出版与人工智能

□程三国

人工智能对出版业到底有什么影响？这些影响意味着什么？众说纷纭，大致有两种说法最有代表性。

一种是悲观派，认为人工智能生成内容带来的影响是全方位的，既是工具赋能，也是入口之争，最终将是范式变革，对于现有内容和出版产业，不是快慢问题，而是生死之争。另一种是乐观派，在他们看来，生成式人工智能给出版业内容生产带来巨大变革，改变内容生成方式，提高编辑工作效率，改变产品形态与服务模式，利于个性化推荐、精准营销、按需出版等。

这两种看法指向的是两种可能比较极端的情况，那么不极端的情况是怎样的呢？科技咨询公司盖特纳6月发布的一份报告指出，虽然已经开发出了许多令人信服的生成型人工智能产品的原型，但在实际应用中，采用这些技术的成功率却较低。高盛7月发布的一份研究报告显示，人工智能对所有工作任务的影响将不到5%，在未来10年内仅能提升美国生产力0.5%，推动美国GDP增长0.9%。智库兰德公司8月发布的一项研究显示，80%的人工智能项目以失败告终，这一比例是非人工智能项目的两倍多。

基于BOOKSGPT打造的AI编辑工作室，已在多家出版单位“上岗”。 本报记者 袁舒婕 摄

几个维度与框架——理解当下人工智能与出版

出版人对人工智能理解莫衷一是可以理解，因为人工智能是当今世界聚集最多风险资本和顶尖人才的热门前沿科技领域，技术每天都在升级和迭代，每天都有无数信息扑面而来，要想从中理出一个相对清晰的头绪不大容易。这里提供几个相对比较简单的框架和维度，希望为理解当下人工智能与出版提供一点基础和铺垫。

人工智能的五级水平。美国人工智能研究公司OpenAI的全新AGI路线图曝光，将人工智能按一到五的数字由低到高排列等级。一级：聊天机器人，能够使用自然语言进行对话的人工智能；二级：推理者，可以解决人类级别问题的人工智能；三级：智能主体，能够代表用户采取行动的人工智能；四级：创新者，能助力发明创新的人工智能；五级：组织者，能够执行复杂组织任务的人工智能。

OpenAI自称其已发布的大模型处于一级，很快会达到二级；五级的AGI最快也要到2027年才能实现。

全球最强的人工智能公司的人工智能能力尚处于初级阶段，照理其他的追随者应该不会更强，因此无论人工智能公司如何吹嘘其功能强大，人工智能当下能够做到的让人类信任的事情还十分有限。

数据与智能。生成式人工智能之所以现在爆发，是数据、算法、算力三者发展合力的结果。智能不再被视为机器本身的特性，而被认为是数据的特性。人工智能的突破性是大语言模型的智力涌现，其连续性在于数字化在线化数据的积累和发展。

相比较一些领先行业而言，出版业数字化程度整体不高也不均衡，数据在线化程度有限，因此限制人工智能开发、应用与能力发挥。在编印发产业链3个环节中，出版最高，印刷最低，发行居中。出版中专业出版数字化程度最高，如励讯集团数字化产品收入占比超过90%；大众出版最低，数字化产品收入最高的也就20%左右；教育出版数字化收入占比居中，最高如圣智集团超过70%，麦格劳·希尔集团和培生集团超过60%。数字化程度与人工智能应用潜力直接相关，励讯集团中国区高管张玉国说，励讯集团迄今推出了12款人工智能应用。

液体与容器。纸质图书的数字化转型第一轮高潮是2007年亚马逊发布Kindle带动的。如果说十几年前的出版数字化是把实体的出版物变成了流动的水，但水依然还都装在各自隔开的容器、池子或水库里，如文本库、图画库、音频库、视频库，各个数据库互相之间不能通融，微弱的联通是文本中加个二维码扫描出音频或视频。而这次人工智能是消除或消融了容器或格挡，比如文本可以一键生成图画、文本一键生成音频视频、音频视频一键生成文本等。人工智能生成内容等信息技术正在重塑信息内容容器、文字容器、音频容器、视频容器，容器的功能隔阂被打破，出版业进入“活容器”时代。因此，有人说生成式人工智能之后的时代是新航海时代，也意味着这次出版业面临的挑战可能比以往任何时候都要大。

自然交互和知识计算。搜狗搜索和百川智能创始人王小川认为，人工智能有两大方向，知识计算和自然交互。知识计算代表智能，其核心的智能来自于语言；自然交互是使得人跟它有更好的接触，包括语音、图像、视频。

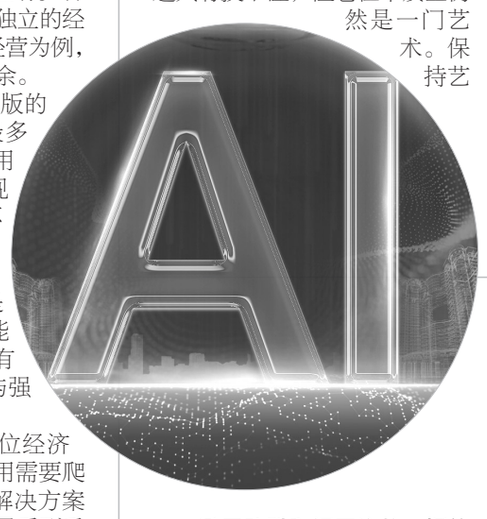
在大语言模型中，大家过分关注第一个关键词“大”，往往忽略第二个关键词“语言”，其实，语言在中间扮演了通向智能最重要的角色，当下机器学习主要还是在充满人类智慧的语言里学习。这意味着，以语言文字系统记录人类智慧的图书，在人工智能大语言模型训练中的价值和可能的市场空间要比我们想象的大得多。

发现与生成。虽然大语言模型涌现的智能被叫作生成式人工智能，但真正用起来其实有点名不副实，如生成文本时，时常产生“幻觉”，会“一本正经地胡说八道”；生成图画时，无标准的想象力发散有余，但有标准的精确性和连贯性却很差。但作为数据洞察工具，它不仅特别靠谱还让人惊艳不断，比如人工智能可以发现蛋白质结构和基因系列，攻克生物领域50年难题，又比如人工智能能够发现新的抗生素药物。因此有人说，生成式人工智能“应该当作一个发现的工具，而不是一个生成的引擎”。

出版业与其说需要发现，不如说更需要高水平系统化知识的生成，而且出版业是一个对编辑出版质量的准确性、连贯性、稳定性有异常高标准和高要求的行业，人工智能算法本身具有概率性，每个人工智能模型都包含一定程度的随机性和不确定性，因此对于目前主要基于概率、其结果不是很稳定的生成式人工智能，要让其深度参与出版流程和输出达标结果，编辑和出版人还是比较慎重的。由此看来，人工智能当下显现出来的特长、能力与出版业特有需求并不是那么匹配。

功能、产品和生意。功能、产品和生意三者虽然在数字化环境里边界时常模糊，但还是有很大区别。功能是现有产品或产品套件的补充。功能可能不会增加任何收入，事实上可能只会增加成本。产品虽然会产生收入和与之相关的成本，但它可能不足以支撑整个生意，一个成功的生意常常拥有一系列产品组合，就像一个出版社往往有多种图书和书系。生意是一种独立的经济存在，以一个出版社的经营为例，它有收入、支出、利润或盈余。

从现有人工智能在出版的应用看，专业出版场景最多的是搜索，大众出版中应用最多的是营销，大多是对现有功能的强化和支持，不能构成独立的产品，形成独立的营收，更不是一个独立完整的生意，即便是励讯集团的12款人工智能应用，大多也主要是对现有产品服务和工具的优化与强化，增加专业用户认可度。



3个阶梯。加拿大3位经济学家研究发现，新技术应用需要爬过3个阶梯，首先是从点解决方案开始，再到应用解决方案，最后到系统解决方案，才能真正使用起来。人工智能作为一种新技术，其应用恐怕也需要爬过这3个阶梯，从人工智能点解决方案，即用人工智能改善现有的决策；到人工智能应用解决方案，即用人工智能改变做决策的方式；再到人工智能系统解决方案，即人工智能促成了新的决策，整个生成模式决策方式发生了改变。

预测与判断。人工智能是一种预测技术，预测是一个填补缺失信息的过程。预测并不是决策，它只是决策的组成部分。预测通过减少不确定性使人们做出决策更加便利，但判断的作用则在于分配价值。人工智能作为预测机器并不提供判断，只有人类才会进行判断，因为只有人类可以表达不同行动带来的相对回报。随着人工智能接管预测，人类会减少在决策中扮演预测加判断的综合角色，而更多地专注于发挥判断的作用。

几个现实推论——提示人工智能与出版的发力方向

依据上述几个框架对人工智能与出版的理解，结合历史上出版与技术的几次相遇，可以得出以下几个初步推论，或许可以提示出版机构在人工智能上探寻发力的方向。

人工智能是出版数字化转型的延长线和深水区。传统出版如果转型到位，对传统出版业的内容生产方式、出版方式、管理方式和经营方式等产业全流程就会带来深刻变革。在这个基础上，人工智能等新型劳动工具的出现会带来革命性突破，数据成为新的生产要素和劳动对象，传统劳动者升级为数字化、网络化、智能化劳动者，出版业的全要素生产率也将大幅提升。

一个出版企业，如果没有良好的数字化基础，只希望借助人工智能弯道超车或一飞冲天，大概率不会有理想的结果。人工智能只是对有准备的、数字化基础好的出版企业锦上添花，绝不会对没有准备的、数字化基础薄弱的企业雪中送炭，尽管人工智能一定程度上也可以为出版数字化转型赋能，但自身的数字化基础是前提。

人工智能提示我们，出版企业需要继续加快数字化转型各项工作，包括产品数字化、流程数字化和营销数字化，而且与之前不同的是，需要借助人工智能来倒逼升级数字化转型，加快速度提升标准，不然人工智能大潮不仅不能借力，很可能被甩开。

技术与内容孰轻孰重的逻辑依然没有变：内容是目的，技术是手段。传统出版每次遇到新技术，都会引发技术和内容孰轻孰重之争，这次也不例外。出版本质上是靠卖内容赚钱，而不是靠卖技术赚钱。普林斯顿大学出版社原社长彼得·J·多尔蒂说，图书出版业是一种商业，而且越来越具有技术性，但它在本质上仍然是一门艺术。保持艺术的中心地位，同时运用科学来支持和推进它，是出版业高管的任务，也是维护声誉的本质。

从电子书胜出者Kindle和许多模仿者、失败者的比较来看，内容版权优先还是内容技术优先，无论对出版公司还是技术公司，不同选择导致不同结果。成功的案例是Kindle，Kindle上市前花的最大功夫是签下美国主流出版社主要在售和即将上市的图书，而且价格是精装本纸质书的三分之一，Kindle胜出是内容优先的结果。

技术推动出版发展不假，但并非技术越多就越好，出版技术含量与其商业成果不成正比。回看上一轮电子书革命，一个有趣的事实值得我们深思：内容含量越高，商业价值越大，如大众出版中技术含量最低的电子书1.0，即转档电子书卖得最好；相反单品技术含量最高、技术投入最大的电子书3.0，即增强型电子书，鲜有成功案例。技术含量越高，价值未必越大。

版权已经成为人工智能与出版各方竞争角逐的新阵地，集体版权与孤儿版权将再度引关注。无论是纸质出版时代还是数字出版时代，版权都是最核心最关键的，所谓“内容为王”，其内在本质是“版权为王”，在人工智能时代依然如此。

其一，一个鼓舞人心的迹象，新闻和专业内容版权率先实现规模级市场价值。

版权是内容行业能够立竿见影带来收入的“低垂的果实”，人工智能合作伙伴关系正成为与传统广告和订阅并列的新收入来源。领先的内容公司已经陆续成为人工智能技术公司的核心内容供应商。生成式人工智能公司需要最高质量的内容，以确保其产品的相关性和准确性。今年2月，谷歌与新闻集团签约，每年付近6000万美元，获得授权使用其新闻内容用于大模型训练。据估计，这项合作在未来5年内价值高达2.5亿美元。

OpenAI也不甘落后，浏览OpenAI合作伙伴名单，数量最多的竟然是内容公司，许多全球知名内容品牌赫然在列。除了新闻内容交易，最近陆续披露出许多专业出版内容交易，比如英国学术出版公司泰勒&弗朗西斯和微软的内容合作，每年付费800万英镑；威利已经和人工智能技术公司达成两笔总额超过4000多万美元的内容授权生意；剑桥大学出版社、牛津大学出版社透露将向人工智能公司出售其作者作品的访问权，赚取收入达5800万英镑。

其二，图书版权，特别是文史哲图书版权将成为下一个版权价值洼地。

内容公司版权内容变现的路径之所以从新闻内容到学术内容，最后再到大众图书，其底层逻辑在于：新闻内容对读者订户而言，只有最新的才最有价值，过时新闻积累而成的新闻内容库打包卖给人工智能公司可以说是白赚；学术出版内容经过20年的开放存取运动，大约一半对读者是免费的，出版商通过向作者收取文章处理费已经收入不菲，因此其内容库授权给人工智能公司并无太多顾虑；大众图书则不然，主要靠向读者收费，因此在内容授权上特别慎重。

虽然现在还没有看到大众图书授权给人工智能技术公司形成有规模收入的消息，但网上一度传过Meta公司想收购美国西蒙&舒斯特出版公司，因为其积累了20多万种优质版权图书。最近美国也出现了一家创业公司对此跃跃欲试。说明各方都很看好这个市场。

中国出版协会理事长邬书林在前不久的一个出版论坛上提醒出版界：“人工智能虽然可以高效地处理大量数据，但在情感表达、文化理解等方面还存在局限性。因此，我们需要思考如何在人工智能的辅助下，保持出版的独特性和人文性，让出版物更具深度和温度。”纯理工知识是形式逻辑，是可以编程的，可以交给外部系统处理。如果说形式逻辑知识像法宝，人工智能流行之后很容易获得，文史哲知识更像内功，需要长时间不断修炼。这意味着，大模型需要特别注重文史哲知识训练，文史哲训练不足的大模型会充满偏见。

现在大模型所训练学习的内容中，深度系统的文史哲知识严重不足，因为文史哲知识主要在图书里，图书的数字化程度整体不高，授权出去的优质内容非常有限，头部新书和再版书是出版社的“现金牛”，一般不会轻易授权，再版长尾图书许多难以找到数字版，更多的孤儿图书根本没有数字版，雄心万丈的谷歌图书数字化计划涉及1.3亿种图书，据说也只是人类所有图书品种的十分之一。

因此，传统出版界未来一项重要工作就是继续将优质图书特别是文史哲长尾图书数字化，因为图书品种多，许多是不再版且版权难以确认的孤儿作品，全世界解决这一版权难题的路径是集体版权管理，包括如何给版权定价，如何补偿作者都是需要解决的问题。

徘徊不前的专业数据库工作找到新动力与新方向。在专业出版数字化转型浪潮中，尽管中国出版界有同方知网、科学出版社的期刊数据库以及社会科学文献出版社的皮书数据库等亮眼成果，但与世界领先的数据库相比差距依然很大。许多立项并投入很多资金的专业数据库面临的问题，不是找不到应用场景就是找不到商业模式。

人工智能的开发与应用为这些专业数据库找到新的应用场景。人工智能的应用特别是在专有领域中的应用，需要跟领域知识深度结合，训练专有小模型也需要专业数据库。因此说人工智能为原来停滞徘徊的专业数据库工作找到新的动力和方向。

这里需要指出的是专业数据库，小不是关键，主题关联性强才是关键。无论数据集的大小如何，数据的相关性才是真正重要的。

什么是相关性的数据？比如数学知识，只有自己出版社出版的少量品种不行，不说把全世界至少也是全中国有代表性的数学图书和文献差不多都装进来才行。比如要训练人工智能会写像样的图书营销文案，至少也得收集国内几十万或几百万种图书的营销文案数据集，这才是相关性数据集。

几个建议——守护和挖掘出版的内容价值

避开陷阱和摆正姿势。据估计，超过80%的人工智能项目以失败告终，兰德公司最新调查报告，人工智能项目失败有五大原因：第一，不清楚需要用人工智能解决的问题是什么；第二，企业往往缺乏足够的数据来训练有效的人工智能模型；第三，企业更注重使用最新的最先进的技术，而不是为预期用户解决实际问题；第四，企业通常没有足够的基础设施来管理数据并部署已完成的人工智能模型；第五，企业将技术应用于人工智能无法解决的问题。这是企业应用人工智能应该留意的陷阱。兰德公司报告建议，企业应用人工智能，应该确保技术人员了解项目的目的和领域背景。在开始任何人工智能项目之前，领导者应该做好准备，让每个产品团队至少在一年内专注于解决特定问题；应该关注问题本身，而不是技术；应该投资于基础设施，为支持数据治理和模型部署而进行的前期基础设施投资，可以减少完成人工智能项目所需的时间，并增加可用于训练有效人工智能模型的高质量数据的数量；应该了解人工智能的局限性，在考虑潜在的人工智能项目时，领导者需要邀请技术专家来评估项目的可行性。这或许是任何企业应用人工智能需要摆正的姿势。

行业层面。无论是发达国家的出版协会还是国际出版协会，在谈到人工智能与出版时，首先说到的是版权保护，强调不能任由优质出版内容未经授权就被人工智能公司拿去训练大模型，需要保障作者和版权人从人工智能的发展中获取应得的利益。

创作编辑出版优质内容有很高的成本，背后是无数专业人士的多年积累与付出。无论是国家还是机构层面，人工智能研究既然有巨额投入，就应该支付优质内容成本，这些事情当然要政府主管部门或行业协会统筹协调，才能够解决有望。中国是参照欧盟还是美国，也需要权衡，在法律出来之前，出版机构可以从合同层面先行改善优化。

中国出版界有各级出版基金，用来扶持耗时长和投入大的出版项目，经过多年发展，成果蔚为大观，许多都是文史哲的优质内容，只是大多数只有纸质版，如果能够进行数字化转换，应该能够形成许多优质的文史哲专业数据库。新的基金项目成果提交也可以考虑加上数字版或数据库要求。

考虑到文史哲图书数据库建设严重滞后，数字化建设成本和孤儿版权处理成本巨大，可以考虑设立专项基金来推进。外文世界是谷歌以一己之力斥巨资投入建设的，中文世界也需要类似的图书数字化计划。

企业层面。当然，更多的专业数据库，需要企业自主投入建立，现在出版企业销售规模最大、利润最多的大多是地方出版集团，地方出版集团恰恰是专业资源最薄弱的出版机构，为了建设系列高水准专业数据库，出版机构有3个思路可以考虑。

其一，独立数据库。自己有足够的专业资源优势，自成一体，就可以独立建数据库，这样的情形应该不多。

其二，联合数据库。就是多家企业合作，突出优势，资源置换，最后形成多个专业资源库。

其三，联邦数据库，也叫联邦学习。这是一个折中的策略，就是各自数据都在自己的库里，但使用时可以在不影响版权和隐私的前提下以专题来集中使用。

联邦学习是一种允许多个参与方在保护数据隐私的前提下，共同训练模型的分布式机器学习方法。它的发展对数据产业产生了重大影响，尤其是在金融等数据敏感性较高的领域。联邦学习可以解决数据孤岛问题，同时保护数据隐私和安全，符合如欧盟通用数据保护条例等法律法规的要求。

这个思路出版业也许可以参考，毕竟是最快而且最折中的解决方案。

（本文是百道新出版研究院院长程三国在浙江出版集团数融发展大会上，题为《出版集团AI与融合出版战略》的演讲中，“出版集团AI战略”部分的整理文字稿，引文与出处有删节。）