“AI竞赛”加剧安全风险需要分层治理

□张 平

今年3月，又一则关于AI安全与治理的行业共识达成，包括图灵奖得主约书亚·本吉奥、杰弗里·辛顿、姚期智等在内的数十位的中外专家在北京联合签署了《北京AI安全国际共识》（以下简称“《北京共识》”）。自2022年末，以ChatGPT为代表的生成式AI 技术快速崛起，掀起产业界的新一轮科技革命的同时，也引发了全球范围对于其安全风险的深切担忧。2023年3月，美国生命未来研究所（Future of Life）发布了一封《暂停大型人工智能研究》的公开信，呼吁所有AI实验室立即暂停比ChatGPT 4.0更强大的AI系统的训练，暂停时间至少为6个月，以呼吁在此期间各国能共同制定全球范围内相关的技术共享安全协议。此后的一年时间中，世界各国政府、学界、企业界均从各自角度出发提出应对AI安全风险的各类解决方案。无论是联合国最新通过的首个关于AI的全球决议，抑或28个国家及欧盟共同签署的《布莱切利宣言》、我国提出的《全球人工智能治理倡议》，以及欧盟最新立法《欧盟人工智能法案》（草案），均强调AI的治理应当重视安全风险，无底线运用AI将会给人类社会带来毁灭性危机。

AI安全风险治理是一项综合性工作，应当分层级、分领域、分阶段、分重点开展。具体而言，“人类安全——国家安全——个体安全”这三个维度的潜在安全风险都不容忽视，应予重点关注，三层风险的解决方案也应“对症下药”。

“人类安全”即从全人类角度出发，强调AI技术应以尊重人类权益为前提，其目标应是为人类服务而非取代甚至危害人类。其核心要点即AI技术应当“以人为本”和“科技向善”，打造“可信”人工智能（Trustworthy AI）。这其中主要涉及AI相关伦理风险，如深度伪造（Deepfake）、真实性与准确性、歧视与偏见、AI欺诈等。而这些问题的解决之道主要涉及两方面措施，不仅在AI研发开始就需要“人类干预”和“人类监督”，也即《北京共识》中所强调的“任何AI系统都不应在人类没有明确批准和协助的情况下复制或改进自身”；另一方面，还需保证相关“人类参与”是“善意和正确的”，即需要保障AI研发过程中的透明度和监督机制。AI犹如利刃，若是被邪恶或带有毁灭人类意图的人利用干预，则会成为“杀伤性武器”。故而，AI的发展首先要进行科技伦理审查和承诺，这不仅需要来自技术领域的专家参与，还需要社会科学、人文科学等多领域专家的共同治理和监督。实践中，监管机构可以通过设立标准化评测平台、评估框架或采用沙盒测试等进行AI技术的监管。

“国家安全”即从各国角度出发，保障各国在这场“AI竞赛”中权利平等、机会平等、规则平等，弥合智能鸿沟和治理能力差距。其重点是保障各国在AI的发展和治理过程中拥有平等的权利，不会被“智能鸿沟”淘汰。首先应认识到，不同国家有其不同的立场和利益考量。例如，美国作为已拥有先进AI产品的“领先者”，其目的是确保不丧失领先地位。但对于大多数发展中国家而言，“国家安全”的重点是“防守”，即不被数字鸿沟所淘汰。因而，在当前发展不平衡的格局之下，有必要倡导“人类命运共同体”的理念，应当从地球村的利益出发，本着互助互利精神给欠发达国家留有一定的发展空间。

“个体安全”则是从宏观视角的群体概念转向关注AI对个体权益的影响。其中最突出的问题便是AI应用对个人隐私安全的影响，AI训练数据对个人知识产权的影响、AI欺诈对个人财产安全的影响、AI替代人类部分工作引发的群体性失业风险等。例如，作者们已经在全球范围内开始了对Open AI、Stability AI、DeviantArt和MidJourney等生成式AI服务提供者的诉讼，指控其训练数据侵害了他们的版权。目前各方主要关注的还是前两层风险，而对于“个体安全”则需要各国监管机构、AI服务提供者、AI服务使用者多方合力共治。例如，微软公司已在其Azure AI Studio中推出新工具，并且在AI Studio中推出了AI辅助安全评估、风险和安全监控功能，旨在加强AI模型的安全性。

数据显示，在2023年，基于AI的深度伪造欺诈暴增了3000%，基于AI的钓鱼邮件数量增长了1000%，已有多个有国家背景的黑客组织利用AI实施了十余起网络攻击事件。生成式AI的安全威胁正在急剧增长。而已经出现的大模型幻觉问题，随着信息智能大规模应用延伸至物理智能、生物智能时，风险也会规模化叠加。故而，无论是《暂停大型人工智能研究》还是《北京共识》，这些来自科学家的呼吁虽然不具有法律强制性，但却为世界敲响了警钟。在这场如火如荼的“AI竞赛”中，各方需要保持冷静与克制，加强对AI安全风险的重视，警惕在未做好充分的安全评估和保障之前释放“猛兽出笼”。

【作者系北京大学法学院教授、博导，北京大学人工智能研究院双聘教授，北京大学武汉人工智能研究院副院长，中国科学技术法学会常务副会长兼秘书长，中国法学会知识产权法学研究会副会长】

面对多重不确定，人工智能立法如何应对

□张凌寒

人工智能产业的突出特点之一是飞速发展带来的不确定性和不可预见性，没有人能够断言未来人工智能产业会走向何方，带来何种新的风险与挑战。与传统立法相较，中国人工智能立法需要应对三方面的不确定性，包括技术飞速发展导致调整对象的不确定性、技术产业迭代给社会关系调整带来的不确定性以及多种类高风险的不可预见性。对此，我国人工智能立法必须进行相应的适应性制度设计，确保监管能够充分应对人工智能的未来发展及不确定性。

首先，调整对象的不确定性在以往的技术治理过程中已有所体现，我国曾以“小、快、灵”的立法思路予以应对。从《互联网信息服务算法推荐管理规定》到《互联网信息服务深度合成管理规定》再到《生成式人工智能服务管理暂行办法》，调整对象顺应技术发展经历了从算法到深度合成再到生成式人工智能的转变。然而，一部完整的人工智能立法不同于低位阶法规，立法周期可能以年为单位计，这就要求立法为调整对象的不确定性预留充分的适应空间，有效包容未来可能的技术发展。

其次，人工智能产业迭代带来的社会关系的不确定性对立法提出了较高要求。随着生成式人工智能的广泛应用，通用人工智能已初现端倪，大模型可广泛赋能千行百业，形成“基础模型——专业模型——服务应用”的分层业态。这不仅意味着技术的迭代革新，还意味着法律调整的社会关系的深层变革。过去十年间，数据、算法、平台是我国人工智能治理的主要对象，但技术的飞速发展不断更新着现有的社会关系，从深层次挑战着既往监管模式。因此，人工智能立法应能够包容产业迭代带来的社会关系的不确定性。

最后，人工智能多种类高风险的不可预见性增加了立法难度。目前，人工智能技术发展的速度远超社会预期，社会治理体系并未做好充分的准备。人工智能技术的不透明度从面向普通人的“黑箱”，升级为即使专业人士也无法理解的“黑箱”。社会担忧强大的人工智能系统未能与人类价值观“对齐”进而带来灾难性风险。既有的经验和制度并不足以全知全能地应对各种高度不可预见的风险，我国人工智能立法在不过度限制技术发展的同时，需要妥善应对已知和未知风险，确保人工智能始终朝向有利于人类文明的方向发展。

相比于传统立法活动，应对高度不确定性将是我国进行人工智能立法制度设计应遵循的底层逻辑之一。人工智能不应再被视为单纯的技术或工具，而是社会生产方式与社会关系重大变革的驱动力。对此，法律需基于“生产关系适应生产力发展”这一原理对前述问题作系统性回应。

从立法思路来看，为应对技术的不确定性，我国的人工智能立法可以借鉴网络立法经验，以具有高度开放性的科技伦理入法以及具有迭代演化能力的制度来应对风险、技术和规制需求的不确定性。近年来，法律已密集地将科技伦理要求纳入规范，如2021年颁布的《新一代人工智能伦理规范》要求促进公平公正、保护隐私安全、确保可控可信等，这一情形表明科技伦理规范与法律及行政法规制度逐步趋于融合。同时，以风险评估、影响评估及安全评估等评估制度为基础具有迭代演化能力的网络法，注重从事前到事中激励相对人预先构建技术机制来预防和规避技术风险，这些经验均可为我国人工智能立法所吸纳。

从架构上来看，人工智能法作为一部综合性法律，可以通过“立总则，留接口”的方式，解决容纳和化解高度不确定性的问题。具体而言，我国人工智能法应确立“总则式”的框架体例，定位为人工智能治理的提纲挈领制度，并为应对未来技术发展的治理规范预留充分的制度接口。人工智能法的调整对象是人工智能技术、产业和应用全链条以及在此基础上形成的复杂系统与社会生态，立法不能抱持“毕其功于一役”的理念，试图涵盖所有人工智能技术、要素和应用等不同维度的治理规则。无论社会各界是否支持启动制定人工智能法，“总则”部分都应在重要问题上达成最大共识，为人工智能时代个体权利的保护制度体系定调，绘制人与技术的合理关系图谱。而“接口”部分不仅应当为已经或未来可能影响到的教育、医疗、金融、交通等社会领域留足“领域接口”，还需将关于技术、要素与产品应用的低位阶制度授权有关部门细化，留足“授权接口”，当未来产业发展变化后可相应地调整法律义务与责任设置。

综上所述，不确定性是未来人工智能技术发展的突出特点，中国的人工智能法必须为未来技术发展的不确定预留包容空间，以高超的立法智慧充分加以应对，实现对人工智能的科学、有效治理。

【作者系中国政法大学数据法治研究院教授、博导，联合国高级别人工智能咨询机构（UN High-Level Advisory Body on AI）专家组中方成员，中国人工智能产业发展联盟政策法规工作组组长】