2.1.1 知覚・運動系のAI技術

（1）研究開発領域の定義

知能を知覚・運動系と言語・知識系という2面で捉え、ここでは前者を俯瞰する（後者については次節2.1.2で俯瞰する）。

知覚系は実世界からの入力、運動系は実世界への出力として、知能の実世界接点の役割を担う。研究開発領域として、知覚系は画像・映像などのパターン認識、運動系はロボットなどの動作生成が中心的に取り組まれてきたが、近年、機械学習（Machine Learning）、特に深層学習（Deep Learning）の発展によって、知覚系・運動系それぞれの精度・性能が向上したことに加えて、知覚系と運動系を統一的に扱う取り組みが進展しつつある。また、状況を知り、判断し、行動するという一連のプロセスは、知能において、知覚系と言語・知識系と運動系の連携によって熟考的に実行されることもあれば（ここでは熟考的ループと呼ぶ）、知覚系と運動系の間で即応的に実行されることもある（ここでは即応的ループと呼ぶ）[1]。

本節では、知覚・運動系の研究開発動向として、機械学習技術をベースとしたパターン認識と動作生成、および、それらを統合した即応的ループを中心に取り上げる。なお、機械学習技術は、人工知能（Artificial Intelligence：AI）の研究分野全般にわたって用いられる共通技術となっているが、その主要な研究開発動向については、本節に記載する。

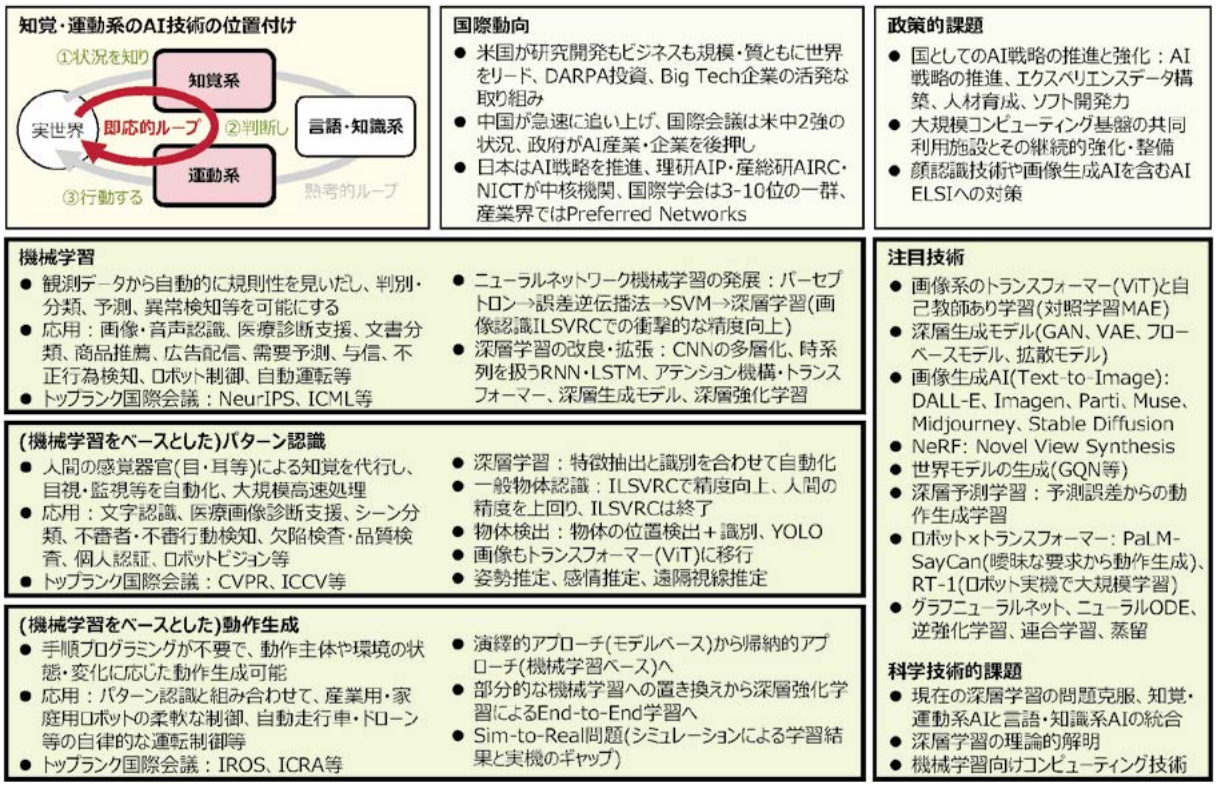

図2–1–3　　領域俯瞰：知覚・運動系のAI技術

1　人間の思考は、直感的・無意識的・非言語的・習慣的な「速い思考」のシステム1と、論理的・系列的・意識的・言語的・推論計画的な「遅い思考」のシステム2とで構成されるという「二重過程理論」（Dual Process Theory）がある。社会心理学・認知心理学などの心理学分野で提案されていたが、ノーベル経済学賞を受賞したDaniel Kahnemanの著書「Thinking, Fast and Slow」[1)] でよく知られるようになった。本稿ではシステム1を「即応的ループ」、システム2を「熟考的ループ」と呼んでいる。

（2）キーワード

機械学習、画像認識、映像認識、パターン認識、一般物体認識、物体検出、顔認証、行動認識、深層学習、ニューラルネットワーク、敵対的生成ネットワーク、動作生成、ロボット制御、即応的知能、二重過程理論、基盤モデル、世界モデル

（3）研究開発領域の概要

［本領域の意義］

機械学習は、経験からの学習により自動で改善するコンピューターアルゴリズム[2]もしくはその研究領域である。事象や対象物についての観測データを集めて機械学習にかけることで、そこから（人間がルールを書く必要なく）データの背後に潜む規則性を自動的に見いだし、判別・分類、予測、異常検知などを行うことを可能にする。ビッグデータの時代と言われる今日、さまざまな事象や対象物について大量の観測データが得られるようになり、機械学習は幅広い分野・目的に利用されるようになった。例えば、画像認識、音声認識、医療診断支援、文書分類、スパムメール検出、広告配信、商品推薦、囲碁・将棋などのゲームソフト、商品・電力などの需要予測、与信、不正行為の検知、設備・部品の劣化診断、ロボット制御、車の自動運転など、多数の応用例が挙げられる。

このように機械学習はさまざまな応用が可能であるが、ここでは特に機械学習を用いたパターン認識と動作生成について述べる。

パターン認識は、カメラやビデオレコーダーなどで撮影された画像・映像・音声を、機械学習によって判別・分類して、その画像・映像・音声の内容、つまり、そこに写っているものや話されていることが何であるか、その位置や状態、あるいはシーン全体の状況を認識する技術である。人間の感覚器官（目・耳など）による知覚の代替となり、人間が行っている目視作業の自動化といった単なる省力化としての価値だけでなく、ヒューマンエラーを低減する判断・診断の支援や、人間では処理しきれないほど大量の画像・映像データの高速処理など、これまで得られなかった新たな価値も提供できる。具体的な応用先は、郵便区分機などでの文字認識、マンモグラフィーなどの医療画像診断支援、監視カメラ映像からの不審者・不審行動や異常状況の検知、インターネット上の画像・動画像検索、カメラ画像のシーン分類、半導体ウェハーやフォトマスクなどの欠陥検査、食品の異物検査、製品の品質検査、衛星画像などのリモートセンシング、出入国管理などでの顔や指紋を用いた個人認証、自動車の安全運転支援や自動運転、ロボットビジョン、スポーツ画像解析、動作認識によるヒューマンインターフェースデバイスなどへと広がっている。

動作生成は、実世界に作用する機器・デバイスに対して、どのような動作をどういう順序で実行させるかを計画し、その実行指示を行う技術である。従来は角度・距離なども含む詳細な動作パラメーターや動作順序をすべて人間が事前にプログラミングする必要があった。しかし今日、機械学習を用いた動作生成によって、ロボットなどの動作主体や環境の状態・変化に応じた臨機応変な動作生成が可能になり、さまざまな運動系タスクを容易に自動化できるようになりつつある。応用分野は、産業用から家庭用までロボット制御への適用はもちろん、自動走行車やドローンなどの移動体・飛行体の運転制御への適用も試みられている。また、このような応用では、カメラ映像から状況・状態を認識し、その状況・状態に応じた動作を計画・実行するという、パターン認識と動作生成を組み合わせた形態（前述の即応的ループ）が取られることも多い。例えば、カメラ映像から対象物の形状・位置・向きなどを認識し、それを把持するためにロボットアームの動作（アームをどう移動し、対象物のどこをつかむか）を決定したり、巨大な対象物をカメラ付きドローンで観測・検査する際に、対象物の一部を観測・検査した結果をもとに、次に観測すべき箇所を自律的に決定したりといっ

2 「コンピュータープログラムがタスクのクラスTと性能指標Pに関し経験Eから学習するとは、T内のタスクのPで測った性能が経験Eにより改善されること」という定義[2)]がよく知られている。

たことが可能になる。

以上のように、機械学習をベースとしたパターン認識と動作生成、および、それらを組み合わせた即応的ループは、人間の知覚・運動系のさまざまなタスクを代行できるようになりつつあり、幅広い産業応用にもつながっている。

［研究開発の動向］

❶ 機械学習の発展[3]

機械学習の基本的な処理構成は、訓練ステップ（学習ステップとも呼ばれる）と判定ステップ（推論ステップや予測ステップとも呼ばれる）に分かれる。訓練ステップは、訓練データ（学習データとも呼ばれる）を与えて、モデルを作るステップである。ここで作られたモデルは、訓練データの統計的傾向・規則性を表したものになる。判定ステップは、新たに入力されるデータに対して、訓練済みモデル（学習済みモデルとも呼ばれる）に基づき、分類・回帰・予測・異常検知などの判定結果を出すステップである。訓練データに判定結果が付与されているケースは教師あり学習（Supervised Learning）、付与されていないケースは教師なし学習（Unsupervised Learning）と呼ばれる。なお、教師あり学習・教師なし学習とは異なるタイプとして強化学習（Reinforcement Learning）があるが、これについては後述する。

機械学習の研究では、訓練ステップのアルゴリズム（学習アルゴリズム）、つまり、訓練データからそこに潜む統計的傾向・規則性をどのようにして見いだすかが、一つの重要なポイントになる。モデルを訓練データにフィットさせ過ぎると、判定ステップで与えられるデータに対して必ずしも高い精度が得られないという問題（過学習と呼ばれる）も生じるため、汎化が適切に行われるような仕掛けが必要である。学習アルゴリズムの研究では、統計解析の手法とともに、人間の脳神経回路にヒントを得たニューラルネットワークを用いた手法が注目されるようになった。

このような研究は、古くは1958年にパーセプトロンと呼ばれる単純なニューラルネットワークモデルが提案され、任意の線形分離関数を学習できることから1960年代に活発に研究された。しかし、単純なパーセプトロンでは排他的論理和のような関数を学習できない問題が指摘され、1970年代には関連する研究は下火になった。この問題はニューラルネットワークに階層構造を持たせれば解決できるのだが、その学習を可能にする誤差逆伝播法（Backpropagation）が提案されたのは1986年であった。これをきっかけにニューラルネットワーク研究が再び活発化し、画像認識、音声認識、ロボット制御など、さまざまな問題に適用されるようになったが、一般に大域的な最適解を求めることができないという弱点があった。これに対して、1992年に提案されたカーネル学習器SVM（Support Vector Machine）は、階層性を持たず、容易に大域的な最適解を求めることができることから注目され、その利用が広がった。

ここからさらに衝撃的な精度向上をもたらしたのが、Geoffrey Hintonらが発表した深層学習（Deep Learning）である。これは、層の数が多い、すなわち、深い層のニューラルネットワークを学習させる手法であり、特徴抽出の自動化も可能にした。その前身として、1979年に福島邦彦が発表したネオコグニトロンがある。畳み込み層とプーリング層の組を複数積み重ねることで、パターンの局所変動に頑健になることが示されていた。1989年にYann LeCunが発表した畳み込みニューラルネットワークCNN（Convolutional Neural Network）では、それを誤差逆伝播法で最適化している。しかし、ネットワーク構造が深くなるほど伝播される誤差が小さくなり、学習が進まなくなる問題があった。この問題は、誤差が深い層まで伝播するように活性化関数や正則化を工夫することで改善された。その結果、深層学習は、2012年の画像認識コンペティションILSVRC（ImageNet Large Scale Visual Recognition

3 機械学習・深層学習の歴史的研究成果についての個々の参考文献は省略する（2021年版の俯瞰報告書[3)]では参考文献を挙げている）。各成果・方式の詳細は、機械学習[2)]、深層学習[4),5),6)]、画像認識[7)]などの教科書的文献が分かりやすい。

Challenge）で衝撃的な精度向上を示して大きく注目され、第3次AIブームを牽引する技術となった。「深層学習の父たち」と呼ばれるGeoffrey Hinton、Yann LeCun、Yoshua Bengioは、2018年度ACM（Association for Computing Machinery）チューリング賞を受賞した。

その後も深層学習の改良・拡張が活発に行われている。CNNの多層化を大きく進めたのは、入力データから出力への変換を学習するのではなく残差を学習するResNet（Residual Network）である。ResNetは、迂回路を含むネットワーク構造を持ち、階層を深くしても効率よく学習が行える。また、時間的構造の表現を扱いやすい回帰型の構造を持つニューラルネットワークRNN（Recurrent Neural Network）が考案された。RNNは過去の入力の影響を受ける構造を持つが、長期的影響と短期的影響を区別しないのに対して、長期の依存関係をモデルに取り込んだLSTM（Long Short-Term Memory）ネットワークも考案された。RNNやLSTMは、自然言語や時系列データなどの解析に用いられたが、その後、RNNやCNNを使わず、アテンション機構のみを用いたトランスフォーマー（Transformer）と呼ばれる多層ニューラルネットワーク（アテンション機構やトランスフォーマーの詳細は「2.1.2 言語・知識系のAI技術」を参照）が自然言語処理の主流になり、次いで画像処理・パターン処理にもトランスフォーマーが用いられるようになった。さらに、多層ニューラルネットワークを用いてデータの生成過程をモデル化する深層生成モデルや、強化学習に深層学習を組み合わせた深層強化学習といった拡張も行われ、これらの研究開発・応用も活発に取り組まれている。

❷ パターン認識の研究開発動向

パターン認識の基本的な処理は、観測、前処理、特徴抽出、識別から成る[8)]。観測は、カメラなどを通して、実世界の事象を処理可能なデータに変換する処理である。実世界は3次元立体であるが、カメラで撮影されるデータは2次元平面のため、被写体の姿勢変動や照明変動の影響で被写体の見えが大きく変化する。センサーの併用など、隠れ（オクルージョン）の発生への対処が課題として検討されている。前処理は、以降の処理にかかる演算量を軽減するための処理であり、具体的にはデータの正規化やノイズの除去が行われる。不明瞭な領域の鮮鋭化、霧などを除去するデヘイズ処理、画像の解像度を上げる超解像処理などの画像処理技術も開発されている。特徴抽出は、前処理後の画像・映像から識別に有効な特徴を抽出する処理である。局所フィルターを用いたエッジやコーナーなどの画像特徴抽出、識別に有効な特徴の組み合わせを選ぶ特徴選択、識別に有効な特徴への特徴変換などが行われる。識別は、得られた多数の特徴値を多次元特徴ベクトルとみなし、あらかじめ設定したクラス（あるいはカテゴリ）に分類する処理である。クラスは目的に応じて人間が設定するものであり、例えば、人物と車両を識別する場合は、それぞれが一つのクラスとして設定され、顔認証の場合は、人物一人一人を識別する必要があるため、それぞれが一つのクラスとして設定される。

観測と前処理は、専門家がこれまでの経験に基づき、目的に応じて設計している。識別は、テンプレートマッチングと呼ばれる単純な手法から機械学習で自動設計する手法に移行した。特徴抽出は、従来、専門家が経験に基づいて設計するのが一般的であったが、深層学習によって、特徴抽出と識別を合わせて自動設計できるようになった。

このように深層学習の導入が進むことになったきっかけは、2012年のILSVRCである。ILSVRCは大規模画像データセットImageNetを用いた画像認識コンペティションである[4]。前述のように、Hintonらは一般物体認識タスクで1位を獲得した。しかも、従来法がエラー率26%だったのに対してエラー率17%と、深層学習の適用によって一気に約10%もの飛躍的な精度向上を達成し、画像認識・機械学習の研究者ら

4 ImageNetは、スタンフォード大学のFei-Fei Liらによって構築され、1400万枚もの画像データが集められている。ILSVRCでは、タスクによって、この部分データが用いられた。

に衝撃を与えた。その後、深層学習はさらに改良と多層化が進み、2015年には人間レベルの精度（5.1%）を超えて、エラー率3.57%となった。2016年に2.99%、2017年に2.25%とさらに改善されつつも、精度はほぼ飽和状態に至った。なお、2015年以降は中国勢が1位を取っている。

一般物体認識は画像に映っている物体を識別するタスクであるが、物体の識別だけでなく、物体の位置も正確に検知する物体検出タスクへの取り組みも進んだ。2014年に、物体の候補領域を抽出する処理とCNNを統合したRegional CNN（R-CNN）が提案されたのをきっかけに、2015年にFast R-CNN、Faster R-CNNと高速化が進んだ。2016年には、画像をグリッドに区切った領域をもとに物体を抽出するSDD（Single Shot MultiBox Detector）、YOLO（You Only Look Once）が提案され、さらなる高速化と高精度化が進んでいる。これまでは十数種類の物体の検出・識別するモデルが多かったのに対し、2017年に提案されたYOLO 9000は、9000種類の物体の検出・識別が可能である。さらに、静止画でなく動画やカメラ映像に対する物体検出への拡張も盛んに取り組まれている。例えば、自動走行車がカメラ映像から周辺状況（他の車、歩行者、道路標識など）を認識するために必要な重要技術である。また、動画・カメラ映像からの人物行動認識も盛んに研究されている。技術的には、従来2次元画像に用いられているCNNを3次元に拡張した3D CNNが開発されている[9)]。また、深層学習が注目される以前に実用化されていた文字認識・音声認識・顔認証などのパターン認識技術も、深層学習を用いた方式に置き換わってきている。併せて、「❶機械学習の発展」で述べたように、深層学習の方式はCNN型からトランスフォーマー型への移行が進んでいる。さらに、より詳細に人物を捉えるパターン認識技術として、人の頭・肩・腰・足・膝・肘といったパーツを検出し、それらの位置関係から姿勢を推定する技術（OpenPose[10)]）、顔の表情や声の調子から人の感情を推定する技術、離れた場所のカメラ映像から人の視線の向きを推定する技術（遠隔視線推定技術）なども開発されている。

❸ 動作生成の研究開発動向

産業用ロボットなどで実用化されている動作生成技術は、伝統的なモデルベースの演繹的なアプローチが主流であるが、昨今活発に研究開発が進められているのは、深層学習を用いた帰納的なアプローチである[11)]。従来の演繹的なアプローチでは、先に環境のセンシングと、環境や操作対象物のモデリングが行われ、環境、操作対象物、操作主体（ロボット）に関する精緻な物理モデルが正確に得られていることを前提に、最適な動作軌道を探索する。しかし、精緻なモデルを得るためには、事前に人手で記述しておかねばならない部分が多く、動作中に環境自体も変化し得ることから、適用できるケースは限定的にならざるを得ない。この改良として、モデル自身の曖昧性を認めた上で、センサーから取得したデータをもとに統計的な修正をかける確率的な手法も提案された。深層学習を含む機械学習の導入方法も当初は、環境や操作対象のセンシングとモデリングの後に、動作軌道を探索・生成するというシーケンシャルな流れの中で、一部のステップに機械学習を適用するというものであった。しかし、深層学習を用いることで新たな可能性が生み出されたのは、環境センシングから動作生成までをEnd-to-Endで学習するという処理形態である。すなわち、途中ステップをどのように構成・モデル化するか、どういう情報に着目して動作を生成するか、といった設計を人間が行う必要なく（モデルフリーで）、環境と操作対象物の状態に応じて操作主体が実行すべき動作の生成（「（1）研究開発領域の定義」で述べた即応的ループに相当する）が、End-to-End学習によって最適化される。

このようなEnd-to-End学習による動作生成で活用が広がっているのが、深層学習と強化学習を組み合わせた深層強化学習（Deep Reinforcement Learning）である。強化学習では、学習主体が、ある状態で、ある行動をしたとき、その結果に応じた報酬が与えられる。行動と報酬の受け取りを試行錯誤的に重ねることを通して、より多くの報酬が得られるように行動を決定する意思決定方策を学習する。この強化学習では、ある状態で、ある行動を取ることの良さを表す評価関数を求める必要があるが、この評価関数や方策を深層学習によって学習するのが深層強化学習である。

この深層強化学習によるエポックメイキングな成果として、Google DeepMindのAlphaGo（アルファ碁）が挙げられる。 AlphaGoは、モンテカルロ木探索に組み合わせて、膨大なプロの棋譜を訓練データとした教師あり学習と、膨大な回数の自己対戦による深層強化学習を用いて訓練され[5]、2016年～2017年に世界トップランクプロに圧勝し、大きな話題になった。AlphaGoでは試行錯誤を通してゲーム空間における行動（碁の打ち手）を学習・生成したわけだが、このような方法は、実世界における行動（ロボットなどの動作）の学習・生成にも応用できる[12)]。既に、ばら積み部品のピッキング作業、衣類を畳む作業、車の運転操作など、さまざまな適用事例がある[11), 13), 14), 15)]。例えば、産業用ロボットによる部品ピッキングタスクを考えると、部品の種類と置き方、照明状態などが固定されていれば、従来の演繹的なアプローチでも、部品のどこをどのように把持すればよいかを事前にプログラミングできる。しかし、部品がばら積みされ、照明状態にも変化があり、多種類の部品にも対応しなければならないならば、想定されるケースがあまりに複雑になり、演繹的なプログラミングはもはや困難である。深層強化学習を用いれば、さまざまなばら積み状態に対して、さまざまなバリエーションで把持操作を試行錯誤し、その成功・失敗から、状態に適した把持方法を学習していくことができる。あるいは、手本となる行動・動作を例示し、それをもとに報酬や方策を学習する逆強化学習・模倣学習も用いられる[11), 12)]。

しかし、AlphaGoで行われたような膨大な回数の試行錯誤を、実際にロボットに行わせることは不可能である。その対策として、コンピューター上でのシミュレーションによる深層強化学習の結果をロボットの実機に適用すること（Sim-to-Real）[16)] が行われているが、実機での動作・作用とシミュレーション上での結果が完全一致するとは限らないために、ドメイン適応やドメイン一般化を含め、何らかの調整が必要になるという課題が生じている。

一方で、自然言語処理や画像処理ではトランスフォーマー型の大規模モデルが成果を挙げていることから、動作生成・ロボット制御でも同様のアプローチが検討され始めた。自然言語による曖昧な要求に対して動作・行動を生成するPaLM-SayCanや、ロボット実機での大規模学習によって、さまざまな動作・行動を学習したRT-1など、その先駆的な研究事例が注目されている。これらについては［注目される国内外プロジェクト］❷❸で取り上げる。

❹ 学会・産業界の動向

各研究分野のトップランク国際会議として、機械学習分野はNeurIPS（Neural Information Processing Systems）やICML（International Conference on Machine Learning）、パターン認識分野はCVPR（Computer Vision and Pattern Recognition）やICCV（International Conference of Computer Vision）、動作生成を含むロボティクス分野はIROS（International Conference on Intelligent Robots and Systems）やICRA（International Conference on Robotics and Automation）が挙げられる。また、これらの研究分野は、AI分野全般のAAAI（Association for the Advancement of Artificial Intelligence）やIJCAI（International Joint Conferences on Artificial Intelligence）においてもホットな研究テーマとなっている。

機械学習の研究開発・ビジネスは、米国が規模・質ともに世界をリードしている。 AI分野の国家戦略・投資では、歴史的に米国国防高等研究計画局（Defense Advanced Research Projects Agency：DARPA）が中心的な役割を果たしてきたことに加え、Google（DeepMindも傘下に含む）、OpenAI、Apple、Meta（旧Facebook）、AmazonなどのBig Tech企業が活発な取り組みを進めている。

その米国を中国が急激に追い上げ、上述の主要国際会議は米中2強という状況になり、AAAIやIJCAI

5 その後、棋譜のような訓練データを必要とせずに自己対戦だけで学習するAlphaZeroや、ゲームのルールが不明でもルール自体を学習するMuZeroへと発展した。なお、ゲームAIの発展については「2.1.6 AI・データ駆動型問題解決」に記載している。

では採択論文数で中国が米国を上回った。中国政府は2017年7月に次世代AI発展計画を発表し、2030年までに理論・技術・応用のすべての分野で世界トップ水準に引き上げ、中国のAI産業を170兆円に成長させるという目標を設定した。これに向けて、政府主導で重点AI分野を定め、医療分野はTencent（騰訊）、スマートシティーではAlibaba（阿里巴巴）、自動運転はBaidu（百度）、音声認識はiFLYTEK（科大訊飛）、画像認識はSenseTime（商湯科技）をリード企業として選定し、政府がAI産業を後押ししている。

画像認識コンペティションILSVRCが深層学習の性能向上に大きく貢献したことは前述の通りだが、2017年に終了する頃には中国勢が躍進し、技術改良・応用における中国の強さを示した。機械学習の応用やデータ分析の分野では、ILSVRCに限らず、共通のデータセットを用いたコンペティションが多数開催されており、技術の性能向上と普及につながっている。また、企業などがデータや問題をネット上で公開して、多数の人々に解かせる場（世界的には米国のKaggle[6]、国内ではSignateがそのプラットフォームとしてよく知られている）も生まれている。日本勢は、米国国立標準技術研究所（National Institute of Standards and Technology：NIST）の顔認証ベンチマークテストをはじめ画像認識関連コンペティションで1位を獲得したり、層の厚みには課題があるもののKaggleで活躍する産業界の人材もいたりと、一定の存在感を示してきた。

日本政府は、2016年4月に人工知能技術戦略会議を設立し、2019年6月に統合イノベーション戦略推進会議決定による「AI戦略2019」を発表し、「AI戦略2021」「AI戦略2022」とアップデートを加えているが、その中で、文部科学省による理化学研究所革新知能統合研究センター（AIP）、経済産業省による産業技術総合研究所人工知能研究センター（AIRC）、総務省による情報通信研究機構（NICT）の三つを中核的なAI研究機関と位置付けている。前述の国際会議の採択論文数において、圧倒的な米中2強の後、日本は欧州各国とともに3位から10位の一群に含まれているが、上記中核研究機関を中心に徐々に論文数を伸ばしつつある。国内産業界で特に注目されるのはPreferred Networksである。深層学習・深層強化学習をコア技術に持ち、交通システム（自動運転、コネクテッドカー）、製造業（ロボット）、バイオヘルスケアを重点領域として、トヨタ自動車、ファナックなどとも共同研究を進めている[13)]。自社開発の深層学習用スーパーコンピューターMN-3は、2020年以降のスーパーコンピューター省電力性能ランキングGreen500で世界1位を3回獲得している。

（4）注目動向

［新展開・技術トピックス］

❶ 画像系のトランスフォーマーと自己教師あり学習

［研究開発の動向］❶❷で述べたように、画像認識などのパターン認識に用いられる深層学習のアーキテクチャーは、従来主流だったCNN型からトランスフォーマー型へと移行が進んでいる。それまでトランスフォーマーで扱っていた自然言語処理では単語（分散表現ベクトル）の系列を入力としたので、2020年に発表されたビジョントランスフォーマー（Vision Transformer：ViT）[17)] では、画像を重なり合わないパッチに分割し、各パッチに位置符号を加えたものをベクトル化して入力系列とすることで、同様に処理できるようにした。ViTでは、それまでCNNで達成されていたスコアを上回り、かつ、計算コストも1桁少なくて済んだことが示され、その後、ViTのさまざまな派生方式が開発された[18)]。

また、トランスフォーマーを含めて深層学習で高い精度を得るためには大量の教師データ（教師あり学習のためのラベル付き訓練データ）が必要だが、それを大量に準備することは容易ではない。そこで、半教師あり学習（Semi-Supervised Learning）、能動学習（Active Learning）、転移学習（Transfer Learning）、ドメイン適応（Domain Adaptation）、データ合成（シミュレーションなどを利用）やデー

6 Kaggle運営会社は、2017年3月にGoogleに買収された。

タ拡張（Data Augmentation）などが試みられてきたが、特に大きな効果を示し、活用が広がっているのが、自己教師あり学習（Self-Supervised Learning）である。自然言語処理分野では、テキストの一部にマスクをかけて（隠して）、それ以外の部分からマスク部分を推測するという穴埋めタスクを設定することで、ラベル付けなしに教師あり学習を可能にする手法が、BERTで用いられ、それ以降のトランスフォーマー型モデルの大規模化を促進した。

画像系のトランスフォーマーで用いられている自己教師あり学習の手法は、主に対照学習（Contrastive Learning）[19]とMAE（Masked Autoencoder）[20]である。対照学習では、画像データに各種変換をかけることで、訓練データ量を水増し（Data Augmentation）する。そして、同じ画像に異なる変換をした画像同士を一致させる特徴量を最大化しつつ、違う画像に異なる変換をした画像同士を一致させる特徴量を最小化するように訓練を行う。これによって、教師あり学習と遜色ない認識精度が得られると報告されている。一方、MAEは、自然言語の場合と同様に、画像においてもマスクをかけて、その部分を推定する穴埋めタスクの学習を行う。

❷ 深層生成モデルと画像生成AI（Text-to-Image）

機械学習でクラス分類を解くための手法には識別モデルと生成モデルがある。識別モデルはデータの属するクラスを同定するが、そのデータがどのように生成されたかは考えない。一方、生成モデルはデータがどのように生成されたか、その過程までモデル化する。深層学習に関して、前述のCNNなどは識別モデルであるが、生成モデルにおいても著しい進展があった。深層生成モデル[21]（深層学習の生成モデル）の主なものとして、GAN（Generative Adversarial Networks：敵対的生成ネットワーク）[22]、VAE（Variational Autoencoder：変分自己符号化器）[23]、フローベースモデル（Flow-Based Generative Model）[24]、拡散モデル（Diffusion Model）[25]がある。

GANは、2014年にIan J. Goodfellowによって発明され、従来の生成モデルではできなかった高精細な画像を生成できることから、大いに注目された。GANは、生成器Gと識別器Dから構成され、Dは訓練データとGが生成したデータを識別するように訓練され、GはDが間違えるように訓練される。 GANは、敵対的なコスト関数を最適化するため、学習の安定性が課題とされている。性能改良・機能追加や応用開発が進み、GANのさまざまなバリエーションが生まれている[21]。

VAEは、自己符号化器（Autoencoder）と呼ばれるニューラルネットワークを用いた深層生成モデルである。自己符号化器は、入力層に入ったデータが隠れ層でいったん変換された後、出力層で入力データが復元されるように構成したニューラルネットワークである。VAEは、その隠れ層にある潜在変数を操作することで、訓練データと類似しつつも異なるデータを生成する。当初はGANほど高精細な画像は生成できなかったが、さまざまな改良が加えられ、十分高精細な画像が生成できるようになってきている。

フローベースモデルは、正規化フローという手法を用いて、確率分布を明示的にモデル化することによって、複雑な分布に基づいた新しいサンプルを生成できるようにしたものである。

拡散モデルは、元データに少しずつノイズを加えていって最後には完全なノイズになるというプロセスが考えられるとき、その逆プロセスをモデル化して、データ生成に用いる。学習に要する計算コストが比較的大きくなるが、学習が安定し、生成結果の品質が高いことから、注目が高まっている。

深層生成モデルを用いることで、架空の人物顔を生成したり、ゴッホ風やレンブラント風など指定した画風に絵を変換したり、幻想的な絵や抽象画風の絵を生成したり、ラフスケッチを写実的な絵に変換したり、自動着色したりと、さまざまなアプリケーションが開発された。さらに、2022年に、テキストから画像を生成する画像生成AI（Text-to-Image）アプリケーションが大きな話題になっている。簡単な文を与えるだけで、まるでプロが描いたようなテイストの画像が生成される。特にインターネット上で使える形でMidjourneyやStable Diffusionといったサービスが提供され、その利用者が急増し、派生サービスも拡大した。

Text-to-Image生成が注目されるきっかけとなったのは、OpenAIから2021年1月に発表されたDALL-E[26]である。2022年4月には改良されたDALL-E2[27]が発表された。ここでは、テキストと画像の類似度を求めるモデルCLIPと、画像を生成する深層生成モデル（DALL-EではVAE系のVQ-VAE、DALL-E2では拡散モデルが使われている）を組み合わせて、Text-to-Image生成を実現している。CLIPは、大量の画像とキャプションテキストのペアをもとに、トランスフォーマーと対照学習によって学習して作られた。また、Googleは2022年5月にImagen[28]、6月にParti（Pathways Autoregressive Text-to-Image model）[29]、2023年1月にMuse[30]という異なる方式で2種類のText-to-Image生成を発表した。 Imagenでは拡散モデル、Partiでは自己回帰モデル（Autoregressive Model）が画像生成に用いられている。 Museの方式はそれらと異なり、事前学習された大規模言語モデルから得られたベクトルを用いて、離散トークン空間でマスク学習を行うことで、高速かつ画像生成をコントロールしやすくなったということである。なお、高品質な画像生成を可能にするため、大量データから大規模モデルを学習することが必要になっており、モデルのパラメーター規模は、DALL-Eが120億個、DALL-E2が55億個、Imagenが76億個、Partiが200億個、Museが30億個ということである。

なお、画像生成AIに伴う倫理的・法的・社会的課題については「（6）その他の課題」❸で取り上げる。

❹ 世界モデルと深層予測学習

人間は外界に関する限られた知覚情報から脳内に外界のモデルを作り、そのモデルを用いたシミュレーションを意思決定や行動に使っていると考えられる。このモデルは「内部モデル」「力学モデル」と呼ばれることもあるが、AI分野では「世界モデル」（World Models）[31]という呼び方が主流である。なお、本節の冒頭で、知覚系と運動系の即応的ループと、言語・知識系まで含めた熟考的ループという2タイプを挙げたが、この世界モデルは即応的ループの中に位置付けられ、無意識的・反射的な行動にも作用すると考えられている。例えば、バットを振ってボールに当てる場合、ボールが飛んでくるという視覚情報が脳に到達する時間は、バットの振り方を決める時間よりも短いので、世界モデルによって無意識的に予測を行い、それに基づいて筋肉を動かしていると考えられている。

知覚情報からボトムアップに世界モデルを作ろうとする試みの一例として、Google DeepMindのGQN（Generative Query Network）[32]がある。 GQNは、異なる複数の視点から見た画像を与えると、内部に世界モデルを作り、別の視点から見た画像を予測できる。そのためにGQNでは、VAEベースの深層生成モデルを用いている。❸で述べた深層生成モデルの研究発展も受けて、世界モデルを取り入れた深層学習研究がホットトピックになりつつある。

また、知覚情報に基づいて作られたモデルを用いてシミュレーション・予測を行うという考え方は、知覚系と運動系を即応的ループとしてつなぐ上で重要なものであり、その一例として、深層予測学習による動作生成が挙げられる。産業技術総合研究所・早稲田大学で開発された深層予測学習によるタオル畳みロボット[11), 33)]は、人間によるタオル畳み操作を手本として、それを模倣する動作を生成しつつ、動作を実行した結果の事前の予測と、実際に実行した結果をカメラで観測して比較し、その差異（予測誤差）から学習する。予測に使われるモデルは、模倣と予測誤差からの学習によって構築される。

「2.1.8 認知発達ロボティクス」で詳細を述べる通り、予測誤差最小化原理に基づくモデルの更新や環境への働きかけが、人間の認知発達に深く関わっていると考えられている。上に述べた世界モデルや深層予測学習を含め、人間の知能の認知発達メカニズムの解明に構成論的にアプローチしている認知発達ロボティクスの考え方は、深層学習の今後の発展と重なりが大きくなってくると思われる。

❺ その他の注目トピックス

本研究開発領域はAI分野の中でも特に活発に取り組まれている領域であり、新たな注目技術・応用が次々に生まれている。また、機械学習はパターン認識や動作生成への適用に閉じず、AI分野全般で幅広く

活用されている。そこで、上述の❶〜❹に含められなかった注目トピックスについても、以下、簡単に触れる。

a. GNN（Graph Neural Networks）[34]：GNNはグラフ構造のデータを扱う深層学習である。 Web・SNS、交通・物流、化合物など、さまざまな対象物がグラフ構造で制約関係を表現でき、そういった関係を踏まえた計算が行える。

b. Neural ODE[35]：多層ニューラルネットワーク構造の層は離散的に扱われていたが、微小化して連続値として扱うことで、常微分方程式（Ordinary Differential Equation：ODE）の枠組みで順・逆伝播が計算でき、メモリ効率なども向上する。

c. 連合学習（Federated Learning）[36]：連合学習は、データを取得する端末側（エッジ）で学習した結果を組み合わせて機械学習モデルを作る手法である。端末側の生データをクラウド側に集めないので、プライバシー保護や処理効率の面で利点がある。

d. 蒸留（Distillation）[37]：訓練済みの大きいニューラルネット（教師ネットワーク）の入出力データを用いて、小さいニューラルネットワーク（生徒ネットワーク）を訓練すると、生徒ネットワークの方が小さなサイズで精度も上がることが多い。

e. メタ学習（Meta Learning）：メタ学習は、学習方法を学習するものであり、ドメインやタスクの異なる複数のデータセットでの学習を通して、ターゲットとするドメインやタスクに合うような学習方法（パラメーターの決め方など）に関するメタ知識を獲得する枠組みである。代表的な手法としてMAML（Model-Agnostic Meta-Learning）[38] が知られている。

なお、トランスフォーマーやアテンションなどの自然言語処理で注目された深層学習関連技術は「2.1.2 言語・知識系のAI技術」で取り上げる。説明可能AI（XAI）、機械学習の公平性、機械学習のテスティング手法や品質保証、自動機械学習（AutoML）など、機械学習の安全性・信頼性を確保するための技術群は「2.1.4 AIソフトウェア工学」で取り上げる。ゲームAIやAI駆動型科学の話題については「2.1.6 AI・データ駆動型問題解決」で取り上げる。AIの倫理的・法的・社会的課題（ELSI）は「2.1.9 社会におけるAI」で取り上げる。

［注目すべき国内外のプロジェクト］

❶ Neural Radiance Field（NeRF：ナーフ）

複数の視点の画像[7]をもとに新たな視点の画像を生成するタスクはNovel View Synthesis（NVS）と呼ばれる。 NeRF[39] は、このNVSタスクをさまざまな撮影画像に対して、新たな視点からとても自然で高精細な画像を生成する。撮影されたものの質感や光の反射や透過などまでリアルに再現されているように見える。 UC Berkeleyの研究者らが2020年に発表し、コンピュータービジョンのトップ国際会議の一つ16th European Conference on Computer Vision（ECCV 2020）において注目され、Best Paper Honorable Mentionを受賞し、その後、多数の派生研究も生まれた。

NeRFでは、3次元座標と視線方向を入力として、その点の輝度と不透明度を出力する「場」（Field）を深層ニューラルネットワークで表現し、それを学習によって求める。新たな視点からの画像生成は、ボリュームレンダリング手法を用いており、光線上の各点の輝度・不透明度を積分することを、画像上の全ピクセルに対して行うことで実現している。

この研究から派生して、天候や時刻の異なる画像を入力としたり、写り込んだ人々を除去したりする研究[40]、複数の静止画を入力する代わりに単一カメラでの撮影動画から生成する研究[41), 42)]、レンダリングを高速化してリアルタイム生成を可能にする研究[43), 44)] など、さまざまな技術改良・拡張が進められている。NeRFのような技術を活用することで、さまざまな対象物の3次元モデルが容易に構築できたり、スポーツ

7 数十枚から数百枚の画像が用いられる。

やゲームを好む視点から観戦したり参加したりと、さまざまな応用が考えられる。

❷ PaLM–SayCan

PaLM–SayCan[45] は、Googleとロボット開発会社Everyday Robots（EDR）の共同研究プロジェクトである。2022年8月に発表されたシステムでは、自然言語による曖昧な要求に対して、その要求に対して何ができるか、ロボットが行動を選択して実行することが示された。例えば「飲み物をこぼしてしまった。助けてくれる？」という問いかけに対して、スポンジを取ってくるという行動を起こす。PaLM–SayCanを搭載したロボットは、自然言語による問いかけを解釈し、事前に定義されたスキルセットの中から、その状況で実行可能で、要求に対して有効な行動を選択する。自然言語処理、画像認識、動作生成などの機能が統合されている。101件の命令に対して、84%は適切な行動を計画し、74%は実行できたということである。

PaLM–SayCanでは、Googleの大規模言語モデルPaLM（Pathways Language Model）[46] を用いている。PaLMはトランスフォーマー型で、パラメーター数が5400億個という、2023年1月時点で世界最大規模の言語モデルである。Pathways[47] という分散学習インフラ上で動作する。

❸ RT–1

RT–1（Robotics Transformer 1）[48] も、PaLM–SayCanと同様にGoogleとEDRの共同研究プロジェクトである。ロボット実機を用いたトランスフォーマー大規模学習によって、深層学習ベースのロボット動作生成の汎用性を高めた。

トランスフォーマーベースの大規模モデルは、自然言語処理や画像・映像処理においてさまざまな機能が実現され、高い汎用性を示している。しかし、ロボット制御では、動作パターンの多様さや複雑さと、リアルタイム処理要求の高さが、大規模モデルの構築や利用の障壁となっていた。

これに対してRT–1では、EDRのロボット実機13台で17カ月にわたって、700以上のタスク[8]をカバーするような13万エピソードの動作データを収集して、大規模学習を実施した。その結果、700種類のタスクで97%の成功率を達成した。画像や動作データをトークン化して圧縮することや、トランスフォーマーのモデルパラメーターを19M個に抑えたことによって、実機での処理速度を確保している。シミュレーションではなく実機で大規模学習を行ったトランスフォーマー型モデルによって、ロボット動作生成の汎用性が高められた先進事例として注目される。コードはオープンソースとして公開されている。

❹ ムーンショット目標3

国内のプロジェクトでは、内閣府のムーンショット型研究開発制度において、「2050年までに、AIとロボットの共進化により、自ら学習・行動し人と共生するロボットを実現」がムーンショット目標3に掲げられ、2020年度に4件、2022年度に7件の研究開発プロジェクトが採択された。特に「一人に一台一生寄り添うスマートロボット」（プロジェクトマネージャー：菅野重樹、2020年度採択、略称：AIREC）では、「柔軟な機械ハードウェアと多様な仕事を学習できる独自のAIとを組み合わせたロボット進化技術を確立し、2050年には、家事、接客はもとより、人材不足が迫る福祉、医療などの現場で、人と一緒に活動できる汎用型AIロボットの実現により、人・ロボット共生社会を実現する」ことを目指しており、深層予測学習を含む先進AI技術をロボットに融合する研究開発が進められている。

8 動作（動詞）と対象物（名詞）のペア（例えば「皿をテーブルに置く」「瓶を開ける」など）を1タスクと数えている。

（5）科学技術的課題

❶ 現在の深層学習の問題克服、知覚・運動系AIと言語・知識系AIの統合

現在の深層学習に対して指摘されている問題をまとめると、次の3点になる[49)]。

a. 学習に大量の教師データや計算資源が必要であること。

b. 学習範囲外の状況に弱く、実世界状況への臨機応変な対応ができないこと。

c. パターン処理は強いが、意味理解・説明などの高次処理はできていないこと。

これらの問題のそれぞれに対して、問題克服のための直接的な対策が検討されているとともに、人間の知能のメカニズムからヒントを得ることAIを進化させようという研究も進められている。比較的短期には前者から個別の成果が得られると見込まれるが、長期的には後者による技術発展が進むことによって、問題a・b・cが合わせて解決されるだろうという期待も持たれる。「2.1.7 計算脳科学」の分析的アプローチによる知能研究と、「2.1.8 認知発達ロボティクス」の構成論的アプローチによる知能研究が、後者の基礎となる。このような取り組みによって目指される次世代AIの姿は、「2.1.1 知覚・運動系のAI技術」と「2.1.2 言語・知識系のAI技術」が統合され、知能の即応的ループと熟考的ループの両方が実現されたものになると考えられている[49)]。そのために具体的にどのような研究開発が行われているかについては「2.1.2 言語・知識系のAI技術」に記載している。

また、本節では詳しく触れてはいないが、深層学習を中心とする機械学習では、ブラックボックス問題、差別・バイアス問題、脆弱性問題、品質保証問題の発生が懸念されている。これらの問題の詳細と克服するための技術開発状況は「2.1.4 AIソフトウェア工学」や「2.1.9 社会におけるAI」にまとめたが、現在の深層学習はデータからのボトムアップなモデル化しか扱っていないことが、これらの問題の原因に深く関わっており、上で述べたような言語・知識系との統合がこれらの問題克服にもつながるはずである。

❷ 深層学習の理論的解明

深層学習は経験的に高い精度が得られているが、その理由は必ずしも明らかになっておらず、その理論的解明を目指した研究が活発に行われている[50)]。一般には、モデルのパラメーターが多くなると自由度が高くなり、訓練データに対する過剰適合（Overfitting）、つまり過学習（Overtraining）が起きて、訓練データに対して高い精度が得られてもテストデータでは精度が低下する（汎化性能が低下する）と考えられている。しかし、深層学習の場合、パラメーターが多くても自由度が高くならず、汎化性能が低下しないらしいということが分かってきた。また、一般に、凸関数では大域最適解を求めるのが容易だが、深層学習が扱うような非凸関数では局所最適解に捕まり、広域最適解を求めることが難しいと言われる。しかし、深層学習の場合、局所最適解が大域最適解に近い値になるらしいということも分かってきた。このように、理論面の知見が徐々に得られてきているが、深層学習の理論的解明は重要課題である。

❸ 機械学習向けコンピューティング技術

機械学習には大規模な計算資源が必要とされ、消費電力の増加も問題となっている。高い精度を得るために大量の訓練データが必要で、学習処理に要する時間も増大している。そのため、機械学習の処理の高速化と省電力化を可能にするコンピューティング技術の研究開発も強く求められている。本節ではその技術内容・開発状況についてほとんど触れていないが、機械学習に必要とされる演算を高速化するGPU（Graphics Processing Unit）などのアクセラレータープロセッサーの開発や、並列処理や省電力化も含めたシステム化技術などの開発も進められている。ニューロモルフィックやレザバーといった新たなコンピューティング技術への取り組みも進められている。本報告書においては「2.5 コンピューティングアーキテクチャー」で関連する研究開発動向を記載している。また、「AI白書2022」でも関連動向[51)]がまとめられている。さらに、量子コンピューティングを活用した量子機械学習の可能性も検討されている[52)]。今後、こういった新たなコンピューティング技術を活用した機械学習の技術開発も進展が期待される。

（6）その他の課題

❶ 国としてのAI戦略の推進と強化

「2.1 人工知能・ビッグデータ」冒頭の総論に書いたように、米中2強と言われる状況において、研究投資規模では米中に追いつくことが困難な日本にとって、日本の社会課題やポジションを踏まえ、日本の強みや勝ち筋を意識したAI研究開発の戦略を持つことが必要である。このため日本政府は「AI戦略2019」（2019年6月統合イノベーション戦略推進会議決定、2021年・2022年にアップデート）を策定した。この中では、AI人材育成やAIリテラシー教育も含めた教育改革、人間中心のAI社会原則、AI中核センター[9]を中心とする研究開発体制強化や「Trusted Quality AI」（信頼される高品質なAI）を掲げた研究開発戦略などが示されている。本節との関わりの深い面では、日本が産業的にも実績を持つ認識応用やロボットなどの強みを生かした実世界適用AIが挙げられている。本節に示した技術群や研究開発の方向性は、この戦略上も重要な位置付けで推進されているが、一層の強化のためデータ基盤や人材育成面で補強・留意したい点を述べる。

まず、本節で述べたような研究開発の推進には、機械学習の訓練・評価用の大規模データの構築・活用が不可欠である。画像認識を中心としたパターン認識については、既に述べたようにImageNetをはじめとする大規模データセットが公開され、利用されている。しかし、動作生成まで含めた即応的ループに関わるデータは未整備である。画像・映像データだけでなく、動作の履歴との対応やその意味情報も付与されたデータ（エクスペリエンスデータと呼ぶ[49)]）の構築を考えていく必要がある[10]。

人材面では、勢いのあるBig Tech企業が、機械学習を専門とする博士学生、ポスドク研究員、さらには大学教授も大量に囲い込もうと躍起になっており、人材獲得競争が熾烈になっている。中国やインドは、トップ人材を組織的に米国に送り、彼らが本国に戻って活躍するという流れを作り、活用してきた。 AI人材の教育・育成とともに、幅広い人材の獲得や引き留めのための施策も重要である。さらに、AI・機械学習はアルゴリズムを適切なソフトウェアとして実装してこそ威力を発揮する。日本は人材育成において、理論・アルゴリズムの基礎研究に加えて、ソフトウェア開発力においても強化施策が望まれる。

❷ 大規模コンピューティング基盤の共同利用施設とその継続的強化・整備

最新の機械学習技術は大量の計算資源を必要とし、その実行環境は大学の一研究室で確保できる規模ではなくなっている。大規模コンピューティング基盤の共同利用施設が不可欠であり、産業技術総合研究所のAI橋渡しクラウドABCI（AI Bridging Cloud Infrastructure）や理化学研究所のスーパーコンピューター「富岳」がこの役割を担っている。この継続的な強化・整備が極めて重要である。

❸ 顔認識技術や画像生成AIのELSI

AI全般のELSI（Ethical, Legal and Social Issues：倫理的・法的・社会的課題）面については「2.1.9 社会におけるAI」にて論じるが、ここでは、本節との関わりが深い問題として顔認識技術と画像生成AIのELSIについて取り上げる。

近年、顔認識技術がさまざまな応用に急速に広がっている。顔認識技術は以前からプライバシー保護の面からの懸念が指摘され、堅牢なセキュリティー確保や画像データを保存しないなどの対策が取られてきた。しかし、従来は顔認識機能の利用が、そのような対策面で意識の高い大手企業に限られていたのが、裾野が拡大し、幅広い人・企業が簡単に利用できるような状況になりつつある。しかも、プライバシー保護の

9 理化学研究所の革新知能統合研究センター（AIP）、産業技術総合研究所の人工知能研究センター（AIRC）、情報通信研究機構（NICT）のユニバーサルコミュニケーション研究所（UCRI）および脳情報通信融合研究センター（CiNet）

10 新型感染症パンデミックによって、さまざまな活動・サービスがオンライン/リモート化されてきており、エクスペリエンスデータを取りやすくなってきたと言えるのかもしれない。

懸念だけでなく、特定の人種やマイノリティーの人々を差別してしまうリスク（訓練データの質・量によっては、そういった人々の認識率が低く、場合によっては犯罪者と誤認識されやすいなど）も指摘されている。さらに、顔の微妙な表情から感情追跡が可能になると、人の内面をのぞき込むような使われ方の懸念も生じる。米国・欧州では顔認識に対する法規制の議論も起きており、技術的な対策検討や日本における政策検討が必要になりつつある。

また、画像生成AIを用いて、一般ユーザーが簡単に一見プロ並みの画像を生成できるようになりつつある。これを悪用したフェイク画像生成（Deepfakes）は社会問題化している。自動生成された画像や画像生成AIの学習に使われた画像の著作権に関わる問題も、現状の著作権の考え方で十分なのかという議論もある。アーティストの創作活動に新たな手法を提供するという側面もあれば、アーティストの仕事を奪ったり収益を減らしたりといった側面もある。ある人の顔画像を少しずつ変形させていったとき、肖像権はどこまで及ぶのかといった議論もある。急速に利用が拡大しつつある画像生成AIについて、ELSI面からの検討が求められる。なお、画像生成AIに限らず、いわゆるDeepfakeなどのフェイク画像・映像・音声の問題と対策については「2.1.5 人・AI協働と意思決定支援」で取り上げている。

（7）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	○	↗	理研AIP、産総研AIRC、NICTのAI中核センターを中心としたAI研究体制強化とともに、「AI戦略」の実行、JST事業・NEDO事業に加えてムーンショットプロジェクトも始まり、国主導の基礎研究推進策が強化されつつある。国際会議における採択率は米中2強には差を付けられているが、徐々に増えつつある。
	応用研究・開発	○	↗	日本の産業界は認識やロボット分野は実用化実績・性能などに強みがあり、特に顔認証ではNECがNISTベンチマークでトップの実績があり、世界的にも大きな存在感を示している。NEC、富士通、日立、パナソニック、NTT、Yahoo Japan!、楽天、リクルートなどがAI分野に積極的な技術開発投資を行っているほか、AIベンチャーも活発になりつつあり、特にPreferred Networksは深層学習・深層強化学習で高い技術力を示している。
米国	基礎研究	◎	↗	大学・企業とも機械学習の研究を非常に盛んに行っており、規模・質ともに世界をリードしている。国際会議における採択論文数も米中2強という状況である。 DARPAによる先進研究投資も注目に値する。また、基礎研究に必要なデータセットの多くが米国の大学・Big Tech企業によって公開されており、研究すべきタスクの設定や研究コミュニティーへの情報発信などでも中心的な役割を果たしている。
	応用研究・開発	◎	↗	Big Tech企業では有能な技術者を全世界から集め、基礎研究も応用研究・開発が盛んに行っている。大学との連携も活発で、大学でも起業を目指した応用研究や開発も数多く実施されている。 Big Tech企業以外にもAirbnb、Uberなど、AI技術を活用したベンチャー企業が次々と誕生し、国際的に成功を収めている。
欧州	基礎研究	○	→	オックスフォード大学、ETH、アムステルダム大学、INRIA、Max Planckなどに優秀な研究者が多数在籍、基礎研究力が高い。 Google DeepMind、Meta Research、Qualcommなどの企業の欧州研究部門での基礎研究もインパクトのある成果を挙げている。
	応用研究・開発	○	↗	ロンドンのGoogle DeepMind、ベルリンのAmazon Machine Learningなど、北米の企業の欧州支社が中心となり、応用研究開発を行っている。特にDeepMindが基礎・応用の両面で存在感を増している。ICMLやNeurIPSなどでの採択率もトップクラスである。

中国	基礎研究	○	↗	清華大学、MSRA（Microsoft Research Asia）などを中心に、国際会議での中国からの採択数が伸びている。画像認識コンペティションILSVRC 2015–2017で中国勢が上位獲得した実績がある。
	応用研究・開発	◎	↗	政府主導で重点AI分野を定め、医療分野はTencent、スマートシティーではAlibaba、自動運転はBaidu、音声認識はiFLYTEK、画像認識はSenseTimeをリード企業として選定し、政府がAI産業を後押ししている。これらの企業に加えてMSRAやHorizon Roboticsなども含め、産業界での応用研究開発が活発に推進されている。
韓国	基礎研究	△	→	ソウル大学、KAIST、POSTECHなどの主要大学にて関連の研究は行われているが、国際的に顕著なものは多くない。
	応用研究・開発	△	↗	Samsungなどで取り組まれていることに加えて、韓国の大企業の共同出資による知能情報技術研究院（AIRI）が2016年に設立され、応用研究が強化されつつある。

（註1）フェーズ

基礎研究：大学・国研などでの基礎研究の範囲

応用研究・開発：技術開発（プロトタイプの開発含む）の範囲

（註2）現状　※日本の現状を基準にした評価ではなく、CRDSの調査・見解による評価

◎：特に顕著な活動・成果が見えている　　○：顕著な活動・成果が見えている

△：顕著な活動・成果が見えていない　　×：特筆すべき活動・成果が見えていない

（註3）トレンド　※ここ1～2年の研究開発水準の変化

↗：上昇傾向、→：現状維持、↘：下降傾向

参考文献

1）Daniel Kahneman, *Thinking, Fast and Slow*（Farrar, Straus and Giroux, 2011）.（邦訳：村井章子訳,『ファスト＆スロー：あなたの意思はどのように決まるか?』, 早川書房, 2014年）

2）Tom M. Michell, *Machine Learning*（McGraw-Hill Science Engineering, 1997）.

3）科学技術振興機構 研究開発戦略センター,「研究開発の俯瞰報告書　システム・情報科学技術分野（2021年）」, CRDS-FY2020-FR-02（2021年3月）.

4）岡谷貴之,『深層学習 改訂第2版』（講談社, 2022年）.

5）Yann Le Cun, *Quand la machine apprend: La révolution des neurones artificiels et de l'apprentissage profond Broché*（Odile Jacob, 2019）.（邦訳：松尾豊翻訳・監修, 小川浩一翻訳,『ディープラーニング 学習する機械：ヤン・ルカン、人工知能を語る』, 講談社, 2021年）

6）岡野原大輔,『ディープラーニングを支える技術：「正解」を導くメカニズム［技術基礎］』『ディープラーニングを支える技術2：ニューラルネットワーク最大の謎』（技術評論社, 2022年）.

7）原田達也,『画像認識』（講談社, 2017）.

8）佐藤敦,「安全安心な社会を支える画像認識技術」,『人工知能』（人工知能学会誌）29巻5号（2014年9月）, pp. 448-455.

9）Kensho Hara, Hirokatsu Kataoka and Yutaka Satoh, "Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition", *Proceedings of the ICCV Workshop on Action, Gesture, and Emotion Recognition*（2017）.

10）Zhe Cao, et al., "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", arXiv：1611.08050 (2016).

11）尾形哲也,『ディープラーニングがロボットを変える』（日刊工業新聞社, 2017年）.

12）有木由香・他,「特集：強化学習最先端とロボティクス」,『日本ロボット学会誌』39巻7号（2021年9月）, pp. 570-636.

13）西川徹・岡野原大輔,『Learn or Die：死ぬ気で学べ, プリファードネットワークスの挑戦』(KADOKAWA, 2020年）.
14）堂前幸康・原田研介,「ロボットラーニングによる部品のピッキング」,『人工知能』(人工知能学会誌）35巻1号（2020年1月）, pp. 25-29.
15）松原崇充・鶴峯義久,「方策を滑らかに更新する深層強化学習と双腕ロボットによる布操作タスクへの適用」,『人工知能』(人工知能学会誌）35巻1号（2020年1月）, pp. 47-53.
16）Wenshuai Zhao, Jorge Peña Queralta and Tomi Westerlund, “Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey”, arXiv：2009.13303 (2020).
17）Alexey Dosovitskiy, et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, arXiv: 2010.11929 (2020).
18）Salman Khan, et al., “Transformers in Vision: A Survey”, *ACM Computing Surveys* Vol. 54, Issue 10s (January 2022), Article No. 200, pp. 1-41. DOI: 10.1145/3505244
19）Ashish Jaiswal, et al., “A Survey on Contrastive Self-supervised Learning”, arXiv：2011.00362 (2020).
20）Kaiming He, et al., “Masked Autoencoders Are Scalable Vision Learners”, arXiv: 2111.06377 (2021).
21）David Foster, *Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play* (O’reilly Media Inc., 2019).（邦題：松田晃一・小沼千絵訳,『生成Deep Learning：絵を描き、物語や音楽を作り、ゲームをプレイする』, オライリージャパン，2020年）.
22）Ian Goodfellow, et al., “Generative Adversarial Nets”, *Proceedings of 28th Conference on Neural Information Processing Systems*（NIPS 2014; Montréal, Canada, December 8-13, 2014), pp. 2672-2680.
23）Diederik P. Kingma and Max Welling, “Auto-Encoding Variational Bayes”, *Proceedings of the 2nd International Conference on Learning Representations*（ICLR 2014; Banff, Canada, April 14-16, 2014).
24）Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker, “Normalizing Flows: An Introduction and Review of Current Methods”, *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 43（November 2021), pp. 3964-3979. DOI: 10.1109/TPAMI.2020.2992934
25）Florinel-Alin Croitoru, et al., “Diffusion Models in Vision: A Survey”, arXiv: 2209.04747 (2022).
26）Aditya Ramesh, “Zero-Shot Text-to-Image Generation”, arXiv: 2102.12092 (2021).
27）Aditya Ramesh, “Hierarchical Text-Conditional Image Generation with CLIP Latents”, arXiv: 2204.06125 (2022).
28）Chitwan Saharia, et al., “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding”, arXiv: 2205.11487 (2022).
29）Jiahui Yu, et al., “Scaling Autoregressive Models for Content-Rich Text-to-Image Generation”, arXiv: 2206.10789 (2022).
30）Huiwen Chang, et al., “Muse: Text-To-Image Generation via Masked Generative Transformers”, arXiv: 2301.00704 (2023).
31）David Ha and Jürgen Schmidhuber, “World Models”, arXiv：1803.10122 (2018).
32）S. M. Ali Eslami, et al., “Neural scene representation and rendering”, *Science* Vol. 360, Issue 6394 (15 Jun 2018), pp. 1204-1210. DOI: 10.1126/science.aar6170

33）Pin-Chu Yang, et al., “Repeatable Folding Task by Humanoid Robot Worker using Deep Learning”, *IEEE Robotics and Automation Letters* Vol. 2, Issue 2（Nov. 2016）, pp. 397-403. DOI: 10.1109/LRA.2016.2633383
34）Ziwei Zhang, Peng Cui and Wenwu Zhu, “Deep Learning on Graphs: A Survey”, arXiv：1812.04202（2018）.
35）Ricky T. Q. Chen, et al., “Neural Ordinary Differential Equations”, *Proceedings of the 32nd Conference on Neural Information Processing Systems*（NeurIPS 2018; Montréal, Canada, December 2-8, 2018）.
36）Jakub Konečný, et al., “Federated Learning: Strategies for Improving Communication Efficiency”, arXiv：1610.05492（2016）.
37）Geoffrey Hinton, Oriol Vinyals and Jeff Dean, “Distilling the Knowledge in a Neural Network”, arXiv：1503.02531（2015）.
38）Chelsea Finn, Pieter Abbeel and Sergey Levine, “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks”, *Proceedings of the 34th International Conference on Machine Learning*（ICML 2017; Sydney, Australia, August 6-11, 2017）.
39）Ben Mildenhall, et al., “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis”, arXiv: 2003.08934（2020）.
40）Ricardo Martin-Brualla, et al., “NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections”, arXiv: 2008.02268（2020）.
41）Zhengqi Li, et al., “Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes”, *Proceedings of the 32nd IEEE / CVF Computer Vision and Pattern Recognition Conference*（CVPR 2021; June 19-25, 2021）.
42）Keunhong Park, et al., “HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields”, arXiv: 2106.13228（2021）.
43）Alex Yu, et al., “PlenOctrees for Real-time Rendering of Neural Radiance Fields”, arXiv: 2103.14024（2021）.
44）Stephan J. Garbin, et al., “FastNeRF: High-Fidelity Neural Rendering at 200FPS”, arXiv: 2103.10380（2021）.
45）Michael Ahn, et al., “Do As I Can, Not As I Say: Grounding Language in Robotic Affordances”, arXiv: 2204.01691（2022）.
46）Aakanksha Chowdhery, et al., “PaLM: Scaling Language Modeling with Pathways”, arXiv: 2204.02311（2022）.
47）Paul Barham, et al., “Pathways: Asynchronous Distributed Dataflow for ML”, arXiv: 2203.12533（2022）.
48）Anthony Brohan, et al., “RT-1: Robotics Transformer for Real-World Control at Scale”, arXiv: 2212.06817（2022）.
49）科学技術振興機構 研究開発戦略センター,「戦略プロポーザル：第４世代AIの研究開発―深層学習と知識・記号推論の融合―」, CRDS-FY2019-SP-08（2020年3月）.
50）今泉允聡,『深層学習の原理に迫る：数学の挑戦』（岩波書店, 2021年）.
51）情報処理推進機構AI白書編集委員会（編）,「開発基盤」,『AI白書2022』（KADOKAWA, 2022年）, pp. 129-158（2.6節）.
52）嶋田義皓,『量子コンピューティング：基本アルゴリズムから量子機械学習まで』（オーム社, 2020年）.