体制复合体理论视角下人工智能全球治理进程*

鲁传颖　　约翰·马勒里

摘　要：体制复合体理论包含多利益攸关方与多边治理等两种治理模式，在应对网络空间治理等复杂性全全球治理议题时有很强的解释力，它同样适用于复杂的人工智能的全球治理。体制复合体理论中所包含的多方和多边治理两种治理模式，可以分别应用在人工智能全球治理中不同领域治理议题中。如在算法歧视、伦理问题、就业问题、数据安全等议题适用于多方治理模式，而与全球经济、安全、政治相关的议题更适用于多边治理模式。最终，通过多方和多边等体制复合体包含的多种治理模式共同构建起以标准、规范、规则和机制等不同层次、不同领域的人工智能全球治理体系。

关键词：人工智能　算法歧视　体制复合体　多利益攸关方

中图分类号：D80　文献标识码：A　文章编号：

人工智能是利用数字计算机或数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。[①]人工智能作为一种通用型技术，将推动人类从信息化时代走向智能化时代，由此而带来的全球治理问题，成为与人工智能发展同样重要。人工智能的跨

* 本研究得到教育部哲学社会科学研究重大课题攻关项目“构建全球化互联网治理体系研究”资助（项目批准号 17JZD032）

① 中国电子技术标准化研究院：《人工智能标准化白皮书》（2018 年版），第 5 页，http://www.cesi.ac.cn/201801/3545.html

专业、跨领域、跨议题给治理问题研究带来了很大的难题。作为一个复杂的全球治理领域，可以借鉴体制复合体理论的分析视角对其进行研究。体制复合体(Regime Complex)理论是应用体制理论(Regime Theory)分析复杂性全球治理问题中产生的一种新的理论视角。网络空间治理等新型全球治理议题及到的治理领域宽泛、行为体多元、议题内涵丰富，没有一种现成的理论可以用来解释这些治理的现象，也没有一种体制可以独自对其进行治理。[①]参照体制复合体在网络空间全球治理领域的定义，可以认为人工智能全球治理的目标是根据不同的议题构建包括标准、规范、规则、条约、法律等不同层次的治理机制，最终由各种机制之间的松散耦合组成的体制复合体。[②]

一、体制复合体理论与人工智能全球治理

传统的全球治理问题往往是由国际政治、安全和经济所主导。约瑟夫·奈认为，体制复合体理论具有一个鲜明的特点就是在关注政治、安全和经济等要素之外，同时强调技术的逻辑的重要性，将技术社群的治理放在同样的地位。以此为视角来看，人工智能的全球治理进程已经在政府、国际组织、技术社群和私营部门之间展开，并在人工智能的伦理、算法、就业、安全、数字鸿沟等领域制定了一系列的标准、规范和规则。

1. 人工智能全球治理的现状

人工智能在名称中包含人、智能等意思，自人工智能作为一门学科诞生的那一天起，关于人工智能治理的讨论就从未停止过。人工智能全球治理与其发展阶段有密切关联。人工智能的发展从学理层面可以定义为三个阶段，即弱人工智能阶段(Artificial Narrow Intelligence, ANI)、类人工智能阶段(Artificial General Intelligence, AGI)和强人工智能阶段(Artificial Superintelligence, ASI)，分别对应机器的智能程度。目前的人工智能发展还处于弱人工智能阶段，意味着机器会在某些模块和方面取代和超越人，但缺乏完全自主意识，只能在设定好的环境中工作，无法应对新的环境并产生新功能，离全面赶上人类还存在较大差距。[③]

---

① Joseph Nye, “The Regime Complex for Managing Global Cyber Activities”, Scholarly Articles, 2014.

② Joseph Nye, “The Regime Complex for Managing Global Cyber Activities”, Scholarly Articles, 2014.

③ Tim Urban, “The AI Revolution: The Road to Superintelligence”, Huffington Post, Feb 10 2015, https://www.huffingtonpost.com/wait-but-why/the-ai-revolution-the-road-to-superintelligence_b_6648480.html

由于人工智能技术的复杂性，关于人工智能治理的讨论基本上也是建立在对人工智能的理解之上。早期关于人工智能的讨论具有很强的伦理导向，跳过弱人工智能阶段和类人工智能阶段，直接关注强人工智能阶段所带来的“机器是否会控制和伤害人类？”等议题，使得讨论逐渐陷入一场持久而无果的争论之中。[①] 随着人工智能被划分为三个不同发展的定义开始出现，并清晰对当前处于弱人工智能阶段的发展特征进行定义，以建立规范、规则、机制为导向的人工智能全球治理开始出现，并将会对人工智能的技术发展和应用产生重要影响。[②]

人工智能的全球治理工作已经在不同领域展开。技术社群是目前最活跃的行为体，已经构建了多个治理机制。以 2017 年 1 月初举行的“受益的人工智能”（Beneficial AI）会议为基础，建立了“阿西洛马人工智能原则”，主要的参与者是人工智能的开发者和公民组织，会议总结了科研问题（Research Issues）、伦理和价值（Ethics and values）、更长期的问题（Longer-term Issues）等 3 大类，23 条规则。[③]其中最核心的是强调应用应当嵌入一些基本的伦理基础如人控制机器而不是机器控制人，并且具有价值观导向，如注重平等性、隐私、人权等。电气和电子工程师协会（IEEE）建立了关于人工智能的全球倡议，并指出要确保从事自主与智能系统设计开发的利益攸关方优先考虑伦理问题，只有这样，技术进步才能增进人类的福祉。IEEE 建立了《人工智能设计的伦理准则》，提出了人权、福祉、问责、透明、慎用五大总体原则，并且依据准则成立了 IEEEP7000 标准工作组，设立了伦理、透明、算法、隐私等 10 大标准工作组。[④]此外，微软、美国信息技术产业理事会、经合组织、斯坦福大学《人工智能百年研究》、英国标准研究院等企业和机构也发布了各种有关于人工智能治理的报告。

联合国系统也高度重视人工智能的治理机制问题，国际电信联盟（ITU）2017 年 6 月在日内瓦召开了“人工智能造福人类”（AI for Good）峰会，ITU 作为联合国机构，敏锐的意识到人工智能的发展和应用会带来新的数字鸿沟，造成新的全球不平等。因此，提出人工智能的发展和应用应当符合联合国可持续发展的目标。会议提出了利用人工智能来促进全球在脱贫、健康、医疗、教育 17 大领域

① Cellan-Jones Stephen, “Hawking warns artificial intelligence could end mankind”, BBC News. Dec 2 2014, Available from: http://www.bbc.com/news/technology-30290540.

② IEEE, Ethically Aligned Design, 2017 年 12 月 23 日，
http://standards.ieee.org/news/2016/ethically_aligned_design.html

③ Asilomar Ai Principles, https://futureoflife.org/ai-principles/

④ IEEE, Ethically Aligned Design, 2017 年 12 月 23 日.

的可持续发展目标。①联合国犯罪与司法研究所在荷兰海牙成立了第一个联合国人工智能和机器人中心，该中心致力于通过提高认知、教育、信息交换和协调利益相关者，来了解和处理与犯罪相联系的人工智能和机器人带来的安全影响和风险。②此外，人工智能已经引起了各国政府的高度重视。自 2015 年以来，美国、中国、法国、欧盟等主要大国和区域组织都积极出台国家战略，支持和引导人工智能的发展。美国政府发布了《为人工智能的未来做好准备》政策报告，并成立了人工智能委员会来促进人工智能的发展和治理，英国政府发布了《人工智能：机遇和未来决策的应用》。这些发展报告都对人工智能的全球治理问题表明了立场和观点。

2. 从网络空间全球治理到人工智能全球治理

体制复合体理论在分析网络空间治理时主要是通过多利益攸关方模式来构建技术领域的治理机制，通过多边模式来应对技术应用所带来的安全、政治、经济等领域的治理机制。人工智能在技术上有跨学科性质，同时也包含多元的治理议题和参与行为体，这为通过体制复合体理论分析人工智能全球治理提供了基础。首先，从技术逻辑上来说，人工智能是一种基于计算机和互联网技术并结合其他学科知识组成的交叉型学科，其发展和治理拥有与生俱来的跨学科、跨领域和多元行为体等特点。除了计算机科学之外，统计学、脑科学、系统辨识、优化理论、神经网络等也成为人工智能不可或缺的基础知识。人工智能既是一个具有颠覆性意义的重要技术，同时也是一个复杂的交叉性学科。③

其次，人工智能全球治理包涵多层级、跨领域的治理议题。人工智能具有通用性和颠覆性等特点，它的应用具有全球性的广度和深度。体制复合体将传统国际关系所忽视的技术逻辑上升为与政治、经济、安全同样重要的地位，反映了网络空间治理领域“代码即法律”“代码即规则”的独特性。人工智能与网络空间一样，治理的内涵不仅要充分考虑技术的逻辑，同样要对技术应用以及对国际体系的影响。如从技术本身来看，人工智能算法所涉及到的伦理问题、价值问题是人类共同面临的挑战，具有普适性；从应用角度来看，各国在人工智能发展和应用上的先后将会造成新的不平等和不平衡，引起新的数字鸿沟；从国际体系的角

① ITU, “AI for Good Global Summit Report”, Geneva, June 2017, p7, https://www.itu.int/en/ITU-T/AI/Pages/201706-default.aspx

② UNICRI: “CBRN National Action Plans: Rising to the Challenges of International Security and the Emergence of Artificial Intelligence”, Oct 7, 2015,http://www.unicri.it/news/article/CBRN_Artificial_Intelligence

③ 同上，第 5-7 页。

度来看，人工智能对于当前国际经济、安全和政治体系的影响会，并最终影响国际体系的稳定。①

再次，人工智能与网络空间全球治理的参与主体都可以被划分为技术社群、私营部门和政府等全球治理行为体。以 IEEE 为代表的技术社群作为人工智能技术的开发者在国际标准的制定和治理的规范形成上具有重要作用；私营部门是规则制约的主要对象，因此对于参与治理有很大的积极性。以微软、谷歌、百度为代表的私营部门不仅具有大量的人才，并且是人工智能技术和应用创新的主要推动力量，另外传统行业也在不断顺应人工智能技术的发展，加入到治理的进程中；政府是人工智能治理的中不可或缺的行为体，当然政府的角色也更加复杂和多元。政府是战略和政策的制定者，同时也是监管者，在涉及到国防、安全等战略技术领域政府是直接的参与者和推动力量。②同时，政府肩负着制定条约和国际规则的责任。因此，关于人工智能的治理必须是各个国家和各个行为体共同参与的全球治理。

最后，从实践层面来看，人工智能的全球治理进程也表现出与网络空间治理进程有相似性。如在数据安全治理议题上，人工智能与互联网治理面临的问题同样都是个人信息安全和国家安全问题。从治理行为体上看，技术社群都发挥着不可替代的作用。技术社群在互联网治理领域同样扮演了关键的角色，如 ICANN 在互联网关键资源的治理上，IETF 在互联网技术标准制定领域的主导性作用。在人工智能领域，以 IEEE 为代表的技术社群在标准和规范制定领域具有重要的作用。以联合国为代表的国际组织在互联网治理领域扮演着重要的规则制定作用，这一点同样适用于人工智能领域，联合国在互联网治理领域成立了信息安全政府专家组，同时在人工智能领域成立了致命性自主武器系统（LAWS）政府专家组（GGE）。③

总体而言，人工智能的治理已经成为全球治理中快速发展的领域。各种不同的行为体都在结合自身的优势和能力在参与治理制度的构建，并最终形成类似与网络空间制度复合体一样的治理生态。联合国、国际电信联盟等政府间国

① 封帅：“人工智能时代的国际关系-走向变革且不平等的世界”，载《外交评论》2018 年第 1 期，第 3 页。

② Gregory C. Allen and Taniel Chan, *Artificial Intelligence and National Security*, Cambridge: Harvard University, July 2017.

③ UNIDIR, *The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches*, Geneva: 2017, http://www.unidir.org/files/publications/pdfs/the-weaponization-of-increasingly-autonomous-technologies-concerns-characteristics-and-definitional-approaches-en-689.pdf

际组织和以及 IEEE 非政府间国际组织积极尝试汇聚各界声音，希望成为未来引领国际规则制定的平台；微软、谷歌等产业界领袖未雨绸缪，通过组建联盟、资助研究等不同形式推动规则制定；一些研究机构和民间团体大力推动广泛开展合作与研讨，期待了解各方诉求，从而在未来的治理平台设计和规则制定中施加影响。

二、多利益攸关方与人工智能全球治理

体制复合体理论认为，人工智能的全球治理由两个层级多个治理议题和治理体制所组成。首先是关于人工智能技术安全（safety）层面的治理议题。最根本的问题是人与机器之间的伦理关系，包括机器是否会取代、控制和伤害人类；人类已有的德道和价值体系如何被机器遵循；①具体的治理议题包括伦理、算法、失业、数据安全等议题，这些议题涉及到技术社群、私营部门和政府等不同的行为体；其次是体系层面的安全（security）治理议题，包括人工智能的大规模应用是否会带来新的不公平，如可持续发展和新数字鸿沟问题等，②人工智能对现存国际经济、安全和政治所带来的影响等③。

从治理模式来看体制复合体理论包含“多利益攸关方”模式和“多边模式”模式(Multi-lateral)。在人工智能领域技术安全层面的治理上，多利益攸关方模式更加适用，它强调治理机制的公开透明、自下而上，强调技术社群的集体身份，不设立参与门槛，超越国家和利益的局限。④体制复合体的理论假设是议题的性质决定了治理模式和构建治理规范的基础。人工智能在伦理、算法、失业、数据安全等技术安全层面议题的性质更多的是强调技术的逻辑，因此，以技术社群为代表的市民社会组织和私营部门是参与治理的主要行为体，治理的目标是将伦理、价值观的考虑嵌入到算法和代码当中，通过全球治理来建立相应的规范和行为准则。

1. 人工智能的伦理问题

关于人工智能伦理治理的目标是要确保机器不会取代、控制和伤害人类；人

① Ed Finn, “What Algorithms Want Imagination in the Age of Computing”, Cambridge, MA : MIT Press, p 12.

② ITU, “AI for Good Global Summit Report”, p15.

③ Julian E. Barnes and Josh Chin, "The New Arms Race in AI," *The Wall street Journal*, March 2, 2018, https://www.wsj.com/articles/the-new-arms-race-in-ai-1520009261

④ 参见鲁传颖：《网络空间治理与多利益攸关方理论》，北京：时事出版社，2016 年版，第 34 页。

类已有的德道和价值体系应当以及如何被机器遵循等。目前关于这一议题的主要治理是以技术社群 IEEE 为代表制定《人工智能设计的伦理准则》和以技术社群与私营部门共同发起《阿西洛马人工智能原则》倡议为主要代表。两者都是都采取的多利益攸关方模式，秉持了公开透明、自下而上等理念，所有感兴趣的人都可以参与，因此两者的规模都达到了上千人。所有的建议都公开地征求参与者的意见，并且都设立了多轮征求意见的过程，确保所有的意见都被认真对待。这两个机制都存在着发展中国家参与程度较小的问题，但是由于中国、印度等国的重视，发展中国家的声音也开始有所体现。

互联网治理领域的“代码治理”是 IEEE 伦理准则的主要理论依据，代码治理认为程序代码代表规则，编写程序就是在创造规则。这对参与治理的行为体提出了非常高的要求，具有编写程序代码的能力的其他领域专家和政府工作人员毕竟数量有限，因此，技术社群是代码治理最主要的支持者和实践者。IEEE 作为专家社群组织，以人工智能的工程师和科学家为对象，提出制定《人工智能设计的伦理准则》(第 2 版)，作为研发人员的参考标准，建立了 10 个标准工作组。其中 IEEEP7000™ -解决系统设计中的伦理问题的建模过程、IEEEP7007™ -伦理驱动的机器人和自动化系统的本体标准、IEEEP7008™ -机器人、智能与自主系统中伦理驱动的助推标准、IEEEP7010™ -合乎伦理的人工智能与自主系统的福祉度量标准等四个标准工作组都是处理与伦理相关的问题。这些标准的制定都是从人权、福祉、问责、透明、慎用五大总体原则来评估每一个指标是否达到了目标。[①]IEEE 由于汇聚了众多人工智能领域的工程师和科学家，并且通过标准工作组将理念落实为行业标准，标准一旦制定完成将会对整个行业发展以及后续各国政策、法律的制定产生重要影响。“阿西洛马人工智能原则”提出的 23 条原则中有很多是对如何处理人与机器之间的关系作出了回应。如从使命角度强调人工智能的发展必须是造福人类，而非不受人类控制，从价值观角度强调，人工智能系统应当与人类保持一致的价值观、接受人类的控制、不能颠覆人类社会秩序。[②]

2．算法歧视

① IEEE, Ethically Aligned Design, p7.

② “阿西洛马人工智能原则”还有一条是“不应当为人工智能能力发展提过高的上限”，即是否需要发强人工智能应当依据实际的需求，而非技术本身。当然这一条没有达成共识。主要原因在于，人工智能的发展目前还处于弱人工智能阶段，后面两个阶段的很多技术还没有突破，现在的很多讨论都是基于各种假设，很难清楚知道类人工智能和强人工智能未来应用的具体场景。因此，对于人工智能发展是否会一定违背上述的规范，还存在不同看法。

算法是指在数学（算学）和计算机科学之中，为任何有着明确定义的具体计算步骤的一个序列，常用于计算、数据处理（Data processing）和自动推理。[①]算法在人工智能的发展中扮演着核心的作用，类似于机器的大脑结构，同时算法的进步带来了也导致了算法歧视问题。算法歧视一方面挑战了公平、平等的价值观，特别是在涉及性别、种族、年龄、职业、国籍平等方面。另一方面，随着基于算法的自主决策系统在金融、法律、安全、就业、教育、消费等方面的应用越来越深入，算法歧视问题对公众的生活带来了很多影响。[②]

算法歧视主要有四类不同的情况：第一类是基于算法滥用导致的针对特定群体的歧视，如针对不同消费者的价格歧视；第二类是缺乏价值观念导致的算法歧视，如互联网企业通过用户的社交圈来评估信用等，如果你的朋友都是富人，你的信用就会高于朋友都是穷人的用户；第三类是由于数据质量缺陷导致的算法歧视，这通常是在数据样本不完善的情况下，自动标签系统（Auto-tagging System）未能对数据属性做出正确的识别。谷歌公司的图片软件曾错将黑人的照片标记为“大猩猩”。[③]Flickr 的自动标记系统亦曾错将黑人的照片标记为“猿猴”或者“动物”。[④]第四类是算法与人的互动中产生的歧视，在有监督的机器学习算法中，机器需要与用户进行大量互动，然后不断的完善。当用户输入了错误的价值观之后，机器的决策也会受到很大影响。2016 年 3 月 23 日，微软公司的人工智能聊天机器人 Tay 上线。出乎意料的是，Tay 一开始和网民聊天，就被“教坏”了，成为了一个集反犹太人、性别歧视、种族歧视等于一身的“不良少女”。于是，上线不到一天，Tay 就被微软公司紧急下线了。[⑤]

算法偏见既有开发者为了获取经济利益主动为之的情况，也有不可知的算法黑盒导致的情况。“与传统的机器学习不同，深度学习并不遵循数据输入、特征提取、特征学习、逻辑推理、预测的过程，而是由计算机直接从事物的原始特征

① Wikipedia contributors, "Algorithm," Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=Algorithm&oldid=830474312 (accessed March 15, 2018).
② Safiya Umoja Noble, “Algorithms of Oppression How Search Engines Reinforce Racism”, New York: New York University Press, p155.
③ BBC: “Google apologizes for Photos app's racist blunder”, July 1, 2015, http://www.bbc.com/news/technology-33347866
④ The Guardian, “Flickr faces complaints over 'offensive' auto-tagging for photos”, May 20, 2015, https://www.theguardian.com/technology/2015/may/20/flickr-complaints-offensive-auto-tagging-photos
⑤ Wikipedia contributors, "Tay (bot)," Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=Tay_(bot)&oldid=825577647 (accessed March 15, 2018).

出发，自动学习和生成高级的认知结果。在人工智能输入的数据和其输出的答案之间，存在着我们无法洞悉的隐层，这就是黑箱（black box）”。很多专家认为，黑箱问题无法解决，人类无法理解机器的逻辑。也有专家认为可以通过算法的优化来进行干预。但无论是通过算法本身的改进还是通过人类在算法做出决策后进行干预和纠偏，算法歧视问题都将是人工智能全球治理的一个重要议题。

针对算法歧视的治理同样需要政府、技术社群、私营部门构建相应的治理机制。马修·约瑟夫（Matthew Joseph）等人在其论文《罗尔斯式的公平之于机器学习》(Rawlsian Fairness for Machine Learning) 中基于罗尔斯的“公平的机会平等”（Fair Equality of Opportunity）理论，引入了“歧视指数”（ Discrimination Index）的概念，提出了如何设计“公平的”算法的方法。[①]在加州大学伯克利分校发布的《人工智能的系统挑战：一个伯克利的观点》（A Berkeley View of Systems Challenges for AI）中，这种关联性被称为“反事实问题”测试。在个人被拒绝贷款的例子中，人工智能系统必须回答如果“我不是女性，是不是就能批贷？”“如果我不是小企业主，是不是就能批贷？”这样的问题。因而数据使用者有义务建构出一套具有交互诊断分析能力的系统，通过检视输入数据和重现执行过程，来化解人们的质疑。[②]

算法偏见既是伦理问题，同时也是公共政策问题，需要建立更多的机制对算法偏见进行监督。社群、私营部门以及政府所做出的努力都在朝着逐步规范算法的方向而努力。IEEE 制定的《人工智能设计的伦理准则》中设计的五大原则，以及谷歌提出的“机会平等”都可以作为应对算法歧视的规范。但是算法歧视问题的解决并非一朝一夕，需要针对不同领域、不同情况的歧视问题构建相应的治理机制。

3．就业问题

人工智能作为一种通用型技术，它给就业市场带来的影响是根本性和全局性的。人工智能引发的失业问题不再是一国的公共政策问题，而上升到全球治理的议题，需要市民社会组织和私营部门发挥更加积极的作用。以往被认为是需要高学历、高智商的工作，也有可能会被人工智能取代，如人工智能已经在新闻、司法、医疗、科研、金融逐步开始应用；在信息化向智能化升级是移动互联网、大

① Mathew Joseph: “Rawlsian Fairness for Machine Learning”, https://www.researchgate.net/publication/309572952_Rawlsian_Fairness_for_Machine_Learning

② 许可：“人工智能的算法黑箱与数据正义”，载《金融时报中文网》，2018 年 3 月 7 日，http://www.ftchinese.com/story/001076567

数据、智能制造、云计算、物联网等新一代信息化和智能化相关技术与产业相互推进的结果，因此，失业问题一旦产生将有可能是突发地并会引发连锁反应。[①] 从全球范围内看，不同的国家应对失业问题的韧性不同，在同样的失业危机面前，发达国家和技术先进国家有更多的资源和能力来应对挑战，而发展中国家很可能由此诱发社会动荡，从而进入失败国家的行列。

因此，以多利益攸关方模式建立一种包括政府、市民社会和私营部门在内的多方参与机制应对人工智能引发的失业问题是必要而且有效的。所需要关注的治理议题包括如何应对智能化时代大规模就业岗位在多个行业同时被机器取代；如何快速的创造新的就业岗位；如何减少失业问题对社会的冲击等议题。如何增加社会和家庭应对失业冲击的韧性（resilience），在失业后依旧可以保持一定的生活水平；各方还可以与政府一道加强职业培训、继续教育和终身学习等方面学习，建立就业和再就业的培训体系。[②]鉴于此，更应该积极的发挥技术社群和私营部门的作用，构建相应的治理规范和机制，探索制定应对失业问题的最佳实践；对新的技术和应用场景可能引发的全球失业问题进行动态评估，提前预警；对技术的集中应用可能带来的连锁失业反应进行评估；甚至可以建立专门治理失业问题的国际组织和非政府组织，帮助各国政府更好的应对失业问题。

4．数据安全。数据是人工智能时代的石油和推动人工智能发展的动力。算法需要数据进行支撑，场景的开发和应用也需要大数据。人工智能时代数据安全问题主要包括隐私和国家安全两个方面。隐私本身是人权的一种，各国的法律都对公民隐私保护有明确的规定，二是个人信息的泄露会对个人的生活、工作、经济和安全带来负面影响，徐玉玉案就是最典型的的代表。另一方面，从国家安全的角度来看，大数据涉及到的是国家的金融、经济风险，国防和国家安全。数据一旦被恶意使用，就会对国家造成重大损失，“斯诺登事件”所揭露出的大规模数据监听就是大数据与国家安全领域的里程碑事件。

数据安全不仅是人工智能面临的挑战，任何涉及到使用数据的企业在使用个人信息或者关键基础设施产生的重要信息时都存在同样的问题。数据安全是政府关注的焦点问题，关于数据安全的治理是各国公共政策的一部分，欧盟出台了通用数据保护规则，中国也制定了个人信息保护规范，并开始探讨就个人信息安全

① William A. Carter, “A National Machine Intelligence Strategy for the United States”, Washington DC: CSIS, p7.
② Rodney Brooks, "The Seven Deadly Sins of AI Predictions," *Technology Review*, October 6, 2017.

进行立法。各国加强对个人信息安全的保护产生的一个影响就是数据的跨境问题。各国基本都是从个人信息保护和国家安全角度限制数据的出境。企业要进行数据出境需要满足所在国家的相关规定，或者国家在双边层面制定相应的规则如美欧隐私盾协议，欧盟的白名单制度，区域性组织，GBPR。 在涉及到人工智能与数据安全时，欧盟规定需要将敏感数据排除在人工智能的自动化决策之外，根据《通用数据保护规则》第 9（1）条，“敏感数据”即有关种族、政治倾向、宗教信仰、监控、性生活、性取向的数据，或者唯一性识别自然人的基因数据、生物数据。由于这些数据一旦遭到泄露、修改或不当利用就会对个人产生不利影响。因此，欧盟一律禁止自动化处理，即使当事人同意也不行，除非法律例外。①

在人工智能的伦理、算法、失业和数据安全的治理议题上上，以技术社群为主导的多利益攸关方模式正在发挥标准和规范制定的作用，会对克服人工智能发展中面临的伦理、歧视、失业和数据安全问题产生重要影响。同时也要注意到，多利益攸关方也存在一定不足之处。特别是技术社群的专家主要是来自西方国家，难以避免的先入为主的把很多价值观带到治理机制中。如在任何相关和不相关的议题中反复强调所谓的“人权”、“言论自由”，忽视发展中国家重视的能力建设和发展问题。因此，人工智能在伦理层面的治理应当在互联网治理的多利益攸关方基础上，更加关注解决平等参与和公平参与问题，要拿出切实的政策鼓励发展中国家的参与，平等采纳发展中国家的主张。

三、多边模式与人工智能全球治理

体制复合体理论的另一种治理模式是主权国家视角下的多边治理模式。人工智能已经体现出在推动经济范式转变、生产方式改变和社会结构变革方面具有的强大动能，并将会逐步对当前的国际体系产生深层次影响。在国际体系层面，人工智能的全球治理更多涉及到国家行为体，更适用主权国家为主体的多边治理机制，强调各国平等和有区分的责任，联合国等国际政府间组织是制定规则的主要场所。治理的目标是通过各种形式的国际机制、国际法来应对人工智能对国际经济、安全和政治领域造成的冲击，规范国家的行为，推动各国政府积极参人工智能的全球治理事务。多边治理的主要议题包括人工智能是否会进一步扩大数字鸿沟，加剧全球经济社会发展的不平衡，并由此导致技术落后国家被排除在全球经济体系之外；人工智能在军事领域的应用，特别是致命性自主武器的使用是否会

① 许可：“人工智能的算法黑箱与数据正义”，载《金融时报中文网》。

对国际安全体系造成冲击；大国主导的国际规则博弈是否会对当前主权国际平等参与治理的国际政治体系造成冲击等。

1．人工智能对全球经济体系的冲击及治理

人工智能对技术、人才和资源的极高要求决定了其发展将会在不同国家之间产生新的数字鸿沟，进一步扩大国家之间在经济竞争力领域的差距，从而影响到当前国际经济体系稳定。人工智能对传统产业的替代会在全球范围内导致大量的失业问题，先发国家和技术大国，可以通过发展人工智能的收益来弥补传统产业的损失，通过将部分收益用于转移支付和社会保障系统来维护社会的基本稳定，并推动社会进步。对于一些未能跟上这一波人工智能发展浪潮的国家而言，大量的就业会转移到国外，国内的产业也会被更先进和有竞争力的外资所接管，最后导致政府缺乏足够的资源来应对经济和社会所面临的冲击，这或许将成为潜在的动荡根源。

上述趋势会进一步推动当前世界经济体系的变革。沃勒斯坦认为当前的世界经济体系是一种中心 - 半边缘 - 边缘的结构，发达国家居于中心地位，发展中国家和不发达国家依靠廉价的土地、资源和人力要素居于半边缘和边缘地位，并且形成了一种对中心国家的依附关系。[①]人工智能可能发生可能会加剧这种不平等的地位。从工业化、信息化到智能化的整个发展路径来看，技术和人才的集中度越来越高，发展中国家面临的形势越来越不利。发达国家通过人工智能在经济领域的应用减少了对发展中国家劳动密集型产业以及自然资源的需求，从而将发展中国家排除在全球产业链与供应链之外，促使发展中国家在世界经济中的地位更加边缘化，并有可能滑落到体系之外。极端的情况下，人工智能的发展可能会改变全球产业分工的格局，世界经济只有中心国家不再有边缘国家。经济体系变革的后果有可能是人工智能推动的全球经济体系变革导有可能产生一批新的失败国家，类似于索马里这样被全球化遗弃的国家，逐步沦落为战争、疾病、暴恐、宗教极端势力的温床，形成对国际体系的极大挑战。

多边治理模式强调国际社会和各国政府的责任，加强人工智能全球治理工作，应当关注人工智能领域新型数字鸿沟问题，解决不平衡的发展问题；另一方面还要积极利用人工智能来实现联合国可持续发展目标，解决当前国际社会面临的各种发展问题；最后从产业和公共政策领域创新角度，探索有效的国际合作机制。

① 参见沃勒斯特著，郭方等译：《现代世界体系》(第 1 卷)，社会科学文献出版社 2013 年第 1 版。

通过多方面的努力来避免人工智能对国际经济体系的冲击。①

第一，通过加大议程设置来提高发展中国家对人工智能发展的意识，并提供必要公共资源来帮助发展中国家制定相应的发展战略，从而构建全球均衡发展的人工智能体系，同时加强人工智能全球治理的议程设置，提高国际社会对于人工智能的关注程度。②人工智能的技术和应用正在不断的突破，但是目前只有发达国家和一些新兴大国在积极跟踪技术和应用的发展，并积极出台各种战略、政策举措。对大多数国家而言，尚未意识到人工智能所带来的机遇与挑战。联合国、二十国集团、亚太经合组织等全球性、区域性国际组织应当高度关注人工智能领域的发展，强调人工智能发展对于各国经济社会的正面作用，引导发展中国家和不发达国家关注人工智能领域的发展和应用，并且根据各自国情制定相应的发展规划。

第二，为发展中国家发展人工智能提供必要的资源。发展中国家在人工智能领域中技术、政策、法律等方面的人才匮乏，并且缺乏必要的公共资源支持，难以依靠自身的力量来培养人才和发展相关的技术和产业。国际社会应当通过各种形式，提供必要的资源，加大对发展中国家在人工智能人才领域的培养，提供必要的技术和设备，帮助其更好的跟上全球人工智能发展趋势。③国际电信联盟、世界银行等联合国机构可以尝试指定人工智能的专项援助，通过项目来帮助发展中国家获取经验、培养人才。

第三，发挥人工智能在应对全球性挑战中的作用，在解决贫困问题、减少疾病传播、降低灾害影响等方面发挥积极作用。国际电信联盟召开的“人工智能造福人类”峰会即是一种通过人工智能来更好的达成 2030 可持续发展目标的尝试。④人工智能在精确分析致贫原因，提供有针对性的脱贫解决方案上可以发挥重要作用；人工智能还可以应用于流行疾病的爆发趋势、传播特征分析，并可以在预防和阻断传染性疾病在全球和地区层面的传播上扮演重要角色。谷歌、IBM、推特都推出了基于人工智能预测流感等流行性疾病的人工智能系统，对于各国的医疗卫生组织应对疾病爆发提供了一定的帮助。⑤最后，人工智能在灾害防灾、救灾等领域已经开始发挥作用。如通过人工智能的信息

① Nicolas Miailhe, Cyrus Hodes, “Making the AI revolution work for everyone”, March 2017, http://ai-initiative.org/wp-content/uploads/2017/08/Making-the-AI-Revolution-work-for-everyone.-Report-to-OECD.-MARCH-2017.pdf
② ITU, “AI for Good Global Summit Report”, p29.
③ Nicolas Miailhe, Cyrus Hodes, “Making the AI revolution work for everyone”, March 2017,
④ ITU, “AI for Good Global Summit Report”, p15.
⑤ GOOGLE, “Artificial Intelligence and Machine Learning at Google”, https://ai.google/

整合处理工作，可以依据气象、灾害信息，结合地理、建筑物、人员分布状况，以最快速度对灾害进行评估，生成灾害地图，并协助政府和救灾人员最快速度进行有效救援。

第四，推动人工智能创新发展的国际合作，构建全球人工智能技术创新平台。推动人工智能在全球均衡发展离不开数据的开放、人才的流动和最佳实践的分享。一是建立对人工智能的数据开放平台。人工智能发展离不开大数据，特别是针对特定应用场景的发展，数据至关重要。因此，拥有各种类型数据的国际组织和国家政府，应当积极开放数据，为人工智能的技术创新创造良好的基础。二是促进人工智能人才的全球自由流动。各国政府应该意识到人才的交流和流动是推动人工智能技术和应用发展的重要基础。各国应当为人工智能的发展制定更好的人才落户政策。三是促进人工智能最佳实践在全球分享。人工智能的技术发展和应用具有全球性，其面临的问题也具有很大的普遍性，对于人工智能在造福社会方面的实践应当积极、快速的分享给国际社会。①

2．人工智能在军事领域的应用及治理

人工智能在军事领域的使用带来了三个层次的问题，包括致命性自主武器系统（LAWS）的发展本身的安全与责任问题；致命性自主武器的使用对战争形态的改变引起的战争门槛下降、成本下降、伤亡减小情况下，人工智能的大规模使用带来的国际法问题以及对国际安全体系的冲击；人工智能军备竞赛问题等。国际社会应当就这些议题尽快达成相应的国际规范和国际法，避免过度地发展人工智能军事武器，控制人工智能武器的安全和扩散，提高使用的门槛，避免人工智能武器的滥用对国际安全体系的挑战。②

致命性自主武器系统是一种以自我导向的方式运行，基于对世界的主动感知进行决策并直接或间接地对人类造成伤害或死亡的系统。完全自主的致命性自主武器系统还需要能够识别和选择目标，确定拟对目标施加的武力级别，并在特定时间和空间范围内对目标实施规定的武力。半自主、全自主系统包括杀伤人员地雷、反舰雷达、各种精确制导导弹、鱼雷、巡航导弹、反卫星武器、空中和海上无人机、无人机群以及网络蠕虫等。致命性自主武器的开发、使用面临着一些列

① Nicolas Miailhe, Cyrus Hodes, “Making the AI revolution work for everyone”, March 2017, http://ai-initiative.org/wp-content/uploads/2017/08/Making-the-AI-Revolution-work-for-everyone.-Report-to-OECD.-MARCH-2017.pdf

② Heather Roff: Meaningful Human Control or Appropriate Human Judgment? The Necessary Limits on Autonomous Weapons. Report. Global Security Initiative, Arizona State University. Geneva, 2016.

的安全挑战。2012 年，美国国防部发布了关于武器系统自主权的 3000.09 指令，从人为控制、合法使用等方面做出了一定的规则探索。要求半自主和全自主的武器系统的设计应允许指挥官和作战人员在使用武力方面做出适当的人为判断。为保证这一要求的实现，需要严格的设计、对设计的测试和评估、操作验证和测试，解决系统错误和故障的安全工程以及避免人为错误的认知工程；负责授权使用、指挥使用或操作自动化和半自主武器系统的人必须遵守战争法、适用的条约、武器系统的安全规则以及适用的交战规则（ROE）。①

以致命性自主武器为代表的人工智能在军事领域的应用还对战场的形态发生了巨大的改变。随着机器取代人走向战场，军人面临的战场条件会变得更安全，伤亡会随着致命性自主武器的使用而降低，武器精确性提高也会导致战争成本降低和杀伤效果提高。某种意义上，战争的内涵将会发生变化。如何定义携带武器的无人驾驶飞机入侵他国领空的行为？是侵略还是侦查？如同把进攻性的网络攻击视为一种行动（Cyber Operation）而不是网络作战（Cyber War），将网络进攻定义为低烈度的冲突。致命性自主武器的使用降低了使用武力的门槛，某种意义上也会鼓励更多国家开发和使用致命性自主武器，造成军备竞赛和武器扩散的问题。

人工智能引发的军备竞赛问题是专家组的一大重要分歧，技术强国与弱国之间存在截然不同的观点。技术弱国认为应当完全禁止致命性自主武器的开发使用，技术大国则持相反意见，认为开发致命性自主武器可以降低人员损伤，有利于打击恐怖主义和维护国家安全，并且很多系统已经在战场上进入实战。军事是推动技术进步的重要因素，互联网就是由美国军方所发明，人工智能的技术发展背后也有军事因素的强力推动。②因此，要完全禁止致命性自主武器的开发并不现实，人工智能的军备竞赛也非技术之福，特别是致命性自主武器的扩散会造成更为严重的后果。因此，从联合国层面制定相应的规范，并且促成大国之间在发展致命性自主武器上达成一定的军控条约是当务之急。③

联合国会经过 2014 年和 2015 年的两次非正式会议后，在 2016 年 12 月 16 日关于特定常规武器公约的会议上成立了致命性自主武器系统（LAWS）政府专家组（GGE），并指派印度担任主席国。该小组的任务是研究致命性自主武器领

① US DOD, Autonomy in Weapon Systems, No. 3000.09, Nov 21, 2012.
② US DOD, Office of Net Assessment. Summer Study - (Artificial) Intelligence: What questions should DoD be asking? Chair and Editor Matthew Daniels. 2016.
③ Miles Brundage, Shahar Avin, *et alia*., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, Cornell: arXiv Archive, Cornell University Library, February 20, 2018.

域的新兴技术，评估其对国际和平与安全的影响，并为其国际治理提出建议。[①] 联合国政府专家组应制定致命性自主武器系统在降低使用武力门槛、意外导致不应有的伤害、造成意外的升级螺旋式增长的问题。致命性自主武器系统治理应该清晰明确，并且在快速的技术变革中保持相关性。即使致命性自主武器系统的操作是合法的，联合国政府专家组也应考虑对国际法此前未曾预见的情况追加法律限制，并以道德或伦理为由尽量减少对平民的伤害。

无论是致命性自主武器本身的安全性、人工智能对战争形态的改变还是军备竞赛问题，最终都会对国际安全体系造成一定的冲击。人工智能条件下，低烈度的冲突是否会常态化？大国是否会无法遏制使用武器的冲动？武器的扩散是否会造成新的地区动荡和恐怖主义袭击等议题都是人工智能全球治理必须要关注和解决的问题。人工智能在军事领域的应用是技术强国谋求战略优势和国际安全体系稳定之间的一种妥协。需要从技术的安全可控、指挥人员的操作守则以及国际、国内法律等方面构建完整的治理体系，和避免技术强国之间的军备竞赛。以联合国专家组为代表的治理机制是国际社会近期努力的一个方面，还需要各国政府根据形势的发展，做出更多的努力来推动人工智能军事应用领域的治理，维护国际安全体系的稳定。

3．人工智能对国际政治体系的影响

以主权平等为基础、联合国宪章为法理以及联合国在处理国际事务中的权威为代表的国际政治秩序是在国际经济体系和安全体系不均衡、不平等的状态下维护整个国际体系稳定的基石。在人工智能时代国际政治体系面临着三重挑战：第一，作为国际政治体系基础的国际经济体系和国际安全体系的不平衡将会进一步加剧，变革速度加快，导致现有的政治体系不适应新的经济与安全需要，产生新的变革动力，国际政治体系的稳定性受到挑战；第二，以人工智能、网络安全为代表的新兴治理议题对于联合国机构的专业性与资源提出了新的要求。需要联合国发挥更加积极主动的作用，展现在人工智能领域的领导力，避免治理议题陷入困境；第三，发达国家以各种理由抵制、阻碍联合国的合法性和权威性，拉小圈子联盟，使得国际社会在治理问题上陷入分裂状态。[②]

---

① UNIDIR: “The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches”, Geneva: United Nations Institute for Disarmament Research, 2017.

② 参见鲁传颖：“新形势下如何进一步在联合国框架下加强国际网络安全治理”，载《中国信息安全》2018 年第 2 期，第 35 页。

人工智能全球治理应当从网络安全治理领域的挫折中吸取教训，避免陷入到不同阵营的分裂对抗中。网络安全治理中，发达国家动辄以各种理由抛开联合国，通过所谓的“观念一致国家联盟”（Like Minded Countries）来制定国际规则，不仅会造成国际社会的分裂，阻碍治理进程，最终也会影响到联合国的合法性和权威性，从而对国际政治体系造成挑战。如第五届信息安全政府专家组就是由于大多数国家不支持美国等西方国家提出的一些关于网络空间军事化的内容，而受到美国政府的抵制未能达成共识。随后，美国以其盟友体系为基础，组织所为“观念一致国家联盟”继续推动在小圈子范围内制定规则。长此以往，现有的国际政治体系将会被逐步弱化，使得发达国家和发展中国家分化为两大不同立场的阵营，并且会进一步加剧各国在经济和安全上的不平等、不平衡状况，造成国际政治体系的不稳定。

应对人工智能对国际政治体系的挑战，需要多方共同努力。发达国家和新兴大国作为主要的受益者应当承担起自身的责任，重视人工智能全球治理问题，从经济上增加对发展中国家人工智能领域的投入，帮助发展中国家提高能力；在人工智能军事领域的研发和使用上应有所克制，特别是在负面影响还未得到全面评估的情况下，应当谨慎使用致命性自主武器，避免人工智能武器的扩散。在政治上，不应动辄就绕开联合国拉小圈子，应尊重主权平等对于现有国际体系的重要作用，维护联合国的合法性。另一方面，发展中国家也应当积极有所作为。抓住人工智能发展的机遇，积极配置资源参与人工智能的全球治理工作。如沙特阿拉伯专门成立了人工智能部，加强政府在人工智能创新和投资领域的职能，营造好有利的人工智能发展环境。[①]积极从国际社会已有的成功经验和最佳实践中获取知识。通过国际社会的援助和发展中国家的努力，使得人工智能的发展和治理保持均衡的状态，从而更好的维护国际体系的和平、发展与稳定。

体制复合体理论视角下的人工智能全球治理包涵了技术安全和体系安全两个层级，具有伦理、算法、就业、数据安全和国际经济、安全、政治体系等多个治理议题，并且通过多利益攸关方和多方参与两种模式构建了 IEEE《人工智能设计的伦理准则》、国际电信联盟“人工智能造福人类”、“阿西洛马人工智能原则”、联合国“致命性自主武器政府专家组”等多个治理机制。除此之外，越来越多的网络空间治理机制开始将人工智能的全球治理纳入新的治理范畴，如信息社会世界峰会、互联网治理论坛、世界经济论坛等都把人工智能治理视为未来重

① Arabian Business, “UAE appoints first Minister for Artificial Intelligence” , 19 Oct 2017.

要的治理工作。在实践中，两者治理模式在议题、参与行为体和治理方式有各自的特点，因此在构建的治理机制的有效性上也有不同的表现。总体而言，多利益攸关方模式更有效率，但公平性不足。多方参与模式，注重公平、平等，但约束机制构建的因素多，不利于发挥机制的有效性。人工智能的全球治理与其发展阶段一样，还处于早期。但是事关人类社会的价值观念、秩序，并与国际体系的经济、安全、政治密切相关。应当根据体制复合体理论的指引，构建有效的治理标准、规范、政策、法律等治理机制，促进人工智能的良性发展，减少对于人类社会的价值观念和国际体系稳定的负面挑战。

（作者简介：鲁传颖，上海国际问题研究院全球治理研究所副研究员，博士，上海 200233，邮编；约翰·马勒里，麻省理工学院计算机与人工智能实验室高级研究员，博士，波士顿 02139）

收稿日期：2018 年月

（责任编辑：）

Artificial Intelligence Global Governance Process from Regime Complex Theory Perspective

Abstract: Regime complex theory consists multi-stakeholder and multi-lateral governance models, it has strong explanatory power in those complicated global governance issues like cyberspace governance. It also applied in complicated AI global governance. Multi-stakeholder and Multi-lateral can be applied in different issues. For instance, multi-stakeholder model more popular in algorithm discrimination, ethic issue, job replacement, and data security. Global economy, international security and political issues , on the contrast, more like to adopt multi-lateral model. Finally, through regime complex theory, to construct a AI global governance system consist set of standards, norm, rules and regime in different dimensions and domains.

hould be constructed through multi-party and multilateral governance models.

Key Words: AI Algorithm Discrimination Regime Complex Multi-stakeholder