[编者按] 随着人工智能研究与开发的日渐成熟及应用的大范围扩展，其在诸多方面都面临挑战。《科普研究》策划了以人工智能为主题的圆桌论坛，期望从人工智能的伦理与治理、人工智能与科学传播、公众理解及科普科幻创作等不同视角展开论述，并探讨可能的应对之策。

构建稳健敏捷的人工智能伦理与治理框架

段伟文 *

（中国社会科学院科学技术和社会研究中心、中国社科院大学，北京 100732）

[摘　要] 在各国和地区纷纷提出人工智能发展规划和战略的同时，人工智能的伦理风险与治理也成为全球共同关注的焦点。鉴于不同区域社会环境和文化背景的不同，很难构建单一的全球性人工智能伦理与治理框架。中国在总体上对于人工智能为经济、社会、企业和个人福祉带来的积极影响更加乐观，但人工智能的伦理风险并非虚构，普通用户在享受各种创新便利的同时，难免对个人数据滥用、算法决策的不透明心存疑虑，开发者也担心伦理缺位会使其为由此带来的风险付出高昂代价。为了消除这种双重焦虑，应该通过技术伦理评估、“技术－伦理”矫正和信任机制的构建，展开必要的价值伦理校准。更重要的是，要在充分考量人工智能的社会影响、对区域与全球的兼容和维护和平底线的基础上，构建一种稳健可行的人工智能伦理与治理框架，实现敏捷治理。

[关键词] 人工智能伦理与治理　人工智能信任机制　“技术－伦理”矫正　敏捷治理

[中图分类号] TP18；B82–057　[文献标识码] A　[DOI] 10.19293/j.cnki.1673–8357.2020.03.002

自 2016 年人工智能围棋“阿尔法狗”（AlphaGo）击败李世石以来，新一轮人工智能热潮以前所未有的态势席卷世界。在各国和地区纷纷提出人工智能发展规划和战略的同时，人工智能的伦理风险与治理也成为全球共同关注的焦点。在全球范围内，人们对人工智能前景的认知明显地呈现出两极化：一方面，对其可能极大促进经济社会发展的预期兴奋不已；另一方面，对其可能带来的破坏民生等危害与风险愈益担忧。尽管各国和地区、相关国际组织、行业机构等纷纷提出了各种人工智能的伦理与治理方面的原则、规范、宣言、公约、共识等，而且大多都强调负责任、可信赖、尊重隐私、公平公正、安全可控等基本理念，但鉴于不同区域社会环境和文化背景的不同，各个国家和地区对具体的伦理权利、责任和是非的认知存在微妙的差异，很难构建单一的全球性人工智能伦理与治理框架。因此，构建人工智能伦理与治理框架的基本思路在于立足本土实际情

收稿日期：2019–10–31
基金项目：国家社科基金重大项目“智能革命与人类深度科技化前景的哲学研究”（17ZDA028）。
* 作者简介：段伟文，中国社会科学院科学技术和社会研究中心主任、研究员，中国社科院大学特聘教授、博士生导师，研究方向：科学哲学、信息技术哲学，人工智能的哲学、伦理及社会研究等，E–mail：duanweiwencass@163.com。

况，同时加强不同地区的框架之间的互操作性和协同性，进而搭建起全球人工智能伦理与治理框架的生态体系。包括中国在内的亚洲地区，尽管人工智能创新与应用发展十分迅猛，但由于受到儒家文化等东方群体导向的文化的影响，公众在认识和法律层面对于隐私、数据权利、歧视等相关价值诉求与权利界定不甚明确，伦理与治理框架也尚未系统构建，由此所形成的张力使人工智能发展中的价值冲突与伦理抉择变得更为突出。

对此，《麻省理工技术评论洞察》最近发布的一份题为“亚洲人工智能议程：人工智能伦理”的报告指出，在亚洲，随着人工智能大规模的工业化和商业化，人工智能几乎无处不在，正在成为经济增长的主要动力，相关研究机构和企业在对前景充满乐观的同时，也希望政府能推动健全的伦理规范和治理体系的建设，使人们从人工智能中获得更多的好处[1]。但我们应该看到，这种乐观和实用的立场对人工智能伦理与治理带来了一系列不容小觑的挑战，应对其予以适当的价值校准。

1 乐观实用立场亟待价值伦理校准

近 30 年来，在追赶信息通信技术和网络数字技术的过程中，中国成为全球科技进步最快的区域。如今，在迈向全球人工智能发展的快车道上，人们无疑比 1/4 个世纪之前跟随美国和欧洲拥抱互联网时更为主动与自信。面对这一波的人工智能热潮，曾以新兴科技带动社会发展的经验，再加上经世致用和民生优先的文化传统，使政府和企业对人工智能的基本立场更为乐观，也更具实用主义色彩，在整个社会也自然地形成了相对有利于人工智能探索的制度文化环境。在此环境下，中国在总体上对于人工智能为经济、社会、企业和个人福祉带来的积极影响更加乐观，这使得虽然相关的研究者、行业组织、标杆企业特别是业界领袖已指出人工智能应用可能带来潜在的伦理风险，但在人工智能发展战略中，伦理考量的优先级相对于欧洲等其他地区较低。由此，一旦出现由人工智能的滥用和恶意，很可能导致偏见、排斥和社会不公等问题。

必须指出的是，人工智能的伦理风险并非虚构。在各种人工智能应用场景中，数据分析、内容推荐、人脸识别等应用直接涉及和影响到人的身份和行为，相关技术的滥用对人造成的危害和负面影响将远远大于传统的网络与数字技术。如果不加反省地固守乐观主义和实用主义的立场，人工智能对社会和个人造成的潜在威胁往往容易被忽略或无视，由此所引发的伦理风险，既得不到事先防范也难以事后纠正，非但难免酿成严重后果，还会破坏整个社会对人工智能的信任，打击普通公众对发展新兴科技的信心，最终必然给创新生态系统中的机构和企业带来巨大的风险和损失。因此，在这种乐观主义立场和实用主义态度背后，实际上潜藏着相关利益群体的双重焦虑：一方面，人工智能用户在享受各种创新便利的同时，难免因为担心个人数据滥用、算法决策的不透明而心存疑虑；另一方面，人工智能开发者也会担心伦理缺位会使其为由此带来的风险付出高昂代价。为了从根本上消除这种双重焦虑，应该通过技术伦理评估、“技术 – 伦理”矫正和信任机制的构建等方面入手，对乐观主义立场和实用主义的态度展开必要的价值伦理校准。

首先，要通过技术伦理评估克服乐观实用立场对人工智能伦理问题的实质性忽视。针对人工智能的伦理风险，有人提出应该强调伦理优先，但不一定要将创新与伦理对立

起来。实际上，各种场景中的价值伦理问题是否应该得到优先考虑，取决于相关群体对具体的技术应用的负面影响及其严重程度的厘清和认知，而这又必须借助各种形式的技术伦理评估。具体而言，促进开发者和使用者切实认知人工智能伦理风险的关键在于：一方面，技术监管者应该要求企业对技术应用在相关群体中造成的负面影响展开预见性的伦理评估，进而对其加以矫正；另一方面，应该通过建设性与参与性的伦理评估，使可能因技术应用而受到严重负面影响的相关群体参与到相关的伦理评估、争论与矫正等环节之中。

其次，运用“技术－伦理”矫正，以包容审慎的监管实现伦理与创新的协同。毋庸置疑，在中国的人工智能研究与创新生态系统内外，正在就人工智能伦理问题进行各种形式的探讨与辩论，目前政府、业界、媒体及公众已经达成的基本共识是政府应该主导对人工智能的监管。但在业界和相关部门的认知中，多少有些吃不准的是，他们希望相关政策、框架和监管应审慎行事，不应扼杀创新。对此，不应错误地将伦理上的相对宽松视为创新的“伦理优势”，而应该认识到，对于那些具有高度价值敏感性的人工智能应用，应该努力寻求一种将伦理价值融合到技术之中的复合创新，从而使创新与伦理相辅相成、协同并进。具体而言，就是通过有针对性的“技术－伦理”矫正，落实对人工智能的伦理治理。在此过程中无疑需要通过科技伦理专家与科技专家的对话，以实现价值诉求与技术需求之间的“转译”。由此，在讨论推荐算法可能导致的“过滤气泡”时，不应停留在对具有“多样性”的推荐算法之类的倡导，而应该通过批评者与设计者的对话，共同探讨如何进行反向的“技术－伦理”矫正。这种使技术合乎伦理的矫正又称为道德敏感设计——将伦理道德内置于技术设计，在信息技术等新兴技术中已有长期的运用实践。早在 21 世纪初，现任哈佛大学教授的拉坦娅·斯威尼（Latanya Sweeney）就曾在她当时任职的卡内基梅隆大学创建了数据隐私实验室，开展隐私设计方面的研究，如通过软件帮助人们堵住网上的隐私漏洞、通过随机换脸应对人脸识别技术的滥用可能导致的隐私泄露和侵权[2]。相关实践探索表明，人工智能的伦理和治理可以贯穿于人工智能创新应用的全过程，为了实现伦理与创新的协同，要将其背后的价值观转换为技术层面的目标与需求。

最后，以信任机制的构建作为人工智能伦理与治理的突破口。我们看到，尽管政府和企业正在更加积极地为人工智能行业的伦理发展制定目标和指导方针，但客观地讲，由于人工智能处于开放性的高速发展之中，目前只能通过个案处理和经验积累，对人工智能滥用和恶意使用的不良后果进行伦理和法律上的规制。因此，一种相对可行的伦理与治理策略是，转而诉诸消费者、人工智能用户和人工智能开发商之间的信任，在人工智能研究与创新生态系统内外构建起可持续的多方信任机制，使整个行业得到全社会的接受，进而为其在多方共治下的健康发展奠定基础。毋庸置疑，可持续的多方信任机制的构建不可能建立在简单的信任与承诺之上，而要求人工智能开发者承担起应有的机构与企业责任，开展负责任的研究与创新，使整个行业和产业以一种负责任和透明的方式发展。以当前的人工智能涉及大量数据驱动的智能应用为例，其中所涉及的与人相关的数据反映了人格特征、行为方式和具体行为，企业必须就如何恰当地收集、使用和共享数

据与客户及利益攸关者展开沟通，只有这样才可能避免用户的疑虑和不信任，使创新与应用高效有序地展开。

2 构建稳健可行的人工智能伦理与治理框架

近 20 年来，发展迅猛且具有高度不确定性的新兴科技不断涌现，由此产生了人工智能、物联网、基因编辑等可能对经济社会与人类未来造成根本性、深远性、普遍性影响的颠覆性技术，如何针对人工智能等技术构建稳健可行的伦理与治理框架无疑是各个国家与地区所必须面临的最大挑战。虽然对此并没有最佳答案和统一标准，但还是有一些基本理念有助于进一步的实践探索。

一是要对人工智能可能造成的社会影响进行更加深入系统的研究，在此基础上构建起考虑到最坏情形的社会风险防范机制。从人工智能对未来社会的发展角度来看，最为重要的问题是人工智能对就业的影响。尽管一般认为，人工智能驱动的技术性失业的困扰将被创新所带来的新机遇和人类创造与适应新的工作的潜力所抵消，但应该看到，由于我国劳动密集型工业和服务业的“可自动化”的比例较发达国家更高，人工智能及其所带来的新一轮的自动化浪潮在短时间内对于中国的具体影响可能是巨大的，特别是那些受到人工智能威胁的低技能职业阶层再培训和再提升技能的能力更弱。因此，在对未来普遍持有乐观态度的同时，政府对所主导的人工智能伦理与治理框架必须看到最坏的情形，充分考虑到可能被智能机器替换的劳动力的出路，以此避免人工智能的颠覆性发展带来新的社会不平等。

二是要从文化出发构建一种兼具区域性与全球性的人工智能伦理与治理体系框架。正如前文所言，由于在表达伦理的“正确”方式上存在着巨大的文化差异，加上科技创新与应用文化本身具有开放性，人工智能伦理与治理并没有“全球性”。毫无疑问，人们对人工智能的不同态度确定了不同区域与社会发展人工智能的不同方法和旨趣。以前卫的性爱机器人为例，具有神道教传统与丰富人偶文化的日本可能更容易接受其研发与应用。在未来的无人汽车等智能无人系统中，个人主义与集体主义等不同社会主流价值导向也会在一定程度上影响机器的价值取向和伦理抉择。因此，文化和价值观上的可接受性是构建人工智能时代的人机信任与人机和谐的关键所在。但与此同时，必须指出的是，人工智能的发展必然会加速推进更深层次的全球化，区域性的人工智能伦理与治理不仅要具有局域的适切性，还应该建立起通达全球治理的机制。

在中国的《新一代人工智能发展规划》[3]中，不仅设想通过人工智能提高社会管理能力，还承诺对隐私与知识产权、信息安全、问责、设计伦理等展开研究，并积极参与全球治理。在此过程中，如何讲清区域性伦理与治理框架的核心价值取向和伦理诉求是关键，不论在什么文化和社会主流价值观下，都必须对人工智能时代的隐私、尊严、个人权利等基本的伦理和法律概念做出明晰与合理的诠释，都要对数据的采集与使用、智能系统对人的行为的影响与干预的限度做出合乎伦理与法律的准确陈述。而对这些问题的深究最终所涉及的与其说是价值和伦理问题，毋宁说关乎更大社会历史背景与文化语境下的社会政治抉择，这就使得区域性的人工智能的伦理与治理框架之间的对话成为其整合为全球性治理架构的难点和关键所在。

三是要在构建人机和谐的未来的同时确保人类和平这一文明底线。在我们试图通过

人工智能发展构建人机和谐未来之时，应该认识到人与机器的关系实质是人与人以机器为中介的关系，对人工智能等技术的伦理规制和治理的关键是避免其对人作恶。因此，人工智能的伦理与治理要对人工智能的恶意使用和军事运用等负面用途加以必要的法律管制、伦理规制全球治理。随着深度造假等技术的发展，人工智能的恶意使用会越来越大，如果不加以及时的法律管制和伦理规制，必将极大地破坏整个社会对人工智能等颠覆性技术的信任，影响其有益的创新与应用。同时，应该看到，人类科技文明发展的悖谬之处恰恰在于包括计算机、互联网和人工智能在内的很多颠覆性的技术与军事对抗和国家竞争密切相关。我们必须承认，在这一波人工智能热潮背后，或明或暗地存在着人工智能军备竞赛的阴影，无人机及大规模智能自动武器系统等人工智能的军事应用，将对人类文明的未来构成前所未有的威胁，必须得到有效的控制和全球性的治理。

3 结语：走向稳健敏捷的人工智能伦理与治理

2019 年 6 月 17 日，国家新一代人工智能治理专业委员会发布了《新一代人工智能治理原则——发展负责任的人工智能》，提出了“和谐友好、公正公平、包容共享、尊重隐私、安全可控、共担责任、开放协作和敏捷治理”八项原则 [4]。其中的敏捷治理强调，应在人工智能的发展中及时发现和解决可能引发的风险、推动治理原则贯穿人工智能产品和服务的全生命周期、确保人工智能始终朝着有利于人类的方向发展。而实现这一原则的关键在于切实将人工智能发展中的所有相关利益群体的认知纳入人工智能伦理与治理的各个环节，诉诸面向人工智能研究创新实践的具体的价值校准和稳健的框架设计，以此形成多相关方的对话与共识机制，落实多方参与，实行多元共治，实现从微观到宏观的反馈、修正与迭代，最终推动人工智能的未来沿着合乎人性的方向和谐发展。

参考文献

[1] MIT Technology Review Insights.（2019）Asia’s AI agenda：The ethics of AI [EB/OL].（2019–07–11）[2020–02–10] https://mittrinsights.s3.amazonaws.com/asiaaiethics.pdf.
[2] 奇普·沃尔特：网络侦探弥补隐私漏洞 [J]. 环球科学，2007(8).
[3] 中华人民共和国中央人民政府 . 国务院关于印发新一代人工智能发展规划的通知 [EB/OL].（2017–07–08）[2020–02–10] http://www.gov.cn/zhengce/content/2017–07/20/content_5211996.htm.
[4] 中华人民共和国科学技术部 . 发展负责任的人工智能：新一代人工智能治理原则发布 [EB/OL].（2019–06–17）[2020–02–10]. http://www.most.gov.cn/kjbgz/201906/t20190617_147107.htm.

（编辑 姚利芬）