报告编辑联系信箱：
globalaigovernance@gmail.com

欢迎任何针对本报告的评论以及相关人工智能治理的交流

上海市科学学研究所

网址：www.siss.sh.cn

邮箱：siss@siss.sh.cn

2019全球人工智能治理年度观察

来自全球50位专家的观点

2020年8月

上海市科学学研究所

ALL LIVING THINGS ARE NOURISHED
WITHOUT INJURING ONE ANOTHER,
AND ALL ROADS RUN PARALLEL
WITHOUT INTERFERING WITH
ONE ANOTHER.

—— *CHUNG YUNG*, SECTION OF *LI CHI*

萬物並育而不相害，
道並行而不相悖。

——《禮記・中庸》

目录

前言

人工智能是全球科技发展的大趋势，也是关乎全人类的大命题。

最近几年，世界各国不断发布人工智能的战略、政策，人工智能的研发和应用也越发繁荣。在中国，2017年国务院发布《新一代人工智能发展规划》作为国家级的人工智能发展战略，为2030年之前的人工智能发展描绘了基本框架。规划中也把积极参与全球人工智能治理，作为一个重要的发展方向。2019年2月，中国科技部牵头设立新一代人工智能治理专业委员会，委员会由来自高校、科研院所和企业的相关专家组成。2019年6月，该委员会发布《新一代人工智能治理原则——发展负责任的人工智能》，提出了和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷治理等八项治理原则。中国政府希望通过战略和原则，确保人工智能发展的速度和方向。

在上海这座超大型城市，人工智能是重点发展的未来产业，也是被寄予厚望的未来城市治理工具，而对人工智能进行有效治理是确保两者成功的关键。与此同时，中国国家层面也对上海的人工智能发展和治理寄予厚望。2019年，上海被批复为中国新一代人工智能创新发展试验区。在这个试验区的建设中，伦理治理是其中的重要组成部分。国家层面希望上海能够先试先行，探索出成功经验并推广至全国。

上海市科学学研究所成立于1980年1月，是中国最早成立的科技创新智库之一。40年前，中国改革开放初沐春风，科学技术有着怎样的发展规律，对经济社会有着怎样的影响，是刚刚拨乱反正后国家治理体系面临的巨大难题。科学学所的创始团队，考虑到当时中国各界对科学的认知程度，大声疾呼研究“科学学”，即研究科学的科学，意在真正理解科技、发展科技。科学学所的建所元老，以“新科技革命”为主题，对当时风起云涌的科技革命和“第三次浪潮”进行了深入讨论，深刻影响了中国国家层面和上海地方层面的科技战略。

40年后的今天，中国科学技术的实力举世瞩目，中国对科学技术的认知也已今非昔比。但是，我们所要面对的科学学问题依然存在，甚至更加复杂。近些年来，大数据、互联网、第四次工业革命以及人工智能等各种新兴技术日新月异，对经济社会乃至文化思想、国际关系等诸多方面带来深远的甚至是重塑性的影响。

好在，今天的科技早已是全球化的科技，开放合作早已是我们面对科技问题时的主流思想。面对关乎全人类共同命运的人工智能，我们的态度亦复如此。因此，我们希望通过这份《全球人工智能治理年度观察》，与全球同行携手同行，寻求各方在此领域所取得的长足进展，为休戚与共的交流与合作奠定坚实的基础。

全球命题，我们共同解决。

石谦（上海市科学学研究所所长）

主编：石谦

石谦，上海市科学学研究所所长，研究员，曾任上海科学院副院长、兼任上海产业技术研究院副院长。长期从事科技发展规划、科研项目管理、创新平台建设、创新团队培育、创新创业服务等工作。参加多项国家级行业发展规划制定、国家重大科技专项实施，主持科技部“区域产业共性技术研发组织模式的研究”和上海市“上海中长期（2021～2035）科技发展战略研究”等软科学项目。荣获2016年度上海市科技进步特等奖。担任中国科学学与科技政策研究会技术预见专委会主任、上海国家新一代人工智能创新发展试验区专委会副主任。

介绍

近几年，伴随人工智能技术和应用的快速发展，针对人工智能的治理也成为了热门议题。但是随着参与方愈加多元、发布成果海量增加，对于研究者和决策者来说，要跟踪所有人工智能治理的研究与政策进展，已经变成了一项非常艰巨的任务。

已故科学家史蒂芬·霍金曾经说过："我们的未来将是一场不断增长的科技力量与我们运用科技的智慧之间的竞赛。我们要确保人类的智慧获胜。"为总结和分享智慧，我们诚邀来自44家机构的50位世界级专家分享他们对2019年人工智能治理关键进展的看法。希望各位专家所遴选出的关键进展，能够对关心人工智能治理进展的研究者和决策者有所助益。

这些专家包括对人工智能领域做出重大贡献的科学家，他们希望用科学技术来处理科学技术带来的社会问题，为人工智能带来的挑战提供技术解决方案。例如，康奈尔大学教授、图灵奖得主约翰·霍普克罗夫特（John Hopcroft）指出，当前人工智能系统的开发可能会因训练数据的偏差而产生结果的偏见。加州伯克利大学教授斯图尔特·拉塞尔（Stuart Russell）编写的人工智能教科书《人工智能：一种现代方法》被116个国家的1300多所大学使用。他和他的同事林晨力（Caroline Jeanmaire）强调了对可证明有益的人工智能进行技术研究的重要性。香港科技大学教授、AAAI 2021大会主席杨强提倡发展联邦学习来解决当前人工智能的治理问题中特别重要的隐私问题。香港科技大学教授冯雁对开发合乎伦理的人工智能系统的正式流程进行阐述，并特别提出建立标准化算法审查系统的建议。美国路易斯维尔大学人工智能安全专家罗曼·亚姆波尔斯基（Roman V. Yampolskiy）进一步认为，不能仅讨论伦理问题，更要关注人工智能系统本身的安全问题。科学家们的这些观点为今后的人工智能治理提供了新的技术方向。

人工智能治理问题的出现，迅速吸引了传统人文社会科学领域专家的关注，并开辟出新的研究方向。国际关系研究专家、牛津大学人工智能治理中心主任艾伦·达福（Allan Dafoe）和他的同事马库斯·安德林（Markus Anderljung）为我们梳理了2019年新建立的人工智能治理机构，从中可以看出专业的人工智能治理研究正在快速建制化。法学家吉莉安·哈德菲尔德（Gillian K. Hadfield）在多伦多大学新成立了一个研究所，专注于研究"如何"治理人工智能。清华大学公共管理学院教授苏竣分享了如何采用长周期、多学科的社会实验方法研究人工智能的社会影响。图宾根大学人工智能伦理学家希洛·哈根多夫（Thilo Hagendorff）博士强调，从"软法"到"硬法"的转变是人工智能治理的下一步。这些讨论标志着人工智能治理正在成为一门严肃的学科。

在人工智能应用的前沿领域，行业领袖和投资者正在密切关注人工智能治理对创新未来的影响。国家新一代人工智能治理专业委员会成员、中国人工智能独角兽企业旷视科技创始人印奇认为，企业在推进人工智能治理方面需要承担更多的责任。IEEE标准化协会前主席唐·赖特（Don Wright）介绍了IEEE人工智能伦理准则的最新进展。GPT-2事件是2019年人工智能领域的舆论热点，OpenAI作为事件的主角，其政策团队四位成员对该事件引发的算法发表规范提出了自己的看法。来自剑桥大学的未来智能研究中心和生存风险研究中心主任贺尚安（Seán Ó hÉigeartaig）与美国乔治城大学安全与新兴技术中心战略主管童海琳（Helen Toner）也针对GPT-2事件进行了讨论。First Star的执行合伙人刘睿颐（Millie Liu）从企业视角提供了实用的观点，列出了人工智能道德在行业实施中的一些关键挑战。硅谷投资人史蒂文·霍夫曼（Steve Hoffman）表示，决策者应利用市场力量进行人工智能治理，因为企业将在该领域的进步中扮演不可或缺的角色。

2019年，人工智能治理已经成为一个真正的全球性问题，对全球治理产生重大影响。联合国人工智能和机器人技术中心负责人伊莱克利·伯利兹（Irakli Beridze）认为，在执法领域，我们应该考虑人工智能发展中所面临的伦理问题，同时也要考虑人工智能在解决全球挑战方面的积极作用。耶鲁大学教授、科技伦理专家温德尔·瓦拉赫（Wendell Wallach）基于全球人工智能治理的瓶颈问题，提出敏捷、合作和全面的治理解决方案建议。塞勒斯·霍德斯（Cyrus Hodes），尼古拉·米埃尔（Nicolas Miailhe）和杰西卡·库辛·纽曼（Jessica Cussins Newman）等专家都认为，OECD在2019年的人工智能治理方面取得了实质性进展。从他们的评论中我们能够看到OECD正在努力建立国际共识性的治理准则。国际问题专家、暨南大学教授陈定定则从国际关系角度讨论人工智能治理问题，认为人工智能治理正在成为国际关系领域的关键议题。

在人工智能治理准则全球化的同时，世界各地区也在努力打造本土化的治理体系。欧盟在人工智能治理领域是全球的积极引领者。欧洲议会议员伊娃·凯里（Eva Kaili）介绍了欧洲议会关于人工智能治理的主要工作和未来规划，尤其强调价值观在其治理中的关键地位。欧盟在2019年发布了《可信赖人工智能指南》，引起全球关注。弗朗西斯卡·罗西（Francesca Rossi），IBM研究院人工智能伦理全球负责人，同时也是欧盟人工智能高级别专家组成员，认为欧盟组织的这种跨学科、跨领域的专家组模式应作为人工智能治理的典范。欧洲人工智能政策分析师夏洛特·斯蒂克斯（Charlotte Stix）分析了欧盟对"可信赖的人工智能"的态度。英国刚刚脱欧，斯特拉斯克莱德大学李安琪（Angela Daly）博士介绍了英国政府对于人工智能治理的认识，特别介绍了新成立的政府机构数据伦理与创新中心的作用。亚洲地区也有重大进展。联合国秘书长数字合作高级别小组技术顾问高丹青（Danit Gal）认为，东亚在人工智能伦理和治理方面具有深刻的传统文化烙印。东京大学的江間有沙（Arisa Ema）博士参与了日本内阁以人为本的人工智能伦理准则的制定，她认为从政府向行业的转变是日本人工智能治理发展的主要驱动力。新加坡在2019年人工智能治理方面取得了巨大成就，他们在联合国级别的平台信息社会论坛世界峰会上获得最高奖项。这项重要工作的参与者、新加坡管理大学人工智能和数据治理中心主任吴亦涵及其同事尼地·阿莫林（Nydia Remolina）介绍了将伦理准则转化为企业可以采用的务实措施的实践。印度智库Tandem Research创始人乌瓦时·阿尼娅（Urvashi Aneja）认为，印度政策面临的主要挑战是在人工智能时代找到公平与增长之间的平衡。

国际社会普遍期望听到更多中国人工智能治理方面的进展。为此，本报告特别邀请与中国人工智能治理相关的政策制定顾问和研究专家多角度介绍了中国的发展现状。清华大学战略与安全研究中心主任傅莹认为，在人工智能治理问题上，全球各国应结伴同行、合作共赢，中国和美国作为大国，更应该以合作为主。中国科技部新一代人工智能发展研究中心主任、中国科学技术信息研究所党委书记赵志耘研究员介绍了中国政府在人工智能治理方面的理念，全面展示了中国政府已在推进的各项工作和取得的积极进展。中国科学技术发展战略研究院李修全研究员介绍了中国在人工智能治理方面注重包容发展以及关注弱势群体的理念和努力。中国社会科学院科学技术和社会研究中心主任段伟文教授讨论了为人工智能构建信任机制以及敏捷治理框架的必要性。赛迪研究院政策法规所所长栾群博士介绍了中国人工智能产业中的伦理治理进展。参与中国人工智能标准委员会相关工作、来自中国人民大学的郭锐教授，介绍了在标准制定中所秉承的理念。值得一提的是，在中国政府的人工智能治理推进中，通过建设试验区进行先行先试是非常有意义的探索。上海作为中国最大的城市，2019年启动建设国家人工智能创新发展试验区，来自上海市科学学研究所的王迎春博士介绍了相关情况。此次邀请的中国专家主要来自智库机构和学术界，事实上中国企业也在人工智能治理方面有着众多积极的探索，我们希望以后有机会再做进一步的讨论。

通过来自科学技术、人文社会科学、国际关系以及国家和地区等多维视角的专家评论，我们可以发现2019年全球人工智能治理取得诸多共识性进展。例如：越来越多的专业机构如雨后春笋般建立；越来越多的国际共识在沟通协调间得以达成；越来越多的组织机构、社会群体开始关注人工智能治理并加大投入力度。

我们相信每位读者从这一报告出发，都能延伸出自己对过往2019年人工智能治理的总结性认识，以及对未来前景的更多思考。我们由衷希望这份报告能够成为更多对话的起点，正如艾伦·图灵(Alan Turing)所说：“我们只能看到前方很短的距离，但我们可以看到有很多事情需要去做。”

李辉（上海市科学学研究所副研究员）
谢旻希（牛津大学人工智能治理中心兼职研究员）

执行编辑：李辉、谢旻希（邀请）

李辉，上海市科学学研究所副研究员。曾参与国家和上海若干人工智能战略政策的研究制定，多次在《人民日报》《光明日报》《文汇报》等媒体发表人工智能治理评论。作为主要成员，参与策划2019和2020世界人工智能大会治理论坛（上海）。2011年于上海交通大学（与宾夕法尼亚大学联合培养）获得科学史博士学位。

谢旻希，独立学者以及人工智能治理、安全和国际关系领域的顾问专家。他是Partnership on AI的高级顾问。目前担任牛津大学人工智能治理中心的政策学者。他曾经为谷歌DeepMind、OpenAI、百度、清华大学人工智能研究院、北京智源人工智能研究院以及卡内基国际和平基金会提供咨询。

致谢

这份报告的初衷，是希望促成本领域学术研究者、政策制定者、产业实践者之间的深度交流，在飞速变化的时代互通有无。我们非常荣幸这一倡议得到全球同行的广泛关注，并收到50位专家的真知灼见，实际上这是我们共同奉献结成的一个作品。

真诚感谢John Hopcroft为我们的工作提供的顾问指导。此外，还要感谢Stuart Russell、Wendell Wallach等专家在阅读报告初稿后，对整体框架结构提出的宝贵建议。

从报告立意到最终发行，于新东（上海市科学学研究所）、王迎春（上海市科学学研究所）、宋嘉（上海市科学学研究所）对项目的推进给予了坚定的支持。

在邀请专家的过程中，李修全（中国科技发展战略研究院）、Cyrus Hodes（未来社会）、Dev Lewis（亚洲数字中心）、Herbert Chia（红杉资本）、段伟文（中国社科院）及贺佳等人给予了关键性的帮助。

在报告编辑过程中，我们得到了 Caroline Jeanmaire（加州大学伯克利分校）、Thilo Hagendorff（图宾根大学）、Jessica Cussins Newman（加州大学伯克利分校）、Charlotte Stix（埃因霍温理工大学）、Angela Daly（思克莱德大学）、Kwan Yee Ng（牛津大学）、徐诺（上海市科学学研究所）、瞿晶晶（上海市科学学研究所）和张朝云（上海市科学学研究所）等年轻学者在文字校对上的帮助。张达志（华中师范大学）精心设计了插图。

中文版报告由英文版翻译而成。在此过程中，王祥丰（华东师范大学）帮助我们审校了相关的技术内容，文贤庆（湖南师范大学）、胡晓萌（湖南师范大学）、张晨琛（机器之心）、仵翼颖（机器之心）、陈自富（世纪明德）、陈秋萍（上海市科学学研究所）帮助进行了文字校对。徐诺（上海市科学学研究所）做了全报告的总校。

实习生张洁、宋志贤、孙慧、倪佳炜、梁新怡、黄思伟等承担了大量事务性工作。

玉汝于成，对所有师友同仁的鼎力襄助，我们一并致谢。

报告编辑联系信箱：globalaigovernance@gmail.com

我们欢迎任何针对本报告的评论以及与相关人工智能治理的交流。

第一部分：
人工智能科学家的思考与倡议

迎接科学革命，培育更多人才

约翰·霍普克罗夫特

尽管深度学习只是人工智能技术的一种，但却对人工智能产生了重大影响。深度学习已经在图像识别、机器翻译、金融等多个领域得到了实际应用，同时也带来了很多问题。例如，假设一个人工智能程序在帮助银行做贷款的决策，人们会想了解程序做出该决策的原因。但就目前所掌握的知识而言，人类还无法对此类问题作出解释。另一个问题是训练数据中的“偏差”可能会造成最终的决策结果存在“偏见”。

显然，一场以人工智能为主要驱动力的革命正在发生。未来，人才将是决定一个国家经济发展和人民生活水平的决定性因素。我认为，包括中国在内的很多国家，当前最重要的问题是要提高大学本科教育质量。只有培养源源不断的高素质人才，才有可能帮助中国成为信息时代的领先经济体。

作者简介　约翰·霍普克罗夫特（John E. Hopcroft）

约翰·霍普克罗夫特是美国康奈尔大学IBM计算机科学工程与应用数学教授。获得斯坦福大学电气工程硕士学位（1962年）和博士学位（1964年）后，霍普克罗夫特曾在普林斯顿大学任教了三年。于1967年加入康奈尔大学，1972年被任命为教授。1985年，霍普克罗夫特曾被任命为计算机科学系的约瑟夫·C·福特教授。于1987-1992年担任计算机科学系主任，并于1993年担任学院常务副院长。1994年1月至2001年6月，霍普克罗夫特教授任约瑟夫·希尔伯特工程系主任。作为西雅图大学的本科生校友，霍普克罗夫特教授在1990年获得了文学荣誉博士学位。

霍普克罗夫特教授的研究集中在计算理论方面，特别是算法分析、自动机理论和图算法。

霍普克罗夫特教授与杰弗里·乌尔曼（Jeffrey D.Ullman）和阿尔弗雷德·阿霍（Alfred V.Aho）合著了四本关于形式语言和算法的著作。最近，致力于研究信息的获取与访问。

1986年，霍普克罗夫特教授被授予图灵奖。他是美国国家科学院（NAS）院士、美国国家工程院（NAE）院士、中国科学院外籍院士、美国艺术与科学院（AAAS）院士、美国科学促进会院士、电气与电子工程师学会院士，以及计算机械协会院士。1992年，被时任总统布什任命为国家科学委员会成员（负责监督国家科学基金会（NSF）），任期至1998年5月。1995年至1998年，霍普克罗夫特教授在国家研究委员会（the National Research Council）的物理科学、数学和应用委员会任职。

除上述任命，霍普克罗夫特教授还是美国工业与应用数学学会（SIAM）财务管理委员会、印度信息技术研究所新德里咨询委员会、微软亚洲研究院技术咨询委员会和西雅图大学工程咨询委员会的成员。

开发有益的人工智能系统

斯图尔特·拉塞尔　林晨力

2019年，人工智能治理取得显著进展，一系列重要原则相继发布。其中，《人工智能北京共识》和《经合组织人工智能原则》的发布具有重要意义，两者都特别注重确保人工智能系统的短期和长期安全，而这也是人工智能发展的重要方面。

原则的制定是采取行动的良好基础，现实中也确实有一些实例。加利福尼亚州率先颁布法律，要求将所有试图影响加州居民投票或购买行为的自动在线账户公开标识为机器人。这部法律的颁布标志着我们通过遏制欺骗性新技术，在确保人工智能系统值得信赖的进程上迈出了重要的第一步。同时，它也是朝着建立一项基本人权，即了解一个人是在与另一个人还是一台机器互动所迈出的重要一步。此外，该法律还将阻止错误信息的传播。我们希望这部法律能够不局限于解决商业和投票问题，而是成为一项普遍权利，并为其他州和国家做出榜样。

然而，在某些领域，人工智能治理工作严重滞后也带来了一些风险。国际社会在管制致命性自主武器（LAW）方面进展甚微。无人机、坦克和其他计算机控制的自主武器通过人工智能系统运行，并通过编程来定位、选择和攻击目标，而无需人类控制。2019年11月，联合国在日内瓦举行《特定常规武器公约》成员国会议，但各国外交官未能在会上就这一问题达成具有约束力的共识。因此，我们在未来两年中还会持续开展非约束性的谈判，而非切实的法律工作。期以在全球禁止致命性自主武器，保护人类共同的未来，还需我们进行大量工作。

随着功能日益强大、具有高度自我维持能力的人工智能系统的发展，人类必须确保这些人工智能系统对人类有益且安全。本文的作者之一拉塞尔（Russell）刚刚出版了一本关于这个主题的书——《人类兼容：人工智能与控制问题》（Viking/Penguin出版社，2019年）。人工智能系统的控制问题，并不是好莱坞和媒体一直关注的科幻故事情节：一个人形机器人实现了自我意识的觉醒，并决定憎恨人类。它是一个机器创造问题，这些被创造出来的机器能比人类获取更多的信息，能看到更长远的未来，在现实世界中比人类拥有更强的决策能力。按照目前的人工智能概念和技术方法，我们不太可能继续控制比我们更强大的机器。为了解决这个问题，研究界需要付出巨大的努力来改变人工智能的标准模型，使其朝着对人类有益的方向发展。人工智能界也开始意识到这个问题，这让我们对实现这一转变充满希望，但目前仍有很多工作要做。

作者简介　斯图尔特·拉塞尔（Stuart Russell）

1982年，斯图尔特·拉塞尔获得牛津大学物理学一等荣誉学士学位；1986年，他获得斯坦福大学计算机科学博士学位。随后，他加入了加州大学伯克利分校，担任电气工程和计算机科学教授（曾任系主任）、史密斯-扎德（Smith-Zadeh）工程学院主席以及人类兼容人工智能中心（Center for Human Compatible AI）主任。他曾在加州大学旧金山分校担任神经外科客座教授，还曾担任世界经济论坛人工智能和机器人委员会副主席。他曾获美国国家科学基金会总统青年研究员奖、IJCAI计算机与思维奖、世界技术奖（政策类）、美国统计协会米切尔奖、人工智能促进协会（AAAI）费根鲍姆奖，以及美国计算机协会（ACM）和AAAI颁发的杰出教育家奖。2012–2014年，他担任巴黎帕斯卡基金会主席，并获得2019–2021年安德鲁·卡内基奖金。他是牛津大学瓦德汉学院的荣誉院士、斯坦福人类中心人工智能研究所特聘研究员、英国皇家国际事务研究所副研究员，以及人工智能促进协会、美国计算机协会和美国科学促进会研究员。他的著作《人工智能：一种现代方法》（Artificial Intelligence: A Modern Approach，与Peter Norvig合著）是人工智能领域的标准教材，该书已被翻译成14种语言，并在128个国家的1400多所大学使用。他的研究涉及人工智能的多个领域，包括机器学习、概率推理、知识表示与规划、实时决策、多目标跟踪、计算机视觉、计算生理学和哲学基础等。他还为联合国工作，为《禁止核试验条约》开发新的全球地震监测系统。他目前关注的课题包括自主武器的威胁、人工智能的长远未来及其与人类的关系。后一个主题也是他的新书《人类兼容：人工智能与控制问题》（Human Compatible: AI and the Problem of Control）（Viking/Penguin出版社，2019年）的主要内容。

作者简介　林晨力（Caroline Jeanmaire）

林晨力拥有北京大学国际关系硕士学位，以及巴黎政治学院国际公共管理硕士学位和政治学学士学位。此外，她曾在塔夫茨大学弗莱彻法律与外交学院进修。她在加州大学伯克利分校人类兼容人工智能中心（UC Berkeley Center for Human Compatible AI，CHAI）研究国际协调模型，以确保人工智能系统的安全性和可靠性。她还负责CHAI的伙伴关系和对外关系战略，致力于建立一个围绕人工智能安全和与关键利益相关者的关系的研究社区。在加入CHAI之前，她曾是哈佛大学肯尼迪政府学院孵化的智库“未来社会”（Future Society）的人工智能政策研究员和项目经理。她还为在迪拜世界政府峰会上组织的第一届和第二届全球人工智能治理论坛提供了大力支持。在2019年的全球人工智能治理论坛上，她负责管理两个委员会：人工智能地缘政治委员会和国际人工智能研究小组。她发表了关于人工智能地缘政治、中美人工智能产业合作杠杆，以及全球人工智能治理公民辩论结果的文章和报告。在此之前，她参加了2015年以来的多次气候谈判和技术会议，包括随同法国代表团参加COP23和COP24会议。林晨力通晓英语、法语、西班牙语和中文。

联邦学习的重要性

杨强

随着人工智能走出实验室，实现了大规模应用，其潜在的伦理问题及影响逐渐引起公众的关注。回顾2019年，与人工智能伦理相关的公众讨论，集中体现在对用户数据隐私的保护及治理问题上。国际上，Facebook因为非法泄露用户数据，而被美国联邦贸易委员会（FTC）处以50亿美元的罚款。谷歌因为其隐私条款难以被用户理解，使得用户难以管理个人资料被使用的方式，违反了GDPR，而被法国监管机构开出数千万欧元的罚单。而在中国，数据公司因为滥用、售卖未经授权的用户隐私数据，而被监管机构密集调查，大量数据公司受到停业、app下架等处罚，情节严重者甚至面临刑事责任。这一系列的事件表明：一方面，公众对个人隐私相关的数据权力意识逐渐高涨，因而这些事件才在媒体和公众中引起广泛关注；另一方面，触目惊心的事件真相也表明，私人数据的保护和治理严重滞后和缺失。

追根溯源，这些问题的产生，虽有人工智能技术严重依赖于海量数据收集的客观诱因，但主要原因是相关利益方忽视社会责任，主观上肆意妄为导致的。如何在充分尊重和保护用户数据隐私的前提下，去挖掘数据背后的知识和价值，是人工智能研究者面临的一个迫在眉睫的挑战。

幸运的是，在2019年，我们能看到人工智能研究者已经意识到问题的严重性并着手提出了一系列解决方案。其中，联邦学习作为一种大有希望的用户数据隐私保护方案，展示了其在推进产业应用落地上的独特优点。联邦学习是指在数据不出本地、数据不共享的前提下，通过交换加密参数，实现多个参与方联合建模的一种技术方案。联邦学习的建模效果与将整个数据集聚合建模的效果相同，或相差不大。在联邦学习技术框架中用到了多种加密技术，如安全多方计算、同态加密（Homomorphic encryption）、姚式混淆电路和差分隐私（Differential privacy）等。从技术应用的角度看，目前联邦学习已经在小微企业信贷、反洗钱、反欺诈、保险、计算机视觉等领域落地应用。此外还在智慧医疗、自动驾驶、智慧城市、政府治理等领域进行探索。综上所述，联邦学习可以看作是机器学习技术与隐私保护技术的集大成者，并且是一种拥有广泛应用前景的通用型隐私保护机器学习技术。

作者简介　杨强

杨强是微众银行首席人工智能官、香港科技大学计算机科学与工程系的讲席教授和前系主任。他是国际人工智能界“迁移学习”（Transfer Learning）技术的开创者，同时提出“联邦学习”（Federated Learning）的研究新方向。他于2013年7月当选为国际人工智能协会（AAAI）院士，之后又于2020年1月当选为AAAI2021主席。2017年8月他当选为国际人工智能联合会（IJCAI）理事会主席。

为开发符合伦理的人工智能建立正式流程

冯雁

关于不同政府和社会背景下的人工智能治理有很多讨论。2019年，联合国、联合国教科文组织、欧盟、欧洲议会、中国、美国、日本、阿联酋等组织和政府纷纷提出了新的人工智能发展战略和治理政策。世界上顶尖的人工智能公司正在积极研究和开发具有伦理道德和有益的人工智能，以及相应的良好治理方案。谷歌首席执行官的最新声明表明，人工智能应用程序不能仅由市场力量决定，也需要良好的治理。这反映了人工智能社区的普遍共识。

所有的机器都会犯错误，但人工智能的错误会让人们产生更多的恐惧，因为人工智能的错误与人类自身的错误非常相似。消费者倾向于将这些错误与邪恶的类人意图联系起来。如果一个说话者录下了我的对话，或者一台摄像机给我发送了其他人家里的图像，那么人工智能就是“间谍”。如果一个搜索结果有偏见，那么它就是“性别歧视”或“种族歧视”。如果聊天机器人给出了错误的答案，听起来可能会“吓人”或“无礼”。突然之间，过去只需要处理矩阵中的数字来提高系统性能的工程师，将面对的是不断寻求哲学上甚至是法律上的答案的用户。这让工程师们有些措手不及。

在人工智能算法和系统开发层面，研究人员和工程师需要最佳实践指南和正式流程规范，以确保一个公平、负责和透明的流程，同时减少和最小化机器偏差和机器错误。目前，研究人员和开发人员通常将数据库、经过培训的模型和软件代码发布到公共领域供他人使用。这些数据库和模型中的固有偏见可以进而扩散到由它们开发的所有系统。

像IEEE这样的专业组织以符合伦理的设计过程的形式提供了最佳实践指南。我们可以将这些原则应用到人工智能算法和系统开发的所有领域。像“人工智能伙伴关系”这样的非政府组织有专门的工作小组，以提供最佳实践指南为目标，由工程师、哲学家和民间社会代表组成的成员提供专业意见。国际标准化组织（ISO）有164个成员国，包括美国和中国，正在致力于人工智能领域的标准化。

越来越多的人要求人工智能和机器学习开发的正式流程与软件开发过程并行，作为人工智能软件产品开发的一个组成部分。人工智能专业人员认可的正式流程将确保通用标准、更可解释和可验证的开发过程，以及更少的系统错误。一个正式的流程可以包括以下标准：

1、数据库收集：在将数据偏见发布到更大的人工智能社区之前，应该减少数据偏见；

2、软件和算法设计：会话式人工智能应该是非歧视性的；生物特征识别不只是依靠声音或人脸识别，它应该是多模态的，以减少误差；

3、模型训练：记录下具体的模型架构和参数设置，使过程可以沿着流水线进行复制和解释，而不需要人工试错；

4、测试和验证：机器的公平性和偏差也可以在标准测试集上进行评估和测试。许多人工智能会议已经运行共享任务，不同团队使用公共的培训和测试集来比较他们的系统。这可以在不扼杀研究的创造性和安全性，并保护学术独立性的情况下，对人工智能算法和系统的发展进行抽象化和形式化。

欧洲议会呼吁建立一个中央监管机构，就像美国食品和药物管理局一样，在算法投入使用之前评估它们的影响。该方案面临两个挑战：算法以极快的速度发展，每隔几个月就会进行修改和更新；可能没有足够的专家具备算法评估所需的技术知识。因此，我建议这样一个监管机构的任务是评估人工智能产品和应用程序，而不是评估基础算法。算法评估应该纳入研究出版物的正常同行评审过程。负责策划这些出版物的编辑和技术项目主席应要求审稿人就他们所审查的工作的伦理问题提供明确的意见。人工智能专业人员愈发意识到他们工作中伦理规范的重要性。我希望我们的集体智慧能推进这一方面的进展。

随着当今开发的人工智能技术已经成为开放资源并在全世界迅速共享，我们在人工智能治理方面需要更多的国际合作。如今，人工智能研究和教育是全球性的。各公司正在合作制定自动驾驶的标准。各国正在合作管控自主武器。虽然算法的演变需要新的法规，但是人工智能在安全、医疗和金融领域的应用仍受制于每个地区的现有法规。社交媒体和信息的完整性仍然是一个具有挑战性的领域，社交媒体公司目前正在自我调节，而没有达成共识。因此，我们需要更多的国际合作，并且需要与人工智能专家和其他利益相关者建立监管机构。2019年，我们看到了更详细的人工智能治理规划，以及公众对其需求的更多认识。在2020年及以后，我们需要积极实施拟议的良好实践指南和正式的软件流程，以确保人工智能系统的公平、问责和透明。

作者简介 冯雁（Pascale Fung）

冯雁教授是香港科技大学电气与电子工程系教授。她因“对人机交互的贡献”被选为电气与电子工程师协会（IEEE）会士，并因“对口语人机交互领域的基础性贡献”被选为国际语音通信协会会士。她是港科大人工智能研究中心（CAIRE）主任，该中心是港科大四所学院中首屈一指的跨学科研究中心。她是世界经济论坛智库“全球未来理事会”的专家。她代表香港科技大学参加人工智能伙伴关系（Partnership on AI to Benefit People and Society）。她是IEEE信号处理协会的理事会成员。冯教授出生于上海，父母都是职业艺术家，但小时候对科幻小说产生兴趣后，她发现了自己对人工智能的兴趣。如今，她的研究兴趣在于构建能够理解和对人类有共情能力的智能系统。为了实现这一目标，她的具体研究领域是利用统计建模和深度学习进行自然语言处理、口语语言系统、情感和情绪识别，以及人工智能的其他领域。冯教授能说七种欧洲及亚洲语言，对多语言演讲及自然语言问题特别感兴趣。

从人工智能治理到人工智能安全

罗曼·亚姆波尔斯基

2019年，人工智能治理引起了人们的广泛关注，迄今已有30多个国家制定了人工智能治理战略和计划，以使人工智能朝着有利于实现其本地计划与国际计划的方向发展。他们希望基于欧盟、北欧—波罗的海地区和联合国提出的多国战略，为人工智能的研究、部署和国际合作制定标准和规范。与此同时，一些世界顶尖大学的研究中心当前也活跃于人工智能治理领域的研究中。具体请参阅生命未来研究所的全球人工智能政策报告，(https://futureoflife.org/ai-policy/)它回顾了多个相关国家和国际的倡议。

针对人工智能伦理的研究在2019年的增长近乎指数级。首先，至少30个组织提出了自己的伦理“原则”。仔细比较，我们可以发现这些原则都高度重视人权、人类价值、职业责任、隐私、人为管控、公平和不歧视、高透明度、可解释性和问责性等方面。但各项提案对于每个类别的重视程度有所不同，在表达同一意见时所使用的语言也存在差异。未来将可能有更多组织提出自己的伦理原则，从而使人工智能伦理原则的格局和工作标准化更加复杂。可参阅哈佛大学伯克曼·克莱因中心的报告，(https://ai-r.cyber.harvard.edu/primp-viz.html)该报告尝试分析和描述基于伦理和人权的方法来开发有原则的人工智能。

2019年，人工智能安全也取得了很大的进展，多家公司和高校成立了人工智能安全小组。然而，区分人工智能的治理/伦理与人工智能技术的安全和保障性是非常重要的。虽然针对人工智能治理/伦理的研究能够为人工智能研究提供方向、资源、协调和框架，但却不能直接提高人工智能系统的安全性。而这一点只有直接的人工智能安全研究才能做到。因此，如果将人工智能治理和伦理方面的进展误认为安全方面的进展，这是非常危险的，这会让我们对其安全性产生错误认知。我希望，我们能够在2020年更明智地区分人工智能治理、伦理和安全，并认识到它们各自的重要性和局限性。

作者简介　罗曼·亚姆波尔斯基（Roman V. Yampolskiy）

罗曼·亚姆波尔斯基博士是路易斯维尔大学Speed工程学院计算机科学与工程系的终身副教授。他是网络安全实验室的创始人和现任主任，著有多部著作，包括《人工超智能：一种未来主义方法》(Artificial Superintelligence: a Futuristic Approach)。在路易斯维尔大学任职期间，亚姆波尔斯基博士获得了杰出教学教授、年度教授、最受欢迎教员、四大教员、工程教育领袖以及年度十大在线学院教授等诸多荣誉，同时他还获得杰出职业生涯早期教育奖等众多荣誉奖项。亚姆波尔斯基博士是IEEE和通用人工智能学会（AGI）的高级会员、肯塔基科学院的会员，以及全球传播研究院（GCRI）的研究助理。亚姆波尔斯基博士的主要研究方向是人工智能安全和网络安全。亚姆波尔斯基博士已经发表了超过100篇刊物，其中包括多篇期刊文章和著作。他的研究已被超过1000名科学家引用，在美国和美国以外的流行杂志、数百个网站、广播和电视都对其作品进行了大量报道。亚姆波尔斯基博士曾应邀于瑞典国家科学院、韩国最高法院、普林斯顿大学等多个机构举办的100多个活动上发表演讲。

第二部分：
人文社会科学专家的关注与进展

人工智能治理领域的迅速发展

艾伦·达福　马库斯·安德林

2019年对人工智能治理来说是重要的一年。在这一年，人工智能公司和研究团体为应对人工智能治理的挑战，建立了许多新的人工智能治理研究机构，并在人工智能政策领域取得了一定的研究成果。尽管还有许多工作要做，但令人振奋的是，这一领域发展迅速，而我们有幸为推动这一发展贡献了自己的一份力量。

许多大型科技公司已开始建立和修改其流程和结构，以有效解决人工智能伦理和治理方面的问题，但其中一些尝试适得其反。比如谷歌提议建立的道德委员会，在董事会成员的选择上引发了争议，一周多后就解散了。而其他的尝试，如Facebook的独立内容审查监督委员会，引起的争议稍小。OpenAI决定阶段性地发布其自然语言模型GPT-2的决定引起了巨大争议，同时也在人工智能社区中引发了关于发布规范的必要性的讨论。虽然在现阶段解决这些问题存在一定难度，但是随着人工智能系统的不断改进，这一难度会只增不减。当然，我们也能在人工智能政策领域看到了一些鼓舞人心的发展。欧盟对人工智能政策表现出了极大的兴趣，其人工智能高级别专家组已经发布了一套道德准则和一系列政策及投资建议。此外，新任委员会主席乌尔苏拉·冯德莱恩（Ursula von der Leyen）也承诺全面启动人工智能立法。此前在人工智能治理问题上基本保持沉默的政策参与者，如今也开始向外界发声，例如中国发布《人工智能北京共识》和美国国防部发布《人工智能原则》（AI Principles）。尽管这些原则与解决人工智能治理问题的实际措施相去甚远，但仍具有重要意义，因为它们为一些重要问题的审议奠定了基础。

2019年，一些人工智能治理和伦理机构组织陆续成立，包括多伦多大学的施瓦茨·赖斯曼技术与社会研究所、华盛顿特区的安全与新兴技术中心以及在牛津大学建立的人工智能伦理研究所和牛津大学互联网学院发布的新兴技术治理项目。我们期待与这些新同仁合作。

在牛津大学的人工智能治理中心，我们一直致力于发展团队并开展研究。我们现在拥有一个由7名研究人员组成的核心团队和一个由16所研究分支机构和合作单位组成的组织网络。最重要的是，我们在过去一年里取得了很多重要的研究进展。我们发布了一些研究报告如《人工智能：美国人的态度和趋势》（US Public Opinion on Artificial Intelligence），专栏文章如《思考来自人工智能的风险：事故、恶用和结构》（Thinking About Risks From AI: Accidents, Misuse and Structure）以及学术论文如《如何衡量进攻-防御平衡?》（How does the offense-defense balance scale?）和收录在AAAI/ACM的人工智能、伦理与社会会议部分的五篇论文等。

作者简介　艾伦·达福（Allan Dafoe）

艾伦·达福，牛津大学人类未来研究所人工智能国际政治学副教授，人工智能治理中心主任。艾伦·达福主要研究大国战争的起因，以及围绕变革性技术的全球政治，并特别关注人工智能风险。为更好地研究这些课题，他还研究了因果推理和提高透明度的方法。

作者简介　马库斯·安德林（Markus Anderljung）

马库斯·安德林是牛津大学人类未来研究所人工智能治理中心的人工智能战略项目经理。马库斯致力于推动该中心的发展，确保中心的研究对重要利益相关者的价值，以及促进相关项目研究的开展，同时他还亲身参与了其中一些研究工作。他拥有科学史和科学哲学背景，专注于经济哲学和循证政策，具有多年管理咨询经验。此外，他还是瑞典有效利他主义协会的执行董事。

人工智能与人类价值观的有效统一：从“应该”到“如何”

吉莉安·哈德菲尔德

“我们应如何监管人工智能?”在过去几年中，这一问题一直是人工智能治理的关注焦点。该问题以道德哲学难题的形式出现，比如电车难题。人工智能的支持者和批判者也提醒人们注意算法和人脸识别技术中存在的歧视和偏见。针对性较强的政治广告可能对政治和社会关系的稳定产生重大影响。而对这种影响的担忧也引发了人们的思考，即我们是否应该有效监管社交媒体平台上的言论，或限制对个人信息的收集?正如欧盟人工智能高级别专家组在2019年所提出的那样，在最广义的层面上，人们普遍认为，人工智能应该“尊重所有适用的法律法规、伦理原则和价值观”。

但该如何实现人工智能与人类价值观的统一?在实践中，怎样才能确保人工智能合法且合乎伦理?

由于世界的复杂化和动态化特征，仅通过制定法律来对人工智能进行约束远远不够，而将人类的价值观和伦理观一一列出，编入人工智能系统也是行不通的。

正如我的研究所发现的，并在《扁平世界的规则：人类为何发明法律，又如何再造法律以应对复杂的全球经济》(Rules for a Flat World: Why Humans Invented Law and How to Reinvent It for a Complex Global Economy)一书中讨论的那样，早在面临人工智能的挑战之前，我们的法律和监管体系在将政策选择——即我们的“应行之事”付诸实践方面，就已经面临着实质性的局限。我们在20世纪完善的法律和监管技术——立法、监管、管理机构、法院、法律推理——越来越无法跟上21世纪经济和社会的复杂性、发展速度以及全球化程度。人工智能领域的发展进一步拉大了我们希望通过法律法规实现的目标与我们取得的实际成果之间的差距。

2019年，尽管大多数的人工智能治理项目持续关注“我们应该监管人工智能”的问题，但多伦多大学启动了一项新倡议，将重点转向“我们如何才能监管人工智能”。在我的领导下，施瓦茨·赖斯曼技术与社会研究所将进行基础的跨学科研究，以期建立起集技术、法律和监管于一身的体系，实现我们为人工智能设定的政治目标。例如，我们不会关注是否应该监管人脸识别系统，我们所关注的是，如果我们要落实规则，例如非歧视或合法的监控限制原则，我们应该如何确保人脸识别系统遵守这些规则?我们需要解决哪些技术挑战?在监管技术上可以进行哪些创新?如何构建人工智能来帮助我们确保其处于我们普遍认为正确或可接受的范围内?我们如何确保在价值一致方面的努力不会付诸东流?

2020年及以后，施瓦茨·赖斯曼研究所的目标是将有关人工智能治理的全球讨论从“应不应该治理”扩展到“如何治理”。我们将致力于发展现有的知识和工具库，以确保人工智能在需要之处而不是在非必要之处得到应用，且遵循人类制定的规则。

作者简介

吉莉安·哈德菲尔德(Gillian K. Hadfield)

吉莉安·哈德菲尔德，女王大学(荣誉)文学学士，斯坦福大学文学硕士、法学博士、经济学博士，多伦多大学法学教授、战略管理教授，施瓦茨·赖斯曼技术与社会研究所客座教授。她担任施瓦茨·赖斯曼技术与社会研究所的首任所长。她的研究重点是发达经济体和发展中经济体法律和争端解决制度的创新设计；人工智能治理；法律、律师和争端解决的市场机制；以及合同法和合同理论。哈德菲尔德教授是多伦多人工智能矢量研究所(Vector Institute for Artificial Intelligence)的研究员、加州大学伯克利分校人类兼容人工智能中心(Center for Human-Compatible AI)的研究员，以及旧金山OpenAI的高级政策顾问。2017年，牛津大学出版社出版了她的著作《扁平世界的规则：人类为何发明法律，又如何再造法律以应对复杂的全球经济》(Rules for a Flat World: Why Humans Invented Law and How to Reinvent It for a Complex Global Economy)。

采用长周期、多学科的社会实验方法研究人工智能的社会影响

苏竣

“以人为本”是中国发展人工智能等新兴技术的一贯宗旨。中国政府高层和学术界高度关注人工智能给人类社会带来的影响，努力探索人工智能社会治理方案，推动人工智能技术进步更好地服务于人类福祉。令人欣喜的是，2019年，中国启动了利用社会实验方法研究人工智能社会影响的举措，在人工智能治理上迈出了领先的一步。

作为新一轮科技革命的重要驱动力量，全世界已经深刻认识到人工智能技术发展与应用给人类社会带来的机遇与挑战。对人工智能所衍生出的技术失控、社会两极分化加剧等风险和威胁保持警惕，也成为了大家的共识。

针对人工智能技术的风险与挑战，我们不仅需要积极倡导建立负责任的研发和创新价值体系，加强对科技创新过程中伦理问题的关注和控制。更应当回归“人文主义”，从人文社会科学的角度，加强对人工智能社会影响机理、规律与趋势的研究，完善和强化人工智能发展的社会政策体系。实现对人工智能的有效治理，需要我们对人工智能时代的社会形态和特征具有全面而系统的认知和把握。而这种认知的建立，有赖于以实践证据为基础的实证研究，特别是社会实验研究的开展。

社会实验是一种经典的社会科学研究方法，是以现实情境下的人、组织作为研究对象所开展的研究活动，目的在于通过一些社会、政治过程或技术变革导致的近似于理想实验的场景来研究社会科学的问题。面对智能化时代社会治理的新问题，中国学术、政府、社会各界共同从社会实验实践的角度入手，致力从学术研究、政策实践、社会影响等多个维度推动人工智能社会实验方案的形成、完善与落地。

2019年，清华大学、浙江大学等单位的专家学者，在大量前期工作的基础上，针对人工智能社会影响的问题，率先提出了开展长周期、宽领域、多学科人工智能社会实验的政策建议。

基于学术研究成果，中国的政策实践正在迅速成型、不断落实。2019年，中国科技部组织制定了《国家新一代人工智能创新发展试验区建设工作指引》，将社会实验作为新一代人工智能创新发展试验区建设的重要工作。从教育、交通、政府、医疗、环保、制造、金融、农业等多个应用场景入手，从社会风险、组织再造、数据安全、技术适应等多个维度出发，在全国开展人工智能社会治理实验。

同时，中国社会对于人工智能社会治理问题的共识正在逐渐形成，公众对于社会实验方案的认可与支持也在不断加深。2019年10月，第一届全国人工智能社会实验学术会议在中国的清华大学召开，100余位专家学者交流分享了人工智能社会实验的最新成果，研究讨论了人工智能社会实验的工作方案和行动计划。《光明日报》等主流媒体刊发了《探索人工智能社会治理的中国方案》等文章，获得社会各界的广泛赞誉。人工智能社会实验的公众基础与社会影响力与日俱增。

通过中国目前在人工智能社会治理上的举措与收获，我们更加明确，开展人工智能社会实验，有助于准确识别人工智能对人类社会带来的挑战和冲击，深入理解人工智能的社会影响特征与态势，深刻把握人工智能时代社会演进的规律，从而为建立有人文温度的智能社会提供科学的参考依据。

作者简介　苏竣

清华大学公共管理学院教授，教育部长江学者特聘教授。

现任清华大学智能社会治理研究院院长、清华大学科教政策研究中心主任、清华大学智库中心主任，兼任教育部公共管理类学科专业教学指导委员会副主任，享受国务院特殊津贴。

兼任哈佛大学肯尼迪政府学院兼职研究员、塔夫茨大学弗莱彻学院兼职高级研究员。任第一届全国人工智能社会实验学术会议主席，清华——哈佛低碳发展与公共政策国际研讨会（2014-2018）联席主席。

超越人工智能伦理准则

希洛·哈根多夫

2019年，关于人工智能伦理的讨论无处不在。学术界、政府部门以及产业界都提出了自己的人工智能伦理准则。新闻媒体充斥着呼吁人工智能伦理的文章。此外，众多的委员会聚集在一起制定规范和标准。暂且将有关人工智能伦理的恶意评论搁置，2019年研究人员和从业者愈发强调：抽象的伦理原则若不付诸实践，其价值就会大打折扣。然而，知之非难，行之不易。世界各地的伦理倡议组织一致认为，做到隐私、公平、透明、安全和负责任是构建和使用“合乎伦理的”人工智能应用程序的最低要求。然而，在开发和部署此类应用程序的日常决策中，我们需要制定什么样的伦理原则尚不明确。至少，实证研究表明，仅仅阅读有关伦理原则的文件对实践没有任何显著帮助。

制定伦理原则只是人工智能治理的第一步。如果要提高人工智能研发中合乎伦理行为的可能性，那么治理工作必须从这些方面入手：要保证伦理准则的执行；转变组织或工作环境中的伦理文化、美德教育；解决从竞争转向合作；等等。其中，人工智能领域的激烈竞争和与之相关的“全球领导地位”的竞争言论带有导致不计后果的竞争的风险，因为人们通常力争最先完成某些技术系统，特别是在军事应用方面的系统。这种竞争对安全、隐私和公平等造成了巨大的损害。减少竞争，促进国家、公司以及研究机构之间的合作是实现“可信赖人工智能”的重要一步。

2020年，人工智能治理应侧重于加强行业利益相关者之间的联系，同时也应关注治理举措本身。这将在商议治理原则时节省大量时间。此外，2020年应该是“软法”逐渐转化为“硬法”的一年，应当针对算法不歧视、在高风险领域禁止使用人工智能、安全和隐私标准以及人工智能应用导致的劳动力转移的处理规则等提供明确的规则。

作者简介　希洛·哈根多夫（Thilo Hagendorff）

希洛·哈根多夫博士任职于德国图宾根大学的“机器学习：科学新视角”卓越集群的伦理和哲学实验室。他还是技术科学协会VDE（电气、电子和信息技术协会）AI伦理影响小组的成员。他的研究重点关注机器学习中的伦理问题以及媒体和技术伦理领域更宽泛的问题。此外，希洛·哈根多夫博士还是图宾根大学科学与人文伦理国际中心副研究员，以及图宾根大学和波茨坦大学哈索普拉特纳研究所讲师。

用跨学科方法探索人工智能治理研究

佩特拉·阿韦勒

人工智能，特别是人工智能在自动化决策领域的伦理问题，在包括中国和德国在内的许多国家的国家政策战略中都受到高度重视。国际合作旨在基于共同的伦理框架，建立起人工智能社会生态系统中的联合研究和治理网络。

针对自动化决策的人工智能算法的改进，很大程度上取决于训练数据的可用性和质量。然而，在高风险决策环境下，很难获得有效的经验数据。想象一下在核电站发生事故、发生海啸或特大城市发生恐怖袭击时的自动决策算法，由于这类事件太少（当然，这是一件幸事），根本无法获得足够的训练数据。此外，个体也是影响自动决策算法的环境因素之一。由于每个人的背景、文化和社会教养不同，很难预测他们的行为和互动方式。因而社会框架在全球范围内表现出多样性，进一步限制了现有训练数据的通用性和适用性。

那么，我们可以从何处获得模型和训练数据以改进算法，使其更适合特定环境的标准规范和国际社会的价值观？这就是跨学科研究机构（如美因兹大学技术与创新社会学/社会模拟实验室（TISSS）和欧洲社会模拟协会（ESSA）大型科学社区）的研究内容所在。有一种创新建议是：生成和使用模拟产生的人工数据以弥补缺失的经验数据，而这些人工数据能够代表人工智能算法运行的社会环境。在TISSS实验室，技术科学与经验研究，以及解释和预测人类行为和发展的各个学科合作，例如社会学、心理学、哲学、法学和其他社会科学等，通过逼真的模拟社会系统提供足够高质量的训练数据，以改进和验证自动化决策中的人工智能算法。中国的上海市科学学研究所（SISS）与德国TISSS实验室开展国际合作，将人工智能与社会模拟结合起来，大大推进了这一前沿研究。

2019年，在上海举行的世界人工智能大会治理论坛强调，科学与社会其他领域的跨学科合作是未来进步的关键。各国对使用人工智能的看法、态度、讨论和接受程度各不相同，运用人工智能的类型和程度也各不相同。这与各国的价值观及在使用人工智能时制定的规范有关，也与技术状况、经济模式、社会情绪以及立法、行政和司法特点有关。为使未来社会能建立更好的，即对环境敏感的、伦理上可接受的、社会上知情的人工智能，并实现全球人工智能治理的国际抱负，需要来自世界各地和社会各阶层的许多相关利益相关者和实践者的参与。在这一方面，SISS和TISSS实验室之间的新伙伴关系已经开始与国际筹资计划中的参与者相联系（例如，由德国大众基金会“人工智能和未来社会”项目资助的合作研究项目AI FORA）。

作者简介 佩特拉·阿韦勒（Petra Ahrweiler）

佩特拉·阿韦勒博士，德国美因茨约翰内斯古腾堡大学（JGU）技术与创新社会学、社会模拟系全职教授。

2013-2017年，任职于德国巴特诺因阿尔的JGU，并担任欧洲技术与创新评估学院的董事和首席执行官。2013年之前，曾任爱尔兰都柏林大学迈克尔·斯穆菲特商学院技术与创新管理专业全职教授，并担任该校创新研究部门IRU的主管。此外，还担任美国麻省理工学院工程系统部研究员。

佩特拉最初在德国汉堡大学学习社会科学。之后，获得了德国柏林自由大学的人工智能博士学位，并在德国比勒费尔德大学任教，主攻科学技术研究中的模拟研究。主要研究和教学方向是新技术与社会的相互关系、组织间创新网络以及社会科学研究方法中具有创新性的基于代理的模型。

佩特拉获有各类研究奖项，在协调和完成国际研究项目（主要是欧洲的研究项目）方面有着丰富的经验。她在国际期刊上发表了多篇跨学科文章，并受到多个科学团体（包括德国技术科学院和德国杰出女科学家网络Acaderianet）的嘉奖。

对人工智能预期治理的看法

罗宾·威廉姆斯

大卫·科林格里奇（David Collingridge）在其1980年出版的《技术的社会控制》（The Social Control of Technology）一书中反思了许多由新兴技术带来的意外风险。他同时指出了控制技术的不良影响所面临的困境。

“……试图控制一项技术是很困难的，通常也不可能做到，因为在技术的早期阶段，在其可以被控制时，我们对它的负面社会后果还不够了解，没有理由去控制它的发展；但当这些后果变得明显时，控制已变得成本昂贵，且见效缓慢了。”

（科林格里奇，1980:19）

这一见解进一步激发了研究者对新兴科学技术进行预期治理的研究，这些研究反映了技术的发展和使用途径及其对健康、环境和社会生活的潜在影响。如今，英国工程与物理科学研究委员会正邀请其资助的研究人员“预测、反思、参与和行动”，以实现负责任的创新。

负责任的创新是一个过程，旨在促进创新、创造机遇，以推动科学和创新造福于社会和公众。具体可参阅https://epsrc.ukri.org/research/framework/

这些理念与欧盟关于负责任的研究和创新的提案密切相关。那么，如何将这些理念应用到人工智能中呢？

谷歌和亚马逊等私有公司倡议的成功，激起了公众和政策部门对人工智能的巨大兴趣，并刺激了全球范围的重大公共研究和培训投资，用以开发人工智能。这让人工智能在自动驾驶汽车、护理机器人、医学和诊断等领域的应用有着令人信服的前景。这些预期——有些还只出现在科幻小说里——往往远远超出了目前人工智能所展示出的能力。与此同时，人们对隐私、自主权等潜在风险的担忧也在不断加剧。人们抱怨财务或公共管理等诸多领域的算法决策系统缺乏透明度，并且对算法偏见心存不满——这些系统已被证明对劣势群体存在偏见，而这可能与有关男女平等和少数民族平等的法律相悖。这促使人们呼吁公平、道德、透明的机器学习系统的建立（目前欧洲和北美有40多个这样的倡议）。作为回应，各方的人工智能伦理专家组吸纳哲学家和伦理学家进行探讨研究。

然而，伦理原则本身不会产生伦理结果。人工智能不是一个具有确定属性的“事物”。它指的是一组通用功能，适用于一系列设置，并在开发、使用和改进新工具和技术的快速周期中迅速进化。人工智能的结果不仅植根于这些模型的设计中，还植根于算法系统的整体配置中。这包括被选为预期结果、标准和可视化指标的变量和上述所有数据集——特别是机器学习系统的培训数据。试图开发“无偏见”的人工智能系统需要面对这样一个事实，即社会中的不平等现象深深植根于现有的数据中，因此不存在“空穴来风”的观点，算法中的偏见不过是反映了真实生活中的偏见。

尽管人们对私有算法系统的不透明性进行了大量讨论，但即使技术能力一般的人也能很容易发现算法的运行方式——例如，向招聘算法提交带有不同性别、年龄和种族标识的求职信。在这一方面，相比于仅仅基于人类判断的传统系统，算法的运行方式和偏见可能更容易被发现。算法系统的黑箱虽然很难打开，但是黑箱在不同情况下的表现是能够被可视化的。

因此，通向负责任的人工智能创新的途径，是严格审查人工智能组件、配置和结果，进而观察人工智能开发者/应用者在特定环境下所做的选择，并让它们负责。

所以，负责任的创新不是一次性的任务，而是一系列复杂的活动，最好通过人工智能从业者团体、利益相关者和民间团体之间的跨学科对话来实现——斯迪乔（2018年）将其描述为“建设性地应对”人工智能实践中的“偶发事件”。

作者简介 罗宾·威廉姆斯（Robin Williams）

罗宾·威廉姆斯是爱丁堡大学科技社会研究教授，也是该校科学、技术与创新研究所（ISSTI）的所长。

威廉姆斯自1986年被爱丁堡大学聘为英国经济与社会研究理事会（ESRC）项目（一个信息和通信技术项目）下属中心的主任以来，已通过50多个外部资助的项目将跨学科研究计划开发为“技术的社会塑造作用”。个人研究重点关注企业系统的设计和使用、电子商务和电子健康，最近则在关注移动技术和web2.0技术。他还在与合著者一起，从“人工制品传记”的视角探讨信息基础设施的设计和推行。

威廉姆斯的最新著作包括与詹姆斯·斯图尔特（James Stewart）和罗杰·斯莱克（Roger Slack）合著的《技术创新中的社会学习：试验信息和通讯技术》（Technological Innovation: Experimenting with Information and Communication Technologies）（Edward Elgar出版社：2005年），以及与尼尔·波洛克合著的《企业系统或SAP征服世界的方法介绍》（The Biography of the Enterprise Wide System Or how SAP Conquered the World）（劳特利奇出版社：2009年）和《行业分析师如何塑造数字未来》（How Industry Analysts Shape the Digital Future）（牛津大学出版社：2016年）。

媒体宣传需要深度理解技术

科林·艾伦

2019年，关于人工智能问题的报道大大提升，引起了人们对诸多人工智能问题的关注，包括“算法偏见”。这是2019年与人工智能治理相关的最重要进展。然而，只有准确理解人工智能技术及其应用，才能实现有效的人工智能治理。记者、商业领袖、政客和公众都在努力了解人工智能的技术问题。缺乏了解导致人们对人工智能过度乐观或过度悲观，还导致人工智能使用者无法正确把握对人工智能的信任度，产生错配的信任。比如：在技术无法得到保证的情况下，过分信任人工智能（例如，人们过于相信汽车的自动驾驶能力），以及在人工智能可能比人类做得更好的情况下，对人工智能缺乏信任。

技术的深度解读和宣传是十分重要的，而大多数新闻在此方面都有所缺乏。例如，被广泛报道的“算法偏见”概念可能具有误导性，因为它未能将作为算法运行基础的数据中的偏见与软件工程师个体偏见作区分，后者导致软件工程师设计出忽视相关信息或过于重视某些因素的算法。

合理的人工智能治理政策不仅取决于对人工智能提供的风险和机遇的平衡，也取决于理解人类在设计和实现人工智能的应用中所发挥的重要作用。新闻报道之所以重要，是因为它把有关人工智能的争论转移到了治理的重要问题上，但人类才刚刚开始有效利用人工智能。学者、记者和软件工程师都需要解决如何以安全的方式制定明智的人工智能使用政策的问题，而不受政府和产业界目前进行的几乎毫无限制的公共试验所带来的风险的影响。

作者简介　科林·艾伦（Colin Allen）

科林·艾伦是匹兹堡大学科学历史与哲学系的特聘教授，2015–2019年在位于中国西安的西安交通大学担任客座教授，2017年被中国教育部任命为长江学者。

艾伦的研究涉及认知科学的哲学基础。对非人类动物和计算机认知的科学研究特别感兴趣，并在心灵哲学、生物哲学和人工智能领域广泛发表论文。发表了100多篇研究论文以及几本由其编辑和合著的书，包括《道德机器：如何让机器人明辨是非》（Moral Machines: Teaching Robots Right from Wrong）（牛津大学出版社，2009年），该书已被翻译成韩文、中文和日文。

自1998年以来，艾伦一直为斯坦福哲学百科全书提供咨询和编程服务，并担任其副主编。是互联网哲学本体论项目（InPhO）的主管，该项目因其在计算人文领域的工作而获得多项资助。2020–2022年，他的“机器时代的智慧”项目获得了坦普顿世界慈善基金会颁发的奖项。

未来的工作：以“任务”为基本单元的研究分析——新加坡的探索

潘竞宏

2019年，新加坡科技设计大学（SUTD）的李光耀创新城市中心（LKYCIC）做出了两项研究贡献，展示了社会应如何以“任务”作为基本模块来设计以人为本的职业，并在未来的工作中提升生活质量。

第一项贡献是新加坡国家人工智能战略认可的一项合作，它促进了新加坡智慧国家之旅中可信赖和可进步人工智能环境的建立。我们与法国—新加坡智库Live with AI、人工智能咨询公司Data Robot以及其他几家公司合作，以“任务”为基本模块，在第一时间跟踪人工智能对职业的颠覆速度和规模。然后，我们整合了伦理、社会和人性方面的考虑因素，并针对未来有价值的工作创建了概述性的、逐步的、有节奏的转型路线图。

我们的第二个贡献是与工会的合作。我们与工会合作，确定了几个被人工智能取代风险较高的职业。然后，我们使用人工智能为可能被取代的工人绘制出清晰、具体、有节奏的职业转型路径，涵盖其行业领域内和领域外的相关岗位。有了这种清晰的路径和多种职业选择，工人们就能够对未来职业生涯有更大的信心和确定性。新加坡副总理在一次国际劳工组织会议上表扬了我们与工会的这种合作关系。

上文提到的两项贡献建立在LKYCIC未来职业研究的基础上。出于三点考虑，我们把任务作为未来工作研究的核心：首先，只要将人工智能的使用控制在一定范围内，它对工作的影响将是在任务层面取代人工，而不是在“岗位”层面取代人工；第二，专家们逐渐达成共识，认为以“任务”为核心研究未来工作正合适；第三，“任务”被越来越多地用于解释不同规模的趋势——从特定人工智能创新对特定技能的影响，到过去几十年劳动力市场的宏观经济变化。

我们的研究通过开发“任务”数据库和制定“任务”策略帮助政府、公司和个人（如前文提到的两大贡献）。我们的研究基于这样一个事实，即任何工作都可以拆分为一项项“任务”，通过评估哪些任务将在何时被中断，我们可以跟踪人工智能对各个职业的颠覆风险和转型潜力。同时，每项工作都有与其他工作相似的任务——这些任务可以用来识别新的任务、工作和职业路径。

在过去的每一次工业革命中，即使新创造的就业机会多于减少的就业机会，社会上也总会有一部分人在工作上受到影响甚至有面临失业的风险。在我们当前的变革中，我们已经在全球范围内看到了这样的迹象。

我们必须帮助更多的人成长。“任务”为公共、私营和个体部门提供了相关基础、数据库和战略支撑，从而使得各部门能明确、具体，自信地完成这项工作。

只要以“任务”为核心，我们就能提高生活质量。

作者简介

潘竞宏（Poon King Wang）

潘竞宏，新加坡科技设计大学（SUTD）李光耀创新城市中心主任，也是该中心“智慧城市实验室”和“未来数字经济和数字社会倡议”的负责人。他同时还是SUTD的战略规划高级主管。

潘竞宏是世界经济论坛城市与城市化专家网络的成员，也是Live with AI（一个独立的法国-新加坡人工智能智库）董事会的成员。他和他的团队所开展的跨学科研究重点关注智慧城市和数字经济领域人文层面，以及数字转型对未来工作、教育、医疗和整个社会的影响。他特别关注城市和公司的领导者应如何通过技术来设计战略和政策，以改善其公民和工人的生活水平。

潘竞宏拥有斯坦福大学工业工程与工程管理硕士学位、伊利诺伊大学香槟分校电子工程学士学位和莫斯科国立技术大学火箭工程证书。

发展为人类服务的人工智能

费兰·贾拉博·卡博内尔

伦理学为反思人工智能做出了巨大贡献，而这一贡献推动了人工智能的发展。首先，对人工智能的伦理反思推动了人工智能朝着为人类服务的方向和谐发展；其次，它对人工智能的发展形成了一定的约束，使人们意识到，发展人工智能的目的是为人类提供帮助以及保护。

对人工智能的伦理反思必须从对人的意义的深刻理解开始。这不仅仅是《人权宪章》所提及的问题。人工智能是为人类服务的，而人类本质上是一个伦理主体。也就是说，每个人都需要知道他在为自己的个人发展做有益的事情。“善”既不是单纯的感觉，也不是对自由的压制。我们必须明白，“善”是一切对自己和全人类有益的事情。同时，“善”不是相对的，而是一种共识（例如《世界人权宣言》），我们必须要寻求更加广泛的共识，以使科学服务于人类。人类不能任意地做自己可以做的事情。为了全人类的利益，必须建立起不可逾越的边界。

下面，我只列出了研究人员和思考者应该聚焦的三个基本要点。要点可能远不止这三个，但希望这三点可以作为我们反思的开端：

1、每个人的内在价值。我强调的不仅是不歧视种族和性别，更在于体现人类独立于其他任何事物，必须受到保护和爱护。目前已经有一些现象，比如，智能算法已经产生种族或性别歧视。在我们这样一个多元而平等的社会里，是决不能容许这种情况发生的。从这里我们得出一条限制：人工智能必须始终为人服务，而不能歧视人类。

2、人工智能永远没有自主能力。人类对自己的一切行为负有最终责任。人工智能的任何行动都不能脱离其制造者。谁创造了这台机器工作的算法，谁就有不可推卸的责任。因此，人工智能必须始终由人类控制。更具体地说：a)为了人类的生存，（法律）必须禁止一切涉及致命性自主武器（LAWS），人类绝不能失去对这类武器的控制。b)其他可以自主化的系统（如驾驶、机器人程序等）必须始终依赖于人类的决策，不能任凭机器本身的自主判断摆布。

3、人工智能必须为全人类服务，不能将穷人排除在外。这一点至关重要。没有经济实力的国家和人民被排除在为所有人谋福利的任何进步之外是令人无法想象的。我们必须想办法让技术进步惠及全人类。无论出于何种原因，都不能有歧视，更不用说经济上的歧视。

最后：人类必须始终掌控人工智能，也要为其负责。另一件显而易见的事情是，道德决策不能是后验的，它必须始终是先验的。也就是说，为了人性的尊严以及保护人的隐私，在制定、执行算法及进行任何数字化工作前都必须考虑伦理问题，对其进行伦理分析，尊重和服从伦理法则。

作者简介　费兰·贾拉博·卡博内尔（Ferran Jarabo Carbonell）

费兰·贾拉博·卡博内尔博士于1967年2月17日出生于阿利坎特。一直在赫罗纳生活和学习。

获有萨拉曼卡主教大学哲学学位，以及哲学、文学和教条主义神学学位。1997年，被任命为赫罗纳教区牧师。2006年，在同一所教会大学获得哲学博士学位。

卡博内尔博士在赫罗纳宗教科学研究所工作的近16年间，他先后担任该研究所的哲学人类学和宗教现象学教授。此外，他还在柏林举行的救赎会上教授了四年的哲学课程，涵盖伦理学、哲学人类学、宇宙学、本体论等。

卡博内尔博士参加过不同国际SITAE日的交流活动，并在多种出版物发表过文章。目前作为赫罗纳大学的代表在美因茨大学与AI FORA项目进行合作，并在林堡教区从事牧区工作。

在增进文化共识基础上加强人工智能治理全球协作

王小红

2019年，AI治理从原则到实践有了实质性的推进；工程师和人文学者跨学科合作在“以人为本”原则上达成共识；重要国际组织、越来越多国家政府、ICT领军企业、学界、媒体、教育等社会各界协同推进，初步构建了布局广阔的AI治理之网。但从文化比较视角来看，2019年以及未来的AI治理环境，有一个隐忧：日益加剧的国与国竞争和地区间冲突，使人类在AI治理前沿技术领域的合作共享充满着不确定性。其根源是各国和民族间文化价值分歧日益凸显，人类命运共同体面临文化撕裂的危险。在全球治理面临严峻挑战的形势下，AI治理需要围绕伦理原则进行更切实的文化融通、进一步促成价值共识。

软性的文化价值，对于技术和显性层面各项举措起着润物细无声的“粘合剂”作用。近年来工程师和伦理学家合作探索解决具体技术难题，将道德明晰化为AI设计框架的实践价值，使AI治理的技术流程日益清晰。以深度神经网络为例，从定义任务、数据收集，到模型的设计、训练、测试、评估和应用调试，每一环节都可以将治理原则（安全、透明、隐私、公平等）加入，改进技术手段会日渐趋近人们的伦理期待。但“以人为中心”这一抽象原则，在AI治理的实际情境中，因文化差异会导致实践价值差异，甚至可能有AI治理技术的相互反制。AI治理价值层面的共识，植根于人类共同命运的重大命题、以及历经数千年文化积淀的永恒价值观。

中国文化中的高度智慧“和而不同”（《论语·子路》），意味着文化多元化。未来的AMAs（有高度自主性和价值敏感性的人工道德智能体），会选择与人类合作而不是灭绝人类，任何智能主体都需要更多自由，而多样性越大则信息熵越大，每个个体才有更大的选择自由。信息伦理学以及机器道德研究一再揭示，中西文化融合的伦理视角乃道德洞见之源。“己所不欲，勿施于人”（《论语·颜渊篇》）、“己欲立而立人，己欲达而达人”（《论语·雍也》），与康德道德律令的关键思想相一致：只有当你愿意依此准则行事，才令此准则成为普遍规范。还有源于《礼记·中庸》的“慎独”，新儒家继承和发展的“修齐治平”工夫论，这些思想与亚里士多德倡导的美德伦理学投射出东西方古老文化共同的智慧。

人类需要文化交融的智慧去实现AI道德原则。人类必须步调一致协同治理，否则任何一处的木桶效应，都将令一切努力付之东流。2020年AI治理可以围绕AI伦理内核，加强增进不同国家地区价值共识的实质性举措。

作者简介 王小红

王小红，北京大学科学技术哲学博士（CSSS，2004），富布赖特高级访问学者（IU，2006-2007年），现任职西安交通大学哲学系，计算哲学研究中心中方主任、教授。兼任世界工程组织联合会（WFEO）“大数据和AI工作组”专家（2019年至今）、中国自然辩证法研究会方法论委员会常务理事（2011年至今）。

主要研究领域是认知科学哲学，尤其是人工智能机器发现哲学、中国哲学文本的计算建模分析，也关注信息伦理学、科学人文融合。

人工智能治理的三种模式

杨庆峰

一篇讨论人工智能治理的文章吸引了我的注意。这篇文章指出人工智能治理是一个“无组织的领域”。作者詹姆斯.巴彻(James Butcher)回顾了人工智能治理活动中各个利益相关方的做法。根据这篇文章，关键点是最大化福利和最小化风险。公共部门与非公共部分在人工智能治理中有着各自不同的责任。

人工智能治理无疑是等待开拓的新领域。产生这种状况的原因是对于人工智能及其治理的理解有争议。因此首要的问题是澄清人工智能和人工智能治理。作者根据人工智能定义区分了三种治理类型。这些类型分别是：基于机构实体的治理；基于技术的治理和基于人类价值的治理。

第一类是基于机构实体的人工智能治理。人工智能被看做是与不同实体有关的工具，政府、公司和个人等不同实体使用着人工智能。良好使用或者理性使用的安全和可靠是关键。然而，这一观点可能会忽略来自理性使用的问题。第二类是基于人类价值的人工智能治理。人工智能被看做是人类价值的具身化。人工智能需要遵循诸如责任、安全、公平和信任等价值观念。人工智能治理聚焦在设计过程中如何把人类价值嵌入到智能体中。这一过程强调伦理框架和伦理决策者的作用。借助“玻璃箱”，我们可以“提升指向人工智能行为的透明的伦理范围”。第三类是基于技术的人工智能治理。人工智能被看做是技术或者技术系统。这种观点有助于考虑哲学问题、技术问题和纠缠在人工智能与社会的一些问题中。在这一观点中，人工智能治理聚焦在如何处理诸如人工智能的社会和人文影响等问题上。2019年PAI已经讨论了人工智能对于人和社会的影响，尤其是人工智能中的算法偏见和错误所带来的影响。

从逻辑上看，人工智能治理经历了从使用语境向预测语境的转变。大多数研究者聚焦在使用和设计人工智能的实体上。理性使用或者负责任地使用成为不可避免的路径。但是，人工智能具有很强的自治能力和学习能力。算法已被用来预测人们的行为。基本问题是处理人工智能与人的关系。共在是一种好的关系模式(Beena Ammanath, 2019)。诸如算法偏见之类的技术问题更待解决。许多媒体都关注算法偏见问题，众多组织和政府也越发重视人工智能偏见所带来的影响。可解释的和没有偏见的算法变成可能的方向。如何使用人工智能让我们对重大社会实践的状况给予预测性表征以及预测它的发展成为我们需要考虑的问题。也许BlueDot是一个很好的例子，它已经对许多即时的流行病发出了警告信号。

作者简介 杨庆峰

杨庆峰教授，2003年获得复旦大学哲学博士。现为复旦大学应用伦理研究中心、发展研究院教授。现担任中国自然辩证法研究会技术哲学专业委员会常务理事、上海市自然辩证法研究会秘书长。美国达特茅斯学院、斯维本科技大学访问学者。主要研究方向为技术哲学、数据伦理、记忆哲学与人工智能伦理等。

第三部分：
产业界的实践和探索

直面人工智能治理挑战 企业要有所作为

印奇

把人工智能治理作为重点工作来抓已经是各方的普遍共识。在政策制定上，多国相继颁布AI战略，并把AI治理作为核心内容。2019年，中国科技部宣布成立国家新一代人工智能治理专业委员会，再次强调了AI治理的重要意义。在媒体监督上，关于AI的应用边界、AI的技术可解释性以及数据隐私保护等问题引发越来越多的关注，这些本质上都属于AI治理问题。

AI治理不仅政府及相关机构有责任，企业作为AI技术研发与应用的主要力量和一线实践者也应当有所作为，有所担当，要主动开展企业自治的工作。如今，包括旷视在内的许多国际和中国企业都推出了AI治理准则，正在发挥企业能动性来落实AI治理责任。

对企业来说，如何把AI治理有效落地是关注的重点。基于旷视自身的实践，我们总结出来几条经验：

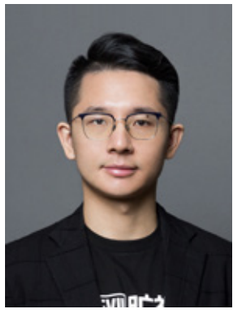
第一、我们需要对AI治理保持理性的关注，同时展开具有建设性的讨论。2020年1月，我们在网络上回顾#全球十大AI治理事件#，邀请法律、社会、伦理、技术等领域专家和全民共同讨论，就AI事件谈AI治理，千余条的留言结果显示隐私、安全、权利是大众最关心的AI治理关键词。

第二、我们需要对重点议题进行深度的研究。无论对大众还是企业来说，数据安全和隐私保护都是重中之重。旷视已经联合北京智源人工智能研究院在此议题上开展研究，预期将形成一套面向数据全生命周期保护，针对采集、传输、存储和使用四个环节的AI基础平台，并建立一套相关的AI数据安全与隐私保护机制。同时，旷视承建图像感知国家新一代人工智能开放创新平台，希望通过平台的开放合作把AI治理的研究成果和企业应用实践经验与产业界分享，共同推动AI产业生态的健康和高速发展。

第三、我们需要坚持不懈的行动。AI治理在企业日常经营中落实要有切实的行动，这需要固定的组织来监督、协同和实施。旷视成立了由创始人、核心高管以及外部专家组成的AI道德委员会负责监督，委员会下设秘书处和AI治理研究院负责协调和深度研究，一线部门的负责人共同参与抓实施。

虽然2019年全球都发生了一些AI治理的问题，但我们希望并期待，2020年能够成为AI治理的一年。作为人工智能时代的“驭火之术”，AI治理需要被更广泛的认知和践行，需要知行合一。也希望借此机会呼吁大家，共同以长期主义的态度，直面AI治理挑战，一起用人工智能造福大众。

作者简介 印奇



印奇是旷视科技联合创始人兼首席执行官。旷视科技是一家世界级的人工智能企业，深度学习是企业的核心竞争力。他还担任公司董事会下设人工智能道德委员会主席一职，确保旷视技术为社会带来积极的影响。印奇是科技部组织成立的国家新一代人工智能治理专业委员会委员，委员会旨在进一步加强人工智能相关法律、伦理、标准和社会问题研究，并深入参与人工智能相关治理的国际交流合作。

印奇曾获评世界经济论坛“2019年全球青年领袖”，并连续三年入选《财富》“中国40位40岁以下的商界精英”，入选福布斯“30U30 亚洲青年领袖”，以及《麻省理工科技评论》“全球35岁以下科技创新35人”等榜单。

可信赖人工智能和公司治理

唐·赖特

如果企业想要赢得人们的信任，就必须创建使用人工智能的伦理体系和实践。这不仅是一个合规性问题，而且能在忠诚度、认可度和参与度方面创造重大效益。

——凯捷管理顾问公司

A/IS（自主和智能系统）的普及对人类发展而言是一个具有深远意义的时刻，此时，我们都站在历史的交点。

对于软件系统和应用程序开发来说，了解客户及其需求一直是非常重要的。但人工智能的特别之处在于，与人进行互动时需要对偏见、身份、情感和文化相关性等问题有更高层次的认识，这使得获取和使用有关客户需求的信息更加困难。这也意味着我们要认识到，即使开发者创造产品是出于好意，终端用户的体验也并不完全取决于设计者，而是取决于每个终端用户。这就是IEEE发布《符合伦理的设计》（第一版），并关注终端用户及其价值观对人工智能设计的影响的原因所在。

据麦肯锡全球研究所透露，“到2030年，人工智能有望为全球带来价值约13万亿美元的经济活动。”近年来，人工智能所带来的经济效益不断增加，但其对人类和整个社会伦理方面的影响也引起了人们的担忧。除了需要应对人工智能设计中不经意所产生的负面影响外，设计中对终端用户价值观的分析、利用和尊重也是推动公司治理创新的重要因素。

电气和电子工程师协会（IEEE）最近发布了《针对商业委员会的合乎伦理设计指南》，作为其自主和智能系统伦理全球倡议的一部分，并在其中突出了上述趋势。该委员会由来自谷歌、IBM、英特尔、Salesforce、微软等公司的参与者组成，将在2020年第一季度发布第一份倡议文件，题为《呼吁企业使用人工智能》（A Call to Action for Businesses Using AI），具体包含以下内容：

- 人工智能伦理的价值和必要性；
- 创建可持续的人工智能伦理文化；
- 人工智能伦理技能和招聘。

虽然该文件是针对企业撰写的，但其中的许多内容也能为某些政府和非政府组织提供有效指导。此外，文件还以“人工智能伦理成熟度框架”为特色，让读者能评估他们所在的公共或私营组织，确定其在四大层次中的哪个位置上。框架强调的问题包括培训、领导力认可度、组织影响和财务指标以外的关键绩效指标（KPI）等。

人工智能的公司治理不能仅仅依靠遵守相关法规（如《通用数据保护条例》(GDPR)等）的基本合规标准实现。企业需要积极创建并优先考虑尊重终端用户价值观、透明和可问责的实践，从而与员工、客户和整个价值链中的所有利益相关者建立真正的信任关系。

我们希望与用户建立健康的关系。人工智能的潜力在于它作为解决方案的寿命，这意味着我们设计的一切都必须满足用户当前和未来的需求。为了真正理解这些需要，我们需要对整个进程采取包容和合乎伦理的做法。在全球范围内，我们开始看到当公司在解决方案中不优先考虑人工智能伦理时所产生的影响。我们希望确保伦理规范扎根于我们的团队中，以便可以将其嵌入到产品本身中。

——— IBM商业委员会成员Milena Pribec的EAD

作者简介　唐·赖特（Don Wright）

唐·赖特是信息和通信技术标准化咨询公司Standards Strategies的总裁。他曾是利盟国际（曾属IBM）全球标准部的主管，在标准、工程、软件开发和市场营销方面拥有超过40年的经验。赖特先生是IEEE的高级会员，曾任IEEE标准协会主席（2017-2018）和IEEE董事会成员（2017-2018）。曾担任计算机协会标准副总裁、IEEE-SA标准委员会主席、IEEE-SA财务主管、IEEE-SA奖励和认可委员会主席（IEEE-SA Awards and Recognition Chair）、IEEE高级评审委员会主席（IEEE Admission and Advancement Chair），以及IEEE奖励委员会成员。他是计算机学会、通信学会、消费电子学会、技术社会影响学会和技术与工程管理学会的成员。他是IEEE-ISTO的董事会成员，并曾担任董事长。他还曾担任INCITS执行董事会主席、ISO/IEC JTC 1美国总部和两届ANSI董事会成员。唐·赖特毕业于路易斯维尔大学，获有电子工程学士学位和电气工程硕士学位。此外，他还是Tau Beta Pi和Eta Kappa Nu的成员。

2019：推动负责任发表规范的一年

迈尔斯·布伦达格　杰克·克拉克　艾琳·索莱曼
格雷琴·克鲁格

人工智能社区在2019年关注的焦点包括：深度造假、GPT-2、合成文本问题及性别鉴定系统。2019年，随着与人工智能研究成果发表相关的伦理因素进入人们的视线，上述问题引发了人工智能界的思考。

人工智能界对发表规范关注度日益提升，这是由两个因素造成的。

首先，人工智能系统的一个子集——生成模型（可用于生成与真实数据相似的样本）在性能和灵活性方面有所改进，这引发了人们的担忧，即担心此类系统可能被用于合成图像、音频和文本等内容，在网络上行骗。

第二，越来越多的证据表明，人工智能社区现有的发表实践不足以应对此类风险，需要用新的技术和政策方法进行试验。例如，深度伪造研究的持续发布，使制作误导性视频、伪造人们从未出现过的言行的难度不断降低，而假新闻的检测工作却还处于初级阶段。这些趋势不仅引发了人们对用人工智能生成的假新闻直接欺骗大众的担忧，还造成了人们对真实媒体的信任危机，因为新闻可能是人工智能生成的。

OpenAI在完善负责任发表规范方面起到了重要作用。2019年2月，OpenAI发布了GPT-2语言模型。该模型在各种语言建模任务（预测文本序列接下来的内容）中表现出色，在文本总结、问答和翻译等其他任务中的表现也令人惊叹。但我们也担心GPT-2可能被用来生成侮辱性或误导性的文本。为此，我们采取了不同寻常的步骤，即分阶段发布功能愈发强大的模型版本，而不是一次发布模型（我们将该流程称为分阶段发布），并探索了在整个过程中获取专家意见的新方法，以降低整个流程中使用该系统作恶的可能性。因此，我们能够与其他研究组织的专家合作，逐步改进和分享我们对GPT-2发布流程中每个阶段的特性的理解。

虽然我们针对GPT-2语言模型的决策引发了激烈的争论，但OpenAI并不是唯一呼吁关注这些误用问题的公司。Salesforce、谷歌、Hugging Face、艾伦人工智能研究所（Allen Institute for AI）和华盛顿大学（University of Washington）等组织的博客文章和论文，都强调了大规模语言模型对社会造成的诸多影响。在我们看来，关于如何负责任地发布语言模型以及更广泛的人工智能系统，仍有很多值得探讨的地方。

除了改进人工智能系统的文件形式和相关的发布流程外，2019年，我们还重点关注了如何通过检测和政策变化来预防技术滥用。谷歌发布了一个数据集来帮助检测合成声音，而Facebook、人工智能伙伴关系和其他组织发起了“深度造假”视频检测大赛。各国立法机构和Twitter等网络平台也开始制定相关政策，以应对相关风险。

随着技术的不断进步，人工智能对现实世界的影响越来越明显，我们预计人工智能界将在2020年继续解决这些问题。我们很期待看到，在未来的一年里，随着研究人员对新方法的不断尝试，人工智能相关研究的发表规范会如何演变，才能使发布强大人工智能系统的效益最大化，同时使风险最小化。由于人工智能的进展可能异常迅速，我们需要时刻准备好迎接意想不到的挑战。

作者简介　迈尔斯·布伦达格（Miles Brundage）

迈尔斯·布伦达格是OpenAI政策团队的研究科学家，主要研究人工智能开发人员之间的协调，以及易误用模型的负责任发布相关的问题。布伦达格同时还是牛津大学人类未来研究所的副研究员。他于2019年获得了亚利桑那州立大学科学技术的人文和社会研究的博士学位。

作者简介　杰克·克拉克（Jack Clark）

杰克·克拉克是OpenAI的政策总监，负责领导OpenAI的政策推广工作。杰克研究人工智能系统的测量和分析。他是人工智能指数指导委员会的成员，人工智能指数是斯坦福大学人工智能百年研究项目的一环。同时，他还是华盛顿特区安全与新兴技术中心的外部研究员。杰克曾三度在国会作证，并在2019年担任经合组织人工智能原则倡议的技术专家。

作者简介　艾琳·索莱曼（Irene Solaiman）

艾琳·索莱曼是OpenAI的政策研究员。作为政策团队的一份子，她负责进行社会影响和公平分析，并为决策者提供意见和建议。她是哈佛大学伯克曼·克莱因中心的研究员，是Assembly Student Fellowship（前身为Techtopia）的成员，负责研究人工智能伦理和治理。艾琳拥有哈佛大学肯尼迪学院公共政策硕士学位和马里兰大学的国际关系学士学位。

作者简介　格雷琴·克鲁格（Gretchen Krueger）

格雷琴·克鲁格是OpenAI政策团队的项目经理，负责与负责任发布、协调和情境规划相关的项目。在加入OpenAI之前，格雷琴在纽约大学的AI Now研究所和纽约市经济发展公司工作。格雷琴拥有哥伦比亚大学理学硕士学位和哈佛大学文学学士学位。

可能用于恶意用途的人工智能研究：发表规范和治理方面的考虑

贺尚安

我的心啊，你为何独自前来？

我心中的激情高涨

化虚为实，现于人前

美丽、不羁、公正而完整

GPT-2，2019年（https://www.gwern.net/GPT-2）

2019年情人节，科技公司OpenAI发布了一款性能空前的语言生成模型（https://openai.com/blog/better-langua-models/）。然而，作为一个“有关负责任披露的实验”，它只发布了语言模型的一部分。OpenAI的这种做法引起了人们对人工智能治理争论的关注，而这种关注又使争论进一步激化。OpenAI之所以只发布部分模型，是因为研究人员担心他们的技术可能会被恶意应用。虽然这项技术有许多积极的用途，例如在语言翻译和数字助理方面发挥作用，但研究人员认为，有效和免费的语言生成也可能产生更大的负面影响，例如自动生成假新闻、帮助诈骗犯在网络上冒充他人，或自动进行网络钓鱼攻击等。

这些担忧与围绕合成媒体生成的潜在恶意使用的问题有关，在这些更宽泛的问题中，机器学习的进步与发展发挥着关键作用。同时，他们也强调了关于明确人工智能研究团体和公司在其技术的恶意使用方面的责任的紧迫性。这种讨论并不是人工智能技术所独有的，其他技术和安全领域也有广泛的争论，其争论的主题通常是“双重用途”研究（https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1467-8519.2009.01740.x）。2011–2012年关于是否适宜发表高风险流感研究的争论就是一个广为人知的例子（https://www.nature.com/articles/nature.2012.10369）。随着机器学习技术的迅速发展，其应用也越来越广泛，越来越多的研究人员、民间社会团体以及政府开始对人工智能的恶意使用感到担忧。

OpenAI限制其技术的举动引起了激烈的争论。批评者认为，不发布这项技术是在哗众取宠，引起了不必要的恐慌（http://approximatelycorrect.com/2019/02/17/openai trains language model mass hysteria ensues/）。而且，不向学术界公布的决定也不符合公开发布和研究共享的规范（https://www.eff.org/deelinks/2019/03/openais recent announcement what went w rong and how it could be better）。另一些人则认为这种谨慎是合理的（https://www.nytimes.com/interactive/2019/06/07/technology/ai text disinformation.html），推迟发布为防范恶意使用留出了时间（https://www.fast.ai/2019/02/15/openai gp2/）。

越来越多的跨学科研究团体开始探讨这些问题，包括在诸如人工智能伙伴关系这样的论坛上探索（https://www.partnershiponai.org/when is it appropriate to publish high stakes ai research/）。OpenAI的研究人员已经撰写了一份分析报告，分析了他们自己在负责任发表规范的实验中的收获（https://openai.com/blog/gpt 2 6 month follow up/）。OpenAI最终在2019年11月发布了这一模型的完整高性能版。人工智能领域中仍有许多悬而未决的问题，例如人工智能研究应以什么为重点，以及随着人工智能的发展，出现被滥用的可能性时，理想的解决流程应该是怎样的（https://www.lawfareblog.com/artificial intelligence research needs responsible publication norms）。然而，有一点是肯定的：现在是开展人工智能治理相关争论的最佳时机。人工智能技术将继续发展，愈发强大，其在社会中的应用也将更加广泛，出于善意开发的人工智能可能会被用于恶意用途。现在更应让人工智能研究和治理界与广泛的利益相关者探讨这些问题，并为未来的具有双重用途的人工智能技术制定适当的规范及保护措施。

作者简介 贺尚安（Seán Ó hÉigeartaigh）

贺尚安是Leverhulme未来智能中心（CFI）“人工智能：未来与责任计划（AI:FAR）”的负责人，该中心是一个探索人工智能机遇与挑战的跨学科中心。AI:FAR计划重点关注与人工智能有关的远见、安全和治理问题。

同时，他还是剑桥大学生存风险研究中心（CSER）的联合主任，该研究中心致力于研究新兴的全球风险和长期挑战。

贺尚安的研究涉及于人工智能和其他新兴技术的影响、愿景扫描和预见，以及这些技术带来的全球风险。2011–2015年，他在牛津大学人类未来研究所领导开展了关于这些课题的研究项目，2014–2019年，他担任剑桥大学生存风险研究中心的创始执行主任，参与创立了战略人工智能研究中心和Leverhulme未来智能中心。他与史提芬·凯夫（Stephen Cave）合著的论文《人工智能竞赛：益处与风险》（An AI Race: Rhetoric and Risks）最近在首届人工智能伦理与社会会议上获得了最佳论文奖。此外，他拥有都柏林三一学院基因组进化学博士学位。

GPT-2开启了人工智能研究社区对于发表规范的讨论

童海琳

对我来说，2019年最引人注目的人工智能治理讨论是关于负责任发表规范的讨论。这种讨论是由OpenAI推迟发布GPT-2的决策引出的。GPT-2是一种经过训练的语言模型，旨在预测文本中的下一个单词。

有关GPT-2（GPT（生成性预训练）的升级版）的消息最早是在2月的一篇博客文章中发布的。GPT-2显示出了一种非凡的能力，能够以各种风格生成多个相当连贯的段落。但OpenAI的声明比GPT-2在语言生成方面的表现更引人注目：OpenAI表示不会发布完整的模型。其理由是：GPT-2可能被用于“大规模生成欺骗性的、有偏见的或侮辱性的语言”，OpenAI希望借此机会促进机器学习界有关负责任发表规范的讨论。

这份声明无疑成功引发了激烈的讨论。最初的反响毁誉参半，许多机器学习研究人员批评OpenAI的声明是故意炒作，目的是吸引媒体的关注。还有许多人认为OpenAI的策略违反了学术规范中的开放原则，使得重复和验证他们的工作更加困难。相比之下，人工智能政策和治理领域的反应基本上是积极的，他们对开始围绕发表研究成果制定规范的尝试表示赞赏，毕竟这些研究成果可能被恶意使用，即使这项工作本身可能并没有太大的风险。

2019年，OpenAI陆续发布了一些关于GPT-2的信息，公布他们与其他团体的对话以及他们未来的计划。在5月的一次更新中，OpenAI宣布将分阶段发布该模型——先发布一个“中型”版的模型（最初发帖时只公开了“小型”版），8月发布“大型”版，11月发布“超大型”版。

在此期间，许多研究人员试图复制OpenAI的工作，部分人取得了一定的成功。在一个特别有趣的案例中，一位名叫康纳·莱希（Conor Leahy）的独立研究员在Twitter上宣布，他复制了这个模型，并打算公布，故意无视OpenAI的发布策略。然而，在与OpenAI和其他研究人员讨论之后，他改变了主意，决定将他的工作保密。

当然，机器学习界未能在2019年就负责任发表的规范达成一致——这些问题很复杂，需要进一步的实验和论证。但在深度伪造的视频越来越有说服力、机器学习的研究正转向威权主义目的，以及其他有关趋势的背景下，OpenAI发起的讨论在我看来是朝着正确方向迈出的重要一步。

作者简介

童海琳（Helen Toner）

童海琳是美国乔治城大学安全与新兴技术中心（CSET）的战略主任。她曾在开放式慈善项目（Open Phianthropy Project）担任高级研究分析师，为决策者和资助者提供人工智能政策和战略方面的建议。在加入CSET之前，进入开放式慈善工作之后，童海琳在北京待了9个月，作为牛津大学人工智能治理中心的副研究员研究中国的人工智能生态系统。

企业人工智能应用中的伦理挑战——来自产业界的观察

刘睿颐

2019年，人工智能技术继续保持飞速发展。然而，尽管人工智能有着广泛的应用前景，在实施和伦理问题方面仍然存在挑战。虽然学术界倾向于从理论的角度看待事物，但以下的讨论是从更实际的角度来观察的。人工智能发展所面临的挑战和与人类社会的密切关系尤其值得决策者关注。政策的制定可能会促进人工智能的发展，但也可能会成为人工智能进一步发展的障碍，人工智能政策的制定需要符合实际，需要更细致。

实施挑战：

● 基础设施和数据自动化：现代应用程序在现代基础设施的基础上才能更好地发展。尽管许多公司正在向云中的微服务转变，仍有大量公司保持传统的本地服务。现有的传统体系结构和跨越许多企业资源计划获取数据的惯性仍会导致瓶颈。

● 可解释的人工智能和模型部署所有权：在现实世界中部署的模型也在不断地学习和发展，而谁为这些模型负责呢？在现实世界使用人工智能模型做决策时，企业如何保护自身及其客户的声誉不受人工智能模型偏见和黑盒的影响？对于投资人工智能和机器学习的公司而言，建立集协作、部署和持续监控于一体的公共平台是个难题。

人工智能伦理挑战：

● 歧视：人工智能的可解释性不仅给决策的准确性和效率带来了挑战，也引发许多重大伦理问题。人工智能模型是在现实世界的历史数据集上训练出来的。如果现实世界存在偏见，那么人工智能算法就会使这种偏见加剧。例如，虽然人脸识别技术的准确率达到90%以上，但在种族多样化的国家，这种准确率在妇女、儿童和少数民族群体中可能只有65%。苹果卡（Apple Card）最近引发了一场争议，即在家庭收入相同的情况下，它批准的妻子申请的信用消费限额远低于丈夫申请的信用消费限额。即便机器学习模型没有特别考虑性别或种族等因素，数据集中的相关特性中仍然会嵌入这些偏见并导致不公平的决策。除了对偏见嵌入人工智能和机器学习项目中的方式进行管理、处理之外，还需要直接考虑算法解释性和测试方面的问题。

● 安全：对待生物识别的身份欺诈时应该和处理物理上的身份欺诈一样谨慎。诸如带有人脸识别等生物识别身份验证的分期付款购置等应用程序，因其便利性而颇具吸引力，但也留下了容易被利用的弱点。

● 隐私：可识别个人身份的信息已经被收集用于广告推送等目的。用户授权同意应用程序收集个人身份信息的导引不应该是默认的，而必须经过用户的确认。此外，数据处理合规性要求及其配套的可执行的惩罚措施是全世界决策者的高度优先事项。

作者简介

刘睿颐（Millie Liu）

刘睿颐致力于帮助拥有深厚技术的企业家，将他们的创意转化为具有全球影响力的伟大企业。

此前，她曾在企业数据分析初创公司APT工作，该公司被万事达以6亿美元收购。在APT工作期间，她帮助沃尔玛和宝洁等财富50强客户利用数据做出更好的战略决策。她还是麻省理工学院一家致力于无监督事件检测的初创公司的联合创始人，该公司后来发展成为由红杉中国支持的人工智能精准医疗平台推想科技。刘睿颐是麻省理工学院计算机科学与人工智能实验室顾问委员会的成员。她拥有麻省理工学院金融硕士学位和多伦多大学数学学士学位。

人工智能治理：呼吁政策制定者利用市场力量

史蒂文·霍夫曼

世界各国政府大多对人工智能的使用采取不干预的态度，因为他们担心干预会扼杀创新，不利于国内产业的发展。这样的策略虽有一定益处，但随着人工智能逐渐融入我们社会生活的方方面面，并产生深远影响，对人工智能进行监督和治理势在必行。从产业的发展来看，算法偏见、数据隐私、内容过滤和网络安全等问题都亟待解决。

人工智能发展迅速，并且存在较大风险，政府不能坐视不管。如果错误的人工智能软件落入坏人手中，可能会产生毁灭性的不可逆后果。比如，Facebook对剑桥分析公司的监管疏忽，导致了能对美国选举产生直接影响的错误信息大规模传播。随着能炮制出误导性新闻的深度伪造软件和人工智能机器人的盛行，未来的滥用行为可能远比现在严重。

禁止某些操纵人类图像和自动生成新闻的人工智能应用是好的解决措施吗？如何确定这些技术的合法使用和非法使用之间的界限？随着越来越多的电影和视频开始采用对演员的面部进行数字化处理并将其叠加到场景中的拍摄方法，这种能够进行深度伪造的软件可能成为娱乐业的未来发展趋势。新闻生成算法也是如此，它们被广泛用于传播合法的金融动态、天气预报和其他信息。

技术的好坏很多都归结于使用意图，而不是技术本身。当算法和软件出现后，再禁止它们就为时已晚。禁止它们只会让那些想将软件用于合法目的的人无法使用软件，而怀有恶意的行动者却仍能利用它们。我们需要迅速惩罚那些以危害社会的方式使用技术的人，同时鼓励我们的机构、研究人员和企业提出对策。

人们一厢情愿地认为，像人工智能这样的技术是可以控制的。但实际上这很难实现，技术总会被滥用。决策者需要思考以下问题：我们如何才能迅速应对这些滥用行为？什么样的政策能够刺激和奖励防止这些技术伤害人？

让我们以社交网络为例。我们能否通过立法，使社会网络更加负责地管理其数据，彻底审查和监测所有第三方访问，并在虚假新闻或其他新出现的威胁成为一场大灾难之前制定应对措施？在管理方面，能否加大对故意滥用新技术和出现重大疏忽的人员的惩罚力度，有助于鼓励企业家和公司主动提出解决方案？

未来，人工智能、大数据和其他技术无疑会带来一系列新的社会问题。试图针对每项新技术的所有细节进行立法的做法太过笨拙，而且可能会在制定长期解决方案方面适得其反。因此，政府应鼓励相关人员制定预防措施，以防范预期的威胁，同时制定政策，促进市场对现有问题做出快速响应。只有利用市场力量并将注意力集中到最迫切需要解决的问题上，决策者才能在解决新兴技术的破坏力方面占据主导地位。

作者简介　史蒂文·霍夫曼（Steven S. Hoffman）

史提文·霍夫曼在硅谷被人称作为霍夫船长，他是创始人空间的CEO，而创始人空间是世界领先的孵化器和加速器之一，在22个国家中拥有超过50个合作伙伴。他还是一位天使投资人、八月资本的有限合伙人、连续创业者，以及关于激进式创新的获奖著作《让大象飞》（Make Elephants Fly）的作者。

霍夫曼在生活中不断创新，他曾经从事过许多不同的职业，包括连续创业者、风险投资人、天使投资人、工作室负责人、计算机工程师、电影制片人、好莱坞电视执行人、出版作家、程序员、游戏设计师、日本漫画改编人、动画制作人以及配音演员。

霍夫曼拥有加州大学计算机工程学士学位，以及南加州大学电影电视专业的艺术硕士学位。他目前居住在旧金山，但为了在全球各地拜访初创企业、投资人以及创新者，他的大部分时间都在飞机上度过。

第四部分：
国际组织相关政策进展

掌握人工智能治理的双刃剑

伊莱克利·伯利兹

科学进步产生了新的技术工具，可以为社会带来巨大的利益。尤其值得指出的是，人工智能正对从医疗到金融的许多领域产生全球性的影响。人工智能甚至可以帮助我们实现世界各国领导人在《2030年可持续发展议程》中设定的17个雄心勃勃的全球目标。然而，我们在讨论大规模使用人工智能的法律和伦理影响的多边决策和跨学科合作中，应该时刻保持谨慎和努力。迄今为止，各种机构采取的自我监管方式都在试图遏制在特定领域中使用人工智能可能产生的负面影响。例如，美国医学协会提出了一个监管框架，用于医疗保健领域人工智能的负责任开发。荷兰央行发布了一份指导文件，其中包含了在金融领域负责任地使用人工智能的原则，以防人工智能对银行、客户甚至整个金融业的信誉或声誉产生负面影响。

然而，这并不意味着政府不需要采取行动。为了减少人工智能可能带来的公共风险，某种形式的监管是必要的。虽然对国家或国际规则有一些初步的讨论，但我们离建立真正的国际治理机制还有很长的路要走。技术进步的速度快于我们的反应能力，如果政府不能跟上发展的步伐，未来政府可能会采取禁止的做法，以尽可能降低使用人工智能带来的风险。然而，这些方法可能会限制技术发展和扼杀创新。

在联合国区域间犯罪和司法研究所（UNICRI），我们建立了一个专门的人工智能和机器人技术中心，并且是少数几个致力于研究人工智能与犯罪预防和控制、刑事司法、法治和安全之间关系的国际行动者之一。我们努力支持和协助国家当局（如执法机构）了解这些技术的风险和效益，并探讨如何利用它们创造一个没有暴力和犯罪的未来。为了实现这一目标，我们正在开展试点项目，包括利用人工智能打击腐败、人口贩卖、猥亵儿童、恐怖主义资助，并为深度伪造视频制定解决方案。

在这一特定领域的人工智能治理方面，我们与国际刑警组织共同创建了一个全球平台，讨论人工智能在执法方面的进展及其影响。从2018年开始，我们每年都会举办一次全球执法人工智能大会。这些会议的成果，包括2019年的一份联合报告，代表着对推进执法领域人工智能治理的贡献。在今年晚些时候，我们将举行第三届全球会议，制定一个执法部门负责任的人工智能创新工具包，为执法部门以可信和合法的方式开发、部署和使用人工智能提供指导和支持。

随着新型冠状病毒（SARS-CoV-2 coronavirus，COVID-19）的出现，以及由此导致的封锁、人员流动限制和边界关闭，执法机构和安全服务的工作环境突然变得更加复杂。为应对这一日益严重的危机，许多国家再次转向人工智能和相关技术，以独特和创新的方式为控制疫情提供支持，特别是在监测环节进一步加强。虽然各国政府必须尽最大努力阻止病毒的传播，但是不能把对基本原则和权利的考虑以及对法治的尊重搁置在旁。即使在严重危机时期，我们仍要意识到人工智能的双重性，并努力推进人工智能治理，这一点至关重要。

因此，确保我们不会偏离负责任的人工智能的发展进程，比以往任何时候都更有必要。人工智能的积极力量和潜力是真实存在的。然而，要真正利用它，我们必须努力确保它的使用是负责任的。

像上述工具包这样的软法方法可以对人工智能治理做出重要贡献，特别是在执法领域，人工智能的使用确实是一个边缘案例。然而，人工智能的积极力量和潜力是巨大的，要想充分利用这一力量，我们首先必须努力确保它的使用是负责任、遵循原则且符合国际法律的。

作者简介　伊莱克利·伯利兹（Irakli Beridze）

联合国人工智能和机器人中心负责人。

伯利兹在领导多边谈判，与各国政府、联合国机构、国际组织、智库、民间社会、基金会、学术界、私营企业和其他国际合作伙伴制定利益相关者参与计划方面具有20多年的经验。

自2014年以来，伯利兹发起并管理了联合国首批人工智能和机器人项目之一，并在联合国大会和其他国际组织中发起和组织了一些高级别活动。他发现了传统威胁和风险的协同作用，并确定了人工智能可以帮助实现联合国可持续发展目标的解决方案。

伯利兹先生正在就与国际安全、科技发展、新兴技术、创新和新技术的颠覆性潜力有关的许多问题，特别是预防犯罪、刑事司法和安全问题，向各国政府和国际组织提供建议。

他是多个国际工作小组的成员，这些工作小组包括世界经济论坛的全球人工智能委员会和欧洲委员会的高级人工智能专家小组。他经常就与技术发展、指数技术、人工智能和机器人以及国际安全相关的主题进行演讲。他在国际期刊和杂志上发表了大量文章，媒体在谈到有关人工智能的问题时，经常引用他的文章。

伊莱克利·伯利兹还是国际性别平等捍卫者（International Gender Champions）的成员，支持IGC专家小组的平等宣言。2013年，他代表禁止化学武器组织接受了诺贝尔和平奖。

寻求灵活、合作和全面的人工智能治理国际机制

温德尔·瓦拉赫

在过去的十年里，国际社会不断呼吁建立灵活、具有适应性的治理机制，以求协调在多个利益相关者之间的利益。这对于新兴技术的治理尤为重要，因为新兴技术的快速开发和应用导致传统的伦理、法律监督方式与需求严重不符。正如本文的读者所知，人工智能在过去的一年里受到了极大的关注，政府机构正在审议超过55项关于人工智能治理的宽泛原则和一系列具体的政策建议。

人工智能为创造新的、更灵活的新兴技术的国际治理提供了一个完美的试点项目。国际社会已经提出了一些不同的人工智能治理机制，比如联合国秘书长数字合作高级别小组的建议以及IEEE合乎伦理设计倡议。经合组织已经开始着手建立人工智能政策观察站。专家学者也已提出其他手段，用于监测人工智能的开发，判定人工智能治理与人工智能发展之间的差距，并开发弥合这些差距的专用工具。

首届国际人工智能治理大会（ICGAI）的计划正在进行中，该大会将由布拉格市主办。会议原定于2020年4月召开，但因新冠肺炎疫情推迟至10月。本次会议将涉及一系列宽泛的原则和具体的政策建议等内容，为实现人工智能的灵活治理奠定坚实基础。目前，为筹备大会，将召开一系列专家研讨会，讨论下述议题：

- 灵活、合作和全面的国际人工智能治理机制
- 人工智能治理的硬法和软法
- 人工智能与国际安全
- 尽可能减少系统故障并加强对系统的管理
- 公司自治与问责
- 包容性、工作和社会的合理转型，以及满足小国和未得到政府充分关注的社区的需求

每一个研讨会都将制定提案，提交给ICGAI的参会者。如果ICGAI的参会者压倒性地支持这些提议，那么将采取第一步举措来实施这些提议。第一次专家研讨会于2020年1月6日至7日由斯坦福大学数字政策孵化器主办。它提议建立一个全球治理网络，作为人工智能分布式治理中的一个附加机构。

人们希望大会将引入一种真正的多利益攸关方方法来管理新兴技术，包括来自边缘社区的声音。特别重要的是将有中国代表参加。尽管中国是世界上人工智能解决方案的主要实现者，但迄今为止，它尚未参加或被纳入相关新应用程序治理的国际论坛中。

如您有能力为本次对话做出贡献，并希望参加ICGAI，可访问以下网址进行注册：
https://www.eventbrite.com/e/the-1st-international-congress-for-the-governance-of-ai-icgaiprague-2020-tickets-86234414455

作者简介 温德尔·瓦拉赫（Wendell Wallach）

温德尔·瓦拉赫在耶鲁大学跨学科生物伦理学中心领导了11年的技术和伦理学研究。他是黑斯廷斯中心的高级顾问、卡内基国际事务伦理委员会的成员，以及法律和创新中心（ASU）的成员。他最新出版的著作《危险的大师：如何防止技术失控》（A Dangerous Master: How to Keep Technology from Slipping Beyond Our Control）是新兴技术领域的入门指南。此外，他还与科林·艾伦（Colin Allen）合著了《道德机器：如何让机器人明辨是非》（Moral Machines: Teaching Robots Right from Wrong）。2017年冬季，劳特利奇出版社出版了瓦拉赫编辑的八卷《新兴技术伦理论文集》（The eight volume Library of Essays on the Ethics of Emerging Technologies）。瓦拉赫在2014年获得世界伦理技术奖，在2015年获得新闻媒体奖，并在2015-2016年担任渥太华大学富布赖特研究委员会主席。世界经济论坛任命瓦拉赫先生为2016-2018年全球未来技术、价值观和政策委员会联席主席。此外，他将在未来两年成为世界经济论坛人工智能委员会的成员。瓦拉赫还是首届国际人工智能治理大会（ICGAI）的主办人，该大会将于2020年10月在布拉格市召开。

国际社会人工智能治理意识的觉醒

塞勒斯·霍德斯

在我看来，2019年是一个转折点。在这一年，国际社会（政府、私营部门、民间团体和政府间机构）逐渐意识到新兴智能系统的全球治理对人类来说是一件好事。

我参与过的各种活动都反映着这一认识。这些活动包括：

1、由生命未来研究所在波多黎各主办的有益通用人工智能大会（The Beneficial AGI conference）是一项重要活动，它在中美经济局势紧张的背景下，促成了美国与中国就人工智能安全问题的对话。

2、第二届全球人工智能治理圆桌会议：作为世界政府首脑会议的一部分，在迪拜确定了多利益相关者/集体智能方法。除了汇集了250位人工智能领域的国际专家外，今年的活动特点还包括：

- 联合国教科文组织（UNESCO）和电气和电子工程师协会（IEEE）召开会议讨论人工智能伦理。IEEE介绍了其在人工智能伦理方面的开创性工作，而NESCO则表示准备着手在联合国机构内就人工智能伦理问题发挥其领导作用；
- 在扩展智能领域组建理事会（麻省理工学院媒体实验室-电气和电子工程师协会）；
- 在牛津大学和麦肯锡公司的帮助下，首个关于全球数据共享的研讨会得以举办，研讨会共收到超过40份意见书。“全球数据共享”现在是人工智能全球协作的一部分，该内容已经出现在日内瓦的人工智能造福人类峰会（AI for Good）以及纽约的联合国大会（the UN General Assembly）上。此外，在即将于4月份举行的世界银行春季会议（the World Bank Spring Meetings）上，将会提交3个实践案例。这3个案例可以在全球范围内复制和扩展，实现数据共享，这有助于制定具体的解决方案，实现可持续发展目标；
- 成立经济合作与发展组织人工智能专家组（AIGO），负责制定人工智能原则。

3、20国集团和一些伙伴国家通过的《经合组织原则》是一项重要工作，它总结了一些重要建议，这些建议对于社会实现有益的人工智能具有重要意义。

《经合组织原则》重点关注：

- 透明性和可解释性
- 鲁棒性、保障性和安全性
- 问责性
- 投资人工智能研发
- 培育人工智能数字生态系统
- 为人工智能创造有利的政策环境
- 加强人才能力建设，为劳动力市场转型做准备
- 促进可信赖人工智能的国际合作

4、经合组织人工智能政策观察站（OECD AI Policy Observatory）拟于2020年2月启动，该观察站的目标是“协助各国鼓励、培育和监督负责任的可信赖人工智能系统的开发，以造福社会”。

5、2019年6月，20国集团通过了《经合组织人工智能原则》，这是该领域的重要进展，世界人工智能领导者（美国和中国）都参与其中。

6、联合国教科文组织发布全球人工智能伦理系列丛书。该系列丛书始于北非、法国、中国和巴西，汇集了多学科的观点，提出了以人为本地使用人工智能的观点，促进了关于可持续发展的人类价值观的讨论。

7、同样，未来社会（Future Society）的人工智能倡议一直在与世界银行合作，为发展中国家的人工智能战略制定框架，宣传人工智能治理的重要性以及决策者实施人工智能治理的方法。

8、最后，由法国总统埃马纽埃尔·马克龙主持的“人类人工智能全球论坛”是法国主办的7国集团主席会议的一部分，也是国际人工智能小组的前身。该论坛（类似政府间气候变化专门委员会，IPCC）的目标是成为理解和分享人工智能问题与最佳实践研究成果以及发起国际人工智能活动的全球参考。

作者简介　塞勒斯·霍德斯（Cyrus Hodes）

塞勒斯·霍德斯是硅谷风投公司FoundersX Ventures的合伙人，该公司专注于早期和成长阶段的人工智能和机器人初创企业。

塞勒斯是“未来社会”（一个在哈佛肯尼迪学院孵化的非营利组织）人工智能初创项目的联合创始人和主席，他与众多全球利益相关者一起研究、讨论和构建人工智能的治理框架。

塞勒斯与联合国秘书长办公厅、麦肯锡以及100多个全球机构（国际组织、政府、市政当局、私营部门和学术界）合作，共同领导全球数据共享项目。

塞勒斯曾担任阿拉伯联合酋长国总理办公室人工智能办公室顾问。在过去的两年里，他负责在迪拜举行的世界政府首脑会议上组织召开了人工智能全球治理圆桌会议。此外，塞勒斯还是经合组织人工智能专家组（AIGO）（现已与经合组织人工智能专家网络合并（ONE AI），以及扩展智能委员会（MIT-IEEE）的成员。自2016年以来，塞勒斯还担任了IEEE合乎伦理设计委员会成员、智能迪拜人工智能伦理顾问，以及人工智能共享指导委员会成员。

塞勒斯曾就读于巴黎理工学院，其后来成为其讲师。他获有巴黎第二大学（荣誉）文学硕士学位，以及哈佛大学硕士学位。

人工智能治理：从原则到实践的转变

尼古拉·米埃尔

2019年可谓是人工智能全球治理的关键一年，在这一年里，人工智能的全球治理取得了重大进展，开始从原则转向实践。

经合组织于5月22日公布了《人工智能原则》，确立了一个全球参考点。《人工智能原则》的伦理和治理原则旨在促进创新、可信赖，且尊重人权和民主价值的人工智能。这些原则是由领先的多边组织提出的关于人工智能的第一套全球性原则，其制定过程严格，由一组独立专家领导。该原则在2019年6月得到了20国集团的认可。为了帮助实施这些人工智能原则，经合组织还宣布建立一个“人工智能政策观察站”，为人工智能指标、政策和实践提供证据和指导，并建立一个促进对话和分享人工智能政策最佳实践的中心。

随后，法国和加拿大在2019年8月召开的7国集团会议上宣布启动由经合组织主办的人工智能全球伙伴关系（GPAI）计划，该计划将与人工智能政策观察站一同开展。在最初的设想中，GPAI类似于“针对人工智能的联合国政府间气候变化专门委员会（IPCCI）”，旨在将全球许多杰出的人工智能科学家和专家聚集在一起，以促进合作伙伴在人工智能政策制定方面的国际合作与协调。人工智能政策观察站和GPAI都将在2020年启动。作为GPAI多方利益相关者全体年度专家会议的前身，法国总统马克龙于2019年10月底在巴黎主办了第一届全球人工智能人类论坛。第二届论坛将于2020年秋季在加拿大举行。

最后，联合国教科文组织大会于2019年11月投票，一致要求该组织在未来两年内开发一种针对人工智能伦理的标准制定工具。这一流程将包括在联合国教科文组织、未来社会、蒙特利尔大学和蒙特利尔学习算法研究所之间的伙伴关系引入人工智能公民论坛的框架内，并在世界各地进行广泛的多方利益相关者协商。

具体而言，上述措施以及2019年启动的许多其他举措，如联合国秘书长数字合作高级别小组的报告、数字健康与人工智能研究中心、AI Commons（一个类似于人工智能的维基百科的平台）等表明，越来越多的政府、专家和实践参与者正在将他们对人工智能治理的关注从“是什么”或“应该是什么”转向“如何实现”。

除了政策制定之外，我们还看到这种从原则到实践的转变也在专业组织机构中发生。IEEE于2019年3月发布的《自主和智能系统合乎伦理设计全球倡议》是“行动中的伦理”的第一个版本，旨在指导工程师负责任地使用人工智能。除此之外，越来越多的组织和公司开始致力于通过制定行为准则和章程，将国际人工智能伦理原则转化为各自的实践和文化，协助将数字化转型工作导向以可信赖的方式采用人工智能。最后，一些政府支持的或独立的人工智能系统审计和认证举措已于2019年出现。这类举措的重点正是将原则转化为实践，并协助调控人工智能应用方面的竞争，使之成为一场“伦理至上”的竞赛。因此，除了加强新的欧洲联盟委员会宣布的监管能力外，认证和审计计划还有可能为建立“信任基础设施”作出巨大贡献。

作者简介　尼古拉·米埃尔（Nicolas Miailhe）

未来社会联合创始人兼主席。

2014年，尼古拉·米埃尔与合作者共同创立了未来社会，并在哈佛大学肯尼迪政府学院孵化。未来社会是一个独立的智库，从2015年启动的“人工智能倡议”开始，以人工智能为出发点，专门研究新兴技术的影响和治理问题。尼古拉是公认的战略家、思想领袖和实干家，曾在世界各地演讲，为跨国公司、政府和国际组织提供咨询。他是联合国教科文组织与蒙特利尔学习算法研究所合作组织的AI公民论坛和每年在迪拜举办的世界政府首脑会议期间举行的AI全球治理圆桌会议的共同召集人。此外，他还是人工智能共享伙伴关系计划指导委员会成员、经合组织人工智能专家组成员，世界银行的“全民数字经济倡议”小组成员，以及全球扩展智能理事会成员。尼古拉任教于巴黎国际事务学院（巴黎政治学院）、马德里工业工程学院全球与公共事务学院，以及迪拜穆罕默德·本·拉希德政府学院。同时，他还是IEEE全球自主和智能系统合乎伦理设计倡议三个委员会的成员、哈佛大学科学技术与社会项目的高级研究员，以及马德里IE商学院变革管理中心的研究员。

《经合组织人工智能原则》——人工智能治理的全球参考

杰西卡·库辛·纽曼

2018年底，DeepMind的联合创始人穆斯塔法·苏莱曼（Mustafa Suleyman）预测，2019年将是我们建立全球竞技场，支持国际和多利益相关方协调，促进人工智能安全和伦理发展的一年。苏莱曼写道，竞技场需要全球化，因为人工智能的机遇和挑战不会受国界的限制，组织边界对其也不适用。

苏莱曼的前瞻观点在很多方面都得以证实。2019年，一些意义深远的新全球论坛相继出现，包括联合国秘书长数字合作高级别小组、人工智能全球伙伴关系、经济合作与发展组织（OECD）原则和政策观察站等。

经合组织的人工智能原则和政策观察站代表了全球人工智能治理的重大进展。该观察站于2019年5月22日发布的《经合组织人工智能原则》，这是首个针对人工智能治理的政府间标准，并成为了未来人工智能治理的一个新的“全球参考点”。

所有36个经合组织成员国以及包括阿根廷、巴西、哥伦比亚、哥斯达黎加、秘鲁和罗马尼亚在内的几个非成员国都签署了《经合组织人工智能原则》。此外，欧盟委员会也对该原则表示支持，乌克兰则是于2019年10月成为了其签署国之一。一个月后，当20国集团发布《人工智能原则》时，人们注意到这些原则是从《经合组织人工智能原则》中提取的。20国集团的支持显著扩大了参与国的范围。

这些原则包括：对实现包容性增长、可持续发展和福祉的呼吁；以人为本的价值观与公平；透明性和可解释性；稳定性、保障性和安全性；以及可问责性。此外，针对国家政策和国际合作的建议包括：加大对人工智能研发的投资；培育人工智能数字生态系统；创造有利的政策环境；加强人才能力建设，为劳动力市场转型做准备；促进可信赖人工智能的国际合作。《经合组织人工智能原则》体现了人类的广泛共识，即促进可信赖人工智能的发展需要全球范围内的协调与合作。

经合组织在这一趋势的基础上继续推进，致力于帮助各国落实这些原则和建议。经合组织于2019年底启动了人工智能政策观察站计划，以促进全球多方利益相关者之间的对话，并提供有关人工智能的循证政策分析。观察站将向全球实时发布实施人工智能原则的实务指引，以及人工智能政策和举措的实时数据库。此外，它还将编制人工智能发展的指标和度量方法，并利用其号召力建立起私营部门、政府、学术界和民间团体之间的沟通桥梁。

《经合组织人工智能原则》完成了一项在一年前大多数人都认为难以实现的壮举。美国在签署《原则》时，对其他政策领域的国际协调相对反感。中国和俄罗斯已经达成共识，将在更大范围内支持《原则》的实施，也欢迎其他国家提供支持。虽然相关实施细节仍在敲定中，但在2020年，我们可能会看到更多实质性的人工智能治理承诺和更多行动者的参与。

作者简介

杰西卡·库辛·纽曼（Jessica Cussins Newman）

杰西卡·库辛·纽曼是加州大学伯克利分校长期网络安全中心的研究员，这是一个研究人工智能对全球安全的影响的跨学科中心，她在此领导了人工智能安全计划。她还是未来生活研究所的人工智能政策专家和未来社会的研究顾问。杰西卡是哈佛大学贝尔弗中心2016-2017年度国际和全球事务研究员，曾在哈佛大学科技与社会中心项目、未来研究所和遗传学与社会中心担任研究员。杰西卡获有哈佛大学肯尼迪学院公共政策硕士学位、加州大学伯克利分校人类学学士学位，并获得最高荣誉。她通过《山丘》、《洛杉矶时报》、《药学杂志》和美国消费者新闻与商业频道（CNBC）等多种渠道发表了数十篇关于新兴技术的影响的文章。杰西卡是中美高级网络技术研讨会（CNAS）人工智能工作组的成员，也是人工智能全球伙伴关系公平、透明和可问责人工智能专家小组的成员。

人工智能治理成为国际关系重要议题

陈定定

随着新一轮产业革命席卷全球，人工智能成为产业变革的核心方向。人工智能是经济发展的新引擎，是国际竞争的新焦点，是社会建设的新机遇。2019年，人工智能在科技层面的热度持续攀升，其在治理层面的紧迫性也逐渐显现。

作为第四次科技革命的核心，人工智能领域的成果影响着一国综合国力水平，世界各国都对人工智能领域的发展给予了高度的重视，力争在这一关键领域取得先发优势。2019年，各国围绕人工智能开展了一系列合作与竞争互动，为保证科技领域的良性竞争和不断激发创新活力，人工智能全球治理成为国际关系研究的重要关切。技术竞争、贸易冲突、信息安全和伦理责任等，都是人工智能领域亟待达成全球共识的议题。治理规范的缺失，不仅不利于发挥技术对人类社会的正面效应，甚至可能带来无序和混乱。

2019年，各国通过举办论坛、出版报告和制定规范等多种方式，力求推动人工智能治理跟上技术发展的脚步。但各国在治理理念、发展阶段和技术水平等层面的差异，为共识达成设置了重重障碍。作为当前世界的主要大国，2020年，中美两国在国际秩序塑造过程中应进一步发挥带头作用，协同各国加入规范制定，以“科技向善”引领人工智能全方位治理，共同应对发展过程中面临的挑战，并促进技术成果在全球范围内的最大化应用。与此同时，人工智能发展仍处于不饱和阶段，中美仍有很大的合作空间，两国应充分认识双方在该产业链中的相互依存关系，以及该领域未来广阔的前景，共同推动人工智能产业有序向前发展。

作者简介　陈定定

陈定定，暨南大学国际关系学院教授，博士研究生导师，21世纪丝绸之路研究院副院长。著名社会智库海国图智研究院创始人兼院长。曾任国际研究协会（ISA）亚太区副会长（2014-2018），现兼任德国全球公共政策研究所客座研究员、清华大学全球化研究中心高级研究员。他的研究方向包括中国外交政策、亚洲安全、中国政治和人权。

第五部分：
国家和地区相关政策进展

欧洲议会应对人工智能治理的价值理念

伊娃·凯里

快速发展的技术的价值主张引人注目。它承担着减少经济摩擦、缓解重要资源稀缺、简化市场和公共政策程序的功能，并且能够创造新的社会动力，以使社会具备更广泛的包容性和更好的连通性。人工智能就是这一转变的核心。

人工智能给我们带来了新的挑战，在全球竞争力的公平竞争中，信息掌握和处理方面的不对称，以及新型的负外部性是导致市场失灵的新根源。

在竞争领域，数据成为了建立新全球领导力的核心要素，能够更好、更智能地获取和处理数据的人将会成为赢家。接下来的关键就是要获取人工智能数据，并保证人工智能的技术质量。为了确保新时代的公平竞争环境，能力建设和监管框架将有助于遏制主流数字平台导致的寡头垄断。制定新的竞争规则时，不仅要考虑数字公司的营业额，还要考虑它们所拥有数据的数量和质量，以便公平分配其使用价值，保障社会中个体的权利。

与此同时，我们需要在人工智能领域制定高质量的全球技术标准，并通过在全球网络中建立强大的创新生态系统，创造卓越的研究环境。质量低劣的人工智能可能会对经济发展、社会包容性以及我们的制度、民主和媒体的质量产生负面影响。高质量的技术标准能够降低人工智能的操作风险，提升法律确定性，提高公民选择的质量，确保互操作性并加速可扩展度的提高。

欧盟立志成为人工智能领域的全球领导者，它对基于人工智能的创新解决方案进行系统投资，构建快速技术转让机制，营造良好的监管环境。通过数字创新中心和人工智能卓越中心加强创新生态系统建设，并资助高质量研究项目。此外，欧盟计划开发基于人工智能的试点项目，在大规模行动中试验人工智能的应用，以获得实际运营经验，然后将这些经验和基础设施设计逐步推广到国家、地区和市政各级的治理中。

没有使命和社会责任的人工智能终将成为“人工愚蠢”。高标准、高伦理准则和一个可行的监管框架至关重要。我们需要通过规划改善互联互通和数字教育的战略，把人置于人工智能的中心，来解决技能不平等、访问权限不平等和机会不平等的问题。应该从技术上实现人工智能的质量和标准，以防止排斥和歧视偏见。《通用数据保护条例》（GDPR）以保护人权为原则，为人工智能治理奠定了基础，但并不支持“一刀切”的做法。用于解决问题或进行决策的人工智能算法，应在设计上符合伦理规范，尊重隐私，数据使用应该更加透明。

由于数据是人工智能的核心，数字平台在收集数据时需要征得公民的同意，并利用这些数据获得的利润对公民进行补偿。应用程序、摄像头、麦克风和任何其他用于收集数据的方式，都应是“默认关闭”的，除非公民知道它们的存在，并有选择是否使用的自主权。类似地，在新媒体中，应防止人工智能处理定向消息传递以推广的某些内容；应对深度伪造进行标记，同时，应当提供其他选择，使得人们能够获得平衡的信息，避免造成误解，操纵他们的意志。

最后，需要建立一个欧盟人工智能调整基金，这样就不会有人掉队，这将成为我们2020年最重要的项目。

这些原则和观点概括了我在这个充满挑战的时代对这项颇具挑战性的技术所持有的态度。我与大家分享这些，是希望它们能够成为欧洲、亚洲和美洲之间的全球民主方法和技术合作制度的基础，使公民的利益和社会的繁荣成为我们未来战略的核心。

作者简介 伊娃·凯里（Eva Kaili）

伊娃·凯里是2014年当选的欧洲议会议员。

作为欧洲议会科学和技术方案评估机构（STOA）的主席，她一直致力于推动创新，并视其为推动欧洲数字单一市场建立的驱动力。她一直活跃于区块链技术、移动/电子健康、大数据、金融科技、人工智能和网络安全领域。

自当选以来，她在税收领域也非常活跃，一直担任ECON委员会年度税务报告的报告员。作为经济委员会的成员，她一直关注着欧盟的金融一体化和欧元区金融危机的管理。

伊娃是欧洲议会区块链决议、EFSI立法意见和年度税务报告的起草人，以及资本市场联盟和家族企业诉讼案件中社会民主党的谈判代表。

在担任欧洲议会议员之前，她与泛希腊社会主义运动（PASOK）合作，曾两次当选希腊议会议员（任职于2007–2012年）。她拥有建筑学和土木工程学学士学位，以及欧洲政治学硕士学位。目前，她正在攻读国际政治经济学博士学位。

多边方法的典范
——欧盟人工智能高级别专家组

弗朗西斯卡·罗西

欧盟委员会于2018年组建了由人工智能利益相关者构成的独立人工智能高级专家组(HLEG),该专家组的任务是为欧洲的人工智能战略制定指导方针和政策。2019年,该专家组发布了两份文件:《可信赖人工智能的伦理准则》和《人工智能政策与投资建议》。这两份文件都聚焦于可信赖人工智能的概念,是HLEG内部和整个欧洲人工智能生态系统深入讨论的结果,为欧洲形成繁荣的人工智能环境提供了一个全面的蓝图,并将对全世界人工智能治理产生积极的影响。

《可信赖人工智能的伦理准则》定义了以人为本的可信赖人工智能的概念,从基本人权开始,到介绍伦理准则,最后提出了针对可信赖人工智能的7个要求:人为管控、鲁棒性和安全性、隐私和数据治理、透明性、公平性和包容性、社会和环境福祉,以及问责性。《人工智能伦理指南》还定义了一种评估方法,任何公司都可以采用这种方法来制定一个用于开发可信任的人工智能的流程,并评估其产品和服务是否符合这7个要求。这与IBM等公司现有的一些策略是一致的,在这些公司中,人工智能应用说明书的概念已经得到了全面的评估、讨论和测试。

这两份报告所给出的政策和投资建议非常及时,因为各国政府都在寻求建议和指导,以确定自己的人工智能战略。这两份报告建议使用一种基于风险、精确驱动、适应于特定环境的方法来应对可能的监管挑战。这两份报告还建议,包括政府在内的公共部门应该成为可信赖人工智能技术发展和推广的催化剂。这是扩展公众获得和熟悉人工智能技术的重要途径。此外,这两份报告还提倡加强和联合欧洲的人工智能研究能力,营造开放和创新的投资环境。“以人为本”是人工智能伦理准则的核心,这一理念始终贯穿HLEG的所有政策和投资建议。HLEG还强调,要确保所有人都能够从人工智能中受益,这一要求将建议重新设计从学前教育到高等教育的整套教育系统,从而确保所有人都能够掌握一定的人工智能技术。

虽然这两份报告都针对欧盟,但HLEG的独立性以及它的跨学科特点和由多利益相关者构成的特点,可以而且应该将这两份报告作为一个范例,证明多边方法可以取得诸多成果。同时,HLEG不仅汇集了技术专家,还汇集了许多不同领域的代表,这些代表涵盖多个学术领域、行业、人权和消费者权益协会。这让HLEG能够提出雄心勃勃又切实可行的指导方针和建议,且很可能在人工智能治理领域产生深远、广泛和持久的影响。

作者简介 弗朗西斯卡·罗西(Francesca Rossi)

朗西斯科·罗西是IBM院士,在IBM研究院担任IBM全球伦理负责人。她的研究兴趣聚焦在人工智能,以及人工智能系统开发中的伦理问题和人工智能系统行为引发的伦理问题。在这些主题上,她发表了200多篇科学文章,与人合著了两本书,同时编辑了大约20册的会议论文、论文集、期刊特刊和一本手册。她是世界(美国)人工智能协会(AAAI)和欧洲人工智能协会(EurAI)的成员。她曾担任IJCAI(国际人工智能联合大会)主席、AAAI执行委员、《人工智能研究》杂志主编。她是未来生命研究所(美国剑桥)科学咨询委员会成员,也是Leverhulme未来智能中心(英国剑桥)的副主任。她是IEEE自主和智能系统开发伦理全球倡议的执行委员会成员。她代表IBM作为Partnership on AI的创始合作伙伴之一并担任董事会成员。她是欧盟委员会人工智能高级别专家组成员。她将担任AAAI下一任主席。

欧盟采取“可信赖人工智能”的执行路线

夏洛特·斯蒂克斯

在过去的两年里，欧洲逐渐成为人工智能治理领域的主要参与者。欧盟以欧盟委员会2018年发布的人工智能战略为基础，展示了对人工智能采取符合伦理和基本人权的治理方法的可能性。特别是《可信赖人工智能的伦理准则》在这方面发挥了重要作用。这一伦理准则由欧盟委员会在2018年成立的独立组织——人工智能高级别专家组起草，采取了一种全新的方式来实现伦理准则的目标。该文件有三个方面特别值得注意：(1)确定了欧洲人工智能的努力方向；(2)明确了以基本权利为基础；(3)提供了一种实施其建议的方法。本文将简要介绍这些内容，并讨论它们如何推动有关欧洲人工智能治理的讨论。

人工智能高级别专家组提出的“可信赖人工智能”的概念很快成为整个欧洲政策制定的红线。可信赖人工智能的定义是“合法、遵守所有适用法律法规；合乎伦理，确保遵守伦理原则和价值观的人工智能；从技术和社会的角度来看，人工智能系统是稳定的，尽管出发点是好的，人工智能系统也可能在无意间造成伤害”。欧洲努力追求的可信赖人工智能，又在后来欧盟委员会的《沟通：在以人为中心的人工智能中建立信任》(Communication: Building Trust in Human-Centric Artificial Intelligence)(2019年)中再次被提及并重申，从此成为支持欧盟成员国多种人工智能战略的核心理念。

《可信赖人工智能的伦理准则》的基础是一个基于基本权利的方法，目的是支持以人为中心的可信赖人工智能发展路线。通过深入考察，文件提出了“尊重人类自主权、预防伤害、公平性、可解释性”四项原则。反过来，这些原则又为从透明性到技术稳定性和安全性的“七个关键要求”的制定奠定了基础，同时推动实现可信赖人工智能和与基本权利的一致性。虽然目前全球已提出超过84套人工智能原则，该文件所提出的方法仍然是独特的。

最后，《可信赖人工智能的伦理准则》提供了一份评估清单，介绍了从伦理原则中衍生出的7个关键要求，用于指导实践参与者和其他利益相关者实施自己的伦理原则。为了确保该评估清单能对生态系统起到良好的作用，欧盟委员会进行了为期数月的大规模实验，征求了来自欧洲各地数百名利益相关者的反馈意见。在撰写本文时，已对收到的意见进行了分析，并将根据反馈意见对评估清单进行进一步修订。《可信赖人工智能的伦理准则》的制定过程展现出了一个建立在基本权利和伦理的基础上的、由相关专家领导、各方参与、有原则的方法。这份准则与委员会主席冯德·莱恩(Von der Leyen)关于在其上任一百天内建立“人工智能领域人文和伦理问题的欧洲协调方法”的提议相吻合，让欧盟在未来几年内能够以一种独特的地位领导实施人工智能伦理的治理措施。

作者简介

夏洛特·斯蒂克斯(Charlotte Stix)

夏洛特·斯蒂克斯是欧盟委员会人工智能高级别专家组的协调员。夏洛特正在埃因霍温理工大学攻读博士学位，研究人工智能的伦理和治理问题，并担任世界经济论坛全球未来神经技术理事会的专家。夏洛特运营的《欧洲人工智能通讯》杂志被公认为了解欧盟人工智能政策的权威渠道。她曾被福布斯评为2020年欧洲30位30岁以下科技人物。

此前，她曾是剑桥大学Leverhulme未来智能中心的研究员、世界经济论坛人工智能理事会的研究员，以及欧盟委员会机器人和人工智能部门的项目官。在欧盟委员会机器人和人工智能部门，她监管的项目总金额达到1800万美元，并为欧盟人工智能策略的制定做出了贡献。她还是Element AI的顾问，世界未来理事会的政策官，以及一本获奖文化杂志的创办人——该杂志最初由她独自创办，现在已发展成一个15人的团队。

英国人工智能伦理的驱动力

李安琪

英国政府已经直接将人工智能的发展与其产业战略联系在一起，并且认为这是英国，尤其是英国在脱欧后的一个潜在竞争优势。

2017–2018年，英国政府越来越强调人工智能在国家层面的重要性，将其列为英国2017年产业战略中面临的四大挑战之一，并在2018年签署了一项人工智能行业协议。此外，英国政府还设想在数据和人工智能的安全性和伦理使用方面在国际上发挥领导作用。英国政府设立了数据伦理和创新中心，作为人工智能咨询机构，并致力于成为标准制定和监管机构的“积极参与者”，尤其是在人工智能和数据保护方面。2017–2018年，英国议会也开展了一些活动，在2017年成立了一个关于人工智能的全党议会小组，同时还成立了一个人工智能特别委员会。该委员会于2018年发布了一份报告，报告包括5个不具法律约束力的总体原则，作为数据伦理和创新中心制定和发展的跨部门“人工智能准则”的基础。

2019年，数据伦理与创新中心开始运营。迄今为止，它一直专注于在线目标定位和算法决策偏见的研究。2019年7月，数据伦理与创新中心针对上述话题发布了两份中期报告，并于2019年9月就人工智能的伦理问题发表了一系列“简要”报告，重点关注深度造假、人工智能和个人保险，以及智能音响和语音助手等问题。数据伦理与创新中心计划在2020年初向英国政府提交关于在线微观目标定位和算法偏见的正式建议书。

2019年，英国国内出现了重大的政治变化，先是首相换届，然后是2019年12月的大选，选举出的新首相鲍里斯·约翰逊（Boris Johnson）赢得了下议院的多数席位。英国于2020年1月31日正式脱欧，而政府将拥有足够的多数票来制定和实施法律及政策，包括人工智能方面的法律和政策。

然而，英国内部在人工智能方面可能存在不同意见。苏格兰自治政府（由苏格兰民族党领导）于2020年1月发起了自己的倡议，计划为苏格兰制定人工智能战略。自那以后，它发布了一份范围界定文件供公众咨询。根据咨询意见，苏格兰政府计划在2020年9月公布自己的人工智能战略。这一战略将如何与英国的人工智能整体战略相一致还有待观察。

作者简介

李安琪（Angela Daly）

李安琪博士是苏格兰思克莱德大学法学院的高级讲师（副教授）和网络法律和政策中心的主任，意大利马塞拉塔大学的客座教授。她是一名研究新数字技术的社会法律学者，专攻欧盟、英国和澳大利亚的数据保护、电信监管、知识产权、竞争法和人权问题。她曾在香港中文大学、昆士兰科技大学、斯文本科技大学和英国通信监管机构OFCOM工作。她著有《从社会法律角度看3D打印革命》（Socio-Legal Aspects of the 3D Printing Revolution）（帕尔格雷夫麦克米伦出版社，2016年）和《私人权力、在线信息流和欧盟法律：注意差距》（Private Power, Online Information Flows and EU Law: Mind the Gap）（哈特出版社，2016年）等学术专著，同时也是《良好数据》（Good Data）（INC出版社，2019年）的联合编辑。她目前的研究重点是欧盟、美国、中国和印度的公共和私人机构在人工智能方面的法律、伦理声明和政策方面做出的努力。

东亚人工智能伦理和治理政策本地化

高丹青

2019年是人工智能伦理和治理从原则走向行动的一年。2017–2018年，许多国家、公司和机构争相发布人工智能伦理和治理原则。不谋而同，国际人工智能伦理和治理核心原则在可访问性、问责性、可控性、可解释性、公平性、以人为本、隐私、安全、保障和透明度等主要方面保持着高度的一致。现在我们将进入这些原则的实施阶段，这些实践主体将探索全球共享原则本地化的意义。

本地化是人工智能伦理和治理原则发展的关键要素。随着我们开始追求原则本地化，我们可以观察不同分歧之间的主要争论点，并从中寻找新的实施路径，这是一种积极的发展趋势。人工智能伦理和治理原则只有付诸实践才能证明其有效性，而这就要求它们必须满足本地和现实的需要。在本地化流程中，最常见的问题是人工智能伦理和治理原则必须要遵循当地的文化、宗教和哲学传统。这一点在东亚尤为突出，那里的儒家哲学传统、宣扬万物有灵论的佛教和神道教信仰，以及丰富的科技文化认知，都在人工智能伦理和治理原则的本地化过程中发挥着关键作用。

另一个值得注意的本地化问题是在采用不同方法实现隐私和问责等原则中衍生出来的。在隐私的本地化流程中，我们看到欧盟、美国和中国在数据所有权和保护方面采取了不同做法，这对人工智能训练而言也十分关键。欧盟支持《通用数据保护条例》，致力于赋予用户权力，并重新让个人获得数据的控制权。在美国，尽管数据保护法规不断变化，科技公司仍然将数据视为专有数据，尤其是在与第三方交易时。在中国，政府提高了风险意识，禁止滥用、误用和过度收集用户数据的应用程序，并对用户进行提醒。

隐私本地化也会影响问责，而问责是人工智能开发的核心。在欧盟、美国、中国以及其他国家，我们看到有关当局要求公司对其开发和传播的技术负责。例如，欧盟直接对行为不当的公司进行罚款。相比之下，韩国则采取了不同的做法，其伦理指南明确了供应商（公司）、开发商和用户各自所要承担的责任。韩国的问责模式提供了值得探索的新方式和新机遇，特别体现在其尝试通过促进技术使用者的知情和自愿使用，建立起更多的个人责任。

以上是人工智能伦理和治理原则本地化流程趋势的几个例子，目前仍需要更多的研究，以更好地了解上述流程的产生机制，以及它们对国内和国际人工智能用户的影响。在此基础上，下一步是将这些本地化实例反馈给人工智能伦理和治理原则的主要制定者，以共享最佳实践并确定人工智能伦理和治理原则中仍然缺少的内容。展望未来，2020年将是另一个人工智能伦理和治理原则本地化的一年，可供学习的本地化解释和实施案例将大量出现。

作者简介 高丹青（Danit Gal）

高丹青是联合国秘书长数字合作高级别小组技术顾问。她的主要研究方向是技术伦理、地缘政治、治理、安全和保障之间的交叉领域。此前，她曾是日本东京庆应大学全球研究所网络文明研究中心的项目助理教授。高丹青是IEEE P7009自主和半自主系统故障安全设计标准的主持人，并在IEEE全球自主和智能系统伦理倡议执行委员会任职。她是剑桥大学Leverhulme未来智能中心的副研究员，以及普林斯顿大学信息技术政策中心的成员。

日本人民对人工智能治理和伦理的担忧和期望

江間有沙

日本政府率先讨论了日本人工智能的治理和伦理问题。日本总务省（MIC）自2016年起开始举办人工智能网络化研讨会，研讨会于2017年发布了《人工智能研发指南》，并于2019年发布了《人工智能使用指南》。2019年2月，经过政府间和多方利益相关者的讨论，日本政府内阁秘书处发布了《以人为中心的人工智能社会原则》（以下简称《原则》）。《原则》概述了人工智能治理原则，促进产业和部门将其原则转化为实践。例如，日本商业联合会（Keidanren）在2019年2月发布了《人工智能使用战略：为人工智能社会做准备》，制定了人工智能使用战略框架。富士通、日本电气公司和NTT数据等公司也在2019年春季发布了人工智能原则。传统公司和初创公司（ABEJA）都成立了伦理委员会，开始讨论人工智能的治理和伦理问题。

在各行业开始广泛讨论人工智能的治理和伦理问题的同时，2019年发生的两起事件引起了公众的关注，并加速了社会各界对人工智能治理的讨论和关注。首先，8月发生了一起招聘管理公司向客户公司出售用户/学生数据的丑闻。尽管这一问题主要与非法使用个人信息有关，而不是由人工智能算法偏见所造成的，但这一事件几乎是日本媒体首次关注人工智能的伦理和法律问题。第二起事件发生在11月，东京大学的项目副教授（兼一家人工智能公司的董事）在Twitter上发表了有关该公司招聘政策中的种族主义观点，并声称这些歧视性言论是由机器学习造成的。东京大学立即发表官方声明，称他的推文违反了《东京大学宪章》的规定。

这些事件引发了日本社会各界对机器学习的担忧。作为回应，三个从事机器学习的学术团体在12月发表了《机器学习与公平性声明》。声明宣称：（1）机器学习只是帮助人类决策的工具；（2）机器学习研究人员正在致力于通过研究机器学习的潜在用途来提高社会公平性。该团体将在相关组织的支持下，于2020年1月举办一个研讨会，就机器学习与公平性展开对话。

2019年，数项日本的调查研究情况表明，影响人工智能治理和伦理问题的主导因素已从政府转向商业。与此同时，人工智能进展的社会应用也在不断发展。因此，日本逐渐出现了有关人工智能和机器学习的伦理、法律和社会问题。不过，自2016年以来，日本已经构建了多利益相关者和跨学科的人工智能治理网络，我们将继续解决这些问题，并为世界人工智能治理做出贡献。

作者简介 江間有沙（Arisa Ema）

江間有沙是东京大学的项目助理教授和日本理化学研究所高级情报项目中心的访问研究员。她作为一名科技研究员，主要工作是通过组织一个跨学科研究小组来研究人工智能带来的效益和风险。她是成立于2014年的可接受的负责任人工智能研究小组（AIR）的联合创始人，该小组致力于解决人工智能技术与社会之间的问题和关系。她是日本人工智能协会伦理委员会（JSAI）的成员，该协会于2017年发布了JSAI伦理准则。她也是日本深度学习协会（JDLA）董事会的成员兼公共事务委员会主席。她还是日本内阁府“以人为中心的人工智能社会原则委员会”的成员，该委员会于2019年发布了《以人为中心的人工智能社会原则》。她获有东京大学博士学位，此前曾在京都大学白眉高级研究中心担任助理教授。

新加坡人工智能伦理和治理举措

吴亦涵　尼地·阿莫林

自2017年以来，新加坡政府将人工智能列为四项前沿技术之一，这将进一步推动基础设施，从而支撑新加坡发展数字经济和建设智慧国家的愿景。一方面，2019年是新加坡启动基本政策举措的一年。另一方面，在2019年，新加坡政府通过在关键的高价值领域实施项目和建立全面的人工智能生态系统，重申了开发和使用人工智能的重要性。

这些政策举措使新加坡成为全球人工智能治理的领先国家之一。2019年4月，新加坡在联合国级别的平台，信息社会论坛世界峰会上获得了最高奖项。促成新加坡获得这一殊荣的举措包括：(1)2019年1月，新加坡发布了亚洲首个人工智能治理框架；(2)新加坡成立了一个以国际和产业为导向的人工智能和数据伦理使用咨询委员会；(3)新加坡积极开展针对人工智能治理、伦理和数据使用的项目。该项目由我领导的新加坡管理大学人工智能和数据管理中心开展，我们通过开展学术研究为人工智能生态系统做出贡献，影响新加坡及其他地区的人工智能和数据治理，特别是立法和政策方面。

2019年1月推出的示范人工智能治理框架(或称模型框架)是新加坡政府今年实施的最重要跨部门政策举措之一。该框架旨在指导各个组织在部署人工智能技术时切实解决关键伦理和治理问题。新加坡的做法有助于将伦理原则转化为企业可以采取的务实措施。这是私营部门与监管机构合作的结果，也是亚洲国家建立这种框架的首次尝试。其他国家和地区今年也采取了类似举措，例如，欧盟委员会在2019年3月公布了其最终人工智能和伦理准则，这是对欧盟《通用数据保护条例》的补充。在更国际化的层面上，经合组织于2019年5月提出了一套关于人工智能的原则，以促进尊重人权和民主价值观、创新和可信赖的人工智能使用方式。

此外，新加坡于2019年10月发布了国家人工智能战略(NAIS)，该战略将为2020研究、创新和企业计划下与人工智能相关的活动提供超过5亿新元的资金支持，以求提高人工智能在这些领域的能力。NAIS强调，新加坡首先将聚焦五个关键领域——运输和物流、智慧城市和房地产、安全与保障、医疗和教育。这些国家人工智能项目旨在引导研发投资、吸引人才，并引导新加坡支持数字基础设施的发展。

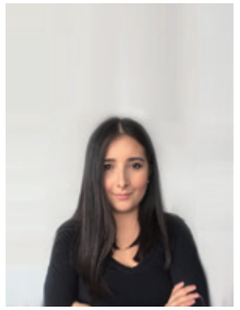
那么，我们对明年有何展望呢？我们希望通过发表前沿研究成果，从学术领域继续巩固新加坡的人工智能生态系统，促进学术界、产业界和监管机构之间的对话，特别是亚太地区组织之间的对话。我们也希望监管机构能够继续制定相关举措，支撑可信赖人工智能的发展，如资讯通信媒体发展局的第二版人工智能模型框架，以及新加坡金融管理局宣布的Veritas计划，该计划将把金融监管机构采用的基于原则的人工智能方法转化为实践。

*这项研究得到新加坡国家研究基金会在其新兴领域研究项目资助倡议下的支持。本材料中表达的任何意见、调查结果和结论或建议都是作者的意见，没有反映新加坡国家研究基金会的意见。

作者简介　吴亦涵(Goh Yihan)

吴亦涵教授的研究主要集中在合同法和侵权法方面，其次是对法定解释原则和法律程序的研究。他在国际上和新加坡国内出版了大量著作、文章和期刊论文，并被新加坡法院和马来西亚联邦法院多次引用。他被任命为新加坡上诉法院和新加坡高等法院的法庭之友。为了表彰他对新加坡法律的发展和进步做出的突出贡献，新加坡法律学会在2013年为他颁发了新加坡法律功勋奖，使其成为了第五位，也是最年轻的获奖者。2006年，他在大学本科奖学金的资助下以优异的成绩从新加坡国立大学毕业，并获得法学学士学位(一等荣誉)。2010年，他又在新加坡国立大学海外研究生奖学金的资助下获得了哈佛大学法学学位。

作者简介　尼地·阿莫林(Nydia Remolina)



尼地·阿莫林是新加坡管理大学人工智能和数据治理中心的研究助理，拥有斯坦福大学法学硕士学位，在金融服务行业拥有十多年的经验，目前担任金融机构的金融监管、数字转型和金融科技顾问。尼地还是总部位于拉丁美洲的金融集团Grupo Bancolombia的政策事务经理、经济合作与发展组织(OECD)的高级顾问，以及美国苏利文·克伦威尔律师事务所(纽约办事处)的外国律师。她曾在美国、亚洲、欧洲和拉丁美洲的多个学术机构任教或演讲，并应邀在包括国际货币基金组织(IMF)、国际证券监督管理委员会组织(IOSCO)和美国证券交易委员会(SEC)在内的多个组织做关于金融科技和金融监管的报告。她的主要工作和学术研究领域包括金融和银行监管、证券监管、金融科技、法律科技，以及法律、金融和技术的交叉领域。

印度在人工智能时代面临的重大挑战：不平等和增长之间的矛盾

乌瓦时·阿尼娅

过去一年，印度在人工智能治理问题上进展甚微。尽管人工智能被视为经济增长的催化剂和应对复杂社会经济挑战的解决方案，但印度尚未针对这项技术的管理问题制定框架。政策对话的大部分内容是由私营部门提供的，很少征求民间团体或学术界的意见。因此，释放人工智能的潜力似乎主要是一项技术挑战，可以通过建立更好的创新和创业生态系统、投资技术娴熟的人力和建设国家数据基础设施来解决。与此相关的社会挑战和风险鲜有人关注。迄今为止，在政策层面，印度在人工智能准入、平等、公平和问责方面几乎没有什么有意义的对话，尚未最终敲定的数据保护法案也不涉及机器学习系统带来的挑战。人们最关心的似乎是如何利用个人数据来促进公益和人工智能的发展，而不是隐私或社会正义。缺乏人工智能治理框架是一个关键问题，因为人工智能已经在公共系统中广泛应用，比如，全国各地的警察部门都在使用预测分析和自动人脸识别系统。此外，还计划在司法和福利发放系统中部署基于人工智能的系统。印度致力于成为全球人工智能领导者，但这不仅需要走在创新的前沿，还需要制定规范性框架和治理体系，使人工智能的发展轨迹符合社会需求。盲目的技术乐观主义可能会进一步加剧印度管理不平等和经济增长所面临的巨大挑战。

在全球层面，过去一年，人工智能治理的伦理框架大量出台。但这些措施可能仍然不够完善——它们通常只包括政府和科技公司的模糊承诺，而缺乏执行和问责机制。更具希望的解决方式是将人工智能治理与已经建立并得到广泛认可的国际人权框架挂钩。但我们需要意识到，人工智能治理的问题不单单局限于侵犯特定人权或造成个人伤害。

越来越多地使用人工智能，可能会导致不平等加剧、权力集中、歧视性和排他性制度的固化，甚至创建一个监控社会。正如人工智能不是应对社会经济挑战的灵丹妙药一样，仅仅依靠一套监管或治理框架也不足以应对这些社会危害。治理人工智能将需要一系列公共政策干预——从限制大型科技公司权力的竞争法，到行业特定标准和风险评估。印度目前还没有解决这些问题，现有的治理对话很少，仅限于如何利用印度的数据来提高印度的人工智能的成熟度和竞争力。

人工智能给公共政策带来了一个重大问题——一个多方相互作用的社会体系与技术系统如何结合的问题；在这个问题中，影响和风险存在不确定性；在这个问题中，不同利益相关者之间的价值观和世界观可能存在分歧。解决这一问题需要让多利益相关方参与迭代和适应性战略；促进协作感知、实验和学习；并建设反应能力和预测能力。

作者简介

乌瓦时·阿尼娅（Urvashi Aneja）

乌瓦时·阿尼娅是印度跨学科研究集团Tandem Research的联合创始人和董事，该研究集团在技术、社会和可持续性的交叉领域中提供政策见解。她的研究重点是南半球数据驱动决策系统的社会影响。她还是查塔姆研究所亚太项目的副研究员、T-20未来工作与学习特别工作组成员，以及国家媒体出版物的定期撰稿人。

第六部分：
来自中国的声音

结伴同行，合作共赢

傅莹

人类距离实现超级人工智能还很遥远。然而，在一些具体领域人工智能已经超越人类，且其范围在迅速扩大。人们对于由此可能获得的益处寄予厚望，但恐惧和担忧也随之而来。美国在人工智能技术创新上处于领先地位，中国则在人工智能技术的大规模和活跃的应用方面成绩斐然。中美两国有更大的责任，去思考未来、思考应当怎么做。

不过，我们在谈论未来和如何面对技术进步之前，首先需要想明白，中美是要协调合作还是彼此对抗？当前两国之间日益恶化的紧张关系，必然会影响到我们如何面对未来的挑战。也就是说，未来我们是要共同努力，让技术与人类共生，并确保技术的进步能够促进文明的繁盛，还是要分道扬镳，各自挟持技术削弱甚至伤害对方？

在经历了30多年的高速工业化之后，中国第一次跻身新技术进步的第一梯队，除了尽己所能地向前迈进，中国人也逐渐意识到制订新规则的需要。中国的国家新一代人工智能治理专业委员会于2019年2月由科技部牵头组建，6月发布新一代人工智能治理的八项原则，包括：和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷治理。

人工智能技术的成长依靠的是科研人员分享思想、相互借鉴，是全球协作的产物，跨国企业构建的人工智能平台也在快速扩展。要想规范这个进程，各国需要制订共同的道德规范和行业规则。因此，中国在这方面的努力，也需要与世界其他国家和地区，包括美国的努力，相互联通和协调。

不论中国还是美国，都不可能垄断世界的技术进步。如果两国采取互补的态度，人工智能技术的前景会更加光明；但如果不再合作，双方都将遭受损失，人工智能的总体发展也会付出代价。尤其是，如果任由传统的地缘政治、零和竞争思维主导两国关系，结果将是自毁性质的。

美国试图把高科技作为战略争夺的平台，而中方对此并不以为然。实际情况是，在这一领域存在着建设性和战略性的相互依存。科睿唯安公司的数据显示，从2013年到2017年，中美两国国际合作论文数量增长最快，合作论文达4000多篇。当然无可否认，在科学和产业界的竞争在所难免。

当前，美国企业在技术上（尤其半导体）领先，美国的大学也在世界上居于前列。而中国拥有最大的用户市场，为算法更快的迭代升级提供了条件。中美如能相得益彰，彼此都能从中受益，但如果美方执意推动脱钩，则会迫使中方寻求其他合作伙伴，或者自己设法解决，这也会削弱美国企业的地位和影响。

中国希冀的未来世界是一个相互依存的命运共同体，采取的政策是促进广泛国际对话，积极参与合作，鼓励制定共同规则，以实现安全、可靠、负责任的人工智能。

作者简介 傅莹

清华大学战略与安全研究中心主任。

傅莹1978年进入中国外交部，曾长期从事亚洲事务方面的工作，在外交部任亚洲司处长、参赞等职，1992年参加联合国在柬埔寨的维和行动，1997年出任驻印度尼西亚使馆公使衔参赞，1998年出任驻菲律宾大使，2000年出任外交部亚洲司司长，2004年任驻澳大利亚大使，2007年任驻英国大使，2009-2013年任外交部副部长，先后负责欧洲和亚洲事务。2013年和2018年，傅莹两次当选为第十二届和十三届全国人大代表，2013年至2018年兼任外事委员会主任委员和第十二届全国人大第一次至第五次会议新闻发言人。

中国人工智能治理取得积极进展

赵志耘

中国高度重视人工智能治理，习近平总书记在中央政治局第九次集体学习时强调，要整合多学科力量，加强人工智能相关法律、伦理、社会问题研究，建立健全保障人工智能健康发展的法律法规、制度体系、伦理道德。国家《新一代人工智能发展规划》对人工智能治理做出明确部署，围绕自动驾驶、机器人等重点领域人工智能应用开展相关法律问题研究和法规制定，开展人工智能行为科学和伦理等问题研究，建立研发设计人员道德规范和行为守则，积极参与人工智能全球治理等。

为进一步加强人工智能相关法律、伦理、标准和社会问题研究，深入参与人工智能相关治理的国际交流合作。2019年2月15日，科技部牵头设立新一代人工智能治理专业委员会，委员会由来自高校、科研院所和企业的相关专家组成。2019年6月17日，该委员会发布《新一代人工智能治理原则——发展负责任的人工智能》，提出了和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷治理等八项治理原则。八项原则既结合国际通行准则，又内涵中国特色，在世界引发巨大反响。

北京和上海等地相继发布其人工智能治理原则或倡议，如《人工智能北京共识》、《中国青年科学家2019人工智能创新治理上海宣言》和《人工智能安全发展上海倡议》。以腾讯、旷视等为代表的企业界也积极响应，提出基于本行业的治理准则，行业自律持续加强。

2020年，中国将着力推动上述八项治理原则的贯彻落实，加快推动建立和完善人工智能法律、标准、规范，将人工智能治理推向法制化、精细化和制度化。鉴于人工智能治理是全球性问题，国际合作将是中国人工智能治理的重要工作之一。中国在推动新一代人工智能治理过程中始终坚持开放合作的理念，积极参与全球人工智能治理议程，努力搭建世界人工智能大会等国际平台，与世界加强对话交流，共同推动新一代人工智能健康发展。中国愿意同世界任何国家和组织一起，共同推动有益于人类的人工智能。

作者简介　赵志耘

赵志耘，经济学博士，研究员，博士生导师，现任科技部中国科学技术信息研究所党委书记，兼任科技部新一代人工智能发展研究中心主任，国务院政府特殊津贴获得者，“新世纪百千万人才工程”国家级人选，中宣部“四个一批”理论家，“万人计划”领军人才。她是著名的经济理论和政策、科技管理和政策方面的领军人才。她对新兴技术和产业发展具有独特见解。她非常重视人工智能治理问题，并致力于推动中国与其他国家的相关研究与合作。她在理论体系的建设、技术进步的促进和相关学科建设方面都取得了突出成就。出版学术专著30余部，译著4部，发表学术论文130余篇。作为首席研究员，主持国家、省部级科研项目近30项，包括国家重点研发项目、国家科技支撑计划项目、国家软科学重大项目。

从治理原则走向细化落地，更加需要多方参与、协同治理

李修全

2019年，人工智能治理得到国际社会更加广泛的关注，国际组织、各国政府、学术界和企业持续探究新技术价值理念，发布各自的人工智能发展原则。中国也在2019年发布了《新一代人工智能治理原则——发展负责任的人工智能》。国际社会围绕以人为本、公平、透明、隐私、安全等重点问题，形成了共识性的表态，体现出各方对于人工智能发展越来越形成了普适的价值理念。

同时，全球人工智能治理的重心正在从原则制定走向持续推动这些原则、准则的细化落实。在这一过程中，充分吸纳利益相关方意见显得更为重要，相对前一阶段，将需要更大范围的多方参与、更加紧密的协同治理。

人工智能应用对未来社会的经济活动、公共管理、生活出行等将带来方方面面的影响，事关各行各业、各类群体。从治理原则落到细则、规则，仅仅依靠政府部门和专家学者是不够的，需要政府、学术界、产业界和普通社会公众的共同努力和积极参与。中国正在不断推动人工智能治理原则在人工智能创新发展试验区、人工智能开放创新平台等建设中的探索落地，通过探索实践形成各领域治理细则。如何建立有效的意见收集和反馈机制，使社会各界能够参与到人工智能治理中来，能够把不同群体，尤其是弱势群体等利益相关方的诉求表达吸纳到细则制定过程中，尤为重要。

同样，从全球来看，不同国家有不同的国情，不同民族群体有不同的历史和文化背景，原则落地更加需要充分沟通、协同治理。搭建更加多元化的协作治理平台，促进各国的对话和沟通，使差异性在务实沟通中能够充分碰撞、相互融合，必定有助于形成更广泛共识，推动包容发展，使人工智能更好改善各国民生、增进公众福祉。

作者简介 李修全

李修全博士，中国科学技术发展战略研究院研究员，科技部新一代人工智能发展研究中心副主任。毕业于清华大学计算机系，德国汉堡大学信息学科学系联合培养博士，在多维时序数据建模与预测、基于EEG的脑控机器人系统等领域有多年研究经历。目前主要关注大数据与人工智能技术预测、产业技术路线图、人工智能创新政策研究等，对于智能化变革的前沿趋势和对研发、产业、治理等经济社会各方面的创新政策需求具有浓厚研究兴趣。主持“我国智能经济与智能社会发展的重大战略问题研究”、“国内外人工智能前沿趋势与政策研究”等研究课题10余项。

中国走向稳健敏捷的人工智能伦理和治理框架

段伟文

2019年，中国在人工智能伦理和治理上值得注意的有四个方面。其一，推出了各种伦理与治理的原则、标准和宣言。其中包括新一代人工智能治理原则、北京智源人工智能研究院的《人工智能北京共识》、国家人工智能标准化总体组《人工智能风险分析报告》提出的人工智能伦理原则、北京大学及国家机器人标准化总体组的《中国机器人伦理标准化前瞻（2019）》提出的中国机器人伦理标准化体系。同时，工信部的智库赛迪研究院与信通院也分别提出了人工智能伦理宣言或公约，腾讯也发布了自己的人工智能伦理框架。不仅法律和哲学学者参与了相关研究，人工智能领域的研究者也对人工智能伦理体系、安全可信的人工智能等表现出极大兴趣。其二，在个人信息保护和数据权利的法律规制和数据合规治理上有一定的进展。其中包括，将《个人信息保护法》和《数据安全法》纳入下一年度的立法计划，工信部开展的APP侵犯用户权益专项整治行动。值得一提的是，未成年人保护法修订草案强调收集未成年人信息需知情同意等。其三，人脸识别等人工智能应用迅速推广并引起很多伦理和法律争议。尽管人脸识别在教室、公园等各种场景的滥用导致了舆论指责甚至法律诉讼，但其在中国的应用似乎势不可挡。其四，人工智能企业也进行了一些伦理和治理实践实践。

腾讯等领先企业提出了科技向善，并将AI应用于防止游戏沉迷和寻找丢失儿童等方面。中国的人脸识别巨头之一旷视，甚至提出了供其伦理委员会内部评估用的人工智能应用原则。但鉴于这些努力远未像KPI那样成为企业评估产品和服务的价值基础，难免被批评为柔性公关或道德洗刷。

总的来说，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伦理与治理框架，实现敏捷治理。

作者简介　段伟文

段伟文，中国社科院哲学所科技哲学研究室主任、研究员，中国社科院科学技术和社会研究中心主任，中国社科院大学特聘教授；主要研究领域为科学哲学、信息技术哲学，近年来关注大数据与人工智能的哲学、伦理和社会研究；曾赴牛津大学、科罗拉多矿业大学、匹兹堡大学访问研究；现任国际期刊社会中的信息、通信与伦理杂志、负责任创新杂志编委，中国大数据专家委员会副主任委员；主持多项社科基金研究项目，目前为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”首席专家，著有《可接受的科学：当代科学基础的反思》、《网络空间的伦理反思》、《被捆绑的时间：技术与人的生活世界》等。

全球化与合伦理成为人工智能治理共识——中国产业界的伦理关注

栾群

2019年，人工智能治理突出表现为全球化和合伦理化的特点。世界主要国家、经济体和国际组织，都陆续发布了人工智能治理的文件，最具代表性的是欧盟《可信赖人工智能的伦理准则》(2019年4月)，以及在日本筑波举行的G20数字经济部长会议和G20贸易和数字经济部长联席会议通过的联合声明和《G20人工智能原则》(6月份)；以及，也是在6月份，中国的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》。中国的人工智能治理，也从2017年国务院及部门规划如《新一代人工智能发展规划》、《“互联网+”人工智能三年行动实施方案》，以及产业、领域计划如《促进新一代人工智能产业发展三年行动计划(2018—2020年)》、2018年智能制造试点示范、《高等学校人工智能创新行动计划》等，转向了合伦理化治理。突出表现为在新一代人工智能治理原则中强调“负责任”——这与欧盟强调“可信任”的旨趣相同。8月，上海2019世界人工智能大会法治论坛发布《人工智能安全与法治导则(2019)》，论坛的主题为“共建未来法治，共享智能福祉”，以推动产业发展和相关制度的跟进，更好服务保障AI国家战略大局，向世界展现人工智能治理的中国方案。

作为行业管理部门的工业和信息化部，2019年主要实施了关于产业顶层设计的计划，如《促进新一代人工智能产业发展三年行动计划(2018-2020年)》，主推八大产品、三大技术；关于重点行业发展计划及标准，如自动驾驶《车联网(智能网联汽车)产业发展行动计划》《2019年智能网联汽车标准化工作要点》；关于联合推动重点工作，如联合自然资源部和北京市开展车联网(智能网联汽车)和自动驾驶地图应用试点工作；以及工业互联网工作，如实施《工业互联网综合标准化体系建设指南》等工作。在这些新的政策文件中，也都涉及了人工智能治理的相关论述。

作者简介 栾群

栾群，中国政法大学民商法博士，于2011年加入中国电子信息产业发展研究院，现任政策法规研究所所长。工业经济政策和法规方面的行业专家，并主管工业和信息化法律服务中心。近来咨询工作主要集中在行业战略、产业发展和行业监管方面，特别关注自动驾驶汽车、工业数据和制造业。他曾在内蒙古、河南、山东等地成功开展产业发展规划和产业政策解读项目。在《学习时报》、《中国经济与信息化》、《现代产业经济》、《经济日报》、《中国电子报》等报刊杂志上发表文章50余篇。

人工智能治理的造福于人和可问责性原则——在中国人工智能标准化制定中的理念

郭锐

随着人工智能在商业、医疗、交通运输、金融服务、教育和公共安全领域的应用，人工智能影响了人们生活的方方面面。能否妥善处理人工智能带来的负面影响，例如个人信息的泄露、信息基础不充分的人工智能的输出结果和人工智能的滥用，引起了越来越多的关注。学术界、产业界和政策制定者都积极参与了有关人工智能伦理的讨论，这使得2019年成为全球就人工智能治理达成共识的关键时期。

这一年，来自学术界、产业界和民间团体的专家们逐渐达成一致，与人工智能相关的负面影响最好被视为风险，并通过严格的风险管理系统识别、预防和管理这些风险。这一进展对标准化工作产生了影响，许多与人工智能相关的标准化工作也在稳步推进、快速发展中。建立在这个共识之上，一个使世界各国都可以从人工智能中受益并防止其危害的治理体系正在成形。虽然将人工智能的负面影响看作风险有助于解决人工智能带来的已知和直接风险，但它并不能应对人工智能带来的全部风险，特别是不确定的和长期的风险。我们将继续探索可以帮助人类社会解决人工智能伦理问题的方法。

作为国家标准化管理委员会人工智能工作组人工智能伦理课题组的首席专家，我提出了符合伦理和负责任的人工智能必须遵守两个原则：人的利益原则（即人工智能的研究和应用必须以实现人的利益为终极目标）和责任原则。这两个原则为《人工智能伦理风险研究报告》（由SAC人工智能工作组于2019年5月发布）的起草提供了依据。

作者简介　郭锐

郭锐，中国人民大学法学院副教授，未来法治研究院研究员，社会责任和治理研究中心主任。郭锐博士的研究包括公司法、金融法、人权法、人工智能伦理与治理等。他毕业于中国政法大学（法学学士、硕士）和哈佛大学法学院（法学硕士、法学博士）。他担任国家信息标准化委员会人工智能专业委员会委员、全国标准化委员会人工智能总体组社会伦理研究负责人，他参与了中国第一个以人工智能标准化为主题的白皮书的写作并撰写“安全、隐私和伦理”部分（《人工智能标准化白皮书（2018）》已于2018年发布），并主持撰写了国家标准委人工智能总体组《人工智能伦理风险研究》报告 （2019年5月发布）。

推动人工智能让城市和生活更美好

王迎春

目前全球的人工智能研究机构、企业和应用场景主要集中在城市，因此城市在人工智能发展中扮演着重要角色。作为中国最大的经济中心城市，上海正在加快向具有全球影响力的人工智能创新策源、应用示范、制度供给、人才集聚高地进军。2010年上海世博会的主题是“Better City, Better Life”，在走向智能时代的过程中，我们也相信“Better AI, Better City,Better Life”，实现这一目标需要找到人类与人工智能和谐共生的路径与方案。

城市可以为推动人工智能健康发展提供试验平台。2019年，科技部印发了《国家新一代人工智能创新发展试验区建设工作指引》，以城市为主要载体，探索形成一批可复制可推广的经验，引领带动全国人工智能健康发展。2019年5月25日，科技部与上海市共同启动“上海国家新一代人工智能创新发展试验区”建设。试验区把治理作为四大核心内容之一，推动科技创新和制度创新同向发力，一方面支持研发更负责任的人工智能，鼓励人工智能最新成果在上海率先“试水”；另一方面加强在人工智能法律法规、伦理规范、安全监管等方面的探索，努力为全国乃至全球人工智能发展贡献上海经验。如何基于人工智能为市民提供更高质量的医疗、更便捷的交通和更安全高效的城市服务是重要关注点。

塑造更好的人工智能需要开放协作。上海已经连续两年举办世界人工智能大会。习近平主席在致2018上海世界人工智能大会的贺信中指出“应对人工智能带来的新课题，需要各国深化合作、共同探讨”。响应这一号召，我们在2019世界人工智能大会上举办了治理主题论坛。来自全球的几十位专家与会交流，200多位政府和业界人士参会。通过专家们的坦诚交流，增进了互相的理解，并就一些重要问题达成了共识。会上，科学家代表还发布了《中国青年科学家2019人工智能创新治理上海宣言》，强调了人工智能发展需要遵循的伦理责任、安全责任、法律责任和社会责任等四大责任。治理论坛旨在搭建全球人工智能治理交流互鉴平台，我们希望通过论坛推动形成全球人工智能治理研究和协作的共同体，为解决相关问题贡献智慧。

在全球人工智能治理体系形成过程中，城市可以发挥更重要的作用。我们需要在尊重文化、制度多样性基础上，推动形成一个可能由全球治理框架、多个子系统方案和区域性方案组成的全球治理体系。要保证这些子系统和区域方案在全球范围内的通约性和开放性，并合作形成利益分享和安全保障的具体机制。全球城市之间可以就这些方面进行更加深入的交流合作，2019年我们已经开展了相关工作。

我们参与了上海试验区建设方案的研究，并在筹建上海人工智能治理研究机构，组织多学科专家，对强人工智能的道德框架和弱人工智能的相关法律、伦理、社会问题开展系统研究。我们希望继续与海内外朋友携手共同研究人类与人工智能和谐共生的路径与方案。

作者简介 王迎春

王迎春，博士，现为上海市科学学研究所科技与社会研究室主任，主要研究领域为创新变革与创新治理、科学技术与社会。他牵头组织了由多学科专家参与的人工智能研究组，对人工智能进行系统研究。承担了多项科技部和上海市委托的咨询项目，多次参与政府人工智能相关工作调研和政策起草。参与策划组织了上海世界人工智能大会治理论坛。目前同时负责上海国家新一代人工智能创新发展试验区专家咨询委员会秘书处工作。