

论人工智能立法的基本路径

林洹民^{*}

内容提要 采用何种路径规范人工智能活动,是人工智能立法的核心问题。风险管理进路存在风险评估与分类困难、放任损害发生等问题,并非人工智能立法的当然选择。与以往科技活动不同,人工智能活动既属于专精科技活动,又具有赋能科技活动属性。以人工智能活动为规范对象的人工智能法不应以单一理论为指导,而应遵从科技法与应用法双重定位。科技法定位下的《人工智能法》应尊重科技自主,将科技伦理内化于人工智能研发活动中,同时打破制度壁垒,设计促进型规则,助力人工智能科技的发展。应用法定位下的《人工智能法》则应关注科技赋能场景导致的功能异化现象,一方面借助抽象的权利义务工具,尤其是通过规定新型权利,构建弹性的规范框架,回应不同应用场景中的价值序列差异;另一方面应推行实验主义治理,通过监管沙箱、授权性立法等设计,动态调整监管策略,满足人工智能赋能应用活动的灵活治理需求。

关键词 人工智能立法 专精科技 赋能科技 科技伦理 实验主义治理

目 次

- 一、引言
- 二、人工智能立法风险管理单一进路之检讨
- 三、因应人工智能活动特性的复式立法路径
- 四、复式立法进路下人工智能规范框架的展开
- 五、余论

^{*} 中央财经大学法学院副教授,法学博士。本文系 2023 年度国家社科基金重点项目“个人数据交易的私法构造研究”(项目批准号:23AFX014)的阶段性成果。

一、引言

随着人工智能技术的发展与应用,人工智能治理正在引起国际社会的高度重视。立法机构已经不满足于伦理宣言式的行业自治,转向出台专门性的法律,从而开启了从“软拘束”到“硬规则”的人工智能立法时代。欧盟于2021年率先推出全面规范人工智能活动的法律《人工智能法(提案)》(The Artificial Intelligence Act)。该提案经过修正后,于2024年3月经欧洲议会表决通过,这标志着欧盟诞生了世界上首部规范人工智能活动的综合性法律。美国在联邦层面尚没有进行类似的横向集中式立法,但也于2023年10月发布“具有里程碑意义的”总统行政命令,要求美国联邦政府各部门和机构制定人工智能活动监管政策。在我国,国务院自2023年开始已经连续两年将人工智能立法列入立法工作计划;全国人大常委会于2023年9月发布的《第十四届全国人大常委会立法规划》也将人工智能立法列入第一类项目(条件比较成熟、任期内拟提请审议的法律草案),并指出“推进科技创新和人工智能健康发展……要求制定、修改、废止、解释相关法律,或者需要由全国人大及其常委会作出相关决定的,适时安排审议”。可以预见,一旦完成准备工作,我国人工智能立法将会提速进行。

围绕人工智能立法的争议问题包括但不限于以下五点:人工智能法的规范对象、人工智能立法的基本路径、人工智能时代的新兴权利、监管机构以及法律责任。其中,对人工智能活动的规范进路问题涉及价值判断与治理观念,反映了立法者应对未知的伦理观。就立法基本路径而言,极端谨慎的态度是,如果不能防范风险,就不允许新技术的应用,但这一方案明显过于保守。与之相对,风险管理进路强调拥抱风险、容忍错误,对技术研发与应用更为友善。欧盟的《人工智能法》就明确采取风险管理进路规范人工智能活动,该法“序言”第14条指出,法案遵循“基于风险的路径”(risk-based approach),即根据人工智能系统可能产生的风险的强度和范围来确定规则的类型和内容。在英国达成的全球第一份针对人工智能的国际性声明《布莱切利宣言》,同样强调应采取“基于风险的政策”调整人工智能活动。我国也签署了《布莱切利宣言》,接受基于风险的人工智能规范路径。然而,风险管理进路未必合理。笔者认为,我国如果简单地照搬他国模式进行人工智能立法,不仅无法为世界法治贡献智慧成果,也将丧失引领人工智能全球治理的良机。是以,研究人工智能立法的基本路径问题,既是发展我国本土法学、树立中国法学界碑的应有之义,也是开展人工智能国际合作、贡献中国智慧、表现大国担当的当然之举。

有鉴于此,本文将在分析风险管理进路利弊的基础上,通过辨析人工智能活动的双重属性,提出人工智能立法的另一种思路与策略,努力打造一种既能满足人工智能动态监管需求,又能促进人工智能创新发展的新型治理框架。

二、人工智能立法风险管理单一进路之检讨

或许是受到“风险社会”理论的影响,人们往往下意识地选择风险管理路径调整新兴科技活动。但风险管理进路不同于风险社会理论,后者是探讨后现代社会家庭、职业、知识以及外部自然环境等因素的风险特征的理论体系,前者则是一套基于“风险—收益”分析的社会管理方法。风险管理进路的模式是:首先,将损害界定为风险;其次,评估可能引发风险的行为;最后,在进行成本收益分析的基础上划分特定的风险分类并配置相应的规范。风险管理进路通过评估风险的危害程度、控制风险的成本以及风险能带来的收益,选择相应的规范设计。^①其中,将损害视为风险是风险管理的正当性基础,风险评估是风险管理的前提,而风险分类则是风险管理的关键。然而,未经充分论证就将损害界定为风险,有轻视个人权益保护之嫌;由于人工智能活动引发的风险具有高度复杂性,对人工智能活动的风险评估与分类也极为困难,风险管理路径未必能有效地调整人工智能活动。

(一) 风险管理的正当性问题

正当性问题是人工智能风险管理进路受到批评的重要原因。风险管理进路强调拥抱风险、容忍错误,对技术研发与应用极为友善。但是,包容审慎的治理思路以不严重侵犯个人权益和产生重大社会风险为前提。^②欲在政策上选择风险管理进路,必须首先论证为什么让个体与社会承受不可知的重大损害是值得的。易言之,既然已经确定地知道人工智能将产生风险,且风险可能会造成严重的损害,为什么要选择一种容忍损害的规范路径?

一种辩护观认为,风险与收益同在,个体虽然承受损害,但也在享受技术发展带来的红利。例如,智能手机极大地提升了人们的生活质量,大部分人已经无法在没有智能手机陪伴下生活。但是,并非所有人都是科技的受益者。在科技转化过程中存在决策者、受益者与波及者三方主体。波及者是不能参与政策制定又难以享受科技福利的群体。在风险管理路径下,偏离一般基准的人,如老人、低智力群体,往往被迫承受技术发展的不利后果,他们的特别需求也常被忽视。^③另有观点认为,为了集体或多数人的利益,有时不得不让少数人承受损失。集体利益确实常被视为承受风险的理由,^④相较于个人,群体和社会才是风险管理进路关注的重点^⑤。但个人利益与集体利益是难以划分的,当海量

① See Margot E. Kaminski, *Regulating the Risks of AI*, Boston University Law Review, Vol.103:1347, p.1347-1411 (2023).

② 参见刘权:《数字经济视域下包容审慎监管的法治逻辑》,载《法学研究》2022年第4期,第37-51页。

③ 例如,在数字化时代,无法适应智能手机的老人遭遇“数字鸿沟”,有被排挤出社会的危险。工业和信息化部印发了《互联网应用适老化及无障碍改造专项行动方案》(工信部信管〔2020〕200号),在全国范围内组织开展为期一年的互联网应用适老化及无障碍改造专项行动。该项举措一定程度上改善了老年人的数字生存困境,但并未从根本上解决问题。

④ See William Boyd, *Genealogies of Risk: Searching for Safety, 1930s-1970s*, Ecology Law Quarterly, Vol.39:895, p.912 (2012).

⑤ See Douglas A. Kysar, *The Public Life of Private Law: Tort Law as a Risk Regulation Mechanism*, European Journal of Risk Regulation, Vol.9:48, p.52 (2018).

个体沦为人工智能客体时,集体利益也在被侵犯。因而,风险管理进路未必对整体有益。

值得庆幸的是,人们已逐渐认识到风险管理路径的弊端,转而考虑采用其他替代策略。例如,在环境保护方面,德国环境保护法就抛弃之前的风险管理模式,转采预防原则。预防原则的宗旨是,面对科学上的不确定性,政策制定者应努力防止危害的发生。即便缺少损害与行为之间充分的因果关系论证,只要损害后果可能极为严重,政府也应马上采取监管措施。目前,预防原则已被写入众多环境保护公约中,取代风险原则成为环境保护的基本原则。^⑥在人工智能治理领域,业界也有要求强化对人工智能活动监管,对特定人工智能活动适用许可制度而非风险管理的呼声。^⑦除此之外,还有美国学者针对人工智能损害的不可逆性,强调应努力构建某种恢复性制度(Resilience),以替代人工智能风险管理模式。^⑧

(二) 风险评估的可行性问题

抛开正当性不论,人工智能活动风险管理进路在应用层面的可行性也存在问题。风险管理以有效的风险评估为前提。风险评估本质是一种成本效益分析,欧盟《人工智能法》第3条就将风险界定为发生损害的概率和该损害的严重程度的组合。风险只有被量化分析,才有可能与收益相比较,从而确定可接受的限度。但是,对人工智能活动引发风险的量化分析极为不易,人工智能风险评估未必能成为决策的基础。

其一,欠缺高质量的数据和有效的模型。首先,缺乏作为量化评估基础的高质量数据。风险评估需要的基本数据包括损害的危害程度、发生的概率、损害的地理和时间分布(普遍性)、损害的持续时间(持续性)、损害的可逆性等。^⑨然而,在新技术被投入应用前,并不存在可供评估风险的相应数据。即便就已投入应用的产品而言,受技术发展、地理、时间等各方面因素影响,所提供的数据也未必具有参考性。尤其是,以ChatGPT、Gemini为代表的人工智能大语言模型具有强大的自我学习能力,依据既有的数据难以预测不断进化的人工智能活动带来的风险。在缺乏高质量数据的情况下,显然难以精确地计算风险发生的概率与危害程度,也就无法借助成本收益法确定人工智能活动可容忍的限度。^⑩其次,目前并不存在评估人工智能风险的有效模型。对新兴技术的风险评估,将因沿用旧有的分析框架而无法精确界定与测量风险。以化学材料为例,“定量结构—活性关系”(QSAR模型)工具可为普通化学品提供相当可靠的风险估计,但纳米材料的毒性受化学结构以外的其他因素(包括尺寸、表面积、表面特性等)影响,因而“定量结

^⑥ See European Parliament, *The Precautionary Principle: Definitions, Application and Governance*, at [https://www.europarl.europa.eu/RegData/etudes/IDAN/2015/573876/EPRS_IDA\(2015\)573876_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/IDAN/2015/573876/EPRS_IDA(2015)573876_EN.pdf) (Last visited on Aug. 3, 2024).

^⑦ 参见黄宁燕:《“ChatGPT之父”提出未来人工智能监管框架》,载《科技中国》2023年第8期,第98页。

^⑧ See Cameron A. MacKenzie & Christopher W. Zobel, *Allocating Resources to Enhance Resilience, with Application to Superstorm Sandy and an Electric Utility*, *Risk Analysis*, Vol.36: 847, p.859 (2015).

^⑨ See Ortwin Renn & Andreas Klinke, *Risk Governance: Concept and Application to Technological Risk*, in Adam Burgess, Alberto Alemanno & Jens Zinn eds., *Routledge Handbook of Risk Studies*, Routledge Press, 2016, p.204-215.

^⑩ See Bridget M. Hutter, *A Risk Regulation Perspective on Regulatory Excellence*, in Cary Coglianese ed., *Achieving Regulatory Excellence*, Brookings Institution Press, 2016, p.101-114.

构—活性关系”并不能有效分析多数纳米材料的风险。^⑪又如,目前对于机动车的检测标准与程序并不适用于智能网联汽车,《医疗器械监督管理条例》《医疗器械生产监督管理办法》也未必能够合乎预期地调整人工智能医学辅助器械(如看护机器人)的生产与使用。可见,如果使用旧有模型评估新型人工智能活动带来的风险,评估结果将不具有参考价值。当评估模型失效时,即便存在有效的数据,风险评估也将缺乏说服力。

其二,难以准确评估技术叠加引发的风险。一般技术引发的风险相对简单,风险发生概率与损害也容易确定(如机动车交通事故风险、断电风险等)。但人工智能大语言模型引发的风险具有叠加性,使得风险评估更为困难。复杂系统中不同活动和事件之间的相互作用使得风险成倍增加,引发协同效应,风险总量远大于各部分之和。^⑫有时两个或更多的故障孤立地看均不具有破坏性,可一旦故障以意想不到的方式结合在一起,就会使得安全装置失效,引发重大系统性事故。在“大模型+具体应用”的产业生态中,上游基础大模型和下游具体应用之间存在着复杂的依存关系。通过海量数据训练出的大模型为具体模型提供底层逻辑支持,具体模型则通过“术业有专攻”的专业优化训练,适配众多的具体的行业和场景。^⑬在这一协同应用关系中,对上游大模型的风险评估无法预计下游应用产生的具体风险,对下游应用的风险评估也无法预测反馈机制对上游大模型自我学习能力的影响。

人工智能风险评估技术越不可靠,评估结果受到外部因素干涉的可能性就越大。近几十年来,学者们对技术风险评估的客观性进行了检验,发现偏见、道德、政策、社会文化等均会影响风险的识别与评估。^⑭其中,政策对于风险评估结果的影响尤其引人注目。温迪·瓦格纳将外部政策环境对风险评估的影响评价为“科学骗局”(science charade),即用看似客观的技术评估为幌子,就技术的应用作出政策决定。^⑮尤其是当人工智能活动风险评估在技术层面遭遇障碍时,风险评估就更易受到外部社会政治环境的影响。此时,财富生产的逻辑总能获胜。^⑯如果风险评估变为隐性的政策选择,风险控制阀门将有失效之虞。

(三) 风险分类的融贯性问题

如上所述,风险分类是风险管理的关键。风险管理路径力求将人工智能活动引发的风险进行分类,进而配置与风险级别相称的规范。目前国际上的风险分类模式主要有三种:风险属性划分模式、风险内容划分模式以及风险程度划分模式。但上述模式均存在

^⑪ See David W. Hobson et al., *Applied Nanotoxicology*, International Journal of Toxicology, Vol.35:5, p.6 (2016).

^⑫ See Morten Broberg, *Risk Regulation and the Future: On the Need for Helping Vulnerable Societies to Adapt to the Consequences of Climate Change*, European Journal of Risk Regulation, Vol.8:101, p.101-105 (2017).

^⑬ 参见张凌寒:《生成式人工智能的法律定位与分层治理》,载《现代法学》2023年第4期,第126-141页。

^⑭ See W. Kip Viscusi & Richard J. Zeckhauser, *Regulating Ambiguous Risks: The Less than Rational Regulation of Pharmaceuticals*, Journal of Legal Studies, Vol.44:387, p.387-422 (2015).

^⑮ See Wendy E. Wagner, *The Science Charade in Toxic Risk Regulation*, Columbia Law Review, Vol.95:1613, p.1627-1628, 1678 (1995).

^⑯ 参见[德]乌尔里希·贝克:《风险社会:新的现代性之路》,张文杰、何博闻译,译林出版社2022年版,第46页。

弊端,不能融贯地划分风险的类型与级别。

1. 风险属性划分模式

风险属性划分模式是一种依据风险属性对人工智能活动进行分类的治理模式。采用此种模式的典型代表为美国。美国商务部国家标准与技术研究院推出的《人工智能风险管理框架》(AI RMF 1.0)根据风险属性不同,将人工智能活动风险分为技术性风险、“社会—技术”风险与指导原则性风险。技术性风险是指影响人工智能运行稳健性和准确性的风险,“社会—技术”风险涉及人工智能对隐私、安全、自由、公平等价值的影响,指导原则性风险则是人工智能的应用可能会影响对“好的”或“可信赖”的人工智能的理解。^{①⑦}这一模式的最大问题是,风险本身难以被精确定性。首先,技术一定会出错,而如何分配错误本身就是一个社会价值判断,并不存在纯粹的技术性风险。其次,“社会—技术”风险与指导原则性风险之间难以区分,因为只有在具体的应用场景中才能判断“好的”或“可信赖”。以深度伪造技术为例:当深度伪造技术被用于合成虚假图片、视频时,我们会觉得人工智能不可信赖;但当深度伪造技术被用以重现消失的艺术时,人们又乐于享受这一应用,认为人工智能是“好的”“可信赖的”。^{①⑧}对人工智能技术的宏观评价,离不开具体语境中的价值衡量。从该意义上说,指导原则性风险本质上也是一种“社会—技术”风险。依据风险属性划分风险类型的思路,在逻辑上并不清晰,也因此欠缺可执行性。

2. 风险内容划分模式

风险内容划分模式是依据风险现实化的后果进行分类。我国信息安全标准化技术委员会在2021年发布的《网络安全标准实践指南——人工智能伦理安全风险防范指引》依据风险后果将人工智能活动风险分为“失控性风险”“社会性风险”“侵权性风险”“歧视性风险”“责任性风险”五类;英国科技创新与技术部在2023年推出的《人工智能管理的创新优先路径》报告也根据风险现实化后果分类,但将人工智能风险划分为“人权风险”“安全风险”“公平风险”“隐私风险”“社会福利风险”与“可靠性风险”六类。^{①⑨}可见,即便采取同样的风险分类方式,中英两国的风险分类也并不相同。由于观察者会因信息获得渠道与关注焦点不同而总结出不同的风险类型,风险内容划分模式很难形成一套具有说服力的风险类型。更何况,列举难免会挂一漏万。立法者若依据该模式确定人工智能风险类型,可能会自我设限,使得一些本应受到充分关注的人工智能活动被排除在法律规范之外。

3. 风险程度划分模式

风险程度划分模式是目前最受关注的风险分类模式。在该模式下,立法者根据人工智能活动引发风险的危害程度划分风险类型。欧盟《人工智能法》为该模式的典型代表。

^{①⑦} See The National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, at <https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf> (Last visited on Oct. 17, 2023).

^{①⑧} Vgl. Markus Thiel, „Deepfakes“ – Sehen heißt glauben?, ZRP202 (2021), S. 203.

^{①⑨} See Department for Science, Innovation & Technology, *A Pro-Innovation Approach to AI Regulation*, 2023, p.12.

该法将人工智能活动引发的风险分为四类,并相应地配置规则:引发不可接受风险的人工智能系统是被禁止的,高风险人工智能系统必须遵守特定要求,限制性风险人工智能系统受到拘束较少,最小风险人工智能系统则完全不受限制。但这一模式也存在明显的局限性。其一,风险程度缺乏明确的判断标准。例如,欧盟《人工智能法》第7条第1款要求结合“是否严重危害健康和安全”和“是否对基本权利、环境以及民主和法治造成严重不利影响”来判断人工智能活动是否属于高风险,但上述标准具有巨大的弹性与模糊性。欧盟试图通过人工智能委员会与人工智能办公室提供相对清晰的指引,但行政机关未必有能力评估风险并进行风险分类。例如,以 ChatGPT 为代表的生成式人工智能可能产生不实资讯,从而直接或间接地损害个人权益,但欧盟对于应将之归入何种类型的人工智能活动存在大量争议。^{②①}美国于2024年5月通过的首部规范人工智能的地方性法律《科罗拉多州人工智能法》(SB24-205),同样以高风险人工智能系统为规范对象。根据该法第6-1-1701条第9款第a项,构成“重大决策的实质性因素”的任何人工智能系统均为高风险人工智能系统。而这一标准可能使得所有人工智能系统都被认定具有高风险。因此,科罗拉多州不得不在该条的第b项规定大量例外,但也会产生“一刀切”“挂一漏万”等问题,并不能为确定是否具有高风险提供清晰的指引。其二,僵化的风险类别与高速发展的人工智能技术之间存在张力。一方面,随着人工智能技术的发展,原本被视为限制性风险或最小风险的人工智能应用可能会变得具有高风险性甚至不可接受。例如,人工智能辅助教学曾被视为教育数字化的重要标志,但随着人脸识别、情绪识别技术的发展,对学生课堂活动的监控有严重侵犯人权的风险。另一方面,原本被认为禁止应用或高风险的人工智能活动,也可能随着安全技术措施的发展与应用风险程度显著降低,但因规范的滞后性,相应的人工智能活动将受到不当限制。例如,无人驾驶引发的风险等级随着自动驾驶技术的成熟而不断降低。无人驾驶汽车越早地被投入到真实场景中进行路测,越有可能占据市场,但若无人驾驶被归入高风险人工智能活动,其路测场所、时间与程度等都将受到严格限制,最终也将制约技术与汽车行业的迭代升级。

风险治理作为一种现代社会治理模式,已成为“一种崇拜”,使得“一种近乎神奇的光环笼罩着对风险评估与风险治理”。^{②②}但若审慎地分析该模式,就会发现风险管理单一路径并非调整先进科技活动的良方。首先,风险管理路径仅适用于可量化的危害,对人工智能活动引发的风险并不适用,而忽视不可量化的危害是“灾难的根源”。^{②③}其次,人工智能活动的复杂性使得全面、妥当的风险分类与归类极为困难。最后,风险管理路径有损害个体和集体利益的危险。欧盟采取风险管理进路规范人工智能活动,是一种有益的尝试,但单一理论是无法规范复杂、多场景的人工智能应用的。我国未来的人工智能立法不应受到“布鲁塞尔效应”的影响,亦步亦趋地遵循风险管理路径,而应在充分了解

^{②①} See Natali Helberger & Nicholas Diakopoulos, *ChatGPT and the AI Act*, Internet Policy Review, Vol.12:1, p.1-6 (2023).

^{②②} See John Durant, *Once the Men in White Coats Held the Promise of a Better Future*, in Jane Franklin ed., *The Politics of Risk Society*, Polity Press, 1998, p.73.

^{②③} See Daniel A. Farber, *Uncertainty*, Georgetown Law Journal, Vol.99:901, p.909 (2011).

人工智能活动特性的基础上,寻找一种更兼顾安全与发展的综合的人工智能治理范式。

三、因应人工智能活动特性的复式立法路径

人工智能活动具有高度复杂性,以之为调整对象的法律难以按照风险管理单一路径进行设计。欲制定调整行为的法律,必须首先了解被规范的对象。如果只是基于一种主观判断进行立法,这种判断未必能够与现实问题相对应,其是否有足够的张力、体现了何种程度的偏差或遗漏,也不得而知。人工智能活动具有双重属性,人工智能立法也应具有复式立法定位。

人工智能活动由人工智能研发活动与人工智能应用活动组成,前者体现出鲜明的科学性,后者则以赋能应用为主,因此具有工具性特征。当然,人工智能研发活动与应用活动也存在交叉。例如,一些研发活动的目的是赋能应用,在应用过程中也可以通过数据训练人工智能系统,实现人工智能研发的质的突破。为了更加清晰地说明问题,下文分别着重在研发活动与应用活动的框架内讨论人工智能活动的科学性与工具性。

(一) 人工智能活动的双重属性

科技大致被划分为专精科技和赋能科技两类。专精科技侧重于深度与专业性,赋能科技则注重于应用与提升能力,二者在性质、功效和实现手段上存在根本差别。以往的科技要么属于专精科技,要么表现为赋能科技。人工智能活动相较于以往的科技活动更为复杂,既具有高度专业性(科学性),又具有广泛的赋能性(工具性),同时属于专精科技活动与赋能科技活动。

1. 作为专精科技活动的科学性

人工智能属于新一代科技革命中的表征性科技。人工智能研究属于一项专精科技活动,旨在开发出可以自我分析、自我总结、自我纠错的人工智能系统。为了实现这一目标,在计算机科学中产生了符号主义、联结主义和行为主义等诸多流派。^{②③}神经网络技术的出现使得人工智能研究有了重大突破(如深度学习算法)。以 ChatGPT 为代表的大语言模型又使得人工智能研发前进一大步。^{②④}目前,人工智能已经具有超强的自我学习能力。在此基础上,有些科研活动直接瞄准超级人工智能,力图使得人工智能呈现出一种人类尚未发现或无法实现的逻辑形式,进而超越人类智能,探索到人类无法触及的高度。^{②⑤}显而易见,人工智能研发本质是一项科学研究,属于专精科技活动。

人工智能研发活动的科学性增加了调整人工智能活动的难度。非专业人士通常无法真正理解作为一项科学研究的人工智能的影响。以作为人工智能三要素之一的算法

^{②③} 参见[英]彼得·本特利:《计算机:一部历史》,顾文天译,电子工业出版社2015年版,第255-284页。

^{②④} See Gary Drenik, *Large Language Models Will Define Artificial Intelligence*, at <https://www.forbes.com/sites/garydrenik/2023/01/11/large-language-models-will-define-artificial-intelligence/?sh=33b50d6ab60f> (Last visited on July. 27, 2023).

^{②⑤} 参见[美]亨利·基辛格、[美]埃里克·施密特、[美]丹尼尔·胡滕洛赫尔:《人工智能时代与人类未来》,胡利平、凤君译,中信出版社2023年版,第15页。

为例,人工智能算法的书写甚至阅读都是一项专业技能。对非专业人士而言,算法是另外一种语言。因此,只有借助计算机科学家对科研活动的自我批判,才能理解与规范人工智能活动。目前也已经存在以控制人工智能为目标的科研项目,例如,谷歌公司、斯坦福大学等都在进行算法可视化的研究,以增强算法的可解释性。当下的研究增加了对图像识别人工智能算法的解释可能,但是仍有很多工作未能完成。^{②6} 只有借助科学系统自身,才能更好地识别与解决问题。

2. 作为赋能科技活动的工具性

当聚焦人工智能活动时,人们更多关注人工智能的赋能应用。以往的科技常被认为是定向的,影响范围基本是领域性的(如化石燃料、通信技术等),即便是通过跨学科合作产生了一定的开放性,也不具备通用性。在此认知之下,科技主要被视为推动经济发展的工具,科技法也主要关注技术的经济意义,一直秉持“技术—经济”的立法范式。^{②7} 人工智能的出现改变了科技的定向性。人工智能是一项通用技术,是科学研究、教育、制造、物流、运输、司法、行政、广告、艺术等众多领域与人类生活各方面的赋能者。通过在具体应用场景中的调整,人工智能可以满足多样的需求。人形机器人就是人工智能技术与场景的叠加和结合的典型。人形机器人根据不同场景需求呈现出不同的表现形式,如陪伴机器人、看护机器人以及亲密机器人等。^{②8} 我国工业和信息化部《人形机器人创新发展指导意见》指出,应首先研发出基础版人形机器人整机,打造“公版”通用平台,然后支持不同场景需求下的结构改造、算法优化以及特定能力强化。

作为赋能工具的人工智能活动,在价值上并不当然具有合理性,可能会带来新的社会风险,导致社会关系异化。仍以人形机器人为例,在大语言模型的加持下,人形机器人将获得强大的操纵能力,损及人的主体性。其一,人形机器人具有智能表象。如果人类发现人形机器人比一般人更有能力,人形机器人迟早会获得高于普通人的影响力。^{②9} 当我们习惯性地首先问询机器人意见时,人类就在丧失自主性。一个显而易见的例子是,高德地图、百度地图等导航程序大大减弱了个人的自主寻路能力。其二,高度拟人化的外观使得人形机器人更容易被视为“同类”。人类对于人形物件产生共情是自然的心理倾向,人形机器人的外观能够引发人类的心理投射,很容易引起共情。如果陪伴机器人或亲密机器人(性爱机器人)几可以假乱真,人类可能会有意无意地将机器人视为同类,并将之视为自己的最佳伙伴。这将使得具身智能进一步获得不同寻常的操纵能力。具有操纵性能力的人形机器人可能会限制个人的自我决定、干扰人的社会化甚至扭曲家庭观念,引发新型社会问题,因而需要法秩序积极应对与规范。^{③0}

^{②6} See Chris Olah et al., *The Building Blocks of Interpretability*, at <https://distill.pub/2018/building-blocks/> (Last visited on Jan. 1, 2024).

^{②7} 参见龙卫球:《科技法迭代视角下的人工智能立法》,载《法商研究》2020年第1期,第57-72页。

^{②8} See Amanda J. Sharkey & Noel Sharkey, *Granny and The Robots: Ethical Issues in Robot Care for The Elderly*, *Ethics and Information Technology*, Vol.14:27, p.27 (2012).

^{②9} See European Parliament, *The Ethics of Artificial Intelligence: Issues and Initiatives*, PE634.452, March 2020, p.18.

^{③0} 参见林涓民:《人形机器人的操纵性风险及规范进路》,载《东方法学》2024年第3期,第159-170页。

（二）双重属性下的人工智能复式立法定位

基于人工智能活动的科学性与工具性双重属性,以之为规范对象的人工智能法也应具有双重定位。人工智能研发活动是一项科学研究活动,以科学活动为规范对象的《人工智能法》当然具有科技法属性;人工智能应用则是技术与场景的结合,因而以场景化应用为规范对象的《人工智能法》也应具有应用法属性。是以,我国未来的人工智能立法应同时具有科技法与应用法双重定位。

1. 作为科技法的人工智能法

科技法导向下的人工智能立法应充分尊重科技自主,关注科技伦理。现代社会的典型特征是功能分化,功能分化创造解决社会问题的诸多子系统,社会子系统无法被一种方式整合,而是始终按照各自的媒介发挥作用。例如,经济子系统无法替代教育子系统的功能,科研活动也只能依托科技系统推进。^{③①} 科研活动一般不愿受到过多的外部限制,对科研活动的规范应优先借助科技系统自身。借助科技伦理规范科技本身,此谓科技系统的自我反身性。长期以来,人们认为科研不应存在禁区,因为科学研究本身是一种求真的活动,依赖一个自由探索的科研环境。如若事先设定禁区,将会破坏科学研究的求真本性。但随着现代科学的兴起,知识的发现不再被认为是最终目的,科学的成功可能就是危害的开始。美国《科学》杂志指出,现代科学引发的科学与社会责任、科学与伦理、科学与现代性等问题,正以前所未有的速度与规模引起人们的注意。^{③②} 科学并非全然有益,一些科学研究的危害性甚至胜过其学术性。人们逐渐认识到,科技并非文明,科研活动也应受到规范,科技伦理应运而生。人工智能活动作为科技活动的一种,当然应遵守人工智能科技伦理。

人工智能法的科技法属性也意味着,我们在规范科技活动的同时应规定适当的科技促进型制度。人工智能的核心要素是算法、算力与数据,但人工智能法并不等同于人工智能要素法。若将人工智能立法简单拆分为人工智能要素的立法,围绕各个要素设计一般性规则,不但会消解人工智能法的立法目的,也会模糊人工智能法与《个人信息保护法》《数据安全法》等法律的关系。不能因为人工智能的核心要素是算法、算力与数据,就围绕各个要素面面俱到地制定规则。所应思考的毋宁是,阻碍人工智能发展的障碍为何、我国未来的《人工智能法》又应如何适当地设计规则以充分发挥制度的破壁效果。目前较为清晰的是,数据资源的汇集与适用是影响人工智能发展的关键性问题。^{③③} 中国社会科学院发布的《人工智能示范法(专家建议稿)》第18条、中国政法大学等高校合作起草的《中华人民共和国人工智能法(学者建议稿)》第20条,均规定了数据要素供给制度,强调国家应支持建设人工智能领域基础数据库和专题数据库,促进数据资源高

^{③①} Vgl. Niklas Luhmann, Rechtssoziologie, 4. Aufl. 2008, S. 236.

^{③②} See Rosemary Chalk ed., *Science, Technology, and Society: Emerging Relationships*, American Association for the Advancement of Science, 1988, p.11.

^{③③} 我国台湾地区于2024年发布的“人工智能基本法(草案)”第15条就专门强调,应建立资料开放、共享与再利用机制,提升人工智能使用资料之可利用性,提升人工智能使用资料之质量与数量。

效汇聚和共享利用,鼓励引导相关主体开展大数据与人工智能技术协同研发等。^{③④}我国未来的人工智能立法,应规定数据使用规则,适度突破既有的个人信息保护规则、著作权保护规则,满足人工智能训练的数据需求。

2. 作为应用法的人工智能法

人工智能复杂、多样的场景化应用增加了统一立法的难度。人工智能赋能不同应用场景,引发的法律关系并不相同。以人脸识别为例,如果人脸识别技术被用于门禁、电子护照、自动支付,涉及的法益包括出行自由与财产安全等;但若人脸识别技术被用于一般性的监控,则主要涉及自由与安全之间的矛盾。^{③⑤}自动决策算法在公私不同领域中的应用,引发的法律问题也不一样。商业领域的自动决策算法应用容易引发算法黑箱与歧视问题。例如,招聘人工智能可能有意识地排斥特定地域的应聘者,将地域歧视程序化。^{③⑥}公权力运用决策型算法的,则会直接对个人的基本权利造成影响。例如,以“秒批”为代表的自动行政算法极大地提升了行政效率,但也在一定程度上侵犯了公民的人身自由和知情权。可见,人工智能在公私领域应用涉及的价值各异,统一的规范设计并不容易。

在很长的一段时间内,思想家偏好一种本质论的思维方式,即通过界定“事物本质”来寻找其发展规律,运用规律规范事物本身。遗憾的是,试图抓住本质、化繁为简的努力往往不易成功,很多时候是表面上似乎对认识对象一目了然了,但是实际上却并没有真实反映复杂社会的实际情况,反倒使得认识对象被过度简化。所以,现代认识范式越来越偏爱非本质论,即通过揭示事物的复杂性、多样性,在全面了解事物不同面相的基础上解释与调整事物。^{③⑦}以经济学为例,复杂经济学理论指出,古典均衡理论过于理想化和理性化,扭曲了现实世界。现代经济学批判以往经济模型的过度简化,认为经济不是确定的、静态均衡的,而是依赖于过程的、有机的、永远在进化的。^{③⑧}笔者认为,欧盟采取风险管理单一进路规范人工智能活动,认为风险是确定的与可预测的,一定程度就是在将复杂世界简单化与静态化。单一理论是无法规范复杂、多场景的人工智能应用的,如果按照单一理论设计规则,难免会产生规制过严或过松的问题。鉴于人工智能多样的应用场景,人工智能赋能应用规范不应寻求简单的、单一的规范框架,而应充分重视事物的复杂性。

我国目前存在两个版本的人工智能法建议稿。中国社会科学院发布的《人工智能示范法 2.0(专家建议稿)》规定了人工智能支持与促进制度,人工智能管理制度,人工智能研发者、提供者义务以及人工智能综合治理机制;中国政法大学等高校合作起草的《中华人民共和国人工智能法(学者建议稿)》规定了发展与促进、使用者权益保护、开发

^{③④} 参见《〈人工智能示范法 2.0(专家建议稿)〉发布》,载微信公众号“网络与信息法学会”,2024年4月16日上传;《〈人工智能法(学者建议稿)〉全文发布》,载微信公众号“人工智能产业发展联盟 AIIA”,2024年3月20日上传。

^{③⑤} 关于公权力在公共场所使用实时人脸识别系统的讨论,参见赵精武:《人脸识别技术应用的利益权衡与合法性认定》,载《法律科学》2024年第1期,第100-110页。

^{③⑥} 参见时婷婷:《女孩因“河南人”应聘被拒案开庭 浙江喜来登被判道歉赔偿1万元》,载澎湃新闻网, https://www.thepaper.cn/newsDetail_forward_5064609, 2024年1月31日访问。

^{③⑦} 参见龙卫球:《人工智能立法规范对象与规范策略》,载《政法论丛》2020年第3期,第95-106页。

^{③⑧} 参见[美]布莱恩·阿瑟:《复杂经济学:经济思想的新框架》,贾拥民译,浙江科学技术出版社2018年版,第4页。

者与提供者义务规范、监督管理、特殊应用场景、国际合作等。^{③⑨}无论是哪一个版本,均未采取单一的风险管理路径或某种单一理论模型规范人工智能赋能活动。笔者认为,不同的人工智能活动所涉问题分属不同的领域,涉及的主体结构、利益关系也有所不同。人工智能赋能应用活动具有高度场景化特征,没有任何一个“正确”的预期模型可以被假定为共同知识。因此,人工智能赋能应用活动的规范,不应过度追求理论化与体系性,而应在充分了解现实问题的前提下,设计能够满足场景化应用与动态调整需求的治理结构。

四、复式立法进路下人工智能规范框架的展开

人工智能活动具有科学性与赋能性,我国未来的人工智能立法也应在科技法与应用法复式定位下展开。一部立法具有双重定位,无疑极为考验立法者的智慧。一个简单的、逻辑清晰的治理框架虽然符合形式美感,但可能削足适履,不能有效地调整规范对象。作为科技法的《人工智能法》应重在将科技伦理内化于人工智能研发活动中,同时打破制度壁垒,促进人工智能科技的发展;作为应用法的《人工智能法》则应关注科技赋能导致的功能异化现象,借助灵活的规范配置,满足多样的规范需求。

(一) 科技法定位下的人工智能研发伴生性规范

人工智能研发活动属于专精科技活动,具有科学性。与之相应,对人工智能研发活动的规范,不能忽视科技系统独有的调整方式。此外,以科技活动为规范对象的人工智能法,应充分发挥科技促进法的功能,破除人工智能发展的制度障碍,促进人工智能科技成果转化。

1. 科技伦理义务化

科技不同于自然,自然中万事万物紧密关联,科技则是操作系统的因果性闭合。^{④⑩}科技系统的内在封闭性意味着,只有本系统成员才能充分理解其内部发生的意义沟通过程。计算机科学家最了解人工智能研发中的危险,也处于控制危害的最佳位置。作为科技法的人工智能法,应关注研发过程,要求科学家遵守科技伦理。

科技伦理是科技系统对科研活动的自我调整方式。科技伦理可以输送更具时效、更加贴合技术与商业模式的行为规范,实现尊重人工智能发展规律、促进人工智能场景创新与立法监管相协调,增强立法的社会适应性。^{④⑪}国家新一代人工智能治理专业委员会于2021年发布的《新一代人工智能伦理规范》就确立了强化自律意识、提升数据质量、增强安全透明和避免偏见歧视四项研发规范。唯须注意,伦理原则只有具体化、义务化,才能真正发挥作用。如果没有通过立法确定具体的人工智能活动规则,人工智能伦理原则将难以落地。^{④⑫}法秩序应通过设计研发伴生性规范的方式,将科研伦理内化于

^{③⑨} 参见前注^{③④},《〈人工智能示范法2.0(专家建议稿)〉发布》;前注^{③④},《〈人工智能法(学者建议稿)〉全文发布》。

^{④⑩} 参见[德]尼克拉斯·卢曼:《风险社会学》,孙一洲译,广西人民出版社2020年版,第129-133页。

^{④⑪} 参见陈亮:《人工智能立法体系化的困境与出路》,载《数字法治》2023年第6期,第11页。

^{④⑫} See Brent Mittelstadt, *Principles Alone Cannot Guarantee Ethical AI*, *Nature Machine Intelligence*, Vol.1:501, p.503 (2019).

科研过程。我国未来的人工智能法应规定记录义务、报告义务、安全管理和伦理委员会等制度。囿于篇幅,下文主要介绍研发人员应承担的记录和报告义务。

记录是科技系统自我观察与调试的基础。人工智能活动引发的风险具有不可预测性,唯有借助科学记录,才能在一定程度上理解正在进行的科研活动。欧盟委员会于2020年发布的《人工智能白皮书——追求卓越和信任的欧洲方案2020》就明确指出,人工智能设计者和运营者应当记录并保存下列信息:训练和测试的数据集,包括对主要特征的描述及挑选数据集的方式;在合理情况下,数据集本身;有关编程和训练方法,构建、测试和验证人工智能系统的过程和技术等,包括避免歧视的相关设计。^{④③}在大语言模型时代,人工智能科研工作者更应承担记录义务。只有以可理解的方式记录科研流程,才有可能“控制”大模型引发的不确定性。

记录义务并非仅存在于实验室,而应贯穿人工智能系统全生命周期。考虑到人工智能的自我学习、自我成长能力,人工智能活动的记录义务应有一定的特殊性。相较于工业时代的产品,人工智能技术在投入应用后更可能实现进化,因为来自现实场景中的数据将不断刺激人工智能的自我生长。基于仿生学设计的“种子人工智能”(Seed AI)能够不断改良自身。种子人工智能首先通过试验和试错、信息采集以及程序员协助等方式理解自己的运行逻辑,然后通过建立新算法与结构实现“递归性自我提升”(recursive self-improvement),从而获得自我进化能力。^{④④}从某种意义上说,整个社会都是人工智能系统的进化场所。欧盟《人工智能法》第12条第1款明确要求高风险人工智能系统应具备在运行时自动记录事件的能力,第2款则强调人工智能开发者将人工智能产品投入市场后需要承担产品追踪记录义务。只要人工智能系统具有自我学习能力,研发者就应承担持续的记录义务,以实现人工智能系统的全生命周期监测。

报告义务与记录义务相辅相成,是同一硬币的两面。如果根据记录发现人工智能模型存在错误或缺陷,研发人员应及时向监管机构汇报。欧盟《人工智能法》第62条第1款要求高风险人工智能系统的提供者发现严重故障时,应立即向监管机构进行汇报,报告时间应在发现严重故障之日起15日内。当强人工智能有望实现时,科研工作者更应及时报告,因为强人工智能技术可能是人类最后一个发明,会给作为发明者的人类带来巨大危险。当然,也有人指出,强人工智能的到来尚遥遥无期,人工智能末世论过于夸大人工智能的危险。^{④⑤}对此,或许只有时间才能分辨对错。真正需要思考的问题是,我们是否愿意接受因技术蝶变而自我毁灭的可能。当奇点时刻来临时,不应仅由科学家决定人类的命运,而应让包括科学家在内的全人类共同作出选择。但要让研发者之外的其他人也获得决定权,就必须使其能够实时了解研发进度。因此,人工智能研发者应负有重大

^{④③} See European Commission, White Paper on Artificial Intelligence - A European Approach to Excellence and Trust, Brussels, Feb.19, 2020, COM (2020) 65 final.

^{④④} 参见[英]尼克·波斯特洛姆:《超级智能:路线图、危险性与应对策略》,张体伟、张玉青译,中信出版社2015年版,第36-37页。

^{④⑤} See Kate Crawford & Ryan Calo, *There is a Blind Spot in AI Research*, Nature, Vol.538:311, p.312 (2016).

事项报告义务,以使监管机关可以据此及时介入人工智能科研活动,防止系统发生整体性崩坏。

目前,互联网监管部门正在尝试通过算法备案的方式强化对人工智能活动的监管。^{④⑥}但相较而言,算法备案不如记录与报告义务那么契合专精科技活动的科学性。通过算法备案的方式规范人工智能活动,并不符合科技法规范模式。算法备案实质是一种权力机关试图深入到科技系统内部、直接监管人工智能设计与运行的有益尝试,旨在获取平台设计部署的具有潜在危害和风险的算法系统的相关信息,以固定问责点,为今后的行政监管提供信息基础。^{④⑦}但是,备案可能成为变相的审批,不当干扰科技研发。实践中,部分监管部门以备案之名行审批之实,过度侵蚀科技系统的自主性。欧盟 1995 年颁布的《个人信息保护指令》(Data Protection Directive)第 18、19 条曾规定个人信息控制者和处理者使用自动化算法的备案义务。但在制定欧盟《通用数据保护条例》(GDPR)的过程中,欧盟委员会就指出备案义务增加了企业的官僚主义负担,并主张废除一般性的备案义务,改为要求全面记录个人信息处理活动;一旦监管机构要求算法使用人提供记录,算法使用人有义务及时、全面提供相应的记录。^{④⑧}2016 年通过的欧盟《通用数据保护条例》采纳了欧盟委员会的建议,于第 30 条规定中以记录和报告义务取代《个人信息保护指令》中的备案义务。此外,人工智能大语言模型具有不断进化的能力,将静态的、一时的算法予以备案,也不足以帮助监管部门了解算法的后续进化。更优的路径是强化人工智能开发者的持续记录与报告义务,借助科学系统自身的控制阀门实现治理目标。

2. 建立促进人工智能科技发展的数据制度

《人工智能法》以科技活动为调整对象,应在规范之余设计促进型制度,助力人工智能科技发展。数据利用问题是制约人工智能发展的卡脖子问题,不论大模型设计的技术水平有多高,数据训练数量与质量均对 AI 性能具有决定性影响。依数据来源主体不同,数据可以分为个人数据与非个人数据。目前人工智能训练使用个人数据的,只有在符合《个人信息保护法》规定的合法性基础的前提下才能实现,这严重限制了人工智能的数据训练;非个人数据的利用虽然无需受到个人信息保护法律的拘束,但也面临“数据孤岛”“数据垄断”等问题,加之知识产权制度的影响,人工智能训练难以汇聚利用商业数据、中国知网等知识平台的高质量数据。^{④⑨}

首先,人工智能立法可以借助可期待性同意规则,允许合理利用个人数据进行人工智能训练。《个人信息保护法》第 13 条第 1 款规定了七项处理个人信息的合法性基础,其中“个人的同意”是处理个人信息的核心规则。^{⑤⑩}依据第 14 条规定,同意必须由个人

^{④⑥} 例如,《互联网信息服务算法推荐管理规定》第 24 条第 1 款规定,具有舆论属性或者社会动员能力的算法推荐服务提供者应当履行备案手续。

^{④⑦} 参见张凌寒:《网络平台监管的算法问责制构建》,载《东方法学》2021 年第 3 期,第 22-40 页。

^{④⑧} Vgl. Spiros Simitis/Gerrit Hornung/Indra Spiecker (Hrsg.), Datenschutzrecht, Nomos Verlag, 1. Aufl. 2019, S. 804.

^{④⑨} 参见杨建军等:《人工智能法:必要性与可行性》,载《北京航空航天大学学报(社会科学版)》2024 年第 3 期,第 169 页。

^{⑤⑩} 参见杨合庆主编:《中华人民共和国个人信息保护法释义》,法律出版社 2022 年版,第 45 页。

在充分知情的前提下自愿、明确作出,但一方面,人工智能训练需要海量的个人数据,无法事先一一获得个人的同意;另一方面,同意机制在人工智能时代也可能成为个人的负担——个人信息主体在无法充分理解人工智能活动的情况下,被迫面临信息过载和决策过频的问题,将很难作出真正有价值的决定。^{⑤①}我国未来的人工智能立法可以尝试通过规定“可期待性同意”的方式,在减轻个人负担的同时,允许人工智能系统在符合期待的情境下收集与处理个人信息。早在2004年,美国学者就提出“情境诚信理论”,认为个人信息保护应当根据不同的情境而有不同的期待。如果行为不合期待,即便有个人明确的同意也构成对个人信息的侵犯;反之,即便个人没有事先明确同意,个人信息处理活动也是合法的。^{⑤②}例如,智能家居机器人在征得购买者同意处理其个人信息时,当然还会处理购买者家庭其他成员(如未成年人)、偶然使用者(如访客)等个人信息,此时人形机器人的数据处理行为应是符合期待的。借助“可期待性同意”规则,可以一定程度上柔化《个人信息保护法》中的同意规则的刚性,推动人工智能通过收集与处理个人数据实现科技蝶变。

其次,人工智能立法应借助数据访问权打破数据孤岛,并适当纾解人工智能利用非个人数据的知识产权限制。数据不会被耗尽,且能够为实现不同目的而被反复利用与共享,数据的这一属性使得数据共享作为常态成为可能。但在技术壁垒与商业模式的影响下,数据集中于大型平台之上,最终容易导致数据孤岛和数据垄断。对于个人数据,尚可以借助《个人信息保护法》第45条第3款规定的的数据可携带权实现数据流转;但对于非个人数据的流转,却不存在明确的规则。监管部门要求大型平台之间“互联互通”,但有过分侵犯营业自由的嫌疑;要求强强联合,却可能进一步促成数据垄断。^{⑤③}2022年公布的《欧盟数据法》专门规定了数据访问权(right of access to data),允许个人携转与产品或服务有关的非个人数据。^{⑤④}我国的人工智能立法可以参照规定数据访问权,将用户携转数据的对象从个人数据扩张到非个人数据,借助个人力量打破数据孤岛,促进数据聚合。此外,著作权也会对人工智能训练数据的汇聚和融合形成挑战。开放性网站是人工智能训练的重要数据来源。但是,开放性网站上的文字、图片、声音等很可能已经受到《著作权法》的保护。例如,微博、知乎上发表的文字或问答,微信、小红书、抖音上用户上传的音乐、图片、视频,在达到独创性的门槛后,都可能获得著作权保护。一旦人工智能企业利用这些内容数据,就有可能构成著作权侵权。^{⑤⑤}为了鼓励人工智能科技发展,日本政府官员表示,日本法律不会保护人工智能集中使用的原材料版权。^{⑤⑥}但笔者认为,

⑤① 参见林涓民:《论个人信息主体同意的私法性质与规范适用——兼论〈民法典〉上同意的非统一性》,载《比较法研究》2023年第3期,第144页。

⑤② See Helen Nissenbaum, *Privacy as Contextual Integrity*, *Washington Law Review*, Vol.79:119, p.127 (2004).

⑤③ 参见周汉华:《互操作的意义及法律构造》,载《中外法学》第3期,第606页。

⑤④ 参见丁晓东:《论数据来源者权利》,载《比较法研究》2023年第3期,第15-17页。

⑤⑤ 参见丁晓东:《论人工智能促进型的数据制度》,载《中国法律评论》2023年第6期,第179页。

⑤⑥ 参见《严苛版权保护阻碍 AI 技术发展? 日本政府重申: AI 所用数据不受版权保护》,载新浪科技, <https://finance.sina.com.cn/tech/roll/2023-06-02/doc-imyvaxapa8336048.shtml>, 2024年6月26日访问。

一概允许人工智能使用他人作品进行训练,未免过于极端了。更为妥适的做法是借鉴欧盟 2019 年《单一数字市场版权指令》第 4 条的规定,一方面规定以人工智能训练为目的使用作品构成著作权保护的例外,另一方面允许著作权人以可机读的方式明确反对人工智能系统使用其作品进行训练。考虑到人工智能将使用海量的作品进行训练,我国未来的《人工智能法》还可以借鉴欧盟《人工智能法》第 53 条第 1 款第 c 项的规定,要求相关主体制定专门性的符合著作权法的政策,帮助著作权人行使反对权。

(二) 应用法定位下的人工智能赋能规制型制度

科学研究的主要目的是发现知识,控制危险仅为其考虑的次要因素,而主要目的会天然地弱化次要因素。例如,企业的主要宗旨是营利,企业文化形成仅为企业的次要追求,当主旨与次要目的发生冲突时,企业会为了营利而放弃建设或改变企业文化。同样,科技系统可能会为了促进科技进步,而忽视权益保护、社会正义等价值,因此,法律作为社会回应新型挑战的有力工具,有必要进行更为积极的干预,人工智能立法因而迫在眉睫。倘若人们将法律系统视为一种应对危机的方式,那么法律系统就是全社会的免疫系统。^{⑤7}但是免疫并不总是有益,有时会徒增烦恼。法律规范可能因过于刚性而阻碍人工智能产业的发展。是以,对人工智能应用的规范既需要借助法律系统,也应留出必要的接口推行实验主义治理,构造一种灵活的人工智能赋能规制型框架。

1. 规范人工智能多场景应用的权利义务机制

合法与非法是法律系统的基本符码,只是人工智能的应用呈现为从场景到场景的现象罗列,增加了设计统一规则的难度。鉴于法律系统借助抽象的权利义务制度调整复杂的生活,人工智能立法应侧重于识别与总结相关主体的权利与义务类型。^{⑤8}

(1) 人工智能系统相对人的应有权利

为了捍卫人的中心地位,人工智能系统相对人应享有一系列保护性权利。欧盟在起草《通用数据保护条例》时就已经明确指出,采取基于风险的规制路径无法充分支撑起欧盟数据保护的法治框架,无论在数据处理过程中产生的风险程度如何,法律都应赋予数据主体权利。^{⑤9}相较于个人信息保护,人工智能涉及的问题更为复杂、更加场景化,因而更应重视私权的作用。

首先,人工智能系统相对人的知情权应受法律保护。知情权是确立以人为本的人工智能善治的前提与基础,不论是在何种场景下,个人的知情权均应受到尊重。如果个人都不清楚是否已经进入人工智能生态系统,就很难产生捍卫个人权利的意识。人工智能相对人应有权了解自己的信息是否在被人工智能系统处理,应有权获知人工智能系统

^{⑤7} Vgl. Niklas Luhmann, *Das Recht der Gesellschaft*, 1. Aufl. 1993, S. 565.

^{⑤8} 除了权利义务之外,人工智能致损的民事责任承担问题当然也是人工智能立法的重要问题,受篇幅所限,本文暂不讨论责任承担问题。目前学界对于人工智能致损的民事责任的讨论,参见吴汉东:《人工智能时代的制度安排与法律规制》,载《法律科学》2017年第5期;冯珏:《自动驾驶汽车致损的民事侵权责任》,载《中国法学》2018年第6期;郑志峰:《诊疗人工智能的医疗损害责任》,载《中国法学》2023年第1期。

^{⑤9} See Article 29 Data Protection Working Party, Statement on the Role of a Risk-Based Approach in Data Protection Legal Frameworks (Adopted on 30 May 2014), 14/EN WP218, p.2.

的预设功能、局限性、不良影响等信息。当人工智能系统具有一定程度上的操纵性时,系统提供者更应将这一情况告知相对人,使其意识到自己的自由意志有被扭曲的可能。即便是治疗性医疗人工智能,在修复心理创伤之前,也应告知病患,尊重病患的自主决定权(《民法典》第1219条)。

其次,个人在了解相关信息后,是否享有解释请求权、应用拒绝权等权利,则应视场景而定。其一,不同类型的人工智能系统对应不同的个人权利。应用场景中的人工智能活动大致可以分为决策型与辅助型两类。决策型人工智能代替人快捷地作出判断与决定,典型如信用评级(贷款或信用卡发放)、防疫健康码、“秒批”、税务抵扣系统等。辅助型人工智能仅提供辅助支持作用,如个性化推荐、医学图像处理、智能客服机器人等。决策型与辅助型人工智能系统在不同应用场景中对相对人的影响不同,权利配置也应有所不同。决策型人工智能系统直接影响相对人的法律利益,相对人应有权利请求解释决策是如何作出的;辅助型人工智能系统并不直接作出影响相对人利益的决定,但可能潜移默化地影响相对人的思维方式,因此相对人应有权利事先拒绝该系统的应用。是以,个人的解释请求权应主要针对决策型人工智能系统,应用拒绝权则主要以辅助型人工智能系统为对象。其二,无论是解释请求权还是应用拒绝权,均应区分公私应用场景分别判断权利的适用空间。以解释请求权为例,公权力借助人工智能系统进行行政或司法行为的,相对人应有权请求解释人工智能系统的具体决策;商业机构使用人工智能系统进行商业活动的,相对人是否享有解释请求权,则应综合衡量解释可能性以及商业秘密保护等因素进行判断。^{⑥0}当公权力运用人工智能系统进行决策时,如果不能解释决策,公共行为的正当性和合法性都将遭受质疑。因为与提升行政与司法效率相比,公民基本权利保护的价值位阶更高。与之不同,商业人工智能系统的逻辑、参数、特征的权重以及分类等是企业的商业秘密,未必应屈从于个人的算法解释请求权。德国联邦最高法院在2014年的判决中明确指出,数据主体不得请求披露权重、评分公式、统计值和参考组等信息,因为这触犯了企业的商业秘密,披露企业的商业秘密也明显不符合立法目的。^{⑥1}当个人要求解释商业自动决策算法的决定时,算法解释请求权如何与保护商业秘密的法律要求相协调,将是司法审判的难点。

最后,我国未来的《人工智能法》还应专门规定请求人工沟通的权利。只有保证个人享有表达意见、获得人为干涉的权利,人才不会沦为机器的客体。欧盟《通用数据保护条例》第22条第3款专门规定了算法相对人请求人工干预的权利。相较而言,我国《个人信息保护法》第24条第3款仅规定请求说明算法和拒绝算法决定的权利,并未规定人工沟通的权利。在人工智能时代,受影响个人有权与机器背后的人进行有意义的沟通,是捍卫人的主体地位的基本要求。我国未来的人工智能立法应专门规定人工智能系统相对人享有人工沟通权。

^{⑥0} 参见林涓民:《〈个人信息保护法〉中的算法解释权:兼顾公私场景的区分规范策略》,载《法治研究》2022年第5期,第48-58页。

^{⑥1} Vgl. BGH, Urteil vom 28. Januar 2014 - VI ZR 156/13, BGHZ 200, 38.

（2）人工智能系统提供者与使用者的不同义务

人工智能系统相关义务主体大致可分为提供者与使用者两类,由于二者对作为赋能工具的人工智能系统的影响不同,所应承担的义务也应有所不同。人工智能系统提供者负责设计开发产品并将之投入市场,应承担信息公开、人工监督等义务;人工智能系统使用者则因可以影响具体的输出结果,应承担谨慎使用、保存日志等义务。

第一,不论是决策型还是辅助型人工智能系统,也不论应用于何种场景,人工智能系统提供者均应一般性地承担特定的法律义务。这些义务包括但不限于信息公开、人工监督、保障系统稳定性等。首先,人工智能系统提供者应承担信息公开义务,以使下游合作者和用户能够及时了解系统相关信息。这些信息主要包括提供者的身份和联系方式、人工智能系统的功能及其局限性、测试结果、对目标人群可能的不利影响以及如何进行系统维护等。人工智能系统提供者应通过说明性文件(技术文件)的方式,将上述信息告知相对人。其次,人工智能系统设计应留有适当的人机交互界面,以便对人工智能系统进行有效的监督。人工监督机制旨在及时发现人工智能系统的功能异常和性能突变迹象,纠正自动化偏差,手动控制或逆转输出结果。当出现重大隐患时,人工智能系统提供者应有义务“一键关闭”人工智能系统,避免发生不可逆的后果。再次,人工智能系统提供者应当维护人工智能系统的稳定性和安全性,尤其是确保人工智能系统具有自我复原能力。恶意第三方可能通过“脏数据”、系统漏洞等试图改变人工智能系统的运作模式、性能以及输出结果。人工智能系统提供者可以有针对性地通过提供备份、补丁和安全解决方案等技术方案保护网络安全。最后,我国作为《联合国残疾人权利公约》(UNCRPD)的签署国,有义务确保残疾人同等地使用人工智能系统。在人工智能时代,弱势群体同样有权不受限制地使用人工智能技术。人工智能服务提供者应当在系统设计时就考虑到弱势群体的需求,提供必要措施,以使弱势群体不至于被技术排除在外。

第二,人工智能系统使用者应承担谨慎使用、保存日志等义务。人工智能系统的提供者与使用者并非同一主体的,人工智能系统使用者直接操作人工智能系统,系处于影响系统结果和监测程序运行的最佳位置,因而应承担与之相称的法律义务。其一,人工智能系统使用者应谨慎使用人工智能系统。使用者输入数据可能影响到系统功能的,应确保输入数据的正当性。例如,生成式人工智能以“用户输入+机器输出”的模式提供服务,生成内容的性质、价值取向很大程度上取决于用户输入指令。人工智能系统使用者应当遵守人工智能系统的预设功能,不能通过模糊词等欺骗性手段实现非法目的。^②其二,如果日志对诊断系统的运行状态和故障是不可或缺的,人工智能系统使用者就有保存相关日志的义务。只是考虑到存储成本问题,日志保存宜设置期限限制(如六个月)。其三,人工智能系统使用者如果发现人工智能在具体场景中可能产生歧视、侵犯人格尊严等问题,应当暂停使用该系统,并将问题反馈给人工智能系统提供者。如果存在

^② 例如,虽然 OpenAI 在使用政策中明确禁止生成恶意软件,很多使用者依然可以通过输入提示欺骗 ChatGPT 为恶意软件应用程序编写代码。

重大隐患,应当同时上报国家监管机构,以防止发生不可逆的后果。

我国未来的人工智能立法除了规定上述一般性的义务外,还可以挑选重要的具有成熟监管经验的人工智能场景,有针对性地规定人工智能系统提供者和使用者的特殊义务。例如,我国未来的人工智能立法可以考虑对公权力使用人脸识别等人工智能系统作出专门规定。反之,如果一些场景中的人工智能相关主体义务尚有争议,不妨适当留白,交由监管机构出台专门性法律文件进行规范。^{⑥3}

2. 调整人工智能复杂应用活动的实验主义治理

新兴科技对治理产生的挑战主要体现为未知与不确定性。未知是指决策者可能并不了解真实的、不断变化的问题;不确定性则是指决策者不具有解决问题的能力,需要不断研究和改进治理方案。^{⑥4}即便在充分调研的基础上配置权利义务制度,人工智能复杂的、多场景的应用活动也可能使得一时有规则的规则转瞬变成科技发展的障碍。是以,法律常常面临与新兴技术的步调失调问题。为了破解这一难题,我国未来的《人工智能法》可以考虑推行实验主义治理模式,构造试错与纠错的动态机制。

实验主义治理(Experimentalist Governance)起源于对欧盟治理政策的总结性思考和提炼。^{⑥5}实验主义治理大致包括四项内容:大致的框架目标、参与者的自由裁量、基于同行评议的动态评估,以及根据评估结果的反馈修正。具体而言,决策机构在充分听取利益攸关者的意见后,确立一个开放式框架性目标;授权相关机构以较大的自由裁量权,根据具体情况调整策略;负责机构应定期汇报治理绩效,并经由同行评估判断妥适性;负责机构如果没有取得良好进展,则应当根据评估结果提出合理的改进计划。在该治理模式下,目标是可变的,既定的规则是不存在的,它们都会基于评估结果被修正,通过在临时性目标设置与修正之间反馈迭代、循环往复,经由“共同学习”最终寻找妥适的应对挑战方案。^{⑥6}实验主义治理强调构建多元性、开放性和互动性的治理系统,能够弥补传统科层式治理模式的主体单一性、体系封闭性和过程单向性等不足,并更好地应对不断发展、调整的人工智能应用活动带来的挑战。

实验主义治理突出体现为一定程度的纵向放权,使得监管机构能够进行监管实验、积累监管经验。在实验主义治理模式下,立法者设计人工智能监管的大致目标,授予监管机构自由裁量权,监管机构定期评估并汇报治理绩效,并根据相关主体的建议调整监

^{⑥3} 以无人驾驶汽车被动接管规则为例。2016年3月修订的《维也纳道路交通公约》第8条增加1款,明确车辆系统需要能够被驾驶员接管或者关闭。2017年6月,德国通过《道路交通安全法第八修正案》,新增的第1a条第2款也规定自动驾驶汽车应能随时被驾驶人接管。但也有研究指出,被动接管规则有悖于自动驾驶汽车作为替代型人工智能的设计初衷,只会徒增事故,且阻碍自动驾驶汽车的商业化进程。参见郑志峰:《论自动驾驶汽车被动接管规则》,载《华东政法大学学报》2023年第3期,第59-71页。

^{⑥4} See Charles F. Sabel & Jonathan Zeitlin, *Experimentalism in the EU: Common Ground and Persistent Differences*, Regulation & Governance, Vol.6:410, p.410-426 (2012).

^{⑥5} See Charles F. Sabel & Jonathan Zeitlin eds. *Experimentalist Governance in the European Union: Towards a New Architecture*, Oxford University Press, 2010, p.1.

^{⑥6} See Jonathan Zeitlin, *EU Experimentalist Governance in Times of Crisis*, West European Politics, Vol.39:1073, p.1073-1094 (2016).

管策略。监管沙箱（Sandbox）是推进实验主义治理的典型设计。监管沙箱是一种由监管机构依据法律规定设立的受控测试环境，在限定时间内允许人工智能系统进行开发与测试。人工智能活动即便违反现行法律法规，只要符合监管沙箱的要求，相关人员也不会被问责。英国科学创新与技术部指出，监管沙箱能够帮助迅速将新产品和服务推向市场，产生经济和社会效益；检验监管框架在实践中的运行情况，揭示需要解决的创新障碍；并确定监管框架需要适应的技术和市场的调整方向。^{⑥7}公权力机关基于一些特殊的考虑，会希望在特定地域或领域加速人工智能产业的发展。例如，日本就专门划出特定地域和场景布局人工智能产业，如福冈（道路交通）、关西（无线电）、京都（数据保护）、筑波（安全治理和税收规制）等。我国未来的人工智能立法应授权监管部门在特定地域与领域设立监管沙箱，从而摆脱“操之过急”或“听之任之”的弊端。监管部门可以根据评估结果，适当调整监管沙箱政策，在摸索中寻找妥适的治理手段。此外，立法还可以通过授权监管机关出台专门规范性文件的方式推行实验主义治理。在高新科技领域，中小企业往往是创新的主力。正如凯文·凯利指出的，未来在人工智能领域大放异彩的企业，大概率不会是现在的大型企业，而是某个不起眼的小公司。^{⑥8}欲使中小企业完成蝶变，就应在制度层面提供额外的支持。如同《个人信息保护法》第62条第2项授权国家网信部门出台专门针对小型个人信息处理者的规则与标准一样，我国未来的人工智能法也可以赋权监管机构出台专门的规范人工智能中小企业的文件，以促进科技创新。例如，监管机构可以授予中小型人工智能企业对人工智能监管沙箱的优先访问权；可以提供专项资金，解决中小企业的融资难题；可以适当降低义务标准，或者在充分保障个人权益的基础上，提供一定程度上的责任豁免等。监管机构应定期评估规范性文件的效果，并根据同行评估不断调整方案，以真正满足中小型科技企业的诉求。除上述规则外，日落规则、智能预期草案等设计，都可以成为贯彻实验主义治理的有益尝试。^{⑥9}

五、余 论

目前世界正处于第四次工业革命全面来临前夕。第一次工业革命和第二次工业革命都是围绕动力提升展开，第三次工业革命是信息革命，第四次工业革命是智能革命，与第三次工业革命既有关联，也有本质不同。第四次工业革命是对人类智能的模拟与提升，围绕智能展开，其核心是人工智能技术的研发与赋能推广。中国未来的人工智能立法必然应以促进人工智能可控发展为目标，否则中国将再次落后于世界浪潮。

^{⑥7} See Department for Science, Innovation & Technology, *supra* note 19, 60.

^{⑥8} 参见〔美〕凯文·凯利：《5000天后的世界》，潘小多译，中信出版社2023年版，第21页。

^{⑥9} 日落规则（Sunset Rules）是指特定法律只在一定时间段内有效，由此立法者可以针对特殊的问题而为“权益之计”；智能预期草案（Intelligent Anticipatory Drafting）是指立法者可以允许进行特定科技开发和应用，但是政府一旦发现其中存在问题，就可以要求企业或研究机构在一段时间内停止开发或应用。参见龙卫球、林涓民：《我国智能制造的法律挑战与基本对策研究》，载《法学评论》2016年第6期，第2页。

欧盟《人工智能法》采取风险管理路径,形成了以监管为核心的规范框架。但欧盟采取风险管理单一进路规范人工智能活动,认为风险是静态的、确定的与可预测的,并不符合复杂、动态发展的世界。更为科学的做法应是基于人工智能活动的双重属性,明确人工智能法的科技法与应用法复式定位,在规范人工智能活动的同时最大限度地促进人工智能技术与产业的发展。笔者相信,尊重科技自主、强调灵活规范的中国《人工智能法》将成为人工智能立法的一种新范式,为世界人工智能治理提供来自东方大国的智慧。

Abstract: Which path to be adopted to regulate Artificial Intelligence (AI) activities, is the central question of AI legislation. The risk management approach suffers from the difficulties of assessing and categorizing risks and letting damages happen, which should not be considered a perfect choice for AI legislation. Unlike previous scientific and technological activities, AI activities have the dual attributes of professional technology activities and enabling technology activities. The AI law, which takes AI activities as the object of regulation, should not be guided by one single theory, but should comply with the dual position of scientific law and application law. Under the dimension of scientific law, the AI law should respect science autonomy and help internalize science ethics in AI research activities, while breaking down institutional barriers and designing facilitative rules to help the development of AI technology. Under the positioning of application law, the AI law should concern functional alienation caused by technology-enabled scenarios. The AI law should, on the one hand, rely on the abstract rights and obligations tools of the legal system, especially through the provision of new types of rights, to build an elastic normative framework to respond to the differences in the sequence of values in different application scenarios; on the other hand, follow the steps of experimentalist governance with the help of regulatory sandboxes, enabling legislation and others, dynamically adjusting regulatory programs to meet the flexible governance needs of AI-enabled application activities.

(责任编辑:任彦)