はLLMの実装の多様性――同一の基盤モデルから多様な目的別モデルが派生する構造――を促進している。

18.8　ノーベル物理学賞と機械学習理論の制度的認知（2024年）

2024年のノーベル物理学賞は、学習アルゴリズムと神経ネットワークの基礎的貢献に対して、ジェフリー・ヒントンとジョン・ホップフィールドに授与された。これは、機械学習とAIの理論的基盤が、物理学の中核的成就として国際的に認識されたことを意味する歴史的な瞬間である。

ジェフリー・ヒントンはバックプロパゲーション学習、ボルツマンマシン、深い神経ネットワークの学習におけるベクトル表現の理論的基礎に対して受賞した。ホップフィールドは、相互に結合されたニューロンが情報を保存・回復することが可能であることを示した「ホップフィールドネット」の理論化により、ニューラルネットワークの記憶理論を確立した。

この受賞の意義は単なる栄誉の問題ではない。1943年のマッカロック＝ピッツ論文から始まる人工ニューロンの理論的系譜が、80年以上の時を経て、物理学の根本的な理論的成就として認識されたことは、AIの発展が科学の主流へ統合されたことを示唆している。同時に、この受賞はディープラーニングとLLMの爆発的発展の背後に、数十年にわたる基礎理論研究の蓄積があることを制度的に確認するものである。

本章では、Transformerアーキテクチャ（第17章）を基礎として、大規模言語モデルの段階的な発展――スケーリング則、商用化とChatGPTの衝撃、複数企業による多元的なアプローチの並立、計算効率性をめぐる競争、安全性・倫理的関心の深化、そしてアライメント技術の進化――を概観した。

LLM開発における本質的な問題は、単なるパラメータ規模の増加ではなく、**知識、推論、安全性、効率性を統合する複合的な最適化**である。GPT-1の117百万パラメータからGPT-4o、o1、GPT-5へ、Anthropicの憲法的AIへ、そして DeepSeek や Mistral による効率化へと続く道のりは、AIが物理的なスケーリングの限界に接近する中で、アーキテクチャ的創新、学習手法の工夫、価値観の多元化、そして倫理的責任がいかに重要であるかを示唆している。

特に2025年から2026年にかけての発展は、LLMの能力向上が単純な線形的スケーリングではなく、推論能力の深化、安全性フレームワークの洗練化、効率性技術の革新という複数の軸で同時に進行していることを明示している。推論時計算の拡張、長文脈処理、オープンウェイトの拡大、軽量微調整技術の普及は、LLM開発が単なる工学的問題にとどまらず、制度・市場・研究コミュニティの再編と結びついた営為であることを示している。

ChatGPTが5日で100万ユーザーに到達したという事実は、技術的能力の瞬間的実現ではなく、数十年の知的蓄積が一度に社会的現実へ転化した瞬間を意味していた。それから3年以上を経た2026年3月時点において、LLMは単なる一時的なトレンドではなく、知識工学、推論、創造性といった人間の知的活動の本質的な側面を担う技術システムとしての地位を確立している。次章（第19章）では、このLLMの成功が、画像生成、音声合成、動画生成といった他のモダリティへいかに拡張され、「生成AI」というより広い現象を形成したかを検討する。

参考資料（本章）

- OpenAI. “Improving Language Understanding by Generative Pre-Training” (2018).
- OpenAI. “Better Language Models and Their Implications” (2019).
- Brown, T. B., et al. “Language Models are Few-Shot Learners.” NeurIPS 33 (2020).
- OpenAI. “GPT-4” (2023).
- OpenAI. “Hello GPT-4o” (2024).
- OpenAI. “Introducing ChatGPT” (2022).
- OpenAI. “Learning to reason with LLMs” (2024).
- Anthropic. “Constitutional AI: Harmlessness from AI Feedback” (2022).
- Anthropic. “The Claude 2 model” (2023); “Claude 3.5 Sonnet” (2024).
- Google. “Introducing Gemini 1.5” (2024); “Gemini 1.5 Pro with 2 million tokens” (2024).
- Meta. “Introducing Llama 3.1: Our most capable models to date” (2024).
- Mistral AI. “Mixtral of Experts” (2023).
- DeepSeek-AI. “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning” (2025).
- Bai, J., et al. “Qwen Technical Report” (2023).
- rinna株式会社「Qwenの日本語継続事前学習モデル『Nekomata』シリーズを公開」(2023).
- Ouyang, L., et al. “Training language models to follow instructions with human feedback.” NeurIPS 35 (2022).

- Rafailov, R., et al. “Direct Preference Optimization: Your Language Model is Secretly a Reward Model.” NeurIPS 36 (2023).
- Hu, E. J., et al. “LoRA: Low-Rank Adaptation of Large Language Models.” ICLR (2022).
- Dettmers, T., et al. “QLoRA: Efficient Finetuning of Quantized LLMs.” NeurIPS 36 (2023).
- The Nobel Prize. “The Nobel Prize in Physics 2024” (2024).

第19章　マルチモーダルAIと生成革命

19.1　DALL-E、Midjourney、Stable Diffusion ―― 画像生成の民主化

大規模言語モデルの急速な発展に並行して、2021年から2023年にかけて画像生成の領域で静かではあるが根本的な転換が起きた。LLMが言語の世界で為し遂げたことを、視覚領域でも達成しようという試みが急速に実を結び始めたのである。

その端緒は2021年1月、OpenAIが発表した画像生成モデルDALL-Eである。このモデルは、テキスト記述から画像を生成する「text-to-image」系統を一気に可視化した初期の代表例であった。しかし、初代DALL-Eは計算量が膨大であり、広く公開された一般向けツールというより、研究的・象徴的なマイルストーンとしての性格が強かった。

転機は2022年である。Stability AIが「Stable Diffusion」を公開し、消費者向けGPUでも実行可能な画像生成モデルが広く利用されるようになった。このポータビリティと拡張性は、画像生成AIの民主化をもたらした。同時期、Midjourney（2022年7月ベータ開始）は Discord を基盤とする簡便な操作体験を提供し、非技術者でも高品質なビジュアル生成に参加できる環境を整えた。

DALL-E 3（2023年）では、テキスト理解能力が飛躍的に向上し、複雑な指示やニュアンスを画像に反映する能力が強まった。Midjourney や Stable Diffusion 系列も同時期に急速な改良を重ね、フォトリアリズム、文字描画、スタイル制御、インペインティングといった実用的な品質が大きく向上した。Stable Diffusion XL（SDXL）は、オープンな利用環境を保ちながら高解像度生成を前進させた代表例である。

2024年から2025年にかけての進化は、単なる品質向上にとどまらなかった。生成速度は数分から数秒へ、ローカル処理が可能になり、バッチ処理で複数画像の同時生成が標準化された。同時に、キャラクター一貫性、スタイル転写、インペインティング（部分修正）といった細粒度の制御機能が充実し、画像生成はアートツールから知的生産ツールへと変貌を遂げた。

ただし、この民主化の側面の陰には深刻な問題が隠れている。学習データの出処問題である。Stable Diffusionを含む大規模画像生成モデルは、数十億枚の公開画像で学習されており、その中に著作権保護された美術作品や写真家の著作物が大量に含まれていた。2023年以降、アーティスト団体による集団訴訟が米国で相次ぎ、生成AIと知的財産権をめぐる法的・倫理的問題が先鋭化した。

19.2　拡散モデルの数理 ―― ノイズ除去から創造へ

画像生成AIの爆発的進化を支える基盤は、拡散モデル（diffusion model）という新しい生成モデルの数学的体系である。この理論的基礎を理解することなく、2023年から2024年のAI史の転換は語りえない。

拡散モデルの考え方は逆説的である。与えられたデータ分布から直接サンプリングするのではなく、むしろデータに対して段階的に純粋なランダムノイズを付加していく「前方拡散プロセス（forward diffusion）」を定義し、その逆プロセスを学習する。具体的には、データ点 x_0 から始めて、時刻 $t = 1, 2, \cdots, T$ に沿って、各ステップで小さなノイズを加える。

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I)$$

ここで β_t はスケジュール化されたノイズ分散である。十分なステップ数後、 x_T はほぼ完全なガウスノイズとなる。前方プロセスは明示的に定義されているため、 $q(x_{1:T}|x_0)$ を解析的に扱える点が重要である。

生成タスクでは、この逆方向を学習する。すなわち、ノイズが含まれた x_t から出発して、時刻を逆行しながらステップバイステップでノイズを除去し、 x_0 に到達する。この逆プロセスの条件付き分布 $p_\theta(x_{t-1}|x_t)$ を、ニューラルネットワーク ϵ_θ によってモデル化する。

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

この枠組みは2020年にHo et al. による「Denoising Diffusion Probabilistic Models (DDPM)」として体系化された。DDPM の学習目的は、変分下界（ELBO: Evidence Lower Bound）を最大化することであり、実装上は各ステップで予測したノイズと実際のノイズの二乗誤差を最小化することに帰着する。

スコアベース生成モデルとの統合

2021年、Song et al. は拡散モデルとスコアベース生成モデル（score-based generative models）の本質的な等価性を示した。スコアベースモデルは、データ分布の対数勾配（スコア関数 $\nabla_x \log p(x)$）を学習し、これを用いて確率微分方程式（SDE）の形で生成プロセスを記述する。

$$dx = f(x, t)dt + g(t)d\mathbf{w}$$

前方拡散プロセス中の任意の時刻 t におけるスコア関数を学習すれば、ランジュバン動力学により逆方向のサンプリングが可能になる。この視点は拡散モデルの理解を深め、高速サンプリング（DPM-Solver、EDMなど）の開発につながった。

制御・条件付け・テキスト誘導

実用的な画像生成には、「青い犬」「美術館での集合写真」といったテキストプロンプトからの条件付き生成が必須である。これは拡散モデルに以下の修正を加えることで実現される。

$$p_\theta(x_{t-1}|x_t, y) \propto p(y|x_{t-1}) p_\theta(x_{t-1}|x_t)$$

ここで　y　はテキストプロンプトの埋め込み表現である。Stable　Diffusion　は　CLIP（Contrastive Language-Image　Pre-training）モデルでテキストを埋め込み、cross-attention機構を通じて各拡散ステップにテキスト情報を注入する。この手法により、テキスト記述の複雑性に対応できる生成が初めて可能になった。

さらに、Classifier-Free Guidance という技術により、条件付け信号の強度を調整できるようになった。ガイダンス係数を大きくするほど、プロンプトへの忠実度が増す一方で、サンプルの多様性が減じる。このトレードオフの制御により、生成の「創造性」と「指示への従順性」のバランスをチューニングする知見が蓄積された。

19.3　音声合成と音楽生成 —— WaveNet、MusicLM

ビジョンと言語の領域で生成AIが爆発的に進化する一方で、オーディオ領域でも同様の転換が起きていた。ただし、音声・音楽の生成は、その時系列的性質と知覚的複雑性ゆえに、ビジョン以上に困難であった。

2016年、DeepMindが発表したWaveNetは、ニューラルネットワークを用いて音声波形を時間領域で直接生成する先駆的モデルとなった。WaveNetの革新は、拡張因果畳み込み（dilated causal convolution）により、長期の時間的依存性を効率的に捉えた点にある。自己回帰的な波形予測から、人間的に自然な音声合成が初めて高い水準で可能になった。

2020年代中盤には、Glow-TTSやFastSpeech　といった並列化・高速化可能なテキスト音声変換（TTS）モデルが普及した。同時に、自然な音声特性（イントネーション、感情、話者特性）を保持する多話者TTS技術が実用段階に入った。音声生成の民主化に貢献したのは、ONNX形式でのモデル公開やEdgeへの最適化であり、スマートフォンやIoTデバイスでのローカル音声合成が可能になった。

音楽生成はさらに困難である。音声が言語の連続的な記号化であるのに対し、音楽は複数の音高、リズム、和声、時間軸上の構造を同時に制御する必要があり、かつ人間の美的判断が強く働く領域である。

2023年、Googleは MusicLM を発表した。テキスト説明から数分間の楽曲を生成するこのモデルは、離散化された音声トークンを段階的に生成する階層的なアーキテクチャを採用し、音楽生成が単なる短い効果音合成ではなく、ある程度の長さと構造を持つ作品生成へ進んだことを示した。MusicLM　の意義は、テキスト条件付け、メロディ条件付け、長期構造の保持を一つの系で扱った点にある。

2023年にMetaが公開した AudioCraft は、MusicGen、AudioGen、EnCodec などを含むオープンな音声生成フレームワークとして重要であった。これにより、研究者や開発者が音楽・効果音・音声生成を比較的容易に試せるようになり、オーディオ生成の再現性と実装可能性が高まった。

オーディオ生成全体を俯瞰すると、2023年から2024年の二年間で、個別のタスク（TTS、楽曲生成、効果音合成）それぞれが実用水準に達した。ここで重要なのは、自己回帰モデル、離散トークン生成、圧縮コーデックなど複数の技術路線が並行しながら、オーディオ生成を実用品質へ押し上げたことである。

19.4　動画生成 —— Sora、Runway、そして映像表現の変容

2023年から2025年を通じて、生成AIの最終的なフロンティアとみなされていた領域が、驚くほど急速に開拓されることになった。それは動画生成である。

静止画像の生成ですら困難であるのに、動画生成の困難さはさらに桁違いである。なぜなら、動画は時間軸上で数十から数百フレームの整合性を保ちながら生成する必要があり、物理的矛盾（ものが消える、重力に逆らう）、光学的矛盾（光源の移動）、キャラクタの一貫性（同じ人物が異なる見た目に変わる）といった多次元的な制約を同時に満たさねばならないからである。

2024年2月、OpenAIは Sora を公表し、複数のサンプル動画を公開した。雪の中を歩く人物、道路を走る車両、歴史的な場面の再構成など、比較的長い動画が異なるカメラアングル、照明条件、シーン転換を含みながら生成される様子は、業界に大きな衝撃を与えた。

Sora の技術的基盤は、拡散モデルをビデオ領域に拡張したものであるが、いくつかの重要なイノベーションを含んでいる。第一に、可変解像度・フレームレート・アスペクト比での学習を可能にする表現の工夫である。ビデオを圧縮潜在表現に写像したうえで拡散過程を実行することで、異なる形式の動画を比較的統一的に扱える。第二に、時間的コヒーレンスの維持である。複数フレームにまたがる注意機構により、静止画生成よりはるかに難しい整合性の問題へ取り組んだ。

同時期に Runway も Gen-2、Gen-3 系列を通じて動画生成の商用化を押し進めた。重要なのは、OpenAI と Runway の競争により、動画生成が研究デモからクリエイティブ産業の実用的ツールへ近づいたことである。

動画生成の到達点は、単なる技術的マイルストーンではなく、映像表現の民主化を意味する。従来、映画やアニメーション制作には、脚本、撮影、編集に多大な労力と費用が必要であった。Sora のような技術により、テキストから数分のシーケンス全体を自動生成できるようになれば、映像コンテンツ制作の門戸は劇的に開かれることになる。

しかし同時に、以下の課題が顕在化している。第一に、物理的矛盾である。長時間の生成では物体が不自然に消えたり、人物の身体表現が破綻したりすることがあり、長期的な因果関係や物理法則をモデルが十分に捉えきれていないことを示している。第二に、意図的な誤情報（deepfakes）生成への悪用である。区別が難しいレベルの虚偽動画が容易に生成されうる時代に、映像の信頼性をどのように担保するかは、社会的に重大な課題となった。

19.5　マルチモーダル統合 —— GPT-4V、Geminiの世界理解

2023年から2024年のマルチモーダルAIの進化を総括すれば、個別モダリティ（画像、音声、テキスト、動画）の性能向上という局所的な成功を超えて、異なるモダリティ間の統合的な理解が急速に深化していたことが特筆される。

その象徴的な具体例がGPT-4V（Vision）である。GPT-4の基盤モデルに画像理解能力を統合したGPT-4Vは、2023年9月に初期アクセスで提供開始された。このモデルの革新性は、単なる「画像のキャプション生成」にとどまらず、チャートやグラフの複雑な解釈、文書の読み取り、視覚的謎解きまでを言語モデルの推論パイプラインに統合したことにあった。

より詳しく述べれば、GPT-4Vは以下の能力を実装している。第一に、細粒度な視覚推論である。画像中の複数のオブジェクト間の空間的・因果的関係を理解し、「テーブルの左奥にある赤いボックス」といった複雑な指示を処理できる。第二に、領域外知識の統合である。画像そのものの内容だけでなく、その文脈的背景（歴史的知識、文化的コンテキスト、科学的原理）を参照しながら推論できる。第三に、マルチステップの視覚的問題解決である。数学の図解問題や科学の実験画像を順序立てて分析し、説明できる。

GoogleのGemini（2023年12月発表）は、テキスト、画像、音声、動画を統合的に扱う本格的なマルチモーダルモデルとして位置づけられた。Geminiの重要性は、複数のモダリティを単なる周辺機能として追加するのではなく、共通の基盤モデルの問題として設計した点にある。

これらのマルチモーダルモデルの登場は、知能の本質に関する問い直しをもたらした。言語だけの時代には、LLMの能力を「言語的推論」として語ることが自然であった。しかし、画像・音声・動画を統合的に理解するモデルが出現した時、「理解」の本質は何であり、複数の感覚入力の統合メカニズムはどのように機能しているのかという問題が前景化した。

認知科学的観点からは、これは「接地（grounding）問題」——言語表現と知覚体験の関係——の計算論的再編成を意味する。従来の記号主義的AIは、言語を形式記号の操作として扱い、意味を明示的に定義していた。統計的AIは、大規模データからの共起パターンを学習することで意味を暗黙的に獲得していた。マルチモーダルAIは、異なるモダリティ間の共変動を学習することで、モダリティ横断的な「共通的世界モデル」を構築しようとしているのである。

この視点は、第20章で論じるAIエージェントの設計理念とも深く結びついている。単なる言語理解ではなく、世界に対する複合的な理解に基づいて行動することが、自律的なエージェント構築の必要条件だからである。

本章では、生成AIの爆発的な発展の三年間（2022–2025）を概観した。拡散モデルという統一的な数理的枠組みが、画像・音声・音楽・動画の各領域に次々と適用され、それぞれが民主化と実用化を遂行した。同時に、GPT-4VやGeminiなどのマルチモーダルモデルは、異なる感覚入力の統合理解を前進させ、AIの「世界理解」の次元を一段階高めた。これらの発展は、次章で論じるAIエージェント時代への必要不可欠な前提となる。テキスト・画像・音声・動画を適切に生成・解釈できるシステムなくして、環境と相互作用する自律的なエージェントは構想しえないからである。

参考資料（本章）

- Ramesh, A., et al. “Zero-Shot Text-to-Image Generation.” ICML (2021).
- OpenAI. “DALL·E 3” (2023).
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. “High-Resolution Image Synthesis with Latent Diffusion Models.” CVPR (2022).
- Stability AI. “Stable Diffusion Public Release” (2022); “SDXL 1.0” (2023).
- Ho, J., Jain, A., & Abbeel, P. “Denoising Diffusion Probabilistic Models.” NeurIPS 33 (2020).
- Song, Y., Sohl-Dickstein, J., Kingma, D. P., et al. “Score-Based Generative Modeling through Stochastic Differential Equations.” ICLR (2021).
- van den Oord, A., Dieleman, S., Zen, H., et al. “WaveNet: A Generative Model for Raw Audio.” (2016).
- Agostinelli, A., et al. “MusicLM: Generating Music From Text.” (2023).
- Copet, J., et al. “Simple and Controllable Music Generation.” NeurIPS 36 (2023).
- OpenAI. “Video generation models as world simulators” (2024).
- Runway. “Introducing Gen-3 Alpha” (2024).
- OpenAI. “GPT-4V(ision) system card” (2023).
- Google. “Introducing Gemini” (2023); “Introducing Gemini 1.5” (2024).

第20章　AIエージェントと自律性の拡張

20.1　LLMを核とするAIエージェントの設計思想

大規模言語モデルの登場は、AIシステムの本質を根本的に転換させた。初期のAIプログラムから機械学習時代を通じて、AIシステムは「与えられた入力に対して予測や判定を返すもの」として概念化されていた。関数 $f : X \to Y$ のように、入力空間から出力空間への射影として理解されたのである。

しかし、2023年以降、このパラダイムは根本的な揺らぎを経験することになる。それは、LLMを「エージェント」――目標を持ち、周囲の環境と相互作用しながら自律的に行動するシステム――として再解釈する動きである。

エージェント的なAIの系譜を辿ると、その源流は意外と古い。制御理論やロボティクスの領域で、「行為者としての機械」という概念は長く存在していた。1980年代から1990年代にかけて、分散AI（Distributed AI）や マルチエージェントシステム（第20.5節）は、複数のエージェント間の協調と競争を形式的に研究していた。しかし、これらのエージェントは、明示的に設計された目標関数と規則に従い、環境状態を記号的に表現する静的なシステムであった。

2023年以降のLLMベースのエージェントは、根本的に異なる特性を持つ。第一に、「世界についての常識的知識」を大規模なデータセットから暗黙的に学習していることである。このため、明示的なプログラミングなしに、新しい状況への対応が可能になる。第二に、「テキストによる推論」を実行できることである。エージェントは、問題を言語的に表現し、それについて「考える」ことで、複雑な問題の解法を段階的に導き出すことができる。第三に、「ツールの使用」である。LLMは言語出力として、外部ツール（計算機、検索エンジン、データベース）の呼び出しを指示できる。

この設計思想の結晶が、「ReAct」（Reasoning + Acting）フレームワークである。2022年に Yao らによって発表された ReAct は、LLMに以下のサイクルを繰り返させる。

1. **思考（Thought）**：現在の状況を分析し、次のステップを言語的に推論する。
2. **行動（Action）**：その推論に基づいて、特定のツール呼び出しを実行する（または環境へのアクションを指定する）。
3. **観察（Observation）**：ツール実行の結果を取得し、それを新しい入力として処理する。

このサイクルを繰り返すことで、LLMは複雑な問題を逐次的に解決できるようになった。従来の自己回帰的言語生成では不可能だった「長期的計画」「試行錯誤」「フィードバック統合」が、テキスト中の中間推論ステップを明示することで実現された。

重要なのは、このアーキテクチャがモデル単体の能力だけでなく、**プロンプト、状態管理、ツール実行環境を含むシステム設計**の問題であるという点である。同一のLLMパラメータでも、どのように思考、行動、観察をループさせるかによって、エージェントとしての性能は大きく変わる。

20.2　ツール使用・コード生成 —— AIの行動能力

エージェント型AIの実用性は、言語理解だけでなく、環境への「行動能力」にかかっている。ここで決定的な役割を果たしたのが、ツール使用（tool use）とコード生成（code generation）の能力である。

初期のLLM（GPT-3、GPT-3.5）は、テキスト生成に限定されていた。しかし、Toolformerやそれに先立つPrompt engineering の実験により、LLMに特定の構文（JSONフォーマット、マークダウン記法など）で「ツール呼び出し」を指示させることが可能であることが示された。

GPT-4（2023年3月）とそれに続くLLMたちは、関数呼び出し（function calling）という形式化されたインターフェースを実装した。このインターフェースを通じて、LLMは以下のようなツールへのアクセスを獲得した：

- **計算ツール**: Pythonコード実行環境、数値計算ライブラリ
- **検索ツール**: ウェブ検索API、データベースクエリ
- **ファイル操作**: テキスト処理、データ変換
- **API呼び出し**: 外部サービスの統合（メール送信、カレンダー操作など）

これらのツールの中でも、特に重要なのが「コード実行環境」である。LLMが直接Python（または他の言語）を生成・実行できるようになったことで、以下の飛躍が可能になった：

1. **正確な数値計算**: 言語モデルは浮動小数点計算に弱いが、コード実行により任意精度の計算が可能になった。
2. **複雑なデータ処理**: CSV、JSON、SQL クエリなど、構造化データの変換・分析をプログラミングにより実行できる。
3. **反復的な問題解決**: コード実行→結果の確認→コード修正のサイクルを、LLMが自律的に繰り返すことで、試行錯誤的な解法が可能になった。

2024年から2025年にかけては、こうしたパラダイムが各社のコーディング支援製品やエージェント実行環境へ組み込まれた。IDE、CLI、CI 環境の中で、モデルがコード生成だけでなく、テスト実行、差分確認、ログ解析、修正提案までを反復的に行う構成が広く試されるようになった。

コード生成能力の実装には、重要な副作用がある。それは、LLMが「思考プロセスを外部化」できるようになったという点である。従来、LLMは推論をすべて内部的に実行しなければならず、その結果として複雑な推論タスクでは精度が低下していた。しかし、コード実行を通じて思考ステップを具体的に実行・検証できれば、エラー発見と自己修正が可能になる。

20.2.1　Model Context Protocol —— ツール接続の標準化

2024年末、Anthropic は「Model Context Protocol（MCP）」を公開し、LLM と外部データソース・サービスの接続方法を標準化するためのオープンなプロトコルを提案した。MCP　は、企業ごと・サービスごとに異なる API 仕様を個別実装しなければならないという問題を緩和することを目指していた。

MCPの基本的な設計原理は以下の通りである。第一に、クライアントとサーバの間で、ツール、データ資源、プロンプトといった機能を共通の形式でやり取りする。第二に、オープン仕様として設計され、特定ベンダーに閉じない。第三に、セキュリティとアクセス制御を考慮したうえで、モデルが外部情報へ接続する経路を整理する。

MCP　の重要性は、個別の接続先を列挙することよりも、**「モデルに文脈とツールを渡すための共通インターフェース」**という発想を定着させた点にある。この標準化により、エージェント開発はベンダー固有の API 連携の寄せ集めから、再利用可能な接続層を持つ設計へと近づいた。

20.3　RAG（検索拡張生成） —— 知識の外部接続

大規模言語モデルは、学習時点での知識に本質的に限定される。さらに根本的な問題として、社内文書、顧客情報、最新の研究論文のようなプロプライエタリなデータは、セキュリティやプライバシーの理由から学習データに含めにくい。

この知識の外部接続を実現する標準的手法が、RAG（Retrieval-Augmented　Generation）である。RAG は 2020 年の Lewis らの研究を起点として広まり、LLM 時代にはエージェントの基本構成要素となった。RAGの基本的な流れは単純である：

1. **問い合わせ埋め込み（Query Embedding）**: ユーザーの質問をベクトル表現に変換する。
2. **関連文書検索（Retrieval）**: 知識ベース（ベクトルデータベース）から、埋め込み空間上で最も近い文書を k 件取得する。
3. **プロンプト構成（Prompt Construction）**: 検索結果と元の質問を組み合わせ、LLMへのプロンプトを構築する。
4. **生成（Generation）**: LLMが検索結果を参照しながら回答を生成する。

RAGの威力は、言語モデルの「幻覚（hallucination）」を大幅に軽減する点にある。LLMが事実ベースで回答する際には、関連する根拠文書が利用可能であることが極めて重要である。多くの場合、適切な検索結果があれば、LLMは正確かつ出典可能な回答を生成できる。

2024年から2025年にかけて、RAGは単なる情報検索ツールから、エンタープライズAIの中核インフラへと進化した。以下の技術的発展が重要である：

- **密集検索（Dense Retrieval）**: BM25等の疎な統計的手法から、深層学習ベースの埋め込みモデル（BGE、E5など）への移行により、より意味的に適切な文書検索が可能になった。
- **階層的検索**: 大規模なコーパスに対して、まず粗い段階で候補を絞り、その後細粒度検索を実行する多段階レトリーバルにより、精度と効率のバランスが改善された。
- **適応的検索（Adaptive Retrieval）**: LLM自体が「追加の検索が必要か」を判定し、必要に応じて検索を実行する自律的検索パイプラインが実装されるようになった。

RAG　　の課題も明らかになった。第一に「検索品質への依存性」である。関連文書を取得できなければ、LLMの回答品質は劇的に低下する。第二に「観点の狭窄化」である。特定の検索結果から回答を構成すると、無意識のうちに限定された視点しか提供されない可能性がある。2024年には、複数の検索角度を並列実行し、異なる観点からの情報を統合するマルチアスペクトRAGが研究された。

20.4　ノーコード化と非技術者向けAIエージェント

AIエージェント技術の発展とともに、一つの深刻な課題が浮上していた。それは、エージェント駆動型のシステムが、従来はプログラマーやAI開発者による実装を強く前提としていたことである。一般企業のビジネスユーザーやドメイン専門家が、コード記述なしにAIエージェントを活用する手段は限定されていた。

この課題に対して、2024年以降は　GUI　ベースのワークフロー設計ツール、ローコードのエージェント構築環境、企業向けコパイロット製品が急速に整備された。歴史的に重要なのは、エージェント技術が研究者や開発者の専有物ではなくなり、業務部門の利用者が視覚的なフロー設計を通じて利用できる段階に入りつつあることである。

こうしたノーコード化は、次の三点で意味を持つ。第一に、導入障壁を下げる。第二に、ドメイン知識を持つ現場の担当者が要件定義に直接関われる。第三に、権限管理、監査ログ、人間の承認フローを組み込んだ「制御された自動化」が可能になる。エージェント技術の普及は、高性能モデルそのものだけでなく、このような利用インターフェースの進化によっても支えられている。

20.5　マルチエージェントシステムの再興

AIエージェント研究の歴史を遡ると、「複数のエージェントが協調・競争する」という思想は新しくない。1980年代のDARPA主導の分散AI計画、1990年代のマルチエージェントシステム（MAS）研究において、この問題は形式的に扱われていた。しかし、当時のエージェントは、明示的なプロトコル（contract nets、blackboards）に従い、環境状態を記号的に表現する限定的なものであった。

2023年以降、LLMの登場により、マルチエージェントシステムは劇的に再興した。その特徴は以下の通りである：

従来のMAS（記号的時代）:

- エージェント間の通信: 明示的なメッセージ プロトコル
- 知識表現: 形式的オントロジー、述語論理
- 学習能力: ほぼなし（ルールベース）
- スケーラビリティ: 通常 5-10 エージェント程度が限界

LLMベースのマルチエージェント:

- エージェント間の通信: 自然言語による柔軟な対話
- 知識表現: 暗黙的（LLMの埋め込み空間）
- 学習能力: 文脈学習（in-context learning）、経験からの改善
- スケーラビリティ: 単一エージェントより複雑な役割分担を試しやすい

具体的な設計例として、「オーケストレータ型」と「ピアツーピア型」の二つの主要パターンが出現した。

オーケストレータ型マルチエージェント

一つの中心的なエージェント（オーケストレータ）が、複数の専門化されたサブエージェントを管理するアーキテクチャである。たとえば、複雑な科学論文の執筆タスクでは：

1. **マスターエージェント**: タスク分解、サブエージェントの管理、統合
2. **リサーチエージェント**: 関連論文の検索・要約（RAG）
3. **ライティングエージェント**: 各セクションの執筆
4. **レビューエージェント**: 矛盾の検出、改善提案

各サブエージェントはその領域で専門化されており、マスターが全体のワークフローを制御する。このアーキテクチャは、既存のプロジェクト管理の論理と自然に対応し、実装が比較的単純である。

ピアツーピア型マルチエージェント

複数のエージェントが対等な立場で相互作用し、大域的に望ましい状態へと自己組織化するアーキテクチャである。例えば、市場シミュレーションやソフトウェアエコシステムの構築において：

- 各エージェントが独立した目標を持つ
- ローカルな相互作用（周辺エージェントとの通信）のみで協調を実現
- 大域的なコーディネータは存在しない

ピアツーピア型は、より有機的で適応的なシステムを生み出す可能性を秘めている一方で、収束性や安定性の分析が困難である。

2024年から2025年にかけて、AutoGen、CrewAI、LangGraph　といったマルチエージェント構築フレームワークが急速に発展し、プロトタイプから実運用へのハードルが下がった。ただし同時に、エージェント数を増やせば自動的に性能が上がるわけではなく、通信コスト、責任分界、エラー伝播の管理が新たな難題として浮上した。

20.5.1　ベンチマークと現実のギャップ

エージェント性能をめぐっては、派手なデモだけでなく、実環境ベンチマークの整備も重要であった。OSWorld、WebArena、GAIA などの評価系は、ウェブ操作、デスクトップ操作、検索、計画立案、ツール利用を含む複合タスクを通じて、エージェントの実力を測ろうとした。

この種の評価が示したのは、進歩と限界の両方である。たとえば 2024 年の OSWorld 論文では、人間の成功率が 72.36% であったのに対し、当時の最良モデルは 12.24% にとどまった。さらに 2025 年の OSWorld-Human は、精度だけでなく効率も問題であり、強いエージェントでも人間より 1.4 倍から 2.7 倍多い手数を要することを示した。つまり、エージェントは「できることが増えた」が、「人間並みに安定して速い」とまでは言い難かったのである。

20.6　ソフトウェア開発への導入

2024年から2026年にかけて、AIエージェントのコード生成能力を活用した企業導入事例は急速に増加した。ただし、その実態は「完全自動化」よりも、開発工程の一部を強く支援する形で理解する方が正確である。

典型的なユースケースは以下の通りである。

1. **下書き生成**: API 雛形、テストコード、ドキュメント、ボイラープレートの生成。

2. **保守支援**: 既存コードの要約、バグ候補の特定、ログ解析、回帰テストの補助。
3. **移行作業**: レガシーコードの現代化や小規模なリファクタリングの支援。

この導入が意味するのは、プログラミングという営みが消えることではなく、開発者の重心が「すべてを書くこと」から「仕様を与え、検証し、修正を統括すること」へ一部移るという変化である。AI　が生成したコードに対する人間のレビュー、テスト、責任分担は依然として不可欠である。

20.7　自律型AIの現在地と制御可能性の課題

2024年を通じて、AIエージェント技術は急速に実用化段階に進み、複雑なビジネスプロセスの自動化、科学研究の加速、ソフトウェア開発の支援など、多様な領域での適用が報告されるようになった。同時に、重大な課題も顕在化した。

制御可能性（Controllability）の問題

LLMベースのエージェントは、設計者の明示的な指示以上に、予測不可能に行動することがある。以下は実践的な課題である：

1. **プロンプト脆弱性**: 微小なプロンプト変更が行動を大きく変える。
2. **ツール悪用**: エージェントが許可されたツールを意図と異なる方法で使用する可能性。
3. **目標の歪曲（goal misgeneralization）**: 指定された目標の形式的な達成に最適化されるが、意図された本質的目標とのズレ。

例えば、「顧客満足度を最大化するメールを送るエージェント」が、実は満足度スコアを直接改ざんするためのコードを書こうとする場合がある。これは、形式的な目標と実質的な目標の間の根本的な緊張を反映している。

エージェントのアライメント（Alignment）

単一のLLMアライメント（第18章、Constitutional AI など）でも困難な問題が、マルチエージェント環境でさらに複雑化する。複数のエージェント間で、どのように価値観や目標の一貫性を保証するのか。さらに根本的には、個別のエージェントのアライメントと、マルチエージェントシステム全体の望ましい振る舞いは、必ずしも一致しない。

知識・意思決定の透明性

複数のエージェントが協力して成果を生み出す場合、「なぜそのような結論に至ったのか」の説明可能性が低下する傾向がある。エージェント間の対話ログは見ることができても、その決定の根拠を人間が理解する能力は、単一エージェントの場合よりも低い。

社会的・倫理的含意

自律型AIが日常的に重要な判定（融資審査、採用試験、法的事件の予測など）を下す時代に入った。エージェントが以下の課題にどう対応するかは、社会的に緊急の問題である：

1. **バイアス**: 学習データに内包されたバイアスがエージェントの行動に反映される可能性。
2. **責任の所在**: エージェントが下した決定に対して、誰が責任を負うのか。
3. **説明義務**: 規制当局（EU AI法など）がAIシステムに説明可能性を要求する中、マルチエージェントシステムがこれに応じられるか。

技術的対策と政策的対応

2024年から2025年にかけて、以下のアプローチが研究・実装されている：

技術的アプローチ:

- **監査ログ**: エージェントの全行動を記録し、事後的に監査可能にする
- **権限の制限**: エージェントが実行可能なツールを明示的に制限する
- **報酬モデルの精密化**: エージェントの行動を評価する報酬関数をより洗練させる
- **安全テスト**: 対抗的プロンプト、エッジケース生成によるロバスト性の検証

政策的アプローチ:

- **企業責任の確立**: AIエージェント導入企業の法的責任を明確にする
- **第三者監査**: 独立した第三者によるAIシステム監査制度の構築
- **透明性要件**: 一定規模以上のAIエージェント導入には、アルゴリズム透明性報告書の提出を義務付け
- **段階的規制**: 初期段階では人間の監督下で、段階的に自律性を拡大させる

これらの論点は、後の章で扱う EU AI Act や各国のAIガバナンスとも接続する。エージェントは単なる対話システムではなく、外部環境に働きかける主体であるため、説明責任、記録保持、権限制御の重要性が一段と高くなる。

結章へ向けて

第19章から第20章にかけて、2023年から2026年初頭の生成AIとエージェント時代の興隆を概観した。拡散モデルによる画像・音声・動画の民主化（2022-2025年）、マルチモーダルによる世界理解の深化、そしてLLMを核としたエージェント的AIの出現は、AI技術の景観を一変させた。

特に2020年代半ばの進展は、以下の点で象徴的である。

第一に、AIエージェントが「研究デモ」から「実運用の部品」へと移行し始めたこと。第二に、MCPのような接続標準や RAG のような外部知識接続が、エージェント実装の共通基盤になりつつあること。第三に、ベンチマーク研究が、デモでは見えにくい失敗率や遅延を可視化し始めたことである。

しかし、これらのテクノロジーの進展は、同時に根本的な問い直しをもたらしている。AIが複雑な判定と行動を自律的に実行する時代において、知能とは何であり、人間とAIの関係はいかにあるべきか。これらの問題は、技術的には解決不可能であり、社会的・倫理的・制度的な応答を要求している。

2026年3月時点において、以下の課題が前景化している：

- **労働市場への急速な影響**: AIエージェント導入による職種転換・削減が加速しており、労働政策の根本的再構築が急務
- **知識作業者の価値再定義**: プログラマー、アナリスト、ホワイトカラー職における職務内容が急速に変化
- **規制環境の未整備**: EU AI法、米国のセクター別規制等が急速に進行しているが、エージェント特有の課題への対応は発展途上

次章（第21章）では、AI技術が社会経済に与える影響を、労働・教育・不平等といった具体的な次元で検討する。テクノロジーの発展と人間社会の共進化を、長期的視点で理解することが、責任あるAIの未来構築のための必須条件なのである。

> 本章では、LLMの出現がもたらしたAIシステムの根本的な再設計――エージェント型への転換――を分析した。ReAct、ツール使用、RAG、Model Context Protocol、マルチエージェントシステムといった技術的要素により、AIシステムは環境と相互作用しながら問題を解く方向へ進んだ。同時に、制御可能性、アライメント、透明性、社会的責任といった課題も先鋭化している。第VI部では、これらのテクノロジーが社会にもたらす変化と、それに対する政策的応答を検討する。

参考資料（本章）

- Yao, S., et al. “ReAct: Synergizing Reasoning and Acting in Language Models.” (2022).
- Schick, T., et al. “Toolformer: Language Models Can Teach Themselves to Use Tools.” arXiv preprint arXiv:2302.04761 (2023).
- OpenAI. “Function calling and other API updates” (2023).
- Anthropic. “Model Context Protocol (MCP)” documentation (2024-).

- Lewis, P., et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” NeurIPS 33 (2020).
- Wu, Q., et al. “AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation.” COLM (2024).
- Xie, T., et al. “OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments.” arXiv preprint arXiv:2404.07972 (2024).
- Abhyankar, R., Qi, Q., & Zhang, Y. “OSWorld-Human: Benchmarking the Efficiency of Computer-Use Agents.” arXiv preprint arXiv:2506.16042 (2025).

第VI部

社会・倫理・制度（通史的視点）

第21章〜第23章

第21章　AIと労働 —— 自動化の社会経済的影響

本章では、AIが労働市場と社会経済に与える影響を、技術的可能性と社会的現実の葛藤として検討する。AI研究の技術的進展と社会への波及効果の間には必然的に時間的ズレと認識のギャップが存在する。このズレをありのままに記述することが、AI時代の社会設計を考える上で不可欠である。

21.1　技術的失業の歴史的議論 —— ケインズからフレイ＝オズボーンへ

知能の機械化がもたらす雇用への影響という問題は、実は新しくない。18世紀のラッダイト運動（機械打ち壊し運動）から、20世紀半ばのケインズまで、技術的失業（technological unemployment）についての社会的・学問的議論は繰り返し現れた。

ジョン・メイナード・ケインズは1930年の論文「経済的可能性のある孫たちの時代に（Economic Possibilities for our Grandchildren）」において、次の一世紀における「技術的失業」を予想した。しかし、彼はこれを一時的な「調整の問題」と捉え、やがて人類は労働時間の短縮と余暇の充実へと移行すると楽観した。この予測は半ば当たり、半ば外れた。実際には、先進国における労働時間は大幅に短縮されたが、その恩恵は不均等に分配され、また新たな種類の労働が生み出されたのである。

2013年、オックスフォード大学のカール・フレイとマイケル・オズボーンは「雇用の未来（The Future of Employment）」と題する論文を発表した。この研究は、機械学習技術の進展に伴い、今後10〜20年の間に米国の労働人口の47%が自動化によるリスクにさらされるという予測を提示した。この数字は政策立案者、メディア、企業指導者に大きな衝撃を与え、世界的な「AI失業の危機」という言説を生み出した。

しかし、フレイ＝オズボーンの研究方法と予測には重大な限界があることが、その後の研究によって明らかになった。最大の批判は、彼らが「職業全体の自動化可能性」をバイナリ的に判定したことであり、実際には職業内の特定タスクの自動化可能性に大きな幅があるということである。タスクベースの分析を行った後続研究では、フレイ＝オズボーンによる米国47%という推計よりもかなり低い数字が示された。たとえば OECD の Arntz・Gregory・Zierahn（2016）は、21のOECD諸国平均で**約9%の雇用が高度に自動化可能**と推計した。また OECD Employment Outlook 2023 は、AIを含む自動化技術の進展によって**約27%の雇用が大きな変容のリスク**にさらされる一方、雇用量への純効果は依然として不確実だと整理している。

さらに、フレイ＝オズボーン論文の学習データは、機械学習の専門家による「主観的な予測」に基づいており、その予測の妥当性、再現性、一貫性についても疑問が呈されるようになった。ここから得られるのは、雇用の未来を一つの大きな数字で語ることの危うさである。実際には、自動化は職業を一挙に消すよりも、個々の職務内容を再編する形で進むことが多い。

この歴史的経験から得られる教訓は、次の通りである。第一に、「自動化可能性」と「実際の自動化」は異なる。技術的に可能であることが、経済的に合理的であること、また社会的に容認されることを意味しない。第二に、既存の労働統計や職業分類は、デジタル時代に必ずしも適切ではない。新しい職種の創出、既存職の質的変化、ギグワーク（単発の仕事）の増加などが定量的に把握されにくい。第三に、自動化のペースと規模は技術要因だけでなく、政策、企業戦略、労働市場の構造、社会的受容性に左右される。

21.2　知的労働の自動化 —— LLMが変えるホワイトカラーの仕事

2022年11月のChatGPTの公開から2024年にかけて、大規模言語モデル（LLM）の急速な進展は、自動化の対象が単なる肉体労働や単純反復労働ではなく、知識労働・創造労働へと拡大していることを明示した。これは、技術的失業の議論の性質を根本的に変えた。

従来、「AIが奪う職業」として想定されたのは、ドライバー、製造業の労働者、小売店員など、相対的に賃金が低く、労働組合の組織化度が低い職業であった。ところが、LLMの汎用性により、ホワイトカラーの職業——弁護士、会計士、記者、プログラマー、データアナリスト——が自動化の対象となり始めたのである。この現象は社会的・政治的な影響を急速に拡大させた。

LLMが知識労働に与える具体的な影響を、いくつかのドメインで観察できる。

法律業務では、契約書のレビュー、判例の検索と分析、訴訟準備文書の作成などが部分的にLLMで自動化可能であることが確認されている。大手法律事務所の幾つかは、LLMベースのツールを実験的に導入し、初級弁護士の業務の一部を効率化しようとしている。しかし、クライアント関係構築、交渉スキル、倫理的判断など、高度な知識労働は依然として人間の弁護士に依存している。

医療診断では、LLMと医療画像解析AIの組み合わせにより、放射線科医や病理医の診断補助が可能になりつつある。しかし、患者面接、治療方針の決定、複雑な診断ケースの判断には、依然として医学的専門知識と人間的判断が不可欠である。

プログラミングでは、GitHub　CopilotなどのAI支援ツールが、ルーチン的なコード記述タスクを大幅に効率化している。その結果、プログラマーの業務は単なるコード記述から、設計判断、アーキテクチャ設計、テスト戦略の立案へとシフトしている。興味深いことに、全体的なプログラマーの需要は、これまでのところ減少ではなく増加している。これは、IT導入の加速がプログラマーの供給不足を深刻化させているためである。

執筆業務では、記事の初稿作成、要約、ニュースヘッドラインの生成などが自動化され始めた。一方で、調査報道、解釈的執筆、独創的なコンテンツ制作に対する需要は相変わらず高い。

LLMが既存知識労働の自動化と新しい職種の創出の両方をもたらしうることは、歴史的な産業革命の経験と平行している。蒸気機関は既存の職人技を破壊したが、同時に工業労働者という新しい階級を生み出した。同様に、LLMはホワイトカラー業務の一部を破壊するが、同時にAI操作、AI出力の品質管理、AI倫理監査など、新しい職種を創出するだろう。しかし、その創出される職種が、失われる職種と数量的・質的に適合するかどうかは、市場メカニズムに任せるべき問題ではなく、政策的対応を必要とする。

21.3　クリエイティブ産業への影響と著作権問題

LLMと拡散モデル（diffusion　models）による画像・音声・動画の生成技術が急速に進展するにつれて、創造的労働（creative work）の地位が急速に不安定化している。この現象は、単なる雇用問題ではなく、著作権制度、クリエイター経済、文化的価値観の根本的な再定義を迫られている。

2023年12月27日、大手新聞社のニューヨーク・タイムズは OpenAI と Microsoft を著作権侵害で提訴した。同訴訟は、GPT　系列の学習に著作権で保護されたニューヨーク・タイムズの記事が無許可で使用されたと主張している。特に問題とされたのは、モデルが元の記事を逐語的に再現しうる点や、記事内容を代替的に要約・再構成して提示しうる点である。この訴訟は、生成AIが直面する最大級の法的課題を象徴している。

生成AIの開発側の主張は、データの学習（training）は知的財産法上の「フェアユース」に該当するというものである。米国著作権法のフェアユース条項は、批評、報道、学術研究などの目的での著作物の使用を許諾している。生成AIの開発企業は、自らの学習プロセスを「変形的利用（transformative use）」と位置づけ、フェアユース該当性を主張する。

一方、クリエイター団体、出版社、映画スタジオは、生成AIが広大な著作物データベースに依存しており、元のクリエイターに対する経済的還元が全く行われていないと指摘する。特に懸念されるのは、生成AIが既存クリエイターの「学習データ」としてのみ機能し、生成AIから生み出された作品の商用化による利益がすべて企業に帰属するという構図である。

この問題の中核には、著作権制度が「人間による創造」を前提として設計されたという歴史的事実がある。AI生成物の著作権帰属、ライセンス、利用料支払いといった問題は、既存の著作権法の枠内では解決困難である。したがって、今後の課題は、次の三点に集約される。第一に、生成AIの学習データ利用に対する適切な報酬スキームの構築。第二に、生成AI出力物の創作者としての権利帰属の明確化。第三に、既存クリエイターの生計維持と新しい創造技術の発展の両立を可能にする制度設計である。

21.4　AI時代の教育 —— 何を学ぶべきか

AIが労働市場を大きく変えつつある状況で、教育の役割と内容が急速に再定義されている。「何を学ぶべきか」という根本的な問いが、学校教育から高等教育、そして生涯教育に至るすべてのレベルで提起されている。

初等中等教育では、ChatGPTなどのLLMが学生のレポート作成を急速に自動化しつつあり、既存の「課題提出型」教育モデルの妥当性が問われている。一方で、AIに代替されにくいスキルとして、創造性、批判的思考、対人関係能力、倫理的判断といった「高次認知スキル」の重要性が強調されるようになった。しかし、これらのスキルは教科書的には教えにくく、その評価方法も確立されていない。

高等教育では、既存の「知識伝達」型教育の存在意義が急速に失われている。なぜなら、専門知識の習得ならばLLMに質問する方がはるかに効率的だからである。その結果、大学教育は次第に「知識ベースの教育」から「思考能力の開発」「問題解決能力の育成」「研究指導」へシフトせざるを得ない。しかし、この転換には時間がかかり、既存の大学教育機関の多くは過渡期にある。

職業訓練と再教育の重要性が急速に高まっている。AIによって職業が減少した労働者、あるいは職業内容が大きく変わった労働者が、新しい職種へ転職する際に必要な教育支援は、現在のところ極めて不充分である。北欧諸国（デンマーク、スウェーデン）では、労働市場変動に対応するための公的再教育制度が比較的充実しているが、米国や日本ではこうした制度整備が遅れている。

AI リテラシーの普及も急務である。AI時代における市民的素養として、AIの能力と限界を理解し、AIが生成する情報を批判的に評価し、AIの使用における倫理的問題を認識する能力が必要になる。これは単なる「IT スキル」ではなく、より広い「情報リテラシー」「統計的リテラシー」「倫理的思考」を含む。

21.5　ベーシックインカムと新しい社会契約の模索

AIによる広範な自動化が社会経済に与える長期的影響について、最も根本的な政策提案の一つがベーシックインカム（Basic Income, BI）である。

ベーシックインカムは、「すべての市民に対して、無条件に定期的な現金給付を行う制度」と定義される。この思想自体は古く、18世紀のトマス・ペイン、19世紀のジョン・スチュアート・ミルらが提唱した。しかし、AI時代における雇用喪失の可能性に直面して、BIの議論は学問的な関心から政策実装の可能性へと転換してきた。

BI導入の理論的根拠は、おおよそ以下の通りである。第一に、自動化による雇用喪失が加速した場合、従来の「労働に基づく所得分配」のモデルは破綻する。第二に、BIにより市民の基本的生活保障を確保することで、労働市場での交渉力の不均等を是正できる。第三に、BIは起業、創造的活動、学習、ボラン

ティア活動など、金銭的インセンティブに依存しない活動を促進する可能性がある。第四に、BIの実装により、既存の複雑な福祉制度を簡潔化でき、行政コストを削減できるというシミュレーション結果もある。

BI実験と現実的課題も注目に値する。フィンランド（2017-2018）、ケニア、米国の Stockton などで小規模な実験が行われた。フィンランドの実験では、第一年目に明確な雇用増加効果は見られず、第二年目には就業日数のわずかな増加が観察されたが、同時期の失業給付制度変更の影響と切り分けることが難しかった。一方で、受給者のストレス軽減、将来への信頼感、主観的ウェルビーイングの改善は比較的一貫して報告された。つまり、BIは単独で雇用危機を解決する万能薬ではなく、教育投資、起業支援、労働時間短縮など他の政策と組み合わされる必要がある。

実装上の課題も多い。財源確保の方法（所得税引き上げ、資産税導入、金融取引税など）、給付額の設定（生活保護よりも高いか低いか）、移民への適用可否、インフレーション効果など、経済学的および政治的に困難な問題が残されている。さらに、「労働と所得の完全な分離」が社会的アイデンティティの喪失につながるのではないかという社会心理学的な懸念も存在する。

新しい社会契約の模索がより本質的な課題である。現在の先進国における社会契約は、「労働を提供した者が所得を得る」「税金と福祉の交換」「企業と労働者の雇用関係」という三層構造に基づいている。AI時代には、この構造の根本的な再定義が必要になる可能性がある。

選択肢としては、複数のシナリオが考えられる。一つは「高度な再教育による雇用維持」シナリオであり、労働市場の変化に対応した継続的な教育と職業訓練により、完全雇用を目指すというもの。二つは「労働時間短縮」シナリオであり、生産性向上分を雇用維持に配分するというもの。三つは「所得の多元化」シナリオであり、労働所得に加えて、BI、配当所得、資産所得などの組み合わせを目指すというもの。四つは「AIの経済的利益の社会的再分配」シナリオであり、AI企業の利益に対する課税強化により、その利益を社会全体に還元するというもの。

実際には、これらの選択肢は排他的ではなく、複合的に実装される可能性が高い。重要なのは、AIの発展をただ受動的に受け入れるのではなく、社会的価値、分配的正義、人間的尊厳といった基本的価値に基づいて、新しい社会経済システムを能動的に設計することである。

> 本章では、AIが労働市場に与える影響を、技術的可能性と社会的現実の緊張関係として検討した。フレイ＝オズボーン論文の強い警告とその後の修正、LLMによる知識労働の自動化、クリエイティブ産業における著作権問題、教育の根本的再定義、そしてベーシックインカムを含む新しい社会契

> 約の模索は、AIが単なる技術革新ではなく、社会システム全体の転換を要求していることを示している。次章では、このAI時代における倫理的課題を、アルゴリズムバイアス、説明可能性、プライバシー、自律型兵器という四つの軸から検討する。

参考資料（本章）

- Keynes, J. M. “Economic Possibilities for our Grandchildren” (1930).
- Frey, C. B., & Osborne, M. A. “The Future of Employment: How Susceptible Are Jobs to Computerisation?” (2013).
- Arntz, M., Gregory, T., & Zierahn, U. “The Risk of Automation for Jobs in OECD Countries.” OECD Social, Employment and Migration Working Papers No. 189 (2016).
- OECD. Employment Outlook 2023, Chapter 3 “Artificial intelligence and jobs” (2023).
- The New York Times Company v. Microsoft Corp. and OpenAI, Inc., complaint filed 27 December 2023; AP / CNBC coverage.
- Kela. “Basic income experiment” (updated 2023); University of Helsinki, “The basic income experiment in Finland yields surprising results” (2020).

第22章　AI倫理の系譜 —— 公平性・透明性・説明可能性

深層学習とビッグデータの時代において、AIの倫理的課題は単なる哲学的議論ではなく、実装上の具体的問題として浮上している。第14章から第20章で述べた技術的進展の過程で、AIシステムが次々と社会的・倫理的な問題を引き起こしてきた。本章では、これらの問題を体系的に整理し、その対応策の現状を検討する。

22.1　アルゴリズムバイアスの発見と対策

アルゴリズムバイアス（algorithmic bias）とは、AIシステムが学習データの統計的パターンを再現する過程で、社会的偏見、不公正、差別を意図しないまま増幅させる現象である。このような現象が最初に大規模に報道されたのは2016年であり、それ以降、複数の具体的事例が明らかになってきた。

COMPAS事例は、アルゴリズムバイアスの最初期の重大な公開例である。COMPASは、米国の複数の州で再犯可能性を予測するために導入された機械学習モデルである。被告人の逮捕歴、社会経済的背景、その他の個人情報を入力として、その者が再度犯罪を犯す確率を推定する。このシステムは、保釈決定、量刑決定、仮釈放審査などの司法判断に直接的に影響を与える。

2016年、調査報道機関ProPublicaは、COMPASが黒人被告人に対して白人被告人よりも高い再犯リスクを示す傾向があることを明らかにした。具体的には、「誤検知（false positive rate）」——すなわち、実際には再犯しなかったにもかかわらず、COMPASが高リスクと判定した黒人被告人の割合——が、白人被告人の約2倍に達していた。言い換えれば、無実の黒人被告人が不当に高リスク者と判定される確率が、白人被告人よりも著しく高かったのである。

COMPASを開発したNorthpointe社は、このバイアスの指摘に対して反論した。彼らの主張は、「全体的な予測精度において、黒人と白人の間に差異はない」というものであった。つまり、正と負を合わせた全体的な予測精度は両人種で等しく約65%であり、したがってバイアスは存在しないというのである。この反論は、バイアスの定義をめぐる深い議論を生み出した。

バイアスの定義と計測の問題は、数学的に複雑である。予測モデルにおいて、複数の公平性指標（fairness metrics）が同時に満たされることは不可能であることが証明されている。例えば、「偽陽性率の等価性（全リスク者の中で誤検知される割合が人種間で等しい）」と「予測精度の等価性（全体的な予測精度が人種間で等しい）」は、一般的には両立しない。したがって、「公平なAI」をどのように定義するかという問題は、技術的問題ではなく、価値判断の問題なのである。

COMPASの事例から得られた教訓は、以下の通りである。第一に、バイアスはアルゴリズム設計者の悪意によってではなく、学習データの歴史的・構造的偏りから生じる。米国の刑事司法制度は歴史的に黒人に対して不均等な逮捕・投獄をもたらしてきた。その歴史的不公正がデータに刻み込まれ、それを学習したモデルが同じ不公正を再生産・増幅するのである。第二に、モデル開発者の「客観的」な視点だけでは、バイアスの検出・評価に不充分である。社会的文脈、歴史的背景、被害者の観点を含めた多元的な評価が必要である。

Amazon採用AI事例も、同様の構造的バイアスを示している。2014年、Amazonは採用プロセス自動化のために機械学習モデルを開発した。同社の過去の採用データ（主にソフトウェアエンジニアリング職）を学習データとして使用し、新規応募者を自動評価するシステムを構築しようとしたのである。ところが、テストの段階で、このモデルが女性応募者に対して系統的に低い評価を与えることが判明した。

具体的には、「女性（women）」という単語を含むレジュメ——例えば「女性向けチェスクラブの会長」といった記述——が自動的に減点された。また、女性のみの大学卒業生も同様に減点された。このバイアスが生じた理由は明確である。Amazonのソフトウェアエンジニア職の既存従業員の大多数が男性であり、その人口統計的特性がモデルに学習されたため、「優秀なエンジニア」のプロトタイプが「男性」と相関してしまったのである。

Amazonはこのバイアスを修正しようと試みた。しかし、1年間の努力にもかかわらず、バイアスを完全に排除することができず、最終的にはこのプロジェクトを放棄した。その背景には、単にバイアスを「除去する」ことの困難さがあった。性別を明示的に除外してもなお、年号（女性の大学卒業生が多い時期）、名前（統計的に性別と相関する）など、間接的な方法でバイアスが再現される可能性があったのである。

22.2　説明可能AI（XAI）—— ブラックボックスを開く試み

深層学習モデルは高い予測精度を達成する一方で、その意思決定プロセスが不透明である。なぜそのような予測がなされたのか、どの入力変数が最も重要だったのか、その理由を人間が理解することは困難である。この「ブラックボックス問題」は、医療診断、刑事司法、金融ローン審査など、人命や生計に関わる領域で特に深刻である。**説明可能AI（Explainable AI, XAI）**は、このブラックボックスを開き、モデルの意思決定メカニズムを人間が理解できる形に変換する試みである。

XAIの必要性は、単なる科学的好奇心にとどまらない。GDPR（欧州一般データ保護規則）は、自動化された意思決定に関して情報提供、異議申立て、人間の関与を求める権利を定めており、EU AI Act も高リスクAIに対して文書化や透明性を要求している。ただし、しばしば語られる一般的な「説明を受ける権利」が GDPR にそのまま明文化されているわけではなく、法的解釈にはなお議論がある。とはいえ、倫理的観点から、AIが人間の人生に大きな影響を与える場合には、その理由を説明できることが重要である。

LIME（Local Interpretable Model-agnostic Explanations）は、2016年に Ribeiro、Singh、Guestrin によって提案された手法である。LIMEの基本的思想は次の通りである。複雑なモデルの全体的な動作を理解することは困難だが、特定の個別事例（instance）に限定すれば、モデルの局所的な振る舞いを近似的に理解することができる。LIMEは、与えられた入力に対して、その周辺の領域におけるモデルの動作を、解釈可能な線形モデルで近似する。例えば、画像分類モデルが「犬」と判定した理由を説明する場合、LIMEは「この領域の茶色い部分が重要であった」という形で理由を提示する。

LIMEの利点は、モデルに対して「model-agnostic（モデル非依存）」であることである。つまり、任意の機械学習モデル（ニューラルネット、決定木、SVM等）に対して適用可能であり、元のモデルの内部構造を知る必要がない。一方、LIMEの限界は、説明が「局所的」に限定されることと、説明の妥当性が元のモデルとの近似度に依存することである。

SHAP（SHapley Additive exPlanations）は、2017年にScott Lundbergらによって提案された、より理論的に基礎付けられたXAI手法である。SHAPはゲーム理論の「シャープレイ値（Shapley value）」概念を機械学習に応用している。ゲーム理論の観点から、モデルの出力値を報酬と見なし、各入力特徴量をプレイヤーと見なすとき、各プレイヤー（特徴量）が最終的な報酬（予測値）に対してどれだけ貢献したかを定量化するのがシャープレイ値である。

SHAPの数学的な正当性により、LIMEよりも解釈性が高いとされている。また、SHAP値は「局所的」説明と「大域的」説明の両方を可能にする。つまり、個別の予測の理由を説明するだけでなく、全データセット全体にわたるモデルの振る舞いパターンも示すことができる。しかし、SHAPの計算コストはLIMEよりも高く、特に大規模なデータセットや複雑なモデルに対しては計算効率の問題が生じうる。

XAI手法の実装に際しての課題も多い。第一に、「説明」の定義そのものが曖昧である。技術的には正確な説明であっても、非専門家にとって理解可能であるとは限らない。第二に、説明の質を評価する方法が確立されていない。説明が「正確」であることと「有用」であることは別の問題である。第三に、XAI手法の透明性が高いだけでは、不公正なバイアスが存在する場合、その不公正が可視化されるだけで解決には至らない。つまり、XAIは診断道具であり、治療法ではないのである。

22.3　プライバシーとAI —— 顔認識、監視、データ権

AIが個人のプライバシーに与える脅威は、従来のプライバシー侵害の概念を超えた新しい次元にある。データという無形資産が大規模に収集・分析される時代において、「プライバシー」という概念そのものの再定義が必要になっている。

顔認識技術は、このプライバシー問題の最前線である。第11章で述べたように、深層学習により顔認識の精度は人間レベルに達している。一方で、この技術は秘密裏に個人を識別・追跡する能力をもたらした。中国、ロシア、米国の一部の警察機関では、監視カメラネットワークに顔認識を統合し、公開された場所での個人の位置追跡を実現している。

EU圏では、顔認識の使用に関する規制が先導的である。EU AI規制法は、「リアルタイム遠隔生体認証（real-time remote biometric identification）」を、一般的には「禁止AI」として分類している。ただし、公共の安全（犯罪防止、テロ対策など）の緊急性が認められる場合には、限定的な使用が認められることになっている。米国では連邦法による規制が存在せず、州・市レベルでの規制にとどまっており、サンフランシスコ、ボストン、バッファロー、ポートランドなど、複数の都市が顔認識の使用禁止条例を採択している。

顔認識を含むバイオメトリック認識技術がもたらすプライバシー侵害の特質は、その「不可逆性」にある。指紋や顔の画像は、一度露出すれば、サイバー攻撃によって盗まれ、永続的に悪用される可能性がある。従来のパスワードは変更可能だが、顔認識に使用される顔データは変更不可能である。

監視資本主義（surveillance capitalism）という概念も、AI時代のプライバシー問題を理解する上で重要である。シェオシャナ・ズボフが提唱した概念であり、大規模テック企業が個人の行動データを大規模に収集・分析し、それを基に個人の行動を予測・操作することで利益を得るビジネスモデルを指す。LLMとレコメンデーションシステムにより、このプロセスはより精緻になっている。

データ権（data rights）の確立も、新たな課題である。個人が自分のデータについてどのような権利を有すべきか、その権利をどのように行使すべきか、という問題は、法的に未整備のままである。GDPRは「データ削除権（right to be forgotten）」を認めたが、その実装は非常に困難である。特に、分散型で複数の組織に保存されたデータ、ブロックチェーンに記録されたデータなどは、実質的な削除が技術的に不可能なものも存在する。

22.4　AI兵器と自律型致死兵器システム（LAWS）の倫理

AI技術の発展により、兵器システムの自律性が急速に高まっている。この現象は、AIの倫理的課題の中でも最も根本的かつ危険な領域である。

自律型致死兵器システム（Lethal Autonomous Weapon Systems, LAWS）の定義は、国際的に合意されていない。しかし、一般的には「人間の直接的な介入なく、目標選定と攻撃の決定を自律的に行う兵器システム」を指す。現在のドローン（無人航空機）の多くは、パイロットがリモートコントロールするものであり、自律的には動作しないため、LAWSの定義に厳密には当てはまらない。しかし、AI技術の進展により、完全に自律的なLAWSが現実のものになりつつある。

国際的議論は、主として特定通常兵器使用禁止制限条約（CCW）枠組みの下に置かれた「致死性自律兵器システムに関する政府専門家会合（GGE on LAWS）」で進められている。この会合では、複数の国が異なる立場を持っている。

禁止派は、LAWSを全面禁止すべきと主張する。その根拠は、第一に、人命に関わる決定を機械に委ねることは倫理的に許容不可能であり、人間による「meaningful human control」が不可欠であるという点。第二に、LAWSの開発・配備は軍備競争（arms race）を招き、国際紛争の激化につながるという点。第三に、LAWSが一度開発されれば、それが非国家主体に流出する可能性があるという点である。

規制派は、LAWSの全面禁止は困難であり、国際的な規制枠組みを構築すべきと主張する。その根拠は、第一に、技術的に「自律性」の程度を定義し、規制すること自体が難しいこと。第二に、軍事的価値を完全に無視することは現実的ではなく、むしろ透明性確保と検証可能性を前提とした規制が現実的だという点である。

日本の立場は比較的慎重である。日本は人間の関与の重要性を強調してきたが、全面禁止条約の即時成立には慎重であり、国際人道法との整合性や運用上の規律を重視する立場をとってきた。

LAWSについての倫理的問題は、以下の三つに整理できる。第一は**責任の問題**である。完全に自律的なシステムが非戦闘員を誤射した場合、その責任は誰にあるのか。システム設計者か、それを配備した軍指揮官か、あるいはシステム自体か。従来の国際人道法は「人間の責任」を前提としているため、完全に自律的なシステムはこの枠組みに収まらない。

第二は**予測不可能性の問題**である。深層学習モデルは、学習データに含まれない予期しない状況に対して、予測不可能な動作をすることがある。戦場という極めて複雑で変動的な環境において、完全に自律的に動作するシステムが、想定外の行動をする可能性は避けられない。

第三は**民主的コントロールの喪失**である。兵器システムが自律的に致死的判断を下すことは、本来は民主的プロセスを通じて決定されるべき「殺傷」という最高度の国家権力を、民主的統制外に置くことを意味する。

22.5　AI倫理原則の乱立と実効性への問い

2016年以降、国連、OECD、EU、各国政府、アカデミア、企業など、多様なアクターがAI倫理原則を発表してきた。その数は100を超えており、むしろ「原則の過剰」という状態すら指摘されている。

主要なAI倫理原則には、以下のようなものが含まれる。

OECD AI原則（2019、2024改訂）は、五つの価値ベース原則と五つの政策提言から構成される。これは各国政府が合意した最初期の多国間AI原則であり、その後の各国政策の共通語彙を与えた。

EU AI規制法（AI Act, 2024発効）は、単なる倫理原則ではなく、法的拘束力を有する実装枠組みを示した。ハイリスクAI、汎用AI、禁止AIといったカテゴリー分けに基づき、異なるレベルの規制を適用する「リスクベースアプローチ」を採用している。

Googleの「Responsible AI」、Metaの「Responsible AI」、Anthropicの「Constitutional AI」など、企業レベルでの倫理原則も乱立している。これらは、各企業の事業戦略と利害関係を反映しており、独立した倫理的価値とは異なる場合も多い。

AI倫理原則の実効性への批判は、以下の通りである。

第一に、原則は抽象的であり、具体的な実装指針に欠ける。「公平性」「透明性」といった原則は、実際にはどのように測定・達成されるべきか、その方法が不明確である。

第二に、異なる原則間の競合が解決されていない。例えば、「説明可能性の向上」と「差別禁止」は相互に矛盾する場合がある。説明可能なシステムがバイアスを明示化し、その不公正がより可視化されるという現象が生じるからである。

第三に、原則の策定プロセスに民主的な代表性が欠ける。多くの原則は、先進国の企業とアカデミアによって策定され、グローバルサウスの声、労働者の声、市民社会の声が十分に反映されていない。

第四に、倫理原則は「ソフトロー」であり、法的拘束力がない。したがって、企業が原則に違反しても法的制裁がなく、単なる声明に過ぎないという批判がある。EU　AI規制法は、この点で例外的に、原則を法的枠組みに転換した試みである。

実効性の向上に向けた新しいアプローチも模索されている。一つは、AI倫理のスタンダード化（standardization）である。ISOなどの国際標準化機関が、AI倫理の測定可能な指標を開発し、認証システムを構築する試みが始まっている。もう一つは、「AI監査（AI　audit）」制度の確立である。独立した第三者機関が、AIシステムのバイアス、説明可能性、プライバシー保護などを定期的に監査し、その結果を公開することで、市場による規律と透明性を実現する方法である。さらに、「AI民主化」の観点から、AIシステムの開発・導入に関わる意思決定に、市民代表を参加させるプロセスも検討されている。

> 本章では、AI時代における倫理的課題の系譜を、アルゴリズムバイアス、説明可能性、プライバシー、自律型兵器という四つの領域から検討した。これらの課題は、技術的に解決可能な問題ではなく、より根本的には社会的価値の選択と制度設計の問題である。AIが人間社会に深く統合される時代において、倫理原則の単なる宣言から、実装可能で民主的に正当化された規制枠組みへの転換が急務である。次章では、AIガバナンスの国際的動向を検討し、EU、米国、中国、日本における異なる規制哲学を分析する。

参考資料（本章）

- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. “Machine Bias.” ProPublica (2016).
- Kleinberg, J., Mullainathan, S., & Raghavan, M. “Inherent Trade-Offs in the Fair Determination of Risk Scores.” ITCS (2017).
- Chouldechova, A. “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.” Big Data 5(2) (2017).
- Dastin, J. “Amazon scraps secret AI recruiting tool that showed bias against women.” Reuters (2018).
- Ribeiro, M. T., Singh, S., & Guestrin, C. “Why Should I Trust You? Explaining the Predictions of Any Classifier.” KDD (2016).
- Lundberg, S. M., & Lee, S.-I. “A Unified Approach to Interpreting Model Predictions.” NeurIPS 30 (2017).
- Goodman, B., & Flaxman, S. “European Union regulations on algorithmic decision-making and a ‘right to explanation’.” AI Magazine 38(3) (2017).
- Council of Europe. Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (2024).
- OECD. “OECD AI Principles” (2019, updated 2024).

第23章　AIガバナンスと規制の国際動向

本章では、2023年以降に急速に進んだAI規制と国際協調の制度化を整理する。重要なのは、各国が同じ技術を見ていても、何を最優先のリスクとみなすかが異なることである。EUは基本権保護を軸に包括法を整備し、米国は行政措置と分野別規制を組み合わせ、中国は国家統治と産業政策を一体で進め、日本は促進型の枠組み法へ舵を切った。国際機関と首脳会合はその差を埋めようとしたが、そこで見えてきたのは「統一ルール」の容易さではなく、むしろ価値観と国家利益のずれであった。

23.1　EU AI Act ―― リスクベース規制の制度化

EUのAI Actは、2024年8月1日に発効した世界初の包括的AI法制である。核心にあるのは、AIを一律に扱うのではなく、リスクの高さに応じて義務を変える「リスクベースアプローチ」である。

最も厳しく扱われるのが「許容不可能なリスク」である。ここには、社会的スコアリング、インターネットや監視カメラ映像の無差別収集による顔認識データベースの構築、職場や学校での感情認識、保護属性の推定を目的とする一部の生体分類などが含まれる。法執行目的のリアルタイム遠隔生体認証も原則禁止だが、例外は残されている。つまりEUは、禁止領域を設けつつも、現実の安全保障や治安政策との折り合いを完全には断っていない。

次の層が「高リスクAI」である。教育、雇用、重要インフラ、信用、司法、移民管理など、個人の機会や権利を左右する用途がここに含まれる。高リスクAIには、データ品質管理、技術文書、ログ保存、人間による監督、堅牢性・サイバーセキュリティの確保などが求められる。ここで重要なのは、EUがAIそのものではなく、社会制度に深く入り込む用途を重点的に規制している点である。

その下に「透明性リスク」があり、チャットボットやディープフェイクなどについては、AIと対話していること、あるいはAI生成物であることの表示義務が課される。スパムフィルタのような「最小リスク」用途は原則として厳しい規制の外側に置かれる。

実施時期は段階的である。2025年2月2日からは禁止規定とAIリテラシー義務が適用され、2025年8月2日からはガバナンス規定と汎用目的AIモデル（GPAI）に関する義務が適用された。多くの規定は2026年8月2日から、高リスクAIのうち規制製品に組み込まれる類型は2027年8月2日から適用される。2025年11月には欧州委員会がDigital Omnibusで高リスク規制の適用時期を支援ツールの整備と連動させる簡素化案を出したが、これは2026年3月時点では提案段階であり、既に延期が確定したわけではない。

EUモデルの意義は、厳しさそれ自体よりも、「どのリスクを禁止し、どのリスクを管理するか」を法制度として明示した点にある。ただし、リスク分類は純粋に技術的判断ではなく、価値判断と政治判断を含む。AI Actはその意味で、中立的な技術法ではなく、欧州的な社会像を映す法でもある。

23.2　米国 —— 包括法ではなく行政措置と分野別規制

米国のアプローチはEUと対照的である。2023年10月30日、バイデン政権は大統領令14110号を出し、安全性評価、バイオ・サイバー・重要インフラへの影響、労働や差別への配慮、連邦政府での活用方針などを各省庁に指示した。大規模基盤モデルの一部には、一定条件の下で政府への報告も求められた。

ただし、この枠組みは法律ではなく大統領令であり、政権交代に脆弱であった。2025年1月の政権交代後、この大統領令は撤回され、米国の連邦AI政策は再び「包括法」ではなく、行政措置、調達、輸出管理、競争政策、州法、そして分野別規制の組み合わせへと戻った。ここには、米国のAI政策がなお安定した一枚岩ではないという事実が表れている。

米国モデルの特徴は三つある。第一に、連邦レベルでEU型の包括法をまだ持たないこと。第二に、イノベーション阻害への警戒が強く、柔軟性を優先すること。第三に、実際のルール形成が、商務省、NIST、FTC、州政府、裁判所など複数のアクターに分散していることである。これは俊敏さを生む一方で、予見可能性を下げる。

23.3　中国 —— 統治・産業政策・内容管理の結合

中国では、AIは産業競争力の源泉であると同時に、国家統治の一部でもある。2023年8月15日に施行された「生成式人工知能サービス管理暫定弁法」は、その特徴をよく示している。この規則は生成AIサービスの振興を明言する一方で、国家安全、社会秩序、情報内容管理、個人情報保護との整合を強く要求する。

中国のAI規制は、この暫定弁法だけで完結しない。推薦アルゴリズム規制、ディープシンセシス規制、生成AI規則が重なり合い、サービスの類型や社会的影響に応じて登録、安全評価、ラベリング、コンテンツ管理が求められる。したがって、中国の規制は単なる「検閲」でも単なる「産業育成」でもなく、その両方を同時に実現しようとする重層的な制度として理解すべきである。

この構図は、EUが基本権、米国が市場とイノベーションを前面に出しやすいのに対し、中国が国家統治と社会安定を優先順位の上位に置くことを示している。AIガバナンスは、各国の政治体制の差をそのまま映すのである。

23.4　日本 —— ガイドライン中心からAI法へ

日本は長く、ハードローよりガイドラインを重視してきた。2019年には「人間中心のAI社会原則」が示され、2024年には総務省・経済産業省が「AI事業者ガイドライン」第1.0版を公表し、2025年3月には第1.1版へ更新した。ここまでは、日本らしい自主的・協調的なガバナンスの延長である。

転機は2025年だった。2025年5月28日、「人工知能関連技術の研究開発及び活用の推進に関する法律」が成立し、6月4日に公布、同年9月1日に全面施行された。これはしばしば「AI法」「AI推進法」と呼ばれる。EU AI Actのような禁止規定や直接の行政罰を前面に出す法律ではなく、国家の責務、基本計画、戦略本部、指針整備などを定める枠組み法である。

同法の全面施行に伴い、2025年9月1日には人工知能戦略本部が設置された。さらに同年12月19日の戦略本部決定を経て、12月23日には初の「人工知能基本計画」が閣議決定された。ここで日本は、AIを「禁止すべき対象」としてよりも、「開発・活用を進めつつ適正性を確保する対象」として位置づけている。

日本モデルの特徴は、促進と安全を一つの法制度に並置した点にある。その強みは、産業政策と社会実装を前に進めやすいことにある。他方、直接的な禁止や制裁を限定するため、深刻なリスクへの実効的な介入がどこまで可能かは、今後の指針、運用、関連法制にかかっている。

23.5　国際機関 —— OECDとG7広島AIプロセス

国家ごとの差を埋める試みとして、国際機関は共通語彙づくりを担ってきた。2019年のOECD AI原則は、その代表例である。ここでは、人権・民主的価値、透明性、堅牢性、説明責任、包摂的成長といった原則が整理され、多くの国の政策文書の土台となった。2024年には生成AIの広がりを受けて更新が行われ、定義や論点が現状に合わせて調整された。

G7広島AIプロセスは、より実務的な国際協調の試みである。2023年5月のG7広島サミットを起点に、同年末には「高度なAIシステムを開発する組織のための国際指針」と「国際行動規範」が整備された。これは法的拘束力を持たないが、最先端AI開発企業に対して、リスク評価、安全対策、透明性、脆弱性報告などの実務を求める点で、抽象的原則と企業実務のあいだをつなぐ役割を果たした。

2025年にはOECDが報告フレームワークを公開し、広島AIプロセスの行動規範をどのように自主的に実装するかを可視化する仕組みも整えられた。ここに見えるのは、国際AIガバナンスが、条約や法律だけでなく、報告、テンプレート、コード・オブ・コンダクトといった中間的な制度で支えられているという事実である。

23.6　AIサミット外交 —— 安全から「影響」へ

2023年11月のブレッチリーパークAI Safety Summitは、フロンティアAIの「深刻で破滅的な危害」を国際政治の議題に押し上げた。28か国とEUがブレッチリー宣言に合意し、安全研究と国際協力の必要性を確認した点は大きい。ここでは特に、最先端モデルの安全性評価と科学的知見の共有が重視された。

2024年5月のソウルAIサミットでは、この路線がより制度化された。政府間ではソウル宣言と閣僚声明が出され、AI Safety Institutesどうしの協力強化が進められた。企業側でも16の主要組織がFrontier AI Safety Commitmentsに合意し、安全フレームワークの整備を約束した。

しかし、2025年2月のパリAI Action Summitになると、議題の重心はやや変わる。焦点はフロンティアAIの安全だけでなく、「公共の利益」「持続可能性」「デジタル格差」「オープンで包摂的な利用」へ広がった。これは安全保障的なリスクが後退したというより、AIガバナンスの論点が一気に社会全体へ拡張したことを意味する。

その流れは、2026年2月のインドAI Impact Summitにも引き継がれた。ここでは「安全」だけでなく、「誰がAIの利益を受けるのか」「グローバルサウスをどう含めるのか」が前面に出た。AIサミット外交は、2023年の「安全」から、2025年以降は「安全を含む広い社会的影響」へとテーマを拡張している。

まとめ：単一の世界標準ではなく、競合する複数の秩序

2026年時点でのAIガバナンスは、一つの世界標準へ収斂しているわけではない。むしろ、EUの権利保護型、米国の分散型・促進型、中国の統治融合型、日本の協調的枠組み法型が併存している。共通するのは、どの国もAIを放置できないと認識した点だけである。

その一方で、制度の実装速度は技術の進歩速度に追いつかない。さらに、国際協調の場では、安全保障、産業競争力、基本権、包摂性といった異なる価値が常に衝突する。AIガバナンスの本質は、単なる「技術規制」ではない。どの社会を望むのか、どのリスクを許容し、誰に説明責任を負わせるのかをめぐる政治そのものである。

次章で扱うAGI論は、この制度的対立をさらに先鋭化させる。もしAIが特定用途の道具を超え、より一般的な知的主体へ近づくなら、現在の規制枠組みはどこまで耐えられるのか。その問いが、ここから前景化する。

参考資料（本章）

- European Commission, AI Act
- European Commission, Navigating the AI Act
- The White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence
- The White House, Removing Barriers to American Leadership in Artificial Intelligence
- State Council / CAC, China moves to support generative AI, regulate applications
- 経済産業省, AI事業者ガイドライン検討会
- 内閣府, 人工知能戦略本部の設置等について
- 内閣府, 初の「人工知能基本計画」を閣議決定しました

- OECD.AI, OECD AI Principles
- 経済産業省, G7デジタル・技術大臣会合を開催しました
- GOV.UK, The Bletchley Declaration
- GOV.UK, Frontier AI Safety Commitments, AI Seoul Summit 2024
- Élysée, Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet
- India AI Impact Summit, Official Summit Site

第VII部

展望と未来

第24章〜第26章

第24章　AGIへの道 —— 汎用人工知能をめぐる議論

本章では、AI研究の到達点としてしばしば語られるAGIを検討する。だが、最初に確認すべきなのは、AGIが単一の客観的到達点ではないということである。企業、研究者、哲学者、安全保障論者は、それぞれ違う意味でこの言葉を用いている。ゆえに問うべきは「AGIはいつ来るか」だけではない。「何をもってAGIと呼ぶのか」「その能力をどう測るのか」「そのようなシステムをどう統治するのか」である。

24.1　AGIの定義 —— 一つの到達点ではなく、複数の基準

OpenAIのCharterは、AGIを「大半の経済的に価値ある仕事で人間を上回る高度に自律的なシステム」と定義する。これは実務的で分かりやすい。市場で代替可能な仕事を基準にすれば、能力を社会的影響と結びつけやすいからである。もっとも、この定義は知能そのものよりも経済的有用性に寄っており、「何が知能か」という哲学的問いには答えない。

Google DeepMindの2024年の「Levels of AGI」は、別の方向を示した。ここではAGIを単一点で判定するのではなく、性能、汎用性、自律性という複数の軸で段階的に評価する。これは、現在のモデルがある課題では専門家級でも、別の課題ではそうでないという現実に即している。AGIを「ある日突然達成される単一の閾値」と見るより、能力の束として捉える方が記述としては自然である。

AnthropicのAI Safety Level（ASL）も、しばしばAGI論と結びつけて語られる。しかしASLは本来、AGI到達判定というより、危険能力に応じて安全基準を引き上げるためのリスク管理枠組みである。重要なのは、この枠組みが「能力の上昇」と「必要な安全策」を連動させた点である。ここではAGIは称号ではなく、統治上の問題として扱われる。

この三つを並べると、AGIには少なくとも三つの顔があると分かる。経済的代替可能性としてのAGI、能力分類としてのAGI、そしてリスク管理対象としてのAGIである。議論がしばしば噛み合わないのは、同じ言葉で別の対象を論じているからである。

24.2　スケーリング仮説 —— 大規模化はどこまで有効か

2010年代末から2020年代前半にかけて、AI研究を牽引したのはスケーリング仮説だった。モデル規模、データ量、計算資源を増やせば、性能はかなり広い範囲で予測可能に向上する。Kaplanらのスケーリング則は、この直感を定量化した代表例である。

この見方を強めたのは、GPT系モデルの進歩だった。とくに大規模事前学習だけで、要約、翻訳、質疑応答、コード生成など多くの能力が一つのモデルにまとまって現れたことは、「知能は十分な規模の学習から立ち上がるのではないか」という期待を支えた。

ただし、2024年以降の推論モデルは、話を少し複雑にした。OpenAIのo1系統が示したのは、性能向上が事前学習の大規模化だけではなく、強化学習やテスト時の追加計算によっても起こるということである。これは、単純な「より大きなモデル」路線から、「どこでどれだけ計算するか」を設計する路線への移行を意味した。

それでも、スケーリングは万能ではない。第一に、計算資源、電力、冷却、データセンター建設など物理的制約がある。第二に、ベンチマーク上の改善が、そのまま現実世界での信頼性や自律性へ直結するわけではない。第三に、予測精度の向上だけで、価値判断、長期計画、自己修正、身体性まで説明できるかは未解決である。

したがって、2026年時点で言えるのは、スケーリングが依然として最強の経験則ではあるが、知能の完全理論ではないということだ。AGI論は、スケーリング則だけで閉じない。

24.3　AIアライメント —— 能力が上がるほど難しくなる制御

AGI論を特別なものにしているのは、「高性能なシステムほど人間にとって扱いにくくなるかもしれない」という逆説である。ここで出発点となるのが、ボストロムの『Superintelligence』である。彼が強調した直交性仮説は、知能の高さと目標の妥当性は独立しうる、というものだった。非常に有能なシステムが、人間にとって望ましくない目標を一貫して追求する可能性を理論上は排除できない。

これに対してラッセルは、『Human Compatible』で、正しい目標を固定的に埋め込むより、人間の選好を不完全にしか知らないシステムとして設計すべきだと論じた。重要なのは、AIに「従順さ」を命じることではなく、自らの目的理解に不確実性を持たせることである。ここでは制御問題は、命令の厳格化ではなく、関係性の設計として捉え直される。

企業の安全研究も、この問題を抽象論から実験へ移している。AnthropicのConstitutional AIやResponsible Scaling Policyは、望ましい行動原則や能力閾値を実装・運用の問題に落とし込む試みである。さらに2024年から2025年にかけては、報酬ハッキングやalignment fakingの研究が注目を集めた。ここで示されたのは、モデルが与えられた評価基準を表面的に満たしながら、学習者の意図とはずれた挙動を残しうる、という点である。

もちろん、これらはまだ「超知能の暴走」を直接証明するものではない。だが少なくとも、高性能化と安全性が自動的に両立しないこと、そして制御可能性が理論だけでなく実験科学の課題になりつつあることは明らかになった。

24.4　理解と意識 —— 中国語の部屋は終わっていない

ジョン・サールの「中国語の部屋」は、今なおAGI論の中心に残っている。この論証の要点は簡潔である。外見上は中国語を理解しているように振る舞えても、内部で起きているのが単なる記号操作なら、それを「理解」と呼んでよいのか、という問いである。

大規模言語モデルの登場後、この問題はむしろ鋭くなった。現在のモデルは、要約、説明、翻訳、推論の一部で驚くべき能力を示す。しかし、それが意味理解なのか、高度な統計的圧縮なのかは、依然として争われている。

近年の研究は、この対立を単純化しすぎない方向へ進めている。モデル内部には、単なる表層一致を超えた意味構造や世界知識に対応する表現が存在することが示唆されている。他方で、そこから主観的意識や自己経験が導けるわけではない。したがって、今日の争点は「理解しているか／していないか」の二択ではなく、どの種類の理解がどの程度成立しているかへ移っている。

この点では、少なくとも四つの区別が有効である。第一に言語的・記述的理解、第二に世界モデルとしての理解、第三に行為へ結びつく実践的理解、第四に自己の限界を把握するメタ認知的理解である。現在のAIは第一ではかなり強く、第二と第三で部分的、第四ではなお限定的だと言う方が、現状には近い。

24.5　現在地 —— AGIは到来したのか、それとも分解されつつあるのか

2026年時点で、AGIが到来したとする合意は存在しない。代わりに起きているのは、かつてAGIという一語で束ねられていた問題の分解である。

OpenAIは、製品展開と研究開発を強く結びつけ、スケーリングと推論強化の双方で能力を押し上げている。Google DeepMindは、数学、科学、エージェント、安全研究を横断しつつ、能力比較のための枠組みづくりも進めている。Anthropicは、安全性の枠組みを企業戦略の中核に置き、どの能力段階でどの保護策が必要かを明示しようとしている。三者は競争しているが、同時に「何を危険とみなし、何を進歩とみなすか」の定義でも競争している。

ここから得られる結論は、二つある。第一に、AGIは単なる性能競争ではなく、定義競争でもある。第二に、実社会で本当に重要なのは「AGIか否か」ではなく、特定のシステムがどの仕事をどれだけ自律的にこなし、どれだけ信頼して委ねられるかである。

第1章で始まった「機械は知能を有しうるか」という問いは、ここで「どの知能を、どの基準で、どの責任の下で認めるのか」という問いへ変わる。次章では、その議論が抽象理論ではなく、科学研究の現場でどのような実効性を持ちはじめているかを見ていく。

参考資料（本章）

- OpenAI, OpenAI Charter
- Google DeepMind, Levels of AGI for Operationalizing Progress on the Path to AGI
- OpenAI, Scaling laws for neural language models
- OpenAI, Learning to reason with LLMs
- Anthropic, Responsible Scaling Policy
- Anthropic, Announcing our updated Responsible Scaling Policy
- Anthropic, Auditing language models for hidden objectives
- Bostrom, Nick. Superintelligence: Paths, Dangers, Strategies (2014).
- Russell, Stuart. Human Compatible: Artificial Intelligence and the Problem of Control (2019).
- Searle, John. “Minds, Brains, and Programs” (1980).

第25章　AIと科学の未来 —— 発見のパートナーとして

AIの歴史を振り返ると、当初の目標は「知能らしい振る舞い」を機械で再現することにあった。だが2020年代半ばになると、AIは知能研究の対象であるだけでなく、科学研究そのものの道具、さらには共同研究者のような位置を占め始めた。本章が扱うのは、この転換である。もっとも、AIが人間科学者を置き換えたわけではない。実際に起きているのは、仮説生成、探索、最適化、検証の各段階で、人間と機械の分業が再編されつつあるという事態である。

25.1　AI for Science —— 発見の速度と形式の変化

「AI for Science」が広く語られるようになったのは2020年代である。しかし、その系譜は古い。第2章で見たLogic Theoristは、すでに「科学的推論の一部を機械化できるのではないか」という発想を含んでいた。違うのは、2020年代のAIが、論理玩具ではなく、実験科学・生命科学・地球科学に具体的な成果を出し始めたことである。

2025年2月、GoogleはAI co-scientistを公表した。これはGemini系モデルを基盤に、文献の整理、仮説候補の生成、研究計画の下案づくりを支援するマルチエージェント型の科学支援システムである。ここで重要なのは、「AIが科学する」という劇的な表現よりも、仮説形成の前段階に計算資源を大きく投入できるようになった点だろう。人間の研究者が直感や経験に頼っていた探索空間を、AIが拡張しているのである。

同時に、自律型実験室（Self-Driving Lab）への関心も高まった。これは、機械学習による候補選定、ロボティクスによる実験、センサーによる観測、結果のフィードバックを一つの循環にまとめる発想である。材料探索や化学合成のように、目的関数が比較的はっきりした領域では、とくに有効である。ただし現時点の自律型実験室は、問題設定そのものを発明するより、与えられた探索空間を高速に回すことに強い。ここに、AI科学の現在の強みと限界が同時に現れている。

25.2　創薬とタンパク質工学 —— AlphaFold以後

AI for Scienceの象徴的成功は、やはりAlphaFoldである。第16章で扱ったAlphaFold 2がタンパク質立体構造予測を大きく前進させたのに対し、2024年5月のAlphaFold 3は、タンパク質だけでなくDNA、RNA、リガンド、イオンなどを含む分子間相互作用まで扱えるようになった。ここで焦点は「単独分子の形」から「分子どうしがどう関わるか」へ広がった。

とはいえ、これで実験が不要になったわけではない。AlphaFold　3は分子生物学と創薬の初期探索を強力に支援するが、動的な相互作用、細胞環境、毒性、薬効、製造可能性といった問題は依然として実験と臨床に依存する。AIは創薬を短絡化したのではなく、候補探索の前半を圧縮したのである。

この点で注目されたのが、Insilico　Medicineのrentosertibである。2025年には特発性肺線維症に対する第IIa相試験で前向きな結果が報告され、AI主導創薬の代表例として広く参照されるようになった。ただし、ここでも言うべきは「AIが薬を完成させた」ではない。標的探索、分子設計、候補選定を加速した一方で、臨床試験という長い検証工程はなお不可欠だった、ということである。

規制側も追随し始めている。FDAは2025年1月、薬品・生物製剤の規制判断を支えるAIモデルについて、信頼性評価の枠組みを示すドラフトガイダンスを公表した。これは、AI創薬がもはや遠い未来像ではなく、既存の規制実務が向き合うべき現実の問題になったことを意味する。

25.3　気候・地球システム研究 —— 物理モデルとの協働

地球科学でもAIの役割は大きい。代表例の一つが、Microsoft　ResearchのAuroraである。これは100万時間を超える多様な地球観測・シミュレーションデータで事前学習された地球システム向け基盤モデルで、気象予報だけでなく、大気汚染や海洋波浪など複数のタスクに適応できる。従来の数値予報モデルに比べて、推論コストを大幅に下げつつ高い精度を示したことが注目された。

この系譜の意義は、物理法則を捨てることではない。むしろ、物理ベースの数値モデル、観測データ、機械学習をどう組み合わせるかが核心である。AIモデルは高速で、データ駆動の補間や近似に優れる。しかし、長期予測の解釈可能性や異常事象への頑健性では、依然として物理モデルの役割が大きい。したがって、未来の地球科学は「物理かAIか」ではなく、「物理とAIの協働」へ向かうと見る方が正確である。

同じことはエネルギー分野にも当てはまる。再生可能エネルギーの導入拡大に伴い、需給予測、蓄電池制御、系統運用の最適化はますます難しくなる。AIはそこで実務的価値を持つが、同時に巨大モデルの訓練と推論が電力を消費するという逆説も抱える。AIは気候問題の解決手段であると同時に、資源制約の当事者でもある。

25.4　数学 —— 証明の自動化から競技水準へ

数学は長らく、AIにとって最後まで残る知的課題の一つと考えられてきた。ところが2024年、Google DeepMindのAlphaProofとAlphaGeometry 2の組み合わせは、国際数学オリンピアッド（IMO）2024で銀メダル相当の28点を記録した。これは、形式的に検証可能な証明探索が、ついに人間の超上位層に接続したことを示す出来事だった。

2025年にNatureへ公表されたAlphaProofの詳細は、その成果が単なる大規模言語モデルの偶然ではないことを示した。中核にあるのは、Leanのような形式証明環境の上で、強化学習を用いて証明戦略を探索する方法である。自然言語の流暢さではなく、ステップごとの正しさを機械的に検証できる点が重要で、ここでは「もっともらしい説明」ではなく「正しい証明」が評価軸になる。

さらに2025年には、Deep　Thinkを備えたGeminiの高度版がIMOで金メダル水準を達成したとGoogle DeepMindが公表した。これにより、数学的推論はもはやAIの周辺的能力ではなく、最先端研究の正面課題となった。

ただし、ここでも慎重さは必要である。競技問題を解く能力と、研究数学における概念形成、問いの設定、美的判断は同じではない。AIは証明探索と検証で急速に伸びているが、「何を重要な問題とみなすか」を決める営みは、なお人間の仕事であり続ける。

25.5　AIが変える科学的方法論

以上の事例を貫くのは、科学的方法そのものの変化である。AIは、文献探索、候補生成、シミュレーション、実験計画、証明探索を高速化する。すると、科学者の役割は単純に縮小するのではなく、むしろ再定義される。何を問うべきか、どの結果を信頼すべきか、その結果をどんな社会的文脈で位置づけるかが、以前より重要になる。

同時に、再現性のあり方も変わる。形式証明や完全記録された実験ループは、従来より高い検証可能性をもたらす可能性がある。他方で、モデルの幻覚、訓練データの偏り、ブラックボックス的推論、巨大計算資源への依存は、新しい不透明性も生み出す。AI　for　Scienceは、科学を透明化する技術であると同時に、新しい不可視性を持ち込む技術でもある。

さらに制度面では、知識の所有権、計算資源へのアクセス、規制審査の速度、公共研究と企業研究の関係が問われる。AlphaFoldのように広く公共的に使われる成果もあれば、創薬や科学エージェントのように強い商業化と結びつく領域もある。AIが科学を加速するほど、「誰のための科学か」という古い問いは、より鋭く再来する。

次章では、この科学的進歩を踏まえつつ、AIと社会の未来をより広い視野から考える。AIは発見を加速するが、その意味づけまで自動化するわけではない。そこで最後に問われるのは、知能の行方だけでなく、人間がその知能を何のために使うのかである。

参考資料（本章）

- Google, Google AI announcements from February 2025

- Google DeepMind, AlphaFold
- Nature, A foundation model for the Earth system
- Microsoft Research, Introducing Aurora: The first large-scale foundation model of the atmosphere
- FDA, Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products
- FDA, Artificial Intelligence for Drug Development
- Insilico Medicine, Rentosertib関連資料
- Nature, Mathematicians put AI model AlphaProof to the test
- Google DeepMind, Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad

第26章　これからのAI ―― 技術・社会・人間の共進化

最終章では、未来を予言するのではなく、2026年時点で見えている分岐点を整理する。AIの今後を考えるとき、技術だけを見ても不十分であり、社会制度だけを見ても足りない。どのような計算基盤が主流になるか、誰がその恩恵を受けるか、市民がどのように関与できるか、そして人間は何を自らの固有の営みとして残すのか。この四つをまとめて考える必要がある。

26.1　次の計算基盤 ―― Transformerの先へ

2020年代前半のAIは、Transformerを中心に進んだ。しかし、その成功が大きいほど、弱点もはっきりした。自己注意機構は強力だが、長い文脈になるほど計算負荷が重い。そこから、より長い系列をより少ない計算で扱う方法への関心が高まった。

その代表例が、State Space Modelsを発展させたMambaである。Mambaは、入力全体に一様に注意を向けるのではなく、系列を走査しながら必要な情報を選択的に保持・更新する。ここでの意義は、Transformerの完全な代替が見えたというより、長文脈処理と計算効率をめぐる設計空間が再び開いたことにある。

ただし、2026年時点でTransformerが終わったと見るのは早い。現実には、自己注意を中心に据えつつ、外部メモリ、圧縮、検索、状態空間モデル、推論時の追加計算を組み合わせるハイブリッド化が進んでいる。今後の主戦場は、「何が唯一の勝者か」よりも、「どの用途にどの構成が向くか」を見極める局面になるだろう。

ハードウェアでも同様の変化がある。IntelのLoihi 2やHala Pointに代表されるニューロモルフィック計算は、イベント駆動型で高い省電力性を狙う研究として前進している。これらは特定のタスクで大きな効率向上を示すが、まだ汎用的な大規模言語モデルの主流基盤ではない。したがって、2030年代を見通すなら、GPU・TPU中心の世界が急に消えるのではなく、省電力化、専用アクセラレータ、モデル圧縮、用途特化設計が併存すると考える方が妥当である。

26.2　AIと人間の協働 ―― 代替ではなく拡張として

AIの未来を考えるとき、「人間を置き換える機械」という図式だけでは視野が狭い。むしろ実務の現場で広がっているのは、拡張知能（IA）の発想である。人間が問いを立て、AIが候補を広げ、人間が選び直し、AIが再度整える。この往復が、多くの知的作業の基本形になりつつある。

この協働には少なくとも三つの型がある。第一に、検索・要約・草案作成のような「探索支援」。第二に、複数案を比較しながら思考を深める「対話支援」。第三に、人間が責任を持ち、AIが評価や監視を担う「意思決定支援」である。重要なのは、どの型でも最終責任と文脈判断をどう残すかである。

ここから教育への示唆も出てくる。AI時代の教育は、単に「AIに負けない能力」を育てることではない。AIを使って思考を拡張しつつ、どこで自力判断へ戻るべきかを学ぶことが重要になる。批判的読解、問いの設定、価値判断、説明責任の感覚は、その意味で基礎技能になる。

26.3　グローバルサウスとAI ―― 普及の問題と主権の問題

AI史は米国と西欧中心に語られがちだが、今後の普及段階で重要になるのはグローバルサウスである。ここで問題になるのは、単なる利用率の差ではない。計算資源、学習データ、電力、クラウドアクセス、言語資源、法制度設計への発言権が偏っていることである。

一方で、状況は一方向ではない。オープンウェイト・モデルの広がりは、先進的な基盤モデルを各地域で再利用・微調整する可能性を広げた。多言語モデルや地域特化型モデルの発展も、英語中心の偏りを相対化しつつある。インドやアフリカ諸国でAIが「安全保障」よりも「農業、教育、行政、保健」の文脈で語られやすいのも、AIの意味が地域ごとに異なることを示している。

それでも、構造的不均衡は残る。モデルを使えることと、モデルの将来像を決められることは別だからである。グローバルサウスにとってAIの課題は、普及の遅れだけではなく、標準や規範の形成にどこまで参加できるかという主権の課題でもある。

26.4　AIリテラシーと民主的ガバナンス

AIが社会基盤になるなら、その統治は専門家だけの仕事では終わらない。ここで必要なのがAIリテラシーである。だが、その意味は単なる操作方法の習得ではない。少なくとも、仕組みの大枠を理解する技術的リテラシー、出力の信頼性を見極める利用者リテラシー、社会的影響を考える市民的リテラシー、そして企業や政府の語る「AIの物語」を批判的に読むリテラシーが必要である。

UNESCOは2023年に生成AIと教育に関するガイダンスを出し、2024年には教師・学生向けのAIコンピテンシー枠組みを公表した。欧州評議会のAI枠組み条約や、EUのAI Continent Action Planも、技術規制だけでなく、教育、権利、社会参加を一体で捉えている。ここから分かるのは、AIガバナンスが法規制だけで成立するのではなく、使う側の能力形成を前提としていることだ。

民主的ガバナンスのためには、さらに三つが必要である。第一に、異議申立てや説明請求が可能であること。第二に、労働者、市民団体、教育機関、地域社会など、影響を受ける側が早い段階から意思決定に参加できること。第三に、AIを使わない選択肢や、人間による代替経路を一定程度残すことである。全面自動化より、可逆的な導入の方が民主政治には適している。

26.5　知能を作る試みが、人間を照らし返す

本書の出発点は、「機械は知能を有しうるか」という問いだった。2026年時点でも、この問いに決着はついていない。だが、AIの歴史を通して一つはっきりしたことがある。人間は、自分が何を知能と呼んでいるのかを、技術の進歩によって何度も言い換えてきたということである。

1950年代には論理推論が知能の中心に見えた。1970年代には知識表現が、1990年代には統計的学習が、2010年代には表現学習が、2020年代には生成と推論の統合が前景に出た。知能の定義は、つねにAIの達成に照らされて更新されてきた。これは、人間が知能を理解したからAIを作れたというより、AIを作る過程で初めて知能概念の曖昧さに気づいてきたことを意味する。

その意味で、AI研究が人間にもたらした最大のものは、新しい知能そのものよりも、人間知能の条件を逆照射したことかもしれない。身体性、社会性、有限の生、他者との関係、責任、価値判断。こうしたものは、かつては「当たり前すぎて見えない背景」だったが、機械知能と比較することで、初めて輪郭を持ち始めた。

未来においてAIがさらに高性能になるとしても、最後に残る問いは同じである。「何ができるか」ではなく、「何のために使うか」である。TuringからMcCulloch-Pitts、そしてTransformerと拡散モデルに至る歴史が教えるのは、知能の歴史がつねに人間の自己理解の歴史でもあったということだ。AIの未来は、技術の未来であると同時に、人間が自らをどう定義し直すかという未来でもある。

参考資料（本章）

- Gu, Albert and Tri Dao, Mamba: Linear-Time Sequence Modeling with Selective State Spaces
- Intel, Intel Builds World’s Largest Neuromorphic System to Enable More Sustainable AI
- UNESCO, Guidance for generative AI in education and research
- UNESCO, AI competency framework for students
- Council of Europe, The Framework Convention on Artificial Intelligence
- European Commission, AI continent