特集
Special Feature

［AI 判断の根拠を説明する XAI を使いこなす］

1 説明可能 AI（XAI）とは？ ～深層学習の説明性向上と XAI の今後の展望～

長尾智晴　横浜国立大学

機械学習と深層学習

DX（Digital Transformation）の潮流とともに企業の業務への AI 導入，特に機械学習の必要性が高まっている．**深層学習** [1] が登場する以前からさまざまな機械学習法が目的や対象に応じて使い分けられてきたが，最近はその精度の高さから深層学習が好んで用いられることが多い．深層学習で何でもできると誤解している人も多いようである．確かに，深層学習は入出力の事例としての学習用データさえ用意すれば，入力の特徴を与えなくても入出力の変換を作ることができる end-to-end の学習が可能であり，さまざまな分野に適用することができるという大きな特徴がある．しかしながら，学習に膨大な数のデータが必要，作られた処理が複雑で人には理解困難といった課題も多い．特に後者の説明性が低いことは，深層学習を企業で利用する際の障害になる．

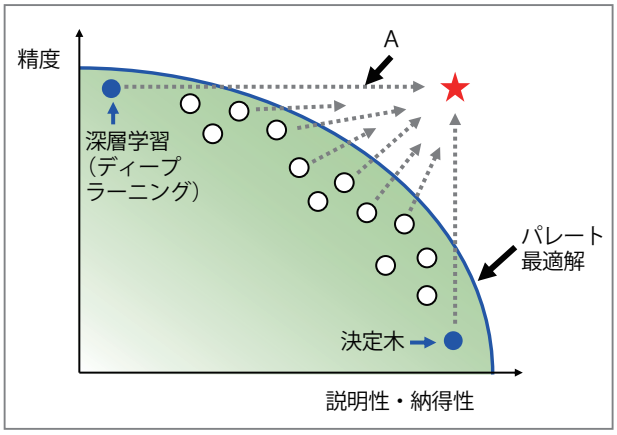
深層学習を含む現状の機械学習技術を説明性と精度の 2 軸でプロットすると図 -1 に示すようになる．深層学習は精度が高いが説明性が低く，決定木は機械の判断基準を理解しやすいが精度が深層学習に及ばない場合が多い．その他の手法も同様の傾向があり，説明性と精度の間にはトレードオフの関係がある．このため，ドメインやタスクやデータに応じて機械学習モデルを使い分けて用いる必要があるが，処理対象のデータや目的に合致する適切な機械学習モデルの選択には，ある程度の経験や試行が必要である．

機械学習を実社会で有効に利用するためには，各手法を図中の★のように説明性と精度がともに高い手法へ改良する必要がある．説明性を高めた AI は **説明可能 AI**（説明できる AI，eXplainable AI，以下 XAI）と呼ばれ，さまざまな方式が研究・開発されている．図中の点線の矢印がその方法であるが，本稿では特に深層学習の説明性向上（図中の A）について述べる．

機械学習の説明性について

XAI では，AI の説明性を高めるために次の事柄を示すことが重要であるとされている．

①入力から出力を得る“判断根拠の可視化”



■図 -1　現状の機械学習の説明性と精度の関係

②入力から出力への“機序（処理手順）の提示”
③未学習の新たなデータに対する応答・出力

ここで③は機械学習の精度保証の観点で論じられることが多いので，本稿では以後①と②について扱う．

機械学習の説明性の前に“説明性とは何か”について考えたい．人がほかの人から何かの説明を受けた際に“何をもって納得の拠り所にするのか？”には，次に示すように考え方の違いや個人差がある．

1）**理論**：物理・化学等の法則・定理による証明．
2）**論理**:演繹推論（A → B, B → C なので A → C）など．
3）**数式**：複雑ではない線形和などの単純な式．
4）**図・グラフ**：視覚に訴える表現．
5）**事例**：類似する学習済みデータの提示．
6）**言葉**：キーワードや文章による説明．　など．

1）はたとえば CAE（Computer Aided Engineering）の分野で要求されることが多い．4）はこれまでにも説明の手段として広く利用されている．また5）はたとえば未来予測などで説明の拠り所とされることが多い．6）は多くの人に受け入れてもらえそうである．

このため，XAI の設計にあたっては，その AI の利用者や説明対象者がどのような説明方法を期待しているのかを事前に調査し，それを考慮して機械学習モデルやその説明方法を設計する必要がある．すなわち XAI はニーズベースで考える必要がある．XAI の設計者が“この説明方法が分りやすい”と考えて作るシーズベースの方法では，説明対象者の納得や理解が得られない可能性があることに注意が必要である．

機械学習に対する説明は特に企業で AI を利用する際に必要である．内閣府「人間中心の AI 社会原則検討会議」[2] では，“(6) 公平性，説明責任及び透明性の原則”において，AI 利用において説明責任を適切に確保することが挙げられている．すなわち，企業で AI を利用した製品やサービスを提供する場合は，その AI を人に説明できることが必要である．

XAI は特に人の生命や財産がかかわる場面などで必要である．たとえば自動運転車両に搭載された人検知処理が深層回路である場合，万一の事故の際に「ブラックボックスの処理を搭載した」と言われて企業の責任問題に発展しかねない．同様に医学的な診断・治療の支援においても XAI が必要である．このように XAI は単に人が納得することができるという心情的な意味だけでなく，企業のコンプライアンスやリスク，訴訟などの法的な問題とも強く関係しており，きわめて重要である．企業での XAI の具体的な利用事例については，本特集の後半で紹介されるので，ここでは割愛させていただく．

深層学習の説明手法

ここでは説明性が低い深層学習（深層回路）のための XAI の例をご紹介する．これまでに多くの手法が提案されてきており，紙面の都合上ここではその一部の手法の考え方しかご紹介できないので，詳細については別途ご確認いただければ幸いである．

判断根拠の可視化例 1：中間層・特徴量の可視化

初期の XAI では，学習済みの深層回路の内部の信号を可視化する試みが多く行われた．たとえば Guided Backpropagation のように，画像のクラス分類において特定のクラスに強い信号を出力する際の深層回路の中間層の出力を画像として提示することで，エッジなどの原始的な特徴量が徐々に組み合わされ，パーツとなり，次第に各クラスの特徴が形成される様子をビジュアルに見せることができる．一方，微分画像のような画像や抽象度が高い画像が表示されることが多く，「よく分からない，直感的に理解しづらい」といった批判も多いようである．

深層回路は，前半部分で入力信号から分類や回帰のための優れた特徴量を形成し，後半の層でそれらの特徴量を組み合わせて最終的に判断していると考

えることができ，深層回路の中間層の信号を入力データの特徴量と見なすことができる．

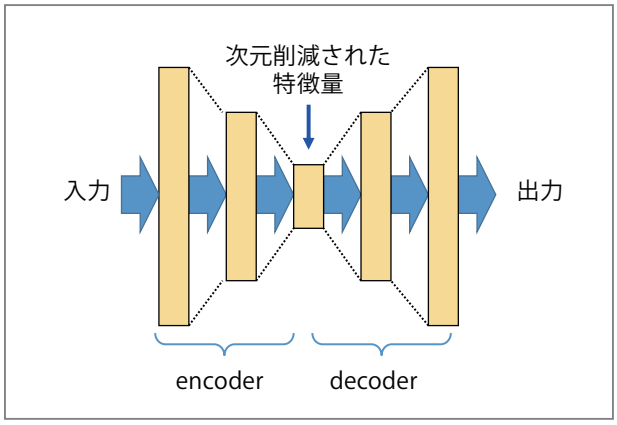
多変量解析の基礎手法として主成分分析（PCA）が知られているが，深層回路の入力信号（の特徴量）を解析する代表的な手法として，AutoEncoder（AE），CAE，VAE などが提案されている．AE は図 -2 に示すように encoder と decoder を組み合わせて恒等変換を行う深層回路であり，中間層に入力信号の有効な特徴量が次元削減されて現れることを期待する．

低次元に削減された特徴量を（必要に応じてさらに低次元化する手法を用いて）2 ～ 3 次元の図として表示することで，そのデータ中のクラスタ（＝集団・グループ）を直感的に理解しやすく可視化することができる．クラス分類の場合は，正解ラベルごとにデータに着色することで，クラスタやデータの分布状態を確認することができる．このため教師なし学習としてクラス分類や異常検知などに用いられることが多い．次元削減後に元の次元に逆変換することはできないが，t-SNE，UMAP なども優れた可視化手法としてよく用いられている．これらの手法を用いてデータのいわゆる disentangled な（＝解きほぐされた）表現を行うことが，データ解析や深層回路の理解に有効であることが示されている．なお，得られた特徴の意味が人の感覚・印象に必ずしも合うとは限らないことに注意する必要がある（よく分からない特徴軸が作られる可能性がある）．

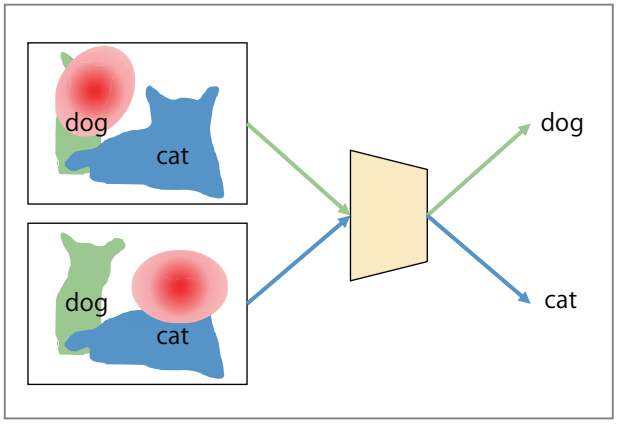
判断根拠の可視化例 2：影響度のヒートマップ表示

特定の出力信号に対する入力信号の影響度を数値や色などを用いてヒートマップ／アテンション表示することで深層回路の判断根拠を示そうとする手法がある．画像認識や自然言語処理などでよく用いられており，前者の手法としては Grad-CAM や Guided Grad-CAM などが知られている．たとえば，図 -3 に例示するように，クラス“dog”に対して入力画像中の犬の領域の影響度が大きいことを示すことで，AI の利用者にある種の安心感を与えることができる．

なお，これらの手法は，判断根拠の参考にはなるが，機序（＝その入力信号をどのように利用して結果を導いているのか）の説明にはなっていない．また，未学習の回路でも類似の影響度表示が行われる場合があることが指摘されている．このため，この種の可視化性能の評価指針として，出力から入力に向かって層を徐々に初期化したときの出力の変化を調べる Model parameter randomization test や，学習時に正しくないラベルを用いたときの出力を調べる Data randomization test などが提案されている．いずれもできるだけ早く可



■図 -2　AutoEncoder（AE）



■図 -3　ヒートマップによる判断根拠の可視化のイメージ

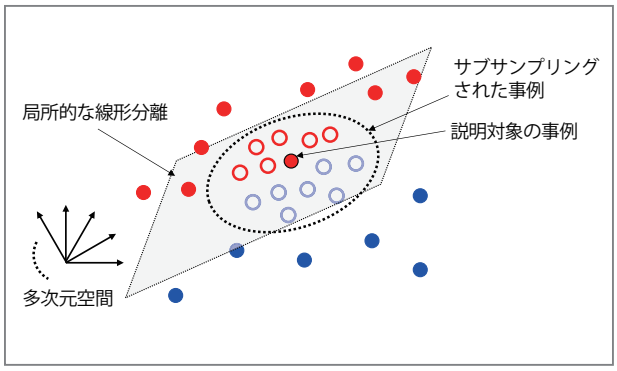
視化結果が大きく損なわれるものが，可視化法として優れていると見なされる．

産業応用において，たとえば製品の製造工程における画像による欠陥検査処理に深層学習を適用する際，単に良品／不良品のラベルを出力するだけでなく，「この部分からそう判断した」という判断根拠領域を示すことで，処理が正常に動作しているかどうかを確認することに利用することができる．

機序の提示例 1：多次元空間の局所線形化

一般に機械学習によるクラス分類に用いられる特徴空間は数十～数百次元，あるいはそれ以上と非常に高次元であり，クラス間の境界面である識別面は非常に複雑な多次元の曲面になる場合が多い．このため，指定されたある 1 つの事例の判断根拠を人に分かりやすく説明することが難しい．これに対してたとえば LIME では，図 -4 に示すように説明対象の事例の近傍をサブサンプリングした後に局所的に線形分離し，その識別面を判断根拠として説明する．これによって説明がシンプルになり，人の納得感が高まる．一方，事例ごとに異なる説明がなされるため，「判断根拠の一貫性が低い」と言われることがある．

LIME は深層回路の入力変数の重要度を解析する際にも用いられ，類似の手法として SHAP などの手法も提案されている．これらの手法は，多入力の機械学習法であれば深層学習でなくても説明可能である点が優れており，説明手法としてよく用いられている．深層回路の入力変数の重要度解析は，次に示す深層回路の構造最適化でも行うことができる．



■図 -4　LIME による事例の判断根拠説明のイメージ

機序の提示例 2：構造の単純化・最適化・自動構築

深層回路は一般に多層の複雑な回路で人が理解することが難しいため，回路網の規模を小さくすることで理解しやすくしようとする試みがある．回路の構造単純化の手法として，学習後の結合荷重が小さい信号線を削除（＝結合荷重を 0 にする）した後，構造変化に伴う精度劣化を補うために単純化後の回路網で再度学習する pruning や，結合荷重に割り振るビット数を削減し，最終的に論理回路まで圧縮する量子化ベースの手法が以前から提案されている．しかしながら，これらの多くは精度を維持して回路規模を劇的に単純化することは難しい場合が多い．

これに対し，2010 年代後半からは，勾配降下法，進化計算法などの最適化法を用いて結合荷重の学習中に深層回路の構造も併せて最適化する手法や，回路網を目的に合わせて全自動で構築する手法が盛んに研究されており，Neural Architecture Search と呼ばれ，Auto-ML (machine learning) の手法の 1 つとして注目されている．なお Auto-ML は回路網の構造最適化だけでなく，使用する機械学習モデルの選択，モデルの学習係数などのハイパーパラメータの調整，データの前処理，アルゴリズム・プログラムの開発などを可能な限り自動化することで，AI や機械学習に不慣れな人でも機械学習を利用しやすくする手段として，現在盛んに研究されているところである．

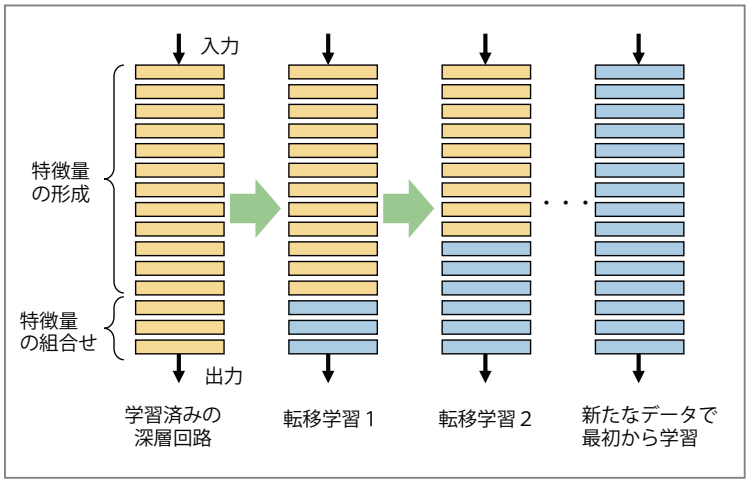
深層回路の構造最適化によって，回路網への入力信号を目的に応じて取捨選択することもでき，入力信号の重要度を求めることも可能である． 筆者らも深層回路を線形回路に変換してレーダチャートや言葉で簡潔に説明する手法を開発している [3]．

他の手法：転移学習・浸透学習

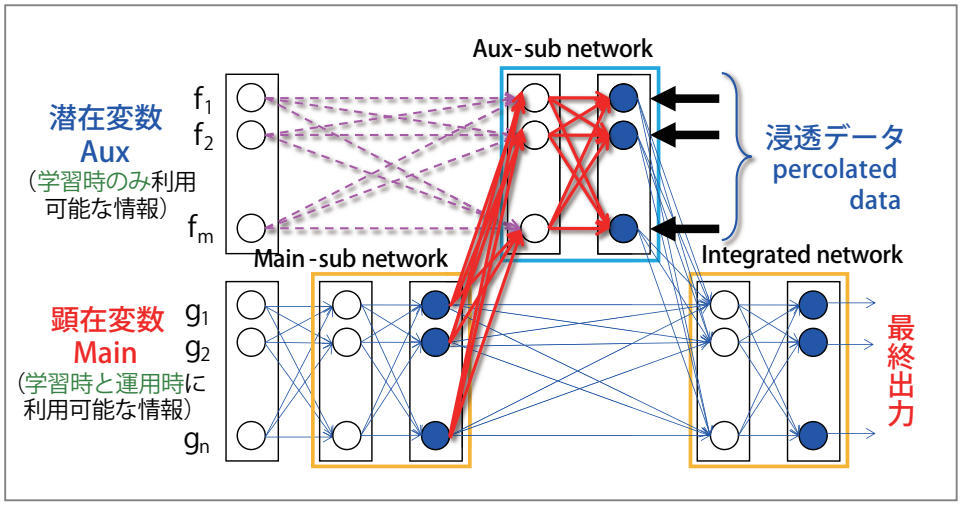
深層学習における転移学習（図 -5）は，XAI の観点からは，すでによく知られている深層回路モデルの知識（＝構造や結合荷重など）を転用することで，利用者にある種の安心感を与える方法と考えることができる．

転移学習では，図 -5 に示すように，類似の目的のために学習済みの深層回路の前段部分（特徴量形成部）の構造・信号の結合荷重をそのまま利用し，後段の特徴の組合せ部を新たなデータで学習することで特徴量の知識を利用する．それによって十分な精度が出ない場合は，初期化する部分を図に示すように入力側に近づければよい．元の深層回路が XAI としても信頼性が高いものであれば，それをベースにしていることで説明性が担保できる．また，Hinton による蒸留（distillation）では，人の教師ではなく，教師の回路が生徒の回路に知識（＝結合荷重や構造）を供与することも試みられている．当然ながら利用する学習済みの深層回路の特徴量が現在の目的にも有効である必要があり，たとえば画像認識回路の知識を音声認識回路にそのまま転移させることは難しい．

扱う問題によっては，学習時には利用できても運用時には利用できない情報（以下“潜在変数”）と，両方で利用できる情報（以下“顕在変数”）がある．通常の深層回路は顕在変数だけを利用しているが，説明性の観点からは潜在変数もできれば利用したいと考える人も多いであろう．筆者らが開発した浸透学習法 PLM（Percolative Learning Method）[4]（図 -6）は，そのようなニーズに応える学習法である．浸透学習では，初めに最終出力に対する勾配降下で全結合荷重を決定した後，図中の“浸透データ”が変化しないよう赤色の結合荷重を調整しながら桃色の



■図 -5　深層回路の転移学習の考え方



■図 -6　浸透学習法 PLM の原理

結合荷重を強制的に減じ，最終的に潜在変数からの信号が存在しない回路を作る．

たとえば，非常に有効だが運用時は高コストで利用できない，あるいはそもそも利用不可の情報（＝未来の情報など）などを顕在変数として学習できるので，深層回路に対する利用者の信頼性・安心感を高めることができる．

XAI の今後の展望

IBL から EBL へ

前項で例示した XAI の手法によって，以前はブラックボックスで説明困難と考えられていた深層回路などの機械学習モデルの判断根拠と機序を人が理解できるようになると，逆に人が機械学習に対して意見を言ったり，要望を伝えたりしたくなる．たとえばある問題で機械学習が“入力変数 No.3 と 4 が最重要”という説明を人に対して行ったとき，仮に人が事前に“入力変数 No.9 が最重要”という知見を持っていると両者の間に齟齬が生じることになる．そして，機械は人に対して No.3 と 4 を組み合わせる方が No.9 より重要である根拠を示すことになる．これが人の“気づき”につながる．逆に人が No.9 を他より重要に扱う深層回路を作ることを機械に要請することも可能である．このように，XAI が人と機械の間の知識の交換や会話（必ずしも自然言語によるものではない）を促すことが期待できる．これによって，多数の事例をただ闇雲にコンピュータに与えるだけの，現在主流となっている“事例に基づく学習”（IBL：Instance Based Learning）から脱却し，人が人に対して説明するように，少数の事例を用いた“説明に基づく学習”（EBL：Explanation Based Learning）へと，機械学習の質点変換を起こすことが期待できる．

人工知能の主流が知識工学であったころのエキスパートシステム全盛の時代に，人の明示的な知識を if ～ then ～型のプロダクションルールなどによって記述して知識ベースに蓄積し，利用することがすでに行われていたが，XAI によって，人と機械が知識を共有することができる新たな形式の知識ベースを創ることが期待される．

XAI から共進化型 AI へ

XAI によって人と機械の間のある種の会話や相互作用が盛んになることで，人と機械が互いに相補的に知能を高め合うことが期待される．これを**共進化型 AI**（CAI：Co-evolutional AI）と呼ぶ．国立研究開発法人 新エネルギー・産業技術総合開発機構（以下 NEDO と表記）では現在，この共進化型 AI のプロジェクトが進行中である[5]．図 -7 にこのプロジェクトの研究開発項目を示す．このプロジェクトでは「人とともに進化する AI システム」の基盤技術を開発し，それらの技術が円滑に社会に適用されるよう，AI システムの評価・管理手法の確立，さらに容易に構築・導入できる AI 技術の開発も併せて行っている．筆者も採択プロジェクトの 1 つの研究代表者として，共進化型 AI の AI 基盤技術開発に携わっており，人の血液中のマイクロ RNA からがんリスクを早期に判定して予防するヘルスケア分野への応用，異常検知処理用 XAI の製造業・金融業などへの応用の社会実装を目指しているところである．

共進化型 AI から職人芸的 AI へ

現在の深層学習は階層型回路網であり，入力を決めると出力が唯一に決まるものがほとんどである．LSTM などの RNN タイプの回路では，過去の入力によって出力が変化するが，これらもある意味では決まったパターンを一定の規則で処理する単純な機構であると言える．今後，進化型ニューラルネットワーク（＝進化計算法などで任意の構造の神経回路網を作る手法）の発展に伴い，非常に複雑な構造を持ち，連想や深い思考などを行う人工脳が出現するかもしれない．そうなると，もはや人が機械の判断を理解することが困難になる．また，同一入力でも

状況に応じて異なる出力をする人工脳は，あらかじめ与えた評価関数の誤差を減らすことで学習させることができなくなる．筆者はそのときに必要なのは“人が人工脳の応答が適切になるように誠実に育てること”であると考えており，このようなAIを勝手に“職人芸的AI（CAI：Craft AI）”と名付けている．大量のデータをもとにして，単純な規則をモデル化するだけの従来型・海外製のAIを人の職人芸による訓練で超えることこそ，Made in JapanのAIが目指すべき道ではないだろうか？

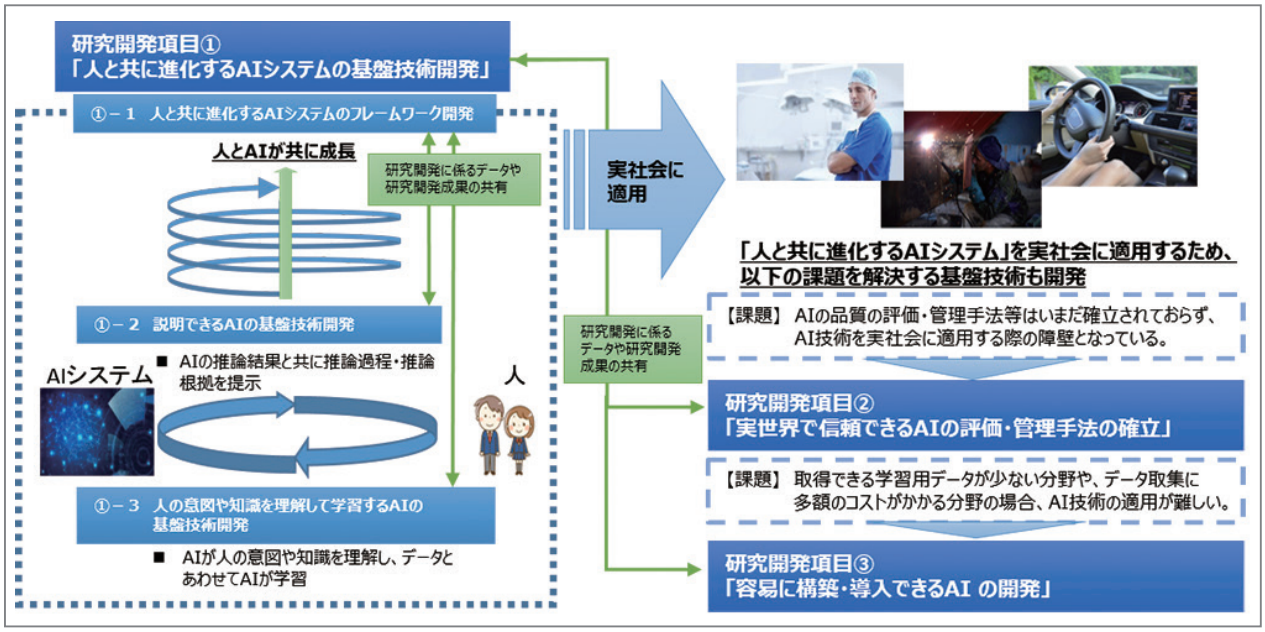
参考文献

1）岡谷貴之：深層学習 改訂第2版（機械学習プロフェッショナルシリーズ），講談社（2022）．
2）内閣府「人間中心のAI社会原則検討会議」，https://www8.cao.go.jp/cstp/tyousakai/humanai/index.html
3）説明生成装置，説明生成方法およびプログラム，国際公開番号WO 2021/020273
4）浸透学習法，国際公開番号WO 2019/031305
5）NEDO「人とともに進化する次世代人工知能に関する技術開発事業」，https://www.nedo.go.jp/activities/ZZJP_100176.html

（2022年4月16日受付）

■長尾智晴（正会員） nagao@ynu.ac.jp

東京工業大学大学院出身（工学博士）．東京工業大学助手，同助教授を経て横浜国立大学大学院環境情報研究院教授（現職）．YNU人工知能研究拠点長．専門は知能情報学．大学発ベンチャーを起業して取締役CTOを兼務中．



出典：https://www.nedo.go.jp/activities/ZZJP_100176.html
■図-7 NEDO共進化AIプロジェクトの研究開発項目