专家知识数据化与文科智能

邱泽奇　李昊林　张平文等*

摘　要：数智技术正在触发知识生产范式变革，文科智能是人文社会科学发展的未来方向。高质量数据是发展文科智能的前提。把人类积累的知识数据化是获取高质量数据的路径之一。其中，专家知识是一类不断迭代、浓缩、有着专门结构的知识，也是更具创新价值的知识。法学领域的专家知识在文科中具有典型性。以法学领域为例，专家知识的数据化可以采取“三步走”路径，即框定专家知识范围、重构专家知识结构以及设计专家知识转化技术路径等。“三步走”的专家知识数据化路径具有可推广性，既有助于夯实文科智能的数据基础，也可为推动文科知识生产范式变革迈出坚实的一步。

关键词：专家知识数据化　文科智能　数据标注　法学领域

作者邱泽奇，北京大学中国社会与发展研究中心教授（北京　100871）；李昊林，对外经济贸易大学法学院助理教授（北京　100029）；张平文，武汉大学人工智能学院教授（武汉　430072）。

一、文科智能及其发展格局

随着数智技术的快速发展，人类知识生产范式正在经历一场智能变革。在这一进程中，自然科学凭借其与人工智能的深度融合，已然走在前列。自科学智能（AI for Science）提出以来，人工智能在自然科学领域的创新和运用已成为世界顶尖高校和头部企业追逐的焦点。①《自然》杂志曾发文探讨智能技术对科学发现的深远影响，2025年又专文讨论科研可用的生成式预训练大语言模型。② 与科学智能相对应

* 第四作者乔天宇，北京大学社会学系助理教授（北京　100871）。

① Stefan Hajkowicz et al.，“Artificial Intelligence Adoption in the Physical Sciences，Natural Sciences，Life Sciences，Social Sciences and the Arts and Humanities：A Bibliometric Analysis of Research Publications from 1960-2021，” *Technology in Society*，vol.74，2023，102260.

② Elizabeth Gibney，“What Are the Best AI Tools for Research? *Nature*’s Guide，” February 17，2025，https://www.nature.com/articles/d41586-025-00437-0，March 12，2025.

中国知网　https://www.cnki.net

的文科智能（AI for Humanities and Social Science，AI for HSS）同样引起了国内外学者的广泛关注和讨论。①

对知识生产的智能变革而言，无论是自然科学还是人文社会科学，都面临一个共同任务：把方便人类阅读、理解、学习和探究的知识表达转化为方便机器阅读、理解、学习和探究的知识表达，以便在知识生产中实现人机协同乃至人机互生。OpenAI的前合伙人之一卡帕西（Andrej Karpathy）甚至认为，应将99.9%的优质资源聚焦于使内容更容易被大型语言模型消化，而非仅仅迎合人类读者。对自然科学而言，绝大多数学科的知识表达均以数学表达式为基础，以数值数据为素材形态，表达易于转化。以此为基础，科学智能的关注点已然从数据化表达转移到运用人工智能协同知识生产。② 然而，文科知识表达的数据化转化要复杂得多，根本原因在于文科知识表达的大多是观念、观点、理论、意义、作品（如艺术作品）等，即便使用了数据也多为非数值数据。只有将其转化为机器易于消化的数据，才有可能讨论与人工智能的协同。因此，把传统表达转化为数据表达是文科智能发展的前提，且需要经历四个阶段：（1）把多模态研究素材转化为数字格式的数字化阶段；（2）通过数据标注将数字格式素材转化为可被机器消化的以数值数据表达的数据化阶段；（3）通过建立算法和优化算法③挖掘数据中蕴含的学科化的事物之间的关系模式，即提炼为学科理论的理论化阶段；（4）把学科理论转化为人类可理解的知识，即把分散的理论整合、精练为不断迭代的人类知识的知识化阶段。每一个阶段都涉及文科与机器的深度协同。④ 其中，最关键的环节是文科知识表达的数据化。

数字人文和计算社会科学的发展为发展文科智能奠定了基础，但整体上尚未完成用数据对知识的表达。⑤ 数字人文发端于20世纪50年代，少数研究者为研究文本建立索引或电子文本库；⑥ 到了21世纪，谷歌建立图书文献数据库、出版机构建立期刊和报纸等各类表达载体文献库，关注改进检索方法以及探索研究方法。⑦ 数

① Igor Grossmann et al.，"AI and the Transformation of Social Science Research，" *Science*，vol.380，no.6650，2023，pp.1108-1109.

② Xuan Zhang et al.，"Artificial Intelligence for Science in Quantum，Atomistic，and Continuum Systems，" February 26，2025，https://arxiv.org/abs/2307.08423v2，March 18，2025.

③ 运用知识图谱是实践专家知识数据化的示例，不意味着专家知识的数据化只有一种路径。恰恰相反，本文认为选择适合素材的数据化方法才是获取高质量数据的方法。

④ 参见邱泽奇：《人机互生：范式革命与知识生产重构》，《探索与争鸣》2024年第11期。

⑤ Laurent Petit，"Humanities and Social Sciences (HSS) and the Challenges Posed by AI：A French Point of View，" *AI & Society*，vol.39，2024，pp.2791-2797.

⑥ Ken Jones，*Education in Britain：1944 to the Present*，Cambridge：Polity Press，2016.

⑦ 参见冯志伟：《数字人文研究的四个层次》，《南京师范大学文学院学报》2023年第3期。

中国知网 https://www.cnki.net

字人文几十年的发展推动了大量文献题录的数字化，极大地便利了文献检索。然而，大多数成果实现的是把多种载体承载的知识转为数字格式存储到存储设备，尚未将其转化为可被机器识读的数值数据。此外，自有大数据积累以来，另一些成果关注对大数据进行特征提取，如运用深度学习方法对文本或多模态数据进行特征判断。① 在社会科学领域，自21世纪初期以来，逐渐形成了“计算＋学科”② 的潮流，如计算社会学、计算政治学等。人们对大数据的关注也从社会科学扩散到人文学科。不过，尽管计算社会科学强调计算方法的应用，③ 使用了可计算数据，依靠的却多是“远读”方式。④ 简言之，由于传统知识表达尚未转化为可被新一代人工智能消化的数值数据，人文社会科学尚未建立起与人工智能协同进行知识生产的路径或方法，把以数字格式存储的文本、语音、图像、影像等多模态素材中蕴含的知识转化为可被机器消化的数据，是文科智能发展的当务之急。

近几十年来，数智技术及其创新在向经济和社会等领域渗透应用的同时，国家也布局了大量与知识生产相关的重大项目和工程，如古籍文本数字化、古代绘画精品数字化、司法案例数字化、智慧城市建设、智慧社会治理探索、智慧司法等。这些重大项目和工程在将适宜人类运用的知识表达转化为适宜机器消化的知识表达方面做出了重要尝试，却也形成了限制文科智能发展的数据割裂局面：一方面是由数智技术应用产生的海量多模态数据由各类机构占有，另一方面是大量以数字格式存储知识的多模态载体也由各类机构掌控。一些数据处于可用状态却未经高质量标注，另一些数据则未被数据化而难以被机器消化。

在新一代人工智能技术带来的历史机遇面前，更有效地将人工智能技术与人文社会科学知识生产协同融合，发展文科智能，是人文社会科学发展的必然趋势，也是构建中国特色哲学社会科学、建构中国自主知识体系的可能路径之一。那么，文科智能发展如何突破高质量数据瓶颈？

回顾科学智能发展的关键节点能为我们提供有益的启示。2007年，李飞飞启动了名为ImageNet的项目。她与合作者通过互联网收集了1500万张包含2.2万个不同类别对象的数字格式图片，招募了来自167个国家的4.8万多名志愿者对图片进行标注，

① 参见卢鹭、郑伊、方一新：《语言模型：汉译佛经考辨的新方法——以安世高译经考辨为例》，《浙江大学学报》（人文社会科学版）2025年第2期。

② David Lazer et al.，“Computational Social Science，” *Science*，vol.323，no.5915，2009，pp.721-723.

③ David Lazer et al.，“Computational Social Science：Obstacles and Opportunities，” *Science*，vol.369，no.6507，2020，pp.1060-1062.

④ 参见陈茁、陈云松：《经典远读：基于机器学习的〈马克思恩格斯全集〉分析》，《西南民族大学学报》（人文社会科学版）2023年第3期。

中国知网 https://www.cnki.net

把图片转化为可计算的数据，即数据化。① 李飞飞认为，若想让机器像人一样“理解”现实，为其提供大规模、高质量的训练数据，是关键一步。

ImageNet的标注数据集被认为是奠定ChatGPT出现之前人工智能技术变革的关键基石。在ImageNet获得成功后，李飞飞团队又建设了一个新的名为Behavior的标注数据集，包括1000种日常活动、50个场景和5000多个具有属性的对象。除了对象及其属性，改变对象状态的动作也被列入标注。除此以外，标注者还需要在对象和动作之间建立关联（如“苹果”与“烹饪”和“切片”相关联，而和“打开”无关等）。因此，建设这样的标注数据集实际是把人们的日常知识带进对图片的标注，用数据构建常识知识库。

李飞飞团队的工作说明了高质量数据不仅是运用算法的前提，更是提高算法效率的核心。事实上，在经历了两年的狂欢之后，生成式预训练人工智能也转向对运用高质量数据的关注。DeepSeek-R1的成功证明，运用精妙算法可以突破算力瓶颈，却依然不能缺少大数据。李飞飞团队的DeepSeek-S1模型尽管没能引起像DeepSeek-R1那样的关注，却证明了使用1000个高质量数据产生的输出可能优于使用10万个普通质量数据得到的结果。

与李飞飞团队训练机器视觉不同，推动文科智能的第一步是把宜人的知识表达转化为宜机的数据表达。面对人类积累的丰富知识，尽管可以依据不同诉求把知识划分为不同的复杂类型，在宜机诉求下，依然可以删繁就简，把知识划分为日常知识和专门知识。其中，日常知识是获得普及且转化为常识的知识，专门知识则是针对专门领域的知识，文科知识便是专门知识中的一类。在文科知识中，除了刻画事实类对象的专门知识以外，还有大量认知类的专门知识，如表达观念、观点、理论、意义等。在两类知识之中，由专家掌握的认知类知识即专家知识是更浓缩、质量更高、对知识生产影响更大、更体现知识主体性的类型。在这里，专家知识指人类专家创造的专门知识。即使机器参与知识生产，知识的价值主体依然是人类。因此，专家知识的另一层所指是人类专家认同的专门知识。此外，专家知识还是处于动态发展、更新、迭代中的专门知识。以此为语境，把数字格式存储的文科知识转化为机器可消化的数据便是推动文科智能发展的关键工作，其中，最值得关注的是专家知识的数据化，即把以人类学习和理解为目标的知识转化为以机器学习和理解为目的的数据。

为探索这一路径，本文以法律规范文本为例。在众多文科知识的题材中，法学领域的法律规范文本是一类典型。一方面，它汇聚了多源专家知识，不仅知识密集度高，还是典型的复杂知识；另一方面，对它的数据化转化更加考验从宜人到宜机

① Jia Deng et al.，“ImageNet：A Large-Scale Hierarchical Image Database，” *2009 IEEE Conference on Computer Vision and Pattern Recognition*，Piscataway，NJ：IEEE，2009，pp.248-255.

中国知网 https://www.cnki.net

的表达转化的困难度，其数据化实践更具示范性。探索法律领域专家知识数据化的路径，可为更大范围专家知识的数据化提供示例。

二、法学研究的方法困境

法律与技术都是典型的复杂系统。① 法律系统的复杂性增长体现在两方面：一是法律数量的增加，以及规制的细分。二是法律之间的相互关联日益密切，法律系统正成为高度互联的规则网络。② 技术系统同样呈现出复杂性增长趋势。数据、算法与算力被视为人工智能的“三驾马车”，共同创造了空前复杂的技术系统。例如，阿尔法（Alpha）系列训练的神经网络模型有近 1 亿参数，生成式预训练大语言模型（GPT）系列则有千亿级参数。其中生成参数的深度学习算法仍被认为是难以解开的“黑箱”，呈现的智能被认为是复杂“神经元”互动的涌现结果。③

应对法律与技术复杂性带来的新实践与新认知需求，一直是法学研究的重点之一。法学与技术的交互也逐渐呈现出两条典型路径。

第一条路径偏重于运用技术对法律系统进行处理与理解，并最终形成新的知识或产品。研究者们运用新技术加工处理法律信息或法律数据，形成针对法律问题的量化研究成果或能够提升司法实践效率的技术产品。法律计量学、法律推理智能系统、法律信息学等均可视为这一路径下的子范畴。④ 随着人工智能的发展，一方面，有研究者利用日臻成熟的自然语言处理技术解析法律规范文本，让计算机不断理解并熟悉法律规范文本的语言特征；另一方面，也有研究者在探索法律推理逻辑的分解和转化，使之能够被人工智能理解和掌握。⑤

第二条路径偏重于从法律视角对技术系统进行理解与规制。这一路径关注新兴技术发展与应用中涌现的新治理议题，旨在利用法律工具规范技术系统的行动主体

① Peter Turchin et al., “Quantitative Historical Analysis Uncovers a Single Dimension of Complexity That Structures Global Variation in Human Social Organization,” *Proceedings of the National Academy of Sciences*, vol.115, no.2, 2018, pp.E144-E151.

② J.B. Ruhl and Daniel M. Katz, “Mapping Law's Complexity with ‘Legal Maps,’” in Jamie Murray et al., eds., *Complexity Theory and Law: Mapping an Emergent Jurisprudence*, New York: Routledge, 2019, pp.23-44.

③ 参见乔天宇、邱泽奇：《复杂性研究与拓展社会学边界的机会》，《社会学研究》2020 年第 2 期；王芳、郭雷：《数字化社会的系统复杂性研究》，《管理世界》2022 年第 9 期。

④ 参见申卫星、刘云：《法学研究新范式：计算法学的内涵、范畴与方法》，《法学研究》2020 年第 5 期；季卫东：《计算法学的疆域》，《社会科学辑刊》2021 年第 3 期。

⑤ 参见雷磊：《人工智能时代法律推理的基本模式——基于可废止逻辑的刻画》，《比较法研究》2022 年第 1 期；廖备水：《论新一代人工智能与逻辑学的交叉研究》，《中国社会科学》2022 年第 3 期。

中国知网 https://www.cnki.net

以及技术应用的方向。个人信息、数据、算法、生成式人工智能等议题均已进入数字法学视域。① 研究者不仅利用既有概念工具与理论框架对新技术与新现象进行描述与理解，也通过对既有法律规范进行重述或创造新的规制工具或规制模式对以技术为枢纽的社会关系进行调整。②

两条路径的成果积累正逐步在法学知识体系中实现跨学科的知识迁移、重塑和更新。③ 两条路径虽有各自的脉络，却也都在试图推动知识生产与数智技术的进一步交叉与融合，可视为对文科智能的初步探索。只是均未实现技术逻辑与知识逻辑的实质性融合。其互动的困境在于，前一路径缺失专家知识，而后一路径形式化不足，各缺一隅。

当前，研究者在运用技术对法律系统进行处理和理解时，由于普遍缺少实现专家知识数据化的法律素材数据集，技术运用层次和知识产出能力都未能获得机器智能的支撑。以美国乔治城大学法律中心构建的现存最大的美国最高法院宪法解释与法律解释数据库为例，其中的案例材料依然只实现了文本的数字化，即以数字格式存储文本，而未能实现法学专家知识的数据化。研究者在利用数智技术进行研究时，大多只能利用自然语言处理技术检索特定语词来进行统计式研究。④ 在涉及对案例的深度分析或部分语段精细理解时，研究者只能先从数据库中检索出所有相关案例，然后再人工进行精读和分析。⑤

在实践中流行的法律科技产品也面临类似的困境。人工智能更多是“解释性智能”而非“理解性智能”，其本质是对人类既有知识语料的模式识别和重组。人工智能或许可以充当更好用的搜索引擎或法律文书语言助手。⑥ 然而，在实质性的法律分析、法律推理等方面，由于其中缺乏法学专家知识，新一代人工智能技术与法律研究和实践的协同始终面临较严重约束。

相对而言，由于技术迭代与法学学者技术知识更新速度不匹配，当研究者试图从法学视角对技术系统进行理解和规制时，其所提供的法学专家知识往往无法被技术系统学习和理解。尽管法学专家提出的许多观点在观念上具有合理性，但在技术实践中却较难找到对应物及着手点。法律与技术各执一端也导致法律与技术的转化

① 参见苏宇：《“信息技术＋法学”的教学、研究与平台建设：一个整体性的观察与反思》，《中国法律评论》2021年第6期。

② 参见张吉豫：《数字法理的基础概念与命题》，《法制与社会发展》2022年第5期。

③ 参见马长山：《数字时代的法学教育转型》，《上海政法学院学报（法治论丛）》2023年第1期。

④ Piotr Bystranowski and Kevin Tobia, “Measuring Meta-Interpretation,” *Journal of Institutional and Theoretical Economics*, vol.180, no.2, 2024, pp. 281-305.

⑤ Anita S. Krishnakumar, “Textualism in Practice,” *Duke Law Journal*, vol.74, no.3, 2024, pp.573-679.

⑥ 参见李超：《人工智能辅助立法：现状、困境及其因应》，《人大研究》2020年第4期。

中国知网 https://www.cnki.net

困难，抑制二者间的有效互动，即便清晰地提出规范性要求，由于不足以将其有效地编码为适宜机器消化的数据，技术体系也会因编码方式、话语体系差异而难以准确把握法律的内容，甚至可能产生望文生义的误解。

综上所述，法学研究跨学科实践困境的症结在于法学知识的表达依然宜人而不是宜机。与文科其他学科一样，把法学知识表达转化为适宜机器阅读、学习、理解和探究的数据，是法学知识生产与实践智能化的关键。有鉴于此，本文基于对《中华人民共和国个人信息保护法》（以下简称《个人信息保护法》）的数据化实践归纳出一条专家知识数据化的“三步走”路径。《个人信息保护法》的法律规范类型较为丰富，既有原则性条款也有规则性条款，既有权利型条款也有义务型条款，既有规范结构相对完整的条款也有规范结构存在缺省的条款，是法律文本典型类型。作为最复杂的法律文本之一，对其数据化实践的探索和方法提炼具有较强的推广性。

三、专家知识的数据化

对文科而言，专家知识数据化是将多模态载体蕴含的专家知识转化为数字编码，即数据，使其不仅可以用机器存储，还能被机器阅读、学习、理解和探究。它与以往的知识工程看似相似，本质却不相同。专家知识数据化的核心目标不只是构建专家系统，还包括对专家知识的范围框定、逻辑重构以及数据表达，① 是运用数据逻辑重新表达内含于多模态载体的专家知识。在法学领域，专家知识数据化的产物是用数据逻辑和形式表达的法学知识。

（一）专家知识数据化的路径

专家知识数据化不仅要遵循专家知识的知识逻辑，还要遵循数据表达的技术逻辑，使生成的以数据表达的知识可以被机器阅读、学习、理解和探究，且数据生成过程应尽可能简单和易操作。在理念模型上，专家知识数据化包含三个关键步骤（见图 1）。

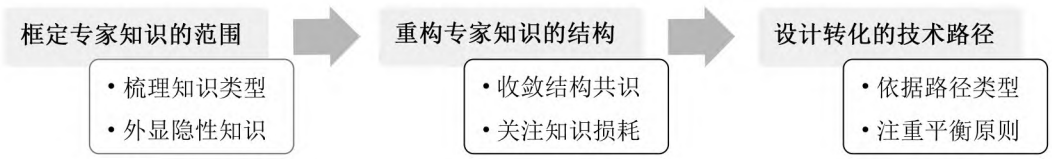

图 1　专家知识数据化的实现路径

1. 框定专家知识的范围

① 参见邱泽奇：《数字化与文科范式革命》，《大学与学科》2023 年第 2 期。

中国知网 https://www.cnki.net

第一步是框定需要且可被形式化表达的专家知识。随着人工智能的发展，人类将被新技术从智力密度较低的重复性知识生产中解放出来，转向从事具有较高创造性和价值规定性的生产活动。一些过去能被称为专家知识的内容，现在已经可以被机器低成本低错误率地提取。这一类专家知识已经没有数据化的必要或很容易数据化。框定专家知识范围的关键是框定出那些不能通过现有技术实现自动获取的专家知识。

在知识管理中，知识常被分为显性知识与隐性知识。前者可以通过符号或自然语言编码、阐述并用于交流，后者则根植于行为、实践和经验，并存在于具体场景之中。① 二者间的界限并不严格，某些隐性知识可通过外显化的方式转化为显性知识，如通过知识获取过程回忆和模拟等方式，尝试用自然语言进行表达。法律系统的专家知识同样既包括显性知识，如最高人民法院发布的指导案例中的裁判要旨；也包括隐性知识，如特定案件中证据链条是否完整的判断标准。

专家知识数据化是将专家知识用数据逻辑和形式进行表达。实现专家知识数据化的前提是专家知识被尽可能清晰地外显化。框定专家知识范围，不仅是就特定领域或针对特定问题对显性知识和隐性知识进行系统梳理，明确各自的内容及范围，也要明晰显性知识表述方式，充分外显化隐性知识。

从一般方法上来看，把隐性知识外显化，通常要尽可能地对隐性知识进行形式化表达。对其中难以表达的部分，或许可以借助大量包含隐性知识的多模态载体，通过归纳与关联的方式建立意义网络或关联逻辑，实现对这部分知识的模拟或趋近。一个可行途径是让不止一位专业人员（如行业或领域专家）参与数据标注或知识图谱构建，通过多位专家在数据化进程中的复式知识对话，让隐性知识外显。例如，对于某类交易中需思考和应对的法律风险这一隐性知识，要使其外显化及形式化，既可以基于大量既有成熟交易合同文本做归纳总结，也可以请多位领域专家从法律规范文本和裁判文书入手，以相关交易为提示词进行归纳，还可以请领域专家对交易流程的法律环节进行拆分，在此基础上对每一交易环节可能产生风险的互动行为形式或表达形式进行归纳，对可能产生风险的法律主体间关联形态进行梳理和表达，并建立与相关法律规范之间的对应。如此，在面对新的交易场景时，便可以借助形式化类比实现对专家法律判断思维的模拟和趋近。

2. 重构专家知识的结构

第二步是重构嵌入于专家知识的表达结构。这既是实现专家知识数据化的关键，也是从以人类自然语言为载体的专家知识表达转化为可被机器阅读、学习、理解和

① 参见 Maryam Alavi and Dorothy E. Leidner：《知识管理和知识管理系统：概念基础和研究课题》，郑文全译，《管理世界》2012年第5期。

中国知网 https://www.cnki.net

探究的编码环节。①

之所以要重构专家知识结构而非识别专家知识结构，是因为这一过程不只是以宜人的方式进行识别，还要在不损失专家知识的前提下，以宜机的方式进行建构。把人类自然语言转化为机器数据语言是一个创造过程，充满了挑战性。其一，并非所有以自然语言表述的专家知识都有清晰的形式结构，此时，识别形式结构即重构结构。其二，对某些专家知识结构，法学专家之间可能没有达成共识，会产生多个侧重点不同的知识结构，因而在重构专家知识结构时需要根据目的在不同形式结构中做选择。其三，能够为法律系统理解的知识结构与可以为机器消化的知识结构可能不尽相同，最终能够服务于专家知识数据化的知识结构必须在技术上具有可转化性。因而重构专家知识的结构需要法学专家与技术专家密切配合，直到最终产出的专家知识结构不仅准确含有专家知识，也能在技术上可行且易操作。在这个意义上，重构专家知识结构也是不断收敛并凝聚法学专家与技术专家共识的过程。

在重构专家知识结构过程中，还可能会遭遇“知识损耗”，即部分知识无法被纳入最终建构的结构。尤其当可能存在多个侧重点不同的知识结构时，这种知识损耗几乎难以避免。对此，或许可以通过重建不同结构的数据予以弥补。即便如此，一个无法回避的问题是，如果某知识尚无法用某结构予以表达，比如无法外显化的隐性知识，便意味着还无法对其实现有效的形式化，这也属于知识损耗。面对这类场景，则需依靠法学专家进行结构迭代，如专家会商。重构专家知识结构需时刻关注可能出现的知识损耗，以此为动力，不断促进建构方案的迭代优化。但是，这也提示我们实现专家知识形式化的可能局限。

3. 设计转化的技术路径

第三步是设计有效的技术路径，将专家知识转化为可被机器识别、存储并处理的数字编码（即数据）。这一过程需要法学专家与技术专家充分互动，在专家知识转化准确性与技术方法易操作性之间建立平衡。

专家知识含量的高低与专家知识是否容易正则表达，是设计转化技术路径的两个重要依据。所谓专家知识含量，即对特定问题，专家理解是否与具有一般知识水平的非专家个体有明显差异。如果差异明显，便应认为专家知识含量较高。在生成式预训练大语言模型不断迭代的背景下，专家知识含量还可进一步被操作化为专家理解是否与由机器内置知识库（如常规词典）产生的理解有明显差异。所谓专家知识是否容易正则表达，指专家知识的识别和提取过程是否可以凭借一套清晰且完备

① 一般认为，数据可以分为结构化数据和非结构化数据。但有研究认为，从计算的视角看，不存在完全非结构化的数据。参见 Aleksi Aaltonen and Esko Penttinen, “What Makes Data Possible? A Sociotechnical View on Structured Data Innovations,” *Proceedings of the 54th Hawaii International Conference on System Sciences*, 2021, pp.5922-5931.

中国知网 https://www.cnki.net

的机器处理规则来实现。若可以，数据化过程便可通过精心设计处理规则并依据规则处理原始信息而实现。由此生成的数据不仅能准确反映专家知识，数据生成标准的一致性也能得到较好的保证。

依据上述两个考量维度，可以将转化技术路径划分为四种类型（见图 2）。对于专家知识含量较高且易正则表达的内容，可以采用正则处理方式，即基于规则将专家知识转化为可嵌入结构的编码，但处理规则的制定需要法学专家介入，我们称之为专家介入的正则处理（图 2 中的 A）。

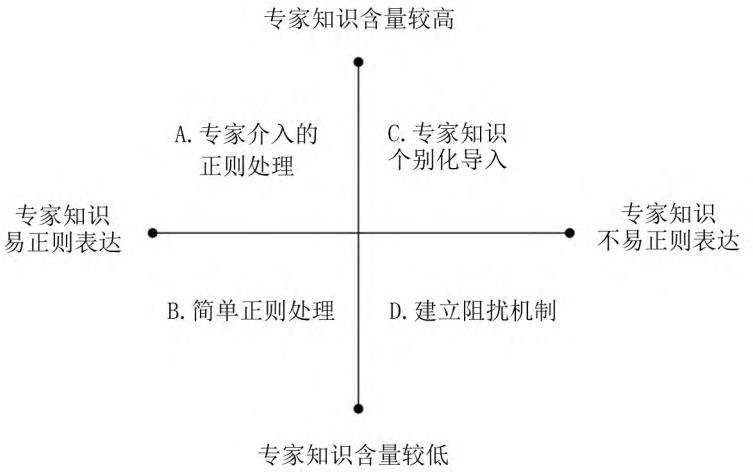

图 2 设计转化技术方法的依据及技术方法类型

对于专家知识含量较低且易正则表达的内容，例如基础语法知识，可以直接调用机器内置的处理规则进行转化，我们称之为简单正则处理（图 2 中的 B）。在处理规则清晰且完备的情况下，A 和 B 既可以通过机器自动批量完成，也可以由经统一培训后的非专家人士完成。

对于专家知识含量较高且不易正则表达的内容，由于知识内容具有较强特异性，故无法制定统一处理规则。但考虑到其重要性，有必要在专家深度介入的情况下做有针对性的知识转化，我们称之为专家知识个别化导入（图 2 中的 C）。

举例而言，对于裁判文书中高频出现的语词类显性知识，在现有技术条件下，通过一部常规词典和检索规则便可完成。相对地，法律法规中常常会出现“法律法规另有规定的除外”或“其他情形”等除外规定或兜底条款。这些条款的具体指向高度依赖专家知识来明确，这就需要进行专家知识的个别化导入。不过，专家知识的个别化导入并不排斥数智技术的辅助。在实践中，专家知识个别化导入常表现为构建一种人机互动的有效方案，通过专家对每一条机器辅助处理数据的审核调整，确保实现素材中专家知识的数据化转化。

对于专家知识含量较低且不易正则表达的内容，如果转化过程中不将这些知识与专家知识含量较高的知识内容清楚界分，便会干扰后续针对专家知识含量较高数

中国知网 https://www.cnki.net

据的处理和分析，造成数据污染。然而，如果直接忽略这部分知识，不对其进行转化，又可能导致知识的不完整，影响其整体性呈现以及后续的理解与运用。因此，有必要设计某种机制，区分性地识别并处理这两类知识内容，我们称之为建立阻扰机制（图 2 中的 D）。

在设计专家知识数据化转化路径的同时，还有必要设计转化结果的评估方案，以检验转化路径的可信性和可靠性。以利用生成式人工智能对数据进行检验为例，如果专家知识的应用场景是实务裁判，则可以先利用转化出的专家知识数据对人工智能进行预训练或后训练，进而要求其对现实裁判场景提出裁判建议；而如果专家知识的主要应用场景是学术研究，则可以考虑围绕某一学术研究议题要求人工智能进行多轮次的对话输出。此外，为检验下文所述阻扰机制的有效性，还有必要特别设置干扰性较强的问题单独进行提问。

（二）专家知识数据化的现实意义

在法学知识生产领域，专家知识数据化将显著提升素材可被分析和处理的广度和深度，① 并逐步实现法学知识的技术编码和数据化重写。② 这有利于突破法学研究跨学科实践的发展困境，③ 破除法学研究与计算机科学等学科研究间的壁垒，促进真正的跨学科交叉融合。同时，专家知识数据化也为整个法学知识体系向智能化发展提供方法储备和数据储备。法学界已经意识到，无论是教义法学还是社科法学都需共同面对数智化，都需要生产可训练人工智能理解人类法律系统运转逻辑的专家知识数据，以及能调整人工智能推理的具体规则或推理框架。数智化在给法学既有知识体系提出挑战的同时，也给法学提供了弥合重塑知识体系的机会。人工智能模型训练要求的专家知识数据提供了学科知识良性互动的平台和检验标准。

在法治实践领域，专家知识的数据化有利于应对法律系统复杂性，真正践行《法治社会建设实施纲要（2020—2025 年）》要求的“推动公共法律服务与科技创新手段深度融合”。受部门法划分的影响，部门法内部体系性得以提升的同时，却产生了部门法之间的知识屏障。④ 已有研究发现，如果法律专业人士只从自己熟悉的部门法领域提出对策及建议，有时反而会陷入依部门法形成的方案间难以协调的窘境。⑤ 专家知识数据化可让法律从业人员借助更加可靠的知识库跨越自身知识局限，

① 参见刘石、李飞跃：《大数据技术与传统文献学的现代转型》，《中国社会科学》2021 年第 2 期。

② 参见马长山：《数字法学的理论表达》，《中国法学》2022 年第 3 期。

③ 参见刘艳红：《从学科交叉到交叉学科：法学教育的新文科发展之路》，《中国高教研究》2022 年第 10 期。

④ 参见熊伟：《问题导向、规范集成与领域法学之精神》，《政法论丛》2016 年第 6 期。

⑤ 参见宋亚辉：《风险控制的部门法思路及其超越》，《中国社会科学》2017 年第 10 期。

中国知网 https://www.cnki.net

其他社会成员也可以通过用户友好的平台界面更便捷且有针对性地查询、学习、运用法律知识。在各类法律大模型蓬勃发展的当下，专家知识数据化产出的数据，可以成为现有各类训练数据集的高质量数据源，从而进一步扩大训练数据的规模与类别，提升法律人工智能在具体应用场景下的表现。

四、法律规范文本的数据化实践

推动专家知识数据化的“三步走”路径落地应用，需针对具体任务场景探讨与之相适配的具体方法。法律规范文本是法律从业者、法学研究者乃至普通社会成员频繁接触并使用的类型，无疑是法学专家知识的聚合点。① 对此，本文以《个人信息保护法》为实例进行了多轮实验。以实验过程和产出获取为例，接下来详细阐释如何通过“三步走”来达成专家知识数据化。

在框定专家知识范围时，不难发现法律规范文本中蕴含着三类专家知识，即阐释学知识、社会科学知识与立法论知识。其中，阐释学知识因其自身特征，最契合形式化与数据化要求。

在构建专家知识结构时，参考阐释学知识特征及其与数据科学知识图谱的契合特性，可以将阐释学知识重构为依据“场景—主体—行动”的结构逻辑而打造的专家知识，并运用“情景条件—法律后果”“主体—行为”“规则—规则”这三重关系，将此类专家知识进行形式化呈现。

在设计转换技术环节时，可以依据上述结构逻辑，重构法律文本标注过程，借助适宜的技术手段（实验中运用了知识图谱），实现专家知识数据化。这一过程宛如在法律与技术系统之间建立高效沟通的桥梁。

（一）与法律规范文本相关的专家知识类型

在法学知识体系中，与法律规范文本相关的专家知识大致有三类：第一类是阐释学知识，亦称教义学知识，主要内容涵盖对法律规范要件的拆解与阐释，对法律规则冲突问题的协调以及对法律漏洞的填补与续造等方面；第二类是社会科学知识，聚焦于法律规范文本在现实场景中的效果及所产生的影响；第三类是立法论知识，涉及法律规范文本应该如何生成、修改或调整等相关内容。②

① 本文使用狭义的法律规范文本概念，指成文法的法律文件的文本内容。考虑到我国法律体系具有较强的大陆法系法典化特质，法律规范文本在我国法律体系中占据核心位置，发挥着重要的制度性作用，因此，对法律规范文本中蕴含的专家知识实现形式化转化是一项重要工作。

② 法学讨论中法教义学知识与社科法学知识的分野，更多是从研究方法或内外部研究视角等方面进行的区分，参见陈兴良：《法学知识的演进与分化——以社科法学与法教义

中国知网 https://www.cnki.net

从专家知识数据化需要来看，上述三类知识目前都无法被人工智能自动处理或构建，都有数据化的必要性。然而，这三类知识在组织形态与外显化程度上却有明显差异。作为阐释学知识的重要生产与积累路径之一，法律规范文本评注使得阐释学知识与具体法律条文之间形成极为紧密的关系，其知识的组织方式与积累方式也大致与法律规范文本的组织逻辑相吻合。[①]因而阐释学知识已具备专家知识数据化的前提条件，也成为下文重点分析与处理的对象。

相较于阐释学知识，社会科学知识虽也已有相当程度的积累，但其组织形态通常围绕法律系统的某个重点个案或制度展开，并非按照法律规范文本的形式体例逐章逐条进行。对这类专家知识的数据化应通过处理法学研究论文文本或法学教科书文本，或在具体应用场景的算法参数调整中实现。至于法律规范文本应当如何生成和编排的立法论知识，其外显化程度相对较低，目前更多表现为立法工作参与人士的一类隐性知识，其知识的组织形态也尚不明确，因而这部分专家知识尚不具备数据化的条件，需要先经历专家知识的系统化与外显化过程。

（二）专家知识的结构重构

遵循“三步走”路径，实现法律规范文本阐释学知识数据化的目标，需要在法学专家与技术专家的双向互动中，让法学专家理解法学知识数据化的技术逻辑，让技术专家理解法律规范文本的知识逻辑，在此过程中重构适于用数据进行表达的专家知识结构。本文借助图类型数据来表示阐释学知识。基于“场景—主体—行动”的结构逻辑，可以将阐释学知识结构化为三类关系。

第一，情景条件与法律后果。在单个法律规范文本规则的内部层次，规则适用的情景条件与规范所规定的法律后果构成结构中的一对关系，这正是法律规范作用的基本模式。[②] 本文将其称为情景条件与法律后果关系。人们通常认为法律规则的结构理应是最大的形式化结构。现实中，尽管法律规范文本中绝大多数条款的核心是规范行为，由于法律规范文本中大量存在“行为规则”与“裁判规则”相分离乃至缺省的情况，故而无法都用“假定—行为模式—法律后果”的理想模式加以刻画。此外，不容忽视的是，法律规范文本中还存在如“定义性规则”“引用规则”等不以行为为核心的规范形式。基于专家知识数据化的现实需求，选取情景条件与法律后果的简单形式逻辑对应关系，实际上具有更强的覆盖性与解释力。从形式上看，它契合知识表示的产生式结构，即“if P then Q”，也符合机器语言的“习惯”，更便

学为视角》，《中国法律评论》2021 年第 4 期。考虑到无论是法教义学知识还是社科法学知识，都含有与法律规范文本相关的三类专家知识，故上述区分的意义较小。

① 参见贺剑：《法教义学的巅峰——德国法律评注文化及其中国前景考察》，《中外法学》2017 年第 2 期。

② 参见雷磊：《法律规则的逻辑结构》，《法学研究》2013 年第 1 期。

中国知网 https://www.cnki.net

于转化和机器处理。

第二，主体与行为关系。在情景条件或法律后果的内部层次，法律规范文本规则所规定的主体与行为之间构成结构中的另一对关系，这明确了法律规范作用的具体指向。① 本文将其称为主体与行为关系。从主体与行为支配关系角度来看，“主体可以做什么”“主体可以不做什么”“主体应当做什么”“主体应当不做（不得做）什么”构成主体与行为的四种基本关系类型。每一个规范性要求，既可能存在具体的受动对象，例如《个人信息保护法》第二十五条“个人信息处理者不得公开其处理的个人信息，取得个人单独同意的除外”，在这一法律规范中，“个人信息”即为“公开”动作的受动对象；也可能没有具体的受动对象，比如《个人信息保护法》第九条“个人信息处理者应当对其个人信息处理活动负责，并采取必要措施保障所处理的个人信息的安全”。主体与行为关系既可能只存在于情景条件之中，也可能只存在于法律后果之中，还可能同时存在于两者之中。

第三，规则与规则关系。在一部法律内部，或扩展至整个法律系统的内部层次，法律规范文本规则之间的关联还构成结构中的一对关系，这体现了法律系统的复杂性，本文将其称为规则与规则关系。法律规范文本中的关系，不仅涵盖了单个法律规范文本规则内部的关系，还囊括了法律规范文本规则之间的关系。已有文献将其细分为三种类型：序列关系、层级关系和引用关系。② 尽管法学研究及法律实践已经积累了一些规则以应对上述三种关系，如“新法优于旧法”“上位法优于下位法”“特别法优于一般法”等法谚；然而在现实中，面对规则冲突，究竟该适用何种规则，依然是一个高度依赖法学专家专业判断的问题。

（三）基于结构逻辑的法律规范文本标注方法

在完成专家知识结构的重构后，接下来是采用何种路径将法律规范文本蕴含的专家知识持续且稳定地转化为数据形态。对于这一技术或工程路径，存在多种可选的方案。例如，可以依照阐释学知识的结构逻辑对法律规范文本进行标注，也可以直接对阐释学经典文本进行标注。本文提供一套基于上文所述法律规范文本阐释学知识的结构逻辑、专门针对法律规范文本进行标注的方法。借由这种标注方法，能够将法律规范文本作为核心枢纽，持续实现阐释学知识的数据化。

1. 专家知识数据化表达的工具选择

鉴于上文重构的知识结构涵盖了多种类型关系，可采用图数据结构对这类错综复杂的专家知识结构进行表达。目前，国内外已有诸多研究应用图数据结构对法律

① 参见王利明：《民法案例分析的基本方法探讨》，《政法论坛》2004年第2期。

② Daniel Martin Katz et al.，“Complex Societies and the Growth of the Law，” *Scientific Reports*，vol.10，2020，pp.1-14.

中国知网 https://www.cnki.net

规范文本进行挖掘和分析，如通过刻画法律条文的结构树，或是构建法条、判例间引用网络等方式，来揭示法律系统复杂结构所蕴含的信息。① 图数据结构同样适用于表达本文探讨的、转化后的法律规范文本阐释学知识。

知识图谱是分析图数据结构的常用技术工具，在人工智能领域被应用于解决复杂的认知推理问题，与生成式预训练大语言模型相结合，则可克服各自的技术短板。② 知识图谱为法律规范文本阐释学知识的数据化提供了坚实的技术支撑。

首先，知识图谱技术方便处理异质性图数据，这与本文对法律规范文本专家知识结构的设想更为契合。在重构的专家知识结构中，情景条件、法律后果、主体、行为和规则属于不同层次、不同类型的“节点”，连接它们之间的“边”也属于不同类型的“关系”，知识图谱技术的应用允许将这些不同类型的节点和边并置。

其次，把法律规范文本的专家知识转化为图数据结构具备极大的可扩展性，理论上可以以此针对整个法律系统建模。构建知识图谱数据集不仅可表达单部法律规范文本的专家知识结构，随着此类法学专家知识数据化工作的逐步开展，还可建立不同法律间专家知识的关联，逐渐形成整个法律系统的“全景图像”。

最后，知识图谱数据集建设以及数据集的积累，也为未来实现法治领域的人工智能自动推理留出技术接口，为新一代法律智能的发展创造必要条件。

在后续知识图谱数据集的建设中，构建知识图谱本体（Schema）几乎是最重要的环节，因为本体将作为接下来数据填充的基本架构。对于法律规范文本阐释学知识图谱的本体构建，在实验中采用自顶向下的模式，实体对象、属性和关系是本体的重要组成要素，要对它们有明确的界定。实体对象是基本单元，在本文描述的任务场景中，如法律规范文本中的法条、情景条件、行为均可用实体对象来表示；属性是一类实体对象具有的性质和特征，如对于行为实体来说，其属性包括行为主体、行为类型等；关系是实体对象之间的连接，可能有多种类型，比如，情景条件可能“引致”了某法条实体，对行为的规定一定会“隶属”于某法条实体，情景条件与行为之间也存在相应的“对应”关系等。

2. 实现专家知识数据化的标注方法

为了更好地通过法律规范文本标注实现阐释学知识的数据化，实验根据上文对转化技术路径的讨论，设计了四种具体方法（见图 3）。

（1）专家词典法

阐释学知识中存在着大量的核心概念，它们又进一步指向法律规范文本中的若

① Corinna Coupette et al.，“Law Smells：Defining and Detecting Problematic Patterns in Legal Drafting，” *Artifcial Intelligence and Law*，vol.31，no.2，2023，pp.335-368.

② Shirui Pan et al.，“Unifying Large Language Models and Knowledge Graphs：A Roadmap，” *IEEE Transactions on Knowledge and Data Engineering*，vol.36，no.7，pp.3580-3599.

中国知网 https://www.cnki.net

干表述形态。由于这部分专家知识含量较高，并可以通过正则表达式进行处理，故而可以先形成清晰统一的标注标准，即专家词典，进而通过标注人员与机器之间的协作标注实现这部分专家知识的数据化。

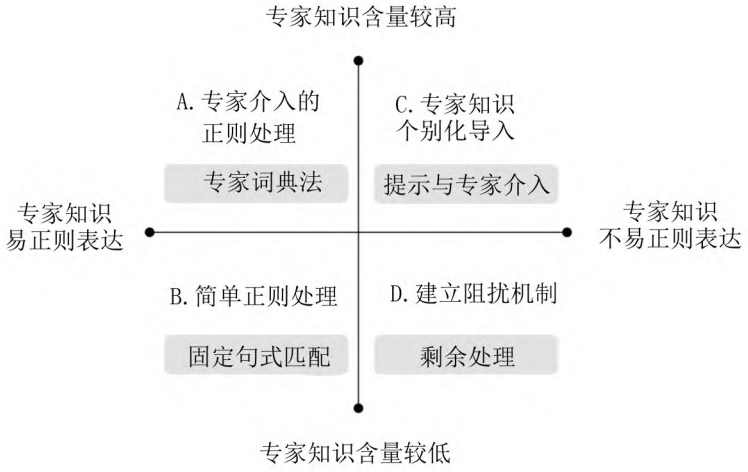

图 3 法律规范文本标注中实现专家知识转化的四种方法

具体而言，由法学专家对需要转化的阐释学知识编写一部关键词词典，澄清文本中相关语词的法律意义，并建立起核心概念与法律规范文本中表述形态的对应关系，从而辅助标注人员对重要法律概念及术语进行识别与分类。例如，在《个人信息保护法》中，“个人信息”与“自然人的个人信息”都对应“个人信息”这一法律概念。“自然人的”这个修饰语并不具备知识内涵，因而应当在标注中统一标注为“个人信息”。类似地，“公开”“明示”等自然语言表达的都是“公开”这一法律概念，这二者之间也没有明显的指向差异，因而也应当统一标注为“公开”。

法学专家在编写关键词词典时，还可根据关键词在法律规范文本中的位置及扮演的角色对其进行分类。例如，主体关键词用来标识法律关系（“主体—行为”关系）中的主体；程度关键词则是用来标识动作、方式或目标等的程度限定语，一般为形容词或副词。区别于常规词典，编制这部关键词词典本身就是凝结法学专家知识的过程。当专家词典的编写达到足够完备的程度后，可由机器基于专家词典中提供的相应规则，对相应关键词进行自动提取和标注，从而提高标注的效率和准确性。

（2）固定句式匹配

在法律规范文本中，大量专家知识是以固定的逻辑句式表达的，如把字句、被字句、条件句式等。例如《个人信息保护法》第二十八条，“只有在具有特定的目的和充分的必要性，并采取严格保护措施的情形下，个人信息处理者方可处理敏感个人信息”，就是一个典型的条件句式。逻辑句式本身体现了法律概念之间的关系，而其中所蕴含的专家知识是由句式本身所规定的，不需要法学专家额外介入进行处理。此时，法律规范文本标注的重点在于准确反映概念之间的逻辑关系，避免因自然语

中国知网 https://www.cnki.net

言表达差异使相同逻辑关系的法律概念被差异化标注。

为应对这种专家知识含量低且易正则表达的情况，需要设置简单正则规则，在法律规范文本专家知识标注中，可以采用固定句式匹配方案。具体而言，在正式标注之前，需对法律规范文本中频繁使用的固定句式进行梳理，并建立一套句式规则作为标注的辅助工具。① 例如，法律规范文本中“以……方式”“通过……”“利用……”等自然语言表述，在标注时都应体现法律概念之间的“手段”与“目的”关系，因而可以建立“以 A 的方式做 B”这样的固定逻辑句式辅助标注。

由于法律规范文本的语言构成相对严谨，其使用的固定句式理论上也可穷尽，内容较易通过基于规则的方式识别并提取。因此，在完善好固定句式列表后，可由经过统一培训、对句式匹配规则熟练掌握的标注人员开展工作，也可通过机器依据固定句式匹配规则批量完成任务。

（3）提示与专家介入

法律规范文本中还存在专家知识含量高但难以运用正则表达的知识类型，即“规则—规则”关系。例如《个人信息保护法》第三十二条规定：“法律、行政法规对处理敏感个人信息规定应当取得相关行政许可或者作出其他限制的，从其规定。”仅通过法律规范文本本身，缺乏专业知识的标注人员无法获知究竟是哪些法律和行政法规对敏感个人信息的处理进行了特别规定。此外，不同条款中所说的“法律、行政法规”可能指向不同的内容，因此这类知识难以通过专家词典或固定句式等正则规则一劳永逸地处理，只能依靠具备专家知识的人员进行个别化处理。

针对这一类专家知识，在法律规范文本标注过程中可以采用“引致关键词”的方式标记，将涉及不明确的规则间引用关系的内容标注出来，提示专家在此处进行介入。在一般标注人员完成初步标注后，再由法学专家介入，对这些标注内容予以逐一核查并完善。在补充具体引致关系时，可以考虑先由机器依据某种用语特征对所有引用该条文的裁判文书进行归纳分析，形成一种形式化的对应关系，而后由法学专家在筛选的同时依据法学知识进行增补，并在这一过程中调整计算机形成对应关系的算法模型。特定标记起到提示性的作用，相当于为专家后续介入留出了适当的“接口”，帮助专家在介入时能迅速定位到需要开展工作的地方。

（4）剩余处理

对于专家知识含量较低且不易采用正则表达的情况，需要建立相应阻扰机制。在法律规范文本标注的过程中，针对情景条件和行为这两种实体对象，均可设置一种“一般关键词”的属性。这类关键词并不存在于专家词典之中，因此，“一般关键

① 在这些句式中，一些是表达逻辑关系的结构，还有一些仅是汉语表达中经常使用的语法结构。在相应的法律规范文本标注实践中，我们一共归纳出八种常见的固定句式类型，分别是因果句式、条件句式、被动句式、对象—动作句式、依据—动作句式、把字句式、判断句式、手段—目的句式。

中国知网 https://www.cnki.net

词”属性是用来存放未被专家词典收录的关键词的。借助“一般关键词”，一方面避免标注人员或机器不当扩大专家知识关键词的数量，进而影响后续对专家知识数据的处理和分析；另一方面实现对法律规范文本的全面标注，从而在转化阐释学专家知识的同时，保留法律规范文本中的剩余信息。

结　语

人类与机器协同互动生产知识早已在进程之中，文科智能是其最新发展阶段。如何更好地运用人类知识来创造性地生产知识、服务人类，以及防止数据毒化和对抗机器幻觉，是新一代人工智能时代知识生产面对的紧迫议题。在科学智能领域，已经出现了一系列人机协同的知识生产模式，如在数字研究中由人类提出猜想，AI完成证明又反向启发人类发展新理论已成为现实。从基础科学到工程科学与实践，运用人工智能进行数据清理、假设生成、假设精练、假设检验、结构预测、材料发现、药物实验等，也已是众多科学智能领域的实践。①

然而，在文科领域要实现人机协同进行知识生产，却被卡在了数据环节。这是因为，让机器协同的前提是先让机器阅读、学习、理解和探究文科知识。机器的食粮是数据，把适宜人类阅读、学习、理解的观念、观点、理论、作品等转化为适宜机器阅读、理解和探究的数据不仅是文科智能实践的前提，也是运用新一代人工智能生产文科知识需要克服的难点，还是突破文科具体领域知识生产瓶颈的重要举措。不仅如此，即使到了机器可以生产知识的阶段，知识生产的责任主体依然必须是人类，因此，专家知识数据化不只在技术上关注文科智能的发展，知识生产的人类性也至关重要。本文聚焦讨论专家知识数据化的法则，尝试推动突破文科智能发展数据瓶颈的路径选择，而不涉及具体工程环节和技术，目标是为人文社会科学普遍需要进行的专家知识数据化提供示例。

第一，文科知识数据化不足，尤其是高质量数据不足是人文社会科学发展面临的困境。这一普遍困境是推动文科智能发展绕不过去的难点。不仅如此，实现专家知识数据化的进程不只会在知识生产中构建新的人机关系，还会形成知识生产分工的新模式，为文科发展创造新的空间和可能。

第二，专家知识数据化“三步走”路径，不只在裁判文书、法学经典著作等法学知识数据化的过程中具有可推广性，作为一种理念模型对人文社会科学领域的专家知识数据化也具有示范性。“三步走”路径的基础是专业知识逻辑与数据技术逻辑的协同与统一，在推动知识生产人机协同的进程中，人文社会科学各领域都会面对

① Lisa Messeri and M. J. Crockett, “Artificial Intelligence and Illusions of Understanding in Scientific Research,” *Nature*, vol. 627, 2024, pp. 49-58.

中国知网 https://www.cnki.net

把宜人表达转化为宜机表达的议题，因此“三步走”是一个普遍路径。在法律规范文本专家知识数据化中运用的结构分析及知识图谱工具对人文社会科学其他学科也具有借鉴性。

第三，文科智能是在理解人类积累的专业知识逻辑与不断发展的数据技术逻辑的基础上实现知识逻辑与技术逻辑的融合。在这一过程中，人文社会科学领域的专家首先要以自然语言等形式将学科知识与素材充分外显化和形式化。在数据化过程中，专家的尽早介入，不仅有利于构建更能反映知识内涵的数据表达结构，也有利于在数据化转换的过程中借由标注或审查等环节反思专业知识并提高数据质量。这是文科智能发展的一条重要路径。随着专家知识数据化在人文社会科学各领域的充分开展与实现，文科智能的其他路径也会涌现、精练和被表达。不过，无论有多少路径，基础依然是知识的数据化表达。可以说，没有数据化表达的知识，会被排斥在文科智能之外，进而有可能被排斥在未来知识生产之外。

专家知识数据化是文科智能向下一阶段发展的前提，文科智能发展也将推动人文社会科学向着整体综合的方向迈进。越来越多的研究者意识到，数智技术正在突破人类社会局限，在人机之间建构数智社会，进而使得现实的诸多方面变得异常复杂和难以预测。面对越来越复杂化的数智社会，以学科分化为基础的知识生产模式，对理解社会总体而言并无助益。① 一如开篇所述，人类的知识生产范式正在经历智能变革，而将研究素材数据化正是智能变革的第一阶段，也是其顺利实现的前提。在文科智能发展的潮流中，人文社会科学研究有望跳出过去分科化困境，走向新的知识综合，最终形成以人工智能为支撑，基于数据的理论发现与检验，输出能够辅助人类理解、判断、决策、行动的综合性知识，深化对数智社会的认识。

在实现专家知识数据化后，发展文科智能还需应对数据理论化与理论智能化的挑战，也还将面对一系列张力，如在物质与精神、科学与人文、实践与理论等一系列张力中蕴含的知识用益张力。一方面，数据化的专家知识在反映人类既有认知成果的同时，也继承了既有的认知偏见和思维模式；另一方面，既有人工智能的决策方式和行动模式主要基于既有数据，尚难做到基于理论的因果推理。因此，让文科智能真正实现向前看的理论创新，还需进一步探索。无论如何，文科智能都将成为推动中国自主知识体系建构的有效探索方向。无论是方法创新的尝试，还是理论创新的愿景，都有望在人机互生时代到来之际，引领新的范式变革，走出一条学术研究与知识生产的新路，为人类知识与文明发展作出中国贡献。

〔责任编辑：余朋翰　责任编审：李凌静〕

① Lorella Viola，*The Humanities in the Digital：Beyond Critical Digital Humanities*，Cham：Springer Nature，2023.

中国知网 https://www.cnki.net

brain, may further inform the architecture of artificial neural networks. Pure consciousness may be understood as a multidimensional manifold, whose inner temporality and a priori correlations offer a basis for constructing a space of meaning encompassing both artificial intelligence and neural activities. The manifold not only reveals a common structure bridging artificial intelligence, neuroscience, and the philosophy of consciousness but also demonstrates a genetic-constitutive sequence of self-knowledge: from the foundational existence of the intentional manifold to the material existence of the neural manifold, and further to the formal expression of manifold learning. This progression offers new possibilities for interdisciplinary unification in cognitive studies.

Datafication of Expert Knowledge and AI for Humanities and Social Sciences

Qiu Zeqi, Li Haolin, Zhang Pingwen et al. · 91 ·

Intelligent digital technologies are reshaping the paradigms of knowledge production, with AI for humanities and social sciences (AI4HSS) emerging as a key direction for future development. High-quality data is a prerequisite for this evolution. One path toward generating such data is the datafication of accumulated human knowledge. Among these, expert knowledge—characterized by its iterative refinement, condensed form, and specialized structure—is especially valuable due to its innovative potential. Legal expert knowledge is particularly representative within humanities and social sciences. Using law as a case study, the datafication of expert knowledge can follow a "three-step" approach: delineating the scope of expert knowledge, reconstructing its internal structure, and designing technical pathways for its transformation. This model is broadly applicable and helps consolidate the data foundations of AI4HSS, while also marking a solid step toward a new paradigm in HSS knowledge production.

The Modern Transformation of Criminal Imputation in the Context of Social Governance

Chen Xuan · 110 ·

The prevailing model of criminal imputation in continental law—prioritizing "dominion" over "obligation"—remains constrained by naturalistic mindset and lacks substantive alignment with the pressing needs of social governance. It also suffers from theoretical superficiality and fragmentation. A modern transformation of this model requires considering obligation as the foundation of imputation. What determines the form that imputation takes is not mere factual occurrence, but the type of obligation involved—such as the obligation to avoid creating risks, the duty to provide care, or the responsibility to maintain essential capacities. On this basis, one can redefine the normative content of "avoidance capability," establish substantive criteria for gradated imputation standards, and reform the attribution model for accomplice liability to respond to new organizational forms and technological conditions. To meet the demands of contemporary social governance, imputation must shift from a focus on outcomes to an emphasis on conduct. It should also broaden its analytical horizon to encompass the integral legal order, drawing on the

中国知网 https://www.cnki.net