特集
Special Feature

［AI 判断の根拠を説明する XAI を使いこなす］

3 制御の根拠を明示できる XAI の取り組み
― DX × UI による AI 説明性向上技術―

毬山利貞　三菱電機（株）　情報技術総合研究所
横須賀佑介　三菱電機（株）　統合デザイン研究所
穂苅寛光　三菱電機（株）　情報技術総合研究所

XAI を取り巻く背景

AI に説明性が求められる時代に

ディープラーニングをはじめとする AI 技術は，画像認識，音声処理，機器制御などさまざまな分野で活用されている．その流れに伴い，国内では AI プロダクト品質コンソーシアムや産業技術総合研究所などにより，AI の品質に関するガイドラインが発行された．欧州においても同様のガイドライン案が発表されており，今後，さらに AI に求められる機能が明確化されていくことが予想される．

これらの動きの中で，AI の推論過程がブラックボックスになっている点が課題の 1 つとして認識されるようになってきた．AI 技術はいずれも，数式としては明確化されており，その動作はすべて可視化することは可能であり，1 つ 1 つの推論過程を数値から追いかけることはできる．しかし，パラメータ数が膨大であり，現実的には推論結果と結びつけることが困難なことが要因として挙げられる．

この解決策として AI の説明性を向上させる手法群を集めた XAI[☆1] という分野が現れはじめた[1]．これらの手法を大別すると，1）線形回帰など可読性の高いホワイトボックスモデルで AI を構成する手法，2）ディープラーニングのように複雑なモデルをホワイトボックスモデルに変換する手法，3）AI の推論に大きく寄与した入力情報や特徴量を明確化する手法，4）知識グラフのように因果関係を明示する手法，5）環境モデルの構築（グレイボックスモデル，シミュレータなど），6）AI が立てたスケジュールやプランの明示，などが代表例としてあげられる[☆2]．AI 開発者は，AI の性能だけでなく，AI の説明性を向上させるために，上記の候補から適切な手法を選択していくことが求められてくる．

製品開発における課題

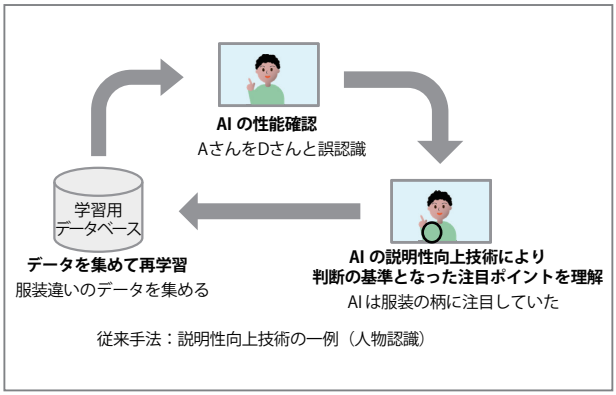
XAI を実現するためにさまざまな手法が提案されつつあるが，この分野には共通とした課題が何点か存在する．その中でも，製品化に向けた大きな課題では，求められる「説明性」がアプリケーションに

☆1 Explainable Artificial Intelligence の略称．
☆2 これらの話題は主要国際会議においても毎年チュートリアルが開催され，Web 上でも一部公開されている．この分野についてくわしく知りたい方は参考文献1）や，学会のチュートリアルも参考にされたい．

よって異なる点が挙げられる．これは，一般生活における「説明性」を考えると容易に理解できる．たとえば，自分の発明物を人に説明するとき，説明する相手によって説明事項は異なる．専門家同士であれば，そのメカニズムや工夫したポイントが説明の中心になり，説明する相手がエンドユーザであれば機能性を中心とした製品の使い方が説明の中心になる可能性が高い．

図 -1 では，XAI 分野では先駆け的な技術ともいえる SHAP[2] をはじめとする，推論に大きく貢献した個所を明示する技術の利用場面を示している．この AI は人物を同定する機能を有する．この図では，AI が A さんを D さんと誤認識した．SHAP のような技術を用いると，今回の推論は主に服装部分によって決定されたことが分かる．説明対象が AI エンジニアであれば，服装違いのデータが少なかったことを理解し，それらの画像データを追加することで性能向上を実現することが可能である．別の使い方として，AI の知見は持っていないが，該当分野の専門家を説明対象とした場合を考えてみる．CT などの画像から病気の原因を推定するような事例を考えた場合，推論に最も寄与した画像個所を明示することで，医師が AI の推論の妥当性も判断できるようなり，AI による誤診断を防止することに貢献することが可能になる [3]．

一方，AI やその機能に対する専門的な知識をユーザが持っていない状況では，推論に一番寄与した画素を指し示されていても，それをもって説明性が向上したとユーザに感じさせることは難しいと考えられる．現状の XAI は AI あるいは対象の専門知識を有することを想定して作られている事例が多い．そのため本稿では，エンドユーザ向けの XAI を組み込んだ事例として，空調機器の制御をテーマとして取り上げ，その技術について紹介する．



■図 -1　XAI の利用シーン例

冒頭でも述べた XAI の手法のうち，AI を制御で用いる場合，スケジュールやプランを提示することが有効な手法の 1 つである．今回は後述するグレイボックスモデルを用いてスケジューリングを可視化する手法を用いた．スケジューリングを行うためには，環境変化をシミュレーションする技術が多くの状況では必要になるため，DX（デジタルトランスフォーメーション）と言われる技術が有効である．DX を活用するためには，そのベースとなるシミュレータと，その各種設定条件や入力条件を特定する必要がある．DX の 1 つの課題は現場にあわせた，この各種設定条件を求めることである．また，物理現象としてシミュレータで再現することが難しい場合には，DX を利用した手法はあっていない．今回想定している状況は，設置環境が既知であれば，シミュレーションによる状態予測の精度は制御を行う上で許容できる範囲内に収まっていることを前提条件としている．

上記の事由から，DX を成立させるための大きな技術テーマとして，シミュレータの設定パラメータの高精度な推測技術，使用環境に依存する入力情報（空調の場合，室内にいる人数，日照条件など）の時系列的な推測技術が挙げられる．以降では，そのための考え方をまず説明し，次に，その解法の一例を示す．最後に，DX にて得られた各種数値を UI（ユーザインタフェース）でエンドユーザに示す方法を示す．

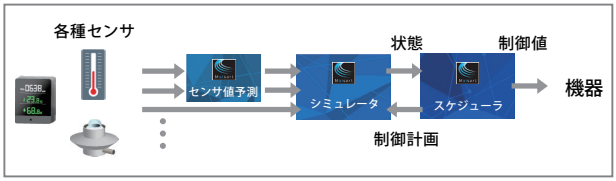
制御の根拠を説明する AI 技術

システム構成

今回紹介する内容は，AI によって空調や換気を制御するシステムにおいて，オフィス内で空調を使用しているエンドユーザに対し，AI による制御が妥当であることを示すための XAI に関するものである．この技術はエンドユーザだけでなく，設備の管理者が空調のクレーム対応などでユーザに制御の根拠を明示しなければならない場合などでも活用できる．

AI は 3 つのモジュールから構成される（図 -2）．それぞれ，センサ値を予測するモジュール，シミュレータ，最適化モジュールである．センサ予測モジュールでは，将来のセンサ値を予測する．シミュレータではセンサ値予測モジュールにて予測した入力値をもとに，環境状態の変化を予測する．また，スケジューラはこのシミュレータを元に最適なスケジューリングを実施する．制御の根拠を示すためには単にシミュレータを持ってくるだけでは不十分だと考えており，その環境の入力（今回のセンサ予測モジュール）も推定する必要がある．DX を実現する上では，これらの要素が欠かせないと考えている．

このシステムでは，AI が機器の設置環境の特性を推定することで，センサで計測できていないシミュレータ上の物理パラメータを特定し数値化することを狙っている．特にシミュレータを利用することで，理論的に意味のある物理量を段階的に計算し，この物理量を見ることで人が機器の周辺で起きている現象を理解することが可能にしようとしている．上記の狙いが達成できれば，制御対象機器のセンサ値など過去の実働データを AI が学習し，将来のセンサ値とともにセンサでは計測できない将来の物理量を予測できるようになる．その結果，将来の機器の設置環境の状態変化をよりシミュレーション可能となり，スケジューラが最適な制御を計画でき，制御計画とそれによる将来の状態を数値化，可視化できるようになる．



■図 -2　システム構成

たとえば，空調機ではセンサが計測していない，設置環境の特性を示す部屋の大きさや断熱性などの物理パラメータを AI が数値化する．次に，過去のセンサ値などの実働データ（部屋の在籍人数など）を学習し，将来の各時刻に出入りする人数や，センサでは計測できていない将来の室内熱量などの物理量を予測する．その結果，空調機が動作した場合に，設置環境の状態を表す室温がどのように変化するのかシミュレーションでき，そのシミュレーション結果を用いてスケジューラが最適な制御計画（機器の稼働率など）を導く．ユーザは，将来の出入りする人数などのシミュレーション結果と制御計画を見ることで，制御の根拠と制御計画の妥当性を理解することが可能になる．

推定方式

本開発では，設置環境ごとのパラメータ推定技術を開発した [4]．方式について空調機器を例として説明する．室内の状況に応じて制御し，室内温度を一定に保つ場合を考える．室内温度の変化は，室内を出入りした熱量に対する微分方程式としてあらわされ，特に一階の常微分方程式で記述できることが知られている．通常のオフィスには複数の空調機や換気機器が設置されており，それらの相互作用により室温が変動するが，今回は簡単のため，部屋が 1 つあり，そこに空調機が 1 つ，換気機器が 1 つ設置されているシンプルな空間をシミュレーションする．その場合，シミュレータは最も単純な形だと以下の支配方程式に従う形で構築できる．

$$\frac{dT_{in}}{dt} = \frac{1}{C}\left(W(t) + Q(t) + \left(\frac{1}{R_{wall}} + \frac{q_v(t)}{R_{fan}}\right)(T_{ext}(t) - T_{in}(t))\right)$$

C，R_{wall}，R_{fan} は時間によって変化しない部屋特有の物理パラメータであり，それぞれ部屋の熱容量，壁面の熱伝導率，換気による外気との熱伝導率である．t は時間であり，$T_{in}(t)$，$W(t)$，$q_v(t)$，$Q(t)$，$T_{ext}(t)$ はそれぞれ時刻 t における室内温度，空調機の仕事量，換気量，室内の熱負荷，室外温度を表す．このうち，$Q(t)$ 以外は空調機および換気機器のセンサから真値を得ることができる．$Q(t)$ は各種の熱負荷の和として現れる．例として日射による熱負荷，在室人数による熱負荷，機器による熱負荷がある．熱負荷の真値は直接得ることが難しく，収集したセンシングデータを用いて値を推定する．

推定には最小二乗法（OLS）を用いることができるが，平時の空調運転データでは得られるデータの多様性が低く，多重共線性☆3 が起こり物理パラメータの推定が正しくできない恐れがある．それを防ぐため，日中の通常運転だけでなく，夜間の日射や人が存在しない時間に，空調機を運転させたデータを収集し活用するなどの工夫が必要である．

物理パラメータが求まれば，それをもとに室内をシミュレーションすることが可能となる．将来の室温変化を予測する際には，将来の熱負荷 $Q(t)$ を推定する必要があるが，それには熱負荷の構成要素である日射量や人数の将来値を予測することで導出する．十分な過去データがあれば時系列予測や，時間帯を説明変数とした回帰モデルにより，それぞれの将来値を推定することができる．ただし推定値と実値にはある程度の誤差が生じるため，後述するユーザへの予測結果の提示の観点でも，予測値は誤差範囲も含めて求めることが望ましい．

将来の予測される熱負荷 $Q(t)$ が分かれば，式を用いて，室内温度を一定に保つ空調機の仕事量 $W(t)$ を求め，スケジューリングすることが可能となる．支配方程式は定まっているが，物理変数が設置環境によって変わるグレイボックスモデルを活用することで，予測から制御の一連の流れは，「人が将来○人になり，日射が△％増加／現象するため，将来の熱負荷がＸＸと予測される．それに合わせて空調機の仕事量をＹＹと変化させることで室温を一定に保つ」といった形で記述することができ，ユーザへの制御の根拠を明示することにつながる．また，将来の熱負荷予測が外れ，室温が一定に保たれなかった場合も，その理由がどこにあったのかを具体的に提示することが可能となる．

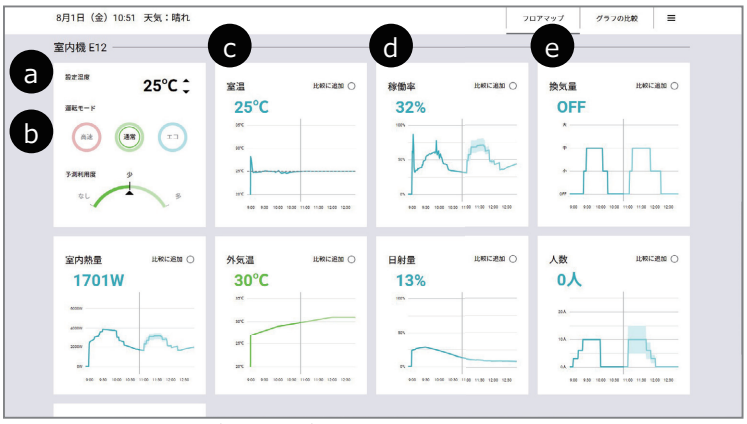
ユーザインタフェース

AI の制御の根拠を，エンドユーザに提示する UI について説明する．UI を検討するために，ユーザインタビュー等の人間中心アプローチを実施した結果，エンドユーザにとっては，判断根拠に加え，以下の３つの要素を提示することで，安心して AI を利用できると考えられた：

(a) AI の今後の振る舞いの見通しが立つこと．
(b) 振る舞いの変化に気づくこと．
(c) 分からない場合に自分で調べられること．

そのため，上記の要件を満たすように，空調機の状態を可視化する画面を図 -3 のようにデザインした．以下では，本画面

☆3 相関係数の高い説明変数が存在している状況．



■図 -3 空調機を例にプロトタイプした UI

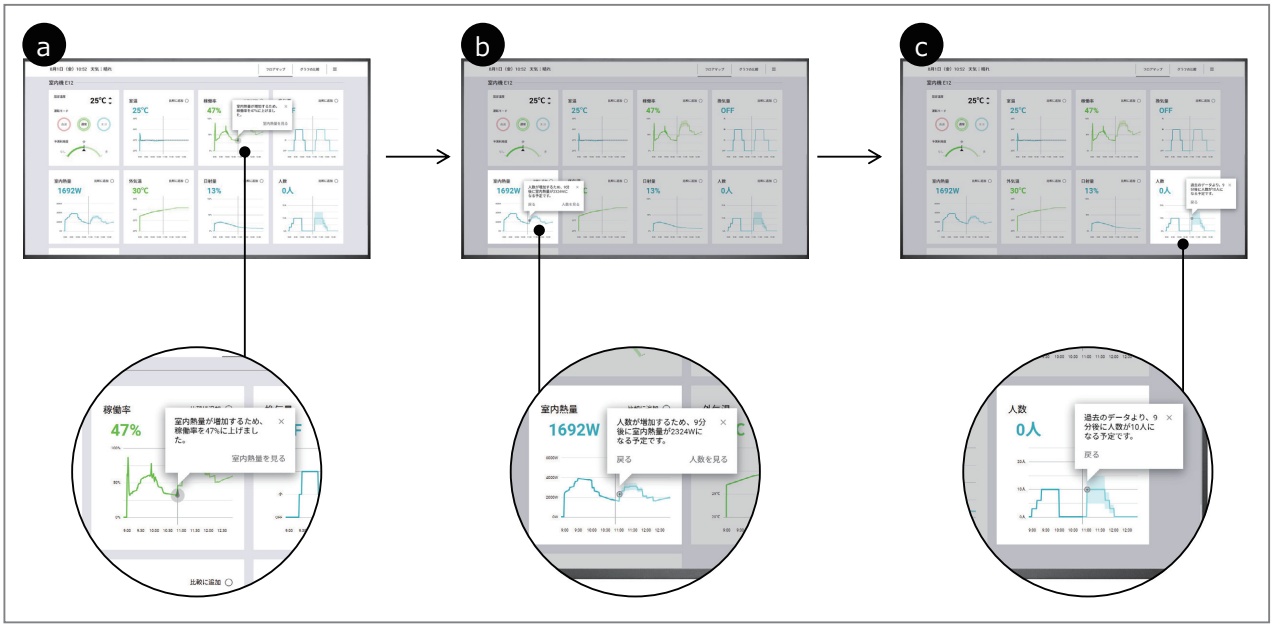
で，上記要件を UI として，どのように実現したのか，具体的に説明する．

実現方法の説明の前に，画面構成を説明する．画面の左上には，現在の設定温度（図 -3a）や運転モード（図 -3b）が表示されたパネルがある．その横には，室温（図 -3c），空調機の稼働率（図 -3d），換気量（図 -3e）など，空調機内部のパラメータや，空調機が利用する外部環境のパラメータを表示するパネルが並んでいる．また，これらのパネル内では，現在の値と，その時系列での変化を表示する折れ線グラフを表示している．

本画面における，上述の 3 つの要素それぞれの実現方法について説明する．まず，「(a) AI の今後の振る舞いの見通しが立つこと」であるが，これは，上述の各パラメータ用のパネルにおける時系列の折れ線グラフで表現している．このグラフの中心にある垂直線は，現在時刻を示しており，その左側が過去の実績値，右側が将来の予測値である．予測値には，予測の誤差範囲も併せて表示している．これらにより，これから空調機がどう振る舞うのか，また，その振る舞いを引き起こす外部環境のパラメータには，どのような変化が起きそうなのか，あらかじめ理解でき，ユーザは安心できる．

次に，「(b) 振る舞いの変化に気づくこと」について説明する．本 UI では，外的要因の変化や AI 予測の不調により，急に振る舞いが変更された場合に，関係するパラメータのパネルに対し，振る舞い変更の理由を説明する吹き出しを表示する．図 -4 に例を示す．図 -4a では，空調機の稼働率のパネルに，値を変更した理由として，「室内熱量が増加するため，稼働率を上げる」ことを自然言語で説明している．パラメータの変更を単に示すだけでなく，自然言語で説明することで，エンドユーザでも理由を理解しやすい．

吹き出しには，「室内熱量を見る」というボタンもある．このボタンを押下すると，図 -4b に遷移し，「人数が 9 分後に増加するため，室内熱量が増加する」ことが説明され，室内熱量が増加する理由を確認できる．さらに，「人数を見る」ボタンを押下すると，人数が増加する理由も深掘りして確認できる（図 -4c）．なお，これらの説明が可能となるのは，前節までで説明したグレイボックスによる推



■図 -4　吹き出しによる振る舞い変化の表示

定方式を利用しているためである．このように，深掘りして理由を確認できることで，単に振る舞いの変化だけでなく，空調機がどのように判断しているのか，という内部の仕組みの理解も促し，ユーザとAIとの間の信頼関係構築を狙っている．

ここで，すべての理由を一度に表示せず，1つのパラメータに対する説明だけを，吹き出しで表示した理由を説明する．空調機について理解が十分でないユーザにとっては，一度に変更の理由をすべて表示することは，認知的な負荷になる．また，本UIに慣れてきたユーザは，最初の吹き出しだけで，その理由まですぐ分かるため，すべての理由を表示すると，かえってわずらわしさを感じてしまう．そのため，一度に表示する情報量を必要最低限に減らし，ユーザに深掘りする選択権を与える形のUIとした．

最後に，「(c) 分からない場合に，自分で調べられること」である．本来，AIでユーザが納得できる情報をすべて提供できることが望ましいが，ユーザによって得たい情報は異なる場合がある．そのため，ユーザ自身がその原因を調べられることが重要である．空調機のように，複数の機器を制御するAIシステムでは，機器の故障発生時などに，ほかの機器と比較することで，原因が探しやすくなる．そのため本UIでは，すべての機器と一覧で比較可能なヒートマップ表示（図-5a）と，時系列のグラフで各パラメータの振る舞いの違いを詳細に比較できる機能（図-5b）の，2つを用意した．

このように，適用分野に合わせて，エンドユーザの要求をUIとして具体化していくことで，エンドユーザの安心感・満足感の醸成が期待できる．

XAIの課題と将来の展望

本稿では，空調制御を例に専門知識を有するとは限らないエンドユーザ向けのXAIの開発事例を紹介した．XAIはアカデミック主導で手法の開発が行われている側面が強いため，本稿では商用化においての注意点を中心に記載した．

システムを実際に開発した後の所感としては，XAIは製品企画段階において説明項目やシナリオを詳細に決定する必要性が従来のAIよりも高まっていると感じている．従来のAI開発にもとめられてきた要求性能（正解率，ノイズ頑健性，データの陳腐化など）は，各種ガイドラインなどで整備されつつあり，明確化されつつある．その一方，XAIで取り扱う説明性の能力に対してはユーザやアプリケーションの依存性が高く，分野として確立できていない状況だと感じている．実際，説明性に対する画一的な評価指標はないのが現状である．

a. ヒートマップでの比較　　b. グラフでの比較

■図-5　比較機能

説明性は AI 技術そのものだけでなく，通常の製品企画に通じる感性が必要となるため，AI の専門家のみで進めていくことは難しい状況だと思われる．UI 部分はユーザを説得するために有力な手法であるため，UI と AI を融合した研究体制を構築していく必要性を感じている．本稿の後半部分では UI を中心に説明しているが，UI に表示したい内容を数値化できるような XAI 機能が今後，活発に開発されていくことになると予想され，UI と AI の複合領域に期待していきたい．

参考文献

1）Molnar, C. : Interpretable Machine Learning (Second Edition) A Guide for Making Black Box Models Explainable, Ruboss Technology Corp (2022).
2）Lundgerg, S. M. and Lee. S.-I. : A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems (2017).
3）Antoniadi, A. M. et al. : Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review, Applied Sciences (2021).
4）Takahashi, A., Hokari, H., Doi, M., Yoshikawa, N., Mariyama, T. and Ueda, N. : Thermal Balance Points Arising from Auxiliary Cooling/heating to Avoid Multi-Collinearity Problems in Estimating Linear Gray-Box Parameters, submitted.

（2022 年 5 月 20 日受付）

■毬山利貞　Mariyama.Toshisada@ab.MitsubishiElectric.co.jp

東京工業大学大学院総合理工学研究科にて脳情報科学の研究に従事し修了（2010）．2016 年より三菱電機（株）にて人工知能に関する研究開発に従事．同情報技術総合研究所知能情報処理技術部部長．博士（理学）．

■横須賀佑介　Yokosuka.Yusuke@bx.MitsubishiElectric.co.jp

中央大学大学院理工学研究科にて，理論計算機科学の研究に従事し修了（2018）．2006 年より三菱電機（株）にて，コンピュータグラフィックス，ヒューマンコンピュータインタラクションの研究開発に従事．同 統合デザイン研究所主席研究員． 博士（工学）．

■穂苅寛光　Hokari.Hiroaki@ap.MitsubishiElectric.co.jp

大阪大学大学院理学研究科にて，理論数学の研究に従事し修了（2012）．2018 年より三菱電機（株）にて人工知能に関する研究開発に従事．同情報技術総合研究所主席研究員．