何时真正迈入“人工智能治理立法时刻”

季卫东

- 正是因为科技和产业的发展与人类社会的安全之间的张力变得空前巨大，不同需求尚未找到适当的平衡点或妥协点，立法机关采取了更慎重的态度。
- 迄今为止，各国的制度设计大致可以分为四种模式，即“硬法模式”“软法模式”“软硬兼施法模式”“技术程序法模式”。中国的经验和机制设计构成“软硬兼施法模式”。
- 只有在语言大模型和多模态大模型的性能与安全度提升形成某种正比关系，而监管转变为并非拘泥于事先明文规定的程序本位和技术本位之际，各国以及全球才有可能真正进入所谓“人工智能治理的立法时刻”。

发展与安全，价值判断尚未形成共识

2023年6月，国务院把人工智能法草案列入2023年度立法工作计划。在此背景下，同年7月，网信办等七部门联合发布《生成式人工智能服务管理暂行办法》。这是世界上第一部关于大模型和AIGC的法规。紧接着，中国社会科学院法学研究所在8月提出了《人工智能法示范法1.0（专家建议稿）》。但是，全国人大常委会9月发布的今后五年立法规划中却没有提及人工智能法。这种显著的温差表明：自从ChatGPT横空出世，人工智能的飞速迭代使我们迅速滑入人机并存的未知混沌之中，科技和产业的发展与人类社会的安全之间的张力变得空前巨大，不同需求尚未找到适当的平衡点或妥协点，立法机关采取了更慎重的态度。今年3月，中国政法大学等七所大学的研究者共同发布《人工智能法（学者建议稿）》，试图掀起新一波热潮，促进立法进程。

实际上，正是因为对发展与安全之间的价值判断和公共选择还没有形成共识，欧洲议会在2023年6月通过《人工智能法》谈判授权草案之后，曾有包括西门子和空客在内的150多家科技企业联名发布公开抵制信，指出法案规定过于严苛的监管机制会大幅度压缩科技创新空间，延误欧盟数字经济的发展；人工智能的治理和立法都应该加强相关行业的参与。11月发生的OpenAI高层戏剧性政变，也充分反映了“发展优先”与“安全优先”两条路线之间的激烈斗争。尽管如此，2024年3月欧洲议会还是以高票正式通过了《人工智能法》。这意味着欧盟的立法机构确立了安全优先于发展的价值排序，要求成员国加强对科技前沿拓展和应用场景的监管，同时也试图对域内的数字空间实行保护主义政策。

几乎与此同时，三十几位海内外技术专家和企业领袖在中国签署了《北京AI安全国际共识》，为人工智能研发划出几条明确的红线，并试图以此为基础形成国际合作机制。北京共识的主要内容包括确保人类对AI系统复制和迭代的控制、反对大规模自动化武器的设计、导入国家注册制以便按照全球对齐的要求加强监管和进行国际审计、防止最危险技术的扩散、开发全面的治理方法和技术、建立更强大的全球安全保障网络。非常有趣而又耐人寻味的对照是，当欧洲科技企业担心偏重监管的欧盟《人工智能法》势必吞噬占投资17%的AI产业发展之际，《北京AI安全国际共识》却呼吁各国政府和企业把AI研发预算的三分之一投到安全保障领域。“33%vs.17%”的成本效益竞争，仿佛构成规则制定话语权的崭新制高点。

AI治理的立法时刻及监管优先的取向

如果说2024年早春的文生视频大模型Sora和文本超长大模型Gemini1.5 Pro标志着现象学意义上的“天人合一”奇点，那么特斯拉斯堡议员的压倒性表决结果和北京科技产业界国际共识就共同标志着“人工智能治理的立法时刻”。为此，有必要审视各国与人工智能相关的法规制定和实施现状。迄今为止，各国的制度设计大致可以分为四种模式，即“硬法模式”“软法模式”“软硬兼施法模式”“技术程序法模式”。

不言而喻，刚刚通过的欧盟《人工智能法》代表了硬法模式，其基本特征是监管高于研发。这种立法的理念可以追溯到艾萨克·阿西莫夫在1950年出版的科幻小说短篇集的导言里提出的机器人三大定律，而新近的学术表达则是W.瓦拉赫和C.艾伦在《道德机器》（牛津大学出版社，2009年）以及M.安德森和S.L.安德森在《机器道德》（剑桥大学出版社，2011年）中的基本主张：1.应为机器人设立伦理标准，包括AI自身道德能力的评价标准和在多种伦理规范发生冲突时进行价值排序和选择的道德原则；2.机器人的自由度越大，相应的AI伦理标准也就应该越严格。正是沿着人工智能性能或自由度与伦理标准或监管力度之间形成正比例关系的思路，欧盟《人工智能法》把AI风险区分为不可接受、高、有限、最小这四个等级，分别规定了不同的规范方式。特别值得重视的是，该法认为极其有害和有违欧洲价值观、一概属于禁止范畴的AI应用还包括通过社会评分系统对个人行为的操纵、实时远程的生物特征识别技术的应用、预测性警务系统的导入。另外，协助法官和律师的法律专家系统以及智能审判项目也被认定为高风险类型，需要重点监管。

不同软法形态及其和硬法的组合

与此形成对照的是美国在2020年制定的《国家人工智能倡议法》，旨在统筹协调以加速人工智能的研发和应用、促进经济繁荣和国家安全、审视AI治理的路径、平衡个体权利与科技创新之间的关系，属于典型的软法模式。2022年10月白宫科技政策办公室发布的《AI权利法案蓝图》、2023年1月美国国家标准和技术研究机构发布的AI风险管理框架等，也都是原则和政策的宣示。拜登政府在2023年7月与人工智能领域的七家头部企业达成关于AI研发和应用的安全自愿承诺，同样不具有任何法律约束力。目前，美国多个州已经制定AI法规，例如2024年3月犹他州的《人工智能政策法》，在规定行政和民事罚款作为制裁措施的同时，为AI创新提供监管缓解的技术豁免权；其他30多个州也有审议AI法案的动向。这些分散式立法的内容，大都侧重数据的隐私保护、算法的可解释性、防止AI歧视以及保护消费者权益等具体问题。美国联邦国会议员的立法提案立场各异，虽然已经通过的都是促进AI产业发展、确保美国科技领先地位的法案，但是，参议员伊薇特·克拉克等提出的《2019年算法问责法案》及其2022年、2023年的更新版本则具有明显的硬法化倾向；众议院能源和商业委员会在2023年7月通过的《人工智能问责法案》则给政府施压，希望在2025年之后采取实质性问责措施来防范AI风险。

一般而言，从P.塞尔兹尼克和P.诺内特关于“回应型法”的构想，转向政府的“敏捷治理（Agile Governance）”方式，规制的主体和规范显然都变得更加多样化。2018年达沃斯世界经济论坛对敏捷治理的定义是“一套具有柔韧性、流动性、灵活性或适应性的行动或方法，是一种自我调适、以人为本以及具有包容性和可持续性的决策过程”。日本政府在人工智能的风险管控方面高度重视敏捷治理，形成了一整套灵活机动而又具体细致的AI研发和应用的行为标准和操作流程。今年4月，日本总务省和经济产业省联合发布的《人工智能产业指引（第1.0版）》中具体界定了利益相关各方进行风险管控的主体责任和敏捷治理的机制设计蓝图，进一步呈现出导向性规范与行政助推措施相结合的软法特征。

中国在2019年率先提倡人工智能的敏捷治理原则，侧重软法模式。但在实践层面主要流于因地制宜的行政裁量权行使，缺乏一套操作自如的规则和弹性结构的制度安排。这样做的好处是可以自上而下随机应变，交替使用软硬两种手段进行风险管理。2021年9月，国家新一代人工智能治理专业委员会发布了《新一代人工智能伦理规范》，把AI全生命周期各环节的责任审查和问责机制作为基本原则之一。2022年3月，为了落实2021年底刚颁布的《互联网信息服务算法推荐管理规定》，“互联网信息服务算法备案系统”正式上线运行，对人工智能的监管形成了算法备案、检查、问责“三位一体”的基本框架。特别是2021年算法滥用治理专项行动和2022年算法综合治理专项行动严厉惩治了“大数据杀熟”“二选一平台垄断”等违规现象，显示了监管的硬派一面。2022年12月颁布的《互联网信息服务深度合成管理规定》开始剑指生成式AI。因此，中国的经验和机制设计构成“软硬兼施法模式”。

技术-程序的路径通往科技企业蓝海

在贯彻人工智能治理的原则和政策方面，新加坡走了一条低调务实的技术路线。2022年5月，该国政府推出全球首个人工智能治理开源测试工具箱“AI.Verify”，把测试和检查融合在一起，通过安全、灵活、透明化、可审计、可问责、互相制衡的动态调整程序达到可信AI的目标。这种测试框架把安全监管与人工智能自身性能提升密切结合在一起，应用基于自愿，因而有利于提高人工智能产品以及监管的接受度。“AI.Verify”又是一个技术性手段的百宝箱，针对不同行业、不同应用和不同产品形态分别制定了有针对性的监管测试方案，并开发相应的测试工具集和数据集。例如在自动驾驶等领域，人工智能应用风险性更大，因而需要采取更加严格、强制性的监管测试举措。这种“技术程序法模式”具有普遍推广可能性。而IBM在2023年12月推出的人工智能治理模型“WatsonX.Governance”把风险防控与自动化的监管工具开发密切结合在一起，可以按照人工智能法规和政策提供AI的“营养标签”、实现主动检测和减少偏差的LLM指标，与“AI.Verify”有异曲同工之妙。此外，还有加强AI自监督学习能力的图像世界模型（IWM），也可以发挥类似的控制作用。

正是在这里，立法者如何在人工智能发展与社会系统安全之间保持适当平衡、相关行政部门如何进行敏捷治理的答案似乎已隐约可见。如果大模型的技术研发不仅是AI治理的对象，也可以反过来为AI治理赋能，科技企业就不会对人工智能立法感到忧心忡忡。如果大模型的安全研究通过技术-程序的进路能够形成测试、评估以及监控的工具箱，其中包括推广数字水印技术、开发AI验证小模型、形成AIGC打假系统、建立AI伦理管理指标体系和认证平台、编制AI安全保障网，那么监管与发展就不再是一种零和游戏，AI治理还能为AI研发开拓出新的投资契机或者市场空间，通过错位竞争构成企业的科技蓝海。换言之，只有在语言大模型和多模态大模型的性能与安全度提升形成某种正比关系，而监管转变为并非拘泥于事先明文规定的程序本位和技术本位之际，各国以及全球才有可能真正进入所谓“人工智能治理的立法时刻”。

（作者系上海交通大学文科资深教授、博导、中国法与社会研究院院长、人工智能治理与法律研究中心主任，中国计算机学会计算法学分会会长，中国法学会法学教育研究会副会长）