Vol.6 Iss.3 2025
https://doi.org/10.37420/j.mlr.2025.023
Modern Law Research

Research on the Authentication of Big Data Evidence

Bingzhang Fan[*]

Faculty of Laws, University College London, London, UK

[*]Corresponding author, E-mail: as11aca212@163.com

Abstract

As one emerging evidence, big data evidence is mostly presented in the form of electronic data. The authentication of evidence based on data technologies requires strengthened examination of the authenticity of the original electronic data, the identity and integrity of the objective electronic data, and the reliability of the presented electronic data. In judicial practice, authentication tends to be defined within the physical ambit regarding the carriers and information, with the emphasis on the "chain of custody" and "uniqueness identification"; however, the technical authentication is still underdeveloped, such as the integrity check value, read-only with mirror copies, and intelligent authentication. Besides, the examination of the reliability of algorithm is overly neglected, which includes the assessment of "the training dataset", accuracy, adaptability, interpretability and reversibility, and further weaken the trial and make the materialized examination of big data evidence impossible. Therefore, it is necessary to clarify the concept of big data evidence, and then to define the object of authentication and construct corresponding authentication rules, which focuses on the comprehensive authentication both on the physical and informational level, thus responding more scientifically to practical needs.

Keywords

Big data evidence; Type of criminal evidence; Authentication of electronic data; Technical authentication; Reliability of algorithm

1 Introduction

Since the advert of information age, the data have gradually become synonymous with the information. Different with those small data which could be calculated accurately and quantifiable easily, the big data, which cannot be exhaustively enumerated, is progressively becoming the focus and difficulty of judicial practice and realistic dissension. As one information medium to prove case facts, the electronic data are generated, stored and transmitted in digital form[1],which also cover the category of big data evidence and then cause numerous challenges for evidence examination in the new era. The proliferation of complex criminal methods promoted by the developing information technology has complicated the identification and test of evidence, and further facilitated the iterative improvement towards the technologies for the examination of big data evidence. Compared with the more concrete authentication rules and simpler technical requirements of real evidence, the technical problems concerning information and algorithms, which are faced by big data evidence, have become the unavoidable obstacles while establishing the authentication rules. Hence, it is necessary to clarify the evidence attribute of big data evidence and discern the internal demands and external norms of establishing the authentication rules, on the basis of which the focus of examining evidence materials at different levels could be strengthened. Furthermore, more specific guidance for the judiciary to conduct the substantive examination could be provided, with the dual emphasis on the substantive and procedural justice well-placed.

2 Theoretical Foundation for the Authentication of Big Data Evidence

Big data evidence often manifests in judicial practice through electronic data, while they can not be equated entirely. The two types of evidence are homologous in terms of information materials, however, the substantive content presented by which differ. Accordingly, it is imperative to transcend the restraints from statutory category of evidence, then adopting a more scientific approach to examining the big data evidence. Nevertheless, the current legal norms relevant to obtain evidence predominantly concentrate on the alignment between the integrity and authenticity of electronic data[1] and excessively emphasize the examination towards the relationship between the original storage media and electronic data, while neglecting the possibility of information distortion inherent in the data. In judicial practice, the transform of technical difficulties in data collection and process into questions of evidence credibility constitutes an overreach of authentication rules. Hereto, the judicial department has attempted to issue new instruments based on practical feedback, while the conflict between the lagging norms and rapidly-evolving technology still plague the status quo. Therefore, in order to delineate the theoretical foundation for big data, two main questions remain to be answered: what is big data evidence and what is the object of the authentication of big data evidence.

[1] Article 1 of the Provisions on Several Issues Concerning the Collection, Extraction, and Examination of Electronic Data in Criminal Cases issued by Supreme People's Court, Supreme People's Procuratorate, and Ministry of Public Security: "Electronic data refers to data generated during the course of the case, stored, processed, and transmitted in digital form, and capable of proving the facts of the case."

2.1 Analysis of Evidence Nature of Big Data Evidence

Both the theoretical and practical fields have acknowledged the evidence capacity of big data evidence, yet no agreement has been reached as to its category.[1]At its root, the pragmatic adaptations to practice and compliance with its technical nature prevail. While such academic discussions do not impede the widespread and diverse application of big data evidence in judicial practice, the clarification towards the alignment between big data evidence and electronic data is essential for establishing authentication rules, which may enable a clearer understanding of whether current instruments governing electronic data can effectively regulate big data evidence and whether authentication rules tailored specifically to the characteristics of big data evidence should be developed.

Evidence categories are closely related to examination rules, with distinct authentication methods applied to different types of evidence.[2] Although the big data evidence and electronic data are homologous concerning information materials, whether big data evidence can be internalized as a subset of electronic data remains to be discussed. However, for the evidence material that is presented in the form of electronic information, it is feasible to use the authentication rules for electronic data for reference, provided that technical scrutiny is guaranteed. Besides, since the existing authentication rules for electronic data are still incomplete, introducing more diverse perspectives into rules construction with the big data as one sample is also one aim of this paper.

It is undeniable that big data is a technological product of digitization and datafication, being consistent with electronic data within its domain.[3] It is argue that the aggregation of specific data constitutes basis of big data evidence,[4] with huge volume of data as the foundation of big data.[2] Moreover, the distinction between big data and small data is not judged by the mathematical measurement, but rather emphasizes those data that is technically difficult to capture, store, manage and analyze.[2]Some scholars even refrain from distinguishing between electronic data and big data evidence in the broad sense, holding that the difference lies only in complexity and magnitude.[3]These opinions, from different spheres, demonstrate that there is one “source-and-outcome” relationship between the electronic data and big data evidence. Certain empirical research also confirms that over half of big data evidence in judicial practice manifests as electronic data.[4]Those directly extracted big data evidence, whose data are cognizable evidence and connected with the facts of case, should be subject to the examination rules applied to the electronic data.

The view that distinguishes big data evidence from electronic data differentiate the relevance between data and case facts from relevancy in evidence law. This indicates that the correlated inference derived from big data analysis reports differs from the causal reasoning from the electronic data to the final conviction. It is contended that what ultimately has probative values is not the data itself, but rather the big data analysis reports generated through data cleansing, algorithmic modeling and computational analysis.[4] Some

[1] This diversity may stem from the multifaceted nature of big data evidence itself. Scholars have proposed various classification frameworks: 1.source material, big database comparison results, and big data analysis reports (Tang Yufu; Cai Neng. Discussion of “Big Data Evidence” as An Independent Type of Statutory Evidence. Journal of Chongqing University of Science and Technology(Social Science Edition), 2025, 03, 47-57, DOI:10.19406/j.issn.2097-4523.); 2. report analysis, identification comparison and predictive inference (Liang Zemin. The Realistic Dilemma of Big Data Evidence Examination and Its Resolution. Jingchu Law Review, 2023, 06, 66-77, DOI: CNKI:SUN:FXJC.0.2023-06-006.); 3. direct and indirect application based on the the degree of technical processing (Cheng Long. On the Formalization of Cross-check of Big-data Evidences and the Path of Its Substantiation. Political Science and Law, 2022 05, 96-114, DOI:10.15984/j.cnki.1005-9512.2022.05.009.)

[2] “Big data refers to data whose scale exceeds the conventional parameters, making it difficult for general software tools to capture, store, manage and analyze.” See Tu Zipei. Big Data: The Coming Data Revolution, and How It Transforms Government, Business, and Our Lives, 3rd ed.; Guangxi Normal University Press: Guilin, China, 2015; p.57; ISBN: 9787549564101.

[3] Some scholars perceive the big data evidence as a higher-order form of electronic data. See Wei Chenshu. On the Assessment of Big Data Evidence in Criminal Trials. Journal of Anhui University (Philosophy and Social Sciences Edition), 2022, 46(02), 77-86, DOI: 10.13796/j.cnki.1001-5019.2022.02.009.

[4] In most cases, big data evidence is presented in the form of electronic data. In one statistical study of cases, as much as 66% of big data evidence appeared is classified under this form. See Xu Hui; Li Xiaodong. Research on Evidence Attribute Verification of Big Data Evidence. Journal of People's Public Security University of China (Social Sciences Edition), 2020, 36(01), 47-57, DOI: CNKI:SUN:GADX.0.2020-01-006.

categorize such analysis reports within the realm of data science and analogize their examination to that of quasi forensic appraisal opinions.[5] However, this categorization is still under discussion considering that compare with the forensic appraisal opinions issued by professionally qualified institutions or specialized reports by individuals with particular expertise, big data analysis reports are mainly the neutral descriptive accounts about specific data processed by algorithms. Unlike the method of forensic appraisal opinions that compares against established standards, big data analysis reports focus on the descriptive presentations of the analysis and process. While the criticism exists against the categorization as forensic appraisal opinions, such algorithm-processed data material could apply the authentication methods for electronic data without conflicts.

The final form of big data evidence depends more on the judicial practice. The same logic exists between the admissible evidence materials in criminal justice and big data evidence, which means that the crucial concern lies in establishing practical and effective examination norms rather than affirming the nature of new evidence. More emphasis should be placed on the fact that the current legislative provisions concerning the authentication rules for big data evidence and electronic data remain superficial and require further clarification.

2.2 Clarification of the Concept of Authentication of Big Data Evidence

Current legislation do not explicitly address electronic data within authentication rules, which is caused by that the authentication itself lacks recognition and that the existing technologies have not fully addressed all difficulties of data process due to the relatively complex and highly technical standards of the authentication of electronic data. Even for the real evidence, where authentication rules are relatively well-developed, the term “authentication” is not universally applied;[1]with the rules concerning the examination of the authenticity of electronic data to be established, there are significant debates over whether the term “ 鉴真 ” should be adopted.[2]Academia accept the “ 鉴真 ” as the name of such examination,[6] since such a custom of term-using do not impede relevant understanding.[3]To establish corresponding authentication rules of different big data evidence, it is necessary to deeply analyze the concept of authentication so as to understand how the authenticity of data is determined in the current digital information era.

The term “authentication” is expressed in Rule 901(a) of the Federal Rules of Evidence as “the proponent must produce evidence sufficient to support a finding that the item is what the proponent claims it is”.[4]The authentication in this rule requires sufficient external evidence to prove that the evidence material is relevant to the case information. This special relevancy applies to the preliminary admissibility standard set by the Rule 104 of the Federal Rules of Evidence, which stipulates that “when the relevance of evidence depends on whether a fact exists, proof must be introduced sufficient to support a finding that the fact does

[1] Explicit wording such as “identification and authentication of physical evidence, documentary evidence, and other evidence” is used in the notice on developing pilot work by Supreme People’ s Court of Uniform Evidence Rules for People's Courts (Draft Judicial Interpretation Proposal). While other instruments, such as the Rules Concerning with the Evidence Examining and Judging in Death Penalty Cases, contain provisions on the inspection, identification, and examination of documentary and physical evidence, but do not employ the term “authentication”.

[2] Unlike “ 鉴真 ” which relies on expert methodologies or authoritative judgments derived from technical testing or empirical logic, electronic data necessitates comparison through standardized procedures or technological means against objective criteria so as to establish its relevancy to factum probandum. This method, which is based on digital identification processes and compared with standards, could be better termed as “ 验真 ” so as to rectify foundational practices, mitigate excessive interventions from forensic agencies towards judicial practice, and avoid investigatory authorities substituting investigation with forensic expert evidence. Here the article adopts “ 鉴真 ” primarily to align with the prevailing academic custom, terminological semantic ambiguity or terminological disputes.

[3] Rules 901(13) and (14) of the Federal Rules of Evidence were newly added in the 2017 amendment. The provisions on electronic data authentication procedures still await further design in different jurisdictions, and it is more significant to discuss the practical application than semantic wording.

[4] Professor Wang Jinxi employs “ 验真 ” as the translation for “authenticating” in this context, which does not affect conceptual comprehension. See Wang Jinxi. Annotation of the Federal Rules of Evidence of the US, 2nd ed.; China Legal Publishing House: Beijing, China, 2023; p.352; ISBN: 9787521633450.

exist”. [7] In other words, authentication requires prima facie proof that the evidence material is relevant to the case information and that the target material is identical with the source material. Authentication neither require a strong causal link or absolute relevancy between the evidence material and the case information, nor confine itself to specific methods of examining authenticity. All the approaches through objectivity, integrity, identity, reliability and other evidence attributes are just the complement to the authentication, and there is no relationship of inclusion or coverage between these dimensions. On the contrary, these evidence attributes supports each other and jointly serve the purpose of authentication int he auxiliary sense.

When it comes to the authentication rules of big data evidence and electronic data, some scholar establishes rules from the perspective of objectivity and is opposed to the reliance on the subjective cognition. Instead, it is required that the judge should determine the authenticity through the comprehensive understanding of the objectivity of evidence during litigation, to determine its authenticity, which indicates that whether evidence aligns with the objective existence should be determined within the proceedings. This approach focuses on the objectivity of the data itself, the connection between the data and the facts carried by them, and even the link between these facts and case facts. It requires not only ensuring that electronic data presented is neither forged nor false, but demonstrating the relationship between such data and the proof facts.[8] However, the exclusive focus on the objectivity might excessively expands the examination by authentication and unreasonably include the judgement of relevancy.

The construction of authentication rules for big data evidence mainly comprises of establishing specialized examinations corresponding to different forms of data, thereby enabling a more scientific test of the authenticity of evidence materials. Source data generally refers to the original state of data in the case and could be perceived as the origin of the data, which also provides the background and context for subsequent evidence-collection activities. Besides, the conversion from source data to target data requires the extraction and preservation from the original medium, which, according to the relevancy to the case facts, delimits the scope of target data from the boundless data and information. Therefore, it is critical to examine the integrity and identity of data at this stage. Integrity examination concentrates on ensuring that data have not been distorted, damaged, added or modified;[9] while the identity requires ensuring that the data remains identical with the original those during whole process of the extraction, analysis and presentation. [10] Furthermore, for data, there is one separation between its forms of physical existence and information expression. Hence the authentication should be applied to its external medium, the binary digits or programming codes, and the final presentation forms after interpretation and conversion. Finally, when big data evidence is presented in the form of analytical reports, the reliability of the data algorithms should also be subject to the authentication as one addition to the above examination. Compared with the double authentication of storage media and the electronic data itself proposed by some scholar,[11] a tiered framework for authentication rules may be more feasible and applicable for big data evidence. Neither the electronic data without original media nor analysis reports after the physical and informational examinations are absolutely authentic. For those data information that inherently relies on algorithms and technology, the authentication is a tiered process of primary and auxiliary examination, gradually providing judicial personnel and litigation participants with a method to believe the evidence materials as authentic.

3 Objects of the Authentication of Big Data Evidence

3.1 the Carrier of Big Data Evidence

Big data evidence takes data as its carrier. Regardless of the final forms as electronic data or analysis reports, the big data evidence can not be completely separated from its information source. The concepts of source data, target data and statements data[1] referenced in this article primarily distinguish between stages that big data evidence has experienced or may undergo from the perspective of data process, which remain dependent on the carrier as data. As for the authentication rules, the authentication of big data evidence may be analogized to the that of electronic data, which is, however, more technical and requires corresponding specialized examinations to be more comprehensive. Some scholar also divide big data evidence into aggregate data in the macro-level and specific data in the micro-level dimensions, thereby examining whether the data source is objectively formed, whether modifications exist during extraction and transmission, and whether each piece of data is authentic.[4] This approach is largely a self-created framework designed to differentiate the various stages of big data evidence. However, when the data are presented to court from the source, they are not constrained by the "aggregate-specific" dichotomy. Indeed, the shift from the whole to the part does not even align with distinction between the macro and the micro.[2]

The carrier of big data evidence is data, with source data as its basic materials; what enters judicial proceedings is those target data which are relevant to case, with statements data as those analyzed by algorithms. All of these exist in the digital form, which distinguishes them from electronic data that emphasizes extraction from original storage media and the following collection and preservation. The authentication rules of the latter prioritize maintenance of the original medium, which is because the medium for electronic data is constant and immutable and should be unaltered. However, the big data evidence is different sine the source data will not be not constrained by the medium. The scope of this type of data, which is generally unexhaustive and immeasurable in view of all physically correlated data, requires to be defined by legal relevancy. The integrity of source data cannot be restrained and investigatory organs can only analyze and process the target data to generate statements data.

In the field of evidence law, not all digitized data constitutes electronic data. The correlations between data are not entirely equivalent to evidence relevancy. Digitized data pointing to the factum probandum presents verifiable weak correlations that establish a causal link between the data and case facts. Since the evidence lacking relevancy should be directly excluded, the data without correlations should also be exclude, which functions as the further screening at the level of evidence law before data enters judicial proceedings. This process could be supplemented with quantitative assessment, rather than relying solely on linear correlations or the automated induction by data processing tools. At this stage, authentication focuses on "the authenticity of the connection between what the exhibit present and the specific case facts".[17] To

[1] Professor Guo Jinxia divides electronic data based on the the stages of generation, collection and extraction, and presentation and statements. This categorization has its advantages. Based on this, the article decomposes the data processing flow as follows: source data refers to the original data from the case; target data refers to the "data relevant to the case information" obtained by investigatory authorities through collecting and preserving the original data; and statements data refers to the process of interpretation of the data through technical means such as analysis and processing. See: Guo Jinxia, Deconstruction of Identifying Authenticity of Electronic Data. Tribune of Political Science and Law, 2019, 03, 56-66, DOI: CNKI:SUN:ZFLT.0.2019-03-006.

[2] Data flows smoothly from being identified as an information source, then bounded by investigatory authorities according to the relevancy to case information, and finally collected, sorted, processed, analyzed and output as conclusions, all of which occur within the same level of observation. When analyzing issues from institutional logic such as micro or macro levels, multiple layers of analysis are usually assumed to explore the interactions between mechanisms and structures. Obviously, attempts to study big data evidence through the "aggregate-specific" dichotomy only complicate the issue further. See Patricia H. Thornton; William Ocasio; Long Sibo (Translator). Institutional logics perspective: a new approach to culture, structure and process; Zhejiang University Press: Hangzhou, China, 2024; pp.19–21; ISBN: 9787308204767.

be more concrete, the target data that collected from the source data could be either directly applied or processed as statements data.

3.2 Objects of the Authentication

How to examine big data evidence constitutes the core orientation of its authentication rules. As the carrier of big data evidence, data manifest in various forms according to the human subjectivity and objective algorithms. Setting aside the source data which is immeasurable and will not completely enter judicial proceedings and disregarding whether the data are generated automatically by systems and storage of externally-inputted information, or produced in other ways, the primary object of authentication should be target data, which is delineated by investigators based on their understanding of relevancy to case facts. This main point could solve many dilemmas, which also explains why many conflate big data evidence with electronic data: target data can be, both internally and externally, understood as electronic data that has not been examined. Furthermore, target data connects the source data to the potential statement data, both of which are also the objects of authentication, with corresponding authentication rules less complex.

The authentication of source data focuses on the elimination of influencing factors, where the ideal conditions are that systems functions normally, operating environments remain stable and free from interference by “unauthorized” users. It is generally presumed that source data exist in a safe, stable and reliable environment; thus, excessive examination of its authenticity is unnecessary. Besides, this presumption might require data keepers to illustrate the physical integrity of storage devices and the internal security measures such as equipment and environmental conditions, defenses against unauthorized access and error detection capabilities.[8] In other words, the authentication of source data authenticity only requires minimal intervention, and challenges to the authenticity of source data should not be regulated by authentication rules, which examination even transcend the limitations of original medium.[1]

The authentication of target data at the stage of collection, transmission and storage focuses on its identity and integrity related to the source data, analysis results and algorithmic codes within the litigation process. The current authentication methods of target data are chain of custody and uniqueness identification, both of which are primarily comparing and examining the data at the level of electronic data. Uniqueness identification compares the target data with other electronic data through their distinctive characteristics. Such uniqueness manifests not only in the physical distinctiveness of the storage medium but also in the expression forms of the electronic data, such as its format or types of code.[11] Chain of custody is to ensure that the target data is not tampered with throughout the evidence collection procedures. This is assessed by examining whether the chain of custody of the storage medium has been broken and whether relevant factors (such as eyewitnesses and audio or video recordings) are standardized, thereby determining whether the electronic data remains identical and integral.

As the reports of analysis results, the statements data manifests in various forms according to the evolving practical demands, which, however, does not affect the corresponding authentication. Particularly, the emphasis should be placed on the authentication of algorithm reliability, which necessitates the examination

[1] It is not necessary for electronic data that cannot be directly displayed to retain its original form; generally, it can be authenticated through a digital signature or digital certificate. However, not all electronic data comes with a digital signature or certificate, the absence of which should not automatically lead to the denial of its authenticity. See Yu Haisong. Commentary on Practical Criminal Procedure Law, 1st ed.; Peking University Press: Beijing, China, 2023; pp.417–421; ISBN: 9787301359433.

towards both of “training dataset”, and the accuracy, adaptability, interpretability and reversibility of the algorithm. Apart from ensuring the identity and integrity of electronic data, examination of the algorithm reliability also contributes to prove the substantive authenticity of big data evidence.

4 Application of Authentication Methods for Big Data Evidence

When discussing the authentication methods, scholars often argue for distinguishing the double authentication of electronic data carriers and the data itself,[11] emphasizing the limitations of scattered legislative provisions with certain practical issues unsolved,[8] or advocating for introducing the concept of “authentication” in Anglo-American evidence law.[12] However, such text-based analyses, which overlook the speed of technological iteration and development, fail to fundamentally resolve the current problems of the authentication of big data evidence. At its roots, technological iteration inevitably necessitate the constant improvements over the legislation, while the current patchwork or stopgap measures have already incrementally addressed practical challenges. Against this backdrop, it is necessary to examine the authentication methods currently employed in judicial practice, so as to design more specialized authentication methods that are better tailored to the the technologically and algorithmically complex big data evidence.

4.1 Authentication Methods in Judicial Practice

The existing authentication methods could be largely divided into the chain of custody and uniqueness identification. For real evidence, chain of custody and uniqueness identification focuses on the inherent or artificially created[13] physical attributes and external characteristics, the specific measures of which are typically the observation, perception, identification and statements in courts of related personnel. However, unlike traditional real evidence, the big data are virtual in existence, separated from the physical carriers and massive in data sources. Consequently, without technical methods, traditional uniqueness authentication can only identify electronic devices, storage medium, displayed graphic-text information and the types of format or code through personal perception.[14] Thus, prior to the enactment of the Provisions on Several Issues Concerning the Collection, Extraction, and Examination of Electronic Data in Criminal Cases, the chain of custody, rather than the uniqueness identification, was employed as the main method in judicial practice.

Both integrated extraction and separate extraction of electronic data related to big data evidence can apply the chain of custody. The former entails jointly extracting, seizing and transferring electronic data together with its original storage medium under seal, while the latter is mainly applied to network-based electronic data, where the data is directly retrieved and extracted from the original storage medium on-site.[6] This authentication method primarily relies on records by related personnel of the custody and transfer of the electronic data and their carriers. The main mechanisms include transcript, extraction procedures, sealing procedures, eyewitness procedures, and audio-video recording procedures. The examination of transcripts ensures the end-to-end records of the whole process of data and directly traces the the information flow. However, this method inevitably bears certain potential risks, including strong subjectivity, low reducibility, and susceptibility to modification.

The extraction procedures serves to standardize the separate extraction, guaranteeing the integrity and identity of electronic data by requirements such as not storing the extracted data in the original medium and not installing new applications in the target system after extraction.[1]The sealing procedures correspond to the integrated extraction, requiring, for example, separate sealing of hard drives, memory cards and so on; photographing when sealing; and adopting signal-blocking measures for wirelessly-communicable medium.[2]The eyewitness procedures requires that eyewitnesses be present during the collection, extraction, seizure and sealing of electronic data, which enhances the authenticity of electronic data by curbing subjectivity in transcripts. Audio-video recording not only performs a function similar to transcripts in preserving evidence by recording the process of collection and extraction,[3]but also independently serve as a method of collecting electronic data. For example, the ephemeral communication information like "burn-after-reading" messages can be preserved through printing, photographing, and video recording.[9]

4.2 Technical Authentication

The application of technical authentication in judicial practice remains relatively underdeveloped, currently focusing on foundational measures such as trusted time-stamp, digital signature and block-chain record.[14] In addition to these widely utilized yet non-mandatory technical authentication methods, judicial authorities also employ read-only and mirroring technologies to verify the authenticity of electronic data, which involve the read-only devices (data access only and no modification) and mirror copies (one-to-one replications of original and associated data), thereby preserving the evidence.[8]

Additionally, whether artificial intelligence can be employed to remedy the current technical deficiencies of authentication methods remains to be discussed. Some studies have attempted to use AI to directly examine whether the presented big data evidence or electronic data has been modified. For example, in the case of AI-processed photographs, algorithms can determine whether a facial image has been swapped without modifying the original image itself.[2] Compared with the more mature and standardized technical authentication methods mentioned above, AI function more as auxiliary tools to sort data. Given that the underlying algorithmic logic diverges from conventional legal reasoning, the judicial applicability of AI still requires rigorous scientific validation.

To sum up, the gap between academic research and scientific development has, to some extent, constrained bold attempts at technical authentication and limited the continuous iteration of investigatory evidence collection. However, considering the current state of practice, it is necessary to adapt to scientific advances, avoid both excessive caution toward technology and blind admiration for it, and establish authentication rules that are more scientific, rigorous and aligned with the operational principles of technology.

4.3 Authentication of Algorithm Reliability

The authentication of algorithm reliability does not emphasize the disclosure of algorithms, as such highly technical programs are not rendered comprehensible to judicial personnel or litigation participants simply

[1] Article 18 of the Rules for Electronic Data Collection in Criminal Cases by Public Security Departments.

[2] Article 11 of the Rules for Electronic Data Collection in Criminal Cases by Public Security Departments.

[3] Relevant provisions include the Article 5 of Provisions on Several Issues Concerning the Collection, Extraction, and Examination of Electronic Data in Criminal Cases, and Article 41 of the Rules for Electronic Data Collection in Criminal Cases by Public Security Departments.

by being public. Instead, reliability authentication is more about observing the entire process of data analysis and process, thereby preventing malicious, highly subjective, or non-compliant tampering. Specifically, rather than challenging issues such as chaotic or modified "training dataset" of source codes, or anomalies in the scope and timeliness of data collection,[2] reliability authentication concentrates more on the repeatable tests on target data or presented data that appears to contradict common sense or general knowledge. This form of authentication should be embedded within the proceedings, which is inherently repeatable, testable and even subject to scrutiny. When examiners are uncertain about whether industry standards have been followed during data pre-processing stages (such as cleansing and annotations), whether operational rules have been adhered to, or whether the engineers are qualified, the procedures of algorithm reliability authentication should be initiated.

In accordance with the scientific evidence principles, there are mainly two examination standards. Where national or industry standards exist, the corresponding standards should be followed; where there is no such official standards, further examination should be conducted under the Daubert standard, which examines whether the scientific technique has been or can be tested and whether the algorithm has been subject to peer review and been published, as well as evaluating the known or potential error rate and the degree of acceptance of this algorithm within the scientific community.[15] As for those highly technical aspects such as algorithmic compatibility and the issue whether the variables selected by the algorithm involve bias, the assistance from experts within the examiners or the engagement of individuals with specialized knowledge are required, which details would not be further elaborated in this article.

Furthermore, the algorithm interpretability needs to be highlighted. Given the high technical threshold explanatory challenges, there have been persistent advocates for authentication through being public. Some scholars argue that disclosure of contested portions of an algorithm should occur only when the dispute exists and the judge deems it necessary, with enhanced confidentiality measures also requisite.[15] Others contend that in the case of disclosure, neither the parties nor the judge is able to understand basic principles, which necessitates the authentication through reverse deduction. They further suggest treating algorithm interpretation as disclosing the external operational processes of the algorithm, thereby ensuring transparency of the computational and then authenticating the algorithm.[16] Crucially, the undisclosed algorithm is not necessarily unreliable, which could be exemplified by GPS analysis reports.[4]

The authentication rules of big data evidence in judicial practice are far more complex than those discussed in academic research, as technological iterations, practical demands and deficiencies in technical examination continually challenge current judicial practice. Compared with emphasizing specific authentication methods or their limitations, one data processing framework better suited for integrating big data evidence into litigation proceedings enables authentication methods to meet practical and procedural requirements more effectively.

5 Conclusion

The practical challenges in authenticating big data evidence stem partly from technological limitations, which may be solved through future upgrades, and partly from the objective loss of data, which could only be addressed through calculations based on supplementary information. Beyond these, investigatory authorities collect data sources they deem necessary and use statistical tools to establish the connection between the data and the evidence for subsequent aggregated proof. However, during this process, there is a shared but implicit consensus: the data investigated and collected are not absolutely reliable, but are good enough in terms of their trends to prove the factum probandum.[18] This dilemma where the source data have been obstructed remains under-addressed,while this article cannot deeply analyze the underlying causes and consequences. The aim of this article, however, is to clarify certain distinction between big data evidence and electronic data, thereby providing a more feasible and scientific perspective for understanding the principles of big data evidence and strengthening authentication.

References

[1]Wu Hongqi. The Legal Orientation and Theoretical Reflection on the Completeness of Electronic Evidence. Journal of National Prosecutors College, 2024, 01, 146-160, DOI: CNKI:SUN:ZJGX.0.2024-01-010.

[2]Wei Chenshu. On the Assessment of Big Data Evidence in Criminal Trials. Journal of Anhui University (Philosophy and Social Sciences Edition), 2022, 46(02), 77-86, DOI:10.13796/j.cnki.1001-5019.2022.02.009.

[3]Xu Hui; Li Xiaodong. Research on Evidence Attribute Verification of Big Data Evidence. Journal of People's Public Security University of China (Social Sciences Edition), 2020, 36(01), 47-57, DOI: CNKI:SUN:GADX.0.2020-01-006.

[4]Liu Pinxin. On Big Data Evidence. Global Law Review, 2019, 01, 21-34, DOI: CNKI:SUN:WGFY.0.2019-01-003.

[5]Huang Jian. On the Limited Function of Big Data Correlation in Criminal Juridical Proof. Tsinghua University Law Journal, 2023, 02, 22-39, DOI: CNKI:SUN:QHFX.0.2023-02-002.

[6]Xie Dengke. The Authentication of Electronic Evidence. Journal of National Prosecutors College, 2017, 05, 50-72+174, DOI: CNKI:SUN:ZJGX.0.2017-05-003.

[7]Siri Carlson. When is a Tweet not an Admissible Tweet? Closing the Authentication Gap in the Federal Rules of Evidence. University of Pennsylvania Law Review, 2016, 164(4), 1033-1065. Available online: https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=9522&context=penn_law_review&httpsredir=1&referer= (accessed on 15 July 2025).

[8]Guo Jinxia, Deconstruction of Identifying Authenticity of Electronic Data. Tribune of Political Science and Law, 2019, 03, 56-66, DOI: CNKI:SUN:ZFLT.0.2019-03-006.

[9]Yu Haisong.Criminal Electronic Data:Regulatory Approaches and Key Issues. Global Law Review, 2019, 01, 35-47, DOI: CNKI:SUN:WGFY.0.2019-01-004.

[10]Chen Zhuoqin. On the Authentication Rules of Electronic Evidence. Shanghai Legal Research, 2022, 4, 82-88, DOI: 10.26914/c.cnkihy.2022.040628.

[11]Liu Yifan. On Double Authentication of Electronic Data.Contemporary Law Review, 2018, 03, 88-98, DOI: CNKI:SUN:DDFX.0.2018-03-010.

[12]Li Xiangyu; Zhang Xiaoling. Construction of electronic data authentication rules-an analysis centered on formal relevance. Evidence Science, 2024, 01, 83-92, DOI: CNKI:SUN:FLYZ.0.2024-01-008.

[13]Liu Pinxin. The Authentication of Digital Evidence:From the QVOD case.Peking University Law Journal, 2017, 01, 89-103. DOI: CNKI:SUN:WFXZ.0.2017-01-007.

[14]Xie Dengke. Technical Authentication of Electronic Evidence. Chinese Journal of Law, 2022, 02, 209-224, DOI: CNKI:SUN:LAWS.0.2022-02-012.

[15]Hong Tao. The Construction of Rules for the Authenticity of Big Data Evidence. Journal of Soochow University(Law Edition), 2024, 01, 69-82, DOI: 10.19563/j.cnki.sdfx.2024.01.006.

[16]Li Guangqi; Chen Bowen. How to ensure that big data evidence to realize its expected probative value. Available online: https://www.spp.gov.cn/spp/xsdfljd/202205/t20220511_565663.shtml (accessed on 15 July 2025).

Books and Book Chapters:

[17]Ronald J. Allen; Richard B. Kuhns; Eleanor Swift; Zhang Baosheng etc. (Translators). Evidence: Text, Problems and Cases, 3rd ed.; Higher Education Press: Beijing, China, 2006; p.308; ISBN: 7040202646.

[18]Howard S. Becker; Wang Fei (Translator). Evidence. Beijing United Publishing Co., Ltd.: Beijing China, 2023; p.27; ISBN: 9787559669292.