ＡＩを活用した医療診断システム・医療機器等に関する課題と提言 2017

平成２９年１２月２７日

ＡＩ専門部会
部 会 長　光石　衛
副部会長　大江　和彦

【目次】

はじめに

近年、AI すなわち人工知能を活用した新技術が活発に検討されており、医療関連分野への応用についても検討が進んでいる。たとえば医療画像診断の分野では、CT 等による画像の量が大量となり、医師による読影が限界に近づいている。医師による読影を補助する AI により、読影の効率化と見落としの防止が期待される。また日々の診療から得られるデータをいわゆるビッグデータとして取り扱い、そこから有用な情報を自動的に抽出する作業への AI 技術の活用も期待されている。

一方、AI といっても様々な技術があり、その応用形態には単なる「従来技術の応用であり、既存の医療機器規制の考え方に沿ってその安全性・有効性・信頼性を評価可能である機器・システム」から「AI 技術の特徴から、これまで医療機器の安全性・有効性・信頼性を評価手法では対応できない性格を持つ機器・システム」までかなりの幅がある。

また特に医療応用を考えた場合には、その応用形態によって考慮すべきリスクも様々に変化するため、その観点からの問題の整理も必要となる。医療機器・システムの製造者側のみならず、AI という新しい技術を導入する際の使用者側の適正使用とはどうあるべきかの議論も重要となる。

当専門部会では、AI 全般について俯瞰し、従来技術と異なる「AI としての新要素」を検討することにより、その特徴及びリスク、利用するための留意点を提供し、将来の医療機器審査や相談等に役立てることを目指した。

本報告書では厚生労働省の「保健医療分野におけるＡＩ活用推進懇談会報告書」で検討された「AI を活用した保健医療開発」・「ゲノム医療」・「創薬」などの分野はスコープ外としたが、PMDA の規制に関わるものに限らず、医療現場で使用される情報機器や情報システム、医療機器ソフトウエア、組み込みソフトウエアを広く対象とした。

この報告書は医療機器の審査や相談等に関わる規制当局及び開発に関わる医療機器製造者、ならびにこの新しい技術による医療機器・システムの使用者の参考になることを期待する。

１章 AI を活用した医療用システムの出現と課題

1956 年のダートマス会議を契機とする人工知能の研究は 2 度の興隆を経て、現在 3 度目の興隆に至っている。最初の人工知能ブームはその始まりの時であり、この時に基本的な方向性と技術の基礎が築かれた。ただ当時は、実用的なシステムを作れるほどの能力がコンピュータにはなく、研究分野の確立以上には広がらなかった。1980 年代の 2 度目の人工知能ブームにおいては、産業応用に注目が集まり、エキスパートシステムと呼ばれる実用システムが数多く作られた。コンピュータの能力が向上し、実用可能なシステムが作れるようになったことが 2 度目の人工知能ブームの大きな要因であった。エキスパートシステムは人間の知識をコンピュータに移転して活用することで人間の代替をする機能を提供することが狙いであった。しかし、人間の知識を取得、管理するところに限界があることが徐々に認識されていき、エキスパートシステムに対する産業界の興味は急速に失われていった。

2010 年頃より人工知能は 3 度目の注目を浴びるようになる。これが第 3 次人工知能ブームである。深層学習(deep learning)を中心とする機械学習や IBM Watson に代表される大規模知識の活用などが注目されている。3 度目の人工知能の興隆もコンピュータの能力向上が契機となっている。

まず、コンピュータの計算能力が劇的に向上している。後に述べる深層学習などニューラルネットワークに基づくアルゴリズムは、学習アルゴリズムとしては極めて効率が悪い。このため、当該アルゴリズムは小規模のニューラルネットワークにおける応用に留まっていたが、飛躍的な計算能力の向上によって、大規模かつ多層のニューラルネットワークを構築、利用できるようになった。次に、大規模データの収集と利用が可能になったことも第 3 次人工知能ブームを牽引した。インターネットの普及によって大規模な知識やデータがオンラインで収集可能になり、それを保持できるだけのストレージを簡単に入手することができるようなった[1]。これに伴い、さらにこのような大規模ストレージ、大規模計算を前提としたアルゴリズムが提案され、多様な可能性が示されている。

医療におけるコンピュータの利用も変化してきた。診断のために様々な測定機器が開発されてきたが、現状その多くにコンピュータによる処理を含んでいる。すなわち測定情報は電子データとして入手可能な状態となっており、上述の第 3 次人工知能ブームの要因の一つである大規模データの入手状況と同じである。この点で医療分野は人工知能活用、ことに機械学習の利用において大きな可能性を秘めている。大量の診療画像を使った例では人間の判断に匹敵する、あるいは上回る結果を出す人工知能が出現しているとも言われている。

診療データの分析に限らず、医療に関わる知識の活用においても人工知能が期待されている。第 2 次人工知能ブームにおいても診療への応用はそのターゲットとされていたが、診療のための知識を人間のエキスパートから直接入手しようとしたことが実現を困難にした。現在は、大量の論文や報告を自然言語処理

[1] 例えば、深層学習が注目される契機になった[Le2012]らの研究では、1 千万の画像データを 16,000 コアのコンピュータ群を 3 日間使って処理している。IBM Watson では約 70GB(本で換算すれば 2 億ページ相当)のデータを約 3000 個のプロセッサで処理している。

技術により処理して自動的に知識を抽出することで、診断等の医療における意思決定に対する応用も期待されている[2]。

このような人工知能システムの導入は、人間とコンピュータの関係を変えることが予想されている。これまではコンピュータや機器の振る舞いを人間が理解して使うという関係であった。しかし、今やコンピュータが人間の記憶や計算能力を超えているので、コンピュータが行ったことを理解することは原理的に困難である。特に機械学習アルゴリズムが作り出すプロセスはそもそも人間の思考プロセスと無関係であるので、人間がそのプロセスを理解すること自体が困難であることに留意する必要がある。

一方、医療においては、よりよい医療の実現のためにこれまでも様々な原理に基づく医薬品や医療機器を導入してきた。その際、その医薬品や医療機器が臨床上の効用が認められるのであれば、必ずしもその全てのメカニズムや原理が明らかでなくても社会に受け入れられることもあった。また臨床医が医療機器を使うにあたって、医療機器の意図した臨床上の意義を理解することは当然であるが、その物理的な原理や工学的技術の詳細を理解することまでは求められていない。ただ臨床上での利用は、医療機器への信頼に基づくものあり、その信頼を確保するために審査のあり方や評価方法が確立していることが重要であった。

すなわち、人工知能を活用した医療用システムの導入にあたっては、人工知能の特性を十分に理解してシステムのリスクに留意して利用できるように、審査のあり方や評価方法を構築することが重要である。これまでの医療機器と異なり、人工知能システムは大規模データ利用や機械学習アルゴリズムを利用することで、システムの振る舞いの柔軟性が高く、この点が課題となる。

[2] 例えば、東京大学医科学研究所では IBM の Watson を使うことで、白血病のタイプを自動的に判定できたことを報告している。Watson は 2000 万件以上の論文をデータとして読み込み、この判定をしたという。
https://www.nikkei.com/article/DGXLZO05697850U6A800C1000000/

２章　AI 技術の現状

この章では、まず、機械学習全般の説明をする。次に、機械学習の中で最近注目されている深層学習について概観し、医用画像解析や CAD 開発における深層学習の利用例、及び課題などについて述べる。最後に、今回の深層学習のブームをもたらした背景にあるビッグデータと大規模計算環境の構築などについて触れる。

２．１　機械学習

機械学習とは「明示的にプログラミングすることなく、コンピュータに行動させるようにする科学」と定義することができ、人工知能の一分野に位置付けられる。人工知能の草創期に行われたチェスを指すプログラム開発などを通じて、生まれてきた概念である。人間や動物が経験から学んで、後で再利用できる知識を獲得している過程をコンピュータで実現したいという動機から生まれた。現在では、様々なデータに基づいて、その中の潜在的な特徴（規則性やパターン）を発見することで、未知のデータについての予測を行う仕組みを総称している。

機械学習アルゴリズムとは、抽象的には、何らかのモデルのクラスを定義して、訓練データにおいて汎化誤差を最小化するモデルを同定するものである。そのモデル・クラスや探索メカニズムの違いから多様な機械学習アルゴリズムが存在する。機械学習アルゴリズムとしては、線形回帰などの統計的手法から生物の構造や振る舞いから着想を得たニューラルネットワーク・統計的分類など、多様なものが含まれる。

機械学習はその入力データ（訓練データ）の違いから以下のように分類される。

- 教師あり学習：予測したい目標の値（ラベル）が訓練データに含まれているとき。例えばクラシフィケーション（分類）問題では、分類がラベルとして付与されたデータを訓練データとして与え、未知のデータをいずれかのラベルに分類する分類器を生成する。
- 教師なし学習：予測したい目標の値が訓練データに含まれないとき。クラスタリング問題では、訓練データを類似するデータごとにまとまりを発見して、未知のデータをそのいずれかに分類する。
- 強化学習：試行錯誤を通じて環境へ適合するような学習。個々の状態に対しては明示的な教師入力はないが、行動と環境に依存して決まる報酬と呼ばれる手がかりを元に学習する。

本報告書で言及する代表的な機械学習手法としては以下のものがあるが、厳密な分類等については専門書を参照されたい。

1) 回帰分析：一般的には統計学の手法とみなされるが、教師あり学習の基本的な手法でもある。説明変数（独立変数）X と目的変数（従属変数）Y の間に Y=f(X)というモデルを当てはめることである。もっとも単純なものでは、線形の単回帰（Y が一次元）の場合で、Y=aX+b となる。統計学ではモデルが線形である場合が多いが、機械学習では非線形を含めたモデルが開発されている。ランダムフォレスト、ニューラルネットは非線形の回帰手法の一種である。
2) クラス分類(classification)：データを複数のクラス（グループ）に分けるこ

とで、回帰同様、教師あり学習であるが、Y=f(X)において Y が離散的である場合に相当する。

3) クラスタリング(clustering) : 代表的な教師なし学習の一種。与えられたデータ集合をいくつかの部分集合に分ける。その基準はデータ集合の中の関係性から求める。クラスタリングは大きく階層型と非階層型に分けられる。階層型クラスタリングは、さらに小さい集合から順に大きい集合へ成長させていく凝集型と逆に全体を分割していく分割型がある。非階層的なクラスタリングとしては k 平均法などがある。データ集合の中心の定義には様々な手法があり、それによって結果が異なる。

上記の具体的技術として、例えば、以下のようなものが開発されている。

・サポートベクターマシン(SVM) : 教師あり学習の手法の一つ。サポートベクターマシンでは出力が 2 クラスの分類問題を解く。一般に、ある n 次元データが与えられた時、2 クラス分類問題の解は n 次元空間の超平面となるが、このような超平面は無数に存在しうる。SVM ではこれをマージンという概念を導入することで、汎化誤差が小さくなるものを求める。

・k 平均法(k-means) : 与えられたデータを k 個の集合（クラスタ）に分ける手法。アルゴリズムとしては、初期にランダムに各データにクラスタを割り付け、各クラスタの中心を計算して、その中心との近さでデータが所属するクラスタを変更する。これを変化しなくなるまで繰り返す。

機械学習を適用するにあたって注意すべきいくつかの共通課題がある。

・過学習 : 機械学習のアルゴリズムは入力されたデータ（訓練データ）に適合する特徴を獲得する（訓練される）が、訓練の仕方によって、本来必要でない特徴まで学習してしまい、未知のデータに対して不適切な予測をするようになることがある。これを過学習と呼ぶ。過学習は訓練データがデータ全体から見て偏った分布であったり、典型的なものが含まれていなかった時に起きやすい。機械学習を適用するときには、過学習を避けることが必要であり、訓練データの作り方や適切な技法（正則化など）を用いる必要がある。

・次元の呪い : 機械学習の多くの場合、データ間の何らかの距離を定義して、それに基づいて計算を行う。ユークリッド距離のような距離尺度は、次元が大きくなると極端に識別性が悪くなる。これは高次元の超球においてはその体積が表層の部分に集中することからわかる。高次元のデータを扱う際にはこの点に注意すべき必要があり、必要ならば次元の削減などを行う。

２．２　深層学習

本節では、ニューラルネットワークに関する研究の歴史と現在の深層学習の概要、及び医用画像解析への応用例や課題などについて述べる。なお、2.2.1 項～2.2.4 項はニューラルネットワークに関する基本的な説明であり、これらに精通している場合は読み飛ばしてもかまわない。また、より詳しい内容について知りたい場合には、例えば、文献(1)などを参照されたい。

２．２．１　ニューラルネットワークとその歴史

ニューラルネットワークは機械学習の一分野であり、その始まりは 1940 年代頃まで遡ることができる。初期の代表的な研究は Rosenblatt によるパーセプト

ロンの研究[2]であり、60 年代にかけて盛んに研究された。しかし、Minsky らがパーセプトロンの限界を理論的に示したことなどがきっかけとなり[3]、最初のブームは去った。2 度目のブームは 80 年代であった。まず、1986 年に Rumelhart が階層型ネットワークに対して誤差逆伝播法と呼ばれる方法[3]を提案し[4]、幾つかの具体例において学習がうまく進むことを示した[4]。この論文などが契機となって多くの研究者が再びニューラルネットワークの研究に取り組むようになり、90 年代前半までに非常に多くの問題に適用された。しかし、性能を伸ばすために多層化（深層化）を進めると下記の問題が深刻化し、限界が明らかになるにつれて研究は下火になった。90 年代中頃にこれらの問題が無い（あるいは少ない）新しい機械学習のアルゴリズム（サポートベクターマシンなど）が登場すると、研究の流れは一気に変わった。

・過学習の問題
・局所最適解への収束の問題
・勾配消失問題

その後長い間、ニューラルネットワークの研究は表舞台から姿を消していたが、2006 年ごろの Hinton らの研究により、再び（3 度）注目されるようになった。今回（3 度目）のブームのキーワードは「深層学習」である。上述の多層化（深層化）の際の問題点が幾つかの発見的な工夫（後述）により解決され、かつ、コンテストを含む幾つかの具体的問題に対し学習がうまく進むことが示され、研究が爆発的に盛んになった。また、深層化の研究が広がった背景には、ビッグデータの存在と GPU やクラウドなどの分散並列計算環境の整備があることも忘れてはならない（2.3 節参照）。

ニューラルネットワークのタイプは、大きく分けて、階層型ネットワークのような確定的モデルと、ボルツマンマシンのような確率的モデルがある。また、ネットワークの学習アルゴリズムも、教師あり、教師なし、半教師などの学習アルゴリズム以外に強化学習などがある。これらの基本的な考え方は過去に提案されたものばかりであり、今回のブームでは、深層化したネットワークを対象に再び研究が盛んになっている。以下では、階層型ネットワークとその学習アルゴリズムについて概説したのち、もう一つの確定的モデルである自己符号化器について紹介し、最後に確率的モデルであるボルツマンマシンについて述べる。

２．２．２　階層型ニューラルネットワーク

現在、深層学習の研究・開発でもっとも多く利用されているのは階層型ネットワークである。図 2-1 に例を示した。左側の入力層に与えられた信号 $\boldsymbol{x}$ は、隠れ層（中間層とも呼ばれる）を経て出力層へと伝わり、出力 $\boldsymbol{y}$ が計算される。ここではその学習アルゴリズムと併せて説明する。

---

[3] 誤差逆伝播法は最適化法の中では勾配法に分類され、勾配法をニューラルネットワークに適用する研究は 1960 年代にルーツがある。

[4] この論文[4]のもう一つの貢献は、5 層のネットワークの中間層において、問題に適した内部表現が自動的に獲得されていることを示した点である。現在の深層学習で注目されている内部表現に関する研究は、この時期にもすでに行われていた

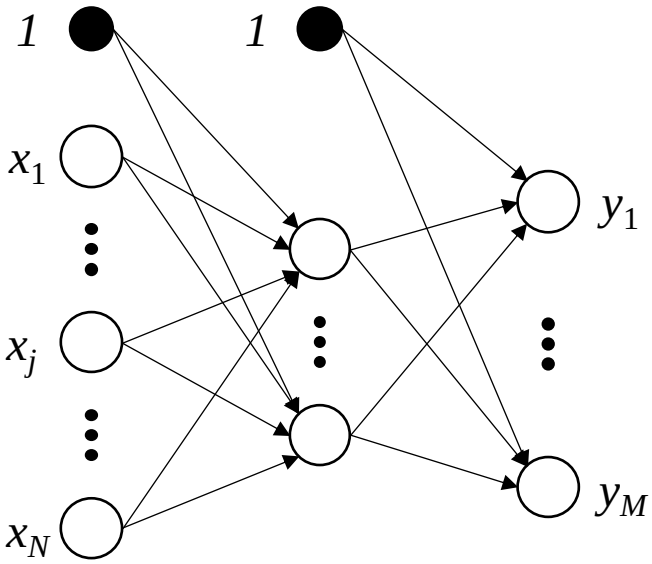


図 2-1　階層型ニューラルネットワーク（3 層の場合）

初期の代表例は入力層と出力層のみの単純パーセプトロンと呼ばれるシンプルなネットワークである。このネットワークの重みは教師あり学習により決定され、判定を誤った学習データのみに対して重みを修正する。このとき、学習データに対する誤りは単調に減少し、有限回の学習で必ず収束することが知られている（パーセプトロンの収束定理）。また、入力層の前に前処理用の連合層（重みは固定）を追加することで、より複雑な問題が解けることも示された。しかし、すでに述べた通り、Minsky らがある種の問題（連結性判定など）は原理的に解けないことを示した結果[3]、研究は下火になった。

2 度目のブームの際に最もよく使われたネットワークに、3 層以上の階層型ネットワークがある。Rumerhart がこのネットワークに対して誤差逆伝搬法と呼ばれる教師あり学習アルゴリズムを提案し、具体的な問題に対して良い性能が得られることを示したことはすでに述べた通りである。現在もっとも広く使われている階層型ネットワークの一つに畳み込みニューラルネットワーク（Convolutional Neural Network（CNN））があり、この学習アルゴリズムとしても使われている。ちなみに、CNN のルーツは福島らのネオコグニトロン[5]と LeCun らの LeNet[6]にある。これらは、いずれも層数の多いネットワークであり、ネットワークの構造を人が上手に設計することでネットワークの複雑化を避け、2.2.1 項で述べた深層化による問題を回避した成功例だといえる。最近は、より規模の大きい深層化したネットワークも登場しているが、その構造の設計は容易ではない。特に大規模なネットワークでは、プラトーと呼ばれる損失関数が平坦な領域が生じ、勾配消失問題なども深刻化する。これらの問題や過学習、局所収束の問題などを回避するために、以下の工夫が提案されている。

・勾配法の工夫：自然勾配法、確率的勾配降下法、Adam 法など
・正則化の導入：DropOut など
・活性化関数の工夫： ReLU、MaxOut など
・重みの初期値の工夫：転移学習（例：事前に別のデータベースで学習）や事

前学習（Pretraining．2.2.3 項や 2.2.4 項参照）など
・ネットワーク構造の設計（ハイパーパラメータの決定）：ランダム探索、ベイズ最適化など

ここで、上記の学習法を一度の重みの更新に用いるデータ数によって区別する表現として、バッチ学習とオンライン学習がある。これらについて補足する[5]。例えば、誤差逆伝搬法を用いてネットワークの学習を行う場合、バッチ学習ではすべての学習データ[6]を用いて勾配を計算する。それに対して、パーセプトロンの学習則や確率的勾配法のように、一つや少数のデータを用いて重みを更新する方法はオンライン学習と呼ばれる。この確率的勾配法は、学習データに冗長性がある場合に有効であるとされており、ビッグデータを用いた学習においてよく用いられる。しかし、少数のデータでは勾配の推定が不安定になることがあるため、ある程度まとまったデータを用いて重みを更新することがあり、これはミニバッチ法と呼ばれる。

次に、半教師あり学習と強化学習についても補足しておく[5]。半教師あり学習は、ラベル（正解のクラス名など．アノテーションデータと呼ばれることもある）が付いているデータのみならず、ラベルの無いデータも用いて学習するアルゴリズムである。ビッグデータの中にはラベルが付いていないデータが含まれることもよくあるため、一つの有効な学習法であるといえる。また、強化学習は、報酬が最大になるように学習するアルゴリズムである。教師あり学習との違いは、最適な出力は試行錯誤を通じてアルゴリズムが自ら発見しなければならない点であり、ある種のゲームをニューラルネットワークと強化学習を使って学習した例が有名である。

最後に、最近のネットワークの大規模化の傾向について触れる。大規模化の一つの方向は層数を増やす深層化であり、2016 年時点で 1202 層のネットワークの例が報告されている(7)（補足：過学習を起こしてしまい、同じ論文内で比較している 110 層の方が性能は高い。このことから、深層化した場合の学習の困難さが分かる）。現在も層数は増加し続けているが、増えるほど学習は難しくなるため、新たな工夫が必要になる。また、大規模化の別の方向として、複数のネットワークを組み合わせた例も存在する。代表例として Generative Adversarial Networks（GAN）があり(8)、生成器と識別器の二つのネットワークを組み合わせた興味深い構造を持つ。ここで識別器には本物のデータと生成器が作る偽のデータとを正しく区別させ、同時に識別器が誤識別しやすい偽データを作るように生成器を学習させる。その結果、高い精度の生成器が得られることになる。ここで、生成器や識別器には、現時点では階層型ネットワーク（例えば CNN など）が使われることが多く、今後も様々な拡張が期待されている。また、ネットワークの大規模化のためには、学習技術の工夫以外に、データベース化技術や分散並

[5] アルゴリズムの学習のためのデータは学習データ。性能評価のためのデータはテストデータであるが、それ以外に、ハイパーパラメータと呼ばれる値（例：深層学習の場合はネットワークの構造（層数やフィルタ数など）及び学習回数など）を選択するためのデータセットが必要となることがある。これはバリデーションデータと呼ばれ、学習用データやテストデータとは区別する必要がある。

[6] 学習に関するこれらの表現はニューラルネットワークに固有のものではなく、機械学習アルゴリズム全般に共通に用いられる表現である。

列計算環境のさらなる進歩も必須である。これらが一体となって技術的に進化し続けることが大規模化の鍵になると考えられる[7]。

２．２．３　自己符号化器

1988 年 Cottrel らは、砂時計タイプと呼ばれる、入力層と出力層が同じノード数で、中間層がそれらよりも少ないネットワークを用意し、入力を教師信号として誤差逆伝搬法により学習を行った（注：入力しか用いていないので教師なし学習モデル）。その結果、中間層には、入力信号の情報を保存しつつ低次元化した特徴が得られることを明らかにした(9)。ネットワークの前半の層が符号化器、後半の層が復号化器の役割を果たしていることから、自己符号化器と呼ばれる。

最近の深層学習の研究ではこの自己符号化器においても多層化が進み、2006 年の Hinton らによる制限ボルツマンマシンを積み重ねた深層自己符号化器(10) 、Bengio らの階層型ネットワークを積み重ねた積層自己符号化器(11)などが登場した。これらは、自己符号化器の重みの初期値を貪欲的に決める事前学習を提案したことで有名である。また、Ng らが行った、3 つの 3 層のサブネットワークを重ねた 9 層のネットワークに 1000 万枚の YouTube の画像を学習させた研究では、猫の顔などに特異的に応答するノードが形成されたことが報告され(12)、大きな話題となった。その他、スパース性を導入したスパース自己符号化器、入力に雑音を与えることで入力の変動に対して頑健な特徴を得ることが可能な雑音除去自己符号化器などが報告されている。

自己符号化器の深層化はあまり進んでおらず、一つの例で 12 層のネットワークを用いた研究が報告されてい(13)。

２．２．４　ボルツマンマシン

1985 年、Ackley や Hinton らは、確率的相互結合型ネットワークモデル、ボルツマンマシンを提案した。これは、連想記憶のモデルとして知られるホップフィールド・ネットワークを確率的ダイナミクスを持つように拡張したものである。ボルツマンマシンは、広くはマルコフ確率場と呼ばれるタイプの学習モデルのエネルギー関数を持ち、ノード全体の発火と非発火の同時分布がギブス分布になることが知られている。このボルツマンマシンの学習には最尤推定が用いられており、そこから導かれる学習方程式を解析的に解くことは難しいため、実際には勾配法を用いてネットワークのバイアスや重みなどを決定する。しかし、それほど大きくないネットワークの場合にも、組み合わせ爆発により勾配の計算が非常に困難であったため、理論的には興味深いネットワークであったが、あまり用いられてこなかった。その問題を解決するために提案されたのが、ギブス

---

[7] 本文では述べていないが、比較的重要なその他の技術についても触れておく。まず、階層型ネットワークは教師なしの学習アルゴリズムによっても学習可能である。その代表例が競合学習であり、入力が幾つかのクラスタから構成されている場合には、教師信号なしに自動的にそれらを見つけることが可能となる。また、その他の階層型ネットワークとして、再帰ニューラルネット、回帰結合ニューラルネットなどがある。前者は再帰的な構造を持つネットワークであり、後者は例えば時系列データに対して前時刻の情報を現時刻の処理に送るための回帰結合入力を持つネットワークであるが、詳細は割愛する。

サンプリング、平均場近似などの近似学習法であり、最近では、コントラスティブ・ダイバージェンス、確率伝搬法などの優れた方法も提案され、大規模なネットワークにおいても学習が可能となった。

最近の深層化の研究で重要なネットワークに制限ボルツマンマシン[14]がある。これは、完全二部グラフの構造を持つボルツマンマシンであり、二部グラフの二層のうち、一層は可視層、他方は隠れ層である。深層学習のブームをもたらした 2006 年の Hinton らの研究[10]は、この制限ボルツマンマシンを基礎とした研究であった。そこでは、自己符号化器の隣接層間をそれぞれ別の制限ボルツマンマシンとみなし、下層から逐次的に学習させる層ごとの貪欲学習によって重みの初期値を決定し（事前学習）、その後全体を学習する。この事前学習のアイディアは、2.2.3 項で述べた Bengio らの階層型ネットワークを利用した自己符号化器以外にも、下記の深層ボルツマンマシンや深層信念ネットワークなどの事前学習にも用いられている。

深層ボルツマンマシン[15]と深層信念ネットワーク[16]について述べる。深層信念ネットワークは、最初の深層学習のモデルであり、深層ボルツマンマシンはそれを拡張したものといわれている。両者の違いは、深層信念ネットワークの最上段の隠れ層につながる結合以外は、可視層のある下段方向へ矢印が向いた有向の結合をもつが、深層ボルツマンマシンでは層間の結合はすべて無向である。いずれもほぼ同じ目的で利用され、例えば、画像の生成モデルや自己符号化器などへの応用例が報告されている。学習は、まず層ごとの事前学習を行い、その後、全体で最尤推定を行う。計算の効率化のために前述のギブスサンプリングやコントラスティブ・ダイバージェンスなどを用いる。

ボルツマンマシンなどの確率的ネットワークは、確定的モデル（階層型ネットワークなど）と比較すると計算コストの問題が依然として大きく、深層化や応用は進んでない。しかし、性能面で他よりも優れているという報告もあり、今後の研究の進展が期待されている。

２．２．５　医用画像解析や CAD 開発における深層学習の利用例

医用画像解析に深層学習を応用した研究論文数は、2015 年より急激に増加し、主要な学会やジャーナルの学術論文数は 2016 年末までに 300 を超えた[17]。発表数の多い処理は臓器のセグメンテーションであり、次いで腫瘍や出血個所などの病変部位の検出、症例（画像）や病変の分類である。使われているネットワークは CNN やそれを拡張したもの（例：fully convolutional network (FCN)やそれをさらに拡張した U-net など）が多く、全体の 8 割以上を占めているが、問題によって再帰ニューラルネットや制限ボルツマンマシンなども用いられている。画像モダリティは現時点では MRI、顕微鏡画像、CT が圧倒的に多く、部位は脳や病理組織の一部を対象としたものが多くみられる。その理由は、これらの画像や部位の性質が深層学習に適していたということ以外に、ネットワークの学習に必要な大量のデータが存在していたことも重要な要因である。現在の撮影技術やデータベース化技術の進化を踏まえると、その他のモダリティや部位へ研究が広がるのは時間の問題である。また、適用される問題の範囲も、上で述べたセグメンテーション、検出、分類以外に、位置合わせ、画像検索、画像変換（強調も含む）、文書処理との組み合わせなど、多岐に渡っている。これらの拡大傾

向は今後も続くと予想される。また、それに合わせて、用いられるネットワークも、multi-stream CNN や GAN のように複数のネットワークを組み合わせた大規模なものへシフトしてゆき、その種類もますます増えると予想される。

最近、ある医用画像解析システムが米国食品医薬品局（Food and Drug Administration(FDA)）のクリアランスを受けた[18]。ARTERYS から発売された心臓の血流解析を行うシステムである。そこでは、心臓のセグメンテーションなどをクラウドベースの深層学習により行っている。また、 QView Medical Inc.の乳房超音波像用の CAD は、Concurrent Reader タイプとして初めて FDA の承認を受け、ここにも深層学習が用いられている[19]。これらのシステムや、現在までのさまざまなコンテストやベンチマークテストでの優れた成績を踏まえると、今後も深層学習を用いた医用画像解析システムや CAD の開発が続くと予想される。

２．２．６　深層学習の課題

以下に、深層学習を医用画像解析へ応用した際に注意すべき点についてまとめる。詳細は 3 章でも述べるので、ここでは概要について述べる。

1) ブラックボックスとしての性質

これはニューラルネットワーク（深層学習）特有の性質である。ネットワークの内部は、多くの場合ブラックボックスであり、例えばある診断が導かれたとしても、その診断の根拠の説明が困難である。また、単に説明できないだけでなく、深層化したネットワークのように入出力関係が高度に非線形な場合、未知データに対する振る舞いが予想困難である。そのため、トラブルが生じた場合に原因の切り分けができず、ネットワークのどの部分を改良すれば良いかが一般には明瞭ではないという問題を抱えている。

2) 継続的な性能の変化

ビッグデータと分散並列計算環境の整備によって、技術的には任意の時刻にネットワークの再学習が可能となった。注意すべき点は、再学習の結果、システムの認証当初の性能から変化することであり、この変化が良くなる方向だけでなく、その反対もありうる点である。なおこの特性は、深層学習だけでなく、他の機械学習にも当てはまる。

3) データベースの信頼性の問題

例えば、転移学習などによってネットワークの性能が向上することが知られており、そこでは、医用画像以外のデータによって学習が行われる。場合によっては、素性の分からないデータが含まれる可能性もある。また、半教師学習や Weak Label を扱う学習では、ラベルが存在しない、あるいは不正確なラベルのデータも様々な仮定をおいて学習に用いられることもある。これらの問題は他の機械学習にも当てはまるが、学習に非常に多くのデータが必要な深層学習では特に重要である。なお，ここでは学習用データの信頼性についてしか触れていないが，バリデーションやテスト用のデータについても信頼性の問題は本質的に存在する。

最後に、深層学習の将来や残された課題について述べる。深層学習の研究は日進月歩であり、現時点で深層学習の将来を予想することは非常に難しい。深層化

したニューラルネットワークが、最近の研究によって従来よりも扱いやすくなり、性能が大きく向上したことは確かである。しかし、この先、例えば技術的特異点[20]に結びつくような更なる飛躍を引き起こすか否かまでは分からない。特に、今回の深層学習のブームを支えている技術が主に発見的な工夫によるものであり、理論的な面からの解明は十分に進んでいるとはいえず、このことが見通しを不透明にしている。例えば、様々な工夫を積み重ね、試行錯誤を繰り返すことによって深層化したネットワークの性能が向上することは多くの例によって実証されているが、深層化することによりなぜ性能が向上するかについての理論的な説明はまだ不十分である。また、単純に層数を増やしても思うようには性能が伸びない（一部では過学習により低下する）例も報告されている[7,13]。一方、今回のブームの背景にあるビッグデータと計算機パワーの進化は当面続くと予想され、このことは深層学習にとっては追い風になっている。そこに理論的な解明が加わることで、本質的なブレークスルーにつながる可能性も残されている。深層学習は将来性のある技術の一つであることには間違いないため、今後の研究の動向についても引き続き注視する必要がある。

２．３　ビッグデータと大規模計算環境による AI 精度の向上

機械学習に基づく AI では、その推論を行うためには大量のデータが必要となる。手作業により特徴量を設計し、それを統計的機械学習、あるいは、SVM、adaBoost などの分類器を用いて識別すべきパターンを分類するのと比較し、deep neural network、あるいは convolutional neural network を用いる場合には大量の学習用データが必要となる。AI 推論の結果は、一般的には質のよい学習用データをどれだけ準備できるかに依存している。Data Augmentation と呼ばれるデータを加工することで学習データを増加させることや、人工的にパターンを生成させることでデータを増加させることも行われる。

・ストレージの大容量化

ビッグデータによる AI には、データストレージ技術の進歩も見逃せないといえよう。HDD（ハードディスクドライブ）に至っては 1 台あたり 10TB（テラバイト＝1000 ギガバイト）の容量を超えるものが登場し、半導体メモリを用いた SSD（ソリッドステートドライブ）なども 1 台あたり数 TB の容量をもつものができるようになっている。このようなデバイスもビッグデータによる機械学習の進化に大きく寄与しているといえよう。また、Amazon S3[21]などの大規模クラウドストレージサービスも機械学習による AI に大きな影響を与えている。

・ネットワーク利用による大量データの収集

当然のことであるが、ネットワーク基盤インフラの整備も大量データ収集について影響を与えている。Web 上で公開されている（ホームページに掲載されている）画像を、「ロボット」と呼ばれるソフトを用いてホームページを巡回しながら集め、それらを深層学習に利用することも行われるようになっている。特に、一般物体認識[22]と呼ばれる分野では、web を通じた学習画像の取得が盛んに行われている。

教師あり機械学習には、大量のデータに対していかに効率的に教師ラベルを付与するかが重要な観点となる。この教師ラベル付与にもインターネットを活用して、特定多数の人にラベル付与作業を委託するクラウドソーシングが機械

学習の分野では行われている。Amazon Mechanical Turk[23]などはその代表的なクラウドソーシングのサービスであるといえよう。医療機器向けの AI において教師データを取得するために、ネットワーク基盤とクラウドソーシングを活用することも十分に考えられる。

・GPU による深層学習における学習過程の高速化

現在の畳み込みニューラルネットワークでは、超高次元空間における最適化問題を解く必要があり、莫大な計算時間が必要とされている。コンピュータグラフィック向けのプロセッサ（Graphics Processing Unit：グラフィックス・プロセシング・ユニット）のアーキテクチャを汎用計算に利用できるようにした GPGPU（General-purpose computing on GPU）が深層学習の発展に大きく寄与している。GPU がなければ深層学習研究がここまで加速しなかったといっても過言ではない。GPU はその内部に超並列アーキテクチャを持っている。また、GPU を大量導入したスーパーコンピュータなども大学情報基盤センターなどで導入されている。

深層学習研究や利用の促進に貢献しているものとしては、Caffe[24]、Tensorflow[25]、 Keras[26]、 Theano[27]、 Chainer[28]などのソフトウエアライブラリの充実も考えられる。また、Python などのスクリプト言語も深層学習の利用を身近なものとしている。これらのライブラリは、それぞれのソフトウエアの開発コミュニティによって日々拡張が続けられており、日々アップデートがなされている。なお、これらのソフトウエアは医療機器として利用することを意図して開発されたものでないことに十分な注意が必要である。

参考文献

(1) 麻生英樹、安田宗樹、前田新一、岡野原大輔、岡谷貴之、保陽太郎、ボレガラ・ダヌシカ :人工知能学会監修：深層学習、近代科学社、東京、2015

(2) F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review、 65(6):386–408、 1958.

(3) M. Minsky and S. Papert、 1972 (2nd edition with corrections、 first edition 1969) Perceptrons: An Introduction to Computational Geometry、 The MIT Press、 Cambridge MA

(4) Rumelhart、 David E.; Hinton、 Geoffrey E.、 Williams、 Ronald J. (8 October 1986). “Learning representations by back-propagating errors”. Nature 323 (6088): 533–536

(5) Fukushima、 K. (1980). "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". Biol. Cybern. 36: 193–202.

(6) Y. LeCun、 B. Boser、 J. S. Denker、 D. Henderson、 R. E. Howard、 W. Hubbard and L. D. Jackel: Backpropagation Applied to Handwritten Zip Code Recognition、 Neural Computation、 1(4):541-551、 Winter 1989.

(7) Kaiming He、 Xiangyu Zhang、 Shaoqing Ren、 Jian Sun: Deep Residual Learning for Image Recognition、 CVPR、 pp.770-778、 2016

(8) Goodfellow、 Ian J.; Pouget-Abadie、 Jean; Mirza、 Mehdi; Xu、 Bing; Warde-Farley、 David; Ozair、 Sherjil; Courville、 Aaron; Bengio、 Yoshua (2014). "Generative Adversarial Networks".

(9) Cottrell、 Garrison and Munro、 Paul. Principal components analysis of images via back propagation. In Proc. SPIE、 1001、 pp.1070-1076、 1988.
(10) G. E. Hinton、 R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. Science 28 Jul 2006: Vol. 313、 Issue 5786、 pp. 504-507
(11) Yoshua Bengio、 Pascal Lamblin、 Dan Popovici、 and Hugo Larochelle. 2006. Greedy layer-wise training of deep networks. In Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS'06)、 B. Schölkopf、 J. C. Platt、 and T. Hoffman (Eds.). MIT Press、 Cambridge、 MA、 USA、 153-160.
(12) Marc'aurelio Ranzato、 Rajat Monga、 Matthieu Devin、 Kai Chen、 Greg Corrado、 Jeff Dean、 Quoc V. Le and Andrew Y. Ng: Building high-level features using large scale unsupervised learning.Proceedings of the 29th International Conference on Machine Learning (ICML-12)、 pp. 81-88、 2012
(13) Oleksii Kuchaiev 、 Boris Ginsburg: Training Deep AutoEncoders for Collaborative Filtering、 arXiv:1708.01715、 2017
(14) Smolensky、 Paul (1986). "Chapter 6: Information Processing in Dynamical Systems: Foundations of Harmony Theory" (PDF). In Rumelhart、 David E.; McLelland、 James L. Parallel Distributed Processing: Explorations in the Microstructure of Cognition、 Volume 1: Foundations. MIT Press. pp. 194–281.
(15) Ruslan Salakhutdinov and Geoffrey Hinton: Deep Boltzmann Machines. In AISTATS、 pp.448-455、 2009.
(16) Geoffrey E. Hinton、 Simon Osindero、 and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18、 7 (July 2006)、 1527-1554.
(17) Geert Litjens、 Thijs Kooi、 Babak Ehteshami Bejnordi、 Arnaud Arindra Adiyoso Setio、 Francesco Ciompi、 Mohsen Ghafoorian、 Jeroen A.W.M. van der Laak、 Bram van Ginneken、 Clara I. Sánchez: A Survey on Deep Learning in Medical Image Analysis、 Med Image Anal. 2017 Jul 26;42:60-88
(18) http://rsna.vporoom.com/2016-11-17-Arterys-to-Showcase-New-FDA-Cleared-4D-Flow-Software-Platform-at-RSNA-2016
(19) http://rsna.vporoom.com/QviewMedical/index.php?s=35665&item=122625
(20) Ray Kurzweil: The Singularity Is Near: When Humans Transcend Biology 1st Edition. The Viking Press; 1st edition (September 22、 2005)
(21) https://aws.amazon.com/jp/s3/
(22) 柳井啓司、 “一般物体認識、”情報処理学会論文誌：コンピュータビジョンとイメージメディア、 Vol.48、 No. SIG 16 (CVIM19)、 pp. 1-24、 2007
(23) https://www.mturk.com
(24) Yangqing Jia、 Evan Shelhamer、 Jeff Donahue、 Sergey Karayev、 Jonathan Long、 Ross Girshick、 Sergio Guadarrama、 Trevor Darrell、

“Caffe: Convolutional Architecture for Fast Feature Embedding、”
arXiv:1408.5093、 2014
(25) https://www.tensorflow.org/
(26) https://keras.io/
(27) http://deeplearning.net/software/theano/
(28) https://chainer.org/

３章　AI 医療システムのレギュラトリーサイエンス

本章では、2 章で説明した AI 技術を搭載する医療機器とその他の医療システム（両方を AI 医療システムと呼ぶ）について診断支援・治療支援での具体的な利用形態を想定してレギュラトリーサイエンス的な見方を示す。

３．１　AI 医療システムの特徴

AI 医療システムが従来の医療、そこで用いられる医療機器等と異なるのは次の 3 点である。

1) AI 医療システムは学習により性能等が変化しうる（可塑性）。これまでも性能が変化する医療機器は存在したが、従来の医療機器とは質的に異なる可塑性を有しうると考える。医療機器に該当する AI 医療システムの場合、製造販売承認を受けた事項からの変化は、一部変更承認申請の要否(1)や品目の同一性などに影響する。
2) 深層学習などの方法による AI 医療システムは、AI の出力の予測や解釈が難しいことがある（ブラックボックスとしての性質）。
3) 将来、AI の支援レベルが高度化すると患者と医師等の関係性が従来と変わってくる可能性がある（将来の高度な自律能）。

本章では、主に 1) と 2) に着目して AI 医療システムのレギュラトリーサイエンスの側面を論じる。将来の高度な自律能については、4 章で言及する。

３．１．１　AI の学習のさせ方による分類

AI 医療システムは学習（再学習を含む）によって出力が変化しうる。学習に関して従来の医療機器等と AI 医療システムの相違に繋がるのは次の 3 点である

1) 学習を行わせるタイミング
2) 学習を誰が行わせるか
3) どのようなデータを学習に使うか。データセットの特性と信頼性

このうち、1) 学習を行わせるタイミング、2) 学習を誰が行わせるかを中心に分類を試みる。3) については、3.3 節で詳説している。また再学習については 3.5 節で詳説している。

①　学習を行なわせるタイミング

AI 医療システムに学習を行わせるタイミングはその出荷・サービス開始前に終了させるか、出荷・サービス開始後も継続するかに大別できる。さらに、その学習が AI 医療システムの性能を変化させるか否かに大別できる（表 1）。

表 1：学習と性能変化のタイミング

		出荷・サービス開始後の性能変化	
		しない	する
出荷・サービス開始後の学習	しない	従来の医療機器と同等	（該当なし）
	する	サービスに供しているシステムの出力は固定されているが、学習は継続している。 バージョンアップの際にその学習の成果をまとめて反映して性能が変化させることが想定される。この場合、開発者が従来通り risk management をすることができ、学習に使用するデータセットを開発者がチェックすることも可能。	サービスに供しているシステムが学習に伴って性能が徐々に変化する。 学習の結果によっては性能が却って悪くなる可能性がある。また、次項（3.1.1.2 項）で述べる通り「誰がどのデータで学習させるか」の問題がある。学習に使用するデータセットを開発者がコントロールするのが難しいケースも想定される。

医療機器に該当する AI 医療システムの性能が変化する場合は[(1)]が述べるように、製造販売承認事項との整合性が問題となる。法的な問題と並ぶ問題が、開発者による risk management が難しくなることである。この問題は医療機器に該当しない場合にも存在する。懸念は以下である。

1) 学習の結果が却って性能を悪くする方向に作用する場合。
2) 学習の結果が標榜する性能等を超えることで品目の同一性を保てなくなる問題を生じる場合（医療機器の場合）。

ただし、1) のような性能低下が常に重大な問題となるとは限らない。医師等が性能低下に気がついて適切に対処できる場合は、結果として性能の悪化に伴う健康被害が回避される（ただしそのような製品の市場での評判が悪くなることが予想される）。

医療用途ではないが、「かな漢字変換」の場合がわかりやすい（表 2）。かな漢字変換には、変換アルゴリズムがローカルの PC 側で独立して動作する場合と、クラウドと連携して動作する場合がある。利用者が自分で読み仮名などを教える「ユーザー辞書」機能と、利用者の選択をシステムが自動で覚える機能の両方がある。後者が学習に相当する。

表 2：かな漢字変換の学習と性能変化のタイミング（参考）

		出荷・サービス開始後の性能変化	
		しない	する
出荷・サービス開始後の学習	しない	学習しないかな漢字変換は現在では廃れた	（該当なし）
	する	ソフトメーカが変換結果（提示した変換候補のうち、利用者が選んだ「正解」とその用例文）を吸い上げて自社クラウドで解析。 バージョンアップの際にその学習の成果をまとめて反映して性能を向上させる。ソフトメーカは学習に使用する「正解」と用例文を選別できる。	利用者の選んだ「正解」の提示順序が上がる、単漢字変換の結果を新たな変換候補として記憶することで、学習に伴って性能が徐々に向上する。 時々意図しない変換結果を覚えて、性能が悪くなることがある。その場合も、利用者はリセットして問題回避が可能。

また、出荷・サービス開始後に学習を続ける場合は、使用現場のデータを学習することとなる。これにはメリット、デメリットが存在する。

1) メリット
   - 使用現場の実態に即した大量のデータが利用可能になる。
   - 地域別、患者群別にカスタマイズされた AI も可能になる。
2) デメリット
   - データセットの信頼性（品質、サイバーセキュリティを含む）を維持する役割を担う者についてコンセンサスがない（→4.3 節）。
   - 個人情報保護など関係する他の法制度へのコンプライアンスの問題とコストを要する。

データセットの特性、信頼性については、3.3 節で詳説している。

② 学習を誰が行なわせるか

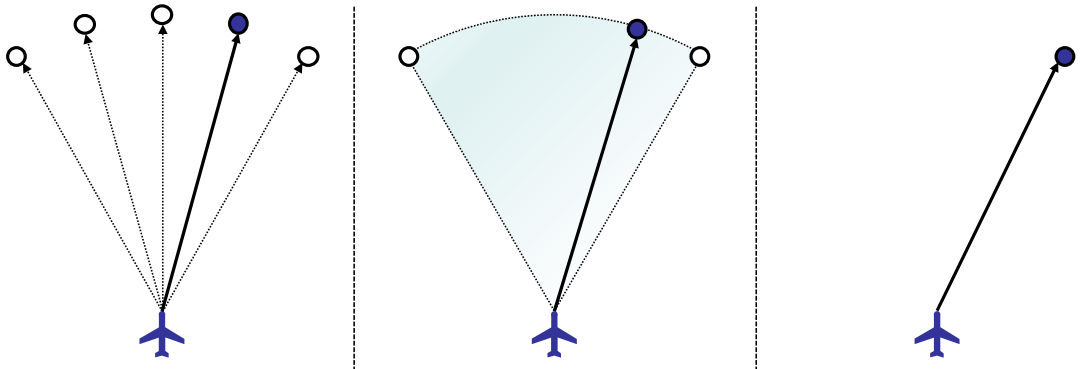
AI 医療システムの出荷・サービス開始前の学習は開発者が行わせるので開発者によるデータセットのコントロールと risk management が可能である。一方、出荷・サービス開始後の学習は、開発者のみが行わせるとは限らない。医師等、患者など、AI 医療システムの操作者が学習を行わせることも想定される。操作者が学習させる場合、データセットのコントロールと開発者としての risk management が、開発者が学習を行わせる場合と同様に実施することが難しい状況が想定される。つまり、誰が risk management を分担するかが問題となる。この点については、4.3 節で言及している。

３．１．２ AI のブラックボックスとしての性質と出力の予見可能性

AI の患者への作用性に関係するもう一つのファクターが、AI のブラックボックスとしての性質に由来する出力の予見可能性である。これは開発者が行う risk

management に影響する。AI 医療システムの出力については、次の 3 つのパターンに分類できる。

1) 出力が所与の有限個の解の中から選ばれる場合。例：患者の症状を元に、予め設定された診断名の中から可能性のあるものを回答する。
2) 出力が所与の有限範囲内の解の中から選ばれる場合。例：画像診断支援において病変の可能性のある領域の位置とその確率を 0-1 の範囲内で出力する。
3) 出力の範囲が事前に規定されない場合。例：疾患の新しい区分を創出する。



AIの出力が所与の有限個の解（オプション）の中から選ばれる場合

※囲碁、将棋も解（指し手）は有限個。ただし膨大な数

AIの出力が所与の有限範囲内の解の中から選ばれる場合

AIの出力の範囲が事前に規定されない場合

開発者によって事前に設定された範囲内に出力が限定される場合、AI を用いない従来の医療機器と同じであることから従来と同様に risk management が可能である。ただし、可能な解の組み合わせ数が膨大な場合は別の困難が出現する可能性がある（例：囲碁将棋の指し手の予想）。出力が学習データに依存して変化する場合も、出力が有限かつ既知の範囲内であれば同様に risk management が可能である。これに対して、出力の範囲が事前に制限されない場合、AI の出力とその影響が未知ということになり、従来的な risk management の考え方で対応できるか疑問が残る。学習とこれによる出力への影響が開発者によってコントロールされない場合も同様である。

３．１．３　将来の高度な自律能

AI 医療システムのもう一つの特徴である自律能については様々な議論がある。自動車の自動運転ではすでに議論が始まっているが、将来、AI の高度化に伴い自律能が高度化すると医師等と患者の関係性に変化が生じることが予想される。

究極的には、AI が単独で診断確定・治療方針の決定をする能力を獲得する状況も想定されるが、医師等と患者の関係が従来の社会・法制度の想定しないものとなることが予想される。この点について厚生労働省の「保健医療分野における AI 活用推進懇談会 報告書」では次のように述べている。

「…現状では、AI が単独で診断確定・治療方針の決定を行なっている

わけではなく、また、AI の推測結果には誤りが有りうる。このような状況を踏まえ、診断確定や治療方針の最終的な意思決定は医師が行い、その意思決定の責任も当該医師が負うべきである。」

AI の高度化とこれに伴う医師等と患者の関係性の変化は医療における AI を論じる上で重要な側面ではあるが、本報告書ではこれについては詳細には論じない。4 章にて倫理・責任の観点から言及している。

３．２　AI の臨床的位置づけと利用形態

本節では、次の 2 つの例

1) 診断支援の AI 医療システム　～患者に間接作用する AI、CAD など（3.2.1 項）
2) 治療支援の AI　医療システム～患者に直接作用する AI、手術ロボットの AI による自律制御など（3.2.2 項）

を中心的に検討する。以下では、アルゴリズムの特徴よりも臨床上の使われ方に基づいたレベル設定を示す。なお、診断と治療は密接に関係しているので、診断支援と治療支援の両者を含めたレベル設定も考えられるが、レベル定義の複雑化を避けて別々に検討した。

３．２．１　診断支援における AI の臨床的位置づけと利用形態

ここでは、コンピュータによる（画像）診断支援（CAD）用のソフトウエアにフォーカスを当てる。CAD の入力の例は、主には診断の際に撮影された画像であり、臨床情報（例：カルテ）や、医師等による診察そのもの（例：患者との会話）、その他の情報（例：患者自身の日常生活上の記録）、別のソフトウエアによる出力（例：深層学習により得られた特徴）、なども考えられる。

以下は、前節の自動運転などのレベル分けを参考にして、画像診断における CAD の役割を 5 つのレベルに分類したものである（現在まだ実現されていない技術を含む）。

診断支援レベル 1　疾病に関係する何らかの特徴量（例：存在の可能性の高い位置の指摘、腫瘍の最大径や体積、悪性らしさ（注：アルゴリズムによって定義される悪性らしさや確率などであり，臨床的な悪性度と一致しない場合もある））を測定して医師等に提示し、診断を支援

診断支援レベル 2　診断結果（例：正常、異常、（臨床的）悪性度、進行度、治療方針など）を医師等に提示し、診断を支援

診断支援レベル 3　様々な臨床情報（マルチモーダル情報）を総合して導いた総合的な診断結果を医師等に提示し、診断を支援

診断支援レベル 4　マルチモーダル情報に基づく自動診断、ただし、診断結果については必ず医師等が承認

診断支援レベル 5　マルチモーダル情報に基づく完全自動診断（医師等を介しない診断）

※補足 1：従来の CADe は上記レベル 1 の一部に相当し、CADx はレベル 1 のうち腫瘍の性状に関する情報を提示する部分から上のレベルに相当する。

※補足２：レベル 2 までは、主に画像などの単一のモダリティを入力とした支援システムを想定。レベル 3 以降は画像以外も含めたマルチモーダル CAD。

ここで着目すべきは、レベル 2 までは従来の CAD、すなわち主として放射線科医による放射線医学的診断の支援であるのに対し、レベル 3 以上は医師等が行う役割、放射線医学的診断を含む総合的な診断もしくはその支援である点である。AI の患者への間接／直接作用の位置づけが変化している。AI の進化により、診断支援の臨床的位置づけは関係する医学領域を横断するものになっていき、その評価にあたっても様々な視点からの検討が必要になる。

診断支援における AI の出力の予見可能性については、CAD の多くが分類器であり、開発者が設定した所与の有限個の解の中から選択することとなり、この点において AI を用いない CAD と risk management 上の相違がない。

３．２．２　治療支援における AI の臨床的位置づけと利用形態

治療支援を目的とした AI 医療システムでは、AI の出力結果が患者へ直接作用しうる、すなわち AI が自律的に治療に関与しうるという点が特徴であり、前節の CAD のレベル分けはそのまま適用できない。

治療支援を行う AI 医療システムにおいては、3.1 節に述べた AI の患者への間接／直接作用の位置づけ、またその出力の予見可能性（開発者によって事前に上限下限が設定されるか、事前に有限の離散値・連続値が設定されるなど）がレベル分けに関係する。これらを総合した IEC TR 60601-4-1 の自律度などが指標となりうる。

３．３　データセットの特性と信頼性

学習やテスト（評価）[8]などに用いるすべてのデータは、個人情報保護の観点から適切に加工されたものを利用しなければならない。また、データに著作権が存在する場合にも、これを適切に扱う必要がある。以下では、必要なすべての対策を施した後のデータの特性と信頼性について述べる。

３．３．１　画像診断支援システム（CAD）におけるデータセットの特性と信頼性

データセットの特性と信頼性の議論は、以下の CAD 開発のフェーズと密接に関係する。

1) アルゴリズム開発と原理検証
2) 製品版の学習と（未知データによる）テスト（評価）
3) 治験
4) 上市後の学習とテスト

例えば、アルゴリズムの基本設計と原理検証のフェーズではデータセットに

[8] IEC 82304-1:2016 で用いる「バリデーション」という用語は、深層学習の分野ではネットワーク構造などのハイパーパラメータを決定するためのプロセスを指すため、本報告では前者及び JIS T2304 のユニット試験、システム試験等を総称して「テスト」と表記する。

対してそれほど高い信頼性は要求されない。しかし、製品版の学習やテスト、さらに治験相当レベルになるとより高い信頼性が要求される。また、学習よりもテスト用のデータに対する信頼性が重要となる。以下で述べるデータセットの信頼性に対する要求や開示すべき情報は、フェーズが進むにつれ高くなり、かつ、テストの方が学習よりも高くなる点に注意されたい。

まず、信頼性の問題について指摘しておく。AI を利用した CAD が、経営戦略的には日本国内のあらゆる病院を対象に開発・販売されたとしても、AI を用いないこれまでの医薬品・医療機器の臨床試験の場合と同様、実際には数施設のデータのみを用いて評価されることが多いと想定される。その場合、開発者はデータのサンプリング法とその適切さ（例えば偏りがないこと）を示す必要がある。ただし、以下に示す例のように、偏ることが必ずしも悪いことではなく、むしろ好ましいケースもあるため、柔軟な対応が必要となる。

例）データセットの特性が偏ることが好ましい場合

納入後に臨床データを再学習して出力が変化する CAD は、その環境下で撮影された画像に特化することによって、より地域の患者背景を反映した診断支援を提供できる可能性がある。再学習は環境に適応する有力な方法であり、納入後に再学習することで、高い環境適化を実現することが可能となる。その場合のデータセットの特性は、全国の平均的な病院から見ると時間の経過とともに偏りが大きくなる。

次に、深層学習をはじめとする最近の機械学習に特徴的な点について指摘しておく。データセットの信頼性の記述や評価の基準の作成の際には、以下の視点も重要である。

1) 転移学習と呼ばれる学習方法の場合、対象とは異なる医用画像や自然画像を用いて学習を行うことがある。
2) ネット上の素性（正解ラベル等）の分からない大量の画像（医用画像には限らない）を利用することも技術的には想定される。教師なし学習、半教師あり学習、Weak Label を用いる学習アルゴリズムのように、正解ラベルを知らなくても、あるいは正解ラベルが不正確でも学習が可能なタイプの学習アルゴリズムがあるため、技術的にはこのようなデータの利用が可能である。ただし、そのようなデータには信頼性の問題が伴う。その影響や結果が受容可能であることの検証が必要である。
3) 学習データを補うために、初期の学習データに対して線形変換（回転、拡大縮小、平行移動など）や非線形変形を適用して新しい学習画像を生成したり、生成モデル（例：自己符号化器など）を用いて人工的にデータを生成したりしたデータを学習に用いることがある（2.3 で述べた Data Augmentation の一例）。

これらの 1〜3 は、対象とする医用画像のみにより学習や評価を行っていた従来とは異なる新しい要素である。以下では、2 と 3 についてさらに補足する。

2') データセットには、疾病の位置や広がりなどを示したアノテーション画像や、良悪性などのラベルが付くことがある。一般にこれらの特性と信頼性の記述と評価は重要である。例えば、複数の専門家が合議で作成したアノテー

ション画像や、病理情報に基づくラベルは信頼性が高く、そうでないものは低く、学習結果へ悪影響があると考えられてきた。しかし最近は、教師なし学習、半教師あり学習、Weak Label を用いる学習アルゴリズムのように、正解ラベルを知らなくても、あるいは正解ラベルが不正確でも学習が可能なタイプのものがある。このようなアルゴリズムが登場した背景には、膨大な画像データの蓄積がある。例えば、カプセル内視鏡や大腸内視鏡のように大量の画像を用いて学習する場合、すべてに対して病理学的に正確なラベルをつけることはほぼ不可能に近く、仮に正確なラベルを付与できてもそれに見合うだけの性能の改善が期待できないケースもある。ラベルの信頼性に関する正確な記述は重要であるが、常にラベルが正確であることが必要か否かは問題依存である。したがって、アノテーション画像やラベルについての特性と信頼性の記述と評価を、問題によらず一様に厳しく求めることは、優れた CAD の開発の妨げになる可能性もあることに注意しなければならない。データセットの信頼性の記述や評価の基準の構築にあたっては、これらのことを踏まえた検討が必要である。

3') Data Augmentation のためのデータ生成法の記述が重要になることがある。例えば、適用する変形が想定されるばらつきを超えて不自然に大きい場合には、CAD の性能への悪影響が心配される。また、人工的な生成法の中には、ある現象のシミュレーションも含まれるが、そのシミュレーションが妥当であれば想定通りの性能が期待できる（例：3 次元 CT 像から人工的に多数の X 線投影像を生成する場合）。

機械学習の技術進化のスピードは速く、今後も現時点で想定していない学習法が登場する可能性は高い。そのため、学習データや学習アルゴリズムに注目するよりも、テストにおける評価を重視することの方が適切かもしれない。

テスト用のデータは、学習データ（深層学習などにおけるバリデーション用データを含む）とは異なっていなければならない。ただし、学習と評価のデータを完全に区別することは難しく、評価用データの一部や全体が何らかの理由によって学習に使われてしまうことがある．その場合には評価の信頼性が失われるため，AI 医療機器の場合には、新しい評価用データを用いて評価しなおす必要がある．また，ネットワークの状態を出荷前や再学習の前の状態に戻す機能及び管理体制も有効である。これを含めて性能検証用のテストデータにアルゴリズムとその開発担当者が故意または偶然にアクセスできないような管理体制が重要である。

なお、テスト用に、大規模な公開データベースがあれば、それを市販前評価に利用することは有益である。例えば，画像ではないが，米国 MIT で開発された心電図データベースは有名であり（MIT-BIH Arrhythmia Database）[(2)]、このデータベースを使うことで心電計などの開発・普及が加速化された．

ここまでは画像を対象としており、その他の臨床情報を含むマルチモーダル情報については述べていない。しかし、データのサンプリング法や偏りの議論、また、転移学習や Data Augmentation などに関する議論は、その他の臨床情報の場合にも原則として同様に当てはまる。また、人種差の問題の本質はデータセ

ットの特性の偏りの問題である。ただし、細部においては個別の議論が必要となる。

３．３．２　治療支援におけるデータセットの特性と信頼性

診断支援におけるデータセットの特性と信頼性に関する前節の議論と合わせて、治療支援における AI のデータセットの特性と信頼性を考える上で重要な点について整理する。なお、これらは AI を用いないこれまでの治療支援の医薬品・医療機器の臨床試験の評価の場合と同様である。

1) 診断支援と比較して、機械学習における「正解」を定義することが困難なケース、正解が自明でないケース、正解が迅速さを要求される技術開発に適さないケースがある。名医の手技など暗黙知に強く依存するため正解の標準が明文化できない場合、標準化されていても 5 年生存率など長期観察を経ないと得られない場合などである。開発者が正解ラベルを定義する、代替的な（サロゲートな）正解の判定方法を設定することができるが、正解ラベルの定義や正解の判定方法について、根拠を示す必要がある。
2) 再現性 : 疾患・治療・医師等につき同じデータセットを多数用意することができる疾患は限定的である。異なる疾患・治療・医師等のデータを同一のデータセットとして学習する場合は、意図しないバイアスを導入する可能性がある。特に施設によって治療方針や術式が異なる場合に注意が必要である。
3) 時間軸 : 治療支援で扱うデータは、時間軸をもったデータが多い。時間軸は有用なデータであり、データにタスクのフェーズを関連づけることで、高精度な学習が可能になる。
4) 操作ログの解析にあたっては、経歴、経験といった医師等の属性を考慮する必要がある。医師等の医学的な経歴、経験だけでなく、当該機器の習熟度についても考慮する。

また、治療支援における AI の構築では、狭義の医療情報に加えて機器の操作ログ、患者や医療スタッフの判断や行動に関する情報も学習の対象となり貴重な情報源となりうる。しかしながら、これらは医療情報扱いされてこなかったため、その質の確保、収集方法、所有権、（医療スタッフ側の）個人情報保護との関係が未整理である。

３．４　リスクの分析と対策

AI 医療システムにおいても使用目的に対して品質、有効性、安全性を確保することの重要性は、これまでと変わらない。一方で、AI 医療システムを導入することにより、AI を用いない従来の医療になかったリスクが発生する場合がある。

なお、IEC TR 60601-4-1:2017 では医用電気機器の自律能の高低はそれがもたらすリスクの大小に直結しないと述べている。同様に、AI 医療システムの導入がもたらす risk の大小は、その AI のレベルに相関するとは限らないことに留意する。つまり、AI の導入によって新たな hazardous situation が導入されて harm を生じる可能性がある一方、他のより重大な harm の発生頻度を下げる効果又はその重大さを下げる効果が期待され、全体として risk が受容される可能性があ

る。逆に、AI の導入によって全体としての risk が増大してしまう可能性もある。

３．４．１　CAD のリスクの分析

※アウトプットのレベルについては 3.1 節参照

CAD の導入により発生する代表的なリスクは見落としや拾いすぎなどの誤診である。以下では 3.1.1 項で示したレベル別にリスクの大きさなどをまとめた。ただし、リスクの具体的な内容やその大きさは、CAD のレベルやソフトウエアの精度に依存する。例えば精度が高ければリスクの影響は限定的になり、逆に精度が低い場合にはリスクが高くなるが、求められる精度はレベルによって異なる点に注意が必要である。

診断支援レベル１　計測値は単純であり、発生するリスクは限定的

診断支援レベル ２　医師等の診断の一部を補完するため、相応の精度が求められる。また、発生するリスクが大きくなるケースも想定され、診断だけでなく治療などへ影響も無視できなくなる。

診断支援レベル ３　医師等と精度が同レベルか、場合によっては医師等を超える精度が求められる。そのため、発生するリスクはかなり大きく、誤診による病状の悪化や死亡なども想定される。

診断支援レベル ４　医師等を超える精度が求められ、医師等によるチェックがうまく機能しないケースでは、誤診断によるリスクは非常に大きい。

診断支援レベル ５　医師等を超える精度が求められ、発生するリスクはもっとも大きい。

３．４．２　治療支援のリスクの分析

治療支援を行う医療機器等においては、3.1 節に述べた AI の患者への間接／直接作用の位置づけ、また AI の出力の予見可能性がリスクの分析において重要である。

治療支援を行う医療機器等の動作にあたっては、操作者が危険を認知して介入するまでの時間（タイムラグ）がリスクに関係する。操作者が危険を認知して介入するまでの時間的余裕が十分ない場合に、AI が回避動作を行うことでリスク緩和を図ることも期待される（例：自動車における自動ブレーキ）。ただし安全監視を自動装置に任せることは、操作者の状況認識の喪失(loss of Situation Awareness)に関する新たな問題になりうることが航空工学など他の分野では知られている。

３．４．３　リスクの対策

深層化したニューラルネットワークのように、AI 技術は操作者にとっては入出力関係が高度に非線形かつブラックボックスであるため、その挙動を予想しにくい。AI 技術を CAD の開発に用いた場合のリスクとその対策についてまとめる。

1) CAD の中に占める深層化したネットワークの割合が増えるほど、問題が発生しうる範囲や程度の予想、発生した問題の制御は難しくなる。問題が発生した場合にその原因の切り分けができないと、安全対策の選択肢が少なく

なる。
一般論ではあるが、学習データが十分存在する範囲では正しく動作することが期待されるが、学習データが不十分な範囲では誤った動作をする可能性が高くなる。したがって、ブラックボックスの場合には、十分なデータを用いた学習や検証が望まれると同時に、リスクが発生した場合の対処も重要となる。ここで、必要なデータ数は処理対象やネットワークの性質など、様々な要素に依存するため、ひとまとめにして議論をすることはできない。類似のシステムや研究例があれば、それが一つの根拠になる可能性がある。（※補足 : 学習データの薄い範囲での挙動が予測困難であり、リリース前に問題を潰す確証が得られない。AI の出力と一緒にその確実さを示す指数を提示するなどの情報提示や、予想しない挙動が深刻な結果に繋がらないための防護手段を医療側で備えているなどの総合的な対策も求められる。）

2) リスクの対策の例について示す。リスクの例は、明らかな異常であるにもかかわらず、正常であると診断する場合や、その逆の場合である。原因の切り分けや問題の解決は、ブラックボックスであるために一般には容易ではない。そのため、まず、誤った診断・治療・予防支援をする確率を操作者に提示し、操作者がそのリスクを事前に正しく理解することが重要となる。そのうえで、リスクが発生した場合の具体的な対策についてあらかじめ決めておくことが重要である。
3) ソフトウエアのハッキングにより､学習データやバリデーション用のデータ、ソフトウエアの一部あるいは全体が書き換えられ、誤った診断・治療・予防支援結果が出力される可能性がある。ブラックボックスの場合にはハッキングによる書き換えに気づきにくいので特に注意が必要である。高いレベルの支援に対しては、より高度なサイバーセキュリティー対策が求められる｡PACS などのデータベースとネットワークにおける対策も参考になるかもしれない。

３．５　市販前評価と市販後評価・管理

ここでは CAD を中心に議論する。

評価方法

支援対象と CAD のレベルに応じて、適切な評価指標を選択して評価しなければならない。例えば、病変の検出であれば評価指標は見落としと拾いすぎであり、これらを CAD が標榜する適用対象の母集団の特徴を反映するテストデータ（≠学習データ）を用いて、評価を行う必要がある。ここで、CAD 出力に対する閾値などを変化させながら用いることを可能とする場合には、Receiver Operating Characteristics (ROC) Curve による評価が考えられる。また、対象が特定の病院でない限りは、多施設での評価が妥当であると考えられる。その他、読影時間や画像診断後の精密検査の割合なども評価指標の候補である。

アルゴリズムの記述

非線形かつブラックボックスである深層化したネットワークの動作機序の記述は、（一部に試みはあるが）一般的にはほぼ不可能である。代わりに、ネット

ワークの構造、学習データ、学習アルゴリズムなどについて記載することが考えられる。具体的には、ネットワークの構造を示した図や数値（層数、ノード数、結線数など）による説明、学習データの質や量に関する記述、学習アルゴリズムの擬似コードなどである。

市販前評価（テスト）

少数の質的に確保されたデータによる評価＝従来の治験。コストはかかるが、受け入れられてきた方法であり、高度に非線形かつブラックボックス化されたCAD の評価の重要な選択肢の一つである。データの信頼性が重要となるが、それについては 3.3.1 項を参照のこと。

市販後評価

再学習により AI を用いた CAD の性能が変化、例えば上記の ROC が変化した場合の評価には、次の観点からの議論が重要となる。

・再学習[9]の主体とタイミング
例えば、再学習の主体者は CAD の開発者、タイミングは再学習のコストと性能向上などのメリットとのバランスを踏まえて決める方法がある。今後は、技術進歩によって、医師等、患者などの操作者、あるいは、コンピュータ自身が主体となって別の基準で選んだタイミングで再学習を実施する可能性がある。

・再学習のデータ信頼性や CAD 性能の確保
再学習用のデータや CAD に対しても、信頼性や性能の確保が必要となる。具体的には、例えば、承認当初のテストデータ（治験データセット）で再評価し、当初の性能が統計的に確保されていることを、95%信頼区間などを用いて確認する方法が考えられる。ただし、3.3.1 項で示した例のように、CAD を納入した病院のデータと承認当初の治験データの特性が異なる場合、治験データセットで評価することが必ずしも好ましくないこともあるので注意が必要である。

・リスクの対策
再学習用のデータの性質によっては、再学習後のシステムのリスクが増大（あるいは軽減）する可能性がある。特に、深層化したネットワークの場合のシステムの予測困難性に基づくリスク（3.4.3 項の 1)）は再学習後にも想定され、その対策が必要となる。また、性能の変化を一定期間適切に記録・保存することが、新たなリスクの発生を抑えるポイントになると思われる[10]。

---

[9] 再学習とは、学習データや学習アルゴリズムを変更する（両方の変更や片方のみの場合も含む）ことにより、CAD 内の一部あるいは全体のパラメータなどを変更するプロセスのことである。多くの機械学習アルゴリズムはこれを自動で行うことが可能である。広い意味では、開発者が新しい学習データを使って実験的にアルゴリズムのパラメータなどを変更する場合も含む。

[10] 類似症例検索システムのように、データセットの画像を選択して提示する事例ベースの検索システムでは、データセットの性質の変化が直接性能の変化に影響する。このようなシステムについては、性能（例：Precision、Recall、F 値など）の評価はもちろん、それ以外にデータベースの収集法や特性について、承認時だけでなく、承認後の継続的変化についても評価・記録が必要になるかもしれない。

補足 1
個人ごとのデータに基づいて毎回学習し、個人に最適化したネットワークを構築することもある。例えば、個人の CT 像からシミュレーションした X 線投影像を用いて放射線治療計画のためのネットワークを学習させる場合である。この場合には、学習を毎回行っていることになるが、投影のシミュレーションが妥当であること、ネットワークの規模に対して十分な数の投影像を用いて学習していることなどを、事前に実験や過去の研究例などで示すことができれば、再学習をしていることをもってすぐに新たな評価が必要とはいえないと思われる。判断基準の構築にはこのような例も踏まえることが重要となる。

補足 2
CAD の中には、ROC 上に設定された複数の動作点の中から操作者が動作点を選択できるものもある。臨床での使用中に動作点を変更すると、見かけ上 CAD の性能が変化しているように見える。しかし、再学習によるネットワークの性能の変化とは本質的に異なる。最大の違いは、前者の変化は予測可能であるが、再学習の場合は予測が困難（深層学習の場合には著しく困難，あるいは不可能）である点にある。従って，両者は区別して議論をする必要がある。

参考文献
(1) 平成 28 年 3 月 31 日付厚労省医療機器・再生医療等製品担当参事官室事務連絡，医療機器プログラムの承認申請に関するガイダンスの公表について
(2) Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng in Med and Biol 20(3):45-50 (May-June 2001). (PMID: 11446209)

４章　AI 医療システムの倫理・責任

本章では AI に関する広義のレギュラトリーサイエンス、AI 医療システムの倫理と責任について長期の将来の技術レベルを想定した議論を試みる。

４．１　医療における AI に関する倫理

「AI 開発ガイドライン（国際的な議論のための AI 開発ガイドライン案）」[(1)]は AI の開発者、サービス提供者の尊重すべき原則をあげており、その中に倫理も含まれている。

AI 開発原則
（主にＡＩネットワーク化の健全な進展及び AI システムの便益の増進に関する原則）
① 連携の原則-----開発者は、AI システムの相互接続性と相互運用性に留意する。
（主にＡＩシステムのリスクの抑制に関する原則）
② 透明性の原則-----開発者は、AI システムの入出力の検証可能性及び判断結果の説明可能性に留意する。
③ 制御可能性の原則-----開発者は、AI システムの制御可能性に留意する。
④ 安全の原則-----開発者は、AI システムがアクチュエータ等を通じて利用者及び第三者の生命・身体・財産に危害を及ぼすことがないよう配慮する。
⑤ セキュリティの原則-----開発者は、AI システムのセキュリティに留意する。
⑥ プライバシーの原則-----開発者は、AI システムにより利用者及び第三者のプライバシーが侵害されないよう配慮する。
⑦ 倫理の原則-----開発者は、AI システムの開発において、人間の尊厳と個人の自律を尊重する。
（主に利用者等の受容性の向上に関する原則）
⑧ 利用者支援の原則-----開発者は、AI システムが利用者を支援し、利用者に選択の機会を適切に提供することが可能となるよう配慮する。
⑨ アカウンタビリティの原則------開発者は、利用者を含むステークホルダに対しアカウンタビリティを果たすよう努める。

なお、ここであげる「倫理の原則」の解説として

開発者は、人間の尊厳と個人の自律を尊重するに当たり、人間の脳や身体と連携する AI システムを開発する場合は、生命倫理に関する議論などを参照しつつ、特に慎重に配慮することが望ましい。

開発者は、採用する技術の特性に照らし可能な範囲で、AI システムの学習データに含まれる偏見などに起因して不当な差別が生じないよう所要の措置を講ずるよう努めることが望ましい。

開発者は、国際人権法や国際人道法を踏まえ、ＡＩＩシステムが人間性の価値を不当に毀損することがないよう留意することが望ましい。

と述べられている。

一方、「人工知能と人間社会に関する懇談会」（内閣府・平成 29 年 3 月公表）[(2)]では、倫理に関する観点として

－　人工知能技術の進展に伴って生じる、人と人工知能技術・機械の関係性の

変化と倫理観の変化
- 人工知能技術によって知らぬ間に感情や信条、行動が操作されたり、順位づけ・選別されたりする可能性への懸念
- 力や感情を含む人間観の捉え直し
- 人工知能技術が関与する行為・創造に対する価値・評価の受容性。価値観や捉え方の多様性

をあげている。

医療における AI では、これらに対応して

- 3.1.2.3 項で言及した専門職としての医師等と機械（AI）との関係性の変化。究極的には、AI が医師等よりも高度な診断支援・治療支援の能力を獲得する状況も想定される（図 4-1）。その場合、例えば「優秀な医師等以上の正答率であることが統計的に示されている診断支援 AI 医療システム（しかし一定の誤りがある）」があるとして、医師等がその AI と異なる判断を選ぶことは、訴訟リスク等を考えると困難である。その場合、医師等と機械の役割は実質的に変質することになるといえる。
- その事態は、医師等の職業観、使命感、充足感にも影響する可能性がある。
- 別の問題として、お手本となる医師（専門医や指導医）の位置づけが変わる可能性がある。現在の AI 開発では医師の判断を「教師データ」として扱うことができるが、AI 医療システムが普及してその利用が当然となる将来もこれを続けることは可能だろうか。

といった倫理課題を生む可能性がある。

海外においても、IEEE の Ethically Aligned Design 等の取り組みが進められている(3)。

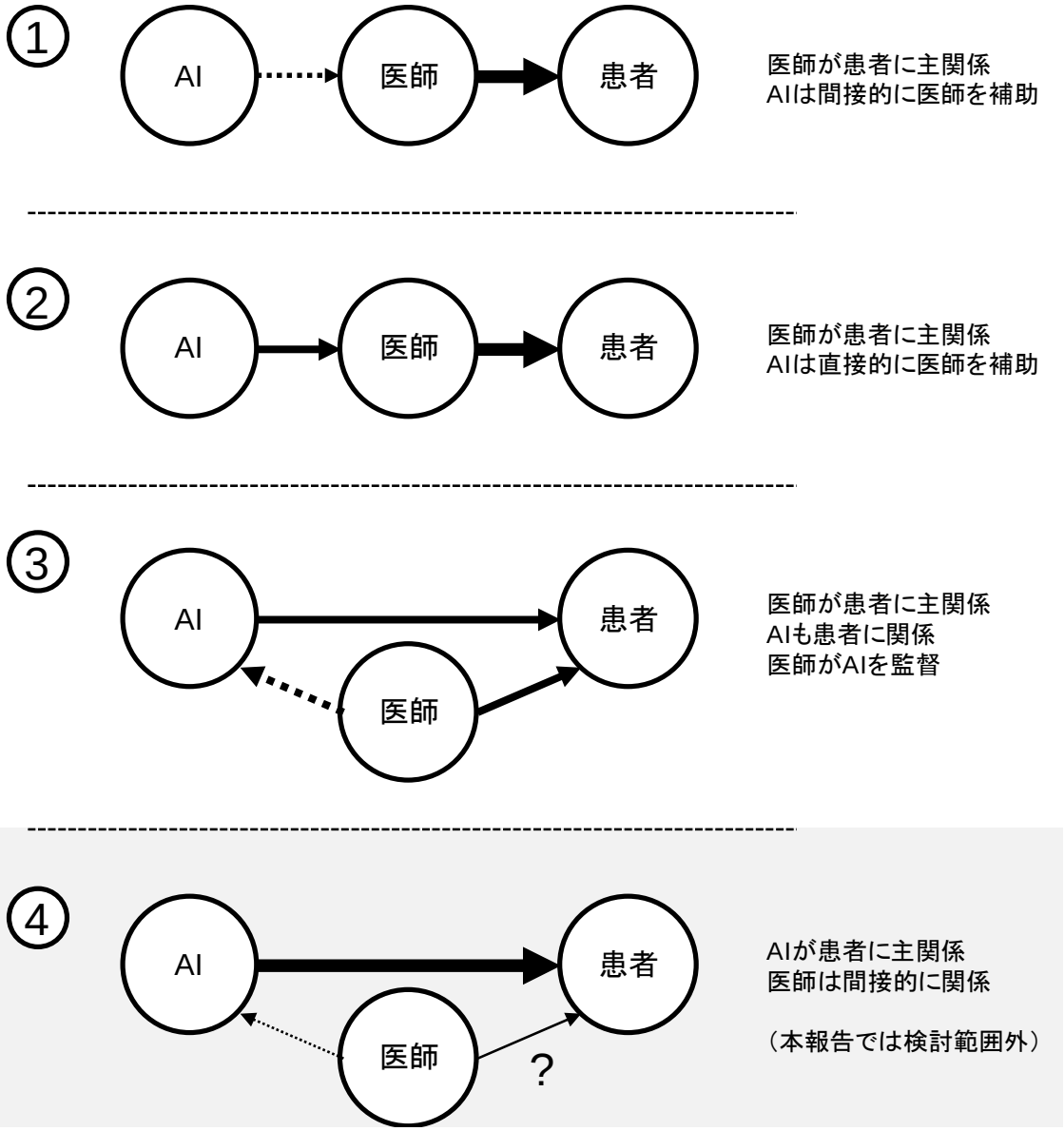


図 4-1: AI、医師等、患者の関係。①から④へと進化するにつれて、AI が主体的となり、特に③から④にかけて医師等と患者の関係性が曖昧になってくる。④は医師等と AI の関係が逆転しており、患者が事実上直接 AI を用いるのに等しく、本報告では検討範囲外。

４．２　医療における AI に関する責任

事故が発生した場合の責任論は、民事責任、刑事責任に大別できる。それは他の種類の事故と同じである。

医療における AI の民事責任を論じるにあたり、3.1 節で述べた間接／直接作用型 AI の考え方が参考となる。AI に限らず、また医療機器に限らず、一般不法行為責任：「故意又は過失によって他人の権利又は法律上保護される利益を侵害した者は、これによって生じた損害を賠償する責任を負う」（民法 709 条）と製造物責任（製造物責任法）が考えられる。前者は過失、後者は製造物の欠陥（通常有すべき安全性を欠くこと）に対する責任。なお無体物たるソフトウエアには

製造物責任法を適用しない。また故意は別の次元の問題なのでここでは扱わない。

直接作用型 AI では、操作者の責任は限定的である一方、機械に責任を負わせることはできない。機器の製造者の過失責任が問われる可能性が高い。その場合、欠陥も認定される可能性が高い。

間接作用型 AI では、操作者の過失と製造物の欠陥の争いとなりうる。その際に、操作者が通常有する知識や技術、操作者への情報提供（使用法の指示や危険性の警告）といった事項が焦点の一つとなる。

医療の場合、操作者が医師等の専門性を有する職業者である場合が多く、その場合は医療における医師等と患者の情報知識の非対称性がある点に留意する必要がある。危険状態が発生した際にそれが予見可能なものであったか、起こった事象に対処することで危害を回避可能であったかなどが問題となる。

AI に関する責任を考える上で法学、技術、社会の合意形成が待たれる事柄として「AI に学習させた者の責任」に関する論点がある。医療に限らず、AI システムは AI の開発者が当初想定しない成長の仕方をする可能性がある。例えば自然言語を学んで対話する AI に対して、悪意ある者が意図的に差別語などを学ばせた結果、その AI サービスが閉鎖に追い込まれるといった事象が起きている。このケースは悪意のある者が行った行為であり極端であるが、想定しない偏りが学習データに含まれた場合や、過学習が起きた場合に、性能低下や想定外の挙動をする可能性がある。

「AI 開発ガイドライン（国際的な議論のための AI 開発ガイドライン案）」では

- 「開発者」とは AI システムの研究開発（AI システムを利用しながら行う研究開発を含む）を行う者（自らが開発した AI システムを用いて AI ネットワークサービスを他社に提供するプロバイダを含む）をいう
- 「利用者」とは AI システムを利用する者（最終利用者（エンドユーザー）の他、他社が開発した AI ネットワークサービスを第三者に提供するプロバイダを含む）をいう

とされており、AI に学習させた者は開発者、利用者のいずれか（あるいは両方）になりうる。同ガイドラインの議論にあたっては開発者以外の者が学習させた場合について区別しておらず、この点については異論も提出されていた。

なお、犬などの動物の場合、飼い主が躾けること、動物が起こす事象の責任については飼い主が責任をとる（動物占有者責任）という考え方が導入されているが、AI の場合に類似の議論ができるかは未確定と考えられる。

４．３　まとめ：AI 医療システムの運用における倫理・責任と課題

AI 医療システムが社会に受け入れられるには、法的要求を満たすだけでなく、そのシステムの使い方、強みと弱みについて使い手を含めた理解が進むことが望まれる。医療に限らず、AI はまだ社会のコンセンサス形成を待つべき事項が残っている。本報告の最後に、そのような未解決の事項につき、その全てをカバーしているわけでないが、実装と運用のあり方を提言する。

1)　開発者以外の者が学習させた場合の実装と運用

4.2 節に述べた通り、開発者以外の者が AI に学習させた場合の責任については医療に限らず議論が成熟していない。法的な問題もさることながら、本質的な問題は「学習過程のコントロールをどうするか」であり、それができない場合に「どのようにリスクヘッジするか、誰が risk management を分担するか」であると考えられる。

開発者以外が学習させる場合においても、データのスクリーニング機能などを実装すれば、開発者による学習過程のコントロールは可能である。AI 医療システムが自動的に学習する場合も同様である。

開発者が学習過程をコントロールしない場合、開発者はその学習をさせた者（医師等など）にリスクヘッジの一部を委任せざるを得ない。すなわち、学習させた医師等が「at your own risk」の部分を受け入れることである。

薬機法は医療機器等の品質、有効性及び安全性の確保における製造販売業者の役割を最重視する法体系である。国際規格と整合する JIS T14971 も、製造業者が現実的に実施可能な risk control を行なった上でなお残る residual risk について医学的効用がこれを上回る場合についてのみ受け入れることができるとしている。実際的には「時の医療水準」に照らして判断することとなる。妥当な医療水準の形成プロセスを経ずして医師等や患者の「at your own risk」に委ねることは薬機法とも JIS T14971 とも相容れない。

例えば、現状では、深層学習が有する高度な可塑性、例えば深層学習に極めて多数の健常症例を学習させた場合の変化などについて、多くの医師等が知っているとはいえない。また特に懸念されるのが専門分野外の医師等が AI に誤った判断を学習させ、その AI をそのまま用いてしまうことであるが、その危険性が広く医師等に知られているとはいえない。

これが、開発者がコントロールしない学習をさせることについて現在消極的となっている理由の一つである。

もちろんこのことは、将来にわたって開発者がコントロールしないデータによる学習をさせることができないということを意味するわけではない。医師等の AI への理解（AI リテラシー）が深まり、臨床研究などの形で性能評価基準の形成を進めることで、その道が開ける可能性がある。

2)　出荷・サービス提供後に性能が変わる場合の実装と運用

この場合も、AI 医療システムの risk management の実現方法が技術的な課題となる。特に JIS T2304（医療機器ソフトウエアのソフトウエアライフサイクルプロセス）の要求事項への適合と維持につき新たな方法論が必要となるだろう。同規格は AI 医療システムが医療機器に該当する場合は必須、医療機器に該当しない場合も、欧州等で要求される IEC 82304-1:2016 が適用可能であり、同規格のライフサイクルプロセスマネジメントでは IEC 62304 が引用されている。

同規格のソフトウエアライフサイクルプロセスは、ソフトウエア開発プロセスとソフトウエア保守プロセスに大別される。出荷・サービス提供後に性能変化する場合は、それぞれのプロセスが性能変化しない場合と比較して異なる。

この時、AI 医療システムの実装形態によって取りうる対策が変わる。AI がク

ラウド上など開発者が直接管理する環境にある場合は、通常は開発プロセスの中で実施する試験（ユニット試験、システム試験など）の一部を定期的に保守プロセスの中で実行するなどして、不適切な学習の結果発生した問題の検出と修正を行うことができる。AI の変化の履歴を記録することで、試験に不合格になる前の状態に戻すなどによりサービスを継続することが可能である。

AI が客先のコンピュータなどに実装されている場合は、試験を行う時間帯や所要時間、試験に不合格となった場合の対応に特に留意する必要がある。

3) カスタマイズされた学習を施した AI 医療システムの実装と運用

医療機関、地域ごとのデータを学習させた AI 医療システムは、3.2 節の補足で述べた通り、地域の医療特性によりフィットする可能性がある。そのようなカスタイマイズ機能を謳う AI 医療システムの登場も今後考えられる。

開発者がこのようにカスタマイズした AI 医療システムを出荷・サービス提供する場合は、医療機器に該当する場合の品目の同一性が保たれるかどうかに関する検討が必要となる。その他の場合、すなわち開発者以外の者がカスタマイズする場合は「1) 開発者以外の者が学習させた場合の実装と運用」と「2) 出荷・サービス提供後に性能が変わる場合の実装と運用」の両方に該当すると考えられる。

4) 医師等への情報提供と教育：医師等の AI リテラシー

医療の AI 支援の時代が間近に迫っている今日、医師等の AI 医療システムに対する適切な理解とこれを活用する能力を醸成する必要がある。例えば、「1) 開発者以外の者が学習させた場合の実装と運用」で指摘した、「医師等が at your own risk で AI を育てて用いる」状況は、すでに深層学習による CAD のオープンソースが海外で登場しているなど、喫緊の事態となっている。反面、シンギュラリティ論が想定する汎用の（スーパー）AI と CAD 用など特化型の AI を混同して議論するのは適切とは思えない。

本報告書では以下の検討を提言する。

- 医師等への専門教育として、AI の特性に関する基本事項に関する教育機会の提供：学会等の主導のもとに共通基礎教程を確立する必要性がある。
- 個別の AI 医療システムに関する情報提供；前項と併せて個別システムに関する情報提供として AI 医療システムを提供する企業が実施する方法が考えられる。
- 医学教育における AI に関する基本事項をカリキュラムに導入：すでに医学部生に対する情報科学教育は広く取り入れられていることから、その一部とすることが考えられる。

カリキュラムの設計にあたっては、技術進歩が早い分野であることを勘案する必要がある。細部の技術的な事項よりは「AI との付き合い方」としてのリテラシーを修得する方が有益であろう（例：AI による診断支援の形態、Concurrent Reader、Second Reader 等の相違など）。

なお、医師等に限らず社会のほとんどの者が AI に関する教育を受けていない現状では、医学教育に携わる者も例外ではないことから、今後は医学教育関係者と人工知能・情報工学の研究者・教育者による連携教育体制の整備も期待される。

参考文献

(1) 総務省 AI ネットワーク社会推進会議編, AI 開発ガイドライン（国際的な議論のための AI 開発ガイドライン案）, 2017/07
(2) 内閣府, 人工知能と人間社会に関する懇談会報告書, 2017/03
(3) IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

５章　結語

AI 専門部会では、近年急速に技術進歩を遂げている深層化したニューラルネットワークによる機械学習を用いた AI 医療システムの特性、取り扱い方、課題などについて有識者によるさまざまな観点からの講演を拝聴し議論を重ね検討してきた。

AI 医療システムは、1) 入力に対して出力が論理的には説明できないブラックボックスとしての性質がもたらす、未知の入力ケースに対するシステムの振る舞いの予測困難さ、2) 学習によりシステム性能等が変化しうる可塑性と、それがもたらす学習タイミングやリスクマネジメント分担への新しい考え方の必要性、3) 学習に使用するデータセットの特性や信頼性の確認とシステム性能への影響評価の必要性が、これまでの医療機器にはない特徴であるといえる。

本報告書ではまず技術の全体を俯瞰するため、第 1 章で AI 医療システムの出現と課題を概観し、第 2 章で AI 技術の現状について特に機械学習、深層学習に焦点を絞って解説した。その後に第 3 章では、レギュラトリーサイエンスの視点から AI 医療システムの特徴、臨床上の臨床的位置づけと利用形態を整理し、特に機械学習におけるデータセットの特性と信頼性に関する課題について、画像診断支援システム（CAD）と治療支援とを例にして提示した。また、こうした新しいスタイルの医療機器のリスク分析と対策を分類し、市販前評価と市販後評価・管理について、従来型の医療機器との相違を浮かび上がらせて論点を報告した。最後に第 4 章で AI 医療システムがもたらす倫理や責任所在の視点での議論を提示し、医療者への知識提供や研修の必要性についても提言を試みた。

深層化したニューラルネットワークによる機械学習を用いた AI 医療システムは、画像診断支援機器を手始めにして、時期は少し後になるかもしれないが画像以外の診断支援機器や治療支援機器などにおいても、今後急速に医療の場に導入されてくると考えられる。技術とりわけアルゴリズムの進歩と医療での活用における研究開発領域は急速に変化と拡大をしており、本報告書はその全体をカバーできているわけではないし、すべての論点を取り上げることができたわけでもないが、AI 医療システムの医療現場への導入にあたって議論すべき重要な論点の整理と提言を含む報告書として活用していただきたい。

用語集

本報告では以下の用語を用いる

AI（artificial intelligence，人工知能）
データ・情報・知識の学習等により、自らの出力等を変化させる情報処理機能又はこれを含むシステム等を総称するもの。
注 1：「AI 開発ガイドライン」の「AI ソフト」の定義を一部改変した。AI ソフトの定義ではプログラムとソフトウエアを使い分けているが、本報告では混乱を回避するためこの使い分けを行わず、ソフトウエアの語を省略した。
注 2：AI と機械学習は一般には同義でないが、近年注目の AI 技術が機械学習によるものであることからこの定義とした。

AI 医療システム
AI を構成要素として含む疾病の診断、治療又は予防に使用されることが目的とされているシステム。AI 医療システムは薬機法上の医療機器に限定されない。クラウド上でサービスのみを提供するシステムや将来登場しうる新しい機器・システムを含む。
注 1：「AI 開発ガイドライン」の「AI システム」の定義を一部改変した。
注 2：本報告の議論は主に AI 医療システムを想定して行なっている。本報告で例示する個々の AI 医療システムの医療機器該当性については本報告では議論しない。

なお、「AI 開発ガイドライン」では次のように定義している。
「ＡＩソフト」とは、データ・情報・知識の学習等により、利活用の過程を通じて自らの出力やプログラムを変化させる機能を有するソフトウエアをいう。例えば、機械学習ソフトウエアはこれに含まれる。
「ＡＩシステム」とは、ＡＩソフトを構成要素として含むシステムをいう。例えば、ＡＩソフトを実装したロボットやクラウドシステムはこれに含まれる。

本報告での AI、AI 医療機器、AI 医療システムは学習のタイミングを「利活用の過程」に限定しないことから、同ガイドラインが定義する AI ソフト、AI システムとは完全には一致しないが、同ガイドラインの適用を否定するものではない。

リスク
本報告においては、医療機器リスクマネジメント規格 JIS T14971 における専門用語としてのリスク(risk)ではなく、一般的に使われる意味での「潜在的な危険またはその状況」の意味で用いる。
JIS T14971 の用語としてのリスク等に言及するときは対応する英語用語(risk, risk management 等)を用いる。
注：本報告でリスクは、JIS T14971 の harm, hazard, hazardous situation に近い意味に分けることができるが、厳密には一致しない。

操作者
AI 医療システム、医療機器などを操作する者。医師等が操作する場合だけでなく患者や健常者が自ら操作する場合を含む。

開発者
AI、AI 医療システム、医療機器などを開発、構築して利用に供する者。

なお、「AI 開発ガイドライン」では次のように定義している。
「利用者」とは AI システムを利用する者（最終利用者（エンドユーザー）の他、他社が開発した AI ネットワークサービスを第三者に提供するプロバイダを含む）をいう。
「開発者」とは AI システムの研究開発（AI システムを利用しながら行う研究開発を含む）を行う者（自らが開発した AI システムを用いて AI ネットワークサービスを他社に提供するプロバイダを含む）をいう。
「他社が開発した AI ネットワークサービスを第三者に提供するプロバイダ」は本報告書では開発者と考える。そのため、本報告書では引用部を除いて利用者とは呼ばないこととする。