doi:10.16060/j.cnki.issn2095-8072.2025.02.005

人工智能辅助仲裁庭审阅证据的可控性研究^{*}

武振国

(北京科技大学文法学院, 北京 100083)

摘要: 人工智能可以大幅提升仲裁庭审阅证据的效率, 但其运行过程的自主性和黑箱属性却可能导致失控风险。特别是, 传统观点将过程控制模式作为控制人工智能的核心方案, 其可行性建立在算法透明的基础之上, 其本质是一个技术问题。在现有技术条件下, 算法黑箱的加剧、算法权力的异化和算法歧视的固化很难得到克服, 当事人、仲裁庭和仲裁司法审查法院对人工智能一时难以实现过程控制。重新审视过程控制模式可以发现, 其目标在于通过过程透明实现结果透明, 确保当事人、仲裁庭和仲裁司法审查法院在知情的前提下自主决定人工智能的审阅结果是否生效。因此, 只要当事人、仲裁庭和仲裁司法审查法院对人工智能的审阅结果保持知情, 并有权对其进行解释、修改、接受和拒绝, 就可控制人工智能审阅结果产生效力的通道, 进而对人工智能实现结果控制。

关键词: 人工智能; 商事仲裁; 过程控制; 结果控制; 技术依赖

中图分类号: D997.4/DF974 **文献标识码:** A **文章编号:** 2095 - 8072(2025)02 - 0085 - 11

随着数据记录和存储技术的飞跃式发展, 在个案中提交的证据数量出现了爆炸性增长, 这为仲裁庭审阅证据带来了巨大的工作量。如果当事人将关键证据放置在存有海量电子文件的服务器中, 仲裁庭在数百万份文件中手动发现一份有效证据便无异于“大海捞针”。通过此种方式, 当事人可以在实质上达到隐瞒证据和拖延程序的双重目的, 减损仲裁程序的公平价值和效率取向。为了应对人脑的计算能力和海量文件的处理需求无法快速匹配的情况, 国内外法律界建议由人工智能辅助仲裁庭审阅证据, 以维持国际仲裁在公平和效率方面的优势(张圣翠和田洋, 2024)。

人工智能虽可大幅提升仲裁庭审阅证据的效率, 但若当事人、仲裁庭和仲裁司法审查法院对人工智能的算法原理不知情, 自然无法对人工智能实现有效的过程控制。生成式人工智能技术的算法原理可以自动向未知方向进化, 这种自主性和黑箱属性使得当事人、仲裁庭和仲裁司法审查法院对人工智能的算法原理知之甚少(苏宇, 2023)。当事人、仲裁庭和仲裁司法审查法院对算法原理了解越少, 对算法运行过程的了解也越少, 对人工智能实现过程控制的可能性也越低, 这会促使当事人、仲裁庭和仲裁司法审查法院产生更多的技术失控感和不信任感。事实上, 博闻国际律师事务所(Bryan Cave Leighton Paisner LLP)2023年度发布的国际仲裁调查报告《国际仲裁中的人工智能: 机器学的兴起》(AI in IA: The Rise of Machine Learning, 下称《报

^{*} 基金项目: 本文受2024年度中国博士后科学基金第75批面上资助项目“科学数据商业化利用的法律规制研究”(项目编号: 2024M750168)和2023年度国家社科基金青年项目“司法数据跨境调取的国际规则构建研究”(项目编号: 23CFX070)的共同资助。

告》)也证实了这一点。《报告》指出,绝大多数受访者对人工智能审阅证据持保留态度,只有15%的受访者持满意态度。硅谷仲裁与调解中心2024年发布的《在仲裁中使用人工智能指南》也指出,人工智能的自主性和黑箱属性导致当事人、仲裁庭和仲裁司法审查法院无法对人工智能实现过程控制,这是人工智能失控风险的产生原因。

尽管当事人、仲裁庭和仲裁司法审查法院现阶段无法对人工智能进行过程控制,但只要上述主体可以理解人工智能的审阅结果并有权进行解释、修改、接受或拒绝,也就可以对人工智能进行结果控制,防止人工智能生成的有害结果产生效力(赵精武,2021)。在非制度化的仲裁程序中,人工智能辅助仲裁庭审阅证据的结果需要接受当事人、仲裁庭和仲裁司法审查法院的个案审查和共同控制,这使人工智能的结果控制变得更为严密。首先,每个仲裁庭的权力均源于当事人根据单个仲裁协议所作出的授权,当事人有权对人工智能审阅证据的结果进行控制。其次,仲裁庭是仲裁权行使的主体,有权对是否解释、修改、接受或拒绝人工智能审阅证据的结果做出决定。最后,仲裁司法审查法院是仲裁权的监督者,也可以控制人工智能审阅证据的结果。如果当事人、仲裁庭和仲裁司法审查法院在知情的情况下,决定在个案中接受人工智能审阅证据的结果,就可在肯定意义上实现人工智能的可控性。否则,这三类主体在个案中拒绝接受人工智能审阅证据的结果,便是在否定意义上实现人工智能的可控性。

一、人工智能辅助仲裁庭审阅证据的可控性模式

人工智能的本质是一种技术工具。当事人、仲裁庭和仲裁司法审查法院在了解并顺应技术法则的情况下,可以结合自身的信息处理能力和行动能力选择合理的技术可控模式(曹志平和葛玉海,2015)。如果上述主体能够通过调节设备、系统和流程中的变量,确保人工智能的运行过程按预定方式进行,就可以对人工智能进行过程控制。如果上述主体只是通过监测和评估最终输出结果是否符合预期目标,以决定是否接受人工智能的审阅结果,就是对人工智能进行结果控制。可以说,过程控制和结果控制是当前确保人工智能在可控状态下辅助仲裁庭审阅证据的主要路径。

(一) 过程控制模式

过程控制模式将重点聚焦于形成结果的过程,只要过程是可控的,那么依据该过程所形成的结果也是可控的。为了达到过程控制的目标,当事人、仲裁庭和仲裁司法审查法院只有清晰了解人工智能的算法原理,方能在此基础上对人工智能的具体运行进行控制(黄瑞,2023)。因此,许多支持过程控制模式的学者都要求设置算法透明义务,促使开发者完全披露人工智能的具体技术细节(樊博和李晶晶,2024)。

如果人工智能只具有工具属性,没有自主性和黑箱属性,其算法原理就可以对开发者完全透明。在此情况下,开发者理应对人工智能的运行细节完全知情。例如,对于海量证据的审阅而言,算法通常是将相关文档定义在函数的一侧,将无关文档定义在函数的另一侧,从而对证据进行区分。至于那些处于临界位置的文档,也就是最有可能被错误分类的文档,人工智能会向资深专家寻求帮助。在资深专家的结论和算法

预测不一致的情况下，算法原理就会自动更新函数，使得输出结果更符合资深专家的判断。在理想状态下，开发者能够提前预知人工智能算法原理的迭代规律，还可以精准预测人工智能在不同时间节点所采用的数学模型和函数公式。

在实现算法透明的理想状态下，开发者有能力按照自己的预期对算法进行修改，这也是实现过程控制的关键环节。开发者的意愿往往以自然语言进行表达，算法则是以机器语言进行表达，故开发者修改算法的核心是实现自然语言和机器语言的顺畅转化。在当下的弱人工智能时代，人工智能无法理解自然语言，只能依靠开发者编写代码、标注数据和抽取要素的过程实现两种语言的低效转化。而在多数学者预测的强人工智能时代，如果人工智能能够实现算法语言和自然语言的高效转化，就有可能实现开发者对人工智能的高质量过程控制（赵泽林和程聪瑞，2024）。

假设开发者对人工智能的算法原理完全知情，又可以根据自己的预期随时修改算法，当事人、仲裁庭和仲裁司法审查法院便可以在顺应技术法则的情况下，对人工智能辅助仲裁庭审阅证据的具体细节实施过程控制（刘永谋和白英慧，2023）。例如，开发者可以根据国际仲裁案件所采用的不同证明标准，调整后台参数，确保人工智能筛选的证据与仲裁庭需要的证据基本符合。除此之外，开发者还可以通过调整代码，置入一些内容过滤规则，帮助仲裁庭提前排除非法证据。

（二）结果控制模式

结果控制模式将重点聚焦于结果，只要结果生效的通道被当事人、仲裁庭和仲裁司法审查法院控制，便可以认为结果具有可控性。在结果控制模式看来，算法过程透明并不是控制人工智能的前提，当事人、仲裁庭和仲裁司法审查法院在知情的前提下对审阅结果的控制权才是关键（张安毅，2024）。在仲裁场景下，对于人工智能辅助仲裁庭审阅证据的结果，当事人、仲裁庭和仲裁司法审查法院可以进行共同控制。

当事人意思自治是仲裁的基石，可以决定仲裁的有无、仲裁的实体内容和程序内容，这意味着人工智能审阅证据的结果需要接受当事人的控制（宋连斌和武振国，2024）。在双方当事人明确表示反对的情况下，如果仲裁庭执意采信人工智能审阅证据的结果，就会导致仲裁的程序安排与当事人合意不符，仲裁裁决便会因为违反正当程序原则而面临被撤销或不予执行的风险。即使只有一方当事人明确反对人工智能的审阅结果，仲裁庭也必须虚心听取当事人的异议意见，合理的部分也需要予以采信。

在当事人的授权下，仲裁庭作为仲裁程序的主导者，有权对人工智能审阅证据的结果进行控制。如果仲裁庭认为人工智能审阅证据的结果是恰当合理的，可以进行采信。如果仲裁庭认为人工智能审阅证据的结果是部分合理的，可在修改其中的偏差后予以采信。如果仲裁庭认为人工智能审阅证据的结果是完全不可接受的，可以不予采信。对于人工智能审阅证据结果中的模糊之处，仲裁庭还有权进行解释，以确保其符合仲裁场景下的叙事逻辑。

人工智能审阅证据的结果要得到仲裁司法审查法院的执行，需要通过技术向善的检测，避免人工智能在个案中出现失控和滥用。中国2023年发布的《全球人工智能治

理倡议》和2024年《中法关于人工智能和全球治理的联合声明》也指出，人工智能的审阅结果想要并入人类社会，就必须在具体的应用场景中与人类社会的道德伦理标准对齐。值得注意的是，人工智能本身并不是万能的，也有一定的局限性。在此基础上，技术向善并不是一种“无限的善”，即当事人、仲裁庭和仲裁司法审查法院可以通过规训人工智能以满足自身的各种愿望，因而只是一种“有限的善”，也就是在个案中符合适当的善恶观和是非观。例如，技术向善需要和国际法协会2002年发布的《关于以公共政策为由拒绝执行国际仲裁裁决的最终报告》对齐，确保人工智能审阅证据的结果不违反公共政策中不可逾越的红线（顾维遐和唐艺，2020）。这些红线不仅仅是为了保护当事人的利益，也是为了确保人工智能审阅证据的结果没有突破道德底线。例如，人工智能审阅证据的结果应该受到公共政策的严格限制，只能用于无害目的，不得将审阅结果用于歧视、人身攻击或其他不道德行为。如果人工智能审阅证据的结果不符合技术向善的基本要求，法院可以代表公众，通过仲裁司法审查程序撤销或不予执行仲裁裁决，从而拒绝有害的人工智能审阅结果进入公共领域（武振国和周健，2024）。

结果控制模式始终坚持以人为本的理念，通过仲裁当事人、仲裁庭和仲裁司法审查法院的三重控制，及时剔除有害的人工智能审阅结果，确保人工智能的引入能够对社会产生善的结果。因此，即使当事人、仲裁庭和仲裁司法审查法院无法对人工智能进行过程控制，也可以通过结果控制模式，在受控状态下对人工智能辅助仲裁庭审阅证据的行为进行安全测试。

二、人工智能辅助仲裁庭审阅证据的过程控制危机

传统观点认为，人工智能的算法原理对开发者单向透明，当事人、仲裁庭和仲裁司法审查法院可以在开发者的帮助下，根据算法原理对人工智能进行过程控制。但在目前的技术条件下，人工智能具有自主性和开发者难以勘破的黑箱属性，极有可能产生无法预判的程序和无法理解的智慧。特别是，人工智能内部具有黑箱属性的算法原理、日渐膨胀的算法权力以及算法歧视的风险，都意味着人工智能技术的运行过程难以被开发者有效控制。

（一）算法黑箱的加剧危机

当事人、仲裁庭和仲裁司法审查法院对人工智能进行过程控制的前提是人工智能的算法原理为人所知，即人工智能对于数据的处理是一个确定且可预测的过程。但是，目前人工智能技术的控制难点恰恰在于算法的运算逻辑不透明，处于未知的黑箱状态，在数据输入和结果输出之间存在着断裂带（钟晓雯，2023）。具体来说，算法的黑箱属性主要表现在三个方面。第一，人工智能算法抓取的训练数据存在黑箱问题。在仲裁场景下，为了避免不真实、不合法、关联性不强的信息干扰仲裁庭的判断，只有符合证据规则要求的证据才能成为决策依据。与之相比，人工智能的训练数

据是从各个网络站点上自动爬取的黑箱信息，其中如果混入一些低质量内容或者虚假信息，就会使人工智能对于传闻证据有着先入为主的预判，从而导致人工智能的审阅结果出现失控风险。第二，人工智能算法对训练数据的处理存在黑箱问题。在仲裁场景中，仲裁庭会按照证明力规则对不同的证据进行排序，选择证明力更高的证据予以采纳。但在自监督学习的情况下，人工智能如何抽取数据、如何标注数据以及如何设计权重均处于黑箱状态，这将导致人工智能的运算过程无法被有效控制。第三，人工智能算法的运行逻辑存在黑箱问题。在仲裁场景下，仲裁庭的裁判逻辑以司法三段论为基础，必须解决法律、事实和结论之间的一致性问题。然而，以生成式人工智能为代表的人工智能技术，其算法逻辑是一个“文字接龙”游戏，只解决人工智能根据给定文本自动续句的问题。例如，为了尽可能满足回答字数的要求，生成式人工智能会自动编造信息。如果仲裁庭要求的回答不在训练数据涵盖的范围之内，生成式人工智能也会自动拼凑信息（王静，2023）。在此情况下，人工智能能否按照司法三段论的要求作出严丝合缝的判断，则在过程上完全处于不可控的状态。

随着人工智能技术愈发复杂化和人工智能算力的日益提高，人工智能算法的黑箱状态不仅没有得到改善，反而不断加剧。越是复杂的算法，其算法模型中所涉及的数学公式越复杂，所抽取和标注的数据越多，不透明性也就越高。例如，深度神经网络算法是目前广泛使用的一种人工智能模型，其中包含海量的逻辑层和参数，理解算法内部如何做出预测变得非常困难。强化学习算法尝试通过不同的数据标识和参数设计来学习最佳策略，但在许多情况下，开发者无法准确解释算法为何选择了特定的行动。自然语言处理算法能够通过海量的数学公式来理解预训练所使用的语料库，但算法的内部工作机制并不透明，开发者无法得知它们是如何理解和生成如此逼真的文本的。在此情况下，有学者认为，算法和算命的观感并无区别，毕竟二者都有着极大的不确定性和失控风险（朱瑞，2023）。

（二）算法权力的异化危机

在人工智能辅助仲裁庭审阅证据的场景下，仲裁当事人、仲裁庭和仲裁司法审查法院能够对算法权力进行过程控制的前提，是算法权力符合已知的仲裁权运行逻辑。但在当前的技术条件下，算法权力的底层逻辑与仲裁权不同，其运行过程可能处于失控状态（单纯和叶茂，2023）。具体来说，算法权力的异化原因主要包括以下几点：第一，人工智能由仲裁场景外的平台企业进行设计，可能导致算法权力过分体现平台的意志，成为脱离当事人控制的平台权力。例如，在开发生成式人工智能的过程中，平台企业是从自身的角度出发，利用算法实现自我优待或其他目的，而不是站在当事人的角度实现仲裁程序的公平公正。第二，平台是从“普遍到具体”的角度运行宏观且抽象的算法权力，该种权力与仲裁权“从具体到具体”的运行模式完全不同。例如，人工智能往往是从群体的角度出发，根据争议类型、证据类型等指标对当事人进行分组并打上简单的标签。此种安排忽视了出庭文化、谈判文化、存证文化对案件所产生的影响，不利于保障当事人在仲裁场景中所期待的个性化、多样化和差异化待

遇。第三，平台在开发人工智能时过分追求效率价值和技术理性，可能导致算法权力偏离公众对公正价值和道德理性的合理预期。例如，生成式人工智能旨在按照算法快速生成连贯且统一的文本，至于生成内容是否符合个案公正则在所不问。作为一种只顾形式连贯性而忽视实质道德性的技术理性，生成式人工智能审阅证据的程序极有可能出乎仲裁各方的意料。

人工智能经平台创设之后，便很少依赖人力进行升级，而是在自监督学习下快速进化，最终发展成为自我主宰的智慧体。此时，算法权力的中心，从平台转移至人工智能自身，当事人、仲裁庭和仲裁司法审查法院很难对人工智能的运行过程进行控制。在此情况下，将算法权力引入仲裁程序，无疑会对仲裁权的正常运行造成冲击。一方面，仲裁庭往往选用头部平台开发的人工智能辅助审阅证据，这意味着同一算法权力既适用于A案件，又适用于B案件，成为具有普遍约束力的命令，具备了准行政权属性。该种公法特征，与仲裁权的民间性格格格不入。从根源上看，仲裁权来自当事人的授权，要求仲裁庭根据个案情况作出当事人满意的裁决。但是，算法权力是一种随着人工智能算力提升不断扩张的准行政权力，而准行政权力又有着天然的扩张倾向，要求所有仲裁案件的裁判程序在形式上保持一致。如果将算法权力与仲裁权强行绑定，算法中的准行政权力就会侵蚀民间性权力，最终压缩当事人的意思自治空间和仲裁权的灵活性空间（张凌寒，2019）。另一方面，算法权力以技术的自主学习为中心，与仲裁权的人类中心主义导向完全不同。仲裁权的本质是当事人通过选择自己信任的仲裁员所建构起来的权力，作出决策的主体始终是法律专业人士。算法权力如果演变为人工智能主导的权力类型，作出决策的主体就是技术。在人脑算力难以匹配海量证据处理需求的情况下，当事人、仲裁庭和仲裁司法审查法院只得依赖人工智能算力解决问题。这种依赖，将使得人工智能获得更多的决策权力，甚至在仲裁场景中逐渐占据主导型地位（季若望，2020）。如此一来，仲裁权原有的人类中心主义导向将会被实质性掏空，进而形成算法权力一家独大的格局。也就是说，仲裁权不再集中于人类手中，而是转移到了人工智能的掌控下。

在算法权力异化为准行政权的情况下，人机之间缺乏畅通的沟通渠道，也降低了权利制约权力运行过程的可能，增加了权力的失控风险（查云飞，2023）。具体来说，在人工智能运行的过程中，人机之间沟通不畅主要表现在以下三个方面：第一，人机之间沟通机会的减少。在传统模式中，当事人和仲裁庭以面对面的形式进行沟通，产生异议时可以通过当场陈述、申辩的方式消除分歧、纠正偏差、增强共识。但在人工智能时代，仲裁庭向人工智能输入某类材料之后，人工智能便可即时输出回答。这意味着，当事人几乎没有机会和人工智能进行沟通，自然也被排除在人工智能的运行过程之外，无法通过及时行使权利纠正算法的误解和误判。第二，人机之间沟通质量的降低。当事人、仲裁庭和仲裁司法审查法院使用的是自然语言，人工智能使用的是算法语言，二者的转换质量决定了沟通质量。在现有条件下，当事人的陈述和申辩很难转化为算法语言，算法对人类语言的记录只能起到留言箱的功能。即使未来人工智能进化到强人工智能时代，自然语言和算法语言可以进行高质量转换，自然语

言的歧义性也会阻碍沟通质量的实现。例如，西方文化注重“言传”，属于典型的低语境文化，信息常常被表达得非常明确。而东方文化则注重“意会”，属于典型的高语境文化，绝大部分信息往往内化在信息发出者所处的环境中，以至于相互传递的言语信息，往往模糊而含蓄。东方当事人常常认为，通过言语直接提出要求或者拒绝别人会伤及面子，因此具体用语应该间接而委婉，尽量在体面的情况下使对方知难而退（王钢和曾加，2011）。因此，东方当事人在谈判过程中提出的核心请求被对方明确拒绝后，如果表示“下次有机会见面还可谈判”，往往是指再也不见，而不是存在缔约意向。如果人工智能将东方当事人的自然语言转化为算法语言，显然就与当事人的真实意愿不符。第三，人机之间沟通成本的提高。由于当事人和人工智能之间沟通机会和沟通质量的下降，当事人只能在人工智能作出错误判断后才能申请纠错程序。至于人工智能作出的出乎当事人意料但又不存在明显错误的决策程序，往往会处于当事人权利制约的范围之外。

（三）算法歧视的固化危机

在人工智能辅助仲裁庭审阅证据的场景中，当事人、仲裁庭和仲裁司法审查法院能够按照正当程序原则的要求，消除人工智能的歧视风险，也是实现人工智能过程可控的重要一环。此前有观点认为，人工智能具备客观性优势，不可能存在歧视问题。这种观点有一定的吸引力，但最近的研究表明，预设人工智能不存在歧视风险是一种盲目的乐观。特别是，分析人工智能背后的算法原理，便可得出人工智能也有可能感染认知偏见（梅傲，2024）。在人工智能研究的早期阶段，计算机科学家试图通过理解和复制人脑的认知过程来开发人工智能程序。但遗憾的是，由于科学家对人脑的研究不够深入，人工智能无法完成复杂的逻辑推理，很难以人脑的方式完成学习任务。随着计算机处理速度的提高和数据存储成本的降低，以大数据为支撑的机器学习异军突起。所谓机器学习，并不是模仿人脑的认知过程，而是从学习的功能意义出发，使计算机能够从现有数据中发现规律、建构模型，并随着时间推移和经验积累修改算法。鉴于机器学习的核心是从现有的人类数据中推断出隐藏的因素或模式，只要现有数据中存在人类的认知偏见，这种偏见便会感染人工智能。例如，微软公司开发的人工智能在与人类用户互动后，迅速学会了种族歧视言论。另一组研究人员利用人工智能技术预测欧洲人权法院的裁判过程，发现人工智能重点标准的关键词中包括大量涉嫌地域歧视的用语。这些歧视性信息即使被人为遮盖，算法仍然可以根据代理变量自动还原出来。

作为一种致力于输出公平正义的争议解决机制，仲裁制度不能容忍算法的决策过程存在歧视风险。在传统的仲裁程序中，如果仲裁员对当事人的歧视足以使知情理性第三人合理怀疑其公正性，仲裁员就应当回避。例如，2020年12月22日瑞士联邦最高法院裁定撤销国际体育仲裁院对中国游泳名将孙杨的禁赛处罚，原因是首席仲裁员曾在推特上发表歧视性言论。该仲裁员是一名动物保护主义者，他在推文中严厉指责食用狗肉的中国公民，同时将中国人描述为残忍的“黄脸”。由于“黄脸”一词涉

及种族主义歧视，且与案情没有关联，可以作为客观上质疑仲裁员公正性的合理理由（冯硕，2023）。如果类似歧视存在于人工智能的算法中，并随着算法权力的普遍适用而放大到所有案件，可能导致涉人工智能仲裁裁决均被撤销的后果。对当事人而言，人工智能辅助仲裁庭审阅证据的过程本就是一个不容易解释的算法黑箱，如果叠加算法歧视的风险，必将引发大面积恐慌。

三、人工智能辅助仲裁庭审阅证据的结果控制进路

人工智能具有自主性和黑箱属性的运算过程是当事人、仲裁庭和仲裁司法审查法院的信息盲区，难以实施过程控制。有鉴于此，当事人、仲裁庭和仲裁司法审查法院对人工智能的控制只能从过程阶段后移至结果阶段，确保仲裁当事人、仲裁庭和法院有权接受、拒绝或修改人工智能审阅证据的结果。

（一）当事人对人工智能审阅证据的结果控制

虽然当事人、仲裁庭和仲裁司法审查法院对人工智能的运行过程无法做到完全知情，但当事人在知悉人工智能审阅结果的情况下，通过控制人工智能审阅结果的生效渠道，仍然可以对人工智能审阅证据的结果进行控制。为了达到结果控制的目标，当事人需要充分知悉人工智能的审阅结果，并有权对人工智能的审阅结果提出异议，同时也可以在一定条件下处分人工智能的审阅结果（刘金松，2023）。

第一，仲裁当事人有权对人工智能的审阅结果知情。在人工智能辅助仲裁庭审阅证据的过程中，仲裁当事人对人工智能进行结果控制的前提是其知悉人工智能审阅证据的结果。在这里，仲裁当事人对人工智能的审阅结果知情，不仅要求当事人能够获得人工智能的审阅结果，还要求当事人能够以普通人可读的形式获得该结果。如果仲裁庭没有告知人工智能审阅证据的结果，或者提供了当事人不能理解的代码或者机器语言，都会影响当事人的知情权，损害当事人陈述申辩的正当权利。在此情况下，仲裁程序的进行将违反正当程序原则，这足以导致仲裁裁决被撤销或者不予执行。

第二，当事人有权对人工智能的审阅结果提出异议。在人工智能辅助仲裁庭审阅证据的过程中，不能将证据调查结果完全交由技术主宰，当事人始终享有质疑人工智能辅助审阅证据的权利，这一权利的有效保障应当是人工复核机制。一方面，当事人有权申请仲裁庭对人工智能审阅证据的结果进行人工复核，即将人工复核的启动权归于当事人。另一方面，当事人如果发现人工智能审阅证据的结果存在明显偏差，可请求仲裁庭进行人工修改。当然，仲裁庭逐条对海量证据进行人工复核并不现实，可进行抽样复核。例如，人工智能审阅证据的结果表明，相关性在50%以下的电子文件占比高达80%。则仲裁庭可以认为相关性在50%以下的文件极有可能与事实无关，无需进行人工审核，而只需对相关性在50%以上的文件进行抽样复核。

第三，当事人在一定条件下可以处分人工智能的审阅结果。当事人意思自治是仲裁的基石，当事人可以通过合意对人工智能的审阅结果进行处分，这在本质上也是

对当事人权利的处分。如果当事人均接受人工智能的审阅结果，则人工智能辅助审阅证据可以成为处分当事人权利的依据。如果当事人均拒绝人工智能的审阅结果，则人工智能辅助审阅证据无法成为处分当事人权利的依据。除此之外，如果某项人工智能应用软件风评不佳，当事人可以事前作出约定，拒绝接受该人工智能审阅证据的结果（武振国，2024）。

（二）仲裁庭对人工智能审阅证据的结果控制

仲裁庭是仲裁程序的主导者，对人工智能审阅证据的结果知情是应有之义。虽然仲裁庭在现有的技术条件下无法对人工智能进行过程控制，但其可以决定人工智能审阅证据的结果是否发生法律效力（刘文杰，2024）。为了保证仲裁庭对人工智能实现结果控制，仲裁庭有权对人工智能审查证据的结果进行解释，有权审查人工智能审阅证据的结果，有权对人工智能审阅证据的结果进行修改或表示接受和拒绝。

第一，仲裁庭有权对人工智能的审阅结果进行解释。在人工智能辅助仲裁庭审阅证据的场景中，一个可能面临的问题是，对人工智能审阅结果内容的理解，是以人工智能开发者的判断为准还是以仲裁庭的判断为准？如果以人工智能开发者的理解为准，在仲裁庭反对的情况下，人工智能审阅结果就不可能被仲裁庭采信，更不可能借助仲裁裁决产生效力。这意味着，如果以仲裁庭的判断为准，实际上更符合仲裁权的运行逻辑。一方面，《中华人民共和国仲裁法》（2017年修正版）第54条要求仲裁庭在作出裁决时说明理由。也就是说，如何理解人工智能审阅证据的结果，是仲裁庭而非人工智能开发者的工作。另一方面，现有的排除合理怀疑、高度盖然性等证明标准，均是从仲裁庭的角度进行判断。也就是说，人工智能审阅结果的最终责任人，仍然是仲裁庭。由此可见，对人工智能审阅结果内容的解释权，应当归仲裁庭享有。

第二，仲裁庭有权审查人工智能审阅证据的结果。对于人工智能审阅证据的结果，仲裁庭需要对证据资格和证明力进行审查，以确保其符合相关程序法的要求。例如，仲裁庭可以对证据的取得方式进行审查，以防止明显不合法的证据流入后续程序，或者防止人工智能将其在网络上随机抓取的证据列入审阅结果中。除此之外，仲裁庭还需要根据所在法域的证明标准，对证明力进行审查，以防止人工智能的后台程序在感染偏见风险之后，对证据的证明力进行歧视性打分。

第三，仲裁庭有权接受、拒绝和修改人工智能的审阅结果。经过当事人的质疑和仲裁庭的审查之后，对于人工智能的审阅结果，仲裁庭可以有三种处理方式。如果人工智能的审阅结果符合仲裁庭和当事人预期的，仲裁庭当然可以予以接受和采信。如果人工智能的审阅结果完全不符合仲裁庭和当事人的预期的，仲裁庭需要拒绝采信。如果人工智能的审阅结果部分符合仲裁庭和当事人的预期，且符合部分和不符部分可分的，仲裁庭可以在修改不符内容的基础上，对其进行有条件接受。

（三）法院对人工智能审阅证据的结果控制

对人工智能审阅证据的仲裁司法审查程序位于裁决作出之后，法院只能对人工智

能进行结果控制，无法进行过程控制。在这一过程中，法院确定人工智能的审阅结果能否并入特定法域的关键，是人工智能的审阅结果能否实现技术向善。为此，法院需要根据留存的记录，对人工智能的审阅结果进行审核（冯硕，2024）。

首先，法院对人工智能审阅结果进行司法审查的前提是相关记录被全面留存，如此方能确保法院的知情权。为了方便法院监督，仲裁庭需要对留存记录的具体细节和完整程序进行详细规定，以确保后续程序能够以可发现、可访问、可互操作、可重用的方式读取留存记录。其次，法院需要根据留存记录，对人工智能的技术伦理进行审核。人工智能审阅证据的结果想要经由仲裁裁决获得法院的承认与执行，至少需要证明其不违反技术向善，对社会无害。为此目的，法院有必要对人工智能的训练数据、人工智能的算法原理以及人工智能的审阅结果进行审核，以确保人工智能在作出决策的过程中始终保持中立，并没有对任何一方当事人做出足以影响案件公正裁决的歧视性对待。最后，如果法院根据留存记录证实人工智能的技术伦理中确实存在算法歧视，就需要对人工智能审阅证据的结果进行评价。如果算法歧视的内容与仲裁庭公正判案有极强的相关性，那么人工智能的审阅结果当然不需要予以维持，也就是需要撤销或不予执行仲裁裁决。如果二者关系不大，也就是算法歧视的内容与仲裁庭公正判案无关的，仲裁裁决可以继续保持效力。

四、结语

人工智能内部的运行过程难以被当事人、仲裁庭和仲裁司法审查法院所准确探知，这使过程控制模式在现阶段难以实施，由此放大了人工智能的失控感风险。在此情况下，算法黑箱的延续、算法权力的异化和算法歧视的风险，都对人工智能的过程控制提出了严峻挑战。为了应对上述问题，当事人、仲裁庭和仲裁司法审查法院必须坚持自身的主体性地位，认识到人机控制关系的主导者始终是人，人工智能辅助仲裁庭审阅证据的权力来源于人，人工智能的审阅结果顺利获得法院执行必须以人为尺度。当事人、仲裁庭和仲裁司法审查法院通过确保对于人工智能审阅结果的知情权，维护自身在人工智能面前的理解、思考、判断和调整能力，控制人工智能审阅结果的生效渠道，便可以对人工智能实行结果控制。对比来看，结果控制模式不仅能够维护当事人、仲裁庭和仲裁司法审查法院在人工智能面前的主体性地位，还可以对算法权力形成有效制衡，从根本上实现人工智能辅助仲裁庭审阅证据的可控性。

参考文献

- [1] 曹志平,葛玉海.关于“控制技术”的三种争论[J].科学技术哲学研究,2015(3):73-77.
- [2] 查云飞.算法的行政法属性及其规范[J].法制与社会发展,2023(6):168-185.
- [3] 单纯,叶茂.智慧司法异化风险的反思与规制[J].江西社会科学,2023(3):160-168.
- [4] 樊博,李晶晶.算法透明何以提升公众的政府信任?[J].公共行政评论,2024(1):4-24.
- [5] 冯硕.《中华人民共和国仲裁法》修订的涉外法治之维[J].上海大学学报(社会科学版),2024(5):13-24.
- [6] 冯硕.国际体育仲裁中的规则之治与权利保障[J].华东政法大学学报,2023(5):166-179.
- [7] 顾维遐,唐艺.中国的“一带一路”发展和亚洲的国际商事仲裁融合[J].国际法研究,2020(1):105-128.
- [8] 黄瑞.立信持疑:法官视角下司法信任的生产逻辑[J].法学家,2023(2):74-88.

- [9] 季若望.法律的再生:人工智能时代的凤凰涅槃[J].上海对外经贸大学学报,2020(4):111-124.
- [10] 刘金松.数字时代刑事正当程序的重构:一种技术性程序正义理论[J].华中科技大学学报(社会科学版),2023(2):18-29.
- [11] 刘文杰.何以透明,以何透明:人工智能法透明度规则之构建[J].比较法研究,2024(2):120-134.
- [12] 刘永谋,白英慧.坚持人工智能(AI)发展的有限主义进路[J].科学·经济·社会,2023(6):58-71.
- [13] 梅傲.积极伦理观下算法歧视治理模式的革新[J].政治与法律,2024(2):113-126.
- [14] 宋连斌,武振国.中国仲裁机构的功能定位——以优化营商环境的对外吸引力为视角[J].上海对外经贸大学学报,2024(1):68-78.
- [15] 苏宇.数字时代的技术性正当程序:理论检视与制度构建[J].法学研究,2023(1):91-107.
- [16] 王钢,曾加.调解中的当事人心理倾向探析[J].西北大学学报(哲学社会科学版),2011(4):102-106.
- [17] 王静.生成式人工智能技术嵌入智慧司法的伦理风险及其法治应对[J].上海政法学院学报(法治论丛),2023(4):43-62.
- [18] 武振国,周健. Generative AI生成司法决策的可靠性困境及其应对[J].天津大学学报(社会科学版),2024(6):512-520.
- [19] 武振国.欧盟人工智能的实验主义治理路径及中国借鉴[J].西北大学学报(哲学社会科学版),2024(6):153-164.
- [20] 张安毅.算法自动决策下人类主体性危机的立法应对[J].理论与改革,2024(4):43-52.
- [21] 张凌寒.算法权力的兴起、异化及法律规制[J].法商研究,2019(4):63-75.
- [22] 张圣翠,田洋.人工智能在商事仲裁中的应用:价值、困境与路径[J].上海财经大学学报,2024(2):122-136.
- [23] 赵精武.从过程控制到结果归责:算法透明义务与算法问责机制的分野[J].北大法律评论,2021(2):18-31.
- [24] 赵泽林,程聪瑞.ChatGPT、人类心灵与人工心灵——科辛斯基ChatGPT实验的考察与反思[J].江汉论坛,2024(7):99-105.
- [25] 钟晓雯.从算法“黑箱”走向算法透明:基于“硬法—软法”的二元法治治理模式[J].中国海商法研究,2023(4):53-62.
- [26] 朱瑞.论算法行政的技术性正当程序[J].财经法学,2023(4):103-117.

【作者简介】 武振国: 北京科技大学文法学院讲师, 律师联讯产品专家委员会委员, 上海政法学院国际仲裁学院研究员, 法学博士。研究方向: 国际私法、仲裁法。

A Study of the Controllability of Artificial Intelligence Assisted Arbitral Tribunal in Reviewing Evidence

WU Zhen-guo

(School of Humanities and Social Sciences, University of Science and Technology Beijing, Beijing 100083, China)

Abstract: Artificial intelligence(AI) can significantly enhance the efficiency of a tribunal's review of evidence, but its autonomy and "black box" nature may lead to risks of loss of control. In particular, the traditional view places the process control model as the core solution to regulating AI, and its feasibility is based on algorithm transparency, which is essentially a technical issue. Under current technological conditions, the worsening of the algorithmic black box, the alienation of algorithmic power, and the entrenchment of algorithmic bias are difficult to overcome, making it hard for the parties, the tribunal, and the judicial review courts to achieve process control over AI. A re-examination of the process control model reveals that its goal is to achieve result transparency through process transparency, ensuring that the parties, the tribunal, and the judicial review courts can autonomously decide whether the results of AI's review should be valid, based on being fully informed. Therefore, as long as the parties, the tribunal, and the judicial review courts are kept informed of the AI's review results and have the right to interpret, modify, accept, or reject them, they can control the channels through which the AI's review results take effect, thus achieving control over the final outcomes of AI.

Keywords: AI; commercial arbitration; process control; result control; technological dependence

(责任编辑: 马莹)