

赋能型人工智能治理的理念确立 与机制构建

张吉豫^{*}

内容提要 人工智能治理已成为国家和社会治理的前沿问题和重要领域。然而,当前在人工智能科技创新、风险防控、企业自治、政府监管、社会监督、国际协作等方面都亟需加强能力建设,必须把提升人工智能安全可信发展的能力作为人工智能治理的第一要务,构建“赋能型人工智能治理”的理念和机制。以此为目标,应坚持以人为本、发展导向的赋能型人工智能治理核心理念,以及从中发展出的智能向善、包容审慎、敏捷治理、可持续发展等基本理念。应建设以法治为核心的赋能型人工智能治理机制以及法治统领下的各项具体机制,如完善法律治理与技术治理相统合的机制,建立多元主体沟通协作的共治机制,构建与人工智能发展相适配的“避风港”机制,建立敏捷互动、激励向善发展的动态监管机制,建设人工智能安全保险等社会保障机制。

关键词 人工智能 赋能型治理 法治化 以人为本 发展导向

目 次

- 一、引言
- 二、赋能型人工智能治理的理论依据
- 三、赋能型人工智能治理的理念确立
- 四、赋能型人工智能治理的机制构建
- 五、结语

^{*} 中国人民大学法学院、未来法治研究院副教授。本文系科技部2022年度新一代人工智能国家科技重大专项“可信人工智能立法制度建设研究”(项目批准号:2022ZD0120101)和教育部2022年度哲学社会科学研究重大专项项目“坚持建设中国特色社会主义法治体系,深入推进全面依法治国实践研究”(项目批准号:2022JZDZ002)的阶段性成果。

一、引言

21 世纪以来,人工智能技术迅猛发展,正引领新一轮产业革命,日益成为决定国家竞争力和国家安全的重大战略性技术,同时其广泛使用也给人类带来全新的风险挑战。这使得人工智能治理成为国家治理体系的重要组成部分。党的二十届三中全会通过的《中共中央关于进一步全面深化改革 推进中国式现代化的决定》(以下简称《决定》)指出,“完善推动新一代信息技术、人工智能、航空航天、新能源、新材料、高端装备、生物医药、量子科技等战略性新兴产业发展政策和治理体系”^①。如何科学有效进行人工智能治理,树立什么样的治理理念,建构什么样的治理机制,形成什么样的治理格局,无疑是法学界和科技界必须回答的时代之问和世界之问。

2024 年 7 月 1 日,第 78 届联合国大会协商一致通过中国主提的加强人工智能能力建设国际合作决议,140 多国参加决议联署。^②这充分反映了当前人工智能能力建设的重要意义。对此,本文从不发展是最大的不安全、不充分发展是最大的隐患等判断出发,指出人工智能的安全可信是一种需要建立在高发展水平基础上的能力,并针对现阶段亟需能力建设的切实情况,提出“赋能型人工智能治理”的概念。本文将在论述赋能型人工智能治理理论依据的基础上,指明赋能型人工智能治理的核心理念和基本理念,并提出健全以法治为核心的赋能型人工智能治理新机制。赋能型人工智能治理新理念和机制的结合,必将形成人工智能治理新范式新格局。

二、赋能型人工智能治理的理论依据

以人工智能为首的新兴科技正有力推动着社会各个领域的迭代升级,将在中国式现代化进程中发挥至关重要的作用。在此背景下,2024 年《政府工作报告》明确提出开展“人工智能+”行动的工作规划。^③人工智能将日益广泛地应用在社会生活和国家治理各个方面。随着应用领域和影响力不断扩大,人工智能系统的潜在风险也逐渐凸显。这对人工智能治理提出了促进发展和保障安全的双重需求。现代社会的运行和人的发展建立在信任基础之上,而信任又是建立在对系统的可依赖性的信心之上。^④社会需要引领和保障人工智能的安全可信发展。

在迈进智能时代的关键历史时期,人类社会需要形成与时代相匹配的运行能力,然

① 《中共中央关于进一步全面深化改革 推进中国式现代化的决定》,2024 年 7 月 18 日中国共产党第二十届中央委员会第三次全体会议通过。

② 参见《联大通过中国提出的加强人工智能能力建设国际合作决议》,载中国政府网, https://www.gov.cn/yaowen/liebiao/202407/content_6960524.htm, 2024 年 7 月 2 日访问。

③ 参见李强:《政府工作报告》,2024 年 3 月 5 日在第十四届全国人民代表大会第二次会议上。

④ 参见[英]安东尼·吉登斯:《现代性的后果》,田禾译,译林出版社 2022 年版,第 39、93、102 页。

而,当前人工智能企业创新发展能力、风险识别和防控能力、政府监管能力、社会监督和正确应用人工智能的能力、国际协作能力等都存在明显不足,缺少充分的信息和成熟的治理经验,社会治理面临全新的挑战。应当看到,这样的挑战是人工智能这种引领产业革命的颠覆性、战略性科技发展初期所必然产生的。创新发展、风险防控等能力不可能凭空获得,需要在发展过程中积极建设。可以说,当前人工智能发展中的主要矛盾,是人工智能安全可信发展的巨大需求与人工智能发展及治理能力不足之间的矛盾。面对这种复杂情景,有必要创新人工智能治理理念和机制,将解决人工智能安全可信发展的能力缺口作为治理要解决的重点问题。因而,在回答我国需要什么样的人工智能治理机制时,首先要明确当下的治理目标,即通过科学治理来赋能人工智能安全可信发展,使得人工智能科技与社会治理能力同步提升,最终使社会建立起可以在人工智能时代良好运行的能力,让人工智能真正服务于人民幸福和人类福祉的提升。因此,本文提出建构着眼于社会能力提升的“赋能型人工智能治理”,具体理据包括如下方面:

(一) 人工智能所处的发展阶段

在动态的社会技术变革过程中,科技发展、治理模式与规范观念之间存在相互作用。^⑤治理模式的确定需要把握科技规律、经济规律和治理规律,符合所处发展阶段的特征和需求。当前我国人工智能发展的阶段性特征对人工智能治理模式提出了基本要求。

首先,我国人工智能发展处于世界前列,缺少可以借鉴的成熟治理经验。过去我国在许多领域的发展晚于发达国家,可以参考国外在实践中产生的信息经验来进行科学决策。然而,随着我国人工智能科技进入世界领先行列,从“跟跑”变为现在的“并跑”乃至在一些方面的“领跑”,几乎没有经过实践检验的成熟治理经验可以借鉴。“传统政策工具进入信息‘盲区’,这是更好发挥政府作用面临的严重挑战。”^⑥在此情况下,需要坚持对人工智能新业态新模式的包容和普惠赋能,升级治理模式和治理能力,加强及时获得风险信息、治理机制及其能效信息的能力建设。

其次,当前人工智能仍处于高速发展阶段,并可能出现难以预测的突破性进展,这也使得传统基于充足信息的静态化治理机制难以有效适应人工智能发展。正如2022年末基于大模型的人工智能发展,使得当时已经形成较高程度共识的欧盟《人工智能法案》受到挑战,不得不进行修改,专门增加了一章针对通用人工智能模型的管理要求。^⑦这突出反映了该法案当时的局限性,体现了相对静态的治理机制回应技术发展的能力不足,因而需要建设更为敏捷动态的治理机制,以增强对变化迅速的科技发展的回应能力。

再次,当前人工智能虽然取得了显著进步,但整体而言仍处在发展早期。科技界对

^⑤ See Ronald Leenes, *Regulating New Technologies in Times of Change*, in Leonie Reins ed., *Regulating New Technologies in Uncertain Times*, ASSER Press, 2019, p.8-9.

^⑥ 黄先海、宋学印:《赋能型政府——新一代政府和市场关系的理论建构》,载《管理世界》2021年第11期,第47-48页。

^⑦ See Regulation of the European Parliament and of the Council of Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), at <https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/en/pdf> (Last visited on May 25, 2024).

于大模型“智能涌现”现象的理解仍不透彻,^⑧用于保障人工智能安全和伦理价值对齐的技术发展更是非常不充分。这种阶段性特征一方面决定了无法要求人工智能立即达到理想的安全可信状态,另一方面也提示我们不能仅以当下的人工智能发展情况进行评价,而是应该以动态的、发展的眼光来进行系统研判,特别是重视和推动科研群体及掌握先进科技的人工智能企业不断发展技术能力,以不断解决发展中存在的风险问题。

最后,我国人工智能发展虽处在世界前沿,但仍落后于美国。过往的百余年历史让我们深刻体会到科技发展对国家命运的重大影响。面对国际人工智能科技发展可能带来的影响,自身科技的高度发展是维持国家竞争力和国家安全的根本能力。因此,我国必须重视赋能人工智能发展以及安全能力的提升。

综上所述,当前我国人工智能发展的阶段性特征决定了需要赋能型人工智能治理。过去那种通过借鉴先发展国家的经验、在相对充足的信息基础之上进行静态规则制定和治理的模式难以适应当前的人工智能治理需求,应当在包容和普惠赋能的基础上,发展敏捷动态的治理机制和提升治理能力,特别是注重赋能人工智能合规科技、监管科技的同步发展,以不断提升安全能力。

(二) 人工智能发展的机遇与挑战

人工智能是引领未来的战略性技术。许多国家都积极抢抓人工智能发展的战略机遇,构筑人工智能发展的先发优势。我国在2017年发布了《新一代人工智能发展规划》(国发〔2017〕35号),以此牵引创新型国家和科技强国建设。在党中央科学决策和国务院规划部署下,我国人工智能产业发展迅速,处于良好的战略机遇期。但同时也要看到,我国人工智能发展还面临一系列挑战。美国在新一轮人工智能发展中仍处在引领地位。根据斯坦福大学发布的研究报告,美国2023年在人工智能领域的投资位于世界首位,高出中国将近8.7倍。^⑨由于新一代人工智能仍处于发展初期,其创新发展和应用存在很大的不确定性,依赖数据、算法、算力、产业生态环境、营商环境和社会应用环境等多重因素,人工智能投资必然是高风险活动。只有当期期望收益高于期望成本时,理性投资人才可能对创新进行投资。^⑩如果人工智能创新活动的成本收益衡量缺乏稳定的良好预期,企业会对重要的投资和研发信心不足,面临是将资金投入人工智能科技还是其他项目的选择。供应链可及性和稳定性、进入市场的门槛、知识产权保护情况和竞争环境等营商环境因素都是市场主体在制定规划时需要考虑的制度要素和社会条件,是国家治理能力、体制机制、社会环境等因素的综合反映。^⑪

习近平总书记强调,“科技领域是最需要不断改革的领域”“推进自主创新,最紧

⑧ See Jason Wei et al., *Emergent Abilities of Large Language Models*, Transactions on Machine Learning Research, No.8, p.11 (2022).

⑨ See Stanford Institute for Human-Centered Artificial Intelligence, *Artificial Intelligence Index Report 2024*, p.30, at https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_2024_AI-Index-Report.pdf (Last visited on March 15, 2024).

⑩ See Robert P. Merges, *Uncertainty and the Standard of Patentability*, High Technology Law Journal, Vol.7, p.20-25 (1992).

⑪ 参见中国行政管理学会课题组:《聚焦市场主体关切 持续打造市场化法治化国际化营商环境》,载《中国行政管理》2021年第8期,第6页。

迫的是要破除体制机制障碍,最大限度解放和激发科技作为第一生产力所蕴藏的巨大潜能”。^⑫从国内来看,还存在一些制约人工智能发展能力的制度问题。第一,在数据汇聚利用方面,目前法律制度中还存在机器学习合理使用规则缺失、反不正当竞争法适用中对数据爬取利用过度限制、公共数据流通利用规则不明确不统一等问题。^⑬第二,人工智能产业发展仍存在一些领域市场准入困难、试验区域有限、法律规则模糊、规制成本偏高、侵权责任过重等问题,需要进行责任科学界定和监管机制优化。第三,人工智能创新的知识产权保护制度尚不完善,对于积极研发人工智能价值对齐机制及规制技术的企业缺少有效激励机制。^⑭第四,由于一些人工智能应用曾出现安全风险、算法歧视、泄露个人信息、损害劳动者权益等问题,加上社会公众对失业和科技异化的担忧、风险防御能力和监管能力不足等现状,影响了社会对于人工智能的信任,^⑮限制了人工智能的应用。此外,从国际上看,美国等一些国家以所谓的“国家安全”名义打压中国信息科技企业,限制其产品和服务,不断升级对中国的芯片出口限制,试图遏制我国在人工智能等前沿科技领域的创新发展,也使我国当前人工智能发展面临严重挑战。

无论是利用好战略机遇,还是应对国内外挑战,都迫切需要面向人工智能发展中能力不足的问题要点,构建赋能型人工智能治理机制,为人工智能创新发展提供更加有利的制度条件,为企业投入人工智能创新提供激励和更加稳定的制度预期。同时,在涉外法治建设中,应建立健全能够有效应对外国遏制和打压的新机制,以此提高我国参与全球人工智能治理体系建设的能力,助力我国人工智能企业和行业抢占创新发展的制高点,为中国式现代化提供强大的科技支撑。

(三) 人工智能风险的基本特征

随着人工智能的应用领域不断扩大,人工智能的潜在风险也日益受到社会关注,很多研究对人工智能的风险问题和应对措施展开了积极探索。^⑯在构建人工智能风险治理机制时必须注意,人工智能带来的风险具备现代社会意义上公共风险的一些共同特征。

首先,人工智能风险具有公共性和规模性。这种现代社会的公共风险在很大程度上超出了个体风险承担者的直接理解和控制。^⑰公民个体通常欠缺对这种公共风险的认知能力、预防能力、谈判能力,因此往往难以在充分理解风险的基础上做出理性选择,完全由公民在意思自治的基础上自担风险的合理性基础被动摇,带来了政府介入和依法治理的必要性。

^⑫ 参见习近平:《在中国科学院第十九次院士大会、中国工程院第十四次院士大会上的讲话》(2018年5月28日),人民出版社2018年版,第13页。

^⑬ 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175-191页。

^⑭ 参见严驰:《论人工智能的激励型立法——基于〈拜杜法〉的思考》,载《人工智能》2024年第2期,第95-96页。

^⑮ 参见苏宇:《算法规制的谱系》,载《中国法学》2020年第3期,第167-169页。

^⑯ 参见苏宇:《大型语言模型的法律风险与治理路径》,载《法律科学》2024年第1期;刘金瑞:《生成式人工智能大模型的新型风险与规制框架》,载《行政法学研究》2024年第2期。

^⑰ See Peter Huber, *Safety and the Second Best: The Hazards of Public Risk Management in the Courts*, Columbia Law Review, Vol.85(2), p.277 (1985).

其次,人工智能风险具有两面性。任何一种创新活动都伴随着未知,在带来风险的同时,也提供着新的发展机遇。吉登斯描述了既包括机会与创新,也包括安全与责任的“风险矩阵”。他指出,“风险不只是某种需要进行避免或者最大限度地减少的负面现象;它同时也是从传统和自然中脱离出来的、一个社会中充满动力的规则”;“机会与创新是风险的积极一方”,“对风险的积极参与是社会与经济动员的一个必要成分”。^⑮例如,辅助驾驶和自动驾驶技术在可能带来新型交通事故风险的同时,也可能极大减缓因人类驾驶员的疲劳驾驶、反应迟缓等情况造成的风险,从总体上提高交通安全。^⑯因此,对风险的态度伴随着价值判断,融合着针对一个事物或行为的收益和损害所进行的判断和比较。在对损害可能规模发生之概率的理性计算的基础上,风险可能成为提供给人们的机会。^⑰我们要以辩证观念和辩证思维正视人工智能风险的两面性,促进科技创新对社会的正面价值和积极意义。

最后,人工智能风险具有一定的可控制性。人工智能带来的风险是一种“人为风险”。社会之所以允许这种人为风险存在,一方面是因为其具有两面性,能带来促进社会发展的积极意义;另一方面是因为人们相信这种风险在很大程度上能够通过对于人们活动的引导和规范来从结构上得到控制,社会可以通过有意采取的预防性行动以及制度化措施克服发展带来的副作用。^⑱现代社会生活具有反身性。社会实践总是不断受到关于这些实践本身的新认识的检验和改造,从而在构成上不断改变着自己的特征。^⑲我们应努力提升认识和防控人工智能风险的能力,将其作用于社会并改变风险的样态。

从人工智能风险的特征和发生及控制规律的认知出发,我们“需要抵御风险的保障,但也需要具有面对风险并以一种积极的方式来对待风险的能力”^⑳。风险的两面性决定了治理目标的双重性。我们需要关注风险中蕴含的机遇,通过实施赋能型治理,在发展中提升社会认知和防控人工智能风险的能力,使得人工智能发展成果切实服务于人类福祉的提升。

(四) 人工智能治理能力的现存缺陷

当前社会在人工智能风险的认知和评估能力、社会监督能力、企业内部自治能力、政府外部监管能力等方面普遍不足,需要以“赋能”为核心,加强人工智能时代所需要的治理能力建设。

首先是对人工智能风险的认知和评估能力不足,社会监督能力欠缺。全面认识和把握人工智能风险是进行风险治理的重要基础,但由于人工智能的风险信息不充分、不及

^⑮ 参见[英]安东尼·吉登斯:《第三条道路——社会民主主义的复兴》,郑戈译,北京大学出版社2000年版,第65-66页。

^⑯ See Rebecca B. Naumann, et al., *Examining the Safety Benefits of Partial Vehicle Automation Technologies in an Uncertain Future*, Technical Report, AAA Foundation for Traffic Safety, 2023, p.39.

^⑰ 参见赵鹏:《风险社会的行政法回应》,中国政法大学出版社2018年版,第77页。

^⑱ 参见[德]乌尔里希·贝克、[德]约翰内斯·威尔姆斯:《自由与资本主义——与著名社会学家乌尔里希·贝克对话》,路国林译,浙江人民出版社2001年版,第121页。

^⑲ 参见前注④,安东尼·吉登斯书,第44页。

^⑳ 前注⑮,安东尼·吉登斯书,第67页。

时,企业、社会公众、政府主管部门等各类主体往往对人工智能新技术新应用的风险缺乏清晰认知,以致出现对人工智能风险产生掉以轻心或过度恐惧的两极分化。目前,存在对人工智能风险的认知、评估和监督的能力缺陷,主要有三方面原因:一是人工智能自身的复杂性。人工智能具有“黑箱性”、一定程度的“自主性”、运行结果的难预测性、运行机理的难解释性、可能根据环境变化的自适应性以及高速迭代更新发展等特性,这些都为风险认知、评估、监督和防控提出了挑战。二是人工智能的社会应用广泛。人工智能广泛应用在社会生活的方方面面,其风险识别和评估不仅需要人工智能科技专家参与,还需要应用领域的专家、法学家、伦理学家、社会学家等共同参与,以更全面地评估人工智能在整体上以及具体场景中所产生的影响。三是数字鸿沟问题加剧。当前社会公众对人工智能的理解和运用能力差别非常大,这既影响到人工智能科技的普惠应用,也使得欠缺理解的群体在人工智能风险面前更为脆弱。这三方面都决定了需要面向人工智能治理进行针对性的能力建设。

其次是企业对人工智能风险的自治能力不足。企业作为人工智能的主要研发者、部署者,具有在其研发部署过程中针对可能的风险采取自我治理的机会。然而大量企业特别是中小微企业的自治能力有限,缺少必要的技术措施和管理措施储备,往往难以将人工智能伦理规范和治理原则转化为有效的具体措施。当前,能够提升人工智能的安全性、准确性、稳健性、可解释性、公平性、包容性等的技术还需要大力创新发展,加上技术的研发和实施成本过高,企业也往往怠于自治。还有许多人工智能企业算力资源不足、高质量训练数据欠缺,这也制约了企业有效自治的能力。

最后是政府科学监管和促进发展的能力不足。一方面,科学有效的监管需要理论支撑、机制建设和技术赋能。关于人工智能发展情况及规律、人工智能风险以及有效的风险治理方法的信息不充分、不对称,加上资源不足、工具匮乏等问题,使得政府监管方式的可操作性、有效性和合理性都容易受到质疑和挑战。面对新兴科技的新型治理理念的实现,更加需要国家的基础设施与行政能力予以辅助,包括信息汇集与科学分析能力、风险判断能力、吸纳市场积极要素的能力等。^{②④}另一方面,从政府角度促进人工智能发展也亟待能力建设。算力、数据等要素以及测试、评估、认证等服务的提供都需要必要的制度支撑。政府推动人工智能在民生服务、社会治理、经济发展等领域的融合应用也需要建设相匹配的新型风险治理能力,以保障其可持续发展。

因此,针对人工智能治理的现存缺陷,必须构建赋能型治理,建设高水平治理能力,以此推进人工智能高质效地安全可信发展。

(五) 既有人工智能规制模式的局限

既有的规制模式和理论,为人工智能治理提供了有益参考。但面对新一代人工智能发展态势,其局限性逐渐显露出来。

既有规制模式主要包括命令—控制型规制、建议劝服型规制、回应型规制、元规制

^{②④} 参见卢超:《包容审慎监管的行政法理与中国实践》,载《中外法学》2024年第1期,第159页。

等,可从规制者、规制对象、命令类型和后果类型等方面考察其差异。^{②5}命令—控制型规制通常事先制定具体的特定命令,而违反的后果通常是比较强的制裁。但理想中这种模式下的命令应具有较高的规则精确性,在人工智能仍高速发展的当下难以实现。建议劝服型规制模式主张建立合作而非对抗的局面,^{②6}命令形态往往是一般性的目标,主要依赖企业自我规制,但在企业利益驱动面前往往效果大打折扣,在人工智能领域更是存在算法黑箱等诸多问题。^{②7}因此,这两种规制模式都难以取得理想效果,以致陷入“科林格里奇困境”^{②8}。许多研究者在这两极之间寻找更加综合性、动态性的规制模式。回应型规制旨在弥合强监管和放松监管之间的鸿沟,设想规制活动发生于对话式、交互性的环境中,规制者通常优先采用干预性较低的措施,但如果措施失灵,规制者就会逐步采取更具有惩罚性或强制性的措施。其最突出的特点即后果的动态回应性。^{②9}回应型规制模式可派生出很多具体形式,常与元规制相结合。^{③0}元规制即对企业自我规制的规制。政府通过提出一般性目标的命令来要求和塑造企业内部的自我规制,调动企业的主观能动性,让企业凭借自身掌握的信息和科技能力来制定适当的具体规范,^{③1}并由外部规制中的法律后果来提供行为激励。这种规制模式在数字科技领域得到了较多应用。但其成功条件在于,政府作为外部规制者需要建立获得风险情况、企业采取的规制措施的成本和实效、行业中相关技术发展水平等一系列关键信息的能力,并能够提供合理的惩戒和激励机制,才能在保障发展的同时有效地激励企业积极开展自我规制,同时企业也需要具有在合理成本内有效实施自我规制的信息、资源和能力,并要特殊考虑中小微企业的自我规制能力问题。例如,“信息对称性”原本是自我规制的重要优势,^{③2}但面对高速发展的人工智能,企业实际上也难以充分掌握人工智能新应用的风险情况和有效措施等信息。因此,这类规制模式在人工智能领域的适用对相关能力建设提出了更高要求。

②5 See Cary Coglianese, *Engaging Business in the Regulation of Nanotechnology*, in Christopher J. Bosso ed., *Governing Uncertainty: Environmental Regulation in the Age of Nanotechnology*, Resources for the Future Press, 2010, p.49-51; [英] 罗伯特·鲍德温、[英] 马丁·凯夫、[英] 马丁·洛奇:《牛津规制手册》,宋华琳等译,上海三联书店 2017 年版,第 165-166 页。

②6 See Bridget M. Hutter, *Regulating Employers and Employees: Health and Safety in the Workplace*, *Journal of Law and Society*, Vol.20(4), p.452-470 (1993); Keith Hawkins, *Environment and Enforcement: Regulation and the Social Definition of Pollution*, Oxford University Press, 1984, p.4.

②7 See Thomas Ferretti, *An Institutional Approach to AI Ethics: Justifying the Priority of Government Regulation over Self-Regulation*, *Moral Philosophy and Politics*, Vol.9(2), p.244-256 (2022).

②8 科林格里奇困境 (Collingridge's Dilemma) 是英国技术哲学家大卫·科林格里奇在《技术的社会控制》中提出的“控制困境”,即试图控制一项技术是非常困难的,因为在技术发展早期还可控时,对其有害的社会后果认知不足,难以证明控制其发展的必要性;但当有害社会后果已经很明显了,控制亦已昂贵而缓慢。See David Collingridge, *The Social Control of Technology*, Frances Pinter, 1980, p.19.

②9 See Ian Ayres & John Braithwaite, *Responsive Regulation: Transcending the Deregulation Debate*, Oxford University Press, 1992, p.35-37; Neil Gunningham, *Enforcement and Compliance Strategies*, in Robert Baldwin, Martin Cave & Martin Lodge eds., *The Oxford Handbook of Regulation*, Oxford University Press, 2010, p.125-131.

③0 See Ian Ayres & John Braithwaite, *supra* note 29, 38-39.

③1 参见前注 ②5, 罗伯特·鲍德温等书,第 168-170 页。

③2 参见黄文艺、孙喆玥:《论互联网平台治理的元规制进路》,载《法学评论》2024 年第 4 期,第 112 页。

上述各种规制模式形成了从最严格到最宽松的谱系,为人工智能规制提供了重要的基础,特别是在回应型规制和元规制模式的基础上,研究者们对寻找与人工智能治理相适配的规制模式进行了许多积极探索,^③但在理念上尚缺乏对人工智能安全可信发展所亟需的各项能力建设的突出强调,即缺少对于“赋能”这一维度的着重考虑。在人类社会迈向智能时代、但安全可信发展能力尚待建设的现阶段,有效的规制状态不可能一蹴而就,需要从对基础的规制模式讨论升级至对能力建设的研究,明确“赋能”这一目标,实现良好规制的切实落地。

综上所述,当前人工智能产业发展需要赋能,形成安全可信的技术和保障体系更加需要赋能。安全不只是一种状态,更是一种能力。我国《国家安全法》将国家安全定义为“国家政权、主权、统一和领土完整、人民福祉、经济社会可持续发展和国家其他重大利益相对处于没有危险和不受内外威胁的状态,以及保障持续安全状态的能力”。这种安全能力观在我国数字科技领域的立法中也得到突显。《网络安全法》将网络安全定义为“使网络处于稳定可靠运行的状态,以及保障网络数据的完整性、保密性、可用性的能力”。《数据安全法》将数据安全定义为“确保数据处于有效保护和合法利用的状态,以及具备保障持续安全状态的能力”。从这些定义中可见“能力”的重要性。同时,这三部法律也都明确了促发展的目标,这不仅是出于平衡考虑,更是因为真正的安全应该是也只能是建立在高度发展之上的安全。

在当下阶段,不发展是最大的不安全。人工智能治理的目标不是简单追求一个局部性、临时性的假性安全状态,而应该是追求达到一个拥有高度发达的科技水平和保障持续安全可信能力的状态,应以此为目标构建推动能力提升的有效治理机制。因此,赋能型人工智能治理的核心要义在于以赋能为核心,通过科学有效的治理,实现高度的人工智能发展水平和保障持续安全可信状态的能力,即针对人工智能安全可信发展中的重要能力建设需求,通过治理实现如下方面的赋能:第一,赋能企业创新发展,提高我国企业和人工智能行业整体竞争力,特别是增强发展要素的普惠赋能,包括对数据要素、算力要素、算法要素、制度要素的协同赋能。第二,赋能企业可信规制,特别是注重风险管理能力的普及发展,使得大中小型企业都能够开展有效的自我管理,从源头防控人工智能研发运用的伦理风险和安全风险。第三,赋能政府有效监管,特别是积极发展和推广用于人工智能治理的相关技术,实现人工智能治理技术的创新突破。第四,赋能社会公众平等受益于人工智能的发展,使社会公众有效地参与人工智能治理,提升社会公众对人工智能发展的信任度,营造有利于人工智能可信发展的社会环境。第五,赋能我国深度参与全球人工智能治理体系建设,为推进全球人工智能治理体系现代化提供中国经验、中国智慧和力量。

^③ 参见张凌寒:《平台“穿透式监管”的理据及限度》,载《法律科学》2022年第1期,第107页;张欣:《生成式人工智能的算法治理挑战与治理型监管》,载《现代法学》2023年第3期,第117页;郭小东:《生成式人工智能的风险及其包容性法律治理》,载《北京理工大学学报(社会科学版)》2023年第6期,第93页。

三、赋能型人工智能治理的理念确立

赋能型人工智能治理建设需要确立与之相适应的理念：既要确立核心理念，即明确如何理解赋能型人工智能治理中涉及的重要关系，从而为“赋能”确立正确的方向和界限；也要从核心理念中发展出基本理念，更为具体地指引赋能型人工智能治理机制的构建。

（一）赋能型人工智能治理的核心理念

赋能型人工智能治理的核心理念基于两大关系而形成，即人与物（智能体）的关系、发展和安全的关系。

1. 坚持以人为本，促进人机和谐

将以人为本作为核心理念，明确了赋能人工智能安全可信发展的最终目标是赋能人民，为赋能型治理确立了方向和评价标准。以人为本是可信人工智能的基础，是人工智能治理的首要理念与终极关怀，体现为从人类福祉、人类安全、人的尊严和权利、人的全面发展出发，充分尊重和保障人权，确保人工智能解决方案是以人为主体、人工智能的使用以人为中心。发展人工智能归根结底是为全体人民提供优质的公共服务，确保人工智能技术成果平等普惠共享。

以人为本是全人类的共同价值。世界上人工智能发展和治理走在前列的主要国家的法律法规、政策文件、纲领宣言、伦理指南等几乎都将以人为本作为其核心理念和根本原则。在人工智能治理上，联合国和相关全球性峰会也都主张以人为本，如2017年发布的“阿西洛马人工智能原则”^③、2021年联合国教科文组织（UNESCO）第41届大会通过的《人工智能伦理问题建议书》、2023年中美等28个国家在英国签署的《布莱切利人工智能宣言》、2024年第78届联合国大会协商一致通过的中国主提的加强人工智能能力建设国际合作决议，等等。

我国在人工智能发展与治理上，更是始终坚持以人为本的核心理念。《民法典》《个人信息保护法》等法律法规，都充分体现以人为本的精神，注重对智能时代公民个人信息权益、隐私权、肖像权、声音权、免受自动化决策权等的保护。2023年10月18日，我国发布《全球人工智能治理倡议》，第1条就明确主张：“发展人工智能应坚持‘以人为本’理念，以增进人类共同福祉为目标，以保障社会安全、尊重人类权益为前提，确保人工智能始终朝着有利于人类文明进步的方向发展。”

以人为本不是空洞的理念标签，而是应当落地为人工智能治理的法治原则和法理话语，指引着赋能的核心方向。以人为本首先体现为以人的权利为本，尊重和保障人权；其次体现为公平正义，重视算法公平，注重数字无障碍设计，缩小数字鸿沟，实现普惠发展；最后体现为确保人类在人工智能运行中的主体性、自主性，实现“人在回路”

^③ *Asilomar AI Principles*, at <https://futureoflife.org/open-letter/ai-principles/> (Last visited on March 15, 2024).

(human-in-the-loop) 控制、发展人机协同控制等方法,^⑤ 确保人类能够始终实质性监督和控制人工智能。

2. 坚持发展导向, 统筹发展和安全

发展是人类社会的永恒主题。坚持发展导向, 统筹发展和安全, 是赋能型人工智能治理的另一项核心理念。

世界各国在人工智能治理中都须考虑如何平衡好发展与安全。欧洲议会在投票通过《人工智能法案》前的最后时刻仍在为如何平衡促进创新与防范风险这一问题辩论, 凸显出如何处理好发展和安全的关系是人工智能立法的基本问题和工作难题。

在新时代中国式现代化进程中, 发展和安全的关系演化为高质量发展和高水平安全的关系, 要以高质量发展促进高水平安全, 以高水平安全保障高质量发展。《决定》指出, “加快构建新发展格局, 推动高质量发展” “建设更高水平平安中国, 健全国家安全体系” “实现高质量发展和高水平安全良性互动”。^⑥ 这为人工智能治理确定了根本遵循。要把这一核心理念贯穿到人工智能治理的各环节全过程, 将安全与发展协同部署, 构建有利于人工智能安全可信发展的生态系统。

正确认识和统筹人工智能发展和安全的关系, 要从我国人工智能发展水平和治理能力出发, 对人工智能战略问题开展前瞻性、针对性、储备性研究。发展是安全的基础。历史和现实都证明, “发展是解决我国一切问题的基础和关键”^⑦。我国人工智能技术发展仍处于追赶阶段, 在人工智能科技已经成为国家的核心竞争力乃至大国博弈实力的今天, 人工智能不发展是最大的不安全, 发展不充分是最大的风险隐患。保障人工智能安全可信的能力、保障人工智能时代国家和人民的重大利益处于持续安全状态的能力, 只可能是建立在高发展水平基础上的。人工智能治理应为其创新发展留有充足的空间和时间。相关政策、法律制度应遵循和尊重新一代人工智能研发应用的规律和现实, 以有利于人工智能发展和更好发挥作用为根本出发点, 着力赋予权利、减轻义务、科学监管、提供服务, 防止监管发力过猛对创新发展造成实质性损害。要把握未来发展主动权, 增强我国人工智能的竞争力、发展力、持续力、安全力以及引领力。

安全是发展的条件和保障。没有安全机制和安全措施, 人工智能的可信发展也无从谈起。人工智能治理的重要任务: 一是划清安全底线; 二是在发展中坚持总体国家安全观, 推进人工智能全过程的安全保障体系和能力建设; 三是坚持以共同、综合、合作、可持续的全球安全观为战略武器, 反制一些国家对我国人工智能发展的遏压, 打造安全屏障和安全环境。

坚持发展导向、统筹发展和安全的关系, 要坚持良法善治, 在法治轨道上促进发展和

^⑤ 参见许为、葛列众、高在峰:《人—AI 交互: 实现“以人为中心 AI”理念的跨学科新领域》, 载《智能系统学报》2021 年第 4 期, 第 613 页。

^⑥ 参见前注①,《中共中央关于进一步全面深化改革 推进中国式现代化的决定》。

^⑦ 习近平:《决胜全面建成小康社会 夺取新时代中国特色社会主义伟大胜利——在中国共产党第十九次全国代表大会上的报告》(2017 年 10 月 18 日), 人民出版社 2017 年版, 第 21 页。

安全良性互动,切实做到人工智能产业发展与安全治理之间的稳定平衡,使之相辅相成,以安全可信的新一代人工智能技术推动产业升级和新质生产力发展。

(二) 赋能型人工智能治理的基本理念

基于上述核心理念和国内外人工智能治理经验,赋能型人工智能治理理念可进一步展开为如下基本理念:

1. 智能向善

智能向善既是以人为本理念的要求,也为统筹发展和安全指出具体路径,指引着对安全可信发展能力的不断建设。智能向善是科技向善、数字向善原则在人工智能科技领域内的具化。其核心是,推动从事人工智能研发、提供和使用活动的个人、企业、行业等主体遵守社会公序良俗、社会主义核心价值观和全人类共同价值,向善发展和利用人工智能,满足人民群众美好生活对智能科技的需要,不断增进人类共同福祉;同时推动政府部门和社会组织对人工智能向善而治,防范恶意开发和应用人工智能技术,消除数字鸿沟,促进数字正义和社会公正,推动人类文明进步。我国发布的《全球人工智能治理倡议》特别提出“发展人工智能应坚持‘智能向善’”,其宗旨就在于此。

面向未来,确保人工智能技术开发应用遵循智能向善的理念,是人工智能治理的关键所在。首先,需要在以人为本、发展导向的核心理念引领下,不断完善对人工智能应用领域价值目标(“善”)的认识,并将“善”的目标加入人工智能研发的目标体系,在技术研发全流程中加以实现。其次,需要依托具体机制措施,推动人工智能科技向善发展,特别是推动能够支撑人工智能价值对齐、安全可信的技术措施和管理措施的创新与应用。各类科技的创新发展不是均衡的、步调一致的,要针对企业的“研究偏向”问题,^③加强伦理和法治引导,保障企业在“智能向善”的轨道上运行。要在保障企业创新发展能力的同时,赋能有益于人工智能与社会伦理价值对齐的科技研发和管理制度创新。

2. 包容审慎

包容审慎是坚持发展导向、统筹发展和安全理念的具化,有助于赋能企业创新发展,是我国在科技领域一以贯之且行之有效的先进理念,并体现在法律、行政法规和部门规章之中。《科学技术进步法》第35条规定:“国家鼓励新技术应用,按照包容审慎原则,推动开展新技术、新产品、新服务、新模式应用试验,为新技术、新产品应用创造条件。”2023年,国家网信办等7部门联合发布的《生成式人工智能服务管理暂行办法》第3条规定,对生成式人工智能服务实行包容审慎和分类分级监管。包容审慎与科技赋能是辩证统一的。2024年3月,李强总理在北京调研时指出,人工智能是发展新质生产力的重要引擎,要在守住安全底线的前提下,积极推行包容审慎监管,给予新技术足够的创新空间和必要的试错空间。^④

^③ 即在没有约束或引导的情况下,企业的研发力量往往大量投入于与企业未来盈利具有明显关系的领域,对弱势群体权益保障、数字无障碍技术、安保技术等往往投入不足。

^④ 参见《李强在北京调研时强调 推进科技创新和产业创新深度融合 加快塑造高质量发展新动能新优势》,载《人民日报》2024年3月14日,第1版。

包容审慎作为人工智能治理的一项基本理念,其要义在于,对人工智能新技术新产品新业态采取适度宽容的态度,允许其在守住安全底线的前提下自行纠正研发应用过程中出现的问题,政府在审慎跟踪观察的同时,只进行适时适度的干预。换言之,对人工智能新技术包括颠覆性技术的出现应当秉持积极拥抱、包容审慎的治理理念和规则策略,对其引发的新情况新问题,不必急于管制或惩罚,防止因监管失当而将其扼杀在萌芽状态。从传统规制理论角度观察,包容审慎监管映射出回应性规制的执法策略、合作规制的运行范式以及规制试验主义的演进逻辑。^{④①}坚持包容审慎理念,有利于促进发展与安全、公平和效率、自律和他律的动态平衡,最大限度地鼓励和支持创新,赋能科技产业发展。

3. 敏捷治理

敏捷治理是面向动态发展的产业,统筹发展和安全的一条路径,是建设动态机制提升治理能力的一种理念。敏捷的理念一定程度上借鉴自20世纪90年代软件工程领域的“敏捷开发”^{④②}。2018年,世界经济论坛对第四次工业革命中的政策制定问题进行反思,正式提出了“敏捷治理”(Agile Governance),将其界定为具有适应性、以人为本、包容性和可持续性的政策制定,并且要引导越来越多的利益攸关方积极参与。与会代表认为这是快速驾驭变化、主动或被动地拥抱变化并从变化中学习的持续准备,同时为实际或设想的最终用户价值作出贡献。^{④③}2022年,中共中央办公厅、国务院办公厅印发《关于加强科技伦理治理的意见》,将敏捷治理列为五大治理要求之一,强调“加强科技伦理风险预警与跟踪研判,及时动态调整治理方式和伦理规范,快速、灵活应对科技创新带来的伦理挑战”。《全球人工智能治理倡议》等文件中也都强调了敏捷治理。此前,2019年,中国国家新一代人工智能治理专业委员会发布《新一代人工智能治理原则——发展负责任的人工智能》,将敏捷治理列为八项原则之一,并进行了具体阐释;在其2021年发布的《新一代人工智能伦理规范》中,也在管理规范部分规定了“推动敏捷治理”。可见,敏捷治理已经成为我国人工智能治理中的一项基本理念和原则。其核心意义在于,强调尊重人工智能发展规律并保持跟踪研判,强调治理节奏上的快速回应和尽早介入,治理规则上推进弹性原则与具体类型化规则有效结合,治理关系上的互动合作以及治理方式上的过程快、力度轻的引导性治理。^{④④}

敏捷治理是针对发展迅速、影响广泛的新领域提出的一种治理范式。人工智能的科技发展和产业应用高速迭代,要求在研发和应用过程中,根据发展变化或新的信息来及

^{④①} 参见前注^{②④},卢超文,第143页。

^{④②} Torgeir Dingsøyr et al., *A Decade of Agile Methodologies: Towards Explaining Agile Software Development*, Journal of Systems and Software, Vol.85:1213, p.1213-1221 (2012).

^{④③} See World Economic Forum, *Agile Governance: Reimagining Policy-making in the Fourth Industrial Revolution*, White Paper, Jan. 2018, p.4, at https://www3.weforum.org/docs/WEF_Agile_Governance_Reimagining_Policy-making_4IR_report.pdf (Last visited on March 15, 2024).

^{④④} 参见薛澜、赵静:《走向敏捷治理:新兴产业发展与监管模式探究》,载《中国行政管理》2019年第8期,第31-33页;赵静、薛澜、吴冠生:《敏捷思维引领城市治理转型:对多城市治理实践的分析》,载《中国行政管理》2021年第8期,第52-53页。

时调整治理策略和措施,以保障人工智能的安全、可靠和可控。敏捷治理特别强调前瞻性的视野和方法,尤其是努力尝试在问题出现之前对问题的预测研判。^{④④}在这个意义上,敏捷治理也是一种“预防性法治”。^{④⑤}因此,敏捷治理要求建立稳定良好的信息获取能力和机制,保障面对不确定性问题时能够及时获取广泛、多样、充分的意见,^{④⑥}要求建立健全风险沟通、风险报告和预警机制,并持续加强对风险的研究和预判能力。

4. 可持续发展

以人为本和发展导向的核心理念也决定了应将可持续发展作为一项基本理念。可持续发展强调在推进技术发展的同时,确保这种发展促进经济、社会 and 环境的长期健康和平衡,包括环境友好(绿色发展)、节约资源、减少不平等和数字鸿沟、促进教育和就业、提高公共服务的质量和可及性、尊重和保护文化多样性、鼓励技术开放和共享、支持全球合作等。人工智能科技自身也需要注重可持续发展,理性对待创新和投资,对过往的两次“人工智能严冬”应引以为戒。人工智能治理的可持续发展理念正在成为全球共识。联合国教科文组织发布的《人工智能伦理问题建议书》建议各国政府和机构应充分考虑到人工智能技术对于联合国可持续发展目标的影响。^{④⑦}我国发布的《全球人工智能治理倡议》也提出,积极支持以人工智能助力可持续发展,应对气候变化、生物多样性保护等全球性挑战。这些都应成为人工智能治理应坚持的基本理念、价值标准和行动准则,从而赋能社会公众平等受益于人工智能的发展,并赋能我国深度参与全球人工智能治理体系建设。

四、赋能型人工智能治理的机制构建

在人工智能治理机制设计中,应自觉将“赋能”作为重要目标,将前述核心和基本理念融入人工智能治理相关的各方面机制设计之中。赋能型人工智能治理的具体机制仍需不断研究发展。当前在整体上应建设以法治为核心的人工智能治理机制,并聚焦关键问题,建设法治统领下的各项具体机制。

(一) 构建以法治为核心的赋能型人工智能治理机制

之所以要构建以法治为核心的人工智能治理机制和体系,从原理而言,是因为“法治是治国理政的基本方式”^{④⑧}，“是现代社会治理的基本手段”^{④⑨}，“依法治理是最可靠、最

^{④④} See Helmut Anheier & Edward Knudsen, *Agile Governance*, at <https://intelligence.weforum.org/topics/a1Gb0000000pTDaEAM> (Last visited on March 15, 2024).

^{④⑤} 参见黄文艺:《论预防型法治》,载《法学研究》2024年第2期,第20-38页。

^{④⑥} See *Agile Governance: Managing Uncertainty*, at <https://intelligence.weforum.org/topics/a1Gb0000000pTDaEAM/key-issues/a1G680000004CtBEAU> (Last visited on March 15, 2024).

^{④⑦} 参见联合国教科文组织:《人工智能伦理问题建议书》,载联合国教科文组织官网, https://unesdoc.unesco.org/ark:/48223/pf0000380455_chi, 2024年3月15日访问。

^{④⑧} 习近平:《关于〈中共中央关于全面推进依法治国若干重大问题的决定〉的说明》(2014年10月20日),载习近平:《论坚持全面依法治国》,中央文献出版社2020年版,第84页。

^{④⑨} 习近平:《坚持法治精神,实现公平正义》(2017年9月26日),载前注^{④⑧},习近平书,第183页。

稳定的治理”^{⑤0}，“是中国式现代化的重要保障”^{⑤1}，是贯彻以人为本和发展导向的核心理念、赋能人工智能安全可信发展的关键支撑。

第一，从“软法”治理到以“硬法”为引导和保障、“软法”与“硬法”相结合的治理形态，是人工智能治理的必然趋势。过去人工智能大国的人工智能治理主要依靠“软法”，即依人工智能科技伦理、行业规范和技术标准等，但是，随着人工智能发展特别是人工智能大模型这类颠覆性技术的安全风险的外溢，“软法”的局限性和低效能日益显现，并且难以给予企业明确预期。于是，人工智能治理正在由“软法”治理转向以“硬法”为引导和保障、“软法”与“硬法”协同治理的新形态。这标志着人工智能治理在法治轨道上有序推进。

以“硬法”为引导和保障不代表忽视“软法”的优点和作用，而是在充分发挥“软法”的积极作用的同时，关注“硬法”可以对“软法”起到的支撑作用。在人工智能治理体系中，科技伦理、道德规范、行业规范、技术标准、企业规章、国际宣言等“软法”有其不可替代的积极作用，特别是在具体应用场景的科技伦理、行业规范和技术标准独具指导和规范作用。“硬法”对“软法”的引领和制度性支撑可以使其更好发挥作用，以形成“内外结合、法德共治”的局面。

当前，国际上已开始重视人工智能治理的法治化和规范化。例如，欧盟陆续通过了数字领域的一系列重要法律，并在2019年发布的《可信人工智能伦理指南》的基础上，制定和通过了《人工智能法》。这标志着欧盟在人工智能治理中从“软法”到“硬法”的转变升级。美国则出台了系列行政命令，如《关于安全、可靠和值得信赖地开发和使用权使用人工智能的行政命令》等，对行政部门等作出了明确工作部署和要求。^{⑤2}

我国人工智能治理的“硬法”保障还滞后于人工智能发展和安全保障的需求。尽管我国已经制定实施了《网络安全法》《数据安全法》《个人信息保护法》等与人工智能相关的法律及行政法规、地方性法规，出台了《生成式人工智能服务管理暂行办法》等多部相关部门规章，但还缺少专门的人工智能法律和行政法规，尤其是缺少一部人工智能基本法。2017年国务院印发的《新一代人工智能发展规划》提出，“到2025年初步建立人工智能法律法规，到2030年建成更加完善的人工智能法律法规”。为推进这项立法规划的实施，应当在坚持科学立法、民主立法、依法立法原则和确保立法质量的基础上，加快立法步伐，把“小快灵”和“大部头”立法结合起来，经过五年左右时间形成人工智能法律法规体系，打造有利于人工智能安全可信发展的法治环境。

第二，法治是推进人工智能领域善治的必由之路。人工智能治理需要落实以人为本、发展导向的核心理念，使人工智能领域既规范有序又充满活力。达到此目标，就是我

^{⑤0} 习近平：《依法保障“一国两制”实践》（2014年11月—2019年12月），载前注^{④8}，习近平书，第120-121页。

^{⑤1} 前注^①，《中共中央关于进一步全面深化改革 推进中国式现代化的决定》。

^{⑤2} See Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, at <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/> (Last visited on March 15, 2024).

们期待的“善治”。法治正是通向善治的必由之路。法治能够将必要的规范和有效的治理机制予以确立,以法律的确定性消解人工智能的许多不确定性,使人工智能研发、提供、部署主体可以对自身活动及其结果产生稳定预期,保障和促进企业依法依规经营、人工智能科技向善发展、公民个人依法使用、社会治理能力提升,进而消减不确定因素及关联问题的发生,增强社会对于发展和应用人工智能的信心。

第三,法治是营造人工智能创新发展环境的必然要求。习近平总书记指出,“法治是最好的营商环境”^{⑤③}。根据这一科学判断,可以认为法治是人工智能最好的发展环境。其一,法治以其“权利本位”的价值导向,为人工智能领域的产权等权益保护建构法治体系,激发社会创新活力,提升自主创新能力。其二,维护公平竞争的法治秩序可以“激发市场主体发展活力,使一切有利于社会生产力发展的力量源泉充分涌流”^{⑤④}。其三,法治实施分级分类监管,对于少量的高风险人工智能采取审慎的监管,对大量风险较小的人工智能应用提供尽可能宽阔的创新空间。其四,法治为人工智能发展创造了安全的政治环境、稳定的社会环境、公正的法治环境、优质的服务环境。

(二) 完善法律治理与技术治理相统合的机制

科技与法治的有机融合是人工智能治理的必由之路,是人工智能时代治理能力建设的重点。自互联网和数字技术诞生以来,以莱斯格为代表的研究者们就开始探讨数字科技与法律的关系,推进数字科技与法律的良性互动。^{⑤⑤}推进“符合伦理的设计”(ethically aligned design)、“通过设计保障伦理”(ethics by design)、“通过设计保护隐私”(privacy by design)等被认为是治理数字科技的重点路径。布朗斯沃德提出了“法律3.0”的概念,强调利用技术性方案来实现政策目标,即监管机构应当自己采用(或让他人采用)技术管理措施,将规范性观点转化为实际的设计。^{⑤⑥}在立法和法律适用中应当充分考察科技发展情况,促进智能科技向善发展,推动关键要素和治理技术的普惠可及,并用科技赋能企业自治和政府监管,使整个社会更加高效、可控地迎接人工智能时代到来。针对当前的人工智能发展和治理现状,完善法律治理与技术治理相统合的机制,有两方面重要工作:

首先,从坚持发展导向、统筹安全发展的核心理念出发,需要在法治建设中深度研判科技发展的制度需求,赋能企业创新发展。技术治理能力的提升本身需要以人工智能科技发展为基石。这要求法治建设切实把握人工智能科技的发展要素、发展阶段和发展需要,探求其中的关键性生产要素及其是否存在供给不充分、市场失灵等情况,对相关法律进行系统性解释或完善,合理促进关键性要素的提供和利用效率。目前对于人工智能企业而言,科技要素赋能主要是算法、算力、数据的赋能。这些要素的供给和流通利用的不充分、不公平,可能影响人工智能创新特别是中小企业的公平竞争,数据的可及性和质量

⑤③ 习近平:《为做好党和国家各项工作营造良好法治环境》(2019年2月25日),载前注④⑧,习近平书,第254页。

⑤④ 习近平:《在经济社会领域专家座谈会上的讲话》(2020年8月24日),人民出版社2020年版,第7页。

⑤⑤ See Lawrence Lessig, *The Law of the Horse: What Cyberlaw might Teach*, Harvard Law Review, Vol.113(2), p.514 (1999); 齐延平:《数智化社会的法律调控》,载《中国法学》2022年第1期,第79页。

⑤⑥ 参见[英]罗杰·布朗斯沃德:《法律3.0:规则、规制和技术》,毛海栋译,北京大学出版社2023年版,第1-5页。

问题还会影响人工智能系统的公平性、包容性、准确性、安全性等。需要建立科学有效的法律制度,促进要素质量提升、普惠可及,赋能人工智能科技和产业发展。

其次,从智能向善的理念出发,需要通过法治来促进企业不断提升人工智能技术的安全可信性,推动合规技术和监管技术的创新,赋能企业自身治理和政府监管。一方面,应当将研发、采取必要技术措施与企业注意义务相衔接,激励相应技术的发展;另一方面,还应注重不断分析梳理治理科技和管理措施相关信息,重视规则的“精确度”,^{⑤7}将抽象、弹性的法律规则与明确、具体的标准、行动指南等相结合。这些指南和标准建设也要重视法律、行政规制和技术方案之间的融贯性,^{⑤8}避免技术标准对于法律的误读。

(三) 发展多元主体沟通协作的共治机制

有效、及时的多元主体沟通协作是实现包容审慎、敏捷治理理念,提升风险识别和防控能力的必然要求。风险信息和治理信息的有效交流和共享是社会治理的重要机制。人工智能的风险治理需要建立更加及时有效的沟通机制。我国近年来在新闻推荐、平台劳动算法等领域,都出现了基于社会监督以及监管机构的沟通,推动了智能算法改进的实例。未来应将人工智能风险信息和治理信息沟通进一步前置化、制度化、常态化。

首先,针对目前跨领域主体缺少信息沟通、治理对话困难等问题,需要建立由监管机构主导,人工智能企业、科研人员、行业组织、新闻媒体、社会公众等多元主体参与的制度化、常态化的风险沟通机制。参与主体的专业领域应当涵盖人工智能科技、伦理学、法学、经济学、管理学、社会学、新闻传播学、教育学、心理学等,持续建设和发布有益信息,提高透明度,完善用户及社会公众的便捷反馈渠道和处理流程,推动建设及时、便捷的风险报告和应急处理机制,建立健全立法后的法治效果评估机制。

其次,应当创新拓展试验性监管机制。建立监管沙箱是一种典型方式。在可能具有高风险的人工智能新技术新应用进入市场之前,可建立影响可控的监管沙箱,由相应的监管机构入场观察、沟通和指导,帮助创新者更好地理解合规要求、建立有效控制风险的机制,并将产品安全推向市场。我国市场监管总局、工信部等五部门联合发布的《关于试行汽车安全沙盒监管制度的通告》(2022年第6号)指出,汽车安全监管具有提高应急处置能力、防范和化解重大风险、保护消费者合法权益、鼓励企业技术创新、倡导最佳安全设计实践等重要意义。在试验性监管中,可以充分开展监管方式的创新和实验,综合运用适当的监管科技,开展人工智能社会治理实验,对风险情况进行深入评估,并探索出科学有效的风险防治措施。

最后,应当加快构建人工智能治理和能力建设的国际合作机制,为人工智能安全可信发展提供国际协作和规则保障。一方面在国内立法中明确树立我国的人工智能治理理念,明确我国开展人工智能能力建设的对外援助,增强发展中国家的自主可持续发展能力,推动国际发展合作,坚持尊重他国主权;另一方面,积极推动形成具有广泛共识的

^{⑤7} See Colin S. Diver, *The Optimal Precision of Administrative Rules*, Yale Law Journal, Vol.93, p.66 (1983).

^{⑤8} 参见郭春镇:《生成式AI的融贯性法律治理——以生成式预训练模型(GPT)为例》,载《现代法学》2023年第3期,第88页。

人工智能国际治理框架,建立人工智能重要风险信息的国际流通共享机制和跨国风险防控的协作机制。

(四) 建立与人工智能发展相适配的避风港机制

在互联网发展过程中,避风港规则在保障互联网创新发展、促进社会合作治理等方面发挥了重要作用。在人工智能发展应用初期,赋能企业创新发展,要求法律尊重科技发展规律,明确建立与发展阶段、具体情况相匹配的“避风港”机制和具体规则。

从数据层面看,数据利用对于人工智能产业发展极其重要。有学者指出,我国数据流通交易所需要的主要制度增量在于建立合理的数据交易相关行为法律后果的责任规则,例如,确立在数据交易所进行的满足条件的场内数据交易的避风港规则等。^{⑤⑨}

从算法模型及社会应用层面看,当前主流的基于机器学习的人工智能技术建立在概率的基础之上,存在一定的偏差、错误在所难免。但人工智能的错误在不同应用领域可能引起的风险的类型、程度不同,且人工智能及风险防治措施都仍在动态发展之中。因此,为人工智能产业建立分级分类的、合理的避风港规则,对于稳定产业预期、促进创新投入、推动企业采取合理措施都具有积极意义。而背离人工智能发展客观规律、实行过度严格的规制将会形成对创新的抑制作用,也会影响到社会中的人工智能安全能力建设。当然,过于宽松的责任豁免规则,也可能难以推动人工智能企业自身的“智能向善”实践。因此在建立避风港规则时要结合多元主体的沟通协作机制,建立与技术发展水平相匹配的注意义务,做到宽严适度。

(五) 创建敏捷互动、激励向善发展的动态监管机制

人工智能处于高速发展阶段,为了建设有效的监管能力,可在包容审慎、敏捷治理的理念下,创建分级分类、敏捷互动、激励人工智能向善发展的动态监管机制。艾尔斯和布雷思韦特在回应型规制中提出“执法金字塔”模型,将监管工具划分为干预力度由弱至强的不同层次,并建议监管机构优先采取轻力度的干预模式,根据被监管对象的具体情况进行调整。^{⑥⑩} 人工智能治理可以扩展这一模型,构建风险管理金字塔、监管惩戒金字塔和价值对齐激励金字塔,并基于实践中的动态信息,在风险管理要求、惩戒措施和价值对齐激励措施的不同层级间进行动态调整。

其一,建立科学分级分类、支持动态调整的风险管理金字塔。在人工智能治理实践中,金字塔模型的思路首先可以转化为分级分类的风险管理要求,即要求采取的人工智能风险管理措施应当与风险级别和类别相适配,以确保对高风险人工智能系统采取有针对性、实效性的管理措施,并为大量风险微小的人工智能应用减轻负担。

风险管理金字塔模型已经投映在许多国家和地区的法律法规、公共政策之中。欧盟《人工智能法》将人工智能系统风险划分为不可接受的风险、高风险、有限的风险和极小的风险四个等级,并将通用人工智能模型划分为一般模型和具有系统性风险的模型两个

^{⑤⑨} 参见戴昕:《作为法律技术的安全港规则:原理与前景》,载《法学家》2023年第2期,第43-45页。

^{⑥⑩} See Ian Ayres & John Braithwaite, *supra* note 29, 38-39.

等级,分级分类地规定了不同要求。我国相关法律法规和部门规章也体现了分级分类进行风险管理要求的立法思路。^{⑥1}例如,《互联网信息服务算法推荐管理规定》第23条明确规定对算法推荐服务提供者实施分级分类管理,并延续到对深度合成、生成式人工智能的管理模式。2021年1月,全国信息安全标准化技术委员会发布《网络安全标准实践指南——人工智能伦理安全风险防范指引》,对人工智能伦理安全风险进行了分类,包括“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”,虽然该指引没有对风险等级进行划分,但从其分类术语中可以大体理解其风险级别。我国学者在人工智能立法建议中也提出了将人工智能系统基于其风险程度或关键程度,建立相对应的管理规范 and 监管措施。^{⑥2}

在建立分级分类管理基本框架的基础上,特别要注意人工智能仍处于高速发展阶段,关于风险和有效治理措施的信息都不充分,应当结合前述多元主体共治机制及时获取信息,并设置依法基于风险评估情况进行风险级别敏捷调整的机制,同时为企业提供风险管理措施指引,并留有合理的建设时间。

其二,建立渐进适用的监管惩戒金字塔。我国《网络安全法》《数据安全法》《个人信息保护法》等法律中都建立了从责令整改、给予警告,到没收违法所得、处以罚款、责令暂停相关业务、停业整顿、吊销相关业务许可证或者吊销营业执照,以及构成犯罪的,依法追究刑事责任等不同强度的规制措施。^{⑥3}延续此模式,人工智能领域可以建立惩戒强度逐层递进的监管惩戒金字塔,根据具体风险项的情况,结合渐进适用策略,采取合乎比例的、轻量级干预优先的惩戒措施。

其三,建立有效互动的价值对齐激励金字塔。人工智能治理体系中还应构建与企业推动智能向善实践相关联的“价值对齐激励金字塔”,以法治机制激励企业积极开展与社会价值对齐的人工智能技术研发,不断提升人工智能的安全性、准确性、稳健性、可解释性、公平性、包容性、隐私保护等重要指标,进而提升企业和整个社会在人工智能时代的安全可信保障能力。对于积极创新发展与社会价值对齐的人工智能技术的企业,可以采取给予行政奖励、进行信用联合激励、给予信用优惠或税收优惠等激励制度。^{⑥4}以此弥补传统监管机制中对于支撑人工智能价值对齐、安全可信的技术研发投入激励不足的情况,切实促进智能向善。

⑥1 我国《网络安全法》《数据安全法》《个人信息保护法》等法律中都体现了风险管理金字塔模型的基本思路,包括实施网络安全等级保护制度,对重要数据、敏感个人数据加强风险管理要求等。

⑥2 例如,《人工智能法示范法2.0(专家建议稿)》实际上将人工智能风险划分为高风险、中风险、低风险三个级别,规定了不同风险管理要求和监管机制。参见周辉等:《人工智能法示范法2.0(专家建议稿)》,载上海市人工智能社会治理协同创新中心官网, <https://aisg.tongji.edu.cn/info/1005/1211.htm>, 2024年9月14日访问。《中华人民共和国人工智能法(学者建议稿)》建议区分一般的人工智能系统和关键人工智能系统,对后者规定了更多要求。参见张凌寒等:《中华人民共和国人工智能法(学者建议稿)》,载北京航空航天大学法学院官网, <https://fxy.buaa.edu.cn/info/1143/8452.htm>, 2024年9月14日访问。

⑥3 参见《网络安全法》第59-69条;《数据安全法》第45条;《个人信息保护法》第66条;等等。

⑥4 参见喻文光:《论数字平台的合规监管》,载《法学家》2024年第1期,第126-127页。

（六）建设人工智能保险等社会保障机制

社会保障能力和机制建设是赋能型人工智能治理的重要方面。保险制度是现代风险治理中的一种重要工具,不但可以在灾害之后进行补偿,还可以发挥事先降低与管控风险的作用。首先,保险具有敏捷调整性。特别是面对新型技术,相比侵权责任体系,保险能够更快速地收集数据和进行评估,并且能够更加敏捷灵活地根据新技术及其安全措施的发展情况来持续进行更新和调整。^{⑥5}其次,在新技术发展早期,法律通常需要为发展留有必要的空间,如为人工智能明确设置避风港规则,此时保险可以填补侵权法留下的空白,为利益受损者提供金钱救济。^{⑥6}最后,保险可以通过设置保险范围限制和例外、进行保险核保并提供信息、调整保费等一系列措施来发出信号,激励和引导相关主体调整其行为,例如,采取更有效的风险管理措施,从而减轻与人工智能新兴技术相关的风险,同时为人工智能公司和人工智能用户提供更多的安全保障。这将允许不同的利益相关者继续释放人工智能的力量及其对社会的价值。^{⑥7}

从国际上来看,科技界和法学界对为机器人、自动驾驶等高风险人工智能产品建立保险制度的问题讨论了多年。我国一些地方出台的有关自动驾驶的地方性法规或规章中也明确规定了强制性的购买保险要求。^{⑥8}同时,我国也在积极探索网络安全保险制度,^{⑥9}这既可以为人工智能系统的网络和数据安全风险的保障提供保险机制,也可为更专门化的人工智能保险提供参考。

可以看到,人工智能系统的保险制度在实践中已经得到了一定程度的重视,但具体保险制度建构方案不尽相同,并且存在道德风险、缺少全球性指南、购买主体存在争议、保费估计困难等很多问题。^{⑦0}为实现有效赋能可信发展的目标,应进一步研究探索合适的人工智能保险制度构建方案。可以率先在自动驾驶等高风险人工智能领域开展试点,通过保险制度稳步推动人工智能安全评估、监测、应急处理、救济等能力建设,增强人工智能社会治理能力。

同时,需要推进与人工智能相关的通识义务教育、职业技能培训、就业促进机制等民生保障建设,消除数字鸿沟,赋能社会公众在人工智能发展中平等获益,也为人工智能科技的社会监督、科学运用奠定良好的基础。

^{⑥5} See Kenneth S. Abraham, *The Liability Century — Insurance and Tort Law from the Progressive Era to 9/11*, Harvard University Press, 2008, p.1256-1266.

^{⑥6} See Jeffery L. Vagle, *Cybersecurity and Moral Hazard*, *Stanford Technology Law Review*, Vol.23(1), p.100 (2020).

^{⑥7} See Anat Lior, *Insuring AI: The Role of Insurance in Artificial Intelligence Regulation*, *Harvard Journal of Law & Technology*, Vol.35(2), p.511-519 (2022).

^{⑥8} 例如《北京市自动驾驶车辆道路测试管理实施细则(试行)》第8条第(五)项、《上海市浦东新区促进无人驾驶智能网联汽车创新应用规定》第30条。

^{⑥9} 2023年7月,工信部、国家金融监督管理总局联合发布了《关于促进网络安全保险规范健康发展的意见》;2023年7月,国家市场监督管理总局和国家标准化管理委员会联合发布了《信息安全技术 网络安全保险应用指南》(征求意见稿);2023年12月,工信部发布了《关于组织开展网络安全保险服务试点工作的通知》。

^{⑦0} See Anat Lior, *supra* note 67, 499-504; 刘艳红:《自动驾驶的风险类型与法律规制》,载《国家检察官学院学报》2024年第1期,第119页。

五、结 语

人类社会正在快速进入以新一代人工智能为标志的智能时代。人工智能有望赋能千行百业,但当前其自身发展也需要治理赋能。人工智能所处的发展阶段、人工智能发展的机遇和挑战、人工智能风险的基本特征、人工智能治理能力的现存缺陷、既往规制模式的局限性等,都共同指向“赋能型人工智能治理”。如何营造有利于人工智能创新发展良好法治环境和人文环境,保障人工智能技术科学研究、安全运用、向善发展,增强人工智能企业的合理预期,增强监管能力,增强社会对于人工智能应用的信任,进而有效推动人工智能安全可信发展,是亟待回答的重要问题。应通过依法有效治理来赋能人工智能安全可信发展,并最终赋能社会、赋能人民。我们需要确立赋能型人工智能治理新理念,健全以法治为核心的赋能型人工智能治理新机制,迈向赋能型人工智能治理新格局,将科技发展与国家治理体系和治理能力现代化有机结合,服务于数字中国建设和中国式现代化进程,并为全球人工智能治理作出贡献。

Abstract: Artificial Intelligence (AI) governance has become a frontier issue and an important area of national and social governance. However, there is an urgent need to strengthen capacity building in various aspects, including AI technological innovation, risk prevention and control, corporate self-regulation, government regulation, social supervision and international collaboration, etc. Therefore, enhancing the capacity for safe and trustworthy AI development must be prioritized as the foremost task in AI governance, and the concept and mechanisms of ‘enabling AI governance’ should be established. To achieve this goal, we should adhere to the core concepts of enabling AI governance which are human-centered and development-oriented, as well as the derived basic concepts, including AI for good, inclusiveness and prudence governance, agile governance, and sustainable development. An enabling AI governance mechanism centered on the rule of law should be constructed, along with specific mechanisms under the rule of law, such as perfecting the mechanisms of integrating legal governance and technological governance, establishing co-governance mechanisms that promote communication and collaboration among diverse stakeholders, constructing ‘safe harbor’ mechanisms suitable for AI development, building agile and interactive dynamic regulatory mechanisms that incentivize the development of AI for good, and constructing social security mechanisms such as AI safety insurance.

(责任编辑: 刘海林)