人工知能（AI）の証拠能力に関する一考察（2・完）
―専門証拠の許容性の観点から―

横　山　優　斗

目次

2　AI のブラックボックス性について

(1) AI のブラックボックス性

AI はしばしば、ブラックボックスであるといわれる。前述のように、AI はエキスパートシステムと機械学習に分類される。エキスパートシステムと初期に開発された機械学習は、人によってモデルやルールが用意されることが一般的である。そのため、これらは人が AI の推論過程を容易に理解することができるモデルとなる。その意味で、透明性の高いモデルであると言える(63)。

他方で機械学習は、モデルが自身でデータから学習し、行動・改善する。そのため、人間の関与は著しく減少する。そうすると、人がそのモデルの基

(63)　Grimm et al, *ibid* at 29, Jurs and DeVito, *supra* note 16 at 645, 鈴木・前掲注 19）197 頁。

礎となっているルールや特徴量を捉えることが困難になる傾向がある。例えば、ディープニューラルネットワークは、入力層と出力層の間にある隠れ層で、入力されたデータを何度も値変換し、出力にとって重要な情報を抽出して、最終的な結論を導き出す。その構造は極めて複雑なモデルであり、システムの中で、入力と出力の間でどのような変換が行われているかを理解するのは、非常に困難である(64)。我々は、入力とそれに対する出力は手にしているが、その過程を知ることができないのである。それゆえ、機械学習ベースの AI の多くは、ブラックボックスだとされている。

(2) 専門証拠の許容性基準に与える影響

ここで、専門証拠の許容性との関係で問題が生じる。厳格な許容性基準をアメリカにおいて定立した Daubert 判決は、専門証拠の信頼性の判断において、「当然のことながら、焦点を当てるべきは原則と方法論であり、それらが導き出す結論ではない」(65)と判示している。すなわち、Daubert 判決は、科学的分析の結果の正確性ではなく、それを生み出す理論と手法の信頼性の評価を求めているのである。

だとすれば、AI システムが入力に対して正確な結論を出力するものであることは判明していたとしても、それだけでは AI 証拠の基礎となっているシステムの理論・方法の信頼性が認められることにはならず、そのアルゴリズムやモデルの推論過程のテストが求められることになるだろう。例えば、顔認証技術は、それが 90％以上の精度で結果を出力する、ということが明らかになっているだけでは信頼性を認められず、顔認証技術の基礎にある理論や、顔認証技術が一致結果を出力する手法についての検討が必要だと考えられることになるだろう(66)。しかしながら、機械学習ベースの AI はブラックボッ

(64) Ethem Alpaydin, INTRODUCTION TO MACHINE LEARNING (3rd ed. 2014) at 267-277. 鈴木・前掲注 19）197 頁。

(65) Daubert v. Merrell Dow Pharmaceutical Inc., 509 U.S. 579, 595 (1993).

(66) Jurs and DeVito, supra note 16 at 640.

クスであるがゆえに、アルゴリズムや推論過程のテストが困難なのである。

このようなブラックボックス性は、まず、専門証拠の許容性基準の中心的な要素である、（a）①実際に理論・方法がテストされ、その結果が明らかとなっていること、の要素に影響を与える。すなわち、ブラックボックス性が原因で、AI の推論過程や出力に強く影響する変数が明らかにならない場合には、テスト結果のみが明らかになったとしても、どのようにしてテスト結果が得られたのかが明らかにならない。このような場合、その結果の信頼性に対する疑問が生じるだろう。先に述べたように、テストされ、その結果が明らかになっているという要素は、基礎にある理論・方法の信頼性を認める上で中心的な要素である。この要素が否定されるということは、基礎にある理論・方法の信頼性が認められない結論を導きうるのである。

また、（b）②検査機器の正確性、および④具体的な検査方法の適切性に関しても問題が生じる可能性がある。まず、ブラックボックス性の高いシステムにおいては、検査機器の正確性を評価することが困難になる。例えば、ある事案において使用した顔認証技術が 90％以上の精度を誇っているとしても、それだけでは検査機器の正確性が認められることにはならない。なぜなら、顔認証技術は人種・性別などによって偏った結果を出力することが明らかにされているからである[(67)]。システムの推論過程が不明であれば、当該事案において顔認証技術は何を考慮して結論を出力したのかが明らかにならず、システムが偏った振る舞いをしていてもそれが明らかにならない。

また、特に機械学習ベースの AI システムは、入力と出力の間の数値の変換やデータ処理の手順を理解することは困難である。すなわち、システムがどのように判断を下しているのか、どのような基準に従って分析が行われているかが明らかではない。そのようなシステムにおいては、具体的な検査方法の適切性を判断することは困難であろう。例えば、ブラックボックス性の

(67)　尾崎・前掲注 3）157 頁。

高い顔認証技術においては、特徴量（顔の特徴点、形状、色調など）の抽出や一致度の計算過程が不明確である。そうすると、システムは適切な特徴量を重視しているのか、適切に計算を行っているのかが明らかにならないのである。

３　説明可能な AI

以上のように、AI のブラックボックス性は専門証拠の許容性判断に影響を及ぼしうる。他方で、AI 研究もブラックボックスの問題性を認識しており、近年、AI の透明性を確保することを目的とした技術が開発されてきた(68)。こうした技術は「説明可能な AI（Explainable AI, 以下、XAI）」と総称されている。

XAI について考えるにあたっては、「説明可能性 Explainability」と「解釈可能性 Interpretability」を区別することが有益である。「説明可能性」は、あるモデルの出力結果に対する事後的な説明が可能であることを指す。他方で、「解釈可能性」は、モデル自体が決定や推論の過程を追跡できるようになっていることを指す(69)。以下では、まず説明可能性を確保しようとする技術（以下、説明技術と呼称する）について概説し、その後、解釈可能性を確保しようとする理論を概観する。その上で、それらを証拠の許容性基準に取り込むことを試みる。

（1）説明可能性を確保する試み

AI モデルの説明技術は、非常に研究開発の活発な分野であり、その全てを本稿で紹介することはできない。そこで本稿では、説明技術をその射程に応じて、大域的な説明をする（Global Explanation）モデルと局所的な説明をする（Local Explanation）モデルに分類し、それぞれ代表的なものを概説する(70)。

（68）　恵木正史「XAI（eXplainable AI）技術の研究動向」日本セキュリティ・マネジメント学会誌 34 巻 1 号（2020）20 頁以下。

（69）　Garrett & Rudin, *supra* note 25 at 600. 大坪直樹ほか『XAI（説明可能な AI）－そのとき人工知能はどう考えたのか？』（リックテレコム、2022）を参照。

（70）　Vikas Hassija, et al., *Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence*, 16 COGNIT. COMPUT. (2024) at 55, 恵木・前掲注 68）21 頁。

①大域説明型

大域説明型とは、複雑で解釈の困難なモデルを、より可読性の高い近似モデルで代理することによって、複雑なモデルがどのように振る舞っているかを理解できるようにするものである[71]。

大域説明モデルの代表的な例として、Tree Surrogate がある[72]。これは、深層ニューラルネットワークなどを決定木モデルで代理するものである。Tree Surrogate の基本的な動作は以下のとおりである。まず、ブラックボックスモデルを用いて、トレーニングデータに対する予測結果を生成する。その上で、トレーニングデータの特徴量と予測結果を用いて決定木をトレーニングし、決定木モデルを作成する。これにより、決定木モデルは元のブラックボックスモデルに近似した動きをするようになる。

②局所説明型

局所説明型とは、具体的事例でのモデルの個別の出力において、強く影響した変数を説明するものである[73]。

局所説明型の代表的なモデルとして、LIME（Local Interpretable Model-agnostic Explanations）がある[74]。LIME の基本的な動作原理は、以下のとおりである。まず、説明対象である出力データの一部を削除し、ノイズを加えるなどして、近傍データを生成する。その上で、近傍データに対して、説明対象である AI モデルによる出力を取得する。近傍データとそれに対する出力結果を組み合わせたデータを用いて、線形モデルを獲得する。そうすること

（71） Riccardo Guidotti, et al., *A Survey on Methods for Explaining Black Box Models*, 51 ACM COMPUT. SURV. (CSUR) (2018) at 13. また、局所説明を複数集めて統計分析し、モデルの全体的な傾向を示す大域説明モデルもある。恵木・前掲注 68）21 頁。

（72） Jayaraman J. Thiagarajan, et al. *TreeView: Peeking into Deep Neural Networks via Feature-apace Partitioning*, arXiv:1611.07429 (2016) at 2, Guidotti, *supra* note 49 at 33.

（73） 恵木・前掲注 68）20 頁。

（74） Marco Tulio Ribeiro, et al., *"Why Should I Trust You?" Explaining the Predictions of Any Classifier* arXiv:1602.04938 (2016).

により、説明対象である AI モデルに説明対象のデータを入力した際に結果の出力に対して特に影響を及ぼした要素を明らかにする。

なお、AI の説明技術は、説明対象となるシステムの種類によっては、適用可能なモデルの種類が限定されることもある(75)が、LIME はモデル非依存型（Model-agnostic）であるため、どのような AI モデルに対しても使用することができるのが特徴である。

(2) 解釈可能性を確保する試み

Tree Surrogate および LIME のような説明可能性を確保する技術は、モデル全体的な動作傾向を説明し、あるいは個別の分析結果において重要視された要素を説明することにより、AI の判断過程を明確にしようとするものであった。こうした試みは、いわば AI のブラックボックス性を前提とした上で、事後的に説明可能性を確保するものと言える。

他方で、AI の透明性を確保するために別のアプローチを採る見解もある。これによれば、そもそもこうしたブラックボックス AI を刑事手続において利用するべきではない。私たちがなすべきことは、ブラックボックス AI の説明可能性を確保することではなく、解釈可能性の確保されたモデルを構築することだというのである(76)。

この見解が目指すのは、ブラックボックス AI に対応するホワイトボックス AI である。前述のように、ブラックボックス AI は極めて複雑なモデルであり、その推論過程等について理解するのは困難である。それに対してホワイトボックス AI は、比較的単純な計算式で記述され、誰にでも可読性のあるモデルである。

---

(75) 例えば、Grad-CAM は、画像識別モデルの判断理由が合理的であることを説明する手法である、モデル依存型の XAI である。Ramprasaath R Selvaraju, et al., *Grad-cam: Visual Explanations from Deep Networks via Gradient-based Localization*, (2017) https://openaccess.thecvf.com/content_ICCV_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper, at 618-626.

(76) Garrett and Rudin, *supra* note 25 at 561.

これまで、AI の透明性と正確性はトレードオフの関係にあると考えられてきた。すなわち、AI の透明性を高めようとすると出力の正確性が失われ、出力の正確性を高めようとすると透明性は失われる関係にある、と考えられてきたのである。実際、「高性能かつ透明性の高い人工知能システムは原理的に実現が困難である」とも言われている(77)。

しかしながら、近年、透明性と正確性は必ずしも常にトレードオフの関係にあるとは言えないことが明らかになりつつある。例えば、Goethals らの研究によれば、ブラックボックス AI とホワイトボックス AI の出力の正確性を比較したところ、データセット全体の 70％において、両者は同程度に正確だったことが明らかにされている(78)。

そこで Garrett と Rudin は、人間の生命・自由・身体に関わる AI は、ブラックボックスモデルを用いるべきではなく、解釈可能性を備えたホワイトボックスモデルを用いるべきだとする。刑事手続は、まさしく人の生命・自由・身体に関わる手続である。したがって、この見解によれば、刑事手続における AI は、全てホワイトボックスモデルであるべきだということになるだろう。

ホワイトボックスモデルが備えられれば、システムの利用者は、説明技術を用いることなく AI システムがどのようにして入力に対して出力したかを理解することができる。専門証拠の許容性基準に関連させて言えば、ホワイトボックスモデルは推論過程の透明性を有しているために、そのモデルのテスト可能性が確保され、検査機器の正確性が保障され、検査過程の適切さが認められるのである。

ただ、Garrett と Rudin のようにホワイトボックスモデルではない AI によっ

（77）　鈴木・前掲注 19）200 頁。

（78）　Sofie Goethals, et al., *The Non-linear Nature of the Cost of Comprehensibility,* 9 J. BIG DATA (2022) at 30, François Candelon, et al., *AI Can Be Both Accurate and Transparent,* https://hbr.org/2023/05/ai-can-be-both-accurate-and-transparent (2023)（2024 年 8 月 19 日最終アクセス）.

て生成された証拠は全て許容すべきではないと主張するのは、厳格すぎるように思われる。なぜなら、先に述べたように、既に法科学や捜査には機械学習ベースのAIが取り入れられており、これからもその傾向は続くだろう。そのような状況下で、ブラックボックスAIは全て証拠能力を認めないことを求めれば、事実認定の助けになるような法科学鑑定も、その証拠能力を否定されることになってしまうだろうからである。

４　許容性基準の精緻化

そこで、以上に概観したXAIを、専門証拠の許容性基準に組み込み、精緻化することを試みたい。考えるに、説明対象AIモデルの判断根拠が説明されること、および、AIモデルがGlass-Box/White-Boxであることは、専門証拠の許容性基準のうち、（ａ） 基礎となっている理論・方法を認めるための、①テスト結果の要素に取り入れられるだろう。なぜなら、AIモデルの透明性を確保することは、そのモデルの推論過程を明らかにし、ひいてはテストを可能にすることにほかならないからである。

また、ブラックボックスは（ｂ） 当該事案における検査過程の適切さの要件のうち、②機器の正確性および④具体的な検査方法の適切性に影響を与えることも確認した。前節での検討によれば、まず、機器の正確性を認めるためには、その機器が入力に対してどのような要素を考慮して出力しているかが明らかにされなければならない。また、具体的な検査方法の適切性を認めるためには、単に分析者がAIシステムを適切に操作したということだけではなく、AIシステムがその分析で考慮した特徴量が適切であることも求められるべきである。

したがって、XAIまたはGlass-Box/White-Boxの要素を取り込んだAI証拠の許容性判断基準は、以下のようになる。

【AI 証拠の許容性基準】

（a）　AI システムの理論・手法の信頼性

① AI システムは、XAI によってテストされ、もしくはテストされうるものであること

② AI システムによる分析のエラー率が明らかになっていること

③ AI システムに関するプロトコルが策定・整備されていること

④ AI システムは、関連分野の専門家によって吟味され・承認されていること

（b）　当該事案における検査過程の適切さ

①当該 AI システムを利用した分析者が、当該分野における知識・経験を有していること

②当該 AI システムは、XAI により、その推論過程においても正確であることが明らかになっていること。

③当該 AI システムに入力された資料の同一性・真正性が確認されていること

④当該 AI システムが、XAI により、適切な推論過程で適切な特徴量を考慮して結論を出していることが明らかになっていること。

なお、科学理論に基づく専門証拠／経験則に基づく専門証拠という区別について、成瀬が「専門証拠のほとんどは、……科学理論と専門的経験則を掛け合わせた証拠であろう」[(79)]と認めるように、AI 証拠も、その多くは AI システムとその他の科学技術、もしくは専門的経験則を掛け合わせた証拠になるだろう。そのような証拠の信頼性判断に際しては、「各々の性質に応じて、信頼性判断を使い分けることが合理的」である[(80)]。

(79)　成瀬・前掲注 33）法学協会雑誌（5）37 頁。

(80)　成瀬・前掲注 33）法学協会雑誌（5）38 頁。

Ⅳ　おわりに

以上、AI 証拠が刑事裁判で用いられる場面を念頭に、その許容性について検討してきた。本論で述べたように、AI 証拠は専門証拠の一種として理解されうる。そこで、AI 証拠に対しても、成瀬によって提唱された厳格な許容性基準によってその証拠能力を分析するのが妥当である。ただし、厳格な許容性基準のうちのいくつかの要素・要件が、AI のブラックボックス性を理由として機能しなくなる場合が考えられることを指摘し、こうした問題に対処しうる XAI について概説し、それを厳格な許容性基準に組み込むことを試みた。

AI の発展はめざましく、これからますます刑事手続に与える影響が強くなっていくことが予想される。本稿が、証拠法の観点からみた刑事手続と AI について、今後の検討の一助となれば幸いである。

最後に、本稿では十分に検討することができなかった課題に言及しておきたい。

まず、本論でも説明したように、XAI は AI の説明可能性を確保するための AI 技術である。すなわち、XAI それ自体が信頼性を有していなければならない AI システムなのである(81)。XAI は近年急速に発展している分野であり、新しいシステムが日進月歩の勢いで開発されている(82)。そのような XAI の中には、AI システムの説明可能性を十分に確保できない、今後の技術的な改善が必要とされるものも存在するであろう。未発達な XAI によって、ある AI 証拠を生成するシステムの推論過程、考慮要素等が明らかにされたとしても、

(81)　LIME と SHAP（SHapley Additive exPlanation：LIME に類似した局所説明型の XAI）が AI 証拠を分析する専門証拠として提出された場合を想定して、その信頼性を Daubert 基準に基づいて検討した論考として、Varun Bhatnagar, *The Evidentiary Implications of Interpreting Black-Box Algorithms*, 20 NW. J. TECH. & INTELL. PROP.（2022）がある。Bhatnagar によれば、SHAP および LIME はその信頼性が認められる。

(82)　原聡「説明可能 AI」人工知能 34 巻 4 号（2019）580 頁（「あまりにも多くの説明法が提案された結果、どの説明法が本当に良いのかがわからなくなってきた」）。

それは AI システムの説明可能性を確保するものとはいいがたい。

そのため、信頼性の高い XAI をどのように選択するか、という点については、技術的な観点もふまえたさらなる研究が必要である。

また、本稿で提示した許容性基準は、従来の議論と同じアプローチを採用し、AI 証拠の証拠能力判断を「ゲートキーパー」としての裁判官に任せるものである。しかしながら、同じくゲートキーパーとしての裁判官を前提とする Daubert 基準をめぐっては、そもそも裁判官は AI を含む高度な科学技術、専門技術についてその信頼性、信用性を適切に判断できるのか、ということが、長い間疑問視されてきた(83)。

そこで、アメリカ法においては「コンセンサス・ルール（Consensus Rule）」と呼ばれる考え方が近年提唱されている。この見解は、Daubert 基準を破棄して、ある専門的事実について専門家コミュニティがコンセンサスを形成している場合には、事実認定者はこのコンセンサスを内容とする専門証拠に証拠能力、証明力判断を拘束されるべきだ、とする見解である(84)。このコンセンサス・ルールを参照し、裁判官によるゲートキーパーとしての証拠能力判断から離れ、専門家コミュニティに判断を任せる考え方が、日本法においても検討される余地があるように思われる(85)。

---

(83)　最近では、専門知に関する科学社会学の知見に依拠して、裁判官に専門証拠の内容について判断させることは適当ではないと主張する見解もある。David S Caudill, et al., *Judges Should Be Discerning Consensus, Not Evaluating Scientific Expertise*, 92 U. CIN. L. REV. (2024) 1031.

(84)　Edward K Cheng, *The Consensus Rule: A New Approach to Scientific Evidence*, 75 VAND. L. REV. (2022) 407.

(85)　ただし、専門家コミュニティに判断を委ねることには解決されなければならない問題も当然ある。笹倉香奈「医学的証拠の法廷への顕出のあり方について」後藤昭編『裁判員時代の刑事証拠法』（日本評論社、2021）65 頁以下は、専門家コミュニティの意見を聴く場となりうるカンファレンス鑑定、コンカレント・エビデンス方式の抱える問題点を挙げた上で、こうした尋問方式は、「本質的に刑事裁判とはなじまない」（82 頁）とする。