2019 年第 4 期（总第 4 期）

国际战略与安全研究报告

INTERNATIONAL SECURITY AND STRATEGY STUDIES REPORT

人工智能治理的原则和关键

清华大学战略与安全研究中心
CENTER FOR INTERNATIONAL SECURITY AND STRATEGY
TSINGHUA UNIVERSITY

人工智能治理的原则和关键

傅莹　李睿深

人工智能的广泛应用给人类的生产和生活带来了很大的便利，未来的潜力更是有可能带来颠覆性的影响。与此同时，其风险和挑战也正在引起全球范围的担忧。2015 年 1 月，包括著名物理学家霍金在内的全球数百名人工智能专家和企业家签发了一封公开信，警告说，如果不对人工智能技术进行有效限制，“人类将迎来一个黑暗的未来”。由此引发的担忧和恐惧，已经成为媒体和社会舆论的热门话题，很多国家和组织已经开始考虑构建人工智能的安全治理机制。

2017 年，全球行业领袖制定《阿西洛马人工智能原则》，为技术发展制定了“有益于人类”的自律守则；欧盟委员会也发布了人工智能道德准则；经济合作与发展组织（OECD）于 2019 年正式通过了首部人工智能的政府间政策指导方针，确保人工智能的系统设计符合公正、安全、公平和值得信赖的国际标准；二十国集团（G20）也出台了倡导人工智能使用和研发“尊重法律原则、人权和民主价值观”的《G20 人工智能原则》；中国国家新一代人工智能治理专业委员会发布的《新一代人工智能治理原则》，提出发展负责任的人工智能。

一、治理的六项原则

2018 年 7 月，清华人工智能治理项目小组在世界和平论坛[①]上提出了“人工智能六点原则”，为人工智能的综合性治理提供了一个宏

① 注：1950 年，图灵发表论文《计算机器与智能》（Computing Machinery and Intelligence），提出“图灵测试”的概念，即如果人类测试者在向测试对象询问各种问题后，依然不能分辨测试对象是人还是机器，那么就可以认为机器是具有智能的。

观框架：一是福祉原则。人工智能的发展应服务于人类共同福祉和利益，其设计与应用须遵循人类社会基本伦理道德，符合人类的尊严和权利。二是安全原则。人工智能不得伤害人类，要保证人工智能系统的安全性、可适用性与可控性，保护个人隐私，防止数据泄露与滥用。保证人工智能算法的可追溯性与透明性，防止算法歧视。三是共享原则。人工智能创造的经济繁荣应服务于全体人类。构建合理机制，使更多人受益于人工智能技术的发展、享受便利，避免数字鸿沟的出现。四是和平原则。人工智能技术须用于和平目的，致力于提升透明度和建立信任措施，倡导和平利用人工智能，防止开展致命性自主武器军备竞赛。五是法治原则。人工智能技术的运用，应符合《联合国宪章》的宗旨以及各国主权平等、和平解决争端、禁止使用武力、不干涉内政等现代国际法基本原则。六是合作原则。世界各国应促进人工智能的技术交流和人才交流，在开放的环境下推动和规范技术的提升。

这六项原则为人工智能治理的讨论和共识构建提供了一种可能，在去年底的世界互联网大会和今年的世界和平大会上，国际上很多学者和企业家都对此表达出了兴趣和重视，不少机构希望进一步合作研讨。目前企业界已经出现了一些自律的尝试，如在产品程序中加入“禁飞策略”来规范无人机的使用；又或医疗和交通业界通过数据脱敏，既有效保护了个人隐私信息，又有利于形成数据资源利用的良性循环。现在的任务是，如何在国际社会推动这些原则落地，形成更有加务实、更具操作性的治理机制。

二、治理机制的关键

国际治理机制不仅意味着共识和规则，也应包括确保规则落地的组织机构和行动能力，甚至要有相关的社会政治和文化环境。清华大学战略与安全研究中心正在与一些国家的学者专家、前政要和企业家

一道，对相关问题进行探讨。从现实来看，人工智能国际治理的有效机制至少应包括如下五个关键：

（一）动态的更新能力

人工智能技术的研发和应用都进入快速发展的阶段，对未来的很多应用场景以及安全挑战，目前还有许多不明确之处。因而，对其治理须充分考虑到技术及其应用的变化，建立一种动态开放的、具备自我更新能力的治理机制。

例如，需要向社会提供人工智能“恶意应用”的具体界定和表述，这种表述应该在生产和生活实践中可观测、可区分的，在技术上可度量、可标定。更为重要的是，它应当是持续更新的。只有具备动态更新能力的治理机制才能在人工智能技术保持快速发展的情况下发挥作用。

这就意味着，在推进治理的同时，要主动拥抱人工智能技术的不确定性，做好在思维模式上不断调整的准备。爱因斯坦曾说，“我们不能用制造问题时的思维来解决问题。”颠覆性创新技术与固有思维之间的冲突与激荡，必将伴随着人工智能治理的全过程。在此情景下的治理机制，也应该对各种思潮和意见的交织和反复具备足够的包容之心和适应能力。这一机制将帮助人类携手应对人工智能层出不穷的新挑战。从这个意义上讲，建立一个能够适应技术不断发展的动态治理机制，也许比直接给出治理的法则更有意义。

（二）技术的源头治理

人工智能的应用，本质上是一项技术的应用，对其治理必须紧紧抓住其技术本质，特别是人工智能的安全治理问题，从源头开始实施治理，更容易取得效果。例如当前大放异彩的主要是深度学习技术，其发关键要素是数据、算法和计算力，于是，我们可以从数据控流、

算法审计、计算力管控等方面寻找治理的切入点。

随着人工智能技术的飞速发展，今后可能出现迥然不同的智能技术，例如小样本学习、无监督学习、生成式对抗网络，乃至脑机技术等等。不同的技术机理意味着，应该不断致力于从技术源头寻找最新、最关键的治理节点和工具，将其纳入治理机制之中，以维护治理的可持续性。

另外技术治理还有一个重要内容，就是在技术底层赋予人工智能“善用”的基因。例如在人工智能武器化的问题上，是否可以像小说家阿西莫夫制定“机器人三原则”那样，从技术底层约束人工智能的行为，将武装冲突法则和国际人道主义法则中的“区分性”原则纳入代码，禁止任何对民用设施的攻击。当然这是一个艰巨的挑战，曾在美国国防部长办公室工作、深度参与自主系统政策制定的保罗·沙瑞尔就认为[②]：“对于今天的机器而言，要达到这些标准（区分性、相称性和避免无谓痛苦）是很难的。能否实现要取决于追求的目标、周围的环境以及未来的技术预测。”

（三）多维的细节刻划

人工智能的国际治理必须构建一种多元参与的治理生态，将所有的利益相关方纳入其中。学者和专家是推动技术发展的主力，政治家是国家决策的主体，民众的消费需求是推动各方前进的关键激励因素。这些群体之间的充分沟通和讨论是人工智能治理的意见基础。企业是技术转化应用的核心，学术组织是行业自律的核心，政府和军队是人

② 1990 年，美国未来学家雷·库兹韦尔在《奇点临近》、《人工智能的未来》两本书中，用“奇点”作为隐喻，描述人工智能的能力超越人类的某个时空阶段。当人工智能跨越“奇点”后，一切我们习以为常的传统、认识、理念、常识将不复存在，技术的加速发展会导致“失控效应”，人工智能将超越人类智能的潜力和控制，迅速改变人类文明。

工智能安全治理的核心，这些组织之间的沟通是人工智能技术治理机制能够真正落地的关键。

在这个生态中，不同的群体应该从自身视角对人工智能的治理细则进行更加深入的刻画。例如，今年 8 月亨利·基辛格，埃里克·施密特，丹尼尔·胡滕洛赫尔三人联合撰文提出，从人工智能冲击哲学认知的角度看，可能应该禁止智能助理回答哲学类问题，在影响重大的识别活动中强制人类的参与，对人工智能进行“审计”，并在其违反人类价值观时进行纠正等等。③

如果能将来自不同群体治理主张的细则集聚在一起，将形成反映人类多元文化的智慧结晶，对人类共同应对人工智能挑战发挥正本清源的作用。涓涓细流可以成海，哲学家们对于真理与现实的担忧与普罗大众对于隐私的恐惧一样重要，只有尽可能细致地刻划人工智能治理的各种细节，迷茫和恐惧才能转变为好奇与希望。

（四）有效的归因机制

在人工智能的国际治理机制中，明晰的概念界定是治理的范畴和起点，技术源头的治理是关键路径，多利益相关方的参与是治理的土壤。归因和归责在整个治理机制发挥着“托底”的作用，如果不能解决“谁负责”的问题，那么，所有的治理努力最终都将毫无意义。

当前人工智能治理的一个重大障碍就是归因难：从人机关系的角度看，是不是在人工智能的应用中，人担负的责任越大、对恶意使用的威慑作用就越大，有效治理的可能性就越大？从社会关系的角度看，在各利益相关方都事先认可人工智能具有“自我进化”可能性的情形下，程序“自我进化”导致的后果，该由谁负责？“谁制造谁负责”、

③ 约翰·R·麦克尼尔、威廉·H·麦克尼尔：《麦克尼尔全球史：从史前到 21 世纪的人类网络》，王晋新等译，北京大学出版社，2017 年，第 72–79 页。

“谁拥有谁负责”？还是“谁使用谁负责”？

从技术的角度看，世界上没有不出故障的机器，如同世上没有完美的人，人工智能发生故障造成财产损失、乃至人员伤亡是迟早会出现的。难道我们真的应该赋予机器以“人格”，让机器承担责任？如果我们让机器承担最后的责任，是否意味着，人类在一定范围内将终审权拱手让给了机器？

（五）场景的合理划分

在人工智能发展成为“通用智能”之前，对其实施治理的有效方式是针对不同场景逐一细分处理。从目前的发展水平看，人工智能的应用场景仍然是有限的。在 2019 年 7 月的世界和平论坛上，很多与会学者都认为现在应尽快从某几个具体场景入手，由易到难地积累治理经验，由点及面地实现有效治理。

划分场景有助于我们理解人工智能在什么情况下能做什么，这一方面可以避免对人工智能不求甚解的恐惧，另一方面也可以消除对人工智能作用的夸大其词。例如，美国国防部前副部长罗伯特·沃克（Robert O. Work）一直是人工智能武器化的积极倡导者，但是，具体到核武器指挥控制的场景上，他也不得不承认，人工智能不应扩展到核武器，因为可能引发灾难性后果。④

有效的场景划分，应尽可能贴近实际的物理场景和社会场景，更应该注意数据对于场景的影响。这是因为当前的人工智能技术是高度数据依赖性的，不同的数据可能意味着不同的场景，也就是说，场景至少应该从物理场景、社会场景和数据场景三个维度加以区分。

④ Greg Allen and Taniel Chan, Artificial Intelligence and National Security, (Cambridge: Belfer Center for Science and International Affairs, Harvard Kennedy School, 2017), p. 1.

三、结语

对血肉之躯的人类而言，任何一项新技术都是一把双刃剑，几乎每一次重大的技术创新，都会给当时的人们带来不适与阵痛。但是，人类今日科学之昌明，生活之富足的现实足以证明，人类在新技术治理方面有足够的智慧，只要对新技术善加利用、科学治理，任何由此而来的新威胁都能得到圆满解决。我们相信，国际社会一定能形成良性的治理机制，享受人工智能技术带来的更繁荣、更安全的世界。

清华大学战略与安全研究中心

办公地点：清华大学明斋 217

联系电话：010-62771388

电子邮箱：ciss_thu@163.com