BAAI

人工智能伦理、治理与可持续发展译丛

Translation Series on Artificial Intelligence Ethics, Governance, and Sustainable Development

克服人工智能伦理与治理的跨文化合作阻碍

Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance

北京智源人工智能研究院 人工智能伦理与安全研究中心
中国科学院自动化研究所 中英人工智能伦理与治理研究中心

Research Center for AI Ethics and Safety, Beijing Academy of Artificial Intelligence
China-UK Research Centre on AI Ethics and Governance, Institute of Automation, Chinese Academy of Sciences

译丛引言

人工智能伦理与治理关乎全球人工智能发展与创新的方向与未来。将人工智能作为使能技术推动人类、社会、生态及全球可持续发展是人类进行人工智能技术创新的共同愿景。在这个过程中，来自各个国家、政府间组织、国际组织的人工智能伦理与治理工作通过学术机构、产业、政府等以各种方式积极推动相关原则、政策、标准、法律的制定、技术与社会落地。虽然来自各个国家、组织的努力是在不同文化背景下建立的，但是文化的差异恰恰提供给我们思考问题的不同视角，和相互学习与借鉴的机会。如《论语》中有言“君子和而不同”。建立跨文化互信是全球和谐发展的基石。人工智能伦理、治理与可持续发展将是全球科技、社会领域的持续性重要议题。为此，北京智源人工智能研究院人工智能伦理与安全研究中心携手中国科学院自动化研究所中英人工智能伦理与治理研究中心等单位共同发起《人工智能伦理、治理与可持续发展译丛》，将人工智能伦理与治理、可持续发展领域的重要文献进行遴选，组织翻译，并介绍给全球读者。期待从跨文化、跨语言的交流中各自有所裨益，促进伴随技术发展的文化交流，推动全球人工智能与人类未来的和谐发展。

曾毅

北京智源人工智能研究院人工智能伦理与安全研究中心 主任

中国科学院自动化研究所中英人工智能伦理与治理研究中心 主任

The ethics and governance of Artificial Intelligence (AI) are essential for the direction and future on the development and innovation of global Artificial Intelligence. Using AI as an enabling technology to promote the sustainable development of humanity, society, ecology are the common vision for the global technology innovation of AI. In this process, the ethics and governance of AI from various countries, intergovernmental organizations and international organizations actively promote the development of relevant principles, policies, standards, laws, and their technology and society groundings through academic institutions, industries, governments, etc. Although the efforts of various countries and organizations are established in different cultural contexts, cultural differences provide us with different perspectives and opportunities to learn from each other. As the analects by Confucius say, "Be in harmony, yet be different". Building cross-cultural mutual trust is the foundation of global harmonious development. AI ethics, governance and sustainable development will continuously be an important topic in the advancement of science,

technology and society. Hence, the research center for AI Ethics and Safety, Beijing Academy of AI, together with the China-UK Research Centre for AI Ethics and Governance at the Institute of Automation, Chinese Academy of Sciences, jointly launched the "Translation Series on AI Ethics, Governance and Sustainable Development". Efforts will be put to selection and translation of important documents in the field of AI Ethics, Governance and Sustainable Development, and introduction of them to readers around the world. We look forward to cross-cultural and cross-language exchanges, and promoting the harmonious development of AI for the world, for humanity and for the future.

Yi Zeng

Director, Research Center for AI Ethics and Safety, Beijing Academy of Artificial Intelligence

Director, China-UK Research Centre for AI Ethics and Governance, Institute of Automation, Chinese Academy of Sciences

克服人工智能伦理与治理的跨文化合作阻碍[1]

Seán S. ÓhÉigeartaigh[1,2], Jess Whittlestone[1], Yang Liu[1,2], 曾毅 [2,3], 刘哲 [2,4]

1. 剑桥大学 Leverhulme 智能未来研究中心， 英国
2. 中英人工智能伦理与治理研究中心，中国
3. 北京智源人工智能研究院人工智能伦理与安全研究中心，中国
4. 北京大学哲学与人类未来研究中心，中国

摘要

实现人工智能对全球有益需在人工智能伦理标准与治理的诸多相关领域达成国际合作，同时保证文化观与文化理念的多样性。当前，由于文化间信任缺失及跨地域合作的现实挑战等因素，实现这一目标还有诸多阻碍。欧洲、北美地区与东亚地区的合作将在人工智能伦理与治理发展上产生重大影响，因此本文主要探讨以上地区在开展合作过程中所面临的障碍。

本文认为扩大基于人工智能伦理及治理的跨文化合作前景可观，且有因可循。人们往往低估不同文化或地区间误解的影响，但其实这些误解会瓦解跨文化信任，甚至导致根本分歧。然而，即便存在根本区别，也并不意味着跨文化合作无法有效推进，原因有二：（1）合作并不代表着需要在人工智能所有相关领域达成统一原则或标准；（2）尽管对某些概念性的价值观或原则存在分歧，各方也依旧可能在实际问题上达成一致。本文相信，在促进人工智能伦理以及治理方面的跨文化合作中，学术界扮演的作用至关重要，包括建立相互了解的深厚基础以及阐明何时有必要且有可能求同存异。本文对实际行动与方案也提出了相关建议，包括翻译核心文件、发表核心文件多语种版本，研究人员交流项目，以及推进与跨文化话题相关的学术研究发展。

关键词：人工智能；人工智能伦理；人工智能治理；跨文化合作

---

[1] 本文英文版 2020 年 5 月 15 日在线发表于 Philosophy and Technology 期刊，英文原文可通过以下地址访问：https://link.springer.com/article/10.1007/s13347-020-00402-x。本文中文译文由中国科学院自动化研究所中英人工智能伦理与治理研究中心(http://www.ai-ethics-and-governance.institute/)组织翻译。本文责任作者联络方式(Seán S. ÓhÉigeartaigh 348@cam.ac.uk)，本文两位国内作者请联络：曾毅(yi.zeng@ia.ac.cn), 刘哲(liu.zhe@pku.edu.cn)

1.引言

人工智能用途广泛，因此成为全球诸多国家的核心科技（Brynjolfsson and McAfee，2014）。人工智能技术，特别是机器学习技术的用途尤其广泛，已经成功应用于诸多领域，如语言翻译、科学研究、教育、物流、交通等等。从国家、国际、或全球层面来看，人工智能都显然对于经济、社会以及文化都具有深远影响。由此，各界对于人工智能伦理与人工智能治理的关注日益提升，人工智能伦理指的是，由于人工智能系统对于人类福祉与其他根深蒂固的价值观（如自主权、尊严）有深远影响，人类应当如何发展并应用人工智能系统等相关问题。人工智能治理的问题与实践结合更紧密，指通过基础规则、治理框架、或更“软性”的方式（比如行业规范、道德准则）确保社会中人工智能的应用合乎伦理[2]。

跨文化合作对于实现相关的伦理及治理举措至关重要。 此处“跨文化合作”具体指来自不同文化背景或国家的团体协力合作，确保人工智能技术的发展、应用、治理能够造福社会。 本文主要讨论跨国合作， 具体事例包括（但不限于）： 各国人工智能研究人员合作完成项目，采取安全可靠的方式发展人工智能系统；建立各类沟通渠道，确保着眼于人工智能伦理问题的国际讨论能够平等汲取多样的国际视角；邀请各国利益相关者参与制定实践准则、标准及法规。 跨文化合作应当推广，但这并不意味着参与各方一定要遵循同一套人工智能规范、标准、或法规，也不意味着方方面面至始至终需要订立国际协定。 跨文化合作需要解决的主要问题就在于，确定需要国际准则或协定进行规范的问题，或者确定需要突出文化差异的情况。

跨文化合作的重要性体现在以下几点： 第一，应当确保人工智能可以实现全球社会效益，共享某个地区的先进技术从而带动其他国家的发展，并确保社会共同进步且人工智能为各个地区带来一致的积极效益，要实现上述目标，合作必不可少。 第二，通过合作，世界各地的研究者可以共享专业知识、资源以及实践范例。 由此可以更早实现有益人工智能的应用，也可以合理处理潜在的伦理问题以及关键性安全问题。 第三，如果缺失相应合作，可能会导致国家或不同商业生态系统间的竞争压力，从而在安全、伦理、以及人工智能社会效益发展等方面的发展投入有所缺失。（Askell et al.，2019；Ying，2019）。 第四，国际合作的重要性也体现在诸多实际因素上，如为确

[2] 人工智能伦理与人工智能治理密切相关：治理方案通常为伦理原则的实际体现，而伦理框架也是发展相关政策法规的起点。 如今人工智能伦理更具实践性，衍生出诸多关于人工智能在社会中应用合乎伦理的原则与指导方针，于是很多“软性”治理方案也被纳入人工智能伦理。 因此，本章中的“人工智能伦理与治理”泛指判定伦理问题、归纳问题、以及在治理方法中加以应用的整个流程。

保跨越国家和地区界限的人工智能应用（比如人工智能在主要搜索引擎以及智能驾驶领域的应用）能够有效融入不同监管环境，并与其他地区实现科技互联（Cihon，2019）。

在东亚及欧洲相关领先学者[3]的见解基础上，本文对于如今人工智能伦理及治理的跨文化合作所面临的问题进行了细致分析，并提出可行解决方案。 本文着重关注欧洲、北美与东亚之间的合作， 因为这些区域如今在人工智能伦理及治理相关国际对话中的地位举足轻重（Jobin et al.，2019）。近期，上述地区不同国家间在人工智能伦理及治理领域的竞争及矛盾分析数量众多，尤其是对中国与美国的竞争及矛盾的分析。 本文的讨论以及建议适用于更广泛的人工智能相关国际合作，并希望可以以此促进更多区域间的合作。

人工智能系统愈渐完善，人工智能应用愈发有效、普遍，相应的风险也只会逐渐增高。若跨文化误解、信任缺失在学术研讨以及公开讨论中逐步固化，建立长期合作关系可能会更加困难。出此考虑，全球文化合作更应当尽早建立。因此，对于国际社会而言，就引导人工智能影响培养共识与深远合作关系是当务之急。

2. 北美、欧洲以及东亚在塑造全球人工智能对话中的角色

在企业和政府的资助下，北美、欧洲以及东亚在人工智能基础与应用研究及发展中的投入尤为雄厚（Benaich, Hogarth，2019；Haynes, Gbedemah，2019；Perrault et al.，2019）。诸多文章从竞争角度入手梳理了美国与中国人工智能的发展及应用进程（Simonite，2017；Allen, Husain，2017；Stewart，2017），但此类框架在规范与叙述的层面上都饱受争议（Cave, Ó hÉigeartaigh，2018）。

经深思熟虑，上述地区的学者与政策社群积极响应，从地区与全球两个层面规划人工智能伦理原则与治理建议的发展，并在政府相关举措中可见一斑。例如，欧盟人

[3] 2019 年 7 月，我们举办了跨文化信任研讨会（https://www.eastwest.ai/），与会者有来自英国大学（剑桥大学、巴斯大学）、中国（香港大学、北京大学、复旦大学、以及中国科学院）以及日本大学（庆应义塾大学）与协会（博古睿研究院中国中心）的代表。 这些代表都一直高度参与人工智能伦理及治理相关对话以及跨欧洲、北美、亚洲的合作项目。 本次研讨会主要讨论了学术界在构建人工智能跨文化信任当中的角色，本文也从中有所借鉴。 这并非意味着只有会议关注讨论涉及的地区才具有重要性，也不表明与会代表的观点以及专业知识能够完全代表他们所处地区的各类观点与专业知识。 我们认为本次研讨会推进了发展进程，建立起了重要沟通网络，提出了可行性观点及理论基础，为后续发展搭建了良好基石。该研讨会论遵循查塔姆宫守则。

工智能高级别专家组相关的活动在其第一、二版报告中发表了相应伦理准则、政策，并提出了投资建议[4]。英国政府致力于“与国际伙伴密切合作，就如何实现安全、符合伦理、创新的人工智能达成共识”（2018 年 5 月）。中国政府也采取了类似行动，致力于“积极参与人工智能全球治理，加强机器人异化和安全监管等人工智能重大国际共性问题研究，深化在人工智能法律法规、国际规则等方面的国际合作”（中华人民共和国国务院，2017）。 北美、欧洲以及东亚地区也各尽其能，推进各组织论坛中探讨国际人工智能标准的工作，其中包括国际标准化组织（ISO）[5]、电气电子工程师协会（IEEE）[6]以及经济合作与发展组织（OECD）[7]。

北美、欧洲及东亚地区的显著作用更体现于其对于多利益相关方、非政府组织的领导与组织，如人工智能合作组织（Partnership on AI）[8]、未来学会（Future Society）[9]、人工智能治理国际会议（International Congress for the Governance of AI）[10]及人工智能国际合作组织（Global Partnership on AI ）[11]（Hudson，2019）。诸多重要的人工智能伦理及治理会议都在上述地区举办，例如美国“有益人工智能”（Beneficial AI）系列会议[12]、北京智源人工智能研究院系列年会（Annual

[4] 链接：https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

[5] 在人工智能标准委员会的 28 个成员国中，有 22 个国家来自北美、欧洲以及东亚地区（https://www.iso.org/committee/6794475.html）。

[6] 例：参见电气电子工程师协会（IEEE）人工智能原则文件《人工智能设计的伦理准则》中的协会成员：https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf

[7] 例：参见经济合作与发展组织（OECD）人工智能专家组（AIGO）成员：https://www.oecd.org/going-digital/ai/oecd-aigo-membership-list.pdf

[8] https://www.partnershiponai.org/partners/

[9] https://thefuturesociety.org/our-team/

[10] https://icgai.org/icgai-members/

[11] 原国际人工智能事务委员会（International Panel on AI），由加拿大及法国（而非人工智能合作组织）联合发起（https://www.canada.ca/en/innovation-science-economic-development/news/2019/05/declaration-of-the-international-panel-on-artificial-intelligence.html）。

[12] https://futureoflife.org/beneficial-agi-2019/

Conferences of Beijing Academy of AI）[13]、北京论坛（Beijing Forum）[14]、美国人工智能、伦理以及社会会议（ACM Conference on Artificial Intelligence, Ethics, and Society conference）[15]、先进机器学习会议衍生的治理与伦理研讨会等等。

本文考虑了以下几种情况：

1. 北美、欧洲以及东亚在科技领域的领导地位；
2. 北美、欧洲以及东亚在推进全球伦理以及治理对话上的突出贡献；以及
3. 从竞争角度入手分析发展进程，整合出的深层矛盾以及关于伦理与治理根本问题上的争议。

因此，本文关注了阻碍上述地区与文化实现知识交流及合作的因素。 由于人工智能技术的影响范围应置于全球，需考量各国、各文化扮演的角色，本文并未对人工智能伦理及治理的跨文化合作进行全方位的分析。 对于后续研究而言，重点在于关注科技大国与科技进口国间日益凸显的权力与影响力不平等（Lee，2017），以及科技大国在全球治理与伦理对话中加入科技进口国、增强其自主权的责任。

3.人工智能跨文化合作的阻碍

虽然已有众多关注人工智能伦理及治理问题的国际联盟，但就立足于指导人工智能发展与应用的规范、原则以及治理框架、真正实现跨文化合作，仍有诸多阻碍。

不同地区、文化之间的信任缺失是人工智能伦理及治理实现国际合作的最大障碍。目前，中美学者、技术专家以及决策者之间尤其缺乏信任[16]，原因如下：

（1）近年来，**两个大国间日积月累的政治矛盾日益深化**，如今导致人工智能发展竞争成为“东方”及“西方”国家间的竞赛[17]。

---

[13] https://mp.weixin.qq.com/s/tAGOoqqA6ods9uaigWE7uA

[14] http://newsen.pku.edu.cn/news_events/news/global/9133.htm

[15] https://www.aies-conference.com/2020/

[16] 这种信任缺失的情况尤其可见于七月份的研讨会。 多位参与者表示了解或参加过讨论中国人工智能发展进程地缘政治影响的研讨会，但这些研讨会将中国代表排除在外，还有将（从西方角度出发）“针对中国采取的措施”列为重点的相关活动。

[17] Ess（2005）对“东方”和“西方”两个术语提出质疑，表明这并非刻画国家或文化多样性的准确术语，而是殖民主义的产物。 然而，由于缺少合适的著述描述本文提到的广义文化差异，各类文献中仍大量运用这两个术语。因此，虽然深知两词具有局限性，本文仍将采用这两个术语。

（2）**由于两地区推崇的哲学传统不同**，在数据保密等关键问题上，相关工作认为“西方”与“东方”文化的价值观差异巨大且不可调和（Larson，2018；Horowitz et al.，2018；Houser，2018）。

近期科技以及政治的发展也加剧了这种信任缺失的情况。这包括对美国科技巨头在公众以及政治影响的担忧（Ochigame，2019），对中国社会信用体系的看法与反响（Chorzempa et al.，2018；Song，2019），对人工智能科技应用在争议性领域的担忧，其中广泛讨论且广受争议的案例包括利用人工智能进行入境管制（Whittaker et al.，2018），美国的犯罪风险评估（Campalo et al.，2017），以及中国对维吾尔族穆斯林群体的跟踪（Mozur，2019）。 美国政界及国防领导人的敌对性言论加剧此类矛盾。 近期媒体报道谈及了在人工智能领域成为“威胁”的意图[18]，以及更广泛层面上，由于担心中国科技进程会对美国全球领导地位造成威胁，将中国视为敌对性质的“对手” 的评论（2018 年 6 月）[19]。 若文化间不信任持续加剧，则可能严重削弱人工智能发展及治理全球合作的机会。

此外，如今尚不明确已有跨文化合作与联盟得以发展到何种程度，能否有力规范美国、中国以及两国大型跨国公司等在全球发挥重要作用的行为体。 即使多方利益相关者组织能够在人工智能伦理框架的原则上达成一致，要落实这些原则，实际上也难以直接牵制对人工智能发展与治理起主导作用方的行为。

全球合作亟待推进，而各方因文化、地理因素各异，实施办法也有各自的敏感需求，如何在二者间寻求平衡也是实现有效合作的另一挑战（Hagerty，Rubinov，2019）。我们需要力求避免某个或者几个大国将本国的价值观强加在其他国家或地区的做法（Acharya，2019）。在某些特定领域（比如利用人工智能支持医疗服务），不同文化对于不同利益平衡的理解也千差万别（Feldman et al.，1999）。以区域为单位落实标准以及治理方案较为困难，但十分必要。 应用在不同文化区域的人工智能系统也会产生不同的效应，因此需要匹配不同的治理方案（Hagerty、Rubinov，2019）。

---

[18] “很多人担心人工智能带来威胁，而我们想要成为威胁。” 美国国防部副部长 Patrick Shanahan 在写给国防部部员的邮件中如是说道（Houser，2018）。

[19] 2019 年华盛顿举办的安全论坛引用了美国国务院政策规划主任 Kiron Skinner 谈及中国的言论：“此次斗争的对手与美国拥有截然不同的文明与意识形态，这是前所未有的”，“在非白种人中，我们头一次碰到这样势均力敌的对手”（Gehrke，2019）。

在人工智能发展与治理的某些方面，合作是至关重要的。 例如，人工智能在军事领域的应用（如自动索敌与自动攻击）可能会触及人权与国际人道主义法的底线（Asaro，2012）。 此外，如果在战场实现自动化信息搜集、决策制定以及反馈，可能反而事与愿违，使冲突不断升级。因为事件发生频率变快，也需要反应速度加快，而人类难以作出相应的有效判断或实行监管（Altmann，2019）。 在这两种情况下，追求军事科技进步的国家可从人工智能获益，但是如果缺乏国际协定或标准，其整体影响很可能极不稳定20。 对于人工智能科技在单一地区研发、其他地区加以应用或部署的各类情况而言，国际协定至关重要。 因此跨文化合作的一项关键任务就在于确认国际协定对于哪些地区来说尤为重要，并将这些地区与其他无需特定方式约束的地区加以区分。

两国合作还面临诸多现实问题，包括语言障碍、地域距离、移民限制等，这些问题限制了不同文化及研究群体进行交流与合作。此外，虽然科学无国界，但是大多科学期刊却仅有英文版本。

4.克服合作阻碍

实现人工智能伦理与治理的跨文化合作困难重重，但本文认为仍有推进进程的可行方案，且当下并不需要解决更深刻的问题，例如无需在所有基础伦理与哲学问题上寻求不同文化间的共识，也无需当即解决国家间数十年以来的政治矛盾。

建立深度相互理解，包括围绕分歧达成共识

对于人工智能伦理与治理的未来发展而言，国家间的不信任是一个严峻问题。本文所指的不信任一定程度上由相互间的误解与错误认知引起。因此，为建立更牢固的跨文化信任，首先要建立正确认知、消除误解，增进不同文化与国家之间的互相理解。

显然，东西方在人工智能领域上建立信任的主要阻碍在于这些地区的基本价值观存在本质差异。因此，双方对于如何发展、应用以及治理人工智能的伦理考量存在不同理解，有时甚至相互冲突。 虽然不同文化间确实存在价值观差异，但是有关这些差异如何体现的论断通常只是尚待检验的理论，也存在根深蒂固的成见，缺乏实践证据（Whittlestone et al.，2019）。认为“东方”和“西方”的伦理习俗有本质冲突只

[20] 若缺乏国际网络安全实践与准则的共识，受人工智能密切影响的（Brundage et al.，2018）数字安全领域也同样可能面临更大挑战（联合国大会， 2015）。

是对双方关系进行一概而论，具体而言，还需研究双方地区内部诸多不同的哲学传统，例如中国、日本、韩国在相关哲学观点上也存有重大差异（Gal，2019），并且“西方”哲学价值观和理念随着时间的推移也有所改变（Russell，1945）。概言之，双方的伦理及文化价值观都在不断发展，这些变化在《世界价值调查》[21]以及《亚洲晴雨表调查》[22]等项目中都有所体现。

人们通常认为，不同地区伦理、文化传统的差异构成了不同治理方法的基础。例如，隐私问题通常被视为东西方价值观重大不同之处，因此人们认为，相较于美国和欧洲而言，中国对于数据隐私的管控相对宽松。但是这些论断十分空泛，没有足够的论证或者实证分析支撑（Ess，2005；Lü Yao-Huai，2005），因此造成了双方的重大误解。首先，美国与欧洲在隐私概念（Szeghalmi，2015）和相关法规（McCallister，2018）上也有很大的差异。而中国对于西方社会的理解往往忽略了这些内部差异，过于关注美国的特点[23]。其次，西方对于中国数据保密问题的理解并不与时俱进。早在2005年，吕耀怀（2005）就指出，中国信息伦理文献没有美国成熟，但是受到西方学者的强烈影响，其发展进程也日新月异。诸多中国学者、决策者发表了相关的学术论文及报告，强调了数据隐私在人工智能伦理与治理发展中的重要性（北京智源人工智能研究院，2019；Fu，2019；Zeng, Lu, Huangfu，2018；Ding，2018b）。中国相关管控行动开始提出保护个人数据隐私的原则，中国政府禁用了100个违反个人数据隐私标准的应用程序，并要求数十个应用实行整改，调整数据收集及储存方式[24]。这并非说明这些国家在数据隐私价值观、准则及管控发行不存在重大差异，而是为了表明我们对于这些差异的认知过于笼统也不够了解。

观念差异也体现在对于中国社会信用体系（SCS）的理解上。西方媒体、政策界以及学者对于该体系关注度极高，并视其为中国政府奥威尔式社会管控的实例（Botsman，2017；Pence，2018），代表与西方世界文化和政府截然不同的价值观（Clover，2016）。但是中国与西方都有大量资料表明这是对该体系的误解。许多学

[21] http://www.worldvaluessurvey.org/wvs.jsp

[22] http://www.asianbarometer.org/

[23] 诸多中国学者在前述7月研讨会中提到此种误解。

[24] 2019年11月，中国公安部禁用了100个不满足个人数据隐私标准的应用程序；在2019年调查了683个应用（国家网信办 2019）。2019年11月，中国工业和信息化部公示了41个应用程序，它们必须在2019 年年底前进行整改以达到数据法规需要整改要求（中华人民共和国工业和信息化部，2019）。2018年7月，中国山东报道一起涉及11家侵犯个人信息公司的案件（Ding，2018c）。

者指出，人们通常误认为中国社会信用体系是统一衡量中国 14 亿公民的唯一平台，而其实是一个提供不同角度的信用解读的个人网络平台，主要来自金融机构的社会信用评分（而非基于大数据的综合评价）（Mistreanu，2019；Sithigh 、Siems。2019）。Song（2019）指出，社会信用体系提供的衡量标准多用于解决诈骗以及当地政府的腐败问题。 Chorzempa 等人（2018）也提出，“诸多社会信用的核心要素，例如黑名单、广泛监控等，在美国这样的民主国家中已经存在。”中国社会信用系统日益发展壮大，其当下以及未来具体实施还仍需得到重视。若能更明确理解该体系的工作原理、应用方式、对中国公民的影响，则能够建设性推进伦理与治理等问题的相关讨论。

由于美国、欧洲、中国等地长久以来缺乏共有知识和文化语境，地区间存在误解也就不足为奇。因此我们应当谨防因操之过急而产生无法调和的根本性分歧。目前双方都存在误解。例如，舆论调查数据的分析表明，美国和中国民众对于对方国家的特质和特性存在诸多误解（Johnston, Shen，2015）。如前所述，报道中也显示出中国也常常将西方社会的多样性简化为单一的美国生活模式。同时，美国和欧洲长久以来对于中国的了解也不够全面（Chen , Hu，2019），屡屡未能预测中国的自由化（或自由化缺失）或经济增长周期（《经济学人》，2018；Cowen，2019；Liu，2019）。语言障碍也是西方国家了解中国人工智能发展、伦理以及治理进程的一大障碍（Zhang，2017；Ding，2019）。Andrew Ng 在 2017 年《大西洋月刊》的一次采访中指出：“语言问题造成了一种不对称：中国研究学者通常掌握英语，能充分利用所有的英文文献。 但是对于英语为母语的人群而言，想要接触中国人工智能团体则难上加难”（Zhang，2017）。例如，腾讯发布的人工智能策略著作（Tencent Research Institute et al.， 2017）就涉及到关于伦理、治理以及社会影响的深度分析，但却鲜见英文报道提及（Ding，2018a）。甚至在诸如中国在人工智能研发的公共投资水平这类实证问题上，美国广泛报道的数据也可能出现大范围不准确的问题（Acharya、Arnold，2019）

近期发布的《人工智能北京共识》（北京智源人工智能研究院 ，2019）以及其他世界范围内推进的类似原则（Cowls, Floridi，2018）在核心挑战方面的论述有大量重合之处（Zeng, Lu, Huangfu，2018；Jobin, Ienca, Vayena，2019）。《人工智能北京共识》明确提到了其他文件涉及到的核心概念与价值观，包括人工智能应当“造福全人类”，尊重“人类的隐私、尊严、自由、能动性、权利”；做到“尽可能公正，减少系统中的歧视与偏见；提高系统透明性，增强系统可解释度、可预测性”。

此外，《人工智能北京共识》和《新一代人工智能治理原则》均提出人工智能发展需要公开及合作，后者尤其鼓励“跨学科、跨领域、跨地区、跨国界的交流合作”（Laskai, Webster，2019）。然而在实践中，不同文化背景的国家对于同样的准则会有不同的解读，重视程度也不同（Whittlestone et al.，2019），这可能是造成误解的更深层原因。 例如，我们不能直接认为与“东方”文化相比，“西方”文化更加注重隐私。相反，我们应当细致入微地研究，在隐私与安全等其他重要价值冲突时，不同地区如何进行取舍（Capurro，2005）。同样，虽然诸多文化重视人类能动性，但是我们也应当探究在不同背景下这种价值观背后传递的深层寓意与哲学理念（Yunping，2002）。

长久以来，国家间的误解已然根深蒂固，为了建立更加牢固的跨文化合作，我们首先需要认清与人工智能伦理联系紧密的误解，在争议集中点以及最能影响治理方案的争议达成相互理解。 在此过程中，需要明确伦理上的分歧，而不是治理方面的分歧。因为在某些情况下，虽然不同群体各具伦理观点，但也仍会在某些治理准则上达成一致。这一点我们会在后文进行讨论。这样做也利于区分与人工智能直接相关的误解（比如其他国家的科技投资或者数据保护法），以及化解更广泛的社会、政治、哲学等与人工智能间接相关的误解，毕竟不同的问题需要不同的对策。

强调误解的重要性并非意味着人工智能伦理及治理相关的所有跨文化矛盾本质上都基于误解。在个人、社会、国家关系，民用部门、私营部门、军事部门融合的程度和特质，以及社会政策相关的各类具体问题上，根深蒂固的分歧一直难以更改。但是，若从一开始便着眼于减少误解，将有助于更清楚地定位这些根本差异，同时找到能够达成充分共识并开展有效合作的环境。 这是解决人工智能伦理及治理跨文化合作挑战的首要任务。

求同存异建立合作途径

人工智能伦理、治理以及更广泛社会问题上的分歧无法根除，但这并不影响达成共识及合作。如前所述，人工智能伦理及治理面临的关键挑战是明确准则、标准以及制度达成跨文化共识尤为重要的领域，这些领域还需容纳或鼓励不同的理解以及方法。这一挑战本身就需要通过跨文化合作解决。该挑战所依据的信息来源于不同文化背景下对人工智能影响的理解，以及不同群体的需求与愿望。《新一代人工智能治理原则——发展负责任的人工智能》中便阐述了上述对策，并提出： “开展国际对话与合作，

在充分尊重各国人工智能治理原则和实践的前提下，推动形成具有广泛共识的国际人工智能治理框架和标准规范”（Laskai, Webster 2019）。

地域以及文化在抽象层面的伦理认知以及高层次原则的差异未必会阻碍在规范与治理的具体方面达成共识。如果没有在根本伦理问题上达成共识就无法签订实质性协议，《禁止核武器条约》等许多重要的国际协议也不可能实现。法学领域的“未完全理论化协议”（Sunstein， 1995）指出，即便人们的根本观点或抽象观点持不同意见，但对于如何解决具体问题往往会达成一致意见。这一点是执法的关键所在，对于广义上的多元社会也一样重要。多位学者在关于跨文化信息伦理的论述中提到旨在达成“重叠共识”（Rawls ，1993）的相关概念，即不同群体或者文化体支持同一套规范或实践指导方针的出发点可能不尽相同（Taylor，1996； Søraker，2006；Hongladaro，2016）。例如，Taylor（1996）探讨了如何在不同的文化传统中引入国际公认的人权规范。虽然西方哲学与佛教等其他哲学体系在人类主体的重要性及其在宇宙中的独特地位等本质问题上观点相异，但是两种哲学体系最终都揭示了同样的人权规范。

Wong（2009）批判就跨文化信息伦理求同存异达成共识并不现实，认为这样制定的规范可能会过于“薄弱”，缺乏全面的规范内容。Søraker（2006）针对信息伦理的“实用”方法也提到一种类似的反对意见，即由于这样的共识并非充分基于实质规范内容，可能会较为脆弱。然而，从 Søraker 对这些反对意见的回应可见，“重叠共识”旨在达成一致的规范以及实践指导方针，这些共识由诸多哲学及规范观点产生、支撑，因此**更为牢固**。但应当与此类情况明确区分的是，某一种文化企图通过实用论据将自己的价值观凌驾在他人之上，或是数个文化体达成共识，但出于某些原因没有形成规范内容。本文在后一种情况上认同 Wong 的观点，即认为这样的情况值得担忧。Taylor 提出的案例表明人权受诸多哲学观点支撑，可以体现此种重叠共识几经证实，存在合理性。

本文认为，与就共同认同的根本价值观达成国际共识相比，更为重要的是探索在规范及实践性指导方针方面存在重叠共识的领域，从而确保人工智能对人类有益。这也是近期诸多提案所遵从的目标[25]。对高层面伦理原则达成共识并不意味着这些原则充

[25] 例如，Florid 等人于 2018 年提出了“统一框架”；Awad 等人于 2018 年探讨“为机器伦理发展全球以及社会接受的原则“；Jobin、Ienca 与 Vayena 于 2019 年调查了关于构建符合具有伦理的人工智能的“全球协议”。

分正当（Benjamin，1995）。若要树立牢固可靠的人工智能规范、标准以及法规，最佳方式即为提出不同价值体系都推崇的共识。

在实践中落实原则

理论上可以促进互相理解、达成治理方法共识之处也可能遭到反对，反对者认为在实践中影响人工智能的发展及应用比较困难，因为要达成这样的共识必须影响大国与实力雄厚的公司的行为，而这些国家或公司的合作意愿并不强烈。

国家、企业与其他主体间复杂的权力变动与人工智能伦理及治理所涉及的问题也息息相关，但是本文不在此做深度讨论（然而这一现象值得深入研究），仅简要解释这层障碍为何不会影响本文论点。历史先例表明，虽然利用公众以及学术团体的影响力促使权力主体解决全球性重要问题较为困难，但是仍然可行。实证表明，大范围、跨文化的“专家组”（比如某个领域的专家网络）可以促进国际政策方面的有效合作（Haas，1992）。例如，军备控制专家组在冷战期间推进对于核武器控制问题的国际共识，帮助美国与前苏联建立合作关系（Adler，1992），还有生态学专家组成功协调了国家政策，保护了平流层中的臭氧层（Haas，1992）。

在人工智能领域，员工的积极行动、国际学术研究与各项运动早已共同影响到了大型企业、国家所做出的承诺，其中在军事领域应用人工智能的案例中这样的影响最为显著。国际学术界及民间团体的专家发起诸多声势浩大的运动，表达对于在战争中应用人工智能的担忧，其中包括国际机器人武器控制委员会（ICRAC）以及禁止杀伤性机器人运动。这些运动推进了联合国特定常规武器公约会议（CCW，该会议就禁止集束弹药、激光致盲武器以及地雷进行协商）对致命性自主武器（LAWs）的讨论（Belfield，2020）。90 个国家对致命性自主武器表达立场（多数是在联合国特定常规武器公约会议上提出的），28 个国家同意禁止致命性自主武器的使用[26]。2018 年，4000 多位谷歌员工签署抗议请愿书，部分员工辞职，以此抗议谷歌参与五角大楼 Maven 人工智能项目。这个军事项目旨在探索利用人工智能实现录像分析（Conger，2018）。部分来自美国、欧洲、日本、中国、韩国等地的学者也组成运动组织，公开

[26] 中国支持禁止在战场上使用全自主武器，但不反对全自主武器的研发。美国、俄罗斯、英国、以色列及法国则反对这一禁令（Kania，2018）。

发表文章以及公开信，支持谷歌请愿员工（ICRAC 2018）[27]。谷歌随即宣布不会与 Maven 项目续签合同，也将退出美国国防部 100 亿美元云计算合同的竞标（Belfield, 2020）。

在更广义的层面上，国际学术界及民间团体的努力有助于规范原则，为日后制定更具约束力的法规提供了坚实基础。例如，欧盟委员会发布了《人工智能白皮书——欧洲追求卓越和信任的方法》（欧盟委员会，2020），提出“未来欧盟监管框架的政策选择，将决定何种法律要求适用于相关主体，尤其侧重于高风险应用”（欧盟委员会，2020b）。《白皮书》受到了欧盟人工智能高级别专家组（由来自欧洲学术界、行业及民间团体的 52 名专家组成）工作成果的极大影响[28]。无独有偶，与美国学术界、工业以及政府利益相关方经过历时 15 个月的讨论后，美国国防部正式提出了人工智能的伦理准则（美国国防部，2020），并雇佣专员具体执行工作（Barnett 2020）。虽然在这两个案例中咨询的团体都限制在某一区域，但是不同地区在制定准则上的一致性以及重叠度表明，与更大范围知识群体的跨区域交流实质上提供了诸多想法和建议。这表明，跨文化合作和共识中提取的见解可以纳入区域和国家层面的法规框架。

5. 建议

学术界在支持人工智能伦理和治理的跨文化合作中发挥着重要作用，学术研究可以探索亟需合作的领域以及合作类型，制定方案可以克服更多合作中的实践障碍。本文提出了许多问题，需要不同领域的学术专家进行解答，问题包括：阻碍区域间合作的最大误解为何？何处需要基于人工智能伦理和治理的国际协议？在保留伦理问题分歧的同时，如何对具体的治理标准达成一致？。

以下建议指出了学术中心、研究机构和研究人员可以采取的若干步骤，从而促进基于人工智能伦理和治理的跨文化理解与合作。在这些领域中，部分优质项目已经进入正轨。然而，本文认为人工智能在新领域和新地区的应用速度值得关注，呼吁学术界加强重视在更广泛的伦理和治理研究项目中建立跨文化桥梁，并将跨文化专业知识纳入其中。

---

[27] 国际机器人武器控制委员公布了一封由上千名学者及研究者联名签署的公开信，禁止杀伤性机器人运动成员也通过发表公开文章以及公开信的方式联系企业领导，以此支持谷歌请愿员工，参见：https://www.stopkillerrobots.org/2019/01/rise-of-the-tech-workers/。

[28] 进程以及组成信息参见：https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence。

制定基于跨文化合作的人工智能伦理及治理研究议程。提升跨文化合作研究项目对于建立支持国际政策合作的国际研究社群而言至关重要（Haas，1992）。

符合此类合作要求的研究项目应进行比较性、前瞻性的实践，探索在不同文化中对于人工智能社会影响的积极愿景和突出担忧有何不同。这有助于形成国际化视野，明确人工智能发展中应当达成或避免的方面，指导关于伦理和治理框架的实际讨论。达成共识很可能可以力挽狂澜，安全与保障是全球人类文化的基础，因此制定协议、规避文化威胁成为更容易着手的切入点。然而，本文认为重视积极愿景也尤为重要。跨文化学者协力打造人类共享的美好未来，也是一种深入研究共同价值观中细微差别的良策。

与发展中国家的专家共同探讨人工智能对这些国家的持续、预期影响也具有重要意义。此类研究过程中应确保在当地专家指导下制定发展中国家的科技应用决策，这样才能将决定权落在当地群体手中（Hagerty 、Rubinov，2019）。从更切实的层面而言，国际研究组织可以实现高效合作，从而建立人工智能安全、保障和避免社会危害的研究、专业知识及数据集的国际共享框架。

不同地区和文化体研究人员之间的合作对于进一步推进跨文化合作本身也至关重要。本文中的讨论（尤其是第四节）指出了许多亟需进一步探索的研究领域，包括以下几点：

- 探索、明辨并改变人工智能伦理及治理的价值观、假设和优先级在不同文化中的既定差异，要做到[29]：
  - 分析基于不同哲学传统衍生的科技伦理的异同点，探索这对于人工智能的发展、应用、影响力及实际治理的影响；
  - 探索各个文化核心价值观差异论点的实证。例如，有项目发现、探索东西方文化在人工智能治理的既定价值观差异，包括数据隐私、国家与个人的角色以及对技术进步的态度等方面；
  - 理解不同社会中人工智能实际应用的优先权和限制等方面存在的地域差异，以及这些差异对人工智能研发的影响。

[29] 跨文化信息伦理领域已经存在一些探讨此问题与相关主题的精彩著述，如 Capurro（ 2005，2008）、Ess（2006）以及Hongladarom等（2009）。但本文更希望可以围绕这些主题达成跨文化研究合作，并在人工智能伦理的实际讨论和行动中给予更高的关注度。

- 深入分析，明确人工智能治理需要达成全球共识的领域，并将这些领域与包容或鼓励跨文化差异的领域区别开来；
- 从跨文化层面助力国际以及全球在需要人工智能标准关键领域制定人工智能标准；探索灵活治理模式，确保地区特色与国际标准相匹配；
- 探索模式及方法，包容根本或抽象伦理问题分歧，在具体案例、决策及治理标准上达成共识。借鉴其他领域的成功案例，在人工智能伦理及治理方面加以应用。

在这些领域中，部分优质项目已经进入正轨。然而，本文认为人工智能在新领域和新地区的应用速度值得关注，呼吁学术界加强重视在更广泛的伦理和治理研究项目中建立跨文化桥梁，并将跨文化专业知识纳入其中[30]。

翻译关键论文及报告。与其他科学领域相同，语言也是阻碍人工智能发展、治理及伦理的跨文化理解的主要实际障碍（Amano et al.，2016）。因此，若能将人工智能伦理与治理以及人工智能研究领域蓬勃发展的著述翻译成多语版本，则将创造极大价值。虽然很多亚洲人工智能领域的领先学者能够熟练运用英语，但是不具备此能力的学者仍占多数，掌握中文普通话或者日语的西方学者更是少之又少。

此外，上文论述的一些误解与其他地区学者理解、引用某地区主要文献的方式也密切相关。西方媒体将中国于 2017 年发布的《新一代人工智能发展规划》描述为中国为在人工智能经济与策略领域巩固全球主导地位的手段（Knight， 2017；Demchack， 2019）。但对中国而言，国家人工智能发展目标主要是出于中国经济与社会的发展需求（中国国务院 ， 2017），而不一定是国际竞争的优势（Ying ， 2019）。有些对于关键术语及要点的不当翻译造成了误解。例如，该《规划》的中文原文指出“中国力争到 2030 年成为世界主要人工智能创新中心”（Ying ， 2019）31。然而有些英文译文将此句翻译为中国即将成为“世界最主要的人工智能创新中心”（如 Webster et al.，2017）。随后，谷歌母公司 Alphabet 前任执行董事 Eric Schmidt 进而将这句话理解并表述为“到 2030 年，中国将会主宰人工智能产业。真的是这样吗？这反正是（中国）政府说的。”（Shead， 2017）。虽然从某种意义上来说这句话的措辞并未经历过大的变化，但是这句话毕竟承载着重要寓意。原文中的

[30] 本文的数位作者也参与了倡议计划，该计划旨在支持此类的英国与中国跨文化研究，参见：https://ai-ethics-and-governance.institute/%e5%85%b3%e4%ba%8e/。

[31] 在 2019 年 7 月的研讨会上，也有与会者提供了这份译文。

措辞较为婉转地表达了中国的领导能力和发展进程，而非国际霸主的地位。重要文件的多语种高质量翻译版本可以让学者辨别语言与上下文的细微差别，而这些信息很可能在转述中有所缺失。

提供文章、报道的多语种高质量翻译版本也体现出一种尊重及参与跨文化交流的意愿，可能会鼓励进一步合作。为学术以及政策材料提供高质量译文较为复杂而且耗时很久，但本文希望这方面的翻译能得到更有力的支持与认可。本领域有众多工作都值得赞扬，也正在蓬勃发展。比如，Jeff Ding 翻译了许多中国人工智能的主要文献（Ding，2019），中国海国图智研究院关于国际关系、科技及其他话题的著作发行了五个语言版本（http://www.intellisia.org/），Brian Tse 将英文版 OpenAI 的组织纲领等文献译成了中文32，New America 将中国工业和信息化部提出的《三年行动计划》译成了英文（Triolo et a.，2018）。

确保主要人工智能研究会议、伦理及治理会议在不同大洲轮流召开。为了提高人工智能发展、伦理及治理的全球参与度，本文建议围绕这些话题的主要会议及论坛应当在不同大洲轮流召开。这一做法可带来众多益处：避免来自一些地区的学者总因赶赴别处参加会议而投入大量时间和金钱；避免签证限制对于不同地区全球研究团体具有不同程度的影响；鼓励当地组织者积极参与，带动不愿出国参会的当地研究团体参与进来；鼓励组织者举行多语种会议而非单一语种会议。

此外还有其他积极措施。人工智能研究会议中的国际人工智能联合会议（IJCAI）于 2019 年在澳门举行，并于 2013 年在北京举行。这是头两个在中国举办的国际人工智能联合会议（此前该会议已在日本举办了两次）。国际机器学习大会（ICML）于 2014 年在北京举行，并将于 2021 年在首尔举行。国际表征学习大会（ICLR）将于 2020 年在埃塞俄比亚举行，这是首次在非洲举行的顶级重要机器学习会议。由于人工智能伦理及治理相较而言是新领域，所以明确关于这类话题召开的大型会议并不多。但若可以实现，还是要确保这些会议在不同大洲轮流举办，鼓励全球性参与，这一点尤为重要。例如，人工智能、伦理与社会会议是人工智能促进会（AAAI，原称美国人工智能学会）主办的会议，因此目前在美国举行。若要针对这些话题建立国际学术团体，则需要改变这样的现状。在中国，关注人工智能伦理及治理的会议迅速发展，其中包括北京智源大会。有一些会议也囊括了人工智能伦理及治理相关的话题（不过并非明确围绕此类话题），都能够进一步提升国际参与度。这些会议包括信息社会世界

[32] https://openai.com/charter/

峰会论坛（多在日内瓦举办）、互联网治理论坛（Internet Governance Forum）及权利大会（RightsCon）（后两者的举办地都非常多样，包括南美、印度以及非洲，但是均未在东亚举办过）33。

建立博士生及博士后联合和/或交换项目。在不同文化背景下的研究人员职业生涯初期鼓励其参与跨文化合作将有助于加强合作和相互理解，推进研究进程。如今不同国家间建立了诸多国际奖学金以及交换项目，最多见于中美之间的项目（例如，美国与中国达成的“知行中国”奖学金项目以及苏世民学者项目），此外还有英国和新加坡之间的项目（例如，英国伦敦大学国王学院与新加坡国立大学建立的哲学或英语联合博士生项目）。据悉，这些计划都并不是明确针对人工智能的。目前已知的人工智能伦理及治理项目仅有博古睿研究院学者计划及“天下学者”国际奖学金项目（Bauch，2019）34。建立更多类似的项目可以推动人工智能未来发展的国际合作，如今已经有许多现存模式及倡议值得我们借鉴。

从更宽泛的角度而言，本文支持人工智能合作组织提出的相关建议，包括建立特定签证办理途径，简化、加速签证程序并确保公正标准流程，从而支持人工智能与机器学习多学科专家的国际交流与合作，支持该组织就以上方面向政府提出建议。 这些建议的覆盖范围囊括从事或计划从事人工智能伦理和治理工作的专家（有时不属于“技术工作”范畴）（PAI Staff， 2019）。

6. 局限与未来方向

本文认为学术界在推进人工智能伦理及治理的跨文化合作中发挥着重要作用。无需消除所有基本价值观差异即可有效建立基于相互理解与合作的社群，并且最重要的是减少文化体间的误会与误解。

笔者们也意识到，本文的建议无法完全消除跨文化合作的障碍，未来还需要投入诸多努力以确保人工智能对全球有益。本文将简要点明指导这一目标的两个较为宽泛的未来研究方向。

细致分析跨文化合作的障碍，尤其是分析与权力分配、政治矛盾相关的内容。历史成功案例的分析表明，基于人工智能伦理及治理的跨文化举措可能在很大程度上影响实践中准则、标准和法规的制定，但在实施和执行层面仍然存在许多本文没有谈及

[33] 产业集群也可以参与通过国际研讨会及会议推进跨文化合作。本文推荐参考人工智能产业发展联盟的工作进展，以作为该领域近期倡议的范例，参见：http://www.aiiaorg.cn/。

[34] 长远议题智库创立的天下学者计划也包含人工智能安全及治理的主题。

的障碍。未来的研究可以回顾历史，探索过去成功影响全球性准则及制度的事件发生在哪些时期、如何发生。这样的研究将具有重大价值。

前文已指出，除了价值观差异与文化间的误解外，与权力关系及政治矛盾相关的各类问题很可能也对跨文化合作造成主要阻碍。关于这些问题如何阻碍人工智能伦理及治理的跨文化合作的研究将具有重大价值，有助于认识到学术项目在推进合作中的局限性，以及了解将这些方法与权力和政治变动分析相融合的措施。

考虑未来强大的人工智能系统在跨文化合作上遇到的挑战。人工智能的未来发展可能为全球合作带来更大规模的全新挑战，一些学者也提出，人工智能的未来发展可能会像工业革命或农业革命一样产生变革性的影响（Karnofsky，2016；Zhang, Dafoe，2019）。如果缺失严谨的全球方向把控，这种技术进步将导致技术领先的国家和落后的国家之间产生前所未有的鸿沟，体现于财富和权力的不平等之上。另一些学者的研究更加面向未来，提出开发具有超级人工智能系统的可能性（即超越人类智慧的通用智能；Bostrom， 2014）。如果不考虑后果和安全性，这种系统的强大功能可能会对人类文明构成灾难性的风险。有学者提出，避免灾难性后果的关键举措在于达成统一价值，设计符合人类价值观的系统（Russell，2019）。就共同价值观和原则达成全球共识、设计尊重多元价值观的系统已经成为了当务之急。

众多专家对这种技术的未来发展存在很大分歧，大多数人认为还需数十年。但是为了达成有效的协作成果，发展合作关系、达成必要协议也可能需要通过数十年的工作。这表明当今的合作计划不仅必须解决当前人工智能系统在伦理和治理方面的挑战，而且还应为预测和应对未来挑战奠定基础。

7. 结语

在全球社会实现人工智能对人类最大程度的有益有赖于跨领域、跨学科、跨国以及跨文化的深度合作。当前美国、欧洲与中国之间的紧张态势与不信任尤其对合作造成限制。误解更是加剧了这种不信任，社会与政治优先权的差异也受到过分强调或误解。

此外，如果依然认为这些地区能够在人工智能涉及到的所有重要伦理原则上达成一致，并将这些原则纳入规则和标准中，这样的观点未免太过轻率。即便如此，在全球范围内，这些地区在塑造人工智能伦理及治理方面也不应过分占据主导地位，人工智能影响到的所有全球社群都应当囊括其中并获得相应权力。努力实现“人工智能超级大国”间的相互理解也有两点益处：第一，可以减少全球人工智能治理领域内的主

要矛盾。第二，可以提供参考经验协助发展伦理及治理框架，从而支持多元价值观并达成适当共识。在运转良好的人工智能全球合作体系之中，挑战就在于开发新模式，既要包括由国际共识构建并支持的原则和标准，也要有研究和政策社群满足不同社会需求而提出的不同方法。

从实际的角度而言，国际人工智能研究和人工智能伦理及治理团体必须仔细思考组织的活动如何支持全球合作，或是如何促进对不同地区社会观点和需求的理解。广泛的跨文化研究合作与交流、不同地区举办的会议以及多语种刊物有助于化解合作障碍、化解不同观点和共同目标的理解障碍。如今政治风向愈加偏向孤立主义，研究人员跨越国家和文化鸿沟、致力在全球范围内实现有益的人工智能显得尤为重要。

致谢

在本文完成之际，笔者要感谢 2019 年 7 月 11 日至 12 日“建立对有益人工智能的信任机制”跨文化研讨会中的所有与会者。他们在会上提供了很多与本文相关的重要讨论。还要感谢 Emma Bates、Haydn Belfield、Martina Kunz、Amritha Jayanti、Luke Kemp、Onora O’Neill 以及两名匿名审稿人，各位对本文初稿给予了有益指点。

参考文献

Acharya, A. (2019). Why International Ethics Will Survive the Crisis of the Liberal International Order. *SAIS Review of International Affairs*, 39(1), 5-20.

Acharya, A., & Arnold, Z. (2019). Chinese Public AI R&D Spending: Provisional Findings. Centre for Security and Emerging Technologies Issue Brief. Available at: https://cset.georgetown.edu/wp-content/uploads/Chinese-Public-AI-RD-Spending-Provisional-Findings-2.pdf Accessed 23 December 2019

Adler, E. (1992). The emergence of cooperation: national epistemic communities and the international evolution of the idea of nuclear arms control. International organization, 46(1), 101-145.

Allen, J. R., Husain, A. (2017). The Next Space Race is Artificial Intelligence. *Foreign Policy*. Available at: https://foreignpolicy.com/2017/11/03/the-next-space-race-is-artificial-intelligence-and-america-is-losing-to-china/ Accessed 21 December 2019.

Altmann, J. (2019). Autonomous Weapon Systems–Dangers and Need for an International Prohibition. *In Joint German/Austrian Conference on Artificial Intelligence* (Künstliche Intelligenz) (pp. 1-17). Springer, Cham.

Amano, T., González-Varo, J. P., & Sutherland, W. J. (2016). Languages are still a major barrier to global science. *PLoS biology*, 14(12), e2000933.

Asaro, P. (2012). On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross*, 94(886), 687-709.

Askell, A., Brundage, M., & Hadfield, G. (2019). The Role of Cooperation in Responsible AI Development. arXiv preprint arXiv:1907.04534.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F. & Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729), 59.2018

Barnett, J. (2020). DOD hires policy team to implement AI principles. Available at: https://www.fedscoop.com/dod-hires-new-ai-policy-team/ Accessed March 12 2020.

Bauch, R. (2019). Berggruen Institute announces 2019-2020 class of Fellows in U.S. and China as international cohort of Berggruen Thinkers to Study Great Transformations. Berggruen Institute. Available at: https://www.berggruen.org/news/berggruen-institute-announces-2019-2020-class-of-fellows-in-u-s-and-china-as-international-cohort-of-berggruen-thinkers-to-study-great-transformations/ Accessed 27 December 2019.

Beijing Academy of Artificial Intelligence. (2019). Beijing AI Principles. Available at: https://www.baai.ac.cn/blog/beijing-ai-principles Accessed 24 December 2019

Belfield, H. (2020). Activism by the AI Community: Analysing Recent Achievements and Future Prospects. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 15-21).

Benaich, N., and Hogarth, I. (2019). State of AI Report 2019. Available at https://www.stateof.ai/. Accessed 19 December 2019

Benjamin, M. (1995). The value of consensus. *Society's Choices: Social and Ethical Decision Making in Biomedicine*. National Academy Press

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies* (Oxford Univ. Press).

Botsman, R. (2017). Big data meets Big Brother as China moves to rate its citizens. *Wired UK*, 21.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Campaign to Stop Killer Robots (2018). Country Views on Killer Robots. Available at: https://www.stopkillerrobots.org/wp-content/uploads/2018/11/KRC_CountryViews22Nov2018.pdf Accessed 11 March 2020.

Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). AI Now 2017 report. AI Now Institute at New York University. Available at: https://ainowinstitute.org/AI_Now_2017_Report.pdf Accessed 18 December 2019

Capurro, R. (2005). Privacy. An intercultural perspective. Ethics and information technology, 7(1), 37-47.

Capurro, R. (2008). Intercultural information ethics. The handbook of information and computer ethics, 639.

Cave, S., & Ó hÉigeartaigh, S. (2018). An AI race for strategic advantage: rhetoric and risks. *Proceedings of the 2018 AAAI/ACM Conference on Artificial Intelligence, Ethics and Society.*

China State Council (2017). New Generation Artificial Intelligence Development plan. Available at: http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm (translation: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/). Both accessed 20 December 2019.

Chen, D., & Hu, J. (2019) No, There Is No US-China 'Clash of Civilizations' The Diplomat. Available at: https://thediplomat.com/2019/05/no-there-is-no-us-china-clash-of-civilizations/ Accessed December 2016.

Chorzempa, M., Triolo, P., & Sacks, S. (2018). China's social credit system: A mark of progress or a threat to privacy? *Policy Briefs PB18-14*, Peterson Institute for International Economics.

Cihon, P. (2019). Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development. *Future of Humanity Institute Technical report*. Available at: https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf. Accessed 20 December 2019.

Clover, C. (2019). China: When big data meets big brother. *Financial Times*. Available at: https://www.ft.com/content/b5b13a5e-b847-11e5-b151-8e15c9a029fb.

Conger, K. (2018). Google employees resign in protest against Pentagon contract. Available at: https://gizmodo.com/google-employees-resign-in-protest-against-pentagon-con-1825729300 Accessed 11 March 2020

Cowen, T. (2019). What If Everyone's Wrong About China? *Bloomberg*. Available at: https://www.bloomberg.com/opinion/articles/2019-08-19/china-s-liberalization-shouldn-t-be-ruled-out-just-yet

Cowls, J., & Floridi, L. (2018). Prolegomena to a White Paper on an Ethical Framework for a Good AI Society. SSRN preprint.

Demchak, C. C. (2019). China: Determined to dominate cyberspace and AI. *Bulletin of the Atomic Scientists*, 75(3), 99-104.

Ding, J. (2018a). ChinAI #1 Available at: https://mailchi.mp/b945e27a35ff/chinai-newsletter-1-welcome Accessed 30 December 2019

Ding, J. (2018b). Deciphering China's AI dream. *Future of Humanity Institute Technical Report*. Available at: https://www.fhi.ox.ac.uk/wp-content/uploads/Deciphering_Chinas_AI-Dream.pdf Accessed 19 December 2019

Ding, J. (2018c). ChinaAI #19: Is the Wild East of big data coming to an end? A turning point case in personal information protection. ChinAI Newsletter. Available at: https://chinai.substack.com/p/chinai-newsletter-19-is-the-wild-east-of-big-data-coming-to-an-end-a-turning-point-case-in-personal-information-protection Accessed 28 December 2019

Ding, J. (2019). ChinAI #48: Year 1 of ChinAI. ChinAI Newsletter. Available at: https://chinai.substack.com/p/chinai-48-year-1-of-chinai Accessed 26 December 2019

Ess, C. (2005). Lost in translation?: Intercultural dialogues on privacy and information ethics." *Ethics and Information Technology* 1: 1-6.

Ess, C. (2006). Ethical pluralism and global information ethics. *Ethics and Information Technology*, 8(4): 215–226.

European Commission (2020). On Artificial Intelligence - A European Approach to Excellence and Trust. White Paper. Available at: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf Accessed March 11 2020.

European Commission (2020b). https://ec.europa.eu/digital-single-market/en/artificial-intelligence Accessed March 11 2020.

Feldman, M. D., Zhang, J., & Cummings, S. R. (1999). Chinese and US internists adhere to different ethical standards. *Journal of General Internal Medicine*, 14(8), 469-473.

Gal, D. (2019). Perspectives and Approaches in AI Ethics: East Asia. *Oxford Handbook of Ethics of Artificial Intelligence,* Oxford University Press, Forthcoming.

Gehrke, J. (2019). State Department preparing for clash of civilizations with China. *The Washington Examiner*. Available at: https://www.washingtonexaminer.com/policy/defense-national-security/state-department-preparing-for-clash-of-civilizations-with-china Accessed 22 December 2019.

Gries, P. H. (2009). Problems of misperception in US-China relations. *Orbis*, 53(2), 220-232.

Haas, P. M. (1992). Introduction: epistemic communities and international policy coordination. *International Organization*, 46(1), 1-35.

Hagerty, A., & Rubinov, I. (2019). Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence. arXiv preprint arXiv:1907.07892.

Haynes, A. & Gbedemah, L. (2019). The Global AI Index: Methodology. Available at: https://www.tortoisemedia.com/intelligence/ai. Accessed 21 December 2019

Hongladarom, S., Britz, J., Capurro, R., Hausmanninger, T., & Nakada, M. (2009). Intercultural information ethics. International Review of Information Ethics, 11(10), 2-5.

Hongladarom, S. (2016). Intercultural information ethics: A pragmatic consideration. In Information Cultures in the Digital Age (pp. 191-206). Springer VS, Wiesbaden.

Houser, K. (2018). US military declares mandate on AI. Futurism. Available at: https://futurism.com/the-byte/jaic-militarys-ai-center. Accessed 22 December 2019

Hudson, R. (2019) France and Canada move forward with plans for global AI expert council. *Science Business*. Available at: https://sciencebusiness.net/news/france-and-canada-move-forward-plans-global-ai-expert-council. Accessed 27 December 2017

International Committee for Robot Arms Control (2018). Open Letter in Support of Google Employees and Tech Workers. Available at: https://www.icrac.net/open-letter-in-support-of-google-employees-and-tech-workers/ Accessed 11 March 2020.

Jervis, R. (2017). *Perception and Misperception in International Politics: New Edition*. Princeton University Press.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389-399.

Johnston, A. I., & Shen, M. (Eds.). (2015). Perception and misperception in American and Chinese views of the other (p. 63). Washington, DC: Carnegie Endowment for International Peace.

Jun, Z. (2018). The West exaggerates China's technological progress. *Nikkei Asian Review*. Available at: https://asia.nikkei.com/Opinion/The-West-exaggerates-China-s-technological-progress Accessed 30 December 2019.

Kania, E. (2018). China's Strategic Ambiguity and Shifting Approach to Lethal Autonomous Weapons Systems. *Lawfare*, *April*, *20*.

Karnofsky, H. 2016. Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity. Available at: https://www.openphilanthropy.org/blog/potential-risks-advanced-artificialintelligence-philanthropic-opportunity. Accessed 9 March 2020

Knight, W. (2017). China plans to use artificial intelligence to gain global economic dominance by 2030. *MIT Technology Review*. Available at: https://www.technologyreview.com/s/608324/china-plans-to-use-artificial-intelligence-to-gain-global-economic-dominance-by-2030/ Accessed 26 December 2019

Laskai, L. & Webster, G. (2019). Translation: Chinese Expert Group Offers 'Governance Principles' for 'Responsible AI'. *New America, DigiChina*. Available at: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-expert-group-offers-governance-principles-responsible-ai/ Accessed 30 December 2019

Lee, K. F. (2017). The real threat of artificial intelligence. *The New York Times*, 24. Available here: https://www.nytimes.com/2017/06/24/opinion/sunday/artificial-intelligence-economic-inequality.html
Accessed 30 December 2019

Liu, M. (2019). 30 Years After Tiananmen: How the West Still Gets China Wrong. *Foreign Policy*. Available at: https://foreignpolicy.com/2019/06/04/30-years-after-tiananmen-how-the-west-still-gets-china-wrong/ Accessed 19 December 2019

May, T. (2018). Transcript of keynote speech at 2018 World Economic Forum. Available at: https://www.weforum.org/agenda/2018/01/theresa-may-davos-address/. Accessed 27 December 2019

Matsakis, L. (2019). How the West Got China's Social Credit System Wrong. *Wired*. Available at: https://www.wired.com/story/china-social-credit-score-system. Accessed 19 December 2019

McCallister, J., Zanfir-Fortuna, G., & Mitchell, J. (2018). Getting ready for the EU's stringent data privacy rule. *Journal of Accountancy*, 225(1), 36-41.

McDonald, H. (2019). Ex-Google worker fears 'killer robots' could cause mass atrocities. The Guardian. Available at: https://www.theguardian.com/technology/2019/sep/15/ex-google-worker-fears-killer-robots-cause-mass-atrocities Accessed 11 March 2020.

Ministry of Industry and Information Technology of People's Republic of China. (2019). APP (first batch) notification on infringement of user rights. Available at: http://www.miit.gov.cn/n1146290/n1146402/n1146440/c7575066/content.html Accessed 29 December 2019

Mistreanu, S. (2019). Fears about China's social-credit system are probably overblown, but it will still be chilling. *Washington Post*. Available at:

https://www.washingtonpost.com/opinions/2019/03/08/fears-about-chinas-social-credit-system-are-probably-overblown-it-will-still-be-chilling/ Accessed 20 December 2019

Mozur, P. (2019). One Month, 500,000 Face Scans: How China Is Using AI to Profile a Minority. *The New York Times*. Available at: https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html Accessed 20 December 2019

National Cyber Security Advisory Centre. (2019). Available at: https://mp.weixin.qq.com/s/smT4RbHsA_x0vIZjEKV_yg? Accessed29 December 2019

Ochigame, R. (2019). The invention of "ethical AI": How Big Tech manipulates academia to avoid regulation. *The Intercept*. Available at: https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/ Accessed 22 December 2019

Oppenheimer, M., O'Neill, B. C., Webster, M., & Agrawala, S. (2007). The limits of consensus. *Science*, 317(5844), 1505-1506.

PAI Staff. (2019). Partnership on AI Calls for Visa Accessibility Globally to Accelerate Responsible AI Development. Available at: https://www.partnershiponai.org/the-partnership-on-ai-calls-for-visa-accessibility-globally-to-accelerate-responsible-ai-development/ Accessed 21 December 2019

Pence, M. (2018) Remarks by Vice President Pence on the Administration's Policy Toward China. United States White House. Available at: https://www.whitehouse.gov/briefings-statements/remarks-vice-president-pence-administrations-policy-toward-china/. Accessed 31 December 2019

Perrault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz, B., Lyons, T., Manyika, J., Mishra, S. & Niebles, J. C. (2019). *The AI Index 2019 Annual Report*, AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA.

Rawls, John.(1993) *Political Liberalism*. Columbia University Press, 1993, pp. 134–49.

Russell, B. (1945). *A History of Western Philosophy*. Allen & Unwin.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Shane, S., & Wakabayashi, D. (2018). 'The Business of War': Google employees protest work for the Pentagon. *The New York Times*, *4*.

Shead, S. (2017). Eric Schmidt on AI: 'Trust me, these Chinese people are good'. *Business Insider*. Available at: https://www.businessinsider.my/eric-schmidt-on-artificial-intelligence-china-2017-11/ Accessed 30 December 2017

Song, B. (2019). The West May Be Wrong About China's Social Credit System. *New Perspectives Quarterly*, *36*(1), 33-35.

Søraker, J. H. (2006). The role of pragmatic arguments in computer ethics. Ethics and Information Technology, 8(3), 121-130.

Simonite, T. (2017). AI could revolutionise war as much as nukes. *Wired*. Available at: https://www.wired.com/story/ai-could-revolutionize-war-as-much-as-nukes/ Accessed 20 December 2019

Sithigh, D. M., & Siems, M. (2019). The Chinese social credit system: A model for other countries?. EUI Department of Law Research Paper, (2019/01).

Stewart, P. (2017). U.S. weighs restricting Chinese investment in artificial intelligence. *Reuters*. Available at: https://www.reuters.com/article/us-usa-china-artificialintelligence/u-s-weighs-restricting-chinese-investment-in-artificial-intelligence-idUSKBN1942OX Accesse 20 Decmber 2019

Sunstein, C. R. (1995). Incompletely theorized agreements. *Harvard Law Review*, 108(7), 1733-1772.

Szeghalmi, V. (2015). The Definition of the Right to Privacy in the United States of America and Europe. *Hungarian Yearbook of International Law and European Law*, 397.

Taylor, C. (1996). Conditions of an unforced consensus on human rights. Available at: http://people.brandeis.edu/~teuber/Taylor,%20Conditions%20of%20an%20Unforced%20Consensus.pdf

Tencent Research Institute, China Academy of Information and Communications Technology, Tencent AI Lab, and Tencent Open Platform. (2017). *Artificial Intelligence: A National Strategic Initiative for Artificial Intelligence* (人工智能：国家人工智能战略行动抓手). China Renmin University Press.

The Economist. (2018). How the West got China wrong. Available at: https://www-economist-com.ezp.lib.cam.ac.uk/leaders/2018/03/01/how-the-west-got-china-wrong Accessed 13 December 2019

Triolo, P., Kania, E., & Webster, G. (2018). Translation: Chinese government outlines AI ambitions through 2020. *New America, DigiChina*, 26.

US Department of Defense (2020). Release: DOD Adopts Ethical Principles for Artificial Intelligence. Available at https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/ Accessed 11 March 2020

UN General Assembly (2015). Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the context of International Security. Seventieth Session, Item, 93.

Webster, G., Creemers, R., Triolo, P., & Kania, E. (2017). Full Translation: China's 'New Generation Artificial Intelligence Development Plan' . *New Ameria DigiChina*. Available at: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/ Accessed 26 December 2019

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019) Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: a roadmap for research. London: Nuffield Foundation.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S.M., Richardson, R., Schultz, J. & Schwartz, O. (2018). AI Now Report 2018. AI Now Institute at New York University.

Yao-Huai, L. (2005). Privacy and data privacy issues in contemporary China. *Ethics and Information Technology*, 7(1), 7-15.

Ying, F. (2019). Understanding the AI challenge to humanity. China US Focus. Available at: https://www.chinausfocus.com/foreign-policy/understanding-the-ai-challenge-to-humanity. Accessed 29 December 2019.

Yunping, W. (2002). Autonomy and the Confucian Moral Person. *Journal of Chinese Philosophy 29:2, 251-268*

Zeng, Y., Lu, E., & Huangfu, C. (2018). Linking Artificial Intelligence Principles. arXiv preprint arXiv:1812.04814.

Zhang, S. (2017). China's Artificial-Intelligence Boom. *The Atlantic*, 20170216, 20170924.

Zhang, B., & Dafoe, A. (2019). Artificial intelligence: American attitudes and trends. Available at SSRN 3312874.

Research Center for AI Ethics and Safety,
Beijing Academy of Artificial Intelligence
Tel：+86-010-68933383
E-Mail：aies@baai.ac.cn
Website：https://www.baai.ac.cn/research/ethics-and-safety-research-center

China UK Research Centre for AI Ethics and Governance,
Institute of Automation, Chinese Academy of Sciences
Website: http://www.ai-ethics-and-governance.institute