刑事司法科学与治理
Criminal Justice Science & Governance
2025 年 第 6 卷 第 2 期

风险社会视角下人工智能辅助电子数据取证的挑战与策略研究

叶振鑫

中南财经政法大学刑事司法学院，武汉

摘　要 | 人工智能辅助电子数据取证不仅是现实之需，还是未来必然。在为电子数据取证提供发展机会的同时，人工智能亦不可避免地引发一系列取证风险。本文立足风险社会理论，厘清人工智能辅助取证的风险类型，从风险主体、风险节点与风险程度三个维度，系统分析人工智能辅助电子数据取证的风险生成机制，指出当前取证风险具体表现为直接与间接取证风险，前期—实施—应用阶段的流程性风险，以及潜伏期—爆发期—扩散期—平复期的周期性风险特征。在剖析以上风险对法律规制挑战的基础上，本文提出以人的主体地位为核心价值导向，以技术可信化、取证规范与权利救济协同化为目标，构建兼顾技术发展与法律风险平衡的电子数据取证体系，以期为刑事诉讼中人工智能辅助电子数据取证的合规应用提供理论参考，推动技术与法治的良性互动。

关键词 | 风险社会；人工智能；电子数据取证；风险生成机制

一、问题提出

电子数据这一概念首次出现在1991年，美国召开的第一届国际计算机调查专家会议首次提出“计算机证据”（Computer Evidence）的概念，并将其定义为“可以识别、恢复、提取、保存并形成报告使之成为法律证据的电子形式存储的信息”。[1] 在刑事诉讼中，电子数据[2]等表现形式丰富的新型证据，正在逐步取代传统证据在案件证据中的地

［1］刘浩阳，李锦，刘晓宇．电子数据取证［M］．北京：清华大学出版社，2015：3.

［2］电子数据是案件发生过程中形成的，以数字化形式存储、处理、传输的，能够证明案件事实的数据。2016年最高人民法院、最高人民检察院与公安部联合发布的《关于办理刑事案件收集提取和审查判断电子数据若干问题的规定》，中华人民共和国最高人民检察院网站信息，网址：https://www.spp.gov.cn/zdgz/201609/t20160921_167425.shtml，2025年4月19日。

位[1]，电子数据取证[2, 3]面临的技术需求也越来越大，而人工智能技术为其提供了广阔的应用前景，在提升调查取证效率等关键环节具有较大潜力。例如，通过语音自动转换文字，可以快速查阅手机即时聊天记录中涉及的语音对话，使用图像识别技术，可以快速筛选视频中的涉案信息并锁定嫌疑人等。在对数据不断学习、修正的基础上，人工智能不断提高取证的便利性、效率性和准确性。[4]然而，不可否认的是，人工智能的迅猛发展并非没有风险，特别是在电子数据取证这一敏感领域，风险问题愈加复杂多变。人工智能取证的过程中，可能会逐渐引发弱化人在侦查中的主体地位、放大技术缺陷和解释纰漏所带来的偏差、所得证据的法律效力存疑、侵犯或弱化当事人权利等问题，形成技术风险、法律风险与伦理风险交织的局面。因此，如何在保障司法公正与促进技术创新之间找到平衡，成为当前亟需回应的核心问题。

基于此，本文立足风险社会理论，分析当前人工智能辅助电子数据取证面临的主要风险，深入探讨风险生成机制及其对法律规制的挑战，进而提出有针对性的规制策略。这一探索旨在推动人工智能技术在电子数据取证中的应用朝着规范化、科学化方向发展，实现技术与法治的良性互动。

二、人工智能辅助电子数据取证的风险及生成机制

自从人类文明的曙光初现，人类社会的发展便始终与各种挑战和不确定性并行。无论是古代、近代还是现代，每一个历史阶段都伴随着不同形式的风险。这些风险不仅深刻影响着社会的生存与发展，还渗透到社会结构、文化传统和科技进步等多个层面，成为推动社会变革的重要力量。在现代社会的发展过程中，新兴智能技术既具有建设性又具有破坏性[5]，推动着现代社会在创造文明新形态的同时，也带来了全新的风险模态，形成了以“风险”为显著特征的现代社会。[6]随着人工智能等新兴技术的迅速发展，处于现代社会的人类所面临的风险，在结构、特征、影响范围和程度上都发生了深刻变化，产生了现代意义的风险，即所谓的现代风险。[7]德国社会学家乌尔里希·贝克（Ulrich Beck）认为，“风险”的概念直接与反思性现代化的概念相关，是指为系统地处理现代化自身引致的危险和不安全感的方式[8]，且有别于传统的风险，现代风险是现代化威胁力量和全球化所引发的后果，并强调新型科技风险已经具备新的风险特征，用常规的风险概念应对现代生产力和破坏力所引发的后果是一种错误的手段。[9]贝克还将“风险”的概念引入社会学研究领域，提出著名的“风险社会”理论，为深入理解与分析这个充满复杂性与不确定性的现代社会提供了重要的理论视角。

（一）风险

在人工智能辅助电子数据取证过程中，不可避免地会涉及风险社会理论的相关议题，因此将问题置于风险社会理论的研究范畴进行讨论是可行

[1] 参见吴照美，曹艳琼，杨海强．人工智能技术对取证的影响及应用价值［J］．华东理工大学学报（社会科学版），2019，34（3）：93.

[2] 本文中“电子数据取证”“人工智能取证”“取证”皆属于刑事诉讼中人工智能辅助电子数据取证。基于此，三者将不做严格区分，行文中交替使用。

[3] 电子数据取证就是采用技术手段，获取、分析、固定电子数据作为认定事实的科学。刘浩阳，李锦，刘晓宇．电子数据取证［M］．北京：清华大学出版社，2015：7.

[4] 范鑫．电子数据取证介质及取证技术演变概述［J］．网络安全技术与应用，2024（11）：136–138.

[5] ［美］乔治·萨顿．科学史和新人文主义［M］．陈恒六，等译，上海：上海交通大学出版社，2007：45–60.

[6] 参见刘骥，丘霖．生成式人工智能嵌入教育应用的风险生成及其规制［J］．现代远距离教育，2024（4）：12–13.

[7] 陈嘉鑫，李宝诚．风险社会理论视域下生成式人工智能安全风险检视与应对［J］．情报杂志，2025，44（1）：3.

[8] ［德］乌尔里希·贝克．风险社会［M］．何博闻，译，南京：译林出版社，2004：19.

[9] ［德］乌尔里希·贝克．风险社会：新的现代性之路［M］．张文杰，等译，南京：译林出版社，2018：7.

的。[1]乌尔里希·贝克强调，在现代社会中，技术的发展和全球化进程难以避免地会带来风险，这些风险通常具有深远的系统性和不可预测性。尤其在社会各个领域的广泛应用中，技术风险[2]的影响变得愈加显著且难以规避。在电子数据取证过程中，人工智能技术的嵌入在提升效率的同时，也带来了新的风险：首先，可能弱化人在侦查中的主体地位，尤其是当侦查决策过度依赖技术时，人类的判断和经验可能被边缘化，同时技术自身的缺陷或算法中的偏差可能会被放大，进而影响证据的准确性；其次，人工智能所生成的证据在法律效力上存疑，导致证据的可信度受到质疑；最后，也可能非法触碰当事人的隐私保护范围。

1. 弱化人在侦查中的主体地位

大部分取证人员[3]在使用人工智能进行取证时，一般都对人工智能有一定的了解，但由于自身认知的局限性及人工智能领域的专业性，使其缺乏足够的理性认识，很容易产生“技术乐观主义”，对人工智能技术产生“技术崇拜”，将人工智能理想化、绝对化[4]，以至于在算法自动处理和分析大量数据后，高度信任取证结果。这种现象可能使算法从一种取证工具逐渐成为“取证主体”。目前的算法模型无法确保百分之百的准确性，数字痕迹中的错误与偏差并不明显，可能需要细致观察与大量实验才能发现，而取证人员通常仅对算法输出的结果进行确认，忽视对数据来源、算法运作等全过程的全面审查。这就导致人工智能取证结果对模式化处理产生依赖，缺乏对个案细节的深度分析，逐渐削弱人在案件分析中的主动性与判断力，消解了人的主体性。

尽管人工智能技术具有强大的数据处理能力，但它并非完美无缺，尤其是在面对复杂、动态的法律情境时，人工智能系统的缺陷尤为突出。例如，AI系统在处理数据时，算法的设计和优化可能出现漏洞，或者可能会基于训练集中的某些假设或偏见进行分析，导致其分析结果并非绝对准确。而且，人工智能的算法本质上是“黑箱”式的，许多决策和判断过程难以被完全理解和追踪。当人工智能产生错误的结论时，很难追溯问题的根源，再加上人在分析判断过程中的主动性与判断力逐渐被消解，容易忽略特定情境中的人性化考量，在这些因素的共同作用下，将会误导侦查人员和法庭判断，影响案件结果。

2. 人工智能辅助取证所得数据的证据能力存疑

证据能力，即证据资料在法律上允许其作为证据的资格。[5]人工智能取证得到的数据在证据能力上的争议，主要集中于其可信性存疑。[6]人工智能的“算法黑箱”和“算法偏见”问题，是可信性存疑的核心。

在人工智能系统输入的数据和其输出的结果之间，存在着人们无法洞悉的“隐层”，这就是“算法黑箱”[7]，其核心问题在于信息不对称和不公

---

［1］参见吴汉东．人工智能时代的制度安排与法律规制［J］．法律科学（西北政法大学学报），2017，35（5）：129.

［2］技术是使主体在竞争中获得优势并取得收益的一种手段。“技术风险”是指特定主体从事某项技术研究、开发和应用必然面对的遭受损失的可能性。郭瑜桥，王树恩，王晓文．技术风险与对策研究［J］．科学管理研究，2004（2）：60–63.

［3］刑事诉讼中的侦查取证一般由公安机关、国家安全机关和检察机关完成，本文中统一使用“取证人员”。刘浩阳，李锦，刘晓宇．电子数据取证［M］．北京：清华大学出版社，2015：7.

［4］徐奉臻教授认为：“在技术社会学中，学者们通常将技术的正面作用与影响归结为‘技术的社会功能’。技术乐观主义产生于人类对技术的社会功能有所了解但又缺乏理性认识的特定历史条件下，其实质是‘技术崇拜’或‘技术救世主义’，其基本特征是把技术理想化、绝对化或神圣化，视技术进步为社会发展的决定因素和根本动力。”徐奉臻．梳理与反思：技术乐观主义思潮［J］．学术交流，2000（6）：14–18.

［5］参见杨波．由证明力到证据能力——我国非法证据排除规则的实践困境与出路［J］．政法论坛，2015，33（5）：109–122.

［6］参见何邦武．网络刑事电子数据算法取证难题及其破解［J］．环球法律评论，2019，41（5）：69–70.

［7］徐凤．人工智能算法黑箱的法律规制——以智能投顾为例展开［J］．东方法学，2019（6）：78–86.

开。“黑箱”掩盖了算法对于数据的利用，用户无法了解算法如何使用数据、作出决策的逻辑和依据，不可能得知算法设计者、实际控制者及机器生成内容的责任归属，对其进行评判和监督更是难上加难。[1] 例如，侦查机关使用人工智能工具对犯罪嫌疑人的电子邮件进行筛查，试图通过关键词提取、情感分析等技术，自动分析电子邮件内容，标记出可能与案件相关的邮件，确定是否存在犯罪相关的线索，并生成一份证据报告。那么，在这个过程中，人工智能是如何判定哪些电子邮件与案件相关？它是基于什么算法选择关键字、分析情感或识别潜在信息的？如果某些与案件无关的私人邮件被错误标记为证据，或某些关键信息被忽略，侦查机关能否清晰判断由此产生的归责问题？因此，黑箱性导致人们根本无法了解算法究竟是基于何种原因得出的结论，这些原因既有可能是人类较为熟悉且可以接受的，也有可能是人类无法接受的，还有可能是人类无法观察到的。[2]

在“算法偏见”方面，尽管算法决策可以在速度、效率甚至公平性方面带来好处，但有一种常见的误解，即算法会自动生成无偏见的决策。算法看似是无偏计算，因为它们采用客观的参考值并提供一个标准结果，但这些输入和输出仍然存在许多问题。[3] 在电子数据取证过程中，算法依赖于大量历史数据来分析和识别证据，然而这些数据难以避免地存在偏见性。例如，在使用人工智能进行电子邮件内容或社交媒体数据分析时，如果历史数据中过度集中于某一特定群体或偏向某一特定行为模式，算法可能会放大这些特征，使得在分析挖掘时偏向一部分数据。算法偏见还可能源于设计者的无意偏向，人工智能算法的开发者在设计和编程时，难免将个人的主观判断或社会文化背景融入其中。例如，设计用于筛查犯罪相关数据的算法时，设计者可能会无意中设定某些偏向性规则，导致特定群体或行为被过度关注，甚至错误地将无关数据识别为证据。具体来说，某些算法在筛查聊天记录时，可能会根据性别、年龄等因素，错误地将特定群体的行为标记为高风险，从而出现偏差的取证结果。

3. 对当事人隐私保护范围的非法侵入

在取证人员运用人工智能取证时，无论是海量数据分析还是特定电子数据收集，均可能涉及当事人的个人信息处理，极易对当事人的隐私权产生“越轨行为”[4]，引发公权力行为干预公民合法权益的正当性与合法性问题。[5] 这些风险主要源自电子数据，难免涉及大量个人信息的处理，且具有海量性与无形性，一旦不当操作，甚至操作得当均可能侵犯当事人的合法权利。同时，也缺乏明确的法律或规范条文提供参考，指导他们的行为。一旦法律模棱两可的地方过多，就必然会产生一个“自由的荒野”。但不可否认的是，会存在一定的无意识偏差行为，即因为取证人员缺乏关于人工智能的知识、信息和责任，并非故意发生这些行为。[6]

在传统的刑事诉讼中，证据的收集通常局限于特定范围和明确目标，在当事人知情并在应当公开的情况下进行；而在人工智能辅助的电子数据取证中，技术手段可能导致取证范围的过度扩展。以远程取证为例，利用人工智能进行远程搜查取证时，当事人通常不知情，这使得这种电子数据搜查从“公开措施”变成了“秘密措施”。与公开的侦查措施相比，秘密侦查措施对公民基本信息的检索与获取，受到的阻力将大大减小，因为当事人无法及时知晓相关情况，且在未被告知的情况下，一旦出现“越轨行为”也无法进行必要的法律救济，严重侵犯当事人的隐私权。

（二）风险生成机制

在风险社会的视阈下，电子数据取证的风险链由技术风险和法律风险两个子系统组成，准确识别

---

［1］王聪．“共同善”维度下的算法规制［J］．法学，2019（12）：69.

［2］马国洋．论刑事诉讼中人工智能证据的审查［J］．中国刑事法杂志，2021（5）：158–176.

［3］张涛．自动化系统中算法偏见的法律规制［J］．大连理工大学学报（社会科学版），2020，41（4）：92–102.

［4］人超越了一点社会范围的规范的行为，就是越轨行为。张兴杰．现代社会学新编［M］．北京：北京大学出版社，2012：333.

［5］刘品新，谢登科，裴炜，等．电子证据的法治化路径［J］．数字法治，2024（4）：23.

［6］［美］亚历克斯·梯尔．越轨：人为什么干“坏事”？［M］．王海霞，等译，北京：中国人民大学出版社，2014：273.

其中的风险种类及其表现形式，并进行针对性的管理和应对，成为解决由技术进步所引发的社会风险结构、风险节点与影响程度变化的关键措施。基于此，为有效治理这些风险，厘清人工智能取证风险生成机制的复杂结构与内在逻辑，本文将从风险主体、风险节点与风险程度三个层面，剖析人工智能辅助电子数据取证的多重风险，这有助于推动取证风险治理从模糊到精细、从抽象到具体的转变。

首先，在风险主体上，根据风险的递归（Recursiveness）特性，可将取证风险分为直接取证风险与间接取证风险。在取证的实际应用中，就直接取证风险而言，“算法黑箱”掩盖了人工智能处理数据的逻辑路径，其解释力直接影响到证据的透明性与可信性，而这些正是法官确定此证据是否合法的重要依据；同时也可能引发证据生成过程被操控、算法偏见影响证据内容、技术鸿沟限制辩方获取关键证据等问题。就间接取证风险而言，主要体现为人工智能生成的电子数据对法庭审判与取证秩序产生的负面影响。人工智能的训练数据中可能含有偏见性与歧视性内容，这些偏见可能难以被法官察觉且难以缓解，或悄无声息地被放大并渗透到司法决策中，影响案件判决的公平性。此外，美国管理学家巴斯比（Busby J）认为，风险的递归性指风险演进是一个随着时间推移而不断变化的过程，对原始风险（Primary Risk）的反应本身就会产生次生风险（Secondary Risk）[1]，而人工智能生成内容的偏见性，不仅会污染电子证据本身的质量，还可能对未来人工智能取证系统的发展造成递归风险或次生风险。如在风险叠加效应的作用下，技术预警与风险控制机制的不完善，将会导致刑事诉讼中取证秩序的进一步复杂化和失控。同时，在此过程中，人工智能技术可能因其高度依赖数据资源，而成为一种强化社会不平等和司法不公的工具，引发范围更广、后果更严重的次生风险。

其次，在风险节点上，本文根据风险发生的时间进行分类，分为前期训练风险、实施阶段风险与应用阶段风险。从前期训练风险来看，在数据模型的训练阶段，算法开发者可能因使用未经授权的数据而导致数据侵权问题。由于算法的偏见性和透明度不足，取证模型在开发时可能内嵌不公正的分析逻辑，增加后续取证活动中误判的可能性。这种“算法先天缺陷”一旦进入刑事司法程序，可能造成取证工具的不可信任，加剧取证活动的合法性危机。围绕实施阶段风险展开，电子数据取证在执行过程中，可能面临取证人员过度依赖人工智能的“唯科技论”“唯效率论”问题，过度信任人工智能生成的结果，而忽视对取证过程的人工审查，从而引发技术偏见放大的风险。更为严重的是，实施阶段还可能因滥用技术手段导致对隐私权的侵害，这种风险不仅可能削弱公众对司法程序的信任，还可能引发公权力与个人合法权益之间的冲突。针对应用阶段风险，一方面，人工智能的“算法黑箱”特性使得生成的证据缺乏可解释性，法官和辩护人可能难以全面理解证据形成的逻辑与依据，从而降低证据的公信力；另一方面，因人工智能生成报告的结构化内容缺乏灵活性，可能掩盖案件中细微但重要的事实，影响案件的客观公正审理。

最后，在风险程度上，风险的生命周期被划分为潜伏期、爆发期、扩散期与平复期，风险程度经历了由弱到强、由强到广再到消退的演变阶段。[2]潜伏期主要体现在技术开发和应用初期的隐性风险积累。人工智能取证技术的应用尚未完全成熟，由于技术人员对模型特性和潜在问题缺乏全面认知，可能导致模型设计时隐含算法偏见或技术漏洞，这些问题在早期可能因未被实际应用或未显现直接后果而被忽视。由于这种隐蔽性，监管政策和法律法规的制定基本处于滞后状态，使得潜在问题也缺乏有效预警和治理手段。爆发期标志着潜伏期风险在司法实践中被引爆，人工智能取证得到的证据可能因各种原因在法庭上被质疑甚至被排除，在这一阶段，公众对人工智能取证技术的信任迅速下降，案件的公正性也会因此受到质疑。扩散期是人工智能带来的风险从个别案例逐渐蔓延至广泛司法实践的阶段，它可能进一步弱化控辩双方的力量平衡，控方因技术资源的优势而获取更多证据，而辩方则因技术缺失而被迫处于不利地位，从而加剧司法程序的失衡。平复期是风险受到有效控制并逐渐消退的阶段。通过完善的法律法规、技术改进和

[1] 刘骥，丘霖. 生成式人工智能嵌入教育应用的风险生成及其规制[J]. 现代远距离教育，2024（4）：13.

[2] 杨波，孙白朋. 基于风险生命周期的企业反竞争情报机制模型构建[J]. 现代情报，2019，39（11）：30–37.

伦理约束，人工智能在电子数据取证中的风险逐步降低。与此同时，风险规制经验的积累，也将推动技术和政策的迭代优化，进一步规范人工智能在刑事诉讼中的应用，确保其在助力司法实践的同时，最大限度地降低对社会公平与个人权益的侵害。在不同阶段，风险的表现形式与影响范围各异，这也要求在潜伏期及早预警，在爆发期迅速应对，在扩散期系统治理，并在平复期进行经验总结与规制优化，以确保技术发展与司法公正协同并进。

三、人工智能辅助电子数据取证的风险对法律规制的挑战

第一，由于人工智能的多变性及发展的不确定性，目前难以确定一个合适的时间点来进行规制。虽然人工智能对人类构成的威胁可能还有一段距离，但如果当AI真的发展到对人类产生威胁，那么采取任何措施可能都已经太晚了。[1]但如果过早介入，不仅容易抑制科学技术的发展，而且当技术发展走向不同的方向时，原有的框架也会失去效用。

第二，以机器学习为代表的第三次人工智能发展浪潮，成功突破了“波兰尼悖论”的限制，通过基于大数据的自我训练与自我学习过程，机器学习算法能够自动完成参数调整与模型构建。[2]这导致在人工智能应用于取证的过程中，难以进行检测和监测，其自我优化能力极易导致技术行为超出原有的技术框架，使得监管者难以预测其潜在风险和影响，规制无法落到实处。

第三，许多人工智能技术尚处于实验、萌芽或初步发展阶段，其未来的发展方向并不完全明确，应用模式缺乏完整定型，当前针对人工智能的行业标准和技术规范也还在发展中，缺乏统一且明确的标准来指导如何对人工智能进行有效规制，这就使得在规制的时机和内容上缺乏共识与统一性，大大降低了规制的效率与作用。

第四，人工智能作为一项新兴技术，在监管机构与被监管对象之间存在明显的信息不对称，这将大大影响监管机构对人工智能科技企业进行规制的有效性。[3]人工智能的创新能力和迭代速度，远远高于可能用于管理其监管工具的适应能力，以至于可能会与现有的监管框架不兼容，甚至可能推翻现有的法律和监管标准，即“监管性破坏”。[4]且传统的法律政策与监管框架的制定往往需要较长的时间周期，人工智能技术的突破可能在短时间内就超出了法律的适应范围，使得立法进程难以及时回应技术进步。

第五，人工智能的黑箱特性，即其决策过程中缺乏可解释性，导致在出现问题时，难以追溯责任、确定规制对象。风险是客观存在的，但法律无法直接管控风险，只能通过规制主体及其行为的方式，起到规制风险的目的。在电子数据取证中，人工智能设计者与使用者通常不是同一人，而在利用人工智能取证后出现风险造成损失时，由于取证过程中缺乏可解释性，导致很难明确责任主体。如果追责有误，会严重影响人工智能在司法领域的使用与发展。例如，取证主体严格按照标准进行取证，结果得到的证据出现问题，导致影响司法审判而被追责，这样会大大打击取证主体使用人工智能的积极性；同理，如因取证主体的失误造成损害，导致人工智能设计者被追责，也将造成难以挽回的损失。

第六，有些学者将算法透明原则[5]作为算法规制的首要原则，但本文认为在刑事诉讼中有待商榷：首先，人工智能在刑事诉讼中的应用，通常涉及侦查环节（如数据挖掘、证据分析）或证据审查（如电子数据真实性验证）。如果这些技术过程过于透明，可能泄露侦查方法和策略，给犯罪分子反侦查或设计对策规避侦查提供机会，影响案件侦

---

［1］Guihot M，Matthew A，Suzor N P. Nudging Robots：Innovative Solutions to Regulate Artificial Intelligence［J］. Social Science Electronic Publishing，2017.

［2］贾开. 人工智能与算法治理研究［J］. 中国行政管理，2019（1）：17-22.

［3］周学峰. 论人工智能的风险规制［J］. 比较法研究，2024（6）：48.

［4］Cortez N. Regulating Disruptive Innovation［J］. Social Science Electronic Publishing，2014（29）.

［5］算法透明原则是指一种对于算法的事前规制模式，它要求算法的设计主体或者使用主体公开和披露包括源代码在内的算法要素。沈伟伟. 算法透明原则的迷思——算法规制理论的批判［J］. 环球法律评论，2019，41（6）：23.

破效率与质量；其次，电子数据通常包含大量个人隐私、商业秘密或其他敏感信息，尤其是大数据分析难以避免涉及与案件无关的第三方数据，公开人工智能处理过程极有可能泄露当事人的隐私，与法律对隐私权的保护相违背；最后，虽然公开透明原则是人工智能应用的重要原则，但由于算法的复杂性与专业性，公众可能在理解人工智能技术过程中受到较大阻碍，导致对技术细节产生误解，进而对案件审判结果产生不满或异议。因此，过于透明反而可能引发不必要的争议，提高司法成本。[1]

四、人工智能辅助电子数据取证的风险规制策略

风险是面向未来的，但正如尼克拉斯·卢曼所言，未来发生的一切取决于现在采取的决定，而只有在决定至关重要、不采取决定便会招致损失时，人们才会谈及风险。[2]对于人工智能应用下电子数据取证可能带来的风险，是影响社会稳定性、破坏司法公正性的风险，这种风险一旦发生，带来的损失很可能是巨大且不可逆的，是个体难以承受的人身或财产损害。所以应当在风险带来损失之前采取行动，而不能等到风险造成损失再来应对，必须在问题发生之前进行有效的规制，以确保取证过程的合法性与安全性。

（一）确定人的主体地位

在刑事诉讼中，确立人的主体地位，并以理性且积极的态度面对人工智能，是规制人工智能取证风险的核心策略。要保持理性，人类的特殊性在于能够进行哲学性思考，而人工智能的滥用或将助长现代社会的反智主义。不可否认，人工智能能显著提升取证效率，但它也会因自身技术缺陷等潜在风险，产生难以解释的问题。

需要明确的是，人工智能辅助电子数据取证的核心，不在于能否替代取证人开展工作，而在于思考如何利用人工智能减轻取证人的负担。尽管近年来人工智能技术得到飞速发展，但人们目前甚至在短期未来内，都是为了一定目的创建与开发人工智能程序，它并不是万能的，也无法像人类一样思考，其核心作用仍在于辅助。英国牛津大学副教授Michael A. Osborne和Carl Benedikt Frey博士指出，涉及复杂智力活动、创造性任务和社会协作的工作，在未来几十年内仍难以被人工智能取代。在电子数据取证领域，人工智能的主要价值在于通过模式识别和大数据分析，发现取证人难以察觉的证据细节。同时，人工智能无法完全替代取证人，因为每个案件的特殊性决定了取证工作中许多问题，需要创造性思维和法律判断，这恰恰是人工智能的短板。人工智能可以快速筛查海量数据，但不擅长应对完全没有经验或未学习过的特殊情况，所以是否采纳某些数据为证据，仍需取证人根据法律规范和案件背景做出最终决策。

人工智能与取证人的相互协作，能够在多个方面实现优势互补，不仅显著提高了取证效率，也降低了人为疏忽的风险。人工智能可以利用其强大的计算能力和算法优势，快速处理影像、声音等信息，尤其是在涉及复杂模式识别时具有不可替代的优势。人工智能生成的分析报告可为取证人提供参考，帮助其发现潜在的遗漏点或异常数据，而取证人必须对这些报告进行全面的评估和解读，以确保结论的准确性和合法性。所以在现阶段，二者相辅相成才能最好地发挥效用。例如，在筛查犯罪嫌疑人的通信记录时，人工智能可以自动分类并标注可能与案件相关的内容，但这并不意味着所有标注的数据都具有证据效力。取证人需要结合案件背景，对这些数据进行筛选和分析，最终确定其证据价值。这种人机协作模式既能减轻取证人的工作负担，又能确保人工智能在法律框架内的规范应用。

在理性的同时也要保持积极心态，即在规制风险的同时也需要考虑技术的发展，虽然现阶段将人工智能置于辅助地位是最优解，但在未来的发展中若一直处于完全辅助地位，则有可能限制其更新优化。本文认为，可以借鉴SAE分级体系对人工智能

［1］司法资源的短缺是各国普遍存在的基本事实，但司法资源的增加并不必然导致正义产出随之增量，而且这种昂贵的需求还可能被不断地刺激出来，甚至可能使司法陷入恶性循环或危机。范愉．司法资源供求失衡的悖论与对策以小额诉讼为切入点［J］．法律适用，2011（3）：14–19.

［2］［德］Niklas Luhmann．风险社会学［M］．孙一洲，译，南宁：广西人民出版社，2020：34–35.

在电子数据取证中的应用进行分级管理，并针对不同级别设定相应的应用范围和责任分配机制，如表1所示。

表1 人工智能辅助电子数据取证分级体系

级别	辅助地位	功能
一级	无辅助	取证人完全依赖传统技术手段完成取证工作，不涉及人工智能参与
二级	工具级辅助	人工智能被用于单一且基础性任务，如数据去重、格式转换等，不涉及复杂的分析或推断
三级	分析级辅助	人工智能参与数据分析，提供初步的证据筛查结果，但由取证人对分析结果进行验证和调整
四级	协作级辅助	人工智能与取证人深度协作，可以将人工智能置于同一地位，视为另一取证人，共同完成复杂的多任务取证工作。虽然是深度合作，但人工智能的输出结果仍需取证人最终确认
五级	高度自主	人工智能能够在特定取证任务中实现自主操作，但必须接受取证人的实时监控和审批，每一步操作都需取证人授权
六级	完全自主	人工智能完全独立完成取证任务，从数据采集到分析报告生成均不需要人为干预。这一级别仅适用于技术完全成熟且法律框架完备的未来场景，目前尚未可行

（二）促进人工智能取证可信化发展

1. 标准化人工智能取证工具

为了保障人工智能取证工具在实际应用中的合法性和可靠性，重要措施之一就是必须进行标准化建设，人工智能取证工具应通过具有法定资格的专门机构的认证与验证，以确保其技术能力和法律合规性。同时，要明确取证工具的操作规范、操作范围和操作条件，确保取证人员严格遵守技术规范，防止滥用工具或误用算法导致数据偏差或证据无效。对人工智能取证工具的精确性、鲁棒性[1]和重复性要求进行定期测试，确保其在处理多样化电子数据时，能够保持一致性和可靠性。此外，人工智能取证工具必须具备操作记录和溯源功能，要建立“数据收集—保存—分析—更改—删除—使用”链条，确保每一项分析或决策都可被追踪和验证。例如，记录取证过程中使用的模型版本、输入数据来源及处理方法等关键信息，以便在发生争议时提供可靠的技术依据。

2. 推动质证实质化发展

面对“算法黑箱”或歧视的问题，需要先明确几点：第一，算法透明、可知，也不代表算法问题必然能被发现。单就算法漏洞而言，就包括了输入漏洞、读取漏洞、加载漏洞、执行漏洞、变量覆盖漏洞、逻辑处理漏洞和认证漏洞等；第二，即使算法完全透明，计算机工程师也不能确切预测算法与外部运行环境的交互；第三，即便算法透明，在执行算法的过程中，仍然无法保证排除第三方干预，从而影响最终结果。[2]

因此，强制要求披露其模型结构、训练数据来源及关键参数，使得算法透明公开并不是最优解。这一方式可能更适用于数据量较少、模型结构较简单且易理解的人工智能算法模型，能发挥出较大效用。但当面对数据量庞大、模型结构复杂的算法时，强制公开的效果很微弱：第一，公开成本较大，一些复杂的模型在公开时，可能需要更高的时间和金钱成本，并且公开后更易被不法分子攻击，导致相关算法无法使用；第二，即便能够公开，呈现出来的数据兼具庞大性与复杂性，当事人可能处在“数字鸿沟”中，难以理解，以至于公开并不能解决当事人的问题。而在这种情况下，应当推动质证实质化，允许对方当事人在自身能力范围内无法判断生成数据的合法性，且提供还原生成过程也难以达到效果时，要求取证人员出庭质证，而取证人员应当出庭质证，并针对法庭上的问题进行合理回答。

（三）由规范促进保护：取证主体规范化与隐私权救济的协同发展

1. 强化对取证主体的资格认证与监督

当前电子数据取证多由侦查机关或由侦查机关委托第三方执行，虽然人工智能数据的复杂性要求取证主体具备更高的专业技术能力，但也要防止滥用权力或技术操控。

首先，2021年3月，电子数据取证分析师已入选人社部、国家市场监督管理总局、国家统计局联合发布的新职业信息名录，首次电子数据取证分析师职业技能等级认定考试已在福建国投智能（美亚

[1] 鲁棒性是指在发生偶然事件时对结构造成局部损伤的条件下，结构体系具有不发生整体失效后果与局部损伤原因不成比例破坏的一种能力。吕大刚，宋鹏彦，崔双双，等. 结构鲁棒性及其评价指标[J]. 建筑结构学报，2011，32（11）：44–54.

[2] 参见沈伟伟. 算法透明原则的迷思——算法规制理论的批判[J]. 环球法律评论，2019，41（6）：31–32.

柏科）、中南财经政法大学司法鉴定中心等地陆续开展，显示出电子数据取证分析师培训体系目前已具备一定基础。本文认为，可以依托现有的电子数据取证分析师专业资格培养体系，将人工智能应用于取证纳入电子数据取证分析师的考核内容，并且要定期复核更新资格，进一步强化对取证主体的资格认定。其次，完善授权体系，规定人工智能应用于取证在侦查环节的授权范围和操作规范，防止侦查人员随意扩展使用范围。再次，设立专门监督机构，对具有资格的取证主体在案件中的行为进行全流程监督，建立“监督数据链”，并对采集过程是否遵循合法性标准进行评估。最后，可以适当加强法官技术培训，针对人工智能自主或辅助生成的数据，建议对法官提供技术培训，并不需要达到能够独立取证的水平，能够看得懂与理解即可，以此提升其对电子数据合法性的判断能力。

2. 加强隐私权的事前保护兼顾事后救济

在人工智能取证中，事前保护是隐私权保护的核心环节。在取证操作前，需要明确界定电子数据取证的法律授权范围，严格限定取证主体的权限和程序，避免超范围取证或滥用技术手段。同时需保障公众的参与权与知情权，对取证过程中涉及的隐私或商业机密内容，允许数据提供方及相关当事人参与取证过程，并赋予其充分的知情权与异议权。此外，人工智能设计者可以在模型训练和数据处理阶段，引入隐私增强技术：差分隐私（Differential Privacy），即通过假设攻击者拥有最大背景知识，确保隐私保护不依赖攻击者已知的信息量，从而解决传统隐私保护模型难以应对背景知识不确定性的缺陷。同时，差分隐私基于数学理论，对隐私保护进行严格定义和量化评估，提供了不同数据处理情况下隐私保护水平的可比性。[1] 在刑事诉讼实际应用中，取证主体可能接触大量敏感信息。一旦攻击者[2]掌握了部分背景信息，传统隐私保护措施就可能失效，而差分隐私能够通过添加噪声等方法，确保即使攻击者掌握最大背景知识，也无法准确推断目标数据。比如，在分析大量通话记录以识别嫌疑人的同时，通过差分隐私对数据进行预处理，可确保无关个体的信息不被锁定与泄露。此外，差分隐私的量化评估能力，有助于在取证过程中找到隐私保护与数据可用性之间的平衡，既满足法庭证据的需求，又维护当事人的隐私权。

而在事后救济方面，我国目前缺少对隐私权等基本权利的救济机制，即便有更完善的事前保护，仍可能因过程中存在的算法黑箱或歧视导致隐私权遭受侵害，为此，构建有效的事后救济机制是必要的。第一，建立隐私侵害申诉渠道，针对因电子数据取证导致的隐私权侵害，应设立便捷、高效的申诉机制，如在司法系统内设立专门的隐私权保护委员会，审查取证行为是否符合程序规范，并及时纠正侵权行为；第二，对于违反法律、伦理规范或因违规操作等使用人工智能生成数据导致隐私泄露的行为，应通过司法程序明确责任主体与责任追究机制，并对责任主体实施相应惩罚，要求其承担赔偿责任，尤其是在涉及大数据批量取证的案件中，更需建立清晰的责任链条。同时，也应当赋予当事人申请数据删除或采取其他补救措施的权利。例如，可引入“被遗忘权”机制，以此保障公民对个人信息的自主权。

五、结语

人工智能辅助电子数据取证，既为刑事诉讼提供了向现代化转型的广阔空间，又带来了风险生成与规制的重大挑战。本文通过揭示人工智能取证风险的生成机制与阶段特性，进一步明确了规制策略：确定人的主体地位，并以理性且积极的态度面对人工智能、促进人工智能取证可信化发展及取证主体规范化与隐私权救济的协同发展等。展望未来，希望人工智能在电子数据取证中的应用，继续秉持技术创新与法治保障协同推进的原则，沿着技术赋能与制度引领相辅相成的理性发展轨道前行，在有效规避潜在风险的基础上，全面释放人工智能的取证潜能，推动电子数据取证迈入以高效与可信共同发展为核心的新境地。

（责任编辑：王梦华）

［1］熊平，朱天清，王晓峰．差分隐私保护及其应用［J］．计算机学报，2014，37（1）：101-122.

［2］这里的“攻击者”不仅包括利用人工智能进行取证的取证人员，也包括数据处理、存储或传递过程中可能窃取或分析数据的任何内部或外部主体。

The Challenges and Strategies of AI-Assisted Electronic Data Forensics from the Perspective of Risk Society

Ye Zhenxin

School of Criminal Justice, Zhongnan University of Economics and Law, Wuhan

Abstract: Artificial intelligence (AI)-assisted electronic data forensics is not only a practical necessity but also an inevitable trend for the future. While AI provides development opportunities for the advancement of electronic data forensics, it also inevitably gives rise to a series of forensic risks. Grounded in risk society theory, this paper clarifies the categories of risks inherent in AI-assisted forensics and systematically analyzes the risk-generation mechanism from three dimensions: risk subjects, risk nodes, and risk levels. It identifies that current forensic risks manifest as both direct and indirect risks, procedural risks arising in the stages of preparation, implementation, and application, and cyclical risks characterized by the phases of latency, outbreak, diffusion, and mitigation. Based on an analysis of the challenges these risks pose to legal regulation, this paper proposes constructing an electronic data forensic system that balances technological development with legal risk control. With the primacy of human agency as its core value orientation, the system aims to achieve synergy among trustworthy technology, standardized forensics, and effective rights relief. The research seeks to provide theoretical reference for the compliant application of AI-assisted electronic data forensics in criminal proceedings and promote the positive interaction between technology and the rule of law.

Key words: Risk society; Artificial intelligence; Electronic data forensics; Risk-generation mechanism