观之,人工智能的“算法”也并不能提供绝对的客观性,因为驾驭人工智能技术的人类拥有自身的主观能动性。

史学理论与人工智能的双向互动*

——以大语言模型为例

朱 悦 (同济大学法学院助理教授)
林 展 (中国人民大学清史研究所副教授)

当前,对人工智能造成的冲击,历史学者的态度并不一致,或热烈拥抱,或批判吸收,或警惕反思,总体而言,是在被动地响应。① 但是,仅仅被动响应是不够的。历史学及其理论不仅应当重视人工智能,更应该介入和影响其发展。早在 1976 年,约瑟夫·魏岑鲍姆(Joseph Weizenbaum)就提出,历史学有必要介入人工智能,特别是保存那些无法成为人工智能输入数据的史料。② 2019 年,在斯坦福大学成立了“以人为本的人工智能”研究所,③该研究所认识到,人工智能在改善人类状况方面有着非凡前景,但前提是能够成功引导其朝着负责任的方向发展。为达此目的,研究所吸纳了不少历史学者和其他人文学者。这一机构之所以能赢得人工智能领域的世界级声誉,正是其理念和举措的结果。近年来,也有中国学者注意到这个问题,主张将知识、技能、感性经验、价值观念融入数字人文研究方法,其中包括人工智能的方法,④惜未充分展开。

历史学及其理论之所以能够介入和影响人工智能,特别是当前代表人工智能前沿方向的大语言模型(以下简称“大模型”)的发展,⑤很重要的一个因素在于两者之间工作原理的相似和相通性。在一定意义上说,历史学就是用语言将“数据、记忆、关于过去的证据性遗迹、文献和遗物”变成历史的自觉省思。⑥ 大模型实际上是对人类社会积累的语言材料的吸收和综合,或者说,是以兼收并蓄、有时可能是杂乱无章的方式吸收综合其可以获得的所有文本。由此可见,大模型的工作方式与历史学

* 本文是香港研究资助局卓越研究计划(项目编号:AoE/B-704/22-R)的阶段性成果。

① “热烈拥抱”的代表性文献,参见 Joshua Sternfeld,“AI-as-Historian”,*The American Historical Review*,Vol. 128,No. 3,2023,pp. 1372-1377;“批判吸收”的代表性文献,参见王涛:《大语言模型时代历史书写的路径与局限》,《澳门理工学报》2023 年第 4 期;“警惕反思”的代表性文献参见 Wulf Kansteiner,“Digital Doping for Historians: Can History,Memory,and Historical Theory Be Rendered Artificially Intelligent?”,*History and Theory*,Vol. 61,No. 4,2022,pp. 119-133。

② Joseph Weizenbaum,*Computer Power and Human Reason:From Judgment to Calculation*,W. H. Freeman & Co,1977,p. 238.

③ 关于该研究所的介绍和研究团队,参见 https://hai.stanford.edu/about/people[2024-04-07]

④ 曾军:《数字人文的人文之维》,《中国社会科学报》2020 年 8 月 28 日。

⑤ 此处的大模型,特指 GPT-4、Gemini、“文心一言”等在转换器(Transformer)架构基础上建立起来的模型。有关转换器架构的提出,参见 Ashish Vaswani et al.,“Attention Is All You Need”,*Proceedings of the 31st International Conference on Neural Information Processing Systems*,Curran Associates Inc.,2017,pp. 6000-6010。值得一提的是,Sora 等新兴的多模态大模型也开始采用转换器作为其模型架构的关键组件,参见 William Peebles and Saining Xie,“Scalable Diffusion Models with Transformers”,*Proceedings of the IEEE/CVF International Conference on Computer Vision*,Institute of Electrical and Electronics Engineers,2023,pp. 4172-4182。

⑥ 南希·帕特纳:《根基:关于“过去”知识的理论框架》,南希·帕特纳、萨拉·富特主编:《史学理论手册》,余伟、何立民译,格致出版社 2017 年版,第 1—9 页;彭刚:《叙事的转向:当代西方史学理论的考察》,北京大学出版社 2017 年版,第 279—280 页;陈新:《史学理论的性质、对象、价值与方法》,《史学月刊》2021 年第 1 期。

者的研究存在很大的相似性,即基于过去的各种史料,进行加工和提炼,从而获得比较系统化的知识。正如约书华·斯特恩菲尔德(Joshua Sternfeld)所说,收集模型的训练数据就像是收集史料,模型根据输入产生输出就像是根据证据产生历史,总而言之,“我们越是深入人工智能(大模型)学习的过程,事实上,这个过程就开始和做历史研究的工作越发相似”。①

因此,历史学及其理论与人工智能(大模型)之间存在深度交融的可能和必要,但具体如何实现,既需要认识前者,也需要深入了解后者的运作机理。

一、大模型的技术原理和工作过程

在历史学及其理论的角度下检视,大模型吸收综合过去积累的语言材料的过程包含三个值得关注的环节。一是搜集加工过往的文本材料,形成海量训练数据;二是吸收综合经过加工的训练数据;三是让人工智能和人类的价值保持对齐(AI alignment),从而让大模型更好地服务于人类的价值、目标和利益,防范人工智能的失控、失范。

大模型能够发挥今日所见的强大能力,离不开对过往数千年间全人类所产生的海量文本材料的搜集加工。ChatGPT 的核心研发者之一伊尔亚·苏茨克维(Ilya Sutskever)直陈,大模型背后起作用的关键原理,很可能是高效的压缩。② 或者说,ChatGPT 等大模型所实现的最主要的技术突破,就是能够以一个相对于所有这些海量文本材料来说尺寸很小的模型,以小见大地凝练过往产生的这些材料。这也得到了最近的计算史研究工作的佐证。过去 50 年来人工智能的大幅增强,离不开数字化文本材料的大幅增加,特别是高质量的文本材料的积累。③ 由古登堡计划(Project Gutenberg)等项目数字化的古籍,构成了高质量文本的中坚部分。④ 高质量文本越多,大模型的能力上限就越高。反之,随着高质量文本逐渐用尽,大模型的发展也面临“数据用尽”的瓶颈。⑤

尽管高质量文本如此珍贵、近乎短缺,但目前来看,人工智能研发者主要还是从工程方面的经验出发,总结一些较为直观朴素的技巧来加工这些文本,形成用来训练大模型的海量数据。谢恩·朗普雷(Shayne Longpre)和杰森·韦(Jason Wei)等研发者分别总结了一些最常用的技巧。简言之,要通过区分新旧、区分来源、去粗取精等方式筛选文本,进而通过抽取、替换、移除等方式深入加工文本。⑥

---

① Joshua Sternfeld, “AI-as-Historian”, p. 1372.

② https://the-decoder.com/openai-co-founder-explains-the-secret-sauce-behind-unsupervised-learning/[2024-04-07]

③ Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho, “Will We Run Out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning”, arXiv, 26 Oct. 2022, https://arxiv.org/abs/2211.04325[2024-01-12]

④ Martin Gerlach and Francesc Font-Clos, “A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics”, *Entropy*, Vol. 22, No. 1, 2020, p. 126.

⑤ Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn and Anson Ho, “Will We Run Out of Data? An Analysis of the Limits of Scaling Datasets in machine learning”, arXiv, 26 Oct. 2022, https://arxiv.org/abs/2211.04325[2024-01-15]

⑥ Shayne Longpre et al., “A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity”, arXiv, 13 Nov. 2023, https://arxiv.org/abs/2305.13169[2024-01-22]; Jason Wei and Kai Zou, “EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks”, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 6382-6388.

虽然朗普雷和韦完全是在技术研发的情境下进行讨论,但究其内在,这些技巧的目的都是让大模型尽可能聚焦于高质量的语料。

大模型吸收海量训练数据的能力,很大程度源自其采用的转换器(transformer)这一结构。大模型实际上就是由许多个转换器组成。转换器的优势和局限都很突出。优势是对于内部构成一定顺序的数据,比如说经过数字化以后的文本数据,转换器加以吸收综合的能力非常强大。① 局限是对于内部缺乏一定次序的数据,比如图像、气味甚至体感等,转换器的这项能力就没有那么强了。从增强大模型生成结果的准确性、全面性和深入程度的角度加以检视,对照历史学对史料的要求,可以针对性地提出两方面的批评。一是大模型照单全收的吸收综合委实太过粗糙,很可能在推理生成过程中将不尽准确甚至虚假错误的文本作为依据,造成大模型所谓"胡说八道"的"幻觉"(hallucination)问题。② 在历史学者针对 ChatGPT 展开的测评中,这一缺陷充分暴露——大模型时不时就"幻想甚至编造很容易识别为谬误的事件"。例如,对于简单的、有正确答案的事实性问题,有的大模型只能达到一半左右的正确率。③ 二是大模型的兼收并蓄很大程度上仅限于数字化的文本。这不仅忽略了原始形态的文本所携带的某些信息,也丢失了以文本之外的其他形式存在的某些信息。用历史研究的眼光来审视,我们可以说,大模型采取了一种记忆超群、反应敏捷,但只依靠"史料"的点校整理本,而忽略刊刻影印(或原始史料)和历史现场感的利用"史料"的方式。这样的方式当然可以解决不少问题,但也注定难以解决需要体悟史料物质性和史料现场感的问题。④

大模型具备广泛、强大的生成能力,也蕴涵一定的风险。早在 1960 年,诺伯特·维纳(Nobert Wiener)即指出:一个学习能力和进步速度特别快的人工智能,如果其价值观和人类不完全一致,那将是十分危险的事情。⑤ 鉴于此,人类需要"确信我们输入机器的目标确实就是我们想要的目标"。⑥ 诚然,当时人工智能模型的能力远远比不上今天的大模型,但 60 年前的技术先驱已经注意到让人工智能与人保持一致、服务人类目标的重要性。在大模型能力更加突出的今天,如何保持人工智能与人类价值观充分对齐,已经成为最为棘手的问题之一。⑦

具体而言,实现大模型与人类价值的充分对齐,至少需要做到两点。一是设法识别大模型

---

① Ashish Vaswani et al.,"Attention Is All You Need",pp. 6000 – 6010.

② Vipula Rawte, Amit Sheth, and Amitava Das,"A Survey of Hallucination in Large Foundation Models", arXiv, 12 Sep. 2023, https://arxiv.org/abs/2309.05922[2024 – 01 – 16]

③ Giselle Gonzalez Garcia and Christian Weilbach,"If the Sources Could Talk: Evaluating Large Language Models for Research Assistance in History", arXiv, 16 Oct. 2023, https://arxiv.org/abs/2310.10808[2024 – 01 – 18]值得补充的是,不同的模型在两位学者的测试中显示出显著的能力差异。ChatGPT 尽管无法完全避免幻觉,但能够正确地回答大部分问题。

④ 例如,这样的方式也许可以吸收综合经过拼接整理的简牍内容,甚至加以发挥,但注定难以解决需要通过仔细观察编绳方式和简背划线才能取得突破的释读难题。有关这一巧妙利用史料物质性取得重要发现的实例,参见孙沛阳:《简册背划线初探》,《出土文献与古文字研究》第 4 辑,上海古籍出版社 2011 年版,第 449—462 页。

⑤ Norbert Wiener,"Some Moral and Technical Consequences of Automation: As Machines Learn They May Develop Unforeseen Strategies at Rates That Baffle Their Programmers", *Science*, Vol. 131, No. 3410, 1960, pp. 1355 – 1358.

⑥ Nobert Wiener,"Some Moral and Technical Consequences of Automation: As Machines Learn They May Develop Unforeseen Strategies at Rates That Baffle Their Programmers", p. 1355.

⑦ Jiaming Ji et al.,"AI Alignment: A Comprehensive Survey", arXiv, 26 Feb. 2024, https://arxiv.org/abs/2310.19852[2024 – 04 – 05]值得指出的是,即使这一综述长达近百页,也很难说已经概括了问题的全貌。

的价值观,二是如果识别出来的价值观与人类的价值观不一致,设法加以介入和纠正。目前,在第一个层面上,人工智能研发者主要是通过问卷访谈的方式来测试和识别。比如说,向 ChatGPT 或者文心一言提出一组标准化的、用于判断人格特征和价值倾向的问题,根据答案得到测试结果。① 综合现有的测试和识别结果来看,大模型确实会受到训练文本、吸收过程和用户提问等环节蕴涵的价值观的影响,可能在性别、职业、种族、意识形态等多个方面生成包含偏见的文本。②

第二个层面的问题更加复杂,无论在理论方面还是实践方面,都还存在很多有待突破的难点。简言之,为了实现纠正大模型价值观的目标,首先需要将人类的价值观用大模型能够学习的形式表示出来,然后尽可能鼓励大模型产生和人类价值接近的回答,抑制其产生不符合人类价值的回答。③ 当前,无论人类价值的厘定和表达,还是引导大模型的立场和导向,几乎完全由技术研发者来决断。④ 或者说,几乎完全由研发者依赖其研发技术的工程经验,用技术手段加以书写。这样一来,如果技术研发者对大模型生成文本的价值体察不够充分,或者其所持有的价值观和大模型所服务的对象的价值观不完全一致,人工智能和人类价值观就很难说实现了对齐。面对这一局限,史学理论有关价值、立场、意识等如何将文本碎片"拼合起一幅可以辨识出轮廓的图景"的认识,应当可以为研发者体察和引导大模型的价值对齐提供一定帮助。⑤

总之,大模型的成就固然引人瞩目,但各个主要技术环节都还有显著的改进空间。一是文本语料的收集加工需要更加深入的考订辨正。二是在吸收综合语料时,大模型不仅需要学习历史学的考证功夫,还要面对"纸上得来终觉浅"的局限。三是在价值对齐方面,不仅需要体察大模型所隐含的价值,引导其向正确的方向改进,还要对其中的偏差加以纠正。2023 年以来,技术领域正在积极地解决这些问题。一是文本语料的筛选加工越发受到重视,技巧也越来越繁杂。⑥ 二是对具身人工智能(embodied AI)和世界模型(world model)的探讨越发热烈:具身人工智能以赋予人工智能身体感为其愿景,世界模型则关注人工智能对日常场景和生活常识的认知。⑦ 三是价值对齐成为热门课题,取得越来越多的研究进展。因此,此时正是引入史学理论的最佳时机。

---

① Jérôme Rutinowski et al. ,"The Self-Perception and Political Biases of ChatGPT",*Human Behavior and Emerging Technologies*,2024, Article No. 7115633.

② Yupeng Chang et al. ,"A Survey on Evaluation of Large Language Models",*ACM Transactions on Intelligent Systems and Technology*, Vol. 15,No. 3,2023,p. 1;Jintang Xue et al. ,"Bias and Fairness in Chatbots: An Overview",*APSIPA Transactions on Signal and Information Processing*,Vol. 13,No. 2,2024.

③ Yufei Wang et al. ,"Aligning Large Language Models with Human: A Survey",arXiv,24 Jul. 2023,https://arxiv. org/abs/2307. 12966[2024 - 01 - 25]

④ Yuntao Bai et al. ,"Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback",arXiv,12 Apr. 2022,https://arxiv. org/abs/2204. 05862[2024 - 01 - 10]

⑤ 彭刚:《历史事实与历史解释——20 世纪西方史学理论视野下的考察》,《北京师范大学学报》2010 年第 2 期。

⑥ Zige Wang et al. ,"Data Management for Large Language Models: A Survey",arXiv,26 Dec. 2023,https://arxiv. org/abs/2312. 01700[2024 - 01 - 11]

⑦ Jiafei Duan et al. ,"A Survey of Embodied AI: From Simulators to Research Tasks",*IEEE Transactions on Emerging Topics in Computational Intelligence*,Vol. 6,No. 2,pp. 230 - 244;David Ha and Jürgen Schmidhuber,"Recurrent World Models Facilitate Policy Evolution",*Proceedings of the 32nd International Conference on Neural Information Processing Systems*,Neural Information Processing Systems Foundation, Inc. ,2018,p. 2455.

二、史学理论对大模型发展的价值

(一)考订辨正大模型的文本语料

相比历史学来说,当前大模型挖掘人类的海量语料这座富矿的手段依然不够精细,甚至可以说很是粗糙,在这方面,史学理论有关史料运用的深刻认识能够为大模型语料的去粗取精提供有益的指导。

大模型吸收综合文本应该遵循一些历史学关于史料运用的基本原则,而不能仅仅依赖缺乏坚实理论根基的工程技巧。这些原则大致有三。首先是"论从史出"。直白地说,就是推理结论始终要以史料作为依据,避免"以论带史"。大模型为什么会出现从史学角度看来啼笑皆非的"一眼假"幻觉?部分原因在于,吸收综合的过程中没有始终将其论述与文本出处相关联,没有始终将论述建立在文本的基础上。正因为如此,要求大模型注明文本语料来源才有助于提升其生成能力。① 其次,史料需要考订和分辨。显然,不是所有的史料都一样可信。有流传有序的史料,也有伪造假托的"史料";有第一手的史料,也有第二手甚至辗转更多手的史料;有秉笔直书的史料,也有隐笔、曲笔甚至有意无意漏载、误载的史料;有带着意图形成的史料,也有无意形成的史料;有早出的史料,也有晚出的史料;有同源的史料,也有源头彼此分化,可以彼此相互佐证或者订正的史料;等等。不是所有史料都一般可信,可以据而立论;也不是借助单一史料就足以定论,通常需要全面细致的考证。同理,语料文本也需要考订和分辨。对浩如烟海的驳杂文本,大模型基本上照单全收,目前来看还远远没有达到考订辨正"史料"的基本要求。秽史之所在,则谬论难免。这也很有可能是大模型幻觉的成因之一。最后,如果以更高的理论标准来要求大模型,那就不仅要明辨"史料"的是非正误,还要注意"史料"形成的源流。或如苗润博所论,不仅要"从文本中剥离出源自不同系统、不同时代、不同主体的历史叙述(文本单元)",还要"对其生成衍化过程加以剖析、对比,窥视以往被遮蔽、被掩藏的复杂图景"。② 这是比仔细分辨考证还要困难得多的要求了。

这些史料考订和运用的原则可以直截了当地转化为加工文本语料的原则。"论从史出"要求大模型的文本语料列明出处、来源有自;考订辨正要求大模型的语料文本去劣存精、区别利用;源流批判要求大模型对其海量语料加以程度更高的反思,将论证建立在对来源文本构造衍流的全过程的体悟之上。大模型文本语料的筛选和加工,因而需要相应加以改进。为了使文本语料来源有自,不能再杂乱无章地以自动化的方式一把抓,而是要对每个文本单元尽可能保留其源流信息,特别是作者、发表时间、引注信息、链接信息和版次更迭。为了使语料文本存精去芜,不能再将泥沙俱下的海量文本照单全收,而是要将优先采用可信史料的原则贯穿于语料加工和吸收综合的过程当中。多用真实,避免伪托;多用一手,避免辗转;多采直笔,细察曲笔;重视无意,精审有意;多用早出,慎用晚出;同源不赘,异源不遗等,都是大模型应当遵从的基本原则。为了让大模型具备批判源流的初步能力,在筛选加工优质语料的基础上,还要利用大模型对人类指令的理解能力,指示其在生成过程中充分

① Jie Huang and Kevin Chen-Chuan Chang,"Citation: A Key to Building Responsible and Accountable Large Language Models",arXiv,31 Mar. 2024,https://arxiv.org/abs/2307.02185[2024-04-06]

② 苗润博:《〈辽史〉探源》,中华书局2020年版,第374页。

考虑引证来源的衍变关系。也就是说,注重对文本本身的“结构、层次和缝隙”的把握,将整体的逻辑脉络作为生成的背景框架。① 这样,既使得大模型生成文本的理据更加深厚,又可以更加充分地暴露生成文本和复杂源流之间未必正确、全面、充分的相互关系,令其免于被轻信。

若要将以上理论落到实处,需要进一步论证如何实施在海量的文本上。用技术的概念来表达这一点,就是能够实现批量化(scalable)改进。② 如果不能满足这一点,提出的改进只能由历史学家在数以亿万字计的文本上逐一实施,最终只会是无法实现的美好想象。故此,在加工凝练文本语料的过程中,对来源有自、存精去芜和批判源流的原则性要求应当分别转译为既忠实于理论指导,又能够在海量文本上批量化实施的具体建议。或者说,转化为技术研发者熟悉且擅用的技巧。首先,为了实现批量化的来源有自,需要对自动化的抓取手段加以事前设计、事中监控和事后质检。事前在抓取代码中写入抓取源流的要求;事中监测文本语料中源流信息的缺失率,在信息缺失率较高时调整抓取方式;事后统计源流信息的缺失率,结合缺失情况设计数据筛选加工的策略。其次,为了实现批量化的存精去芜,需要将史料运用的原则转化为五点可批量化数据加工技巧。一是提高权威语料的权重;二是提高早出语料的权重;三是类似校勘中采择底本,将权威和早出语料作为语料加工的“底本”,或者说教材(textbook);③四是降低晚出、辗转甚至抄袭形成的语料文本的权重,或者加以去重;④五是移除伪造语料,或者设定专门的处理规则。最后,为了实现批量化的源流批判,需要通过系统提示(system prompt)的方式,让这一困难但有价值的思维方式成为大模型的默认设置。⑤也就是说,将明辨正误和剖析源衍的思维方式总结为简洁的一小段指令,然后以隐性的方式嵌入大模型的系统中,成为大模型每一次生成文本之前默认遵从的思维规则。这些都是能够批量化实施的改进。

(二)弥补大模型物质性和现场感的缺失

大模型能够凝练的范围始终只是经过数字化的文本。只能凝练文本的,是单一模态的大模型;能够凝练文本、图像和视频的,是多模态的大模型。尽管将基于转换器结构的多模态大模型加以拓展,对图像、气味、触感等不同形式的数据一并加以吸收的探索越来越多,也取得了一定进步,但所达到的水平和文本大模型仍然存在一定的差距。⑥ 即使假设这一差距未来得到一定的弥补——这是一个技术上非常雄心勃勃的假设,大模型很难完全解决史料物质性和历史现场感缺失对其能力的限制。

正如拓本不能完全替代碑刻,影印本不能完全替代原始的纸本史料,整理史料不能完全替代影印史料,大模型所依赖的文本语料也远远不能全然替代人类身体对原始物质史料的感受。具体而

① 苗润溥:《〈《辽史》探源〉题外话》。https://www.thepaper.cn/newsDetail_forward_22216665[2024-04-07]

② Edward Luke, “Defining and Measuring Scalability”, Mississippi State University and National Science Foundation (U.S.), ed., *Proceedings of the Scalable Parallel Libraries Conference*, IEEE Computer Society Press, 1994, pp. 183-186.

③ 有关数据加工中引入教材的方式和意义,参见 Suriya Gunasekar et al., “Textbooks Are All You Need”, arXiv, 2 Oct. 2023, https://arxiv.org/abs/2306.11644[2024-02-01]

④ Katherine Lee et al., “Deduplicating Training Data Makes Language Models Better”, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, 2022, p. 8424.

⑤ Anthropic, “How to Use System Prompts”, https://docs.anthropic.com/claude/docs/how-to-use-system-prompts[2024-04-07]

⑥ Peng Xu, Xiatian Zhu, and David A. Clifton., “Multimodal Learning with Transformers: A Survey”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 10, 2023, p. 12113.

言,大模型失却了所有通过观看、触摸、嗅闻乃至倾听史料才能得到的信息。纵然 Sora 等多模态大模型已经能够一定程度上模拟相当复杂的历史场景,在当前可见的将来,也很难充分再现物质性的各个感知侧面。① 举例来说,假如不只关注纸面上的权力制度,也关注其所嵌入的宫禁格局,不只关注碑刻文字,也在意其间造型美学,不只满足一重证据的内证,也求索二重证据间的彼此引证,这些历史学所熟稔的理论和手艺,大模型目前都做不到。

推而广之,凡是需要观看、触摸、嗅闻或倾听才能言尽其意的那些事,大模型都很难尽善尽美地做到。比如,大模型无法钩沉之前城市生活的风尘给人带来的"视觉、听觉、触觉、味觉与嗅觉皆有的"感受。② 进而言之,大模型不仅在当下损失了史料各个物质性侧面的信息,如果不加更新,还遗失了随着未来的技术发展从物质中提取更多信息的可能性。原始的物质史料不是"死"的,而是随着科技发展的手段不断呈现新的信息。譬如,在碳 - 14 技术和光学成像技术出现之后,既有的物质史料就表现出人类之前无法发现的新信息;在提取分析年轮、冰芯、孢粉、石笋和沉积岩土的技术出现之后,就有更多原本司空见惯的物质转化成为推动史学进展的新史料。大模型能够在事后反刍挖掘这些新的信息和新的史料,但很难在脱离了物质感受的前提下事前提出这些发现。总之,大模型对物质性的感知缺失导致了两方面的能力局限。一是依赖眼耳口鼻之感知的事大多做不到。二是用新技术检视老物件、发掘新史料的事大多也做不到。

即使假设未来发展出了全息影像一般的数字化技术,能够充分保存史料物质性的许多侧面,其始终难以完整还原史料所处的初始现场。即使更加强大的大模型能够吸收综合所有的这些侧面,其始终缺乏亲临历史现场的田野经验。计算机的针脚不像人的双脚,能够深深扎进现场的"泥巴"里。因此,有的材料大模型始终不可能看到,有的阐释材料的视角大模型始终无法想到,有的问题大模型始终难以完满回答。这些挑战,包括如下方面。首先,大模型缺乏行动能力。如历史学家科大卫(David Faure)关于做历史人类学研究的忠告:"这类材料在庙宇的墙壁上,在私人的收藏里多得是,除非你到现场考察,你是没有机会读到它们的。"③这句忠告原本的对象当然不是大模型,但用在没有双脚的大模型上,却十分地妥帖。问题越是接地气,越需要进入现场,大模型没有机会读到的信息就越多。其次,大模型无法感知"史料"在原始现场的存在方式,而这一点可能携带了相当关键的信息。例如,对遗址中的画像和物件的考察,除非掌握其"相互交错的关系和位置",所建立的理论"将很难具有坚实的基础";对礼仪制度的研究如果只关注文本,可以选择和回避不懂与不感兴趣的部分,特别是使用过程、实景布局和仪节实操中的细节,如果要求复原其现场,则"必须全过程、全方位展开……无一处可逃遁";即使对文学的研究,也只有重建和还原文学的历史现场,才能"透彻了解作者之用心,体会作品之奥秘"。④ 大模型没有现场的感受,也就无法沿着需要进入现场才能产生的研究视角去思考。当然,这一要求有可能是过分苛刻了,毕竟很多时候历史学家也未必能够充分地进

① 截至 2024 年 4 月 7 日,Sora 尚未对公众开放使用,因此还很难准确评估其能力是否像所宣称的那么强大。

② 邱仲麟:《风尘、街壤与气味:明清北京的生活环境与士人的帝都印象》,刘永华主编:《中国社会文化史读本》,北京大学出版社 2011 年版,第 433—434 页。

③ 科大卫:《历史人类学者走向田野要做什么》,程美宝译,《民俗研究》2016 年第 2 期。

④ 巫鸿:《武梁祠:中国古代画像艺术的思想性》,柳扬、岑河译,生活 · 读书 · 新知三联书店 2006 年版,第 81—82 页;彭林:《纸上得来终觉浅——为何〈仪礼〉需要做复原研究》,《光明日报》2020 年 12 月 16 日;方星移:《重建文学历史现场》,《光明日报》2020 年 7 月6 日。

入和体认现场。最后,回到历史现场和文献现场互证,常常是开创性的问题意识所不可或缺的来源。只有进入现场的学者才能理解:“置身于古人曾经生活与思想过的独特的历史文化氛围之中,常常会产生一种只可意会的文化体验,而这种体验又往往能带来更加接近历史实际和古人情感的新的学术思想。”①进而言之,“往往不是‘历史’使得我们更能认清‘现在’,倒是‘现在’常常使得我们更易于理解‘历史’”。② 失去了现场感,对文本的理解很可能流于浅薄。

由于大模型不能充分体认文本的物质性和现场感,其生成能力因而存在五道很难逾越的“天堑”。一是对文本的吸收综合始终无法达到全面。比如,从整理本到原刻本之间损失的信息,大模型同样会遗漏。二是不能将文本及其载体放置到原始的现场中去思考。凡是需要感受现场情状才能阐发的问题,都是大模型的弱项。三是不能主动用新技术、新方法和新视角审视其文本,始终需要人来代为弥补其思维时效性的缺乏。四是但凡需要身体进入现场细细体察才能引发的新思想,大多落在大模型的能力范围之外。五是无法超越这些数字文本,提出新的问题、视角和分析。或者说,大模型在根本意义上总是保守、狭隘的,总是拘泥于过去的文本里的一小部分。除非有朝一日人工智能真的科幻般地化身为人,这些弱点都是很难克服的。鉴于此,对于大模型的信息遗漏、现场缺失和时效缺乏的弱点,需要在使用时加以鉴别。如果对其期待很高,所问的问题需要利用这些信息和方法才能回答,那么不要轻信大模型。对于很难充分意识到这一点的用户,历史学者有必要和技术研发者合作加以提醒。

对于大模型根本意义上的保守和狭隘,不只是不能轻信,还要从为人类守护历史的角度加以警惕——不是只有能够数字化的才是历史。魏岑鲍姆 1976 年的疾呼至今振聋发聩:

> 计算机(人工智能)正在成为毁灭历史的工具。这是因为,当社会只承认那些“标准格式的”“能够很容易地告诉机器”的“数据”时,历史,以及记忆本身,也就此湮灭了……当然了,只有作为设定标准的机器的副产品的那些容易导出的数据,才会得到这个(机器)系统的承认。随着这个系统用户数量的上升,随着这些用户越来越依赖于“适合于这个系统的消息”……距离一切事实都由这个系统来判定,而其他所有的知识,所有的记忆,一切都被宣布为非法,还有多久呢……很快,一个超级系统就会建立起来,“历史学者”会根据这个系统来推断“究竟”发生过什么,历史上谁和谁之间存在联系,以及事件之间的“真实”逻辑。很多人到现在都没有发觉这里面究竟有什么不对劲。③

(三)感知和引导人工智能的价值倾向

如前文所述,大模型隐含了复杂的价值观念,甚至价值偏见。感知、测量和纠正大模型隐含的价值和偏见并不容易。技术研发者对文本语料中的价值观念的理解至今仍然过于单薄,很少具备史学

① 陈春声:《走向历史现场》,《读书》2006 年第 9 期;傅衣凌:《我是怎样研究中国社会经济史的?》,《文史哲》1983 年第 2 期。傅衣凌在这篇文章中提及进入现场所引发的问题意识:“1939 年我曾居住在永安的黄历乡,中间有一个很大的碉堡,四围则是一些矮小的平屋,佃户环之而居。我置身于这样的情景中,使我恍惚联想到中世纪的封建城堡制度,是不是还存在于今天的中国社会?那么,在社会史论战中所提出的那么多的问题,现在要不要重新加以检讨呢?”

② 胡宝国:《怀念周一良师》,《将无同:中古史研究论文集》,中华书局 2020 年版,第 388 页。

③ Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation*, p. 238.

所表现出的敏感性。亟待通过史学理论对文本史料的视角性和价值性的理解,丰富技术研发者的理解,并因此形成足以批量化引导对齐的原则和技巧。

自后现代主义兴起,史学经历"记忆的转向"开始,历史叙事的价值和视角始终是史学理论予以密切关注的议题。① 大模型广泛吸收综合语料所生成的文本,同样不能免于对其价值和视角的质疑。大模型生成文本蕴涵的价值和视角可以在三个层次上展开分析。首先,大模型所生成的文本在价值判断上展现出明确的取向。ChatGPT 等大模型能够很好地对齐一部分国家的文化,却又无法很好地和其他国家的文化保持一致。② 这些大模型在面对具体的价值判断时更多地遵循少数发达国家通行的观念,与大部分发展中国家的价值取向有着明显的距离。③ 大模型不止在国家和民族等维度上有所偏颇,在性别、职业等其他维度上也暴露出相当程度的偏见。④ 其次,除了价值取向和偏见的展现,大模型还隐含了特定的视角。视角未必直接蕴涵偏见,但在选择、编排乃至叙述其生成文本的过程当中,大模型完全有可能引入更加深刻,也更难察觉的偏见。⑤ 进而言之,如此所形成的文本完全有可能"布满权力的运作",建构和强化不平等的社会制度和文化。⑥ 对于历史上的重要事件,例如纳粹德国对犹太人的大屠杀,大模型有可能生成加害者视角的"支撑扭曲性或者否认主义叙事"的历史记忆。⑦ 这样的视角性既有可能深藏于大模型的内部,也有可能由用户有意或无意地引入。⑧ 这些视角在大模型中服务于什么样的权力结构,有待史学理论加以分析和甄别。最后,大模型接收的是文本,生成的也是文本。即使假设文本在其价值取向和叙事视角上都绝对地客观——实际并不可能做到,对于那些"在历史的文本中失势的人"来说,大模型本身就是不平等的来源。⑨ 过去积累的文本成为今天再生产权力的媒介。没有留下文本的人、不能正确使用文法的人、不能写出规范格式文本的人都在不平等的权力结构中处于弱势的地位。如果大模型听不进这些人的声音,也不能生成这些人所熟悉的语言,这又是对既有的不平等制度和文化的进一步建构和强化。

与前文对考订辨正大模型文本语料的要求类似,感知和引导大模型所隐含价值和偏见的要求,同样需要能够批量化地实施。当前,技术研发者实现对齐、纠正偏见的方法主要有来自人类反馈的

---

① 彭刚:《历史记忆与历史书写——史学理论视野下的"记忆的转向"》,《史学史研究》2014 年第 2 期。

② Yong Cao et al. ,"Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study",*Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*,Association for Computational Linguistics,2023,p. 53;Reem I. Masoud et al. , "Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions",arXiv,25 Aug. 2023,https://arxiv. org/abs/2309. 12342[2024 - 04 - 07]

③ Yan Tao et al. ,"Auditing and Mitigating Cultural Bias in LLMs",arXiv,23 Nov. 2023,https://arxiv. org/abs/2311. 14096[2024 - 01 - 17];Noam Benkler et al. ,"Assessing LLMs for Moral Value Pluralism",arXiv,8 Dec. 2023,https://arxiv. org/abs/2312. 10075[2024 - 04 - 07]得到相近结论的研究数量还在迅速增长。

④ Yupeng Chang et al. ,"A Survey on Evaluation of Large Language Models",arXiv,29 Dec. 2023,https://arxiv. org/abs/2307. 03109[2024 - 01 - 20];Jintang Xue et al. ,"Bias and Fairness in Chatbots: An Overview",arXiv,10 Dec. 2023,https://arxiv. org/abs/2309. 08836[2024 - 01 - 24]

⑤ 有关历史书写的"视角论",参见彭刚:《叙事的转向:当代西方史学理论的考察》,第 215—217 页。

⑥ 向燕南:《史的回顾与批判:传统历史书写中的女性与传统女性的历史书写》,《郑州大学学报》2014 年第 5 期。

⑦ Mykola Makhortykh et al. , "Shall Androids Dream of Genocides? How Generative AI Can Change the Future of Memorialization of Mass Atrocities",*Discover Artificial Intelligence*,Vol. 3,No. 28,2023.

⑧ Ameet Deshpande et al. , "Toxicity in ChatGPT: Analyzing Persona-Assigned Language Models", *Findings of the Association for Computational Linguistics:EMNLP 2023*,2023,p. 1236.

⑨ 有关"在历史的文本中失势的人",参见 He Xi and David Faure,eds. ,*The Fisher Folk of Late Imperial and Modern China:An Historical Anthropology of Boat-and-Shed Living*,Routledge,2016,p. 25。

强化学习(Reinforcement Learning from Human Feedback, RLHF)、直接偏好优化(Direct Preference Optimization,DPO)和更加前沿,很大程度上仍然处于探索阶段的弱到强泛化(Weak-to-Strong Generalization)三种。① 概言之,人类反馈的强化学习是用人的价值判断逐步引导和调整大模型,使其按照正确导向来行事;直接偏好优化则是直接将人的价值判断作为吸收综合过程当中的一个优化目标,大模型在学习语料文本的同时学习正确导向。二者在实践中各有优劣。无论采取何种方法,史学理论都可以更多地进入用来调整或者指导大模型的基准当中。史学理论有关文本如何隐含价值的认识不仅可以深化技术研发者对大模型价值对齐的认识,其来自人文视角的价值判断亦可补充和丰富技术研发者的判断,从而更有利于克服偏见。

三、史学理论与人工智能如何相互促进

人工智能的研发包括多个环节,这些环节共同构成了人工智能的一个完整周期。从最开始的数据加工,到之后的吸收综合,再到最后的价值对齐和风险监测,每一环节当中都有史学理论介入的空间。值得关注的是,像 OpenAI 这样的前沿技术研发者已经开始将史学理论运用到人工智能的研发和治理当中。由于这些实践很多只是刚刚开始,绝大部分信息还没有完全公开,暂时只能管中窥豹。集中体现这一点的是 OpenAI 在人工智能价值对齐方面的尝试。简单来说,OpenAI 和法国政治思想史学者海伦·兰德摩尔(Helene Landemore)和希腊城邦组织(Polis)等个人和组织合作,尝试从历史上的民主政治的哲学思辨和实践经验中,为今天的人工智能价值观对齐问题寻找解决方案。② 循此,从 2023 年下半年开始,越来越多来自不同国家和地区、学科领域与职业背景的人加入为大模型的行为写作“标准答案”的努力当中来。此外,OpenAI 针对一些具体问题上的研发思路同样体现出来自史学的影响。OpenAI 开设了长期运营的研究员项目,持续地吸纳包括史学在内的各个学科的学者,共同探索交叉角度的治理方案。③

中国的大模型研发和治理当然不能简单照抄照搬这些探索。不过,虽然我国的大模型和 OpenAI 的 ChatGPT 在具体的价值判断上不必强求一致,但这些为大模型融入史学理论的尝试有着很强的参考意义。如果要对错误的价值取向和技术的不当应用“进行实时、高效的智能回击”,在人工智能大模型的研发过程中充分取用历史资源、深挖史学理论、延纳史学人才等,都是值得探索的尝试。④

与此同时,史学理论要在任何一个环节中发挥作用,都离不开真实的历史学者与技术研发者之间的密切配合。在筛选加工文本语料的环节,不仅可以将考订辨正的原则转化为事前设计、事中监

① Yuntao Bai et al., “Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback”, arXiv, 12 Apr. 2022, https://arxiv.org/abs/2204.05862[2024-01-18]; Rafael Rafailov et al., “Direct Preference Optimization: Your Language Model is Secretly a Reward Model”, arXiv, 13 Dec. 2023, https://arxiv.org/abs/2305.18290[2024-01-20]; Colin Burns et al., “Weak-to-Strong Generalization: Eliciting Strong Capabilities with Weak Supervision”, arXiv, 14 Dec. 2023, https://arxiv.org/abs/2312.09390[2024-01-20]

② OpenAI, “Democratic Inputs to AI”, https://openai.com/blog/democratic-inputs-to-ai[2024-04-07]

③ OpenAI, “OpenAI Scholars 2018: Meet our Scholars”, https://openai.com/blog/openai-scholars-2018-meet-our-scholars[2024-04-07]

④ 陈甜:《十年来历史虚无主义演变的理性审思》,《史学理论研究》2023 年第 4 期。

控和事后质检,还要在出现没有覆盖在既有规则之内的情形时,及时处理例外情况和更新完善规则。在现实的工程研发中,即使提前制定的规则再完善,也必然存在数量巨大的例外情形,此时就需要历史学者和技术研发者实时进行合作判断。在为已经完成筛选加工的不同语料设定权重时,历史学者也可以和技术研发者共同摸索。权重的绝对大小和相对比例,数据"底本"的类型和范围,去重的判断条件和彻底程度,同样要在实践中共同一步步地试验出来。之后,大模型如果出现有待纠正的偏见问题,无论采取人类反馈的强化学习还是直接偏好优化的方法,历史学者和其他人文学者可以与技术研发者共同制定一套包含正确导向目标的"标准答案"。① 这套答案覆盖哪些价值维度,在每个价值维度上选择什么问题,答案是保持开放还是恪守标准,答案的表述怎样才最合适,同样需要历史学者的在场。

在促进人工智能发展的过程中,史学理论也将经历进一步的发展、深化与重塑。首先,大模型可以作为史学理论的实时"演武场",将史学理论在各个时代、各个地域、各个群体产生的史料上是否成立形成直观的展示。将理论运用于不同时代、不同地域、不同人群产生的海量史料之上,常常是焚膏继晷、煮海为盐般的任务。大模型汇聚了纵然不全面,但数量和范围足够广阔的语料。大模型的使用可以通过日常对话般的提示来进行,不一定需要数理的方法。即使需要运用数理的方法,大模型本身就是便捷易用的数理方法知识库。由此,通过将史学理论及其假设转译为准确、易用的自然语言提示,可以让大模型在数量和范围极尽广阔的文本史料上快速检验理论假设。尽管大模型内在机理不够透明的特性使其检验结果需要接受进一步的解释,但仍然可以为理论的好与坏提供快速的初步证据,形成初步的判断和筛选。

其次,对史学知识生产过程的理解将得到深化。大模型是史学理论的"演武场",也是史学知识生产的试验场。也就是说,对于既定的文本史料和史学知识,通过不断地用大模型去追问为何如此,或者为何并非如此,从而发生新的理解。大模型的照单全收是其局限,但在需要克服思维定势时,没有单一且牢固的前见,甚至是过度发散的"幻觉",却有可能发挥出乎意料的作用。历史学家的前见和不同来源的史料在一则具体的史学判断中发挥了哪些作用?如果引入不同的预设,或者以不同的方式比对权衡史料,判断的结论是否发生变化?既有的预设和史料是否遮蔽了容易遭到忽视的"无"——或者说某时代无某史事或无某观念的"默证",从而导致史学研究产生疏忽甚至盲点?② 史学的每个领域都可以用这样的问题去反复叩问大模型,从而在一则则"确实未必如此"当中找到理论得以深入的缝隙,推动史学研究的丰富和发展。恰如诸雨辰所述:大模型好比是《红楼梦》中的"风月宝鉴","正面去照可能平淡无奇,可是换一个视角,它确有可能照出文本源流中的某些特异点,发现新问题"。③

最后,将大模型作为全新的研究对象,史学理论因此会得到重塑。大模型不仅是史学研究的工具,也是理论研究的绝佳对象。吸纳了人类海量史料的大模型一定程度上就是一部人类的历史,甚至在未来可能成为每一个体接触历史最简便、也最常见的方式,史学理论当然应该将其作为研究对象。就研究主题而言,尽管计算史(history of computing)领域已经习惯了处理数字形式的史料,大模

① OpenAI,"GPT-4 Technical Report",arXiv,4 Mar. 2024,https://arxiv.org/abs/2303.08774[2024-04-07]

② 《胡宝国论学短札》,微信公众号"中华书局1912"。https://mp.weixin.qq.com/s/i5zidzGJiFb2f2zhWCZFUw[2024-04-07]

③ 诸雨辰:《自然语言处理与古代文学研究》,《文学遗产》2022年第6期。

型、参数和代码应当拓展成为更多史学领域的研究题目。[①] 就研究方法来说,既然技术研发者用以训练、调整和对齐大模型的理论和方法,很可能会形塑和改写对人类历史的认识,那么这些理论和方法应当成为史学理论研究的工具箱的一部分。在研究结论方面,随着研究主题的扩展和研究方法的吸纳,自然会引起更多不同形式的研究发现,这些发现将会辅助历史学者深度介入辅助大模型的研发。

结 语

史学和人工智能的交叉融合已经走过了接近八十年的历史。从最早加以尝试的罗伯特·布萨(Robert Busa),到今天遍布各个学科领域的交叉研究,很长一段时间内,史学是在人工智能技术浪潮的冲击之下被动响应,很少对技术的向善发展施加影响。[②] 这一单向的冲击—响应模式正在发生改变。史学理论对多元史料,特别是实物史料的关注,能够提醒我们注意单纯利用文本史料的大模型的能力局限;史学理论对史料如何隐含视角和导向的探讨,则能够警示我们注意大模型所蕴涵的价值和立场,甚至是偏见和错误。

展望将来,大模型将取得更长远的发展,甚至成为承载人类过往诞生的绝大部分文本所蕴涵知识的数字公地(digital commons),历史学者和技术研发者的双向互动既有必要,也是必须。[③] 当这一天到来,大模型将会成为每一个人接触"数据、记忆、关于过去的证据性遗迹、文献和遗物",从而形塑其历史认识的主要媒介。一方面,既然大模型可能成为历史的主要载体,成为绝大部分个体接触历史的主要途径,史学理论和历史学者理应深度介入其研发应用,确保其中存续的是丰富的、鲜活的、尽可能全面且客观的历史,而不是浅薄的、苍白的、保守狭隘且充斥偏见的历史。另一方面,既然史学理论意味着对从证据形成历史的自觉省思,当大模型成为历史乃至史学的主要媒介,对大模型的自觉省思甚至有可能成为史学理论研究的主体内容之一。这也意味着,利用史学理论改进大模型不只是求知驱动的科学探索,更是智识上和道义上的责任。

人工智能激荡整个人文学科的时代已经来临,需要历史学者和史学理论深度参与其中,贡献理论洞见。

(责任编辑:尹媛萍)
(责任校对:敖 凯)

① 有关将代码本身作为研究对象的著作,参见 Mark C. Marino, *Critical Code Studies*, MIT Press, 2020。

② Roberto Busa, "Half a Century of Literary Computing: Towards a New Philology", *Historical Social Research*, Vol. 17, No. 2, 1992, p. 124.

③ 有关数字公地这一概念,参见 Saffron Huang and Divya Siddarth, "Generative AI and the Digital Commons", arXiv, 20 Mar. 2023, https://arxiv.org/abs/2303.11074[2024-01-27]