人工知能（AI）の証拠能力に関する一考察（1）
―専門証拠の許容性の観点から―

横　山　優　斗

目次

Ⅰ　はじめに

近年、人工知能（以下、「AI」とする）技術の発展に伴って、刑事法分野でも AI をめぐる議論が活発に交わされるようになっており(1)、また実際に、国内外で刑事手続に AI 技術が実装され始めている。例えば、被疑者・被告人のリスクアセスメントツール(2)や、顔認証技術(3)は、とりわけアメリカ合

（1）　日本においては、とりわけ刑法学を中心に、AI に関する議論が盛んになされてきた。例えば、石井徹哉編著『AI ロボットと刑法』（成文堂、2022）、渡辺豊ほか『AI と分かりあえますか?:ブラックボックスが生まれるしくみ』（信山社･2024 年）、日原拓哉『AI の活用と刑法』（成文堂、2024）など。

（2）　リスクアセスメントツールとは、被疑者・被告人の逃亡や再犯等の危険性をスコアにして算出する AI ツールである。例えば、ウィスコンシン州、ニューヨーク州などでは、COMPAS（Correctional Offender Management Profiles for Alternative Sanctions）というリスクアセスメントツールが使用されている。

衆国を中心として広く利用されている。日本においては、例えば神奈川県警察が、ビッグデータを利用して犯罪のホットスポット分析に基づいた犯罪予測、問題解決型警察活動を行っている(4)。

こうした状況に鑑みると、AIの関わるシステムにより生み出された証拠（以下、こうした証拠を「AI証拠」とする）が、将来的には公判廷で重要な地位を占めるようになることは、容易に想像されるだろう。他方で、とりわけ日本における議論では、AIシステムが証拠として公判廷に提出されることを想定した検討がなされることは、少なかったように思われる。しかしながら、「〔刑事裁判におけるAI証拠の〕信頼性・信用性を担保するための新たな証拠法規則について考える必要がある」(5)と指摘されているように、AI証拠に対する規律の検討は、刑事証拠法の重要な課題である。

そこで本稿は、証拠法の視点からAIと刑事法に関係する議論に貢献することを試みる。証拠法におけるAIシステムをめぐる論点はさまざまなものがある。例えば、AI証拠に関する議論が積み重ねられ始めているアメリカでは、AIと対面権(6)、AI証拠の真正性(7)、AI証拠と専門証拠の許容性(8)、な

（3） アメリカの捜査機関によって利用されている顔認証技術として、Clearview AI、FBI's Facial Analysis, Comparison, and Evaluation (FACE) Services などがある。日本における顔認証技術を用いた捜査の問題に関する研究として、水野陽一「顔認証技術を用いた捜査手法に対する規制方法－EU、ドイツにおける議論を参考に－」北九州市立大学法政論集49巻1・2号（2021）85頁以下、尾崎愛美『犯罪捜査における情報技術の利用とその規律』（慶應義塾大学出版会、2023）を参照。

（4） 守山正「犯罪予防に焦点を当てたAI活用による刑事司法制度の将来」ICD NEWS 95号（2023）6頁以下。当然のことながら、AIシステムを用いた証拠の収集過程（捜査）の適法性／違法性をどのように考えるかは重要な問題である。しかし、この問題は、専門証拠の許容性判断においては、結論を左右するような影響力を持たない。そこで本稿では、AIシステムを用いて証拠が収集されたことを前提として、その証拠能力について検討する。

（5） 宍戸常寿ほか『AIと社会と法：パラダイムシフトは起きるか？』（有斐閣、2020）311頁〔成瀬剛発言〕。

どがこれまで検討の対象となってきた[9]。本稿は、AI 証拠と専門証拠の許容性について、日本法の文脈で検討する。

Ⅱ　AI と専門証拠の許容性

1　AI の定義について

(1) AI の定義

AI 証拠の証拠能力を論じるにあたっては、AI 証拠とは何か、そもそも AI 証拠を生成する AI とは何か、が明らかにされなければならない。

AI とは「知覚、言語理解、合成、推論、創造性、感情といった認知能力を必要とするタスクにおいて、人間の能力に匹敵しまたはそれを上回る、コンピューターの仮説上の能力」と抽象的に理解される[10]。この AI の定義は、広範なコンピュータープログラムを含むものであるため、本稿の検討対象を

（6） Andrea Roth, *Machine Testimony*, YALE L. J. (2017) , Gabrielle M Haddad, *Confronting the Biased Algorithm: The Danger of Admitting Facial Recognition Technology Results in the Courtroom*, 23 VAND. J. ENT. & TECH. L. (2020), Ian Maddox, *Artificial Intelligence in the Courtroom: Forensic Machines, Expert Witnesses, and the Confrontation Clause*, 15 J. L. TECH & INTERNET (2024), Benjamin Welton, *Meaningful Machine Confrontation*, 76 STANFORD L. REV. (2024).

（7） Daniel Mason Seng, Stephen, *Artificial Intelligence and Evidence Law and Technology*, 33 SACLJ (2021), Palmiotto, F. *Detecting Deep Fake Evidence with Artificial Intelligence- A Critical Look from a Criminal Law Perspective* (2023). Available at SSRN: https://ssrn.com/abstract=4384122

（8） Patrick W Nutter, *Machine learning evidence: admissibility and weight*, 21 U. PA. J. CONST. L.（2018）.

（9） AI と証拠法に関する重要な先行研究である Paul W Grimm, et al., *Artificial Intelligence as Evidence*, 19 NW. J. TECH. & INTELL. PROP. (2021) at 9 は、特に AI との関連が深い証拠法上のテーマとして、証拠の関連性、証拠の真正性、専門証拠の許容性を挙げている（Part Ⅶ A ～ D）。

(10) Grimm et al., *supra* note 9, 14. ただ、AI の普遍的な定義は存在せず、論者により相当の相違があるとされている。日原・前掲注（1）16 頁。

明らかにするにあたってはさらなる限定が必要である。

AIには、その機能に応じて大別すると２種類あるといわれている。「狭いAI（Narrow AI）」[11]と「汎用AI（General AI, AGI）」[12]である。狭いAIとは、特定のタスクを実行するように設計されたAIシステムを指す。それに対して汎用AIとは、人間が遂行できるあらゆる知的作業を適切に実行できるAIシステムを指す[13]。汎用AIは現時点では実現していないとされている[14]。本稿は、現在刑事手続に実装され、もしくは近い将来に実装されうるAIシステムによって生成された証拠を検討対象とする。そうすると、本稿におけるAI証拠とは、「特定のタスクを実行するように設計されたシステムAIによって生成された証拠」になろう[15]。

---

(11) 一般的には、狭いAIは「弱いAI（weak AI）」と呼ばれることが多い。しかしながら、人間より遥かに処理能力の高いAIの性質を考慮すれば、強力ではないという否定的な意味合いの含まれる「弱い」という語よりも、AIの機能する範囲が特化しているという意味で「狭い」という語が妥当だということが、認識され始めている。Kathleen Walch, *Rethinking Weak vs. Strong AI*, https://www.forbes.com/sites/cognitiveworld/2019/10/04/rethinking-weak-vs-strong-ai (2019)（2024年８月20日最終アクセス）。

(12) 汎用AIは弱いAIとの対比で「強いAI（strong AI）」とも呼ばれる。なお、狭いAIと汎用AIは截然と区別されるものではなく、その関係性は相対的であることについては、Walch *ibid* を参照。

(13) 具体的には、(1) ある領域から別の領域へと知識を一般化し、ある分野の知識を別の分野に応用する能力、(2) 知識と経験に基づいて将来の計画を立てる能力、(3) 環境の変化に適応する能力、を備えたものを指す。Walch, *ibid*.

(14) 宍戸ほか・前掲注５）14頁〔佐藤健発言〕。

(15) もっとも、AIは現在もなお急速に発展し続けている技術である。そうすると、汎用AIが将来的には開発される可能性もある。本稿は狭いAIの証拠能力を論じるものの、今後、汎用AIにより生成された証拠の証拠能力を検討する必要性が生じるかもしれない。その場合には、以下で論じるAI証拠の特徴、およびそれに即した許容性基準論も再検討を迫られるであろうことには、注意が必要である。ただし、汎用AIの開発をそもそも目指すべきではないという指摘もある。鈴木貴之編著『人工知能とどう付き合うか』（勁草書房、2023）13頁〔鈴木〕。

なお、AI は、エキスパートシステムと機械学習にも分類されることがある[16]。エキスパートシステムとは、ある問題について人間の専門家が行う判断を模倣したコンピューターシステムである[17]。人間の判断を模倣するので、そのモデル構築、ルール作成等は人間の手によって行われなければならない。そのため、そのモデルの仕組みや基礎となるデータは人間に理解可能であり、透明性が高い。

それに対して機械学習とは、コンピューターがデータに基づいて自身でモデルを作成し、予測や行動を生成･改善する一連の手法を指す[18]。機械学習は、次のような手順で構成される[19]。

---

(16)　Michael Wooldridge『AI 技術史：考える機会への途とディープラーニング』（インプレス、2022）71 頁以下。エキスパートシステムと機械学習システムの詳細については、Andrew W Jurs & Scott DeVito, *Machines Like Me: A Proposal on the Admissibility of Artificially Intelligent Expert Testimony*, 51 PEPP. L. REV. (2024) at 616-639 も参照。

(17)　エキスパートシステムの代表的な例は、ある判断材料を基に人間が下す判断を if…then ～からなる文で構成したプログラムである、ルールベースシステムである。例えば、「１．（if）歩道が濡れており、スプリンクラーが作動していない場合、（then）雨が降ったことになる。」と「２．（if）歩道が濡れておりスプリンクラーが作動している場合には、（then）雨は降っていないことになる。」というルール構築した場合に、ユーザーが「歩道が濡れておりスプリンクラーが作動していない」と入力すると、上記のルールを適用して「雨が降った」という結論を出力する。こうしたルールを数多く構築して、入力に対して適切な出力をするようにするのである。Jurs & DeVito *ibid* at 616.

(18)　Christoph Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, https://christophm.github.io/interpretable-ml-book/（日本語版：https://hacarus.github.io/interpretable-ml-book-ja/index.html）, 10. いわゆる第三次 AI ブームを引き起こしたニューラルネットワークおよび深層学習は、機械学習の一類型である。

(19)　鈴木貴之『人口知能の哲学入門』（勁草書房、2024）71 頁。機械学習には、教師あり学習、教師なし学習、強化学習の３種類が存在する。教師あり学習とは、入力と出力からなる正解データから関数を学習する。教師なし学習では、入力だけからなるデータから、システム自身がパターンを学習する。強化学習では、システムの行動に対して与えられる報酬から学習する。詳細は、同 71-72 頁、日原･前掲注１）19-20 頁を参照。

１．データの収集と前処理

機械学習モデルを構築するために、大量のデータを収集し、データの正規化などを行う。

２．モデルの学習

収集された学習データを用いて、モデルがデータからパターンや関係性を学習し、最適なパラメータを調整する。この学習プロセスにおいて、特によく用いられるアルゴリズムとして、人間の脳の神経細胞（ニューロン）の働きを模倣したニューラルネットワークがある。ニューラルネットワークは、入力データを複数の層（入力層、隠れ層、出力層）を通じて処理し、最終的な出力を得るものであり、複数の隠れ層を持つものをディープニューラルネットワークといい、複雑なパターンを学習することができる(20)。

３．モデルの評価と適用

構築されたモデルを用いて、テストデータによってその精度を評価する。評価の後、モデルは実際のデータに対して適用される。

機械学習は、以上のプロセスのうち、特にモデルの学習と評価・運用をシステムが自身で行う。そのため、エキスパートシステムに比べて人間の関与が著しく減少するためにシステムの効率が向上する一方で、モデルの学習プロセスの透明性が減少する。これを、ブラックボックス問題という。

本稿は、エキスパートシステムと機械学習を区別せずに、どちらもAIとして論を進める。ただし、専門証拠の許容性基準との関係で本稿が特に問題視する、AIシステムのブラックボックス性は、とりわけ機械学習に顕著に表れる問題であることは、後述のとおりである。

(2) 刑事手続におけるAIの利用：DNA混合資料の解析

AIはさまざまな形で刑事手続において利用されうる。本稿では、２つの例

(20) 鈴木･前掲注19）96-110頁､日原･前掲注１）21-22頁。ディープニューラルネットワークの中には、数十から数百の隠れ層を持つものも存在する。

を取り上げる。

最も容易に想像され、かつ実際に実装が進んでいるのは、法科学分析である。例えば、複数人の DNA が混合している資料（DNA 混合資料）は、対象者の DNA 分析をするために混合資料に関与している者をまず特定しなければならず、DNA 鑑定における困難かつ重要な課題だと捉えられてきた。

DNA 混合資料の解析はこれまで、数学や統計学を基礎とするソフトウェアを利用して専門家によって行われてきた(21)。ところが近年、AI の発展に伴って、DNA 混合資料の解析に機械学習を応用する技術が開発されている(22)。この技術は、現在の解析ソフトウェアに用いられているエキスパートシステムと機械学習を組み合わせた手法（ハイブリッドニューラルネットワーク）である。この手法では、従来のエキスパートシステムでは考慮に入れられなかった要素（DNA の劣化、サンプル沈殿・収集・輸送中の環境条件、サンプル収集の場所など）を考慮して分析結果を出力することができる(23)。これにより、DNA 混合資料のより迅速かつ精度の高い分析が可能になると評価されている。

また、混合資料の分析以外にも、機械学習は法科学のさまざまな分野で活

（21） アメリカでは TrueAllele や STRmix、日本では Kongoh といったソフトウェアが主に使われている。成瀬剛「DNA 混合資料の解析結果の証拠能力―アメリカ法・オーストラリア法の動向―」刑事法ジャーナル 65 号（2020）35 頁以下。DNA 混合資料の解釈手順の詳細については、日本 DNA 多型学会「ヒト DNA 混合資料の DNA 検査の結果の解釈におけるガイダンス」（2022）を参照。これらのソフトウェアは、エキスパートシステムの一種であるから、本稿の定義によれば AI の範疇に入る。

（22） Michael A. Marciano & Jonathan D. Adelman, *PACE: Probabilistic Assessment for Contributor Estimation— A machine learning-based assessment of the number of contributors in DNA mixtures*, 27 FORENSIC SCI. INT. GENET. (2017), Bernardi D., *Forensic Scientists Design the First Machine Learning Approach to Forensic DNA Analysis* (2021), https://news.syr.edu/blog/2021/07/28/forensic-scientists-design-the-first-machine-learning-approach-to-forensic-dna-analysis/（2024 年 8 月 5 日最終アクセス）. なお、Mark Barash, et al., *Machine Learning Applications in Forensic DNA Profiling: A Critical Review*, 69 FORENSIC SCI. INT. GENET. (2024) を参照。

（23） Marciano, M., & Adelman, J., *U.S. Patent No. 10,957,421*, (2021) at 8.

用される可能性を有していることが示されている[24]。法科学にAIを取り入れることが「客観的かつ不可逆的な進歩」[25]であるとすれば、今後、AI証拠が公判廷頻繁に用いられるようになることも十分に想定できるだろう。

（3）刑事手続におけるAIの利用：顔認証技術と顔貌鑑定

もう１つの例が、顔認証技術の活用である。顔認証技術とは、映像や画像内の顔を特定し（顔検出）、その顔の目と目の間や口の大きさなどの特徴点を抽出し（特徴点検出）、データの検索・個人の識別を行う（顔照合）ことにより、人を特定する生体認証技術である[26]。そのアルゴリズムは、膨大な顔データを用いてトレーニングされる、機械学習を基礎とするものである。例えば、この技術は、犯行現場の防犯カメラ映像等から取得した顔画像を用いて、その人物が特定のデータベース内に存在するかどうかを自動的に照合し、顔の微細な特徴を解析し、精度の高い一致判定を提供することができる。顔認証技術は、アメリカにおいては捜査の「きっかけ（lead）」として既に広く利用されており、日本でも捜査段階で利用されている[27]。

他方で現在、防犯カメラ映像等に映った人物の顔画像と被疑者の顔貌が一致するかどうかの鑑定（以下、「顔貌鑑定」とする）は、形態学的検査法、スーパーインポーズ法、人類学的検査法の３種類が主に実施されている[28]。これらの鑑定方法はいずれも、分析対象となる写真および照合される人物が特定された上で行われる。また、（関与の程度は異なるものの）分析者による観察、

（24） Barash et al. *supra* note 22 at 3. 例えば、Barashらは、そうした可能性の例として、犯行現場の遺留資料から被疑者の顔写真を合成すること、部分的な遺体の骨格サンプルに基づいてその人物の顔貌を復元し身元確認のために役立てること、などを挙げている。

（25） Barash et al. *ibid*. at 13.

（26） 国立国会図書館調査及び立法考査局『生体認証技術の動向と活用』（国立国会図書館、2019）１頁、水野・前掲注３）85-86頁、尾崎・前掲注３）153頁。

（27） Brandon L Garrett & Cynthia Rudin, *The Right to a Glass Box: Rethinking the Use of Artificial Intelligence in Criminal Justice*, 109 CORNELL L. REV. (2024) at 581および水野・前掲注３）86頁を参照。

分析、判断等が鑑定のいずれかの段階で必要である。

しかし、これらの従来の方法は時間と労力を要し、また鑑定者の主観に依存する部分が少なからず存在する。近年では、遺体の個人識別において、機械学習をスーパーインポーズ法に導入する試みが進んでおり、その有用性が明らかにされている(29)。この試みが拡大すれば、顔貌鑑定にも機械学習を取り入れる可能性があるように思われる。そうすると、顔認証技術の照合結果が、顔貌鑑定として公判廷に提出されることが考えられるのである(30)。

2　専門証拠の厳格な許容性基準

(1) AI 証拠の分析枠組み

それでは、AI 証拠はどのような証拠法上の枠組みによって検討することができるのだろうか。先に述べたように、アメリカ法の下では、専門証拠の許容性基準が、AI 証拠の証拠能力を検討するひとつの重要な枠組みとなっている。なぜなら、連邦証拠規則 702 条は「知識、スキル、経験、訓練や教育によって専門家として適格とされた者」が専門家証人（専門証拠）にあたると規定しているところ、AI システムは「知識、経験、訓練や教育」がなければそれを利用することができないからである。

他方で、日本の刑事訴訟法には、連邦証拠規則 702 条に相当する専門証拠の許容性に関する規定が存在しない。そのため、AI 証拠がアメリカ法の下で専門証拠の許容性と関連して検討されているということだけを根拠に、日本法の下でも専門証拠の許容性の枠組みで論じることができる、と即断することはできない。

---

(28)　小川好則ほか「三次元画像からの人類学的計測法の検討」法科学技術 21 巻 1 号（2016）95 頁以下、宍戸ほか・前掲注５）310 頁〔成瀬剛発言〕。各鑑定技法の詳細については、船戸宏之・福間匠「画像鑑定が関連する事案のうち、顔貌など人の同一性が問題となるものを巡る刑事実務上の諸問題」判例タイムズ 1513 号（2023）37 頁以下を参照。

(29)　今泉和彦「遺体の個人識別への Ai 画像と AI（人工知能）の活用」Rad Fanl8 巻３号（2020）61 頁以下。

(30)　宍戸ほか・前掲注５）310 頁〔成瀬剛発言〕。

この点について、日本においても、専門証拠の許容性に関する議論は積み重ねられてきた。そして、専門証拠の許容性の問題は、刑事訴訟法 317 条の解釈論としての証拠の関連性の一部として位置づけられてきた(31)。また、成瀬も、「AI〔システムによって生成された証拠は〕……事実認定者である裁判官・裁判員が通常有する経験則によってその内容を評価できないという意味で、『専門証拠』と位置づけることが可能」と指摘している(32)。そこで、本稿は、専門証拠の許容性の観点から AI 証拠について論じたい。

現状、専門証拠の許容性に関する日本の議論は大きく２つの考え方に分かれている。１つは、専門証拠は、証明力判断に委ねる前にその証拠能力を厳しく審査する必要があると考え、厳格な許容性基準を課すものである。他方で、専門証拠の証拠能力はある程度緩やかに認めた上で、裁判官・裁判員の証明力判断に委ねるという見解がある。

(2) 専門証拠の厳格な許容性基準

専門証拠に対する厳格な許容性基準を提示する有力な学説として、成瀬の見解が挙げられる(33)。成瀬は、アメリカ法、イギリス法、オーストラリア法との比較法研究によって、日本における専門証拠に対する厳格な許容性基準が採られるべきであると主張した。成瀬の見解の基本的な枠組みは以下のとおりである(34)。

まず成瀬は、科学理論に基づく専門証拠と、経験則に基づく専門証拠を分類する。その上で、それぞれに対する許容性基準を展開する。科学理論に基づく専門証拠に対しては、満たさなければならない要件として、（ａ） 科学

(31) 司法研修所編『科学的証拠とこれを用いた裁判の在り方』（法曹会、2013）。

(32) 宍戸ほか・前掲注５）311 頁〔成瀬剛発言〕。

(33) 成瀬剛「科学的証拠の許容性（1）～（５·完）」法学協会雑誌 130 巻１号～５号（2013）、同「科学的証拠の許容性」刑法雑誌 53 巻２号（2014）34 頁以下。

(34) 以下の概要については、成瀬・前掲注 33）法学協会雑誌（5）36 頁以下および、同・前掲注 30）刑法雑誌 47 頁以下を参照。

理論・方法の信頼性と、（b） 当該事案における検査過程の適切さ、の２つを提示した。

成瀬は、１つ目の許容性要件である科学理論・方法の信頼性を認めるためには、①テスト結果、②エラー率、③プロトコルの作成・整備状況、④他の専門家による吟味・承認といった４つの要素を検討すべきだとしている。次に、当該事案における検査過程の適切さという２つ目の要件が認められるために、①専門家の知識・経験、②検査機器の正確性、③検査資料の同一性・真正性、④具体的な検査方法の適切性といった要素を考慮するべきだとする。そうして、証拠能力が問題となっている科学理論に基づく専門証拠の、基礎にある科学理論・方法の信頼性および当該事案における検査過程の適切さが認められた場合に、証拠能力が認められるとするのである。

なお、成瀬は、（a） ①「テスト結果」が、「基礎にある科学理論の信頼性を検証する最も直截かつ有効」[35]であるとして、科学理論・方法の信頼性判断における中心的な要素だと考えている。

他方で、経験則に基づく専門証拠は、その基礎にあるのは科学理論というよりも高度な経験則である。そこで、経験則に基づく専門証拠に対しては、（a） 専門的経験則・手法の信頼性と、（b） 当該事案における検査過程の適切さ、という２つの要件を定立した。そして、前者を検討するために、①専門家の経験の集積の程度、②判断実績、③事実認定者に対する判断基準の説明、④専門家の知識・経験、⑤他の専門家による吟味・承認、という５つの要素を提示した。後者については、科学理論に基づく専門証拠と共通する要件、すなわち、①専門家の知識・経験、②検査機器の正確性、③検査資料の同一性・真正性、④具体的な検査方法の適切性、を考慮して決するものとした。

なお、成瀬が指摘するように、（a） 科学理論・方法の信頼性または専門

(35)　成瀬・前掲注 33）法学協会雑誌（5）36 頁。

的経験則・手法の信頼性を認定するためのそれぞれ４つないし５つの要素は、考慮「要素」であって不可欠の「要件」ではない[36]。それに対して、（ｂ）当該事案における検査過程の適切さは、４つの「要件」であるから、それら全てが認められなければ、専門証拠は許容されない[37]。

(3) 専門証拠の緩やかな許容性基準

他方、司法研究『科学的証拠とこれを用いた裁判の在り方』は、以上の学説に比べて緩やかに証拠能力を認める基準を提示している[38]。司法研究は、厳格な許容性基準を課す学説の目指すところに理解を示しつつも、そうした目的は、「必ずしも証拠能力や関連性といった証拠の入口の要件でしか達成できないものではな〔く〕……信用性や証明力といった証拠の実質の判断において十分に達成することが可能」[39]だと主張する。

そこで司法研究は、科学的原理や技法、当該事案における検査・判定方法に重大な欠陥や大きな疑問がなければ、自然的関連性が認められるべきだとした。その上で、証拠調べを行う価値（証拠調べの必要性（狭義））と証拠調べを行うことによる弊害（証拠調べの相当性）から構成される証拠調べの必要性（広義）を判断する（刑事訴訟規則 189 条の２）。こうした証拠調べの必要性は、裁判官の合理的な裁量で判断されるものである[40]。

証拠調べの必要性の観点から一定の証拠の絞込みの可能性があるとしても、「重大な欠陥や大きな疑問」がなければ証拠能力が認められる点で、司法研究の基準は学説に比べて緩やかだと言えよう。こうした緩やかな証拠能力基準は、その後も主に実務家によって支持されている[41]。

---

(36) 成瀬・前掲注 33）刑法雑誌 51 頁。

(37) 成瀬・前掲注 33）法学協会雑誌（5）38 頁、成瀬・前掲注 33）刑法雑誌 49 頁。

(38) 司法研修所編・前掲注 31）37 頁。

(39) 司法研修所編・前掲注 31）25 頁。

(40) 法曹会編『刑事訴訟規則逐条説明　第２編第３章－公判』（法曹会、1989）61 頁、司法研修所編・前掲注 31）37－38 頁、成瀬・前掲注 31）法学協会雑誌（1）61 頁。

3　専門証拠の許容性基準と AI 証拠

以上のように、日本においては専門証拠の許容性に関する動向は二分されており、確立した基準は存在していない。それでは、AI 証拠に対しては、厳格な許容性基準と緩やかな許容性基準のどちらを採用すべきなのであろうか。本稿は、厳格な許容性基準によって AI 証拠の許容性を検討するのが妥当であると考えている。

成瀬は、専門証拠の許容性の判断基準に関して以下のように指摘している。すなわち、専門証拠には特有の危険性があるため、慎重な姿勢で証拠能力・証明力判断に臨まなければならない。他方で、刑事裁判において裁判官と裁判員が証明力を判断する際には、専門証拠も含む全ての証拠を総合考慮して事実認定を行う必要がある。そのため、個別に専門証拠の信用性を慎重に判断しようという姿勢を維持することは困難になる。それに対して、証拠能力段階で証拠が取り上げられれば、裁判官はその証拠だけを見て厳格に審査できるために、慎重な判断がなされる(42)。

この論理は、AI 証拠についてはより強く妥当するように思われる。以下に論じるように、AI システムは、入力とそれに対応する出力（結論）しか明らかにしない場合が多い。そこでは、システムがどのようなアルゴリズムに基づいて計算を行いその出力を生成したのか、出力の際にどのようなデータを重視したのか、などは不明である。

---

(41)　佐伯恒治ほか「座談会『専門証拠』をどのように扱うか」季刊刑事弁護 108 号（2021）55 頁〔成瀬剛発言〕、安東美和子「DNA 型鑑定」大澤裕・川出敏裕編『刑事訴訟法判例百選〔第 11 版〕』（2024）145 頁。司法研究の立場を採用していることを判示において明らかにした裁判例として、大阪高判令和 3 年 1 月 28 日高刑集令和 3 年 8 号 319 頁以下。

(42)　成瀬・前掲注 33）法学協会雑誌（5）28 頁。他方で近年、公判前整理手続の長期化が問題視されている（最高裁判所事務総局「裁判の迅速化に係る検証に関する報告書（第 10 回）」（2023）162 頁以下）。このような状況下で、AI 証拠の証拠能力を公判前整理手続において厳格に審査する場合には、この傾向をさらに促進してしまう可能性があることには、注意が必要である。

このようなシステムの推論過程が不明な証拠の証明力を慎重な姿勢で判断することは、困難である[43]。例えば、ある機械学習ベースの顔認証システムが、防犯カメラ映像に映った人物と被告人を同一人物だと結論づけたことを内容とする顔貌鑑定が提出された。ところが、実は、その結論は誤っていた（映像内の人物と被告人は同一人物ではない）。それにもかかわらずAIが同一人物だと判断したのは、AIが学習したデータが、明るい場所での写真が中心だったため、暗い場所での顔識別は十分にトレーニングされていなかったからである[44]。

このように、機械学習アルゴリズムは特定の状況で誤作動を起こすことがある。そのような場合に、システムの学習データやそれが個別の分析で特に重要視した特徴量が明らかにならなければ、顔貌鑑定の一致結果の証明力を事実認定者が適切に評価することは困難であろう。事実認定者が、「法廷〔引用者注：証明力判断〕で、機械学習の出力にどう反応するかを予測することは非常に困難」[45]なのである。

だとすれば、AI証拠は、事実認定段階に顕出させる前に一定の要件を課して、その信頼性を確保しておく必要性が認められる。したがって、AI証拠の証拠能力を緩やかに認めた上で裁判官・裁判員の証明力判断に委ねるのではなく、証拠能力段階で厳格な規律を及ぼすのが妥当であろう。

なお、本稿では厳格な許容性基準を採用してAI証拠の許容性を検討するが、司法研究の基準によれば、AI証拠は比較的容易にその証拠能力を認められることになるだろう。なぜなら、AIシステムは、「機械的な外観と一見単純な

(43) 学習データが機械学習ベースのシステムにもたらす事実認定における影響について、Nutter, *supra* note 8 at 951-952 を参照。

(44) このように誤って一致の結果を出すことを偽陽性といい、誤って不一致の結果を出す偽陰性と共に、（AIを含む）コンピューターの引き起こす大きな問題と認識されている。Michael I Jordan, *Artificial Intelligence—The Revolution Hasn't Happened Yet*, 1 HARVARD DATA SCI. REV. (2019) at 2.

(45) Nutter, *supra* note 8 at 956.

出力により、客観性と確実性を有しているように見せかける」[46]ために、AI の専門家ではない裁判官が「重大な欠陥や大きな疑問」を発見することは困難だからである。

Ⅲ　AI 証拠の許容性について

1　厳格な許容性基準による AI 証拠の許容性

(1) AI 証拠は科学理論／経験則のどちらに基づくのか

そこで、成瀬の基準によって AI 証拠の許容性を検討する。

検討の前に、まず AI 証拠が、科学理論に基づく専門証拠と経験則に基づく専門証拠のどちらに分類されるかを明らかにしなければならない。この点について、AI 証拠は経験則に基づく専門証拠とみなされる可能性がある。なぜなら、AI システムはデータを「学習」して「訓練」されるという、臭気犬選別をはじめとする経験則に基づく専門証拠と同じ特徴を有しているからである。実際に、機械学習は、アルゴリズムがブラックボックスであり、私たちは「入力について理解しており出力を受け取るが、内部の仕組みがどう機能しているかを完全に理解することはできない」点で臭気犬と同じであるという指摘もある[47]。

しかしながら本稿は、AI 証拠は科学理論に基づく専門証拠として理解すべきであると考えている。なぜなら、まず、端的に、AI は数学、情報工学等のさまざまな科学理論に基づいて開発されるものであり、それが生成する証拠は「基礎にある科学理論なしではその証拠の意味がおよそ理解できない」[48]からである。

また、ひるがえって、（考え方によっては証拠能力段階／証明力段階という違いはあれども）専門証拠が他の証拠に比べて厳格に審査されなければなら

---

(46)　Andrea Roth, *Trial by Machine*, 104 GEORG. L. J. (2016) at 1269-1270.

(47)　Michael L Rich, *Machine learning, automated suspicion algorithms, and the fourth amendment*, 164 PENN. L. REV. (2016) at 912.

(48)　成瀬・前掲注 33）法学協会雑誌（5）35 頁、同・前掲注 33）刑法雑誌 47 頁。

ない理由を考えてみると、それは、専門証拠の有する「危険性」に起因する[49]。すなわち、専門証拠は、非専門家である事実認定者が有していない専門的知識を内容とするために、その内容の理解、信頼性の評価、信用性の評価等において、事実認定者の判断を歪める危険性があるのである[50]。

この「危険性」は、科学理論に基づく専門証拠と経験則に基づく専門証拠で違いがあるように思われる。具体的には、科学理論に基づく専門証拠は、その基礎にある科学理論や手法についての知識がなければ、当該証拠の具体的な妥当性等について理解することが困難である。そのため、科学理論について専門的知識を有していない事実認定者は、その信頼性や信用性を過剰にまたは過小に評価するおそれがある。

他方で、経験則に基づく専門証拠は、一般的な経験則を発展させた高度な経験則がその基礎となっている[51]。事実認定者は、この高度な専門的経験則を有していないので、経験則に基づく専門証拠の信頼性や証明力の判断を誤るおそれがある。この点は、科学理論に基づく専門証拠と同様である。ただし、高度な経験則の基となっている一般的な経験則は、非専門家である事実認定者であっても理解が及びやすい[52]。そうすると、事実認定者は、経験則に基づく専門証拠の内容を一定程度理解することができるために、その証拠に対する慎重な姿勢を崩さずに信頼性や証明力の判断をすることができ、誤った判断をする危険性は、科学理論に基づく専門証拠に比べて相対的に低いように思われる。

この点について、AI システムは一般的に、入力に対して結論を出し、その推論の過程を明らかにしないことが多い。加えて、AI システムはさまざまな

---

(49)　司法研修所編・前掲注 31）8 頁。

(50)　司法研修所編・前掲注 31）8-9 頁。

(51)　成瀬・前掲注 33）刑法雑誌 48 頁。

(52)　例えば、経験則に基づく専門証拠のひとつである筆跡鑑定は、「人によって筆跡は異なる」という一般的経験則を発展させたものだとされている。成瀬・前掲注 33）47 頁。

科学に基づいて構築されるものであるために、その動作は、非専門家である事実認定者の理解が及ばない。前述のように、こうした性質を持つ AI 証拠を事実認定者が過剰に信用し、あるいは敬遠する危険性は、決して「相対的に低い」とはいえないように思われる。このように、AI システムの出力の「危険性」の性質は、経験則に基づく専門証拠よりもむしろ、科学理論に基づく専門証拠に近いように思われるのである。

したがって以下では、科学理論に基づく専門証拠の許容性基準に基づいて、AI 証拠の許容性を論じる。

(2) AI 証拠の許容性基準の具体的検討

(a) 科学理論・方法の信頼性

まず、AI 証拠は、その基礎にある科学理論・方法の信頼性が認められなければならない。注意しなければならないのは、科学理論・方法の信頼性を検討するにあたっては、いくつかの着眼点がありうるということである。例えば、前述の DNA 混合資料の分析についていえば、まず機械学習の信頼性が確認されなければならない。その上で、機械学習が取り込まれた混合資料の分析手法の信頼性が問題となりうる。同様に、顔認証技術についても、顔認証技術の基礎となっている AI ないし機械学習の信頼性と、それを応用した顔認証という手法の信頼性が問題となりうる。AI 証拠の基礎にある理論・方法の信頼性を認めるためには、こうした着眼点のそれぞれについて信頼性を認める要素を備えているかが検討されなければならない。

①テスト結果

成瀬によれば、テスト結果の要素とは、科学的理論・手法について「実際にテストを行い、その結果を確認する」こと、である[53]。この要素を、顔認証技術に関する 2 つの着眼点との関係で検討すると、まず、AI ないし機械学習はこれまで、様々な形でテストされ、その有効性が確認されてきた[54]。また、

(53) 成瀬・前掲注 33）法学協会雑誌（5）36 頁。

顔認証技術そのものの有効性も確認されてきた。例えば、米国国立標準技術研究所（NIST）は、顔認証技術の評価を継続的に実施しており、さまざまな種類の顔認証アルゴリズムの有効性を確認している(55)。

②エラー率

また、AI 証拠は、それを生成するシステムのエラー率が明らかにされているかが考慮される。２つの着眼点との関係では、AI モデルのエラー率とそれが実装されたシステムのエラー率を区別して考えることができるだろう(56)。

AI モデル自体のエラー率は、機械学習のプロセスのうちモデルの評価段階において、テストデータを用いて測定されるものである。ここでのエラー率は、理想的な状況下での性能を反映する。それに対して、モデルが実装されたシステムのエラー率は、現実の環境下におけるエラー率を測定するものである。例えば、顔認証技術に用いられるモデルは、理想的な状況下では高い正解率を誇るかもしれないが、それが実装された顔認証技術は、撮影現場の明るさ、画像の鮮明さなどの現実の使用条件によっては、正解率が下がる可能性がある。

ただし、モデルが実装されたシステムのエラー率は、システムの正確性に直接関係する要素でもある。そのため、これは、AI 証拠の基礎にある理論・方法の信頼性の要素でもありながら、具体的な検査方法の適切性の要件のうちの検査機器の正確性を認めるための要素としても考慮されるべきである(57)。その意味で、基礎にある理論・方法の信頼性と当該事案における検査過程の適切さは重なり合う部分がある。

---

(54) Nutter, *supra* note 8 at 933. 機械学習は 1950 年代に既に登場しており、その後の研究によってさまざまな手法が開発され、

(55) NIST, Face Recognition Technology Evaluation (FRTE) https://pages.nist.gov/frvt/html/frvt1N.html、NEC「世界トップの技術力と取り組みの歴史」https://jpn.nec.com/biometrics/face/history.html（2024 年８月 15 日最終アクセス）。

(56) エラー率について同様の区別をするものとして、Nutter, *supra* note 8 at 933-935.

③プロトコルの策定・整備状況

また、AI 証拠の基礎となるシステムに関するプロトコルが策定されていることも考慮される。ここで、AI そのものに関するプロトコルと、AI モデルに関するプロトコル、それが実装されたシステムのプロトコルを区別することができる。AI そのものに関するプロトコルとしては、各国が策定したガイドラインがこれにあたると言えるだろう。例えば日本では、経済産業省と総務省が公表している『AI 事業者ガイドライン』が、AI の利用に関する基本理念や原則、共通の指針を示している[58]。AI モデルのプロトコルは、システムの開発段階に焦点を当てたものである。これには、モデルの設計、学習、評価に関わるガイドラインや基準が含まれる[59]。他方でモデルが実装されたシステムのプロトコルは、それがどのように運用されるべきか等について定めたものである。

こうしたプロトコルは、AI の基本原則やシステムの使用方法、エラー発生時の対応、データの取り扱いなどを定め、AI システムの信頼性や安全性を確保するための基本的な枠組みを提供する。しかし、現実には、特に AI モデルとそれが実装されたシステムの全てについて、統一されたプロトコルの策定や整備を求めることは困難であることには、注意が必要である。なぜなら、現在、AI は無数のソフトウェアに組み込まれ、日々新たなモデルが開発されているからである。

（57）Nutter も、顔認証アルゴリズムのエラー率は、連邦証拠規則 702 条（ｄ）「事案の事実に対して理論と手法が信頼性を持って適用されていること」、すなわち、成瀬の基準における当該事案における検査過程の適切さの要素との関係で検討している。Nutter, *supra* note 8 at 935

（58）経済産業省・総務省『AI 事業者ガイドライン（1.0 版）』（2024）https://www.meti.go.jp/press/2024/04/20240419004/20240419004-1.pdf

（59）このようなプロトコルとして、例えば Daniel E Ho, et al., *Evaluating facial recognition technology: a protocol for performance assessment in new domains*, 98 DENV L. REV. (2020) がある。

④他の専門家による吟味・承認

最後に、AI システムは関連する分野の専門家によって吟味・承認されているかどうかが考慮される。AI ないし機械学習は、それが開発されてから現在に至るまで、関連する科学界で広く受け入れられていると評価することができるだろう。また、それが応用されたモデルも、開発されてすぐに公表、公刊され、他の専門家による吟味にさらされる。

（b）当該事案における検査過程の適切さ

①専門家の知識・経験

この要件においては、当該 AI 証拠を生成するシステムを利用する分析者が、AI システムの関連分野についてどの程度の知識・経験を有しているかが問われる。

他方で、AI 技術の発達・実装によって、鑑定における人の関与が減少していく可能性がある。顔貌鑑定についていえば、現在はスーパーインポーズ法等に精通した専門家がさまざまな場面で分析に関わっている。例えば、スーパーインポーズ法で異なる画像を重ね合わせた結果を分析し解釈するのは、専門家の役割である。しかしながら、スーパーインポーズ法に機械学習を採り入れた顔貌鑑定技術が実装されれば、システムが自動的に結果を分析し、画像間の共通点や相違点を特定し、一致度を数値で評価することになる。そこでは、専門家の役割は結果の解釈や確認といった限られたものとなり、顔貌鑑定に関わる専門家は、現在ほどの専門性を要求されないことになる可能性もあるだろう。

このように、AI 証拠の許容性判断において、システムを運用する専門家の知識・経験は、なお重要な要素ではあるものの、今後は、その重要性が相対的に下がることも考えられる。

②検査機器の正確性

次に、具体的事例において用いられていた AI システムが正確であったことが確認されなければならない。先に述べたように AI の発達によって人間

の関与が減少していくのに対応して、その AI システムが正確であったかどうかという要件が相対的に重要になっていくことが考えられる。また、前述のように、検査機器が現実の条件下でどれだけの正解率を出すかということについても、検査機器の正確性を認めるひとつの重要な要素として考慮されるべきである。

③検査資料の同一性・真正性

また、AI システムによる分析にかけられる資料は、当該事案と関係のあるものであること（同一性）、時の経過等による変質がないこと（真正性）が確認されなければならない(60)。AI 証拠に関してはその真正性が大きな問題になりうる。例えば、顔認証システムにより被疑者と一致するという分析結果が出力された写真が、実はディープフェイクによって偽造された、全くの別人のものだったというような場合が考えられる(61)。AI 証拠の真正性をどのようにして確保するかについては本稿の検討範囲を超えるが、今後の理論・技術の開発が必要不可欠な重要な問題である。

④具体的な検査方法の適切性

最後に、当該事案において、具体的な検査方法が適切であったかどうかが問われる。

こうして、（ａ）①～④の要素を考慮して AI 証拠の基礎となっている理論・方法の信頼性が認められ、その上で（ｂ）①～④の要件が満たされれば、AI 証拠は許容されることになる。以上の検討によれば、多くの場合に、成瀬の

(60)　成瀬・前掲注 33）刑法雑誌 39 頁。

(61)　ディープフェイクと証拠の問題について、AI 研究においては、ディープフェイク検出システムが開発され、その活用が提言されている。Ebrima Hydara, et al., *Deepfake Detection System for Facial Evidence Verification in Criminal Justice and Its Legal and Ethical Implications*, INTELLIGENT SYSTEMS DESIGN AND APPLICATIONS (Springer 2023)、Hydara Ebrima ほか「刑事司法におけるディープフェイク検出システムの予備的なユーザ評価について」情報処理学会第 86 回全国大会講演論文集 2024 年 1 号（2024）691-692 頁。

基準は、AI 証拠の信頼性を判断するのに役に立つものであると思われる(62)。

他方で、AI システムが有する特徴によって、この厳格な許容性基準の要素・要件のうちのいくつかは、認められない可能性があることには注意を払いたい。そこで以下では、AI 証拠の許容性判断に影響を及ぼすと考えられる AI システムの特徴である「ブラックボックス性」を指摘し、ブラックボックス問題に対する AI 研究の試みを概観する。その上で、厳格な許容性基準を精緻化することを試みる。

(62)　成瀬の基準は、アメリカ合衆国連邦最高裁判所が定立した Daubert 基準に大きく影響を受けたものである。アメリカの議論では、Daubert 基準が AI 証拠に対しても有効に作用するとする見解が有力である。Grimm et al, *supra* note 9 at 93, Nutter, *supra* note 8, 933. 裁判例においても、Daubert 基準に基づいて AI 証拠の証拠能力を判断するものが出ている。エキスパートシステムである STRmix の証拠能力を認めた裁判例として、*United States v. Gissantaner* , 990 F.3d 457 (6th Cir. 2021) がある。