 DOI: https://doi.org/10.12345/bdai.v7i1.35265

A Jurisprudential Study on the Interactive Relationship between Corpus and Artificial Intelligence

Wanning Kong

Xizang Minzu University, Xianyang, Shaanxi, 712082, China

Abstract

In the development of artificial intelligence, corpora provide data support. However, in the process of providing data support, corpora involve legal issues related to data privacy protection, data privacy, and intellectual property ownership. The provider of the corpus, the data controller, must also comply with relevant legal regulations, effectively protect the basic rights and interests of data subjects, and avoid infringement of their privacy rights. In addition, in the development of artificial intelligence, it is crucial to clarify legal issues related to the ownership of intellectual property rights of creations made by artificial intelligence. In the development of artificial intelligence and corpora, the confirmation of relevant legal liability issues cannot be ignored. In this situation, the creators of corpora and artificial intelligence need to determine the responsibility issues in different environments based on their different relationships. At the same time, flexible legal rules should be created according to the specific realities, achieving harmony and unity between technological development and risk control, in order to provide ideas and methods for the utilization of corpora and artificial intelligence.

Keywords

corpus; artificial intelligence; legal regulation; ownership of intellectual property;

语料库与人工智能的互动关系法理研究

孔婉宁

西藏民族大学，中国·陕西 咸阳 712082

摘　要

在人工智能的发展过程中，语料库为其提供数据支持。但是语料库在提供数据支持的过程中，其背后涉及到了数据隐私保护、数据隐私和知识产权所有法律层面的相关问题。语料库的提供者——数据控制者也需遵守相关法律规定，切实保障数据主体的基本权益，并且避免其隐私权被侵犯。此外，在人工智能的发展过程中，关于人工智能的创造物的知识产权归属相关法律问题的明确是极为关键的。人工智能与语料库发展过程中，相关法律责任问题的确认同样不可忽视，在这种情形下，语料库的创造者和人工智能的创造者需要根据其关系的不同确定不同环境下的责任承担问题，与此同时，应该根据现实的具体情况创建灵活的相关法律规则，实现技术的发展与风险控制的和谐统一，以期为语料库以及人工智能的利用提供思路方法。

关键词

语料库；人工智能；法律规制；知识产权归属

1 引言

1.1 研究背景与意义

互联网 + 人工智能和互联网 + 大数据为我国目前互联网、信息技术的进步奠定了良好的发展基础。自然语言处理需要语料库来帮助人工智能中的“语料积累”，如自然语言生成，就要通过语料库建立来避免常识性错误，从而保证生成的文本信息的准确性及可阅读性。同时，人工智能离不开数据量以及高质量的数据与语料，利用人工智能来深度学习也离不开高质量的语料积累与学习。人工智能发展离不开数据的积累，尤其离不开具有针对性的语料库建设，如以人工智能法律为基础的并行学习法，依靠大数据的预训练有助于解决在法律领域的数据积累问题。

以法理为考察对象，探讨人工智能的法律人格及其对社会公共关系的危害，需要承认的是，作为新技术，人工智能绝对无法具备刑法学意义上的犯罪主体资格，但是在其开发、制造、运用的过程中给与公共安全带来危害的，法律应有所规定，在必要的情形下应当给予适当的刑事保护，譬如利用人工智能收集语料库的信息是敏感信息，涉及到侵犯公民权利与隐私的问题，这就需要法律严格限制其利用，并运用法律手段来予以制止，使其所收集的数据必须出于正当理由，其标注结果是否合理公正等等；而语料库的质量对人工智能的运用效果具有极为重要的意义，从人工智能中抽取的

【作者简介】孔婉宁（2002–），女，满族，硕士，从事法律（法学）研究。

数据如果没有足够的合法性和正当性，其标注内容没有正确性、伦理上的合法性不完整等等都会影响到人工智能输出内容的正常运行。

1.2 国内外研究现状

国内外语料库与人机关系相关研究现状呈多元探索研究框架。从人机关系的技术应用角度，部分学者通过语料库建设、语料库分析来构建服务特定情境的智能系统，如以 SELL 语料库为基础的 AI-English-speakingtraining 系统，将语言数据和 ML 技术整合起来进行语言能力的动态评价与回应 [1]。从社会互动的角度，部分学者运用民族方法论（EM/CA）方法研究人机关系，指出了 AI 在任务中心型互动、实验型情景和日常交往的境遇语境特征。针对人机互动的情境即时性问题，有研究提出评分系统的动态框架，将人机互动中实时出现的提示器中的令牌数量、质量标准予以衡量，为使用者给予即时回应，以达到优化即时反应。

哲学上，语料库语言学与人工智能的关联问题引发的是从哲学视角对人机关联的理性思考。既有多篇文献通过对图灵测试和中文房间论证的讨论，指出了规则 / 形式的表述及其语义的关系，以及社会语境等因素在交际双方意义形成中的重要作用，指出用技术与社会规则中权利义务的合法性要结合基于语料库的人机交互社会基础及其技术的有限性进行辩证分析的观点 [2]。然而，从研究实践来看，由于大部分成果侧重于如何“拿来”以及怎样“运用”的问题，而缺乏语料库语言学与人工智能的关联视野下法律上语义权利与义务的关联性、“大数据杀熟”“人脸识别”类语料库语言学信息技术下数据伦理与法律利益的冲突问题、“无人驾驶”“互联网 +”背景下 AI 中机器学习与技术学习能力与作为被学对象的语料库的解释力相互性与不确定性这一原理下语料库学习行为、权利和义务的机制解释等相关法理问题的研究，即尚未有研究关注如何将图灵测试下的技术风险、中文房间论证下的技术局限转化成具有可操作意义的法理机制并加以落实的相关机制探讨，譬如人们在识别出 AI 误用之后，如何将其风险利用后果依据风险评估矩阵转化成法律上的权利和义务以明晰法律上的权责与法律责任问题尚未探讨。因此，从法律视角考虑语料库语言学中语义权利和义务的关联性需要基于语料库语言学技术和人工智能运行逻辑的特点与规律，而不仅仅仅仅是指 AI 与传统的人工知识和智慧以及传统社会中人际或自然人之间的数据、知识等载体的权利义务。

2 法律视角下的语料库与人工智能关系

2.1 数据保护与隐私

数字经济时代下，随着语料库利用人工智能的深入程度不断提高，语料库数据保护与隐私法律问题是限制人工智能发展的核心因素之一。语料库中数据（包括文本、图像、音频、视频等）的数据量和复杂性均不断提高，其中涉及的个人身份标识、位置信息、行为轨迹数据和商业信息、涉密信息存在被不法分子窃取和利用的风险，不仅严重危及公民个人隐私权，也将导致企业商业秘密被泄露、破坏、市场失衡等严重后果，而影响社会公共和网络生态的有序。因此，语料库的设计过程中要高度重视《个人信息保护法》《数据安全法》的法律规定，在技术的开发和运行过程中严格遵照执行数据保护原则。

权利保护是法律的核心构成部分。个人作为数据的生产者依据法律规定具有对其个人信息所享有的知情权、同意权、查询权、更正权和删除权等权利。要保障数据主体权利在语料库建设方面可以通过技术手段和制度建设两个方面来达到。知情权就是数据处理者在数据采集的过程中应当将数据的用途、数据收集范围、数据保护手段等向数据主体进行公示告知；同意权是应当通过符合法律规定的明示同意来保障实现的，尤其是在与敏感数据处理相关的部分，需要数据主体单方面授权同意；而访问权和更正权的保障就是语料库系统应当能方便用户对系统中的语料进行查询与更正，以方便数据主体及时查询及更正其个人信息；而删除权的实现则是要求语料库系统能够具备删除应答，以便在数据处理目的实现和主体撤回同意时在一定时间内能够对数据进行完全的删除 [1]。

数据控制者作为数据处理活动的发起和主导者，必须承担起法律义务和社会责任。一方面从技术上运用数据脱敏、密码存储、权限控制等技术措施实现对数据存储、传输、处理等环节的安全防护，避免数据被未授权者访问；另一方面从制度上建立贯穿数据全生命周期的安全管理制度体系，如数据分类分级制度、数据使用流程控制、数据安全风险管控体系，以及从每个数据处理活动中实现合规定位的数据安全管理制度体系，如隐私影响评估、数据安全官制度等。另外，数据控制者在跨国语料库项目上不仅要满足来源地数据主权立法要求，还需满足目标语境数据隐私保护相关法律条款要求，否则造成管辖交叉的法律冲突，就可能涉及违法数据合规风险。平衡利用与保护的协调之道在于技术与法律协同创新，人工智能技术的应用为保护数据安全提供工具与路径选择，如联邦学习技术是在不转移个人信息原始数据基础上进行模型训练的一种有效机制，差分隐私算法也能部分实现去标识化处理后使个人数据被识别的风险概率显著降低，但前提必须与法律法规要求相结合，如将数据“匿名化”处理必须符合《个人信息保护法》中有关“去标识化”的法律规定。与此同时，数据控制者必须建立常态化的动态合规工作机制，不断评估技术使用和遵守的数据保护法律要求之间的相关性与兼容性，做到既能以科技创新驱动法律规范发展和改革，又能确保科技创新不触及法律红线 [2]。

由此可见，语料库建设和人工智能发展背景下关于数据保护与隐私权的问题，归根结底是一个数字时代的权利保护与技术发展的平衡问题。

2.2 知识产权归属

在法律层面来看语料库与 AI 的法律关系，知识产权归属属于语料库和 AI 法律关系所涉及的重要方面之一，既有智力成果的保护，也有新技术条件下的法律问题。语料库属于法律、语言学以及计算机等领域的知识库，语料库的形成也是耗费大量人力、物力、财力的过程。所以语料库在法理层面上的知识产权归属以语料库建造的“功利主义”观点为主，遵循“谁付出、谁所有”的原理。语料库建设的知识产权归属以建造者在语料收集、标注、分类中的事实投入为准，归其所有。

为此，需要在解决上述问题的技术层面和法律层面之外，构建出灵活且适用的人工智能时代规则。首先，在现有的知识产权规范下细化人工智能生成物的权利归属规则，如在语料的生成过程中发挥主导作用的人类作为生成物的权利主体，并在有关法律规范下遵守“实质性贡献”原则，或建立“数据供作者报酬分享条款”，在尊重语料库和语料构建者的初始权利下，建立技术提供者和数据提供者分享利益的规则。

2.3 法律责任界定

语料库与人工智能技术的结合应用也是法律规制内容的重要方面，对于技术的应用，由于不确定因素与复杂场景下技术本身的难以厘清，导致责任主体确定、责任要素的构成以及应负责任的形态都具有不同于一般法律关系的特殊性。从法律层面来看，技术中的数据控制者与技术的设计者作为技术中的重要主体，往往也会因为不同的使用环境负不同范畴和强度的责任，责任的内容与程度依赖于所实施技术行为的性质、损害结果的发生可预测性和行为人主观过错的考量。语料库在数据安全管理这一角度，数据控制者的法律责任体现在语料库的运行过程中对数据造成的泄露所引发的法律责任的界定。依据《个人信息保护法》《数据安全法》，语料库所处的法律关系中的核心地位是“数据控制者”。“若因未能履行依照本法和其他有关法律、行政法规规定以及合同约定的义务导致重要敏感数据泄露的，通常应首先依据意思自治原则将数据安全事件定性为损害事实，并在此基础上承担责任，通过一定的方法与手段完成受害者的权利救济，如经济赔偿以及消除数据负面影响”；“如果其违反数据安全保护义务的行为达到刑事立案标准则构成犯罪，此时需要追究相关人员的刑事责任”；当然，行政责任也不以损害发生为条件，行政监管部门仍然可以根据没有做好数据安全保护的行为实行相应的行政处罚。

因此，数据控制者应构建技术上的加密存储、分级访问权限以及操作日志留存制度及管理上的员工培训、合规审核机制相结合的全方位多层次立体化防控模式，以规避法律责任风险。

滥用：算法应用带来的新责任形式。在算法模型由于训练数据不精准、训练方式不妥当、应用领域不正确造成的决策失误或侵害事件出现的情况下，算法责任主体的认定要结合技术开发者、应用者、用户等多个环节逐个追溯。开发者在技术本身安全可靠的前提下要承担初始责任，在因为算法黑箱而产生的歧视或重大决策结果错误下要承担侵权赔偿责任；用户在明知其技术产品缺陷却仍然进行使用的情况下要承担过失侵权责任 [3]。随着技术自主性不断提升，在具体处理开发者责任与使用者责任时，如何界定各自的边界成为实践难点 [3]。

3 结论

本研究通过系统考察语料库与人工智能技术的互动发展路径，揭示了二者在技术创新与法律规制层面的双向作用机制。

本文的学术贡献在于对于技术进步与法律规制关系的认识提升：一是对于技术中立原则的修正，指出语料库的质量及算法设置存在必然的价值层面考量；二是对于弹性法规制体系的推崇，将技术创新与风险防范追求共同价值观引导下的人文关怀式原则立法。在现实应用上，对于语料库中的伦理监管机制设计、人工智能中法律责任承担机制选择，尤其是跨语种、跨地域、医疗健康类语料存在风险进行追责提供有益经验借鉴，同时就语料库在互联网背景下法律性质的界定、算法可解释性与语料可解释性的协调规则、人工智能的伦理委员会设立的职责边界问题，以达到技术进步与法律规制之间的持久共生。

参考文献

[1] 田烨.面向SELL语料库的AI虚拟英语教育训练系统研究[J].微型电脑应用.2020,36(12):42-44.

[2] 陈伟,管灵慧.大模型智能语言交互困境的存在论反思-从“此在”到具身智能的人机“共在”.[J].外语研究.2025,42(06):4-6.

[3] 丁国峰,寿晓明.经营者滥用人工智能生成合成信息不正当竞争的技术治理路径[J].福建论坛.2025(05).