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

报告中各文章观点仅代表撰稿人立场
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

2020全球人工智能治理年度观察

来自全球52位专家的观点

2021年10月

上海市科学学研究所

WHEN PEOPLE PULL TOGETHER,
NOTHING IS TOO HEAVY
TO BE LIFTED.

—— *BAO PU ZI*

眾力并，
則萬鈞不足舉也。

——《抱朴子》

目录

前言

2020年是极不平凡又极具挑战的一年。

新冠肺炎疫情的暴发和流行，对世界各国的经济社会发展造成了极大的冲击。全球的政策制定者和研究者不得不匆忙放下手中所有的计划，全力应对这一“破坏性”的新课题。正如我们所看到的，人工智能治理是这一重大课题中的子课题。

实际上，在看似各项工作处于停滞状态的2020年，与人工智能治理相关的工作并没有停滞，反而因疫情带来的一些棘手问题，得到了更多人的关注和更深入的讨论。比如，数字追踪技术的应用就在很多国家引发了广泛的讨论。在中国，“健康码”（一种数字追踪技术）的使用得到了广泛的推崇，成为了疫情防控的有力工具。而在有些国家，即便在疫情最严重的时期，人们对数字追踪技术的态度也是复杂的甚至是排斥的。

这些问题至少提醒我们，人工智能治理事关人类命运，同时也具有极大的讨论空间。

2020年对于上海市科学学研究所而言，也是非常重要的年份，这一年是我们成立的40周年。我们原本计划的一系列庆祝活动，都因为疫情而取消或者小范围举行。由我们主办的世界人工智能大会（上海）治理分论坛，也不得不改为以线下和线上相结合的方式举行。

在40年的发展经历中，我们始终认为，借鉴学习国际经验能够有效地帮助我们国家科技政策的发展，甚至能够起到事半功倍的效果。在上海市科学学研究所40年的成长经历中，我们通过多种方式，不断介绍国际上关于科技发展规律、科技与经济的关系、科技对社会的影响等方面的研究成果以及相应的政策制定。这些知识的引进为中国的科技政策发展作出了一定的贡献。近些年，在人工智能治理方面，我们也一直在关注世界各国的进展，我们希望通过这样的努力来学习经验、增进了解。

当然，学习不应该只是单向的，而应该是相互的。在新冠疫情或者人工智能这种对全人类都将产生影响的挑战面前，每一个人的想法都至关重要。互相学习，不仅仅是后发国家学习先进国家的经验，先进国家也需要了解后发国家的思考。建立在互相学习基础上的互相理解对于最终取得有效的全球共识至关重要。

2019年，是我们第一次联合全球专家共同编制《全球人工智能治理年度观察》报告，我们的邀请得到了广泛的支持，这让我们意识到这项工作的意义超过了预想。报告发布后，我们惊喜地发现，因为这份报告而衍生的更多合作已经形成。我们相信这份报告本身也成为了推动各相关方互相学习的一个平台。

我们希望今年的《全球人工智能治理年度观察》报告能够延续这一全球交流平台的功能，同时，也能够为这不平凡一年中的不平凡思考和行动，留下值得不断回忆的痕迹。

上海市科学学研究所　石谦

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、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

去年是我们第一次做这样一份报告，我们当时的设想是：从众多的人工智能治理研究中，找到真正关键的“进展”。出乎意料的是，有许多专家回复了我们的邀请，最终有50位专家参与了供稿。在报告发布后，我们同样收到了大量的反馈。我们看到了全球同行在各种社交媒体、专业社区的大量转载和评论。我们还收到了联合国大会主席办公室高级顾问的来信、蒙特利尔人工智能伦理研究所在其发布报告中推荐评论等等。这些正向的激励，让我们更加相信，全球人工智能治理的有效推进，需要这样一份可以互相交流、互通有无的年度观察报告。有鉴于此，今年我们延续初衷继续编制这份报告。

2020年注定会在人类历史上留下沉重一笔，新冠疫情的暴发和流行，让全球经济社会按下了暂停键。我们一度怀疑2020年的全球人工智能治理会不会是比较空白的一年。但是在我们的邀请发出后，我们依然收到了大量的反馈，大家都普遍有兴趣继续完成今年的年度观察。最终，与去年相比，今年参与报告的作者（及机构）数量上略有增加，一共有52位专家（47家机构）参与报告的编制。52位专家向我们展示了在这一特殊年份里全球人工智能治理的大量进展。

我们惊喜地发现，即便在疫情之下，很多机构仍在有条不紊地推动人工智能治理工作。有些作者连续两年参与了本报告，因此从他们的文章中我们能够发现一些工作的连续性进展。比如OpenAI，2019年其GPT-2的发布方案，引发了相当大的风波，而在2020年推出GPT-3的发布方案，相较而言，接受度似乎更高一些。又如欧盟在2019年发布《人工智能伦理框架》（AI Ethical Framework）后，2020年又继续发布了《人工智能白皮书》（White Paper on AI），提出了关于人工智能伦理的相应监管规则。

疫情的特殊场景之下，也让各相关方在如火如荼地部署人工智能和建设人工智能治理制度的过程中，更加认真地反思人工智能和人工智能治理。经过这被迫冷静思考的一年，或许会为未来的发展提供新的思路。比如，一直深入参与人工智能治理研究的剑桥大学的贺尚安（Seán Ó hÉigeartaigh），在疫情的背景下，开始反思关于人工智能的过度炒作是否值得。他认为技术为改善疫情所能提供的帮助是比较有限的，人们还是应该回归问题的本质，比如政府在公共卫生方面需加大投入。还有一些专家提出，对于人工智能治理的研究已经有很多，但是在疫情的场景下，已有的研究无法有效地转化为政策。

继续编制这份全球观察报告也让我们有机会邀请去年未邀请到的专家和机构。比如，在去年报告发布后，有些学者批评我们忽略了拉丁美洲、非洲的发展中国家的声音。因此，今年我们特别邀请了数位拉丁美洲的专家，介绍这个有数亿人口的地区正在为人工智能治理所开展的工作。我们也有幸邀请到了非洲的代表，来介绍他们的思考和相关工作。

我们对52位专家的观点简短介绍如下：

技术社区

技术专家始终是人工智能治理的重要力量，图灵奖获得者约翰·霍普克罗夫特（John Hopcroft）教授今年继续从科学家的角度提出看法。他提到了目前人工智能治理所面临的七个方面的问题，并强调了“监督”的重要性。

人工智能促进协会（AAAI）主席巴特·塞尔曼（Bart Selman）提醒了大家一个重要的科学问题：人工智能技术通常以一种我们人类完全陌生的方式运行。基于这样一个科学基础，人工智能的治理必须通过政策专家与技术专家的密切合作才能完成。

世界工程组织联合会（WFEO）龚克主席介绍了世界工程组织联合会正在积极推动人工智能伦理治理的相关工作，他还着重强调了人工智能绿色发展的概念。

OpenMined团队这次的观点令人惊艳，他们提到了继安全计算、联邦学习和差分隐私之后一种新的隐私解决方案——结构化透明度。

汉堡大学的张建伟教授提到了疫情时期人工智能发挥的积极作用，并进一步介绍了其团队在跨模态系统方面的研究进展。

科学家同样考虑技术引发的政治和人文问题。麻省理工学院的亚历克斯·彭特兰（Alex Pentland）教授提到了一个正在浮现的新问题，那就是跨越主权和制度边界的数字平台如何治理。

科学人文畅销书作家布莱恩·克里斯汀（Brian Christian）提出了一个根本性的问题，既是科学问题，也是人类社会长期存在的问题，那就是如何在组织内部和组织之间达成“一致性”。

专注于人工智能安全的专家罗曼·亚姆波尔斯基（Roman V. Yampolskiy），似乎表达了一种相对悲观的未来情景，那就是人工智能也许是无法治理的。

产业界

普华永道作为一家大型国际咨询公司，一直观察国际产业界在人工智能领域的动态。该公司合伙人安纳德·劳（Anand S. Rao）提到了一个有意思的现象：疫情背景下人工智能依然是产业界的宠儿，全球AI风险投资资金在2020年持续上升，但是很少有企业实现了完全嵌入式和自动化的AI风险管控。

蒙特利尔人工智能伦理研究所的创始人阿布辛克·库布塔（Abhishek Gupta）是企业家出身，更能理解人工智能治理原则为什么无法有效落地，所以他提出了基于从业者可行的“实践模式”。

艾琳·索拉曼（Irene Solaiman）代表OpenAI这一全球有影响力的组织，解释了GPT-3的发布方案。吸取了GPT-2发布时的经验，他们采用了一种通过API发布的全新方式，经过批准的用户可以使用API密钥进行访问。

商汤联合创始人杨帆介绍了商汤作为一家全球领先的人工智能公司，正在为“可持续发展”所做的切实努力。

作为人工智能社会影响的观察家，鲁子龙（Danil Kerimi）提到了人工智能正在推动社会契约的改写。

硅谷的风险投资人史蒂文·霍夫曼（Steven Hoffman）讨论了具体问题，他认为数据的商品化也许是一个错误的问题解决方案，人们应该把重点放在遏制对于数据的滥用上。

智能医学专家奥马·科斯提拉·雷耶斯

（Omar Costilla-Reyes）提出社会制度如何适应人工智能的发展，以及在医学领域，如何采用新的认证方式。

政策研究界

牛津大学的艾伦·达福（Allan Dafoe）和亚历克西斯·卡利尔（Alexis Carlier）提到了具体的人工智能治理落实制度，比如国际人工智能顶级会议，已经要求所有参与者提交论文时签署责任声明。

来自生命未来研究所的贾里德·布朗（Jared Brown）提出了一个具体的问题，人工智能的风险到底是什么，以及如何识别、评估和管理。

社会学家彼德拉·阿维勒（Petra Ahrweiler）和马丁·纽曼（Martin Neumann）提到了制定人工智能治理法案，需要对人类价值观建立知识库清单，并用人工智能技术来实现这一清单。

法学家马拉维卡·杰亚拉姆（Malavika Jayaram）强调，数据去殖民化工作应该是让每个地区努力维护数据的主权和自主权不用完全符合西方的标准。

同样是法学家的娜塔莉亚·苏姆哈（Nathalie Smuha）则强调，治理的原则要转化为可执行的法律条款。

知名科技伦理专家温德尔·瓦拉赫（Wendell Wallach）谈到人工智能在疫情这一特殊场景，以及复杂的地缘政治形势下，全球的团结合作是至关重要的。

牛津大学的方淑霞（Marie-Therese Png）积极维护发展中国家在全球人工智能治理中的话语权，她担心人工智能的发展可能会带来新一轮的殖民化。

逻辑学家马库斯·克纳夫（Markus Knauff）从认知心理学的角度讨论了人工智能未来的发展方向。

对于发展中国家而言，不仅是有没有考虑到人工智能治理的问题，而且还要考虑发展中国家的监管能力。印度Tandem研究院创始人乌尔瓦希·阿内贾（Urvashi Aneja）强调，低下的监管能力和机构能力使基于风险的治理方法的适用性受到了进一步的挑战。

国际组织

联合国区域间犯罪和司法研究所（UNICRI）人工智能与机器人中心主任伊拉克利·贝里泽（Irakli Beridze）介绍了人工智能治理实操化的一个可喜进展，他们和国际刑警组织环球创新中心已开始着手合作开发注重操作性的工具包，以帮助执法机构负责任地使用AI，并支持和引导负责任的AI设计、开发和部署。

曾担任联合国技术顾问，负责领导联合国秘书长数字合作路线图中人工智能相关工作的高丹青（Danit Gal），介绍了她参与的这一工作。她敦促人工智能领域必须让发展中国家和代表性不足的群体尽可能公平地参与到全球人工智能合作当中。

来自剑桥大学的贺尚安（Seán Ó hÉigeartaigh）是人工智能全球合作伙伴关系的人工智能与疫情工作组的成员，全球疫情的特殊场景激发了他的反思，有关AI的炒作如此之多，但是实际为疫情提供的技术性解决方案却不多。过于关注技术问题反而会令我们忘了问题的本质，即加强医疗和公共卫生的投入。同时，在疫情中是否使用数字追踪技术的争论提醒了我们治理问题的复杂性。

国际人工智能治理的观察家塞勒斯·霍德斯（Cyrus Hodes）总结了在疫情这一年中代表性组织的代表性工作。他提到的一些工作在本次报告的其他参与者中也有所提及。他强调，切实行动是制定原则之后的关键。

国家和地区

曾任联合国大会主席办公室高级顾问的尤金尼奥·加西亚（Eugenio Vargas Garcia）先生认为，人工智能如此重要，因此全球需要团结，从而共同应对。他介绍了联合国教科文组织以及联合国秘书长数字合作路线图关于人工智能伦理治理的相关进展。加西亚近些年大声疾呼在国际人工智能治理的讨论中，发展中国家的声音不能被忽视。

欧洲是人工智能治理的积极推动者，欧洲议会议员伊娃·凯莉（Eva Kaili）介绍了欧洲确定人工智能治理全球领导者的努力，她强调了欧洲的人工智能治理模式：监管机构负责制定原则，市场则通过定义产品或服务的标准来实施这些原则。

作为欧盟人工智能高级别专家组的协调人，夏洛特·斯蒂克斯（Charlotte Stix）介绍了在疫情这一年中，欧盟的人工智能治理框架仍有条不紊地在推进。她介绍了这一年欧盟发布的《人工智能白皮书——通往卓越和信任的欧洲路径》，并着重介绍了这一监管提案对可信人工智能的七大关键要求。

加州大学伯克利分校的林晨力（Caroline Jeanmaire）介绍了美国前总统特朗普在其任期结束之际发布的关于AI设计、开发和应用的九项原则，以及美国和欧盟在确保人工智能风险可控方面的进展。

日本的江间有沙（Arisa Ema）介绍了第二届法德日人工智能研讨会在2020年召开的情况。第一届会议的联合声明强调以人类为中心，而在疫情暴发之年举办的第二届会议，联合声明强调了全球性问题合作解决的重要性。

印度这一信息技术大国，在2020年针对人工智能治理，也有了可喜的进展。拉杰·谢卡尔（Raj Shekhar）介绍了印度在个人数据保护和负责任AI的各方面讨论。

新加坡科技与设计大学的李光耀创新型城市中心潘竞宏（Poon King Wang）团队在帮助新加坡政府解决人工智能时代的就业问题，他们提出了“跨行业GPS”方案，是把这一问题从学理讨论落实为具体政策的一个极佳案例。

在非洲工作的维克多·法姆波德（Victor Famu-bode）提到了一个可喜的现象，在2020年，非洲各种政府开始重视AI的道德影响。

拉丁美洲有超过6亿的人口，但是在人工智能国际讨论中的声音一直不够。为此，今年我们邀请了多位来自南美的代表。来自阿根廷的奥尔迦·卡瓦利（Olga Cavalli），来自巴西的埃德森·普雷斯特（Edson Prestes），来自哥伦比亚的康斯坦萨·戈麦斯蒙特（Constanza Gómez-Mont），来自多米尼加共和国的让·加西亚·佩里希（Jean García Periche），来自智利的何塞·古里迪·比斯托（José Guridi Bustos），他们从各自的角度介绍了拉丁美洲在人工智能治理领域所取得的进展。他们有一个共性的观点，那就是拉丁美洲应该在全球人工智能治理讨论中发出自己的声音。他们还一致认为，虽然拉丁美洲在人工智能的技术研发和应用上并没有一些国家那么先进，但是他们的治理模式一定不能盲目参照先进国家。

国家和地区（中国）

中国在人工智能领域发展迅速。清华大学战略与安全研究中心主任傅莹谈到了人工智能对国际安全带来的冲击，并提出中国愿意和各方展开对话和合作。

中国科技部新一代人工智能发展研究中心主

任赵志耘介绍了中国在疫情之下，稳步推进2019年发布的治理原则的进一步落地。

清华大学的苏竣教授正在牵头推动一项综合性的社会实验，其目的不仅是要对人工智能的社会影响进行实验性的全面评估，更重要的是期望为中国“建设具有人文温度的智能社会”奠定理论基础。

中国科学技术发展战略研究院的李修全谈到了从技术和制度两方面推动人工智能治理的思考。

复旦大学的王国豫教授介绍了中国的计算机专家与哲学家们，通过中国计算机学会等组织，在2020年所推动的人工智能治理工作。

上海举办的世界人工智能大会在全球同行中有着广泛的影响。上海市科学学研究所的王迎春介绍了2020年世界人工智能大会治理论坛的相关情况。

基于2019年的年度观察报告，我们总结认为，全球人工智能治理体系正在孕育形成。基于2020年的年度观察报告，我们仍保留这一观点，但是这一年有着特殊的属性，就是对全球人工智能治理体系的“反思”。

每位读者或许都可以从这52位专家的文章中，归纳出技术社区、产业界、政策研究界、国际组织、国家和地区等在2020年的进展，也可以从所有的进展中，发现新趋势的浮现。当然，全球人工智能治理是一个全球性的议题，即使有52位专家的参与，也不可能全面介绍所有的进展。能够提供一定的交流平台正是我们的初心所在。

执行编辑：李辉、谢旻希(特约)

李辉，上海市科学学研究所副研究员。国务院发展研究中心创新发展研究部访问学者（2020年5月-2021年4月）。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、2020和2021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，牛津大学人工智能治理中心政策研究员、北京智源人工智能研究院面向可持续发展的人工智能协作网络协调员、上海市科学学研究所全球人工智能治理研究项目联合负责人、海国图智研究院人工智能顾问、人工智能合作组织(Partnership on AI)高级顾问。谢旻希致力于有益安全人工智能的发展和合作，曾为顶尖的人工智能企业提供咨询服务，包括百度、OpenAI、谷歌DeepMind。他担任多个人工智能安全学术会议的委员，包括AAAI、IJCAI和AI&FM。

致谢

感谢所有作者的认可和支持。感谢参与2019年度观察的50位专家，让我们这份报告得以“开启”。感谢参与2020年度观察的52位专家，让我们这一份报告得以“延续”。感谢两年全球80位专家的共同努力，为这份报告奠定了基础，相信在新的一年，我们的报告可以真正成为一份“系列”报告。

还要感谢所有作者的辛苦努力。虽然名义上这是上海市科学学研究所发布的报告，但实际上是由所有作者共同发布的。这份报告的一个特色就是，作者同时还承担着编辑甚至组稿人的角色。我们署名的“编辑”团队对文章的内容基本上不做修改。每位作者既要为自己的文章内容负责，也要为自己的语法词句负责。很多作者实际上还是这份报告的组织者，帮助邀请新的适合的作者加入。比如曾在联合国大会主席办公室担任高级顾问的尤金尼奥·加西亚（Eugenio Vargas Garcia）先生，帮助我们引荐了本次报告所有的南美作者。

感谢上海市科学学研究所同事们的支持。虽然这份报告的制作是由所长亲自领衔的团队完成，但是在项目立项、资金支持以及后期推广等方方面面，上海市科学学研究所的领导团队、科研管理部门、行政主管部门，都给予了坚定的支持。

感谢资助方。这份报告的制作，有赖于上海市科学学研究所的经费支持。在去年发布了报告之后，我们有幸得到了一些基金的关注和认可。其中，我们还有幸得到Skype联合创始人让·塔林（Jaan Tallinn）的支持，为我们今年的报告制作提供了资助。

感谢志愿者。去年报告吸引了一些志愿者的加入。来自加州大学伯克利分校的林晨力，帮忙进行了框架上和介绍部分的润色。来自亚洲数字中心的戴逸司（Dev Lewis）、来自中国社会科学院的陈雅琨、来自湖南师范大学的胡晓萌也提供了很多咨询建议。

来自哥伦比亚大学的聂蕴哲，担任本次报告的项目主管，负责联系各位专家，付出了大量心血。

感谢英文审校团队。来自牛津大学区域研究的博士生季采璇、牛津大学材料工程的张凯博士以及伯明翰大学硕士李海龙，都投入了大量精力对所有英文文章进行了逐字逐句校对。

感谢中文审校团队。来自上海市科学学研究所的徐诺、华东政法在读博士生陈志豪对中文版报告进行了逐字校准。湖南师范大学孙保学副教授和牛津大学博士生蒲星月对有关专业内容进行了审校。

感谢上海市科学学研究所科技与社会研究室王迎春、陈秋萍、博士后瞿晶晶、实习生肖迪，他们为了中文版报告的最终定稿，进行了细致的修改。

对所有全球同仁的帮助，我们一并致谢！很荣幸能与所有人一起合作，为如此重要的事情而共同奋斗。

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：
技术社群的探索

关于人工智能治理的思考

约翰 · 霍普克罗夫特（John E. Hopcroft）

人工智能（AI）将为人类带来极大的裨益，并成为我们生活的重要组成部分。监管是为了确保各组织负责任地使用AI，并使其成为加快经济增长的重要手段。以下七个方面是我们需要考虑的问题：

1. 更新法律制度：例如，无人驾驶汽车发生事故时，应由谁来负责？是车主、制造商，还是AI系统的开发商？特别是对公司而言，法律制度的更新对其投资决策具有重要意义。

2. 明确数据所有权：我们在进行网络搜索时， 通过搜索引擎公司会按照IP地址储存我们搜索的所有关键词。对这一数据库进行分析后，该公司就可以推断出我们的身份、购物地点，以及我们所购买的商品。甚至根据某一个IP地址对应的搜索记录，就能判断出我们的性别、爱好和其他个人信息。我们是否同意搜索引擎公司出售这些信息呢？

3. 公平性：AI应受到监管，并确保所有阶层的人都能享受其带来的乐趣，而不是只有富人和当权者受益。

4. 消除偏见：AI系统是通过数据进行训练的，因此源数据中的偏见会影响AI系统存在同样的偏见。例如，如果在数据库中高级职位大多为男性，那么AI系统可能只会推荐男性作为高级管理者。

5. 可解释性：许多AI系统都用的是黑箱算法，只会根据查询内容提供答案，而不会对答案做出解释。如果一个求职者被面试系统刷下来，他可能会坚持要求得到一个解释。那么他被拒绝到底是因为没有相关资质，还是其他的原因呢？

6. 人脸和情绪识别：人脸和情绪识别技术有很多重要应用，但不应被用于识别种族或民族血统、人格特质或心理健康状况等。

7. 社交网络传播：人们利用Facebook和Twitter等平台向用户传递大量真实或虚假的信息。随之可能对一个国家产生正面或负面的影响。AI可以被用于操控信息传播，故对此进行治理极其必要。

这些只是需要考虑和监管的一小部分问题。还有一些问题涉及人机交互的安全性、保障性和伦理困境等。人们还需要考虑应在什么层面上考虑相关问题。哪些问题应由政府负责，哪些问题应该由产业界处理？换言之，AI治理政策的制定需要政府、学术界和产业界代表的共同参与。

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，图灵奖获得者。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。他于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 · 福德（Joseph C.Ford）教授。于1987年至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年获得了人文荣誉博士学位。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，他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（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。

人工智能治理从理解开始

巴特·塞尔曼（Bart Selman）

很高兴看到最近AI治理活动的水平显著提高。有效的AI治理对于AI技术在社会中的良好应用至关重要。AI治理的目标是以保障人类利益为核心，确保AI技术能真正造福社会。

在本篇报告中，我将重点强调实现有效AI治理的关键挑战。良好的AI治理需要我们深入理解AI方法所带来的机遇及其局限性。在一系列认知型任务方面，AI系统的表现正逐渐赶上甚至超过人类。然而，我们必须意识到，AI系统获得这种能力的方式与人脑处理这类任务的方式大不相同。其中一项根本性的困难是，我们在考虑AI技术时，倾向于赋予这些系统人格化的特征。换言之，我们会假设AI系统执行认知型任务时采用的方法与人类类似。

基于深度学习方法的机器翻译近年来取得了长足的进步。当前的模型能在数十个语言对之间进行合理有效的翻译。令人惊讶的是，这些译文并非是在真正理解原文的基础上所获得的。我们甚至都很难想象在未理解语言的前提下如何对两种语言进行翻译。但这就是语言翻译系统的现状——以一种近乎陌生的方式运作。

在为各类数据阐释与决策任务训练的高度复杂的深度模型中，这种“陌生的”运作方式也十分明显。要求AI系统对其决策做出解释（即“有权获得解释”）似乎并无不妥。然而，研究人员发现，这类提供解释的尝试容易导致伪解释，即让使用者自身满意，但却并未准确体现系统的内部决策流程。

大体而言，未掌握人类视为理所当然的知识和常识的新兴AI系统正在不断涌现。有鉴于此，以紧急医疗情形为例，我们可以指示自动驾驶汽车忽视某些交通规则，将我们尽快送往最近的医院。人类驾驶员能够意识到这种做法应以不会给他人带去危险为前提。这在人类的眼中是一项常识，却需要明确编写进自动驾驶汽车的代码中。为了实现合理有效的AI治理，我们需要注意，人类对于AI技术的运作方式往往是极为陌生的。因此，有效的AI治理需要AI研究人员与政策制定者的紧密合作才能实现。

作者简介　巴特·塞尔曼（Bart Selman）

巴特·塞尔曼是美国康奈尔大学计算机科学和工程系约瑟夫·福德（Joseph C. Ford）特聘教授，也是人工智能促进协会（AAAI）主席。AAAI是面向AI研究人员和从业者的主要国际专业协会。此外，他还是一项国家级研究项目的联合主席，该项目旨在确定AI研究的路线图，以指导美国政府的AI科研投资。塞尔曼教授此前曾任职于贝尔实验室。他的研究领域包括人工智能、计算可持续性、高效推理程序、机器学习、深度学习、深度强化学习、线性规划、知识表征学习，以及计算机科学与统计物理学之间的关系。塞尔曼教授曾编著或合著超过150篇作品，其中6篇获得了最佳论文奖，3篇获得了经典论文奖，这些论文发表于《自然》《科学》《美国科学院院刊》以及一系列有关AI和计算机科学的会议和期刊上。他曾获得康奈尔大学斯蒂芬·迈尔斯优秀教学奖、康奈尔大学杰出教育奖、美国国家科学基金会（NSF）事业奖和阿尔弗雷德·斯隆研究奖。塞尔曼教授是美国人工智能协会（AAAS）会员和美国计算机协会（ACM）会员。他曾获得首届国际人工智能联席会议（International Joint Conference on Artficial Intelligence，IJCAI）约翰·麦卡锡研究奖，这是对处于职业生涯中期的人工智能研究人员所能获得的最高奖项。

从工程视角看人工智能的发展与治理

龚克

如今，人工智能（AI）已经从基于科学论文和技术实验室的研究，发展到工程应用阶段，进入了人们的日常生活。这无疑是AI历史发展的转折点。应用和工程师已经成为了影响AI应用进一步发展，以及AI治理的关键因素，对于确保AI能够造福地球和全人类而言至关重要。

AI是一种助力人类发展的新工具，AI发展和治理的最终目标是确保AI有利于可持续发展。

人类区别于其他动物的重要特征是能够制造工具。从简单的石器和铁器，到杠杆、滑轮、各类机械和电力，再到计算机和现代信息网络，人类已经利用这些工具提高了自身的劳动能力、智能水平、生存能力以及幸福指数。因此，人类历史通常是人造工具的进步史，人造工具则是时代的标志，例如石器时代、铁器时代、机器时代、信息时代等。从这个方面来看，AI也不过是一个推动人类发展的新工具，正将人类带往智能时代。

必须要指出的是，人类发明和使用工具的目的在于解决生存和发展问题。目前，可持续发展是人类面临的最大问题。因此，AI应该成为一种能推动人类可持续发展的有力工具。这同时也是AI发展和治理的最终目标。2020年，AI领域的主要进展在于，研究人员分析了AI在实现可持续发展目标方面的作用，而这种作用也已经引起了许多国际组织的关注。为了引导和支持AI发展与治理，我们亟须通过更加全面和富有见地的研究来探索AI和可持续发展目标之间的关系。

工程界应担起责任，成为多利益相关者共同发展和治理AI的重要参与者。

世界工程组织联合会（WFEO）意识到了其有责任推动AI造福人类和环境，并通过其下属的创新技术委员会常务技术委员会成立了一个由产业界和学术界专家组成的跨学科工作小组。2020年3月4日，WFEO在首个促进持续发展世界工程日上发布了一份题为《促进大数据和AI创新与应用在工程中的责任管理方式》*(Promoting Responsible Conduct of Big Data and AI Innovation and Application in Engineering)*的立场声明。该声明中提出了七项原则，分别是：

- 造福人类和人类的生存环境
- 确保包容性、公平性、公众意识和自主权
- 在尊重隐私和数据完整性的基础上实现开放共享
- 透明性
- 可问责性
- 和平、安全和保障性
- 合作

许多政府、非政府组织和公司等都已声明了AI应用原则。虽然这些原则在很大程度上是相同的，但我们仍需要通过大量努力达成全球共识。总体而言，工程尚未在AI治理中发挥重要作用，而不同行业和社会团体之间的交流也有所欠缺。为此，我们亟须开展涵盖多个利益相关者的全球对话。

AI的绿色发展应成为AI发展和治理的重点。

AI在提高各类工程项目的生产力、质量、安全性和效率方面潜力惊人，因此可以说AI正在从各个方面增强和改变工程应用。

我们惊喜地看到，许多AI应用能够有效提高各类产品和家用电器的能源效率。然而，AI产品的功率效率要远低于人类员工也是众所周知的事实。但是，绿色和低碳发展尚未在AI研发和应用中得到重视。因此，必须呼吁业界充分了解AI开发的严格限制，并更加重视AI的绿色发展和应用，而这也应该成为AI治理的重点之一。

此外，还需要注意的是，虽然许多针对AI安全和保障性的研发工作已经展开，但是与大约20年前家用电脑在开始被大规模应用时所出现的与之配套的杀毒软件和服务市场相比，用于保障AI安全的产品和服务仍然缺乏。

作者简介

龚克

龚克教授是世界工程组织联合会（WFEO）主席和中国新一代人工智能发展战略研究院执行院长。龚教授是信息通信技术方面的专家，获有奥地利格拉茨技术大学的电工电子系通讯与电波传播专业博士。他曾于1987-2006年在清华大学任职，担任清华大学副校长及信息科学技术国家研究中心主任；此后，他又先后任职天津大学校长（2006-2011年）和南开大学校长（2011-2018年）。龚教授于2009年加入WFEO，工作至今，曾先后任职信息和传播委员会主席和创新技术委员会主席。2019年，他当选WFEO主席，任职至今。自2017年中国工程院与天津市政府联合成立中国新一代人工智能发展战略研究院以来，龚教授一直领导着该院的发展。2013-2017年，龚克教授曾任联合国秘书长潘基文的科学顾问委员会的委员。此外，他还曾在中国科学技术协会等组织机构担任过重要的行政职位。

如何善用数据和人工智能

OpenMined 团队

我们如何将数据用于实益用途，同时避免其潜在的危害？随着人工智能的发展，这个问题变得越来越重要。解决这一问题的办法是实现所谓“结构化透明度”的目标。

无数的现代活动和服务都要求获取敏感信息和个人信息，以提供有益的服务。这种交换有时会导致对用户不利的交易、忽视数据隐私以及数据滥用造成的伤害。

然而，数据的使用是人工智能算法的应用和发展基础。

面对这一挑战，诸如安全计算、联邦学习和差分隐私等许多隐私增强技术（PET）应运而生，旨在保护隐私的同时使用数据。然而，目前尚没有一种技术能够彻底解决数据隐私问题。

但是，在某些特定的场景使用时，这些技术可以保证数据只能被用于有预先目的或被批准的用途。这些技术可以帮助人工智能研究人员和治理机构对特定人群对象看到的特定信息进行非常精确的技术或社会安排。

在本文中，我们提出了一个实用的框架，有助于思考新兴技术如何帮助我们实现所需的信息流。为了实现结构化的透明度——换句话说，在减轻危害的同时确保数据仅用于实益用途——必须考虑在何时以何种方式与谁共享哪些信息。为此，我们建议考虑以下五个要素。这五个要素构成了有助于思考结构化的透明度框架，重要的是，指明了哪些新兴技术可以帮助您实现这一目标。

1. 输入端的隐私保护：允许在不泄露内容的情况下使用隐藏信息。输入端的隐私保护主要来自密码学领域——公开密钥密码学、端到端加密、安全多方计算、同态加密、函数加密、乱码电路、不经意内存、联邦学习、设备分析和安全飞地是几种流行（和重叠的）技术，能够保证输入端的隐私得到保护。

2. 输出端的隐私保护：允许读取和接收信息，并防止数据反向溯源，以便隐藏输入数据。技术输出隐私工具（主要是差分隐私和相关技术）可以针对数据点反向溯源算法改造可能性规定严格的上限。

3. 输入验证：确保数据源的稳定性和可靠性。输入验证技术大都依赖于公钥基础设施（SSI、密钥透明性等）、加密签名、具有动态安全性的输入隐私技术和零知识证明的组合应用。

4. 输出验证：确保对给定信息流进行的计算的合法性。输出验证工具的主要限制是验证人员必须检查数据，以执行验证。

5. 信息流治理：保证满足上述四点，从而确保结构化的透明度。换而言之，信息流的隐私安全需要每个环节的保护。从技术上讲，信息流治理的最佳例证是安全多方计算（SMPC），因为信息流受到了公证机构的监控。然而，信息流从根本上来说也受到系统激励措施的驱动，无论是通过优化标准（屏幕时间或诊断准确率）的算法激励措施还是当事人的动机。

随着数据治理和数据隐私问题日益严重，我们希望这一框架能为如何将新兴的隐私技术应用于特定用例提供一个清晰的思路。

作者简介

Claudia Ghezzou Cuervas-Mons, Emma Bluemke(个人意愿不披露个人信息），Pengyuan Zhou和Andrew Trask —— OpenMined写作团队成员

Claudia Ghezzou Cuervas-Mons

Claudia是一家合同研究组织（CRO）的临床数据协调员。她主攻生物化学（BSc）和神经科学（MSc），目前正在帝国理工学院Bentley博士的指导下开展其硕士论文研究，选题为CT扫描中缺血事件的自动量化。

Claudia自身还对机器学习如何促进医学和科学发展，以及如何通过工具优化疾病检测和治疗效果抱有兴趣。此外，她还乐于探索AI进步的社会意义，以及如何运用AI技术造福人类社会。

Pengyuan Zhou

Pengyuan是在赫尔辛基大学计算机科学系和香港科技大学系统与媒体实验室的博士后研究员，Pan Hui是他的指导教授。2020年5月，Pengyuan在协作网络实验室的Jussi Kangasharju教授的指导下取得了赫尔辛基大学的博士学位。他重点研究移动边缘计算与通信系统、智能网联汽车以及边缘人工智能。对网络边缘的计算和网络资源的大量需求是推动其研究的关键动力。

Andrew Trask

Andrew是牛津大学的一名在读博士生，从事保护隐私的AI方面的研究。他还是OpenMined团队的负责人，一个致力于降低使用隐私保护技术壁垒的超过1.1万名成员的集体。

后疫情时代下以人为本的人工智能和机器人研究与发展

张建伟

2020年，新冠疫情肆虐。在这极不寻常的一年里，AI在设法稳步推进在由单一标准优化指导下的受监督深度学习方法中纳入更多自上而下的方法、跨模态学习和无监督架构，同时，不得不面对疫情所带来的全球挑战。我们看到，AI在基础研究和新应用层次方面取得了显著的进展。这些进展大多涉及基于数据/算法的居家办公模式，影响了AI治理的理论和实践。

我们亟须通过实用的AI和机器人系统遏制新冠病毒的迅速传播，并恢复正常的工作和生活节奏。这些要求使AI治理的重点聚焦在针对病毒感染途径的精确建模、实时检测方法的快速开发、有效的疫苗和药物，以及能够协助医生/护士，并代替人类完成配送、收货、工厂装配等劳动密集型任务的自主机器人系统。

我们已相应地开发了一种AI方法，通过基于互动的持续学习和对个人病毒携带率的推断，来确定出现无症状新冠病毒携带者的可能性，从而评估传染几率。与传统的接触者追踪方法相比，该方法大幅减少了寻找潜在无症状病毒携带者所需的筛查和隔离时间，降幅高达94%。同时，我们正着手集成人机互动、多层次学习和机器人决策应用技术，以开发一种安全自主的智能机器人系统，用于进行鼻咽拭子和口咽拭子采样，从而保护医务工作者，大幅降低其接触感染风险。通过与德国初创企业PixelBiotech公司合作，我们已经开始用AI增强复合smFISH（单分子荧光原位杂交）探针的成像数据分析，将新冠病毒的检测时间缩短至15分钟。总体而言，我们相信，AI和机器人技术将通过促进更精确的传播过程建模和高效的自动病毒检测，为全球抗疫带来越来越多的积极影响。

我们将基于中德跨学科合作项目TRR169“跨模态学习的自适应、预测和交互”，继续研究和开发人类增强技术，以促进人机物理协作。此外，我们还正致力于与各神经科学与心理学实验室携手开发人类增强和支持应用，如神经计算表征和神经疾病的治疗应用。此类以人为本的长期研发工作代表了AI最重要的垂直应用之一。

在TRR169项目中，我们通过来自汉堡和北京的合作伙伴之间的成熟合作，进一步提升了对跨模态系统的理解、建模和应用水平，以及对跨模态学习的神经、认知和计算机制的理解和集成水平。我们在一些领域，尤其是深度学习领域（如算法、软件和应用）取得了重大进展，这激励着我们追求实现一系列综合性更强的目标，作为面向未来的强AI研究课题，包括新的学习架构/策略、稳定性、预期与预测、泛化与转移，以及标杆管理。总体而言，我们的使命是制定透明且可解释的类脑AI方法，这是AI治理的一项重要技术特征。

作者简介 张建伟

张建伟是德国汉堡大学信息学科学系教授及多模态技术研究所（TAMS）所长。他于1986年以优异的成绩获得清华大学计算机系工程学士学位、于1989年获得该系的工程硕士学位，此后又于1994年获得德国卡尔斯鲁厄大学计算机系实时计算机系统与机器人研究所的博士学位。其研究方向包括认知传感器融合、稳定机器人感知、智能机器人架构、多模态人机交互、机器人的灵巧操作以及类脑机器学习等。他在这些领域发表了超过400篇期刊论文、会议论文和技术报告，并出版过6本著作。他在智能组件和系统领域拥有超过40项专利。他的研究成果已被应用于现实场景中的机器人系统，例如医疗助理系统、康复系统、先进3C组装系统、工业流程在线质量监控系统等。张建伟是德国研究联合会（DFG）与国家自然科学基金委员会（NSFC）联合发起的跨区域合作研究中心SFB/TRR169“跨模态学习”项目，以及一些欧盟、德国和工业AI项目的协调人，获有多项最佳论文奖。他是2012年国际电气电子工程师协会（IEEE）多传感器融合与集成智能系统国际会议、2015年IEEE/RSJ智能机器人与系统国际会议，以及2018年以人为本的机器人与系统国际研讨会的大会主席。此外，他还是德国汉堡科学院的终身院士。

数字平台的人工智能与数据治理

亚历克斯·彭特兰（Alex Pentland）

对国家与国际间的贸易往来进行现代化和数字化治理，从而达到更高效、更透明和更包容是全球的关键性优先事项，且全球范围内目前已就此开展了大量工作，但是目前这些工作大多零碎分散且进度缓慢。

随着国家开始试点部署能够为金融、贸易和物流、身份验证、欺诈检测和分析（例如AI）提供统一框架的数字平台，对数字平台的治理变得异常迫切。例如，中国正在将其对“一带一路”的现有投资转移到自己的数字系统上，相比西方系统而言大大提高了灵活性，并且显著降低了成本。新加坡已为其在淡马锡主权财富基金内的投资开发了一个与中国类似的数字贸易和物流架构，而瑞士最近也在麻省理工学院连接科学项目的帮助下推出了瑞士信托链（Swiss Trust Chain）平台。最后，大多数主要经济体已经推出或正在认真考虑推出国家数字货币。我们已经参与推出了两种此类货币，不久还将协助推出一种主要交易货币的数字版本。

这些系统有望将世界上大多数的贸易整合到高效、统一的框架之中，并且这些框架可以跨越主权和制度边界的限制，无缝进行相互操作。但是，它们的可问责性、包容性和治理水平可能无法令许多国家满意。所有国家的当务之急是参与到这些数字治理系统的标准规范制定和部署之中。

也许所有新的数字治理系统要解决的第一个挑战便是拯救这个世界破败不堪的经济状况。如果国家间无法达成合作，我们将会面临“逐底竞争”的风险，体量较小的国家将承担最多的损失。而且，与第二次世界大战结束时不同，这些新数字贸易平台的部署将使各国可能通过比法定贬值更为隐蔽的方式乞求邻国帮助。

这表明我们需要应用新的“布雷顿森林体系”多方举措，以运用类似于中国、新加坡和瑞士开发的更有效、更安全和更具包容性的数字平台，对多边制度进行革新。与第二次世界大战时所采取的举措不同，这种协作不仅必须围绕银行和金融业，而且还必须紧密依赖于数字技术标准（例如IEEE创建的标准）以及衡量和预测金融、可持续发展和社会因素之间的相互作用所需的计算社会学。

作者简介 亚历克斯·彭特兰（Alex Pentland）

亚历克斯·彭特兰教授是麻省理工学院连接科学实验室与人类动力学实验室的主任，此外还是麻省理工学院媒体实验室与印度亚洲媒体实验室的联合创始人和前主任。他是全球被引次数最多的科学家之一，最近与谷歌创始人和美国首席技术官一同入选《福布斯》“全球最具影响力的7位数据科学家”。彭特兰教授获有诸多奖项，例如《哈佛商业评论》麦肯锡奖、美国国防部高级研究计划局（DARPA）颁发的互联网成立40周年纪念奖，以及因在数据隐私方面做出的贡献而获得的布兰迪斯奖。他是谷歌、美国电话电报公司、日产汽车和联合国秘书长咨询委员会的创始委员，同时也是一名连续创业者，与他人合伙创办了十几家公司。他还是数据透明度实验室、哈佛大学—ODI—麻省理工学院联合Data-Pop联盟（the Havard-ODI-MIT Data-Pop Alliance）和数据驱动设计研究所等社会企业的联合创始人、美国国家工程院院士，以及世界经济论坛领导人。多年来，彭特兰教授已指导了超过60名博士生，其中有近半的博士生成为了一流机构的终身教员，25%成为了产业研究小组的负责人，另25%则自行创办了公司。他和他的学生共同引领着计算社会科学、组织工程、可穿戴计算技术（谷歌眼镜）、图像理解和现代生物测定学领域的前沿发展。他最新的著作包括由企鹅出版社出版的《社会物理学: 社交网络如何使我们更聪明》（*Social Physics: How Social Networks Can Make Us Smarter*），以及由麻省理工学院出版社出版的《诚实的信号：它们是如何塑造我们的世界的》（*Honest Signals: How They Shape Our World*）。彭特兰教授还有一些有趣的经历，包括与英国皇室成员以及印度总统共同就餐，在巴黎、东京和纽约举行时装秀，以及开发一种从太空计算海狸数量的方法。

一致性：人类永存的问题

布莱恩·克里斯汀（Brian Christian）

在过去的十年里，我们见证了深度学习如何使网络能够执行复杂的任务，如在不需要任何手动的“特征工程”的情况下识别人脸。在最近十年中，也陆续出现了类似的有价值的技术。

以Twitter这样的社交网络为例，其员工无须手动识别平台上“健康”参与的评论、分享和点赞比例，即可简单确定特定社区的“健康”参与与否。然后，系统就可以推断出哪些具体的指标可以预测一个社区的健康度，接着相关人员便能以一种可促进“健康”参与的方式对平台进行优化——这一切都可能在没有任何人明确定义到底是哪些指标构成了“健康”参与的情况下完成。（见 https://arxiv.org/pdf/2008.12623.pdf）

与许多技术能力的进步一样，这在缓解一个问题的同时又强调了另一个问题。我们解决了指标的问题，但判断的问题依然存在。例如，谁来决定什么是“健康”参与？谁负责监督？相比于用数字指明优先事项的工程问题，相关治理问题在当前更为重要。

在目前的情况下，词源学往往能提供意想不到的帮助。2014年，计算机科学家斯图尔特·拉塞尔（Stuart Russell）首次在人工智能中使用“一致性”这个词，但他是借用了经济学和管理科学领域的术语。这些领域几十年来一直在谈论如何在组织内部和组织之间实现“一致”价值观与利益的问题。

随着一个系统的工程师和他们创建的人工智能系统之间的“一致性问题”开始得到“解决”，我们必须把注意力转向更大的原始意义上的一致性问题：工程师团队和他们的经理之间、经理和高管之间、高管和股东之间、整个公司与其监管机构及其用户之间的一致性问题。

词源学提醒我们，“一致性”是人与人之间一直存在的一个问题。随着我们迈入人工智能时代，这将比以往任何时候都更符合现实情况。

作者简介　布莱恩·克里斯汀（Brian Christian）

克里斯汀教授是加州大学伯克利分校的访问学者，著有三部关于计算机科学的人文意蕴的畅销纪实作品：《最有人性的“人”》(*The Most Human Human*)、《算法之美：指导工作与生活的算法》(*Algorithms to Live By: The Computer Science of Human Decisions*)[与汤姆·格里菲斯（Tom Griffiths）合著]和《一致性问题》(*The Alignment Problem*)。他的书作入选了《纽约时报》编辑推荐书目、《华尔街日报》畅销书、纽约人最喜爱的年度书籍、亚马逊年度最佳科学书籍，以及《麻省理工技术评论》年度最佳书籍，并被翻译成了19种语言。其曾在谷歌、Facebook、微软、圣塔菲研究所和伦敦政治经济学院举办讲座，并获有布朗大学和华盛顿大学的哲学、计算机科学和诗歌方面的学位。

人工智能的可治理性

罗曼·亚姆波尔斯基（Roman V. Yampolskiy）

为了使未来的人工智能造福于全人类，人工智能治理倡议试图推进世界各国政府、国际组织和跨国公司间开展合作，以建立监管框架和行业标准，对人工智能进行治理。然而，对人工智能的直接治理是没有意义的，人工智能治理这个词所暗示的是对人工智能研究人员和创造者的治理，即对他们可以开发的产品和服务以及相关的方法进行管理。能否对从事人工智能研究的科学家和工程师实施管理取决于创建通用人工智能（AGI）的难度。

如果创建AGI所需的计算资源和数据收集工作在成本和人力资本方面与美国致力于开发核弹技术的曼哈顿项目相当，那么各国政府就可以采取诸多软硬兼施的策略指导研究人员，并按照自己的标准要求塑造未来的人工智能。另一方面，如果事实证明有一种更有效的方法来创建第一个AGI或“种子”人工智能，该人工智能可由多种途径成长为一个全面的超级智能，例如，一个青少年在车库里用一台价值1000美元的笔记本电脑就可达成这一目的（可能性较低，但并非为零），那么政府所进行的监管尝试将可能是徒劳的。我们注意到，历史上在软件治理方面所做的尝试（例如垃圾邮件、计算机病毒、深度伪造）只取得了非常有限的成效。仅仅依靠AGI是无法实现治理的，因为传统的责任分配方法和基于惩罚的执行方法并不适用于软件。

即使假定一个资源丰富、有利的治理案例，我们仍在人工智能的可预测性[1]、可解释性[2]和可控性[3]方面面临一些既定的技术限制。因此，为实现人工智能的治理，我们至少需要这三种能力以确保成功监管，且同样也只能部分实现，这意味着比人类更聪明的人工智能在某些重要方面是我们无法治理的。最后，即使人工智能治理是可以实现的，那些负责人也可能不愿意为其失败承担个人责任[4]，即便是在既定治理框架的背景下执行经过深思熟虑的行为举措。由此会出现一个具有出色能力和创造力但不受控制的AGI，最终可能会以隐性甚至显性的方式控制一些被我们委托以管理这些智能软件的机构和个人。

参考文献

[1] R. V. Yampolskiy，《人工智能的不可预测性：关于精确预测智能体所有行为的不可能性》，《人工智能与意识期刊》，第7卷（1），第109-118页，2020年。

[2] R. V. Yampolskiy，《人工智能的不可解释性和不可理解性》，《人工智能与意识期刊》，第7卷（2），第277-291页，2020年。

[3] R. V. Yampolskiy，《论人工智能的可控性》，预印本网站arXiv：2008.04071，2020年。

[4] R. V. Yampolskiy，《从历史实例预测未来人工智能的失败》，《预见》，第21卷（1），2019年。

作者简介　罗曼·亚姆波尔斯基（Roman V. Yampolskiy）

亚姆波尔斯基博士是路易斯维尔大学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《超级人工智能：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、年度十大网络大学教授、杰出职业生涯早期教育等诸多荣誉奖项。此外，他还是电气与电子工程师协会（IEEE）和通用人工智能协会（AGI）的高级会员、肯塔基州科学院的成员以及上海交通大学全球传播研究院（GCRI）的副研究员。亚姆波尔斯基博士的主要研究兴趣领域是人工智能安全和网络安全。他发表了100多篇论文，包括多篇期刊文章和书籍，并且还曾应邀在瑞典国家科学院、韩国最高法院、普林斯顿大学等机构举办的100多场活动中发表演讲。

第二部分：
产业界的实践和探索

人工智能伦理实践的挑战和机遇

安纳德 · 劳(Anand S. Rao)

新冠疫情的暴发，令大多数政府、企业和公民猝不及防。新冠疫情对地球上几乎所有人的生活都造成了影响，这种影响涉及方方面面。从直接影响（例如死亡、住院和感染）到间接影响（例如失业、居家办公和心理健康），新冠病毒几乎影响了地球上的每个人。

新冠疫情的暴发是全球所有国家各行各业数字渠道使用率大幅上升的直接原因。在这种浪潮下，高级分析、自动化和AI的使用率日益上升。我们于2020年11月对全球1018名高管进行了一次调查，以了解这次危机对各大企业及企业对AI的态度造成了何种影响。在这份简报中，我们重点强调了这次调查中的两大关键的发现。

1. 虽然出现了经济危机，但AI投资仍呈现出上升趋势。

在我们的调查中，44%的调查对象表示新冠疫情对其业务造成了负面影响，但令人惊讶的是，也有人表示新冠疫情对其业务起到了积极的促进作用，其数量与前者相当。有趣的是，企业规模越大（收入超过100亿美元），新冠疫情就越有可能对其产生显著的积极影响。此外，有近四成的大型企业在疫情暴发前对AI研发投入也较多，而且正从AI应用的试验阶段转向实际使用阶段。这些企业对AI的投入在疫情期间得到了回报，其增加AI使用率的比例（38%）、探索全新AI应用案例的比例（39%）和培训更多员工使用AI的比例（35%）远超其他企业。这不但适用于大型企业，也同样适用于在疫情暴发前就已大力投资AI的小型企业。此外，全球AI风险投资资金在2020年第三季度上升至719亿美元，打破了2018年第四季度690亿美元的纪录。

2. 通过使用负责任的AI来管理和消减AI风险正变得至关重要。

新冠疫情的蔓延使AI在人脸识别、接触者追踪、员工行踪监测等应用中的使用率日益增长。对企业而言，识别、消减和管理AI风险（包括偏见、隐私、透明度、可问责性、可解释性、稳定性、安全性和保障性）是他们通过国际研究协会部署AI模型和系统来解决和管理问题时需要面临的重大挑战之一。只有12%的企业实现了完全嵌入式和自动化的AI风险管控。还有37%的调查对象拥有应对AI风险的策略和政策，但并没有自动化的解决方案。在完全嵌入式AI领域，这一数字分别上升至29%（实现嵌入式和自动化AI风险管理的部门）和38%（拥有相关策略和政策的部门）。在所有AI风险中，重点关注算法偏见的调查对象的比例接近36%。对着手扩大AI应用规模的企业而言，其他主要AI风险还包括可靠性、稳定性、保障性和数据隐私。

综上所述，AI应用的增长、AI投资的增加，以及AI带来的风险让我们有机会采用负责任的AI实践来管理和消减上述风险。

作者简介 安纳德 · 劳(Anand S. Rao)

安纳德是普华永道咨询公司的合伙人之一，拥有超过32年的行业经验和研究经验，负责领导普华永道的全球人工智能工作，是新兴技术小组的创新负责人。安纳德负责带领一个与高层管理人员合作的专业团队，为他们提供有关全球增长战略、营销、销售、分销和数字战略、行为经济学和客户体验、风险管理，以及统计和计算分析等一系列主题的咨询建议。他曾在亚洲、澳大利亚、欧洲和美洲工作和生活，从业经验主要涵盖金融服务、保险、电信和医疗健康领域。

安纳德负责与专注于新型创新大数据和分析技术的学术机构与初创企业建立研究和商业关系。凭借其人工智能博士学位和人工智能研究经验，以及后续获得的管理咨询经验，他独到地结合了商业领域的经验与统计和计算分析专业知识，对“数据科学”理论和实践拥有独特的见解。

在开始从事战略咨询服务前，他曾引导并有效使用了针对空战建模、空中交通管制、客户服务决策支持和电信网络管理的创新人工智能方法。此外，他还在墨尔本大学负责教授和指导博士生，并担任智能决策系统中心的项目主任。

安纳德与他人合编过四本著作，并在各类会议和期刊上发表过五十多篇同行评议论文。他经常在关于分析学、AI、行为经济学和创新等主题的技术和商业论坛上发表演讲，也曾在国际期刊、会议和研讨会的组织计划委员会任职。

实践模式：人工智能治理的成功基础

阿布辛克·库布塔（Abhishek Gupta）

人工智能治理在2020年无疑取得了不错的发展势头，有很多声音呼吁要采取行动，利用法律和技术领域的专业知识，提出人工智能系统开发和部署的治理框架。这些举措有很多共同点，大都侧重于透明度、可问责性、偏见、隐私、非歧视等领域，以及人工智能伦理领域100多套原则中受到普遍认同的其他价值观，其中大多数举措中至少有部分内容侧重于人工智能治理。

2019年，人们已经开始谈论人工智能治理，但当时更多涉及的是抽象的想法。而到了2020年，人们看到了更多的推动力，想将这些想法付诸实践，这种转变十分明显。然而，虽然我们看到了很大的转变，但是仍有一些短板阻碍了这些治理机制的部署。特别是在2020年，我们看到这类系统在尚未完备的情况下被草率推出，用于追踪口罩合规性[1]、给学生评分[2]、发放失业救济金[3]等。那么，我们能在此基础上做出哪些改进呢？

正如我在题为《绿色照明ML：部署中的机器学习系统的机密性、完整性和可用性》[4]的论文中所详述的那样—我与我的合著者埃里克·加林金（Erick Galinkin）在2020年的几次会议（包括ICML）上发表过这篇论文——我们发现，人们几乎没有关注这些想法的实际表现。具体而言，就是缺少对一线设计师和开发人员的需求和实践模式的关注，而他们在实施这些想法时至少要承担部分责任。这并不是说政府对组织的授权和管理不会在人工智能系统的治理中起到重要作用。事实上，我们必须考虑采取一些补充措施，特别是这些措施将有助于增强任何其他应用于人工智能治理的组织规模机制的效用。

从从业者的角度来看，他们在遇到抽象的原则、业务压力以及按时交付高质量产品和服务的限定期限时会面临许多挑战。人工智能治理机制的实际运作正是在这些方面出现了问题，需要修复。根据我的经验，有助于缓解此问题的首选方法是努力将各条治理要求融入设计师和开发人员的现有工作流程中，而不是跳到创建新机制这一步骤。这样做的好处是能降低接受这些新需求时产生的摩擦，并加快部署速度，然后收集证据来判断它们是否有效。有了这些证据，人们就可以在更广泛的层面上为它们的合并提供更有力的依据。

第二是要将从业者纳入这些机制的开发过程中，这可能也是创建人工智能治理解决方案时最重要的一点。具体而言，主要有两方面要求：第一，在实践人工智能治理解决方案时，需要能够根据从业者的经验指出人工智能治理解决方案在哪些方面可能无法发挥作用；第二，还要与上述从业者建立信任，确保他们不但能对相关要求心中有数，还会积极贡献，以强烈的主人翁意识，致力于促进解决方案取得成功。上述一些见解在我所著的《可操作的人工智能伦理》（*https://atg-abhishek.github.io/actionable-ai-ethics*）一书中也有所体现，该书以非常实用和实际的方式讨论了如何将人工智能伦理学付诸实践，试图解决我在此所强调的一些挑战。

因此，如果我们想要切实推进人工智能治理并使其发挥作用，那就不是再花几个月或几年的宝贵时间来讨论抽象的想法，而是要牢记这些实践模式。我们现在就应该采取行动，首先便是关注这些系统在实践中的设计和开发路径。

参考文献

[1]https://www.theverge.com/2020/5/7/21250357/-france-masks-public-transport-mandatory-ai-surveillance-camera-software

[2]https://www.wired.co.uk/article/gcse-results-alevels-algorithm-explained

[3]https://www.usnews.com/news/best-states/articles/2020-02-14/ai-algorithms-intended-to-detect-welfare-fraud-often-punish-the-poor-instead

[4]A Gupta & E Galinkin，《绿色照明ML：部署中的机器学习系统的机密性、完整性和可用性》，预印本网站arXiv：2007.04693，2020年。

作者简介　阿布辛克·库布塔（Abhishek Gupta）

阿布辛克·库布塔是蒙特利尔AI伦理研究所（MAIEI）的创始人，也是微软的机器学习工程师，在微软的CSE负责任AI董事会任职。他的著作《可操作的人工智能伦理》（*Actionable AI Ethics*）（*https://atg-abhishek.github.io/actionable-ai-ethics*）即将出版，该作将是一本实用且实际的人工智能伦理操作指南。

阿布辛克是合作参与美国国务院国际访问者领导项目的人工智能伦理访问研究员、美国西北高校委员会数据咨询委员会的负责AI主管、道森学院人工智能咨询委员会成员，Linux基金会下属LF AI基金会的准会员，以及歌德大学法兰克福大数据实验室的教员。其研究侧重于应用技术和政策方法，在不同领域使用人工智能来解决伦理、安全和包容性问题。他建立了世界上最大的社区驱动型人工智能伦理公共咨询小组，为负责任人工智能领域中许多倡议的制定做出了重大贡献。

基于经验的负责任发布：GPT-3的安全部署

艾琳 · 索拉曼（Irene Solaiman）

2020年，世界上前所未有的事件与有先例的人工智能挑战接踵而来。人工智能研究界正在努力寻找最佳实践，以实现负责任的发布和安全部署，但现在主要是在居家工作环境中进行远程协调。同时，现有的担忧问题（如造谣）显示了当今疫情响应和政治制度等方面的后果。

人工智能系统，特别是生成模型，已经变得越来越强大，因而需要安全保障。生成模型根据文本或图像等数据进行训练，并试图生成与该数据相似的输出结果。对于语言模型来说，这可能意味着可以被预测。我们对风险的担忧在工业界、学术界和公众中都存在。值得注意的是，蒂姆尼特 · 格布鲁（Timnit Gebru）博士在与人合著的一篇论文中提出了对强大语言模型的关注，强调了诸如嵌入式偏见等担忧。OpenAI和其他研究人员已经证明了有害的偏见、潜在的误用和造谣的可能性，以及检测合成文本的难度。

认识到这些风险后，OpenAI在2019年实施了一项不同寻常的负责任的发布战略。我们以分阶段的方式发布了我们的语言模型GPT-2，采用分段式发布是为了便于在每次发布之前研究模型特性，确保每一个新版本都比上一个旧版本更加强大。我们在2020年发布的规模更大、性能更高的语言模型GPT-3还有待进一步验证。和GPT-2一样，GPT-3同样具有灵活的功能，从文本摘要和翻译，到问答甚至三位数的算法，均在其能力范围之内。GPT-3还具有很强的“小样本学习”能力，即根据少量给定的范例解决问题的能力。GPT-3被误用的可能性比GPT-2更高，并且仍然表现出歧视性偏见，因此必须谨慎使用。

为了安全起见，我们选择通过API发布GPT-3。OpenAI托管了该系统，经过批准的用户可以使用应用程序编程接口（Application Programming Interface, API）密钥进行访问。我们为用户提供了一个用于进行实验、开发新应用程序或进行研究的界面。公司决定安全地将该系统产品化，从而为我们的研究提供资金。本次发布是该决策的一环，但包括一个针对研究人员的学术访问计划，旨在协助识别模型特征，特别是在偏见、误用和检测这几个关键领域。

应用程序编程接口是一种确保安全性和可访问性的手段。我们在没有向公众发布完整系统的情况下部署了GPT-3，这使得我们能够轻松对误用做出响应，并在了解情况后对系统做出相应改进。我们概述了使用指南、审查了所有用例，并终止了任何造成伤害或缺乏足够保障的情况。我们托管了运行成本高昂的系统，减轻了成本压力，同时也使得这类系统能为小型企业和组织所用。

系统灵活的功能决定了我们无法预测所有的潜在用例，因此我们对用户做出了限制，并逐渐扩大访问范围。蒙特雷国际研究院、华盛顿大学和艾伦人工智能研究所的研究人员分别帮助我们研究谣言和偏见风险。我们的内部研究有助于建立我们的使用指南和模型改进。

最近在发布规范方面的基础性工作证明了随着系统的增强而不断进行投入的必要性。人工智能研究人员，包括人工智能伙伴关系和谷歌的研究人员，为更广泛的影响分析和模型文档等实践提供信息。系统增强带来的影响取决于环境，减轻影响不仅需要技术，还需要社会科学和社会技术研究。这项研究的继续开展需要与不同的团体进行合作，并适应人工智能的快速发展和前所未有的事件带来的意外挑战。

作者简介 艾琳 · 索拉曼（Irene Solaiman）

艾琳 · 索拉曼是OpenAI的政策研究员，负责引导政策制定者参与相关事务，并进行社会影响和公平性分析。在加入OpenAI之前，她是哈佛大学伯克曼 · 克莱因中心（BKC）的研究员，作为大会学生奖学金的成员研究人工智能的伦理和治理问题。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国际关系学士学位，她在马里兰大学被选为美国大学优等生协会会员。

人工智能治理需遵循可持续发展的理念

杨帆

随着人工智能（AI）技术的不断发展，AI的应用已不仅仅局限于科技领域，AI逐步赋能各行业，对全球经济发展起到了举足轻重的作用。AI在全行业、多领域的不断深入也使得各界对AI人才的需求从“技术优先”转向“人工智能+X”的跨学科复合型人才。同时，关于未来AI可能导致的风险讨论也愈来愈多。因此，我们已经不能用单一的技术视角去看待AI，而应该同时看重AI在赋能时带来的社会价值。

当我们从社会属性的角度出发谈论AI，可以看到“可持续发展”是建设AI生态的必由之路。AI可持续发展的理念来源于联合国17个变革世界的目标，换言之，实现AI的可持续发展能够实现经济目标之外的社会价值，有助于实现共筑人类命运共同体的愿景。2020年6月，联合国发布“数字合作路线图”，阐明了全球互联互通、数字包容等八个关键行动领域，这一路线图更加深刻地解释了AI和大数据时代背景下可持续发展的价值和内涵。

2020年6月22日，商汤科技（以下简称“商汤”）智能产业研究院联合上海交通大学清源研究院，共同举办“AI可持续发展论坛2030”，并正式发布《AI可持续发展白皮书》，这是中国AI产业界首次讨论可持续发展的理念和规划。《白皮书》提出以人为本、共享惠民、融合发展和科研创新的价值观，以及协商包容的AI伦理原则、普惠利他的 AI 惠民原则、负责自律的 AI 产融原则、开放共享的 AI 可信原则，为解决AI治理问题提出新观念和新思路。

在推进可持续发展框架的过程中，商汤以实际行动践行了可持续发展的各项原则。2019年7月，商汤设立AI治理与伦理委员会，并规划设计了AI风险评估与管理框架，各项目落地之前必须通过委员会审核。在细分产业方面，商汤亦做出了许多努力。例如2020年9月23日，商汤成立“商汤教育”子品牌，通过教育平台等一系列产品以及教师培训项目降低AI数字鸿沟，为AI领域输送更多创新型人才，体现了联合国关于优质教育的可持续发展目标；在医疗领域，商汤的SenseCare® AI辅助诊疗系统在提升医疗工作者工作效率的同时有效地帮助遏制新冠疫情扩散，为落实联合国关于良好健康与福祉的可持续发展目标作出了贡献。诸如此类的案例不胜枚举，是商汤目前在实践AI赋能产业时的重要思路和实践。

作为计算机视觉领域的领头羊，商汤将在AI可持续发展和治理领域中持续发挥作用，深刻理解全球科技创新的新趋势，开拓商汤在伦理与治理领域发展的新路径，进一步推动商汤“负责任的AI”的企业理念并在实践中落实可持续发展的各项原则，将企业的发展与社会责任深度联系起来。

作者简介

杨帆

杨帆是商汤科技副总裁、联合创始人，商汤科技AI伦理与治理委员会主席。目前负责商汤的生态体系与产业协同的规划建设。加入商汤前，杨帆曾长期就职于微软亚洲研究院，期间参与研发的技术均被广泛应用于微软全球主要产品中，积累了丰富的尖端技术产品化的路径规划经验和实践经验。

自商汤成立以来，杨帆带领团队在手机、零售、智慧城市等行业成功实践“AI+产业”的创新业务模式，并赋能商汤工程技术团队沉淀从组建基础架构到技术产业化的整体方法论。行业实践中，他主导推动AI在金融科技领域及互联网娱乐领域的产品化过程，决策推动AI在智慧城市行业裂变增长的整体布局，首提并落地城市级视频分析中心概念，成功打造行业第一品牌。商务推进中，杨帆带领团队先后与中国移动、华为、小米、本田等战略客户达成重大合作，并决策主导集团成功申请智能视觉国家新一代人工智能开放创新平台项目。

21世纪的社会人工智能契约

鲁子龙(Danil Kerimi)

可以肯定地说，2020年与我们的预期截然不同。计划被打乱、行程被推迟、会议完全在线上进行。国际和国内，城市和农村之间的贫富差距进一步扩大，金融断层线的清晰程度更胜以往。作为第一个真正意义上的全球性的数字疫情，目前的危机是对我们的政策、制度和战略的完美压力测试。它迫使我们抛弃旧有的假设，提出了种种难题，在各方面摆脱此前想当然的思维习惯。

然而，生产力的增长幅度在发生大动荡之后往往是最大的。在关注危机带来的短期混乱及其中期后果的同时，我们也不能忽视人类面临的长期挑战。

城市化、全球化和数字化是20世纪最后几十年和21世纪初影响社会的三大趋势，这种转变在亚洲最为明显。

2020年及其带来的第一次全球数字疫情标志着这一演变的新里程碑。这场危机不仅迫使我们重新思考世界上许多城市的宜居性，而且也标志着新一轮的全球化浪潮。这是一种质变，但很快就会波及更大范围。

2020年也标志着技术后浪开始正式席卷亚洲许多地区。全球许多政府和公民都在密切关注算法带来的权力动态变化。随着越来越多基于人工智能的设备和服务走进数百万人的生活，针对不稳定人工智能政策进行的辩论不再局限于大学和智囊团，而可能发生在部长级走廊、公司董事会会议室、街头市场乃至私人起居室等各种环境。

人们习惯于区分农业革命和工业革命及其对我们的生产和消费系统的影响，而这反过来又影响我们的生活环境、经济和政治制度。在过去的几个世纪里，我们至少目睹了三次农业革命和四次工业革命。在人口增长、全球变暖、超级大国竞争和消费者意识不断增强的背景下，随着物理、网络和生物世界的交叉，人工智能领域的突破和由人工智能推动的突破相继出现，我们将看到一场新的农业－水产－工业革命。2021年，社会契约将在我们眼前被改写。这一次，它将由人工智能推动，同时也将推动人工智能的发展。

作者简介　鲁子龙（Danil Kerimi）

鲁子龙是一名经验丰富的技术高管，擅长在复杂的多司法管辖区范围内细致入微地创建和表达市场敏感的议程和信息。他是一位鼓舞人心的领导者，在数字化转型过程中建立了高效的团队。鲁子龙在发达市场、新兴市场和前沿市场积累了丰富的经验，并据此在战略、业务发展、公共和投资者关系、ESG、技术和经济外交领域帮助公司打造具有影响力的项目。

作为投资者、内部、外部、政府关系方面的专家，鲁子龙正在与国家、地区和市政府、大型企业、初创企业和投资者团体合作，以加速和保障世界各地的数字化转型，并应用新发现的能力解决健康、食品和环境生态系统方面的难题。鲁子龙最热衷于发掘个人和组织的内外部创业潜力，从而推动世界发生积极变化。

我们的数据应为谁所有？

史蒂文·霍夫曼（Steven Hoffman）

依靠人工智能（AI）处理和利用大量数据的能力，世界各地的决策者正寻求方案，以限制人们对个人数据的滥用。越来越多的立法者提议，个人应对其在网上生成的数据享有专属权。该提议理论上听起来不错，但它是否解决了不对称市场的根本问题，即大公司从这些数据中获利的方式是否可能损害个人和社会？

为了纠正这一问题，许多决策者认为，社交媒体和其他公司应该向用户支付数据许可费。这样，生成数据的人就可以获得补偿并控制数据的使用方式。这听起来是个不错的主意，但实际上，大多数人最终会放弃他们的隐私权利，以换取很少的金钱补偿。事实上，绝大多数人不会阅读这些许可条款的细则，即使阅读了，也难以理解相关的影响后果。这意味着，个人数据受控的想法，不过只是幻想。

在信息流中引发摩擦也是一个问题。世界经济的繁荣得益于数据的快速交换。如果我们用大量复杂的规章制度使数据系统陷入困境，它将严重影响所有依赖这些数据的企业和消费者。从物流、医疗保健、广告到电子商务、媒体和公共服务，各个领域都将受到影响。人们对个人信息的需求之所以如此之高，是因为它对社会极为重要，且用途广泛。

请考虑一个人的姓名、地址、生日和其他基本数据能有多少种用途。许可和限制这些信息意味着什么？人们是否需要与他们使用的每个应用程序签署许可协议？这些数据真正意义上归谁所有？事实上，数据就是信息，把它当作商品来对待，从根本上改变了它的性质，这将影响我们整个社会体系。在我们采取这一激进措施之前，我们需要评估这类立法的影响。目前，我们理所当然地认为大多数公共和私人服务都依赖于这些数据的自由流动。

相反，我们应该把重点放在核心问题上，这不是人们是否能获得一小笔许可费的问题，而是如何规范这些数据的共享和使用的问题。大多数人都关心自己的隐私，他们希望确保自己的数据不会落入坏人手中，以免对自己的生活或社会产生负面影响。

我们需要认识到，问题不在于人们无法从他们的数据中获得报酬。人们可以免费注册Facebook，正是因为人们同意让该公司使用他们的数据来盈利。向人们支付额外的许可费来授予Facebook相关数据信息的使用权并不会改变这一点。当公司没有对数据的处理和管理进行足够的控制时，问题就会出现。

决策者应该把重点放在遏制滥用数据上，而不是将数据商品化。大多数国家已经对如何使用和共享人们的医疗数据有了严格的规定。对于公司如何使用和共享个人数据，我们也需要类似的规定。最终，好的立法需要在保护个人隐私和实现信息自由流动之间找到平衡。只有这样，我们才能在既不妨碍经济增长，又不限制我们所依赖的服务的前提下，同时保证个人与社会利益。

作者简介　史蒂文·霍夫曼（Steven Hoffman）

史蒂文·霍夫曼，在硅谷被称为霍夫曼船长（Captain Hoff），是全球领先的孵化器和加速器公司Founders Space（FoundersSpace.com）的董事长兼首席执行官。同时，他还是天使投资人、August Capital的有限合伙人和连续创业者。史蒂文·霍夫曼著有多本备受赞誉的著作，其中包括《让大象飞》（*Make Elephants Fly*）、《穿越寒冬》（*Surviving A Startup*）以及《五种力量》（*The Five Forces*）。

后疫情时代数字医疗人工智能的有效治理需要多方利益相关者协同合作

奥马·科斯提拉·雷耶斯（Omar Costilla-Reyes）

过去几年，美国和全球数字健康领域的活动增长惊人。该领域的初创企业、研究和投资空前活跃。持续不断的全球疫情和寻求创新型远程医疗保健方案的需求促进了数字医疗应用的迅速普及。

负责批准数字医疗技术临床应用的监管机构是美国食品和药物管理局（FDA）。“数字疗法”是首批通过FDA针对数字医疗软件的“De Novo”认证模式的产品之一。依据数字疗法联盟的定义，数字疗法是“为患者提供的循证治疗干预，这些干预措施由高质量的软件程序驱动，旨在预防、管理或治疗广泛的生理、心理疾病和行为异常”。数字疗法联盟是一家由数字疗法领域的行业领先企业和利益相关者组成的非营利贸易协会。目前已有不少医疗软件获得了FDA的批准，包括药物成瘾和失眠治疗软件，以及首款通过视频游戏治疗儿童注意力缺陷障碍的软件解决方案。

虽然临床数字干预在缓解重大医疗保健问题方面前景光明，但如何提高其接受度、将其融入医疗体系以及推行相关政策改革一直以来都是难题，且目前才刚刚起步，任重道远。部分数字疗法模式遵循了类药报销模式，该模式具有数字化特性，因而尚未被政策制定者和保险公司广泛接受。

对于数字疗法领域的新兴组织机构而言，数字疗法的监管路径也很难遵循。想要监管部门批准临床研究，就必须研发出循证产品，但这类产品必须通过大型临床试验进行检测，成本高昂且费时费力。

目前，数字疗法领域的解决方案主要专注于对循证医疗保健解决方案进行数字化处理，以此作为干预措施，例如对针对心理健康问题的数字化认知行为治疗项目进行数字化处理。

新一代数字医疗产品的目标在于整合能基于从移动电话和可穿戴数据传输设备获得的纵向数据，并进行持续学习和改变，不断提高解决方案的效用和个性化程度的人工智能。针对该类人工智能的研究还处于起步阶段，需要科研界、政府和行业的协同合作。

数字医疗领域的研究同样取得了不错的进展。例如，在心理健康领域，有两项主要的数字健康举措。其一，美国加州大学洛杉矶分校与苹果公司合作，发起了一项抑郁症大挑战，研究如何通过移动设备客观检测抑郁症状。其二，欧盟发起了中枢神经系统雷达监测倡议，旨在探索可穿戴设备和移动设备在预防和治疗抑郁症、多发性硬化症和癫痫症方面的潜在效用。该倡议同样以合作形式开展，合作对象为杨森生物技术公司。

SaMD是一种以人工智能为核心组件的医疗器械，能根据患者的需求不断学习并做出调整。FDA目前正在研究如何为其软件进行立法，并在采用一种预认证模式，即不断对公司和产品进行评估，以获得监管部门的批准，从而提供SaMD AI解决方案。评估内容涵盖患者安全、产品质量、临床责任、网络安全以及积极的文化氛围。

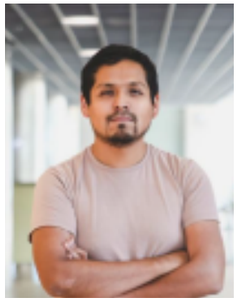
全球各国都应参照美国的SaMD认证模型，依据自身的文化、社会和经济背景，制定相应的立法认证体系。

政策制定者当前面临一项重要挑战，即如何在给予数字医疗领域的行为主体创新自由的同时，保证患者安全和产品功效。后疫情时代正是运用数字疗法提供有效数字医疗解决方案的最佳时机。当下，医疗保健解决方案正以前所未有的速度得到数字化应用。

推动全球数字医疗的彻底改革，需要心理学和计算机科学等多个学科领域协同配合，并同时与政策制定者密切合作，以建立安全有效的解决方案，显著改善全球数百万人的生活。

作者简介

奥马·科斯提拉·雷耶斯（Omar Costilla-Reyes）



雷耶斯博士是麻省理工学院医学工程与科学研究所（IMES）的研究员，同时还是麻省理工学院博士后协会现任副主席。他的研究领域涵盖医疗保健、人工智能、教育和拉丁美洲。在目前着手进行的研究中，他专注于设计经过数字验证的新一代人工智能医疗保健解决方案。奥马拥有英国曼彻斯特大学博士学位，他的博士论文主要研究人工智能在医疗保健和生物测定领域的应用。他是人工智能拉丁美洲峰会的创始人和主席，该峰会的目标是激励拉美地区政府、工业界和学术界的领导人投资和开发人工智能，以造福社会。

第三部分：
政策研究的进展

新兴人工智能治理制度

艾伦·达福（Allan Dafoe）
亚历克西斯·卡利尔（Alexis Carlier）

许多AI治理工作都会涉及为迎接一个立宪时刻做准备，那就是一个建立持久有效的决策机制的机会。这并非易事，需要取得精妙的平衡。这种制度必须对我们目前的行动加以限制，但同时还需留有空间，以随形势变化和我们智慧的增加而进行调整。尽管有着上述困难，但我们仍必须对此做出决定，况且在2020年诞生了多个值得关注的制度。研究界现在也迎来了通告这些决定的机会。

作为一项国际人工智能顶级会议，神经信息处理系统大会（NeurIPS）要求所有论文提交需包含关于“其工作潜在的广泛影响，包括其道德方面和未来社会后果”的声明。这项振奋人心的制度创新可以增加机器学习社区在技术治理方面的参与度和专业知识积累。在AI治理中心（GovAI），我们编写了一份关于如何撰写NeurIPS影响说明的（非官方）指南。在即将发表于最新一期《自然·机器智能》（*Natural Machine Intelligence*）的论文《通过更广泛的影响要求实现AI伦理的制度化》*(Institutionalizing Ethics in AI through Broader Impact Requirements)*中，我们针对能够增加这项创新的成功机会的措施提供了建议。隶属GovAI的卡罗琳·阿瑟斯特（Carolyn Ashurst）举办了一场NeurIPS研讨会，研究潜在的有害影响对研究界造成的必然影响；GovAI主任艾伦·达福（Allan Dafoe）（本文的合著者之一）在自然语言处理国际顶会EMNLP的全体会议上就这一要求的挑战和机遇进行了讨论。

Facebook的独立监督委员会也开始了运作。作为Facebook内容审核的“最高法院”，委员会的决定是有约束力的。这是一项值得称赞的举措，是一家技术公司为改善技术治理（自愿）受外部约束条件控制的罕见个例。研究人员可以通过增进自身的专业知识对委员会进行评估并提供建议，从而为这项举措提供支持，而更广泛的群体（公众、媒体、非营利组织、Facebook员工）则可以帮助监督委员会和公司的相关责任的落实。

合作组织Partnership on AI创立了“AI与共享繁荣倡议”，鼓励私有AI机构致力于构建一个包容性经济未来。GovAI为推动该倡议撰写了报告《意外收获条款：AI收益分配》（*The Windfall Clause: Distributing the Benefits of AI*），提出了减少AI引起的不平等问题的政策工具。该报告的第一作者兼前GovAI研究人员卡伦·欧基夫（Cullen O’Keefe）（现为OpenAI的研究科学家）是该倡议研究小组的一员，隶属GovAI的安东·科里内克（Anton Korinek）则担任指导委员会。

上述制度均由人设计。实现这一目标是“协作式AI”的宏愿之一，协作式AI是最近由艾伦·达福（Allan Dafoe）、托尔·格雷佩尔（Thore Graepel）与DeepMind等机构的其他同仁在其论文《协作式AI尚待解决的问题》（*Open Problems in Cooperative AI*）和相关的《自然》评论NeurIPS研讨会中提出的一个新领域。我们看到该领域在当下可以明确专注于合作，并且由于制度是合作的中心所在，因此制度化设计是具有前景的协作式AI研究方式之一。人类可以确定制度的目标，而AI系统则可用作设计工具和基础框架，为新创新机制的建立提供支持。

我们很高兴能与AI治理界的各位同仁携手合作，共同应对这些挑战。

作者简介　艾伦·达福（Allan Dafoe），亚历克西斯·卡利尔（Alexis Carlier）


艾伦·达福
（Allan Dafoe）

艾伦是牛津大学人类未来研究所人工智能国际政治学副教授与人工智能治理中心主任。他主要研究大国战争的起因，以及围绕变革性技术，特别是关于人工智能风险的全球政治。为帮助科学家更好地研究这些课题，他还对因果推理和提高透明度的方法进行了一定程度的探究。


亚历克西斯·卡利尔
（Alexis Carlier）

亚历克西斯目前在牛津大学人类未来研究所人工智能治理中心担任AI治理项目经理。此前，他曾担任人工智能安全局势联盟的治理助理，并协作参与约翰·霍普金斯大学应用物理实验室和安全与新兴技术研究中心的AI预测项目。在这之前，他还曾在Sensyne Health公司担任机器学习工程师。

人工智能系统的风险管理该如何执行？

贾里德 · 布朗（Jared Brown）

经合组织的人工智能原则正确地指出了人工智能系统的潜在风险“应该被不断地进行评估和管理”。但是政府应该如何做到这一点呢？

为了解人工智能治理方面所面临挑战的程度，可以参考各国政府几十年来为了管理汽车操驾驶险而制定的一些方法，包括关于允许驾驶汽车的人员和条件的规定（如有关驾照、酒驾和醉驾酒精含量判断标准，以及保险要求的规定）；汽车的工程设计和安全功能（如安全气囊、前照灯）；或对汽车驾驶规则（如交通罚单）的持续推行。然而，尽管这项技术实现全球应用的历史早已超过百年，但各国政府尚不能完全控制汽车驾驶的负面后果，因为汽车事故每年仍会造成100多万人死亡。

汽车事故可能造成的后果相对明确，但人工智能系统会产生一系列更难识别和衡量的负面后果。如欧洲联盟《人工智能白皮书》所述，这些后果“既有物质性的（个人的安全和健康，包括生命损失、财产损失），也有非物质性的（隐私权的丧失、对言论自由的限制、人类尊严、就业歧视等歧视问题）”。此外，人工智能系统的风险管理将会特别困难，因为存在一种固有的“风险，即人工智能对其既定目标的追求可能偏离潜在的或原始的人类意图，并导致意外后果，包括对隐私、公民权利、公民自由、保密性、保障性和安全性产生负面影响”（引自美国政府的《人工智能应用监管官方指南》*(Guidance for Regulation of Artificial Intelligence Application)*。正如生命未来研究所（FLI）和我们的合作伙伴从2020年发布的针对政府政策提案的众多正式和非正式回应中所发现的那样（futureoflife.org/policy-work），评估和管理人工智能相关的风险将是一项极其困难和复杂的任务。

尽管大多数政府都投入了大量精力试图去管理风险，但如果没有专家的反馈，来了解能力日益增强且快速发展的人工智能系统的前沿技术，他们将无法妥善处理风险。因此，自2021年起，民间团体组织和人工智能专家将需要积极主动地与政府合作，协助制定与安全带法、限速令和交通法庭等对应的等效政策，以管理人工智能系统的风险。在美国，这将需要与国家标准与技术研究院进行紧密合作，因为国家标准与技术研究院在法律上被赋予了制定“针对可信赖的人工智能系统的自愿风险管理框架”的艰巨任务。在其他地区，例如在欧盟，专家们需要与官员们合作，进一步明确他们关于开展“高风险人工智能应用合规性评估”的提案。在承认存在相当大的风险需要管理，迈出重要一步的前提下，在2021年，我们必须立即着手制定实现这一目标的最佳方案。

作者简介 贾里德 · 布朗（Jared T. Brown）

贾里德是生命未来研究所（FLI）的政府事务高级顾问，也是全球灾难风险研究所的政府事务特别顾问。他致力于研究公共政策和风险管理的交叉领域，此前曾在美国国会研究服务局担任应急管理和国土安全政策分析师，并在美国交通部担任国土安全分析师。贾里德获有乔治城大学公共政策硕士学位和加州大学圣地亚哥分校社会心理学学士学位。

以人为本的人工智能治理

彼德拉·阿维勒（Petra Ahrweiler）
马丁·纽曼（Martin Neumann）

AI治理与马克斯·韦伯（Max Weber）所提出的著名概念“官僚型治理”密切相关。与传统治理或者依靠领袖个人魅力的治理模式相反，这种治理模式通过正式、合法、理性的推荐机制使政治权力合法化。虽然韦伯认可官僚型治理的高效、客观与理性，但他谴责这种社交生活的高度理性化是一种将个体束缚于“寒冰极夜”式系统性控制之下的牢笼。

据分析得知，过去主体与社会之间的脆弱关系是导致主体与社会分裂程度加剧，并最终产生道德沦丧的社会现象的主要原因。而纵观当今的各种危机局面，令人惊讶的是其中的问题与诱因与过往并无差异。许多当代危机皆由主体与相应科技社会之间的紧张关系而生。由于危机现象明显变得更加复杂、相互关联和全球化，因此解决它们需要全球协同，一致采取行动。

乍看之下，算法程序高效、客观、不偏不倚、不受人类情感影响、遵从形式理性的标准，这似乎显而易见。但治理决策是关乎人的，是包含价值观的决策。关于正义、平等、责任、人格尊严以及权利等社会价值主张的差异促使全球范围内存在各种独特的治理制度。当下的AI治理算法并不透明，尚不能反映出制定AI治理相关社会决策所依据的复杂价值观，而这也是应用AI治理会形成明显“牢笼”的原因所在。

与此同时，有充分的研究表明基于算法的治理决策制定会产生歧视与不公，可能有失准确与灵活性，并且会带来严峻的伦理挑战。AI治理正面临关于自反性现代化现象的官僚型理性陷阱，而也正是这种陷阱导致了19世纪欧洲的现代化危机。

因此，如何摆脱官僚型治理的牢笼？答案可能不是“减少AI应用”而是“更好地应用AI”，即以人为本的AI治理。马克斯·韦伯同样可为我们提供一个切入点：为了将文化动态与价值观主张的演变应用于AI治理，可以选择重新研读马克斯·韦伯的《世界宗教的经济伦理》（*Economic Ethics of the World Religions*），并据此将分析对象从宗教延伸至文化，将分析重点从经济拓展至广大社会层面。

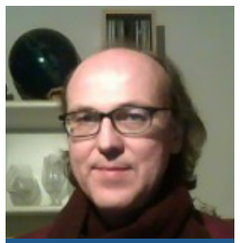
利用机器学习和NLP技术可形成有关全球具体价值观制度的知识库清单，这一包含不断动态变化的价值体系的清单可有助于建立一套可被AI系统识别的指标，用以评估意义建构的可测影响。为了事前评估AI治理，可以进一步共同开发分布式人工智能系统；而为获取人们的价值观，可发展诠释学，以便将文化意义建构和意义架构整合于AI治理中。借此AI治理便可成为负责可靠、有求必应、深思熟虑的“以人为本的AI治理”。

作者简介

彼德拉·阿维勒（Petra Ahrweiler），马丁·纽曼（Martin Neumann）

彼德拉·阿尔瓦勒
（Petra Ahrweiler）

彼德拉·阿维勒教授是德国美因茨约翰内斯·古滕贝格大学技术与创新社会学/社会模拟实验室（TISSS Lab）的负责人，她获有柏林自由大学人工智能学博士学位，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。



马丁·纽曼
（Martin Neumann）

马丁·纽曼博士是彼德拉·阿维勒教授的助理研究员。在取得科学史博士学位后，他开展了一系列重点关注规范、犯罪与冲突的研究，基于智能体的社会模拟，并积累了一定经验。他目前正进行有关人工智能的社会影响的研究项目，重点关注基于AI的治理实践。

从多样性到去殖民性：一个关键转折

马拉维卡·杰亚拉姆（Malavika Jayaram）

虽然我们并不缺乏AI治理框架，但是这些框架的效用和合理性并未受到广泛认可。有人认为层出不穷的原则和指导方针是一种“道德洗礼”，而非对真正变革的承诺；它们进一步推动了行业的自我监管，而不是强制执行具有法律约束力的规则。除了这种声音之外，一系列更加基本的问题也逐渐浮出水面，这些问题有关融入这些框架并在其中具现化的价值观和标准，以及尚未融入并具现化但同等重要的价值观和标准。

为了让有关AI治理的讨论更具包容性，关注诸如神道教或乌班图理念等多种价值观来源的做法日益普遍，甚至成为了一种流行趋势。但是，无论初衷多么好，这种趋势往往都限于表面功夫、经过专门筛选的概念以及趣闻轶事，而没有以更严谨的态度考察这些理念所体现出的权力结构和历史，或AI系统可以内在化和放大的主导叙事和权力模式。这些理念往往由拥有特权的学者和行为主体在相对于他们所借鉴或适合的文化和社区保持封闭的舞台和空间中，或让他们沉迷于追求多样性的舞台和空间中共同挑选和混合。

越来越多的学术研究开始拒绝包容性，因为作者认为，“包容性”是一种不平等的家长式理念——谁在进行融合工作？融入什么之中？融合的条件是什么？它如何巩固和再现既存的权力动态？且包容性这一概念提出了殖民主义理论，让人们透过其理解殖民者与殖民化之间关系的残留，而这种关系正持续影响并产出包括AI在内的数字技术。

殖民的历史建立在对土地的掠夺、对物质资源的剥削和剥夺之上。殖民主义直至今日还在造成影响，继续着侵吞和榨取的过程：这种权力动态和权力结构的留存被称为“殖民性”。随着人们开始挖掘和操纵由数据代表的人类行为和互动以获得物质利益，学者们已经提出了“数据殖民主义”的概念作为这些殖民行为的数字化表现。数据去殖民化工作包括努力维护并不完全符合西方标准的数据的主权和自主权，让当地和边缘化社区管控其信息和案例以及使研究方法和实践非殖民化。

AI大规模嵌入和放大数据殖民主义的程度这一情况对意义的产生，以及现实甚至算法“真相”的建构造成了极其严重的影响。因此，它具有削弱甚至消除其他知识生产和意义建构系统的能力。转向非殖民AI需要的是一次有关抵制和消除使AI系统不公平、不平等和不可持续的主流价值观、假设和偏见，并为其提供替代方案的谈话。随着学者们继续识别和研究殖民性孳生的温床，例如嵌入并延续了有关种族、犯罪和贫穷的殖民历史和刻板印象的预测性警务、福利制度以及身份证，AI治理将从一个恪守狭义、同质的普遍价值观念的领域，扩展为一个能为一套真正多样化的假想和可能性提供支持的领域。这一从多样性走向去殖民性的转变，为以人性化的方式与AI共存或在没有AI的情况下生活提供了更大的机会。

作者简介 马拉维卡·杰亚拉姆（Malavika Jayaram）

马拉维卡·杰亚拉姆是新加坡管理大学法学院的执业助理教授和李光前法学研究员。她还是哈佛大学伯克曼·克莱因互联网与社会中心（BKC）的教员和BKC孵化的独立研究智库数字亚洲研究中心的首届执行董事。她拥有超过15年的技术律师从业经验，曾在伦敦安理国际律师事务所执业，还曾任花旗集团伦敦办事处的副总裁兼技术顾问。

马拉维卡毕业于印度德里国立法律大学，拥有芝加哥西北大学的法学硕士学位。她曾是宾夕法尼亚大学安嫩伯格传播学院的访问学者，还曾在悉尼大学和里约热内卢技术与社会研究所担任研究员。1997年，她负责教授印度首次开设的信息技术与法律课程。

马拉维卡曾是IEEE全球自主和智能系统伦理倡议执行委员会的委员，以及查塔姆研究所（英国皇家国际事务研究所）亚太项目的副研究员。她还是OECD项目“走向数字化：通过数字化转型促进经济增长和福利水平”高级专家顾问组成员。

治理人工智能：从原则到法律

娜塔莉亚·苏姆哈（Nathalie Smuha）

如果说2019年是人工智能治理争辩焦点“从原则转向实践”的一年，那么2020年则标志着这一转变发生了新的转折，即“从原则转向法律”。越来越明显的是，人工智能的开发和使用所带来的一些风险实在过于庞大，仅靠指导方针或私人行为主体善意的自我监管无法解决。这一认识不仅受到媒体对问题重重的人工智能应用的报道和敢于直言的民间社会组织的支持，也受到私人行为主体本身的推动，他们部分开始公开要求制定具有约束力的规则，以提供明确的法律依据，同时增强公民的信任度。因此，政策制定者开始更仔细地审查法律漏洞，这些漏洞会导致某些人工智能应用产生的有害影响无法被有效抵挡。

因此，在人工智能高级专家组工作的基础上，欧盟委员会发表了一份白皮书，描绘了欧盟法律秩序中的一些空白，并为即将在2021年提出的新法规提供了一份蓝图。此外，欧洲委员会人工智能特设委员会（CAHAI）发表了一份可行性研究报告，根据其在人权、民主和法治领域的标准，审查了人工智能发展、设计和应用的法律框架的潜在要素。它概述了适用于人工智能的现有约束性和非约束性文书，列出了它们的优（缺）点，并在此基础上提出了可纳入未来约束性文书（如国际公约）的基本权利和义务。

当然，这些倡议可能需要数年时间才能变为可执行的法律条款。此外，它们在范围上仍然是区域性的，而人工智能带来的许多风险需要通过一种全球性的方法来应对。此外，还需要进一步的跨学科研究来更好地识别、理解和减轻人工智能的潜在不利影响。因此，为人工智能建立一个适当的法律框架以确保能够保护全球公民的工作还远远没有结束。然而，上述政策制定者所走的道路是鼓舞人心的，并有望为其他国家树立榜样。接下来，我想提出值得注意的三点。

首先，人工智能政策和法规应该具有整体性。人工智能系统不是孤立存在的，而是更广阔的社会技术环境的一部分。这个环境包括人工智能系统生命周期中涉及的众多行为主体和流程，以及它们使用的数据和运行所用的基础设施。因此，治理人工智能还需要采取适当的措施来治理数据流和数字基础设施。

其次，必须确保人工智能监管谈判中的参与者足够多样化。通常情况下，最边缘化的个人和社区会首先受到有害人工智能应用的负面影响，进一步加重不平等和问题重重的权力关系。因此，在进行人工智能治理辩论时，不仅要让那些能够得利的主体参与，还应该让那些会有所损失的主体参与，以协助塑造人工智能的法律框架。

最后，政策制定者不仅要考虑使用人工智能系统可能造成的个人危害，还要考虑集体和社会危害，这一点至关重要。从长远的角度来看，人们越来越清楚地看到，人类自治权的下放不仅是在一个领域，而是在我们生活中越来越多的领域里发生，而这种放权累积起来便可能动摇我们道德、民主和社会基础设施的根基。只有承认这一风险，我们才能采取有效措施，妥善应对。

作者简介

娜塔莉亚·苏姆哈（Nathalie Smuha）

娜塔莉亚·苏姆哈是鲁汶大学法学院和鲁汶大学人工智能研究院的研究员，负责研究人工智能（AI）和其他新技术的法律和伦理问题。她的研究重点是人工智能对人权和社会价值观的影响。娜塔莉亚定期就人工智能相关政策问题向各国政府和国际组织提供建议。她是欧洲委员会人工智能特别委员会（CAHAI）的独立专家，也是经合组织人工智能专家网络（ONE AI）的成员。她还曾在欧盟委员会通信网络、内容和技术总局（DG Connect）任职，负责协调人工智能高级专家组的工作，并协助制定欧盟的人工智能政策。娜塔莉亚是一名合格的纽约律师，曾在一家国际律师事务所任职。她拥有鲁汶大学的法律和哲学学位，以及芝加哥大学法学院的法学硕士学位。

新冠疫情与人工智能发展的地缘政治

温德尔 · 瓦拉赫（Wendell Wallach）

除了惨痛的生命损失和筋疲力尽的医疗体系外，2020年的疫情还造成了地方、国家和国际性经济活动大范围停滞。然而，随着工作、教育和娱乐的网络化，数字经济的增长速度显著加快，这包括针对基础设施和人工智能研究的投资。微软等公司的领导们惊讶地发现，原本预计需要3-5年才能达成的目标在下半年就要超额完成了。

在此期间，有关人工智能伦理和治理的工作已从阐明指导人工智能研究和系统设计的原则发展为部署基于人工智能技术的程序和政策。在中国，随着数字信用体系的不断推出，有关人工智能伦理的讨论也越来越广泛，而这些讨论主要集中在隐私问题和个人信息滥用方面。在欧洲和美国，公民自由倡导者带头努力限制人脸识别技术的使用。这些举措在很大程度上是成功的，因为人们越来越意识到，人脸识别系统中使用的算法很可能会产生不准确结果，因此不应该用于执法和其他关键活动。

2020年，特朗普政府倡导的将数字经济划分为两个影响领域的举措进展缓慢，收效甚微。随着曾经初露端倪的独裁和反民主倾向融合成唐纳德 · 特朗普领导下的一股活跃的政治力量，美国的地缘政治影响力遭到了削弱。尽管美国在政治上存在严重分歧，但美国民众普遍不信任中国，他们曲解了中国的一些政策。实际上，拜登政府有机会重新调整美国的对华政策，但鉴于美国民众对中国的意图仍报以一种不信任的态度，改变目前的方针并非易事。

中国的政治制度被证明能有效解决新冠肺炎造成的公共卫生危机。与此同时，在美国，新冠肺炎疫情的政治化带来了令人触目惊心的死亡人数，并让这个以全球领先的医学研究和医疗保健水平而著称的国家的医疗机构倍感压力。这反过来也暴露了美国民主制度的弱点。

我早就指出了美国人对中国的不了解之处，以及中国人对美国的误解，而且，我也一直强调，我们迫切需要在发展新兴技术和制定气候政策方面展开国际合作，以避免造成严重损害今世后代的发展前景的灾难性后果。虽然在人工智能和其他新兴技术的发展方面表现出合作意愿和一定程度的协调配合的国家有所增加，但并未带来实质性的进展。西方国家越来越多地结成治理联盟，中国和俄罗斯要么被排除在外，要么只有在遵守既定规则的情况下才被邀请加入。许多西方国家正在就与中国共享标准和人工智能政策是否符合本国的最佳利益而展开积极的辩论。基础设施标准的竞争，特别是在5G推出之后，将再次把世界分成两派，并可能带来毁灭性的后果。

拜登执政时代的到来还是提供了重新调整中美关系，并在网络安全、生物安全、人工智能武器和地球工程等重大问题上展开合作的机会。但要做到这一点，所有派别都必须在2021年为实现这一目标而积极采取措施。无论在何种情况下，在短期和长期国家利益与地缘政治稳定之间进行有意识的权衡总是很困难的。然而，面对如此情形，我认为未来的国际安全工作必须确保疫情后的恢复，促进各国对全球变暖做出集体反应，以及减轻数字转型加速所带来的不良社会影响。

作者简介　温德尔 · 瓦拉赫（Wendell Wallach）

温德尔 · 瓦拉赫是耶鲁大学生物伦理学跨学科研究中心的学者，拥有11年的研究工作经验，主要负责技术和伦理学研究。同时，他还是黑斯廷斯中心的高级顾问和卡内基国际事务伦理委员会的高级研究员。他的最新著作《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*）是一本新兴技术入门指南。此外，他还与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由瓦拉赫编辑的八卷《新兴技术伦理论文集》（*Library of Essays on the Ethics of Emerging Technologies*）。瓦拉赫于2014年获得世界技术伦理奖，在2015年获得新闻和媒体奖，并在2015-2016年期间担任渥太华大学富布莱特研究委员会主席。瓦拉赫先生曾被世界经济论坛任命为2016-2018年全球未来技术、价值观和政策委员会联合主席，目前是该委员会下属的人工智能委员会的成员。此外，瓦拉赫还是第一届国际人工智能治理大会（ICGAI）的主办人。

减轻历史遗留的不平等问题：发展中国家在人工智能治理上的参与

方淑霞（Marie-Therese Png）

在一个由新冠肺炎疫情和全球反种族歧视抗议所定义的历史性年份，结构不平等的历史轨迹在人工智能治理中受到关注不足为奇。

人工智能治理倡议认识到，它们有责任确保人工智能的部署和监管不会“锁定”内部和国际不平等。因此，我们构建壁垒和原则时，在确保代表方的多样性方面做出了更大的努力。例如，联合国秘书长2020年数字合作路线图呼吁非洲、南美和中亚更多地参与人工智能治理，以重新平衡北美、欧洲和中国的话语权垄断。

正如Jasanoff和Hurlbut提醒我们的那样，我们必须意识到“谁坐在谈判桌旁、哪些问题和担忧被搁置一旁、哪些权力不平衡现象正在影响着辩论的内容”。减轻危害的策略不能由那些受益于人工智能系统的人来确定，而必须由那些了解并付出过代价的人来确定。

例如，尤金尼奥·加西亚(Eugenio Vargas Garcia)博士指出，尽管致命性自主武器可能首先部署在发展中国家的冲突地区，但其中许多地区完全没有监管方面的话语权。通过beta测试将人工智能系统的危害输出到边缘化人群或中低收入国家是有据可查的。

为了召集利益相关者组成有效联盟以减轻危害，我们首先必须查明阻碍发展中国家的利益相关者在政治方面进行有效参与的结构性障碍。如果国家和民间团体的行为主体不能单方面采取行动来保护自己的利益，或进行情境化治理，那么人工智能治理倡议将实施鲁哈·本杰明（Ruha Benjamin）教授所说的“技术恩税”，即旨在解决不平等问题，但却再现或加深了依赖性和榨取主义的干预措施。此外，相关政策还将以不符合发展中国家的目标和制约因素的方式在各司法管辖区推广。

在DeepMind的《去殖民化人工智能》*(Decolonial AI)*论文中，我和我的合著者沙基尔·莫赫德（Shakir Mohamed）博士、威廉·艾萨克（William Isaac）博士提出，如果不看人工智能不平等现象的历史轨迹，我们就无法理解当前的人工智能不平等现象，也无法预测它们的未来。我们今天看到的先发优势和排他性路径依赖，在某种程度上是从我们的殖民历史中遗留下来的。当我们寻求增加发展中国家的代表性时，我们认识到“发展中国家”描述的是殖民主义遗留下来的一种地理现象。

77国集团和不结盟运动等联盟的存在，是非洲、亚洲、拉丁美洲和其他区域非殖民化和独立运动的关键，证实了殖民主义在当代全球不平等中的持续性。如今，这些联盟占联合国成员国的三分之二、全球人口的55%。它们是发展中国家阐明集体利益和促进南南合作的平台。

2020年，乌利塞斯·梅杰斯（Ulises Mejias）教授提出了一项不结盟技术运动，其主要目标是从加强依赖程度的技术过渡到促进发展中国家自主能力的技术。这种精神是发展中国家进行有效参与，并积极讨论如何减轻风险和防止陷入不平等的先决条件。

作者简介

方淑霞（Marie-Therese Png）

方淑霞是牛津互联网研究所的博士生，主要研究自主决策系统治理中的政治去殖民化举措。她曾是联合国秘书长数字合作高级别小组的技术顾问，致力于数字包容、致命性自主武器、网络安全、算法种族歧视等技术政策领域的研究，特别注重多方利益相关者联盟的建立和地域代表性的倡导。方淑霞曾与谷歌 DeepMind在人工智能价值一致性方面展开合作，共同撰写了同行评议的学术论文《非殖民主义理论：人工智能研究的社会技术展望》（*Decolonial Theory as Socio-technical Foresight in Artificial Intelligence Research*）。此外，她还是IEEE合乎伦理的人工智能伦理标准委员会的成员。

方淑霞的研究从数字人权和人工智能治理中的非西方视角着手，对新加坡智慧城市生态系统中人脸识别进行案例研究，并在麻省理工学院媒体实验室和哈佛人工智能倡议开展技术伦理和系统危害方面的研究。方淑霞曾是NeurIPS Resistance人工智能研讨会、麻省理工学院生物黑客运动生物峰会以及世界政府首脑峰会全球人工智能治理圆桌论坛的协办人。她拥有牛津大学的进化生物学和社会科学本科学历，以及哈佛大学的发展认知和群体间冲突硕士学位。

人工智能需要更多的自然智能

马库斯·克纳夫（Markus Knauff）

人工智能（AI）在其诞生伊始主要是一项计算机科学与心理学领域开展的合作项目，旨在创造出具备类人思维的机器。这一跨学科的合作甚至催生出了新的学科——认知科学，该学科研究所有生物和技术系统处理信息的方式。如今，得益于深度学习在商业上大获成功，这一合作继续向AI方向发展。然而实际上，该领域正逐渐背离心理学的规律。深度学习系统受人类学习启发而来，但时至今日，其学习方式已与人类大相径庭，其结果便是它们会很容易受骗并产生许多错误。即便如此，它们能够通过分析大量人类产生的数据，从而发掘其中潜在的关联模式，就这一点而言是成功的。然而，深度学习的局限性已初现端倪，原因之一便是AI系统完全无法进行推断并得出推论。请看以下推论：

1. 如果下雨，街道便会变湿。
   下雨了。
   所以，街道变湿了。

2. 间谍可能出现在柏林或巴黎，但不会
   同时出现在两地。
   间谍在巴黎。
   所以，间谍不在柏林。

处理上述推论是智能的核心所在，尽管这对于大多数人而言不值一提，但如果我们问Alexa、Siri或其他虚拟AI助手系统这类问题，通常只会得到毫无意义的回答。因此，很难想象如果无法解决如此简单的问题，AI如何能取得进一步发展。以下是认知心理学得出的一些结论，能够有助于未来建立具备更接近人类水平智能的AI系统：

1. 当人思考这个世界时，他们会在脑海里模拟真实、假设或虚构的场景。

2. 人类推理是可以推倒重来的，即根据新的证据撤销之前得出的结论。

3. 对于人类而言，诸如如果、那么、所有、一些、没有等逻辑连接词的含义有别于古典逻辑。

4. 人类经常会通过快捷方式或试探性方法得出有用但在逻辑上不成立的结论。

5. 人类更倾向于自行得出结论，而不仅仅是对相关结论进行评估。

6. 一旦发现某种矛盾，人类往往会抛弃之前信以为真的观念，从而避免认知系统得出过多结论。

7. 如果存在多个结论，人在推理时会倾向于选择其中一个而系统性地忽视其他结论。这便是认知经济学理论引导人类诸多思考、推理以及决策过程的实例之一。这种认知心理学方面的洞察力应植入AI之中，以便开发出可与人类互动并能得出用户所能理解和接受的结论的系统。认知科学家已着手应用能够反映自然智能此类核心原则的计算推理系统。人工智能不应落后于这些发展，而应再次顺应心理学的规律，以使技术系统更加智能。

作者简介　马库斯·克纳夫（Markus Knauff）

马库斯·克纳夫是吉森大学实验心理学与认知科学学院教授兼前院长。此前，他先后就职于德国弗莱堡大学、普林斯顿大学以及马克斯·普朗克生物控制论研究所。马库斯·克纳夫还曾任德国认知科学协会会长、《认知科学》（*Cognitive Science*）副主编，并曾主持于柏林召开的第35届认知科学协会年会。在2011至2018年间，他担任“理性的新闻框架”优先项目计划的主任，其中的15项研究项目受到德国研究基金会资助，范围涵盖心理学、人工智能、哲学和逻辑学等多个领域，旨在了解人类理性的本质。他从方法论的角度结合了心理实验、计算模型和神经影像来了解认知与行为之间的神经关联性。他曾从事多年的AI项目研究，但之后因其发展逐渐违背心理学规律又重回人类认知领域。如今他看到了重新通过心理学研究AI的可能（和需要）。

马库斯·克纳夫近期出版的著作包括《推理空间：人类思维的空间理论》（*Space to Reason: A Spatial Theory of Human Thought*）（2013年）和《理性手册》（*The Handbook of Rationality*）（2021年，与哲学家W. Spohn合著），均由麻省理工学院出版社出版。

第三种路径：人工智能治理的对象和目标何在？

乌尔瓦希·阿内贾（Urvashi Aneja）

世界上许多国家，包括印度在内，都在为人工智能治理构建基于风险的框架。基于风险的框架或许能够促进创新，但并不适用于印度等将人工智能视为解决复杂的发展和治理难题的工具的发展中国家。对发展中国家而言，由于人工智能等新兴技术正在影响国家建设和发展的道路，风险和利弊存在差异。低下的监管能力和机构能力使基于风险的治理方法的适用性受到了进一步的挑战。基于风险的治理方法还会导致“监管盲点”，忽视对弱势群体的不同影响和系统性风险。风险评估并非一项客观的活动，它根植于社会文化价值观和首要任务。同时，基于风险的治理方法还面临方法和认知上的挑战——即使一些人工智能应用的风险较低，其累积效应也可能十分庞大。从促进创新的角度来看，以上担忧可能是次要的，但从发展的角度来看，它们的重要性毋庸置疑。

对于世界各国，包括印度的监管机构而言，如何设立监管干预的门槛是一大难题。基于风险的治理方法在这方面具有优势，但在实践中，必须保证围绕风险识别和风险评估的对话做到开放、包容和透明。透明性和专业性本质上是同一问题的两个方面，为确保这一流程的有效性，必须确保民间团体具备评估人工智能影响的知识和能力。目前，印度在上述能力的储备上相当有限，需要在跨学科研究和公众宣传方面进一步投资。此外，提高公众对司法体系和制度的信任度也尤为重要——印度许多基于数字的治理干预措施缺乏完善的申诉机制，不利于建立信任。

相较于以风险程度划分人工智能产品和服务，从而实现人工智能监管，更有效的方式是把人工智能视为一个研究领域，即思考如何推进更负责的人工智能研究和创新。从基本制度方面考虑如何进行人工智能监管也十分有效，有助于我们关注到更多需要管理的问题，包括人工智能创新道路的政治经济方面、促进人工智能产业增长的无形劳动力，以及人工智能的社会影响。这一视角还有助于建立一套用于稳固和指导人工智能治理的价值体系。

最后，伦理框架可能尚且不足以实现产业自治，但从社会角度而言，我们必须围绕自动化算法决策的伦理问题展开进一步的交流——我们需要做出重要的社会选择，决定在何处以何种方式引入人工智能系统。如今，随着全球的监视和专制力度逐步加大，我们应优先针对自动人脸识别和情绪识别系统的使用，为公共和私人行为主体划定明确的红线。

作者简介 乌尔瓦希·阿内贾（Urvashi Aneja）

乌尔瓦希·阿内贾是Tandem研究院的创始人。Tandem研究院是一所位于印度的独立研究院，致力于从科技、社会和可持续发展三个方面提供政策见解。乌尔瓦希主要研究发展中国家的伦理、政治经济以及对新兴数字技术的治理。此外，她还是查塔姆研究所的院士。

第四部分：
国际组织相关进展

2020年人工智能治理回顾：帮助执法机构负责任地使用人工智能的工具包

伊拉克利 · 贝里泽（Irakli Beridze）

2020年对许多人来说无疑是改变生活的一年。在该年，全球暴发了一个多世纪以来最严重的疫情，导致了世界范围的政治、社会和文化层面的剧变。同时，过去需要数年才能取得的技术进展似乎也在短短几个月的时间里相继出现，且第一批新冠肺炎疫苗以前所未有的速度被研发出来，创造了疫苗研发纪录。事实上，人工智能（AI）技术加快了信使核糖核酸（mRNA）疫苗的研发速度，是上述成就的驱动因素之一。

AI应用的进度在新冠肺炎疫情期间着实有所加快，不但被用于医疗研究，而且还被用来限制人口流动，尽管这引发了一些争议。从接触者追踪应用到能监测旅客体温的人脸识别摄像头，AI在追踪和监察方面的应用引发了人们对隐私权等基本自由和权利的担忧。虽然技术可在遏制病毒传播方面起到重要作用，但是AI的使用必须遵循适度、必要和合法的原则。为了避免可能侵犯到基本权利，以及损害民众对国家机构的信任的潜在隐患，各国政府应当努力推进AI治理，并确保AI技术朝着对社会有利的方向发展。

即使关注到各个行业的特点以及AI可提供的不同技术解决方案，全球性AI治理也无法遏制各学科领域受到的潜在负面影响。在道德性、公平性、可问责性和透明性层面，虽然存在适用于不同行业的通则，但面向不同行业的针对性举措可能更加实用和有效。

联合国区域间犯罪和司法研究所（UNICRI）人工智能与机器人中心致力于支持执法机构和刑事司法体系中的其他关键利益相关者了解AI的风险和效益，并探索如何通过AI助力实现没有暴力和犯罪的未来这一目标。人们公认执法机构对AI的使用通常是一个具有高度敏感性且富有争议的话题，而目前又没有具体的指导意见。在此背景下，推进AI治理为执法机构提供支持的重要性日益突显。从这个意义上来说，联合国区域间犯罪和司法研究所人工智能与机器人中心和国际刑警组织环球创新中心已开始着手合作开发注重操作性的工具包，以帮助执法机构负责任地使用AI，并支持和引导负责任的AI设计、开发和部署。INTERPOL-UNICRI全球AI执法会议汇聚全球专家展开了多次探讨，提出了这一远大目标，并为其实现提供了助力。在2020年11月召开的会议上，来自执法机构、学术界和业内的600多位参会者分享经验，相互学习，为这一工具包的打造添砖加瓦。2020年并非完全是黯淡无光的一年，在各方的努力下，许多领域都取得了进展，包括AI治理这一不断发展的重要领域。尽管前路漫长而艰苦，但我们仍在向着实现针对执法机构的负责任的AI迈进。

作者简介 伊拉克利 · 贝里泽（Irakli Beridze）

伊拉克利 · 贝里泽先生在领导多边谈判，以及与各国政府、联合国机构、国际组织、私营公司和企业、智库、民间团体、基金会、学术界和其他国际合作伙伴制定利益相关者参与计划方面拥有20多年的经验。

贝里泽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。他正从有关AI的战略、行动计划、路线图和政策文件等方面为各国政府提供支持。

2014年，贝里泽发起了联合国首批AI项目之一，并管理至今。此外，他还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AI可以帮助实现联合国可持续发展目标的解决方案。

他是世界经济论坛全球人工智能委员会、联合国数字合作高级别小组，以及欧盟委员会人工智能高级专家小组等多个国际工作小组的成员。他经常发表有关技术发展、指数技术、人工智能和机器人以及国际安全主题的演讲。他在国际期刊和杂志上发表了大量文章，媒体在谈到有关AI的问题时，经常引用他的文章。

贝里泽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。

人工智能治理的全球合作：让我们在2021年做得更好

高丹青（Danit Gal）

随着各国继续尝试通过隔离措施来遏制新冠肺炎疫情，且许多产业的发展陷入停滞，人工智能的研究、开发和应用速度却在继续加快，毫无放缓迹象。这进一步加剧了两大紧迫的人工智能治理挑战：全球人工智能使用和监管日益分散的格局，以及人工智能快速部署和监管滞后之间日益扩大的差距。解决这些挑战的一个合适切入点是在人工智能治理方面开展具有包容性的全球合作，但这说起来容易做起来难。

就许多方面而言，2019年和2020年都是在政府层面开展全球人工智能治理合作的好时机，因为各国际机构都在加紧应对上述挑战。2019年经合组织人工智能原则是此类人工智能治理合作的第一个成功案例，影响了后来的二十国集团的人工智能原则。2020年6月，人工智能全球伙伴关系发起，为志同道合的国家提供了一个人工智能合作平台，迄今已有19个国家加入其中。同月，联合国秘书长宣布将建立一个多利益相关者咨询机构，以便进行全球人工智能合作，联合国成员国都将加入其中。

尽管这些倡议的后续发展仍有待观察，但现有的结构性障碍正阻碍着他们的真正全球化应用。在所有案例中，发展中国家以及许多其他边缘化群体的话语权都严重缺失，有失公平。现有的全球人工智能治理合作倡议是由发达国家（通常是西方国家）建立的，随后再由这些国家邀请其他国家参与。它们针对性地遵从少数已在数据集的使用中具有良好代表性的国家的价值观和兴趣，并用其训练对特定领域有益的算法。

代表人数不足的全球大多数国家没有参与进来，预示着遏制全球分裂的任何雄心壮志都将落空。这一现状将发展中国家中的大多数监管机构和用户都排除在了人工智能的加速应用之外，使它们落后于人工智能的监管者以及人工智能监管的受益者。人工智能同样遇到了新技术经常会遇到的问题：人工智能殖民主义使人工智能监管殖民主义日益严重，进一步侵蚀了最需要人工智能的主体的代理权和主权。相较于人工智能等引进技术，从别处引进的政策和法规则更难适应当地的情况。

2021年，我们必须做得更好。在国际层面，全球人工智能治理合作倡议必须让发展中国家和代表性不足的社区都能公平参与，避免标记化。在地区层面，人工智能治理倡议必须加大在共同建设技术监管能力与协调合作方面的投资。在国家层面，监管机构必须征求公众的意见，以确保多样性、包容性和可行性，然后才能在地区和国际上保持一致。应该采用一种分散的、多层次的合作方式，从而让更多需要人工智能治理的人获得自主权并从中受益，而不是采用一种会增强现有权力结构的集中型全球合作方式。

如果我们能够提供机会让更多的人参与国际、地区和国家内部的人工智能治理会议，我们将会在2021年做得更好。我们确保人工智能妥善为全人类服务的总体能力在很大程度上取决于我们对其开发、部署和使用的管理能力。人工智能治理合作的全球化程度越高，我们的总体能力就越强。

作者简介

高丹青（Danit Gal）

高丹青是剑桥大学利弗休姆未来智能研究中心（CFI）的副研究员，主要研究技术伦理、地缘政治、治理、安全和安保等领域。此前，她曾是联合国的技术顾问，负责领导联合国秘书长数字合作路线图执行中的人工智能相关工作。高丹青曾担任IEEE P7009：自主和半自主系统的故障安全设计标准的主席和副主席，是IEEE自主与智能系统伦理全球倡议执行委员会的成员、施普林格《人工智能与伦理》（*AI and Ethics*）杂志的创刊编辑和编委会成员、北京智源人工智能研究院AI4SDG合作网络成员，以及英国研究与创新基金会（UKRI）的可信赖的自主系统可验证节点顾问委员会的成员。

人工智能在疫情应对中的应用：实现承诺

贺尚安(Seán Ó hÉigeartaigh)

2020年，新冠肺炎疫情的应对工作激发了人工智能研究和治理界的热情，因为这场疫情造成的影响是全球性的。人工智能和数字技术早在许多方面就已经被奉为有效应对疫情的关键。事实上，早期迹象给了我们很大的希望。基于人工智能的疫情识别和跟踪系统Blue Dot曾率先对一月的疫情发出警告。数以千计的论文描述了人工智能在预测和应对过程中的应用，包括疫情建模、药物研发、医院后勤规划、症状分析，以及保障隔离工作有序推进的监测。

然而，一年来，人工智能显然受到了一定程度的限制，未能如期发挥其应有效用。原因何在？我是人工智能全球伙伴关系的人工智能与疫情工作组的成员，我们花了6个月的时间分析相关应用、限制、瓶颈和解决方案。我的个人反思与见解分享如下：

1. 炒作与无聊的现实。我们拥有令人惊叹的技术，但疫情应对的核心仍然是久经考验和值得信任的技术——对公共卫生的投资，包括足够的个人防护装备和医院卫生能力、（人工）接触者追踪、隔离等。人工智能可以为这方面的工作提供支持，但目前其作用仍然有限。事实上，过度关注技术解决方案可能会转移人们对基本问题的注意力和相关资源的投入。

2. 新的挑战。人工智能，特别是机器学习系统能够在我们吸取过往经验以预测未来或在未来采取行动方面发挥很大的作用。新冠肺炎是一种新型疾病，它的症状和对各类人群的影响都是前所未见的，并为社会带来了一系列相对较大的挑战和压力。这在许多情况下为妥善训练和部署人工智能系统加大了难度，特别是在以下方面：

3. 数据。许多潜在的人工智能应用，包括基于深度学习的肺部CT扫描分析，以及对人群健康结果的预测，都需要通过大量、多样和具有代表性的数据集来进行训练，以获得足够强大的性能。实际上，当前数据资源较为稀缺，而且分散于不同国家的研究小组之中。

4. 伦理和治理。研究人工智能技术以及为人工智能提供必要数据的数字系统的研究人员和各国政府一直在努力应对法律、监管和伦理方面的挑战。健康数据尤其需要得到具体保护和遵守伦理注意事项（有充分的理由），跨司法管辖区引导数据获取和负责任的使用可能是一个不透明的过程，尤其是对于较小的群体而言。在英国，一种集中式的数字接触者追踪方法最初受到政府的青睐，这种方法本可以为机器学习分析提供丰富的资源；然而，民间社会团体对隐私和数据治理的合理担忧还是让政府最终决定采用分散式的方法。

新冠肺炎疫情危机仍将持续数年，在此期间人工智能可能会发挥更大的作用。我们迫切需要采取一些可以为此提供支持的人工智能治理措施。我特别强调GPAI报告中关于全球研究和数据共享以及数据治理的建议[1]。保护隐私的机器学习和其他数据使用安全保障方法也能发挥效用。为了避免错误和赢得公众信任，我们必须能够按计划更有效地利用远见和伦理，及时处理伦理和治理问题；重点关注和支持稳健性和可靠性；采用快速发展的监督机构和跨社会协商方法[2,3]。最重要的是，我们必须在这场危机的基础上再接再厉，为未来的疫情做好进一步的准备。在这个关键的挑战中，人工智能仍可能为我们提供巨大的助力。

参考文献

[1]http://thefuturesociety.org/wp-content/uploads/2020/12/Responsible-AI-in-Pandemic-Response.pdf

[2]https://www.nature.com/articles/s42256-020-0195-0

[3]《以合乎伦理的方式使用人工智能来应对新冠肺炎疫情》（Cave等人，即将发表于《英国医学杂志》）

贺尚安(Seán Ó hÉigeartaigh)

贺尚安是利弗休姆未来智能研究中心（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和挑战的跨学科中心。AI: FAR计划专注于与人工智能相关的远见、伦理和治理问题。

贺尚安还是剑桥生存风险研究中心（CSER）的联合主任，该研究中心重点关注新兴的全球风险和长期挑战。

贺尚安的研究涉及人工智能和其他新兴技术的影响、前沿扫描和预见，以及这些技术带来的全球风险。2011-2015年期间，他在牛津大学人类未来研究所领导了关于这些课题的研究项目；2014-2019年期间，他曾担任生存风险研究中心的创始执行主任，并参与创立战略人工智能研究中心和利弗休姆未来智能研究中心。他的论文《人工智能竞赛：益处与风险》（*An AI Race: Rhetoric and Risks*）（与Stephen Cave合著）在首届人工智能伦理与社会会议上获得了最佳合著论文奖。此外，他还获有都柏林圣三一大学的基因组进化博士学位。

从原则到行动：造福人类的人工智能治理和应用

塞勒斯·霍德斯（Cyrus Hodes）

2020年，人工智能的全球治理，或者至少是各论坛上关于人工智能伦理的讨论，已成为大势所趋。如今，每个都在（或都应在）致力于将这些原则付诸实践。

与往常一样，经合组织仍是最活跃、最具全球影响力的平台之一。自去年在日本举行的G20峰会采纳其原则以来，今年的沙特阿拉伯S20峰会又重申了这些原则。新推出的人工智能天文台（经合组织，AI）不仅成功地汇集了真正跨学科的观点，深入研究了人工智能政策，探索了各国的人工智能倡议，甚至还通过一个契合背景、数据和输入、人工智能模型以及任务和输出的框架，更深入透彻地对人工智能系统的影响进行了评估。

在欧洲，欧洲议会未来与科学和技术小组（STOA）与经合组织全球议会网络建立了合作关系，重点推广值得信赖和以人为中心的人工智能，并就人工智能的未来发展共同进行反思。该倡议汇集了来自42个国家的议会成员，是人类在人工智能治理的多边合作方面迈出的重要一步。

2020年5月，联合国秘书长根据数字合作高级别小组的建议发布了数字合作路线图，这是2020年人工智能治理的一个重要里程碑。除了关于人工智能的建议3C外，关于数字公共产品的建议1B也证实并强调了全球数据获取框架（GDAF）的工作。特别是，它呼吁利用大数据和人工智能，“以具有可操作性的实时和预测性见解的形式创造数字公共产品，这对包括联合国在内的所有利益相关者都至关重要，因为它们有助于确定新疾病的暴发、反制仇外心理和谣言、衡量对弱势群体的影响，以及应对其他相关挑战”。秘书长接着指出:“全球数据获取框架等措施旨在发展技术基础设施，以促进通过各种方式进行的数据共享并扩大其规模，从而加快创造优质数字公共产品的进程。”

GDAF实践由联合国Global Pulse倡议、未来社会人工智能计划（TFS）和麦肯锡Noble Purpose AI共同领导，涉及超过120个利益相关者，包括主要技术公司、学术机构、非政府组织和联合国机构。我们的目标是在2021年第一季度发布一份蓝图，然后发布一个关于如何共享或获取各类数据的最简化可实行产品（MVP），以便运行人工智能系统，帮助我们实现可持续发展目标。

2020年，世界银行在帮助新兴国家采纳和实施国家人工智能战略方面发挥了重要作用。人工智能倡议很荣幸能参加这项活动，我们也以同样的方式与德国国际合作机构（GIZ）和世界经济论坛合作制定卢旺达国家人工智能战略。

推动国际人工智能治理的另一个相关平台是法国和加拿大发起的全球人工智能伙伴关系（GPAI）。2020年，该合作伙伴关系进一步发挥了其全球职能，促进了关于负责任的人工智能、数据治理、创新和商业化以及未来工作的更深入讨论。在此情形下，负责任的人工智能领域迅速成立了一个小组，重点关注基于人工智能的疫情应对方案。GPAI的15个创始成员为澳大利亚、加拿大、法国、德国、印度、意大利、日本、墨西哥、新西兰、韩国、新加坡、斯洛文尼亚、英国、美国和欧盟。2020年12月，巴西、荷兰、波兰和西班牙也加入其中。目前有一股强大的动力推动GPAI向发展中国家开放。我认为，寻找一种机制让作为世界领先的人工智能大国之一的中国加入其中具有显著意义。

未来社会与GPAI及其成员国密切合作，发表了两份关于负责任的人工智能小组和疫情应对措施的开创性报告，两者都可点击https://gpai.ai/projects/responsible-ai/；https://gpai.ai/projects/ai-and-pandemic-response/ 查阅。

CAIAC（发音为“kayak”）项目是有关本次疫情的一项工作，很好地推动了经合组织人工智能原则的实施，我很荣幸能与斯坦福大学以人为中心的人工智能计划（HAI）、Stability.ai，以及各联合国合作伙伴共同领导这一由联合国教科文组织发起，并得到帕特里克·麦戈文基金会的支持的项目。CAIAC是一个旨在增加我们对新冠病毒的认识，并提供独特决策支持的平台。它通过连接本地化的举措和干预措施，以一种动态的方式绘制有关病毒及其影响的知识图谱。这样一个平台对政策的制定非常重要，它使我们的领导人能够直接获得以人类智能为基础、经人工智能增强的相关知识和现有数据集，以应对本次危机中相关的健康问题，并制定以全球合作的方式摆脱这场危机所需的社会、经济和金融应对措施。我们知道以后还会出现其他疫情，同时也面临着联合国可持续发展目标所强调的潜在危机，首先是气候变化，这是一个明确而恰当的例子，它呼吁全球共同采取行动，并由遵循CAIAC知识收集和共享模式的人工智能系统提供增强性支持。

2020年甚至凸显了危机时期对全球合作的迫切需要。人们只能遗憾地看到，是党派政治限制了通过重要的多边行为体（例如世卫组织）进行全球协调，而不是常识。但我们仍然抱有希望，认为在努力实现联合国可持续发展目标的这十年中，会有更多旨在实现可持续发展目标的人工智能计划蓬勃发展，如获得百度、旷视、依图和滴滴支持的北京智源人工智能研究院（BAAI）的AI4SDGs智库，或由赛德商学院（Saïd Business School）领导、Facebook、谷歌和亚马逊的支持的牛津AI×SDG计划。这也清楚地表明，联合国要发挥更大的作用，不但要将各国政府联系在一起，还要引导东西方的人工智能平台，去推进以人工智能为基础的项目，来应对可持续发展目标这一复杂情况。这是一项可持续发展目标。希望我们能在2021年把这些项目整合到一起。

最后，我们一直在与世界各国政府和联合国合作伙伴（联合国区域间犯罪和司法研究所）合作开展一些具有影响力的项目，部署人工智能系统打击人口贩卖、向执法机构提供关于现有人工智能工具的应用教育，以及删除网络上的儿童性虐待材料（CSAM）。这些都是非常具体的实际案例。针对这些案例，我们可以利用人工智能的力量，帮助政府和决策者应对人类面临的挑战。我们希望在2021年，随着世界各国领导人更好地接受人工智能系统的潜力，并实施各种人工智能原则，我们能在坚实的全球合作下仿效CERN或ISSS项目的合作，大规模部署宏伟的人工智能造福人类项目。

作者简介

塞勒斯·霍德斯（Cyrus Hodes）

塞勒斯·霍德斯最近担任阿联酋总理办公室人工智能办公室主任顾问，过去3年一直负责领导迪拜世界政府首脑峰会的全球人工智能治理圆桌会议（GGAR）。

塞勒斯是硅谷跨阶段风险投资公司FoundersX Ventures的合伙人。他还是哈佛大学肯尼迪学院孵化的未来社会（一个非营利组织）的人工智能倡议联合创始人和主席。他在该项目中与众多全球利益相关者一起研究、探讨和构建人工智能治理框架。

在人工智能倡议方面，塞勒斯遵循联合国主要机构的建议，与斯坦福大学HAI共同领导针对新冠肺炎疫情的集体和增强智能（CAIAC）项目，并与联合国可持续发展目标（Global Pulse）行政办公室和麦肯锡（Noble Purpose AI）共同领导全球数据获取框架（GDAF）项目。

塞勒斯是全球人工智能伙伴关系（GPAI）、经合组织人工智能专家组（ONE AI）、欧洲议会未来与科学和技术小组（STOA）、联合国秘书长数字合作高级别小组建议1B（数字公共产品）和3C（人工智能）、全球扩展智能理事会（MIT-IEEE）的成员，也是IEEE《人工智能设计的伦理准则》的共同作者之一。塞勒斯是AI Commons指导委员会的成员，是其数据倡议的领导者，是智慧迪拜人工智能伦理委员会的成员，也是上海市科学学研究所（SISS）的人工智能治理顾问。

塞勒斯曾在巴黎政治学院就读，后担任该院国际安全讲师，获有巴黎第二大学国防、地缘政治学和工业动力学荣誉硕士学位，以及哈佛大学肯尼迪政治学院硕士学位。

第五部分：
国家和地区政策进展

人工智能治理，绝不能单靠技术专家来应对

尤金尼奥 · 加西亚(Eugenio Vargas Garcia)

在受新冠疫情困扰的过去一年，我们感受了悲伤，也收获了一个重要教训。与病毒十分相似，AI也将给全球各国带来方方面面的影响。疫情证明了我们需要开展国际合作，才能在满足各方利益的前提下成功解决跨国问题。

AI是一项具备多种功能以及长期影响的通用技术，面对其所带来的挑战，我们需要大量资源与专业知识技术，通过确定一致的因素特征加以预防或缓解，从而在确保安全的基础上充分发挥其潜能。

虽然在许多前沿领域都取得了可观进展，但关于AI国际治理的规范、政策、安全措施以及技术标准的制定仍不尽如人意。目前尚无超越高级原则的全球性管控制度或者规范文书。

紧张的政治局势、不断加剧的竞争、两极分化和部落主义在短期内对重大协议的达成并无裨益，由此带来的合作型治理形式的缺失，加上对多边机制的不信任，最终将使得AI风险的解决和应对变得更为困难。

在世界格局分崩离析的情况下，如果各国不能有效应对，未来对于AI的监管将呈现“巴尔干化”。我们应避免对应阵营相关规则互相矛盾的分割网（splinternet）局面出现。

对此，我们不能袖手旁观，坐以待毙。在没有规范的情况下，逐底竞争逻辑可能会驱使政府和私人企业无视法律、道德、安全问题而加大力度推进AI的快速发展。

对于AI治理的需求不言而喻。伴随着科技变得无处不在，着手设立合作机制以预防危害的需求也愈发强烈。通过建立规范实现技术的可预测性可推进负责任战略的制定，从而尽可能减少令人不安的局面的出现。

从现实层面来看，成立AI监管机构似乎遥遥无期。在此关头，举办由多个利益相关方自愿参与、旨在提供相关防范建议而非直接进行监管的论坛可起到一定的积极作用。

多边主义能够起到推进作用。例如，联合国凭借其无可比拟的覆盖范围和普适性，可提供中立、无党无派、合法的平台，以促进相关的协商谈判。

联合国教科文组织（UNESCO）在推进开展迫切讨论，拟定第一份有关设立AI伦理国际文书的全球标准方面进展显著。

联合国秘书长在其对于数字合作的路线图中确定了三大关键任务：提升发展中国家在AI审议事务中的参与度；提高现有倡议的总体协调性；以及推进开展能力建设，尤其是公共领域的能力建设。

2020年，联合国就创立AI咨询机构以提供专业意见并为有效共识的达成提供基础支持发起了多次协商。最终目标应是达成尽可能兼顾所有担忧问题的解决方案。

法国前总理乔治 · 克列孟梭曾说过：“战争过于严肃，不能只交给士兵去应对。”同理，AI也十分重要，绝不能只靠技术专家加以应对。

AI政策的制定需要弥合技术界与政治领袖、政府官员、外交官以及议会议员之间的鸿沟。将双方联系在一起对于AI的成功应用不仅有利，而且至关重要。

作者简介

尤金尼奥 · 加西亚(Eugenio Vargas Garcia)

尤金尼奥 · 加西亚是一名外交官，是巴西利亚大学国际关系学博士，也是AI与全球治理领域的一名研究者，目前担任巴西驻几内亚共和国科纳克里大使馆公使衔参赞兼临时代办。他曾于2018-2020年间担任纽约联合国大会主席办公室和平与安全事务高级顾问，也曾先后就职于巴西驻伦敦、墨西哥城和亚松森大使馆以及纽约巴西常驻联合国代表团。此外，他还在巴西外交部任职不同岗位，包括有关亚太事务和外交政策规划的各项职务、外交部部长顾问（2005-2009年）和外交部副部长（2014-2015年），并在2015-2018年间担任联合国司司长。他在2005-2009年间担任牛津大学客座副研究员，在2004-2005年间担任墨西哥国立自治大学拉美研究学院教授。加西亚共计出版过7本有关对外政策和国际事务的著作，并且还曾在1985年获得巴西青少年象棋锦标赛冠军。他的主要学术研究领域包括AI、新技术对和平与安全的影响，以及多边组织的作用。

欧盟的人工智能治理路径

伊娃·凯莉（Eva Kaili）

人工智能逐渐成为一项关键技术，它不仅有望改变商业模式和消费者的生活，最重要的是还对传统模式和公民观念发出了挑战。欧洲提出的方法是，在不损害公民隐私权、不损害人民生产的数据的所有权和价值，也不在错综复杂的应用以及数据采集过程中侵犯人民的安全和生活质量的情况下，加速创新和人工智能应用。

欧洲引入了一种基于技术中立这一基本概念的治理模式，在这种模式下，监管机构负责制定原则，市场则通过定义产品或服务的标准来实施这些原则。这种模式的颠覆性改变在于标准符合严格的基本原则，包括客户保护、社会包容、人权、隐私和非歧视等，这是一种全球性的保障模式。

在过去的一年里，欧洲的机构系统地开展工作，以建立竞争框架，意图使其成为人工智能的全球标准。该框架旨在提升人们对人工智能的信任，确保人们能在数字时代与智能系统共存，而不必担心被排斥、操纵、压迫或歧视。在以人为本的人工智能中保留选择的自由，将有效防止脑机接口挑战人类的本质和未来。与第四次工业革命走向不平等和非人化的趋势相反，技术和创新的最佳实践现在需要重新转向为人类服务，欧洲可以作为第五次工业革命的全球规则和标准制定者发挥领导作用。基于欧洲原则的人工智能系统框架必须通过法律来诠释和建立，尊重我们在数字时代的权利。在基础层面上，这一框架必须保证更高的透明度和可问责性，并确定人工智能系统的责任，从而建立标准来跟踪、审计、解释、申诉和推翻人工智能在其整个生命周期中做出的决策。

欧洲自动化的快速发展绝不能重复过去的错误，人工智能算法和系统必须通过多样化的数据集进行训练，其目标必须明确可控，以避免偏见和歧视或数据中毒的风险。人工智能做出的决定必须符合整个智能系统生命周期中定义欧洲的集体伦理结构，并划清红线，例如采用基于风险的方法，做到完全禁止致命性自主武器或有意识人工智能的研究。高风险的应用应就我们共同价值观的脆弱性、我们作为公民的权利、我们作为政策制定者的责任，以及我们对后代的义务发出警示。

作者简介 伊娃·凯莉（Eva Kaili）

伊娃·凯莉是欧洲议会议员，自2014年以来一直是希腊科技发展代表团成员。她是欧洲议会未来科技小组（STOA）和人工智能中心（C4AI）主席，以及欧洲议会工业、研究和能源委员会（ITRE）、经济和货币事务委员会（ECON）、预算委员会（BUDG）和数字时代人工智能特别委员会（AIDA）的委员。

伊娃是非加太-欧盟（ACP-EU）联合议会大会（DACP）、阿拉伯半岛代表团（DARP）和北大西洋议会代表团（DNAT）的成员。伊娃一直尽己所能，致力于推动创新，以此推动欧洲数字单一市场的建设。她是区块链技术、在线平台、大数据、金融科技、人工智能和网络安全领域多项法案的起草人、容克计划（Juncker plan）EFSI2的ITRE起草人，以及最近的InvestEU计划的起草人。

伊娃还是欧洲议会驻北约巴勒斯坦代表团的团长，主要负责欧洲的防务和安全。在此之前，作为泛希腊社会主义运动党（PASOK）的一员，她曾被选为2007-2012年希腊议会成员。在从政之前，伊娃还曾是记者和新闻播音员。

她拥有建筑学和土木工程学士学位，以及欧洲政治学研究生学位。

第三种路径：欧洲人工智能治理的后续行动

夏洛特·斯蒂克斯（Charlotte Stix）

美好的愿望要通过具体的行动来实现。显然，这一原则也适用于安全可靠的人工智能开发与部署。近年来，欧盟（EU）内部的人工智能高级专家组（欧盟委员会的一个独立咨询小组）率先对人工智能伦理和政策格局进行了研究，并通过人工智能协调计划对其进行了强化，欧盟成员国同意为该计划在若干相关政策领域进行合作和协调，并采取相关行动。我们现在可以看到这项工作正在被切实推进。最值得注意的是，欧盟已经采取了具体措施来监管人工智能。虽然这将是一个漫长的过程，但首批纲要已经很明显，并已在《人工智能白皮书——通往卓越和信任的欧洲路径》（*The White Paper on AI: A European Approach to Excellence and Trust*）中提出。《白皮书》中综合提出了有关伦理注意事项、法律义务和技术基础设施的提案，形成了欧盟委员会眼中同一事物的两个方面：一个“信任生态系统”和一个“卓越生态系统”。本篇短文将进一步阐述“信任生态系统”。

“信任生态系统”框架旨在概述欧盟委员会将于2021年初巩固的监管框架，目标是监管欧盟内部的所有高风险人工智能系统。作为全球首个基于风险的监管方法，该方法目标远大，倡导一种既能充分应对人工智能系统风险，又能促进其发展并提高接受度的方法。

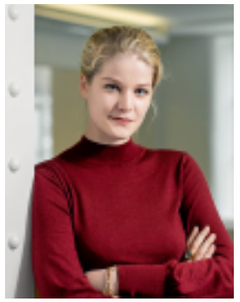
信任和可信赖性持续构成欧盟在人工智能方面活动的红线。事实上，能够促进一个繁荣的生态系统，为用户、公民和环境提供值得信赖的人工智能，被视为欧盟可以利用的一个主要优势，而基于风险的监管方法则可能有助于确保均衡性。

监管提案基于《可信人工智能伦理准则》（*Ethics Guidelines on Trustworthy AI*）对可信人工智能的七大关键要求，使用该非约束性框架来制定针对高风险人工智能系统，即监管范围之内的系统的要求。高风险人工智能系统或案例由两个标准定义：行业本身是否具有高风险（如医疗健康、交通运输等）以及预期用途是否涉及高风险（如伤害、死亡、重大实质性/非实质性损害）。如果上述两个标准均被满足（即累积标准），则适用拟议法律框架中人工智能系统的强制性要求。白皮书还建议，生物识别和将人工智能用于招聘过程等案例应被视为高风险案例。此外，非高风险应用可以选择参与自愿标签计划，该计划也借鉴了可信人工智能的七大关键要求。

在对人工智能进行适当监管并建立相关生态方面，我们还任重道远，但欧盟正在这条道路上不断向前。

作者简介

夏洛特·斯蒂克斯（Charlotte Stix）



夏洛特·斯蒂克斯是一位经验丰富的技术政策专家，专攻人工智能治理。她在埃因霍芬理工大学攻读博士学位，以批判的态度研究考察人工智能的伦理、治理和监管注意事项。

夏洛特是欧盟委员会人工智能高级专家组的协调员。此前她曾担任剑桥大学利弗休姆未来智能研究中心（CFI）的研究员、世界经济论坛人工智能理事会的研究员和Element AI欧洲人工智能项目的顾问，且曾在欧盟委员会机器人和人工智能部门担任政策官，管理过超过1800万美元的项目投资。

夏洛特入选2020年福布斯30位30岁以下（欧洲）精英榜，并被世界经济论坛评为全球杰出青年。

2020年的人工智能治理进一步推进政策落实以取得公众信任

林晨力（Caroline Jeanmaire）

2019年，AI监管框架终于开始陆续出现，超越了单纯的“AI原则”。出现这一变化的基本原因之一在于人们意识到监管并非阻碍创新的绊脚石，而是有助于取得公众信任，进而推动发展的基石。为了保持公众信任和避免另一个AI寒冬，AI治理必须考虑应用“盲点雷达”，以降低意外灾难风险。

美国的监管机构在以建立公众信任为首要目标行事的同时，也始终在关注这一问题。2020年12月，美国时任总统特朗普基于美国国防部和情报体系制定的AI原则发布了一道行政令，概述了各机构在设计、开发或获取AI应用时必须遵循的九项原则。该行政令还要求对机构内部的AI部署进行全面清查，并制定有关改进AI使用政策指导的路线图。这些监管原则有助于减少不确定性，从而鼓励创新。该行政令指出：“公众信任将在很大程度上决定AI能否得到持续应用和认可。因此，各机构必须以一种能够保护隐私、公民权利、公民自由和美国价值观，并加强公众信任和信心的方式来设计、开发、获取和使用AI。”为了监管和实施美国国家AI战略，白宫于2021年1月设立了国家人工智能计划办公室。拜登政府将在未来的几个月内进一步明确开发商和制造商必须对公众尽到的注意义务。

同时，欧盟将在人工智能高级专家小组制定的原则的基础上推出新的监管框架。2020年，欧盟委员会发布了一份白皮书，确定了AI技术应当满足的7大核心要求，以及采用风险管理方法的监管框架。此外，委员会还建议将独立软件纳入产品法规的适用范围，并确保贯穿供应链中各个经济主体的AI系统的安全性。

美国和欧盟的监管机构了解技术后冲风险，知道事故或失败可能导致公众产生恐慌，因为以前也发生过这种情况。1979年的三里岛核泄漏事故对美国核电产业的发展造成了长达30年的深远影响；在欧洲，包括引发疯牛病在内的各种食品安全问题导致公众对转基因食品避之不及。与此相反，强大的安全认证流程和公开透明的事故调查流程让公众对航空业充满信心。

但是，大西洋两岸的监管机构还尚未确定风险评估标准和风险管理策略。他们仍需制定对技术进行监督和重新评估的流程，以应对“无知而不自知”的情况。截至目前，已有超过50位专家在“AI监管展望公开信”上签名，力劝欧盟委员会不要听信一些可能对与AI相关的潜在风险有所低估的团体的建议，而是应该继续实施并强化拟定的法规。AI伦理和安全能促进创新；如果希望通过相关体系来激发公众对这一漫长过程的信念和信心，法规就不能只是一纸空文，而是必须要落到实处。

作者简介　林晨力（Caroline Jeanmaire）

林晨力是加州大学伯克利分校人类兼容人工智能中心（CHAI）的战略研究总监，致力于打造一个以AI安全及其与关键利益相关者的关系为重点的研究社区。林晨力重点研究确保AI系统安全性和可靠性的政策，以及国际AI合作模型。在加入CHAI之前，她曾担任未来学会的AI政策研究员和项目经理。林晨力还曾为与会人数超过200人的第一届和第二届迪拜世界政府峰会全球人工智能治理论坛的组织提供了有力支持，并作为联合国的青年代表随法国代表团工作两年。她拥有北京大学和巴黎政治学院的国际关系双硕士学位和巴黎政治学院的政治学学士学位。2021年，林晨力获得“为AI伦理与责任做出过杰出贡献的100位女性”提名。

从以人为本到解决全球性问题：日本人工智能应用的挑战和前景

江间有沙（Arisa Ema）

日本政府一直在“社会5.0”（Society 5.0）概念下推行信息政策，“社会5.0”是一个以人类为中心的社会，通过一个结合了网络空间和物理空间的系统，同时实现经济发展和解决社会问题。日本内阁府在2019年3月发布了《以人类为中心的人工智能社会原则》。此后，有关部门和机构制定了人工智能相关指南，如医疗诊断成像支持系统开发指南（卫生劳动福利部）、农业合同指南（农林渔业部）、人工智能使用手册（消费者厅）和人工智能教育认证系统［教育、文化、体育、科学和技术部、经济、贸易和工业部以及综合科学技术创新会议（CSTI）］。

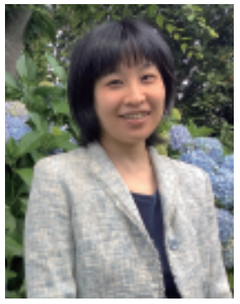
另一方面，我们对新冠肺炎疫情的应对情况表明：我们完全没有准备好向“社会5.0”过渡。事实上，政府和私营企业签署了一项关于新冠肺炎疫情的协议，以促进数据的使用，这可以被视为一种进步。然而，我们发现中央和地方政府以及医疗机构之间的数据共享是低效的。此外，我们还发现东京都政府甚至曾使用传真机报告感染人数。

数据共享的问题在先前就已被指出，并非因本次新冠肺炎疫情而引发。特别是在日本，企业对企业的公司多于企业对消费者的公司。换句话说，在许多情况下，数据采集、人工智能模型开发和服务供应商是不同的。因此，人工智能的安全性和公平性等问题不能仅由一家公司来解决，而是必须由多层次对话的所有参与方共同应对。此外，初创企业缺乏资源，难以应对风险。从这个角度来看，自2020年夏天以来，以初创企业为主要成员的日本深度学习协会除了考虑公司内部的人工智能治理之外，还在考虑如何评估和应对外部环境中的人工智能系统，包括保险、审计、事故调查、消费者保护、举报制度、标准化等。

最后，必须引发国际层面对人工智能治理的讨论。第二届法德日人工智能研讨会于2020年11月举行。两年前举行的第一届会议的联合声明强调了以人类为中心的方针。因此，按照这个方向，第二届会议的主题是以人类为中心的人工智能。然而，我们现在正面临着全球性的挑战，如新冠肺炎疫情、气候变化和社区分裂。因此，2020年的联合声明提出，从以人类为中心的问题到环境问题的各类全球性问题都应该得到解决。我们应该与各利益相关者和组织合作，共同开展人工智能治理工作，以解决全球性问题。

作者简介

江间有沙（Arisa Ema）



江间有沙是东京大学的助理教授，也是日本理化学研究所高级情报项目中心的访问研究员。她是一名科学技术学（STS）研究者，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所带来的效益和风险。她是可接受的负责任人工智能研究小组（AIR）的联合创始人（http://sig-air.org/），该小组成立于2014年，旨在解决人工智能和社会之间的新兴问题和关系。江间有沙是日本人工智能学会（JSAI）伦理委员会的成员，该学会于2017年发布了JSAI伦理指南。她还是日本深度学习协会（JDLA）董事会的成员，担任人工智能治理研究小组的主席。此外，她也是日本内阁府以人类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《以人类为中心的人工智能社会原则》。她获有东京大学的博士学位，此前曾担任京都大学白眉（Hakubi）高级研究中心的助理教授。

印度促使人工智能经济驶入快车道的策略

拉杰·谢卡尔（Raj Shekhar）

2020年，全球政府都面临着前所未有的挑战，特别是在发展中国家，为了遏制新冠肺炎病毒扩散而采取的全国性措施使得这些国家可能不得不将其他公共政策重点事项无限期搁置。但是，莫迪政府成功地推动了一些重要政策的磋商和对话，这些政策旨在提高印度的人工智能成熟度以支持国内公共和私营人工智能企业发展，与印度总理对AtmaNirbharBharat（自力更生的印度）倡议的展望接轨，令人称道。

印度政府在制定印度数据治理框架方面取得了巨大的进展。鉴于2019年《个人数据保护法（草案）》在决定印度数字公民和印度数字经济的命运方面具有无可争辩的重要意义，印度议会联合专责委员会在2020年召开了几次会议，以对该草案进行逐条审查。电子信息技术部（MeitY）拟定了（a）对包括一系列有利于促进国内劳动力和基础设施能力建设，以落实2019年《个人数据保护法（草案）》中的数据本地化授权的法规和项目在内的内容进行了详细阐述的数据中心政策，和（b）在《非个人数据治理框架》中，建议制定中央非个人数据法规，并根据该法规建立中央非个人数据管理机构，以释放数据的经济、社会和公共价值，同时确保国内非个人数据共享有一个公平、有利于创新和尊重隐私的管理制度。印度政府最重要的政策智库改造印度国家研究院（NITI Aayog）提出（a）建立国家数据和分析平台（预计今年推出），以一种对用户友好的开放方式公开政府数据，从而促进研究和创新、基于实证的公共决策，以及印度的参与式治理；和（b）采用数据授权和保护架构，以使公民能够安全无缝地与第三方机构共享其个人数据，轻松便捷地获取众多在线银行和保险产品。

此外，NITI Aayog意识到了随着全球各类AI用例的开发和部署而相继出现的各类伦理问题，因而发布了为全人类服务的有关负责任的AI的工作文件，以及有关为全人类服务的AI的执行机制。上述文件强调了坚持维护为印度负责任的AI治理确定相关原则的宪法基本权利的必要性等内容，并建议设立一个国家级多学科伦理与技术委员会，负责帮助行业监管机构基于风险制定相应的AI法规和其他措施，以协助在印度建立负责任的AI生态系统。

为了让印度的下一代专家能够拥有相应的AI技能，人力资源发展部（MHRD）在第三版国家教育政策中提出要通过在各大院校的课程体系中推出AI课程等方式，改变印度教育体系中的教学方法。

科技部（MST）在2020年末发布了第五项国家科学、技术和创新（STI）政策，总体上对印度的AI成熟度的提高具有一定的影响。该政策详细阐释了印度政府对通过扎实、动态和包容的制度治理机制实现技术自立的展望。上述治理机制旨在实现所有公共资助的研究成果的开放和广泛获取、增加对重点领域STI研究的公共和私人投资，以及为STI生态系统开发必要的人力资本。

尽管MeitY、NITI Aayog、MHRD和MST针对国家政策转变的提案可能存在缺点，但同时它们也大大提高了利益相关者在印度一系列紧迫的AI治理问题上的参与度。这表明，自2019年以来，印度政府在有关AI治理的定位方面有了明显的进步，为印度的AI经济打下了重要基础，使其能够发展壮大，并以百折不挠的姿态参与全球AI市场上的竞争。

作者简介

拉杰·谢卡尔（Raj Shekhar）

拉杰·谢卡尔是AI政策交流所的创始人兼执行董事。AI政策交流所是由从事AI和公共政策交叉领域工作的个人和机构组成的国际合作协会，致力于提供有助于创造AI文化社会，并制定更好AI政策的成果。拉杰还是加州大学伯克利分校社会利益信息技术研究中心政策实验室的关联学者，目前正与布兰迪·农内克（Brandie Nonnecke）博士（创办负责人）合作进行有关AI治理的研究。作为芝加哥大学国际创新小组（IIC）的（数据和AI）顾问，拉杰正为（a）IDFC研究所和IIC发起的开放数据工作小组（ODWG）倡议的运营，和（b）由IIC和电子与信息技术部共同参与，旨在通过实施相关政策和项目来建立数据和AI创新能力的工作提供支持。此外，拉杰还是施普林格·自然集团下属的《人工智能和伦理》（*AI and Ethics*）的创始期刊编委会的成员。

“跨行业GPS”：创造一个全行业通用且以人为本的未来工作格局

潘竞宏（Poon King Wang）

2020年，新加坡资讯通信媒体发展局和个人数据保护委员会在AI与数据道德使用咨询委员会的指导下与“李光耀创新型城市中心”（LKYCIC）联合发布了《AI时代下的职业再设计指南》。

该指南是新加坡首份全行业通用指南，旨在帮助各企业了解应该如何充分利用AI，并同时运用切实可行、以人为本的战略管理AI对员工的影响。

“以人为本”的需求很好理解。2020年的疫情突显了全行业通用的实现以人为本的方法的价值。

新冠肺炎向我们揭示了当各行各业都受到干扰时，我们有必要采取哪些举措以提升技能水平。相关行业的从业人员处境艰难，因为许多现有从业人员的技能提升举措都只适用于特定行业。但是，当相关专业和行业的从业人员几乎没有其他选择时，这类技能提升举措并不足以应对困局。

这一困境在未来将会变得越发普遍，因为疫情同时还加快了数字化进程，而这会导致更多行业由于自动化和远程工作而受到广泛的颠覆性影响（这可能导致外包）。基于同样的道理，特定行业的技能提升举措也不足以帮助受干扰专业和行业的从业人员。

这就是全行业通用战略的价值所在。全行业通用战略将使我们能够帮助从业人员从其所从事专业和行业之外寻找机遇。

LKYCIC的未来工作研究项目便制定了这样一种全行业通用方法，即能够取得有效成果的“跨行业GPS”。通过运用AI和已经建立的“任务－技能”框架，我们可以从已受干扰的专业/行业到正在不断发展的专业/行业，绘制出清晰具体且循序渐进的发展路径。

我们的全行业通用战略立足于科研界和产业界的共识，即任务是用于研究经济和技术对各职业的影响的正确解决方案。我们可以通过确定不同行业中职业的共享任务来绘制跨行路径，并以该路径为基础来制定相关举措，从而为受影响的从业人员提供帮助。我们的“跨行业GPS”借此为各行各业的从业人员拓宽了选择范围，帮助他们更好地应对当前和未来的危机。

我们的“跨行业”GPS结合了AI算法的力量和我们的任务－技能栈。任务－技能栈连接了国内和国际数据库的行业和职业数据。与此相关的研究均吸收和整合了数据科学、工程学、AI、劳动经济学、职业心理学和组织研究等学科的见解。

我们通过深入的研究，创建了新加坡的首份全行业通用指南《AI时代下的职业再设计指南》，并将加大力度和范围，建立一个先进而可信赖的AI环境，造福广大公民、企业和各国政府。

作者简介

潘竞宏（Poon King Wang）

潘竞宏是新加坡科技与设计大学（SUTD）“李光耀创新型城市中心”（LKYCIC）的主任，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、德国阿登纳基金会强大城市2030网络，以及人工智能全球伙伴关系未来工作工作组的成员。同时，他还是未来成人学习国家工作组和未来服务与数字经济国家工作组的成员。他与人合著并出版了《2040数字生活：工作、教育和医疗健康的未来》*(Living Digital 2040: Future of Work, Education, and Healthcare)*，该著作最近被翻译成了韩语。

他的团队的研究成果有助于建立一个先进而可信赖的AI环境，受到了新加坡国家AI战略的认可。他的团队还与新加坡资讯通信媒体发展局下属的个人数据保护委员会（由AI与数据道德使用咨询委员会指导）联合发布了《AI时代下的职业再设计指南》（*A Guide to Job Redesign in the Age of AI*）。该指南是新加坡首份全行业通用指南，旨在帮助各企业管理AI对员工造成的影响。

为人工智能治理做准备：重新思考公共部门的创新

维克多 · 法姆波德（Victor Famubode）

2020年对于人类文明而言是棘手的一年。人类在疫情暴发之际的惊慌失措、焦头烂额明显揭露了这一点。疫情暴露出了全球创新体系中的漏洞，但反过来又通过更广泛的数字化应用加速创造了新的机遇。在各种情况下，人工智能在不同行业和领域的应用都是这种数字化应用的核心。

对于政府而言，协同使用此类通用技术与其他技术和数据阻止新冠肺炎的传播是疫情期间AI应用的一个重要里程碑。例如，埃及政府通过联合国开发计划署（UNDP）推出了一项用手语聊天机器人来自动检测新冠肺炎症状的服务。值得注意的是，在公共医疗健康等高风险领域中应用AI用例有其自身的风险，最终需要非洲各国政府的认可才能实施。

AI应用在整个非洲大陆的快速增长显示出了重新思考公共部门创新的必要性，这需要从使部署和应用AI系统的负责任方式民主化的标记化转向上述系统的实操化。虽然“在53个非洲国家中，只有24个通过了数据保护法律法规（国际隐私组织2020年报告）”，导致数据保护措施进展缓慢，但能否妥善落实和遵循这些法规将成为能否解决AI和数据治理问题的决定性因素。不过，非洲许多国家的政府仍然面临着机构力量薄弱的问题。这意味着要落实和推广AI治理，就需要应对这些机构面临的挑战。

重要的是，2020年，非洲各国政府开始重视AI的道德影响，在这方面取得了一定的进展。各国政府与相关机构合作，让机构协助其制定将道德准则纳入考虑的国家战略，在这方面迈出了重要的一步。“卢旺达信息通讯技术和创新部（MINICT）和卢旺达公共事业管理局（RURA）通过德国国际合作机构实施的FAIR Forward计划让未来协会为卢旺达国家人工智能战略的制定提供支持”是一个重要标杆。

进入2021年，我们必须彻底解决算法偏见、监控、数字鸿沟和隐私等重大道德问题。为了应对这些问题，非洲各国政府在推进AI治理准备时需要考虑以下关键考量因素：

1. 利用社区渠道设计、部署和应用AI系统。这最终将有助于确保为公共消费建立的数据和模型具备多样性和变化性。
2. 重新评估公共采购流程，以大幅提高从私营技术公司采购AI系统时进行的咨询的广泛性和采购的透明度，这有助于在实施部署之前及早发现关键安全漏洞。
3. 帮助政策制定者、开发商和社会规划师构建负责任地使用AI系统的能力。
4. 进一步提高公众对AI系统的优点和局限性的认知。
5. 从设计到实施阶段，将风险和影响评估融入可能应用相关AI系统的所有公共领域。

作者简介

维克多 · 法姆波德（Victor Famubode）

维克多 · 法姆波德现任电气电子工程师学会自主与智能系统伦理全球倡议P7003（对算法偏见的考量）项目委员会委员，是AI治理和政策方面的专家，致力于帮助政府机构构建数据治理框架。他拥有媒体、政府和咨询领域的工作经验，主要致力于确保在技术领域实现更好的公共政策方案。

拉丁美洲人工智能治理及其对发展的影响

奥尔迦·卡瓦利（Olga Cavalli）

拉丁美洲是一个地域辽阔、地貌多样、生物多样性丰富的地区，拥有世界上最大的河流、山脉以及许多自然资源和美丽风景，同时还具有丰富的文化和高水平的人力资源。

最近的新冠肺炎疫情事件对拉丁美洲的区域经济产生了深远的影响。在这种复杂的情形下，可以通过利用人工智能在国家和地区产业中引入变革和创新，提高其生产力。美洲开发银行（IADB）的一份报告显示，未来几十年，人工智能的使用将有效助力拉丁美洲地区最大经济体的GDP增长。

尽管2020年国家优先事项发生了变化，更加注重卫生和经济等领域，但拉丁美洲地区的一些政府正在制定公共政策和国家战略，以促进人工智能在全国的使用和发展。

墨西哥是世界上最早制定国家人工智能战略的国家之一；巴西已经启动了国家物联网计划，其中包括建立人工智能实验室，重点关注网络安全和国防等战略领域；智利正与社会团体和全国的专家合作制定自己的人工智能计划；阿根廷正在制定国家人工智能战略；哥伦比亚已经公布其人工智能伦理框架。

拉丁美洲地区面临的挑战是需要达成一个多利益相关者计划，该计划必须涵盖政府、私营部门、民间团体、技术界和学术界，必须将精力和相关资源集中投入到使用最先进的技术和加强战略领域和行业的教育上。拉丁美洲地区是全球不平等现象最严重的地区，而人工智能的使用将能有效改善这一现象。

拉丁美洲地区的问题之一，是其在确定了全球人工智能治理和伦理框架的领域的发言权有限，参与度低。这些领域通常由发达经济体主导，如果不增强拉丁美洲的发言权，最终的框架可能无法满足该地区的需求。

为了克服这一问题和加强区域专家在国际谈判中的相关参与度，拉丁美洲已经采取了具体的措施。例如，南方互联网治理学校（SSIG）致力于培训学生和青年专业人员，使他们在相关谈判中成为拉丁美洲地区的领导者。该培训免费提供，自2009年创立以来，已先后推出12个版本，培训了来自拉丁美洲地区和其他地区的成千上万的研究员，为他们提供了有助于驾驭国际技术生态系统的工具，并在他们和国际专家之间建立了一个非常强大和有价值的网络。

在如今由高度发达国家不断带来的创新所主导的世界里，能力十分重要，但拉丁美洲地区所有国家都拥有创新型公司和人力资源，这些公司和人力资源具备相关条件，能够根据人工智能应用将自己定位为领导者。

作者简介 奥尔迦·卡瓦利（Olga Cavalli）

奥尔迦·卡瓦利是阿根廷布宜诺斯艾利斯大学经济学院的教授，她专注于研究互联网基础设施。她是南方互联网治理学校（SSIG）的主任，曾参与编辑《拉丁美洲的互联网治理与监管》（*Internet Governance and Regulations in Latin America*）一书，并多次投稿发表有关自己专业领域的论文和章节书。

她曾任物联网与智慧城市研究小组主席（2015-2017年），并被国际电信联盟（ITU）认可为该领域的创新先驱。在2007至2014年期间，她得到联合国秘书长的器重，被选为互联网治理论坛多利益相关者咨询小组（MAG）的成员。奥尔迦还积极参与互联网名称与数字地址分配机构（ICANN）的相关事务，并且在最近被任命为GNSO理事会的成员。奥尔迦参加过国家人工智能战略发展小组。她目前的研究涵盖人工智能和5G技术对拉丁美洲国家发展的影响。

奥尔迦拥有商科博士学位、工商管理硕士学位、电信监管硕士学位和电子电气工程学位。她精通西班牙语、英语、葡萄牙语和德语，能听懂法语和意大利语。奥尔迦现居阿根廷的布宜诺斯艾利斯。

拉丁美洲的人工智能治理现状

埃德森·普雷斯特（Edson Prestes）

拉美地区关于人工智能治理的讨论仍处于起步阶段。尽管该地区有一些关于人工智能的国家计划，但有关治理本身的辩论在广泛性和包容性方面还远远不够。它们仅限于社会的特定领域，不涉及正在进行的多层次和政府间参与。事实上，主要参与者之间存在着巨大的信息鸿沟，阻碍了对彼此的理解，也阻碍了对该领域复杂性的理解。我们显然需要加强人力和机构能力建设，以增强拉美地区公民和国家的能力，从而建立有效和实用的治理机制。为充分开发治理机制，主要利益相关者需要了解技术的局限性、广泛使用该技术会带来哪些影响、基于人工智能的技术可以创造哪些机遇、人工智能领域的多学科性质、对灵活监管机制的需求、对跨国协作的需求、践行可问责性和透明度的必要性，以及公众参与的重要性。

拉美地区的人工智能国家计划受到国际伦理和监管原则的强烈影响，如经合组织、世界经济论坛等制定的原则。然而，这些片面的计划似乎能在一定程度上满足拉丁美洲的现实需求。我们不能简单地引进或复制其他地区发展起来的成熟模式，而不深入思考它们在当地的影响。我们的情况具有特殊性，有些影响是本区域固有的，而另一些影响则已由高收入国家解决。举一些例子：一些尚未通电的社区既无法充分利用人工智能生态系统，也无法在人工智能模型中体现他们的价值观。拉美地区的一些国家并不重视教育，尽管这是人工智能最重要的组成部分，因此，我们看到教育投资逐年减少。此外，拉美地区还需制定相关策略，并建立一个丰富的生态系统，以防“人才外流”。

当然，我担忧的问题非常多，但它们都涉及到一些非常基本的问题：教育和信息。在巴西，人们普遍会为了获得购物折扣而提供自己的信息，或者为了获得福利而被迫提供信息。一些巴西人认同巴西政府目前的观点，即假新闻是言论自由的一种体现，而不是对人权的侵犯。这自然引发了如下问题：如果公民不了解，或者更糟的是没有得到足够的正规教育来了解自己的权利和义务，那么人工智能的治理机制如何保护人们免受人工智能直接或间接造成的滥用问题或人权侵犯？如果在进行相关讨论时不以人为中心，即当主要利益相关者（即政府或公司）抱有以利润为导向或扭曲的观点时，任何治理机制都不可信，也不可靠。

作者简介　埃德森·普雷斯特（Edson Prestes）

埃德森·普雷斯特是巴西南大河州联邦大学（UFRGS）信息学研究所的教授。他在1996年获得了巴西亚马逊帕拉联邦大学计算机科学学士学位，在1999年和2003年分别获得了UFRGS计算机科学硕士学位和博士学位。埃德森是IEEE机器人与自动化协会（IEEE RAS）和IEEE标准协会（IEEE SA）的资深会员。在过去的几年里，他一直致力于投身于与标准化、人工智能、机器人和伦理相关的各种倡议。例如，埃德森是联合国秘书长数字合作高级别小组成员、联合国教科文组织人工智能伦理特别专家组成员、IEEE 技术伦理计划南美大使、IEEE RAS/SA 7007—伦理驱动机器人和自动化系统本体标准工作组的主席、IEEE自主和智能系统伦理全球倡议成员、未来学会顾问、卡内基道德与国际事务委员会卡内基人工智能和平等倡议顾问、联合国教科文组织全民信息计划（IFAP）信息无障碍工作组成员，以及IEEE RAS工业活动委员会的前任副主席。

2020年：拉丁美洲寻求人工智能道德治理的关键一年

康斯坦萨·戈麦斯蒙特（Constanza Gómez-Mont）

2020这一年，数字技术（特别是人工智能技术）对于人们生活的影响加深了有关伦理问题的思考。从短期和长期来看，为了应对健康危机而放弃一定程度的隐私意味着什么？只有拥有数字工具的家庭才能继续学习和工作，这又意味着什么？数字平台使用率的上升如何进一步加剧已经存在的数据垄断现象？在大多数基本服务都迁往线上的情况下，缺乏数字素养意味着什么？

对于拉丁美洲这个社会不平等现象最严重的地区而言，这些问题不能掉以轻心。为了寻求一种更加公正和基于权利的方法，确保数据和人工智能驱动的技术能加以采用，拉丁美洲地区的各个机构纷纷在2020年制定了关于合乎伦理的人工智能治理的关键倡议。

例如，美洲开发银行推出了 fAIr LAC倡议，通过制定相关指南和与公共和私人伙伴合作，在墨西哥、乌拉圭、哥斯达黎加和哥伦比亚携手建立中心，帮助政府和企业家采用负责任的人工智能实践。此外，IEEE全球伙伴关系与C Minds共同创建了Latam Circle，旨在优先考虑运用自治和智能系统提升该地区的人类福祉，并促进拉丁美洲专家为全球人工智能伦理标准的制定做出贡献；联合国教科文组织开设了一个在线社区，用于后续跟踪合乎伦理的人工智能全球工具的区域协商工作。联合国区域间犯罪和司法研究所和Eon Resilience实验室等其他机构也开始努力在开发基于权利的人工智能全球工具包，以用于预防犯罪和司法用途时考虑到区域专家的建议；包括墨西哥的IA2030Mx和具有区域保护伞的AI Latam在内的网络正在加强人工智能生态系统。该行动清单是其他地区推进合乎伦理的人工智能治理时进行的工作之一。

此外，一些政府在2020年也没有落于人后。乌拉圭政府公布了一个用于评估人工智能系统造成的影响的工具；哥伦比亚政府公布了一份人工智能伦理框架草案；墨西哥联邦公共信息获取研究所协助启动了该地区关于人工智能系统透明度和可解释性的首个政策原型；智利政府开始制定人工智能政策，其中包括将于2021年公布的伦理支柱（a pillar of ethics）。

这些只是巴西和阿根廷等国家在经过几年的努力后推出的一部分现行举措。毫无疑问，拉丁美洲在巩固和扩大这些举措的范围，部署更协调的行动方面还有很长的路要走。最重要的是，要让所有人都能享受到数字技术和人工智能带来的好处，特别是拉丁美洲还在抗击新冠肺炎疫情，正处于经济复苏阶段。在这个迫切需要重新思考伦理问题和加快得到具有包容性的答案的时代，人工智能在拉丁美洲的道德意识设计和治理方式，将对该地区的繁荣发展产生深远的影响。对此，拉丁美洲应该不负众望，为发展中国家的领先实践，以及全球人工智能治理进程的发展做出贡献，这一点十分关键。

作者简介 康斯坦萨·戈麦斯蒙特（Constanza Gómez-Mont）

康斯坦萨·戈麦斯蒙特是一名社会影响战略家和实践者，她将政府、公司、跨国组织和当地社区聚集在一起，共同制定相关倡议，加速显现新兴技术的积极影响。

康斯坦萨是C Minds的创始人兼总裁，这是一家由女性领导的行动机构，致力于新技术、社会和环境的融合。作为C Minds领导层的一员，她已在新兴经济体的数字经济、人工智能和社会创新政策和倡议制定领域深耕超过12年，特别关注拉丁美洲。此外，她还是人工智能改善气候全球倡议的联合创始人。

康斯坦萨对建立社区和创建区域平台充满热情，这促使她成为了世界经济论坛人工智能造福人类全球未来理事会人工智能治理工作组的领导者；共同创建并主持了IEEE全球伙伴关系Latam Circle；受邀加入联合国教科文组织人工智能伦理全球文书草案高级小组；并协助美洲开发银行（IDB）建立人工智能促进社会影响区域倡议（fAIr LAC）；以及其他关键倡议。

康斯坦萨的成就贡献曾被国际媒体报道，并得到了巴黎和平论坛、世界经济论坛和英国政府等机构的认可。

走近拉丁美洲区域人工智能战略

让·加西亚·佩里希（Jean García Periche）

2020年的全球危机使世界陷入瘫痪，并让所有人开始重新思考人类的未来。这场疫情迫使我们加快了人工智能（AI）的数字化和采用。随着人工智能对各领域的颠覆性影响越来越普遍，国际社会必须协调采用综合战略，以应对认知技术带来的全球挑战。在拉丁美洲，人们对人工智能治理、实施和部署的认识仍相对缺乏。随着世界其他地区迅速认识到人工智能在治理人类未来方面的核心地位，拉丁美洲正在努力适应一种可使用计算机智能的令人信服的说法。

尽管拉丁美洲地区的科技独角兽公司越来越多，但人工智能实施项目往往规模较小，很少超越试点阶段。人工智能为社会带来的大范围颠覆性影响将给经济带来深刻的结构性变化。如果拉丁美洲不迅速采取行动，其地位就可能变得无关紧要。在人工智能呈现指数级发展的情况下，拉丁美洲的权力结构仍然是垂直驱动的，效率非常低。

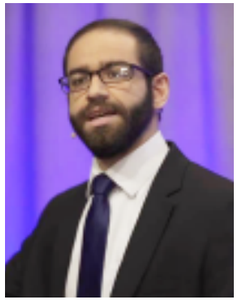
然而，混乱之中仍然存有希望。拥有超过6亿人口的拉丁美洲，是可以大规模开发和部署机器学习系统并充分利用丰富的数据资源的理想之地。如此一来，拉丁美洲就能成为决定这项技术未来的关键，并成为在人工智能全球治理领域拥有实际话语权的新兴主体。人工智能应用就绪水平最高的100个国家中，有15个位于拉丁美洲。

随着一些国家开始制定和实施国家人工智能战略和数字转型框架，一些拉美国家在去年取得了显著的进展。就数字政府而言，哥伦比亚是经合组织成员国中第三发达的国家。同样，墨西哥、乌拉圭、智利和巴西也在人工智能就绪水平和关键治理问题（如数据隐私）方面采取了关键措施。尽管如此，拉丁美洲在全球人工智能治理论坛上的发言权和参与度仍十分有限，令人震惊。如果新兴政策标准中不包含拉丁美洲标准，那么未来建立的人工智能治理框架可能会阻碍拉丁美洲的进步。

需要注意的一个基本要素是，国家倡议是必要的，但并不足以完全解决问题。如果没有统一的标准和标准化的框架，没有一个拉美国家能够独自成为人工智能的领导者，拉丁美洲需要区域凝聚力和良好的政治团结性。在这十年中，拉丁美洲地区需要围绕区域人工智能战略建立一个强大的国际联盟，将这项技术作为向新发展阶段跃进所必需的工具加以整合。

通过将人工智能作为优先事项，拉丁美洲将能实现社会经济和政治制度的升级，使其符合21世纪的情况。为此，拉丁美洲需要团结和国际合作。在人工智能时代，团结在拉丁美洲意味着生存。

作者简介 让·加西亚·佩里希（Jean García Periche）



让·加西亚·佩里希是GENIA Latinoamérica的联合创始人和首席政府关系官。GENIA Latinoamérica是一个区域研发（R&D）平台，其使命是将拉丁美洲纳入全球人工智能开发体系。让·加西亚·佩里希还是美国国家航空航天局埃姆斯研究中心奇点大学的研究员，担任公共部门高级官员的顾问和战略远见研究员。此前，他创立了Global Neo，这是一个通过实施基于区块链的系统和分散的代币经济来制定新型全球治理模式的组织。此外，他还是多米尼加共和国政治创新中心的主席，致力于通过循环经济模式推动智慧城市和社区可持续发展。

制定人工智能政策需要合作共建和共同学习

何塞 · 古里迪 · 比斯托（José Guridi Bustos）

全球所有国家都已经制定（或正在制定）AI政策和战略，好似对如何充分利用第三个AI爆发期进行规划是某种当务之急，但是又对为何以及如何进行规划不甚了解。只要对全球的AI战略加以分析，就能明显察觉到后一个问题，因为你能看到许多重点（例如研究、开发和伦理）、治理措施（例如公共措施、私人措施和混合型措施）、流程（例如自上而下、自下而上、以政策制定者为中心和以学术为中心的流程）等。

考虑到AI是一种正对社会造成极大影响（并受到社会极大影响）的具有普遍用途的技术，所有的小题大做都是可以理解的。在未来的5-10年内，AI将彻底改变我们的信息感知和处理方式、工作方式以及人际互动方式，并为我们生活中的许多其他基本要素带来极大的变化。因此，政策制定者迫切需要促进社会经济发展，并对AI实施监管，以防AI对人们造成伤害。但是，即便这种焦虑感确实存在，政策制定者对如何正确处理这一问题也并没有多少头绪。

制定AI战略时出现了名为“认知层次”的概念，科学家和政策制定者试图在这一概念的框架内通过区分自身掌握的专业知识和“非专业”知识来维护自己的资格和权威，而后者已被证明会对民主制度造成负面影响（Jasanoff，2005年），并导致政策制定工作被限制在一种主要障碍为研发不足的创新赤字模型中（Pfotenhauer、Juhl和Aarden，2018年）。该框架可能没有意识到AI是一种以“作为人类，我们应当如何与AI共存”为主要问题的社会技术结构。此外，这种框架还可能导致政策工具与围绕AI政策的手段和目标进行的全方位政策讨论发生脱节。

在智利，2019年发生的社会动乱和此次的新冠肺炎疫情为设计参与性方法，以建立AI国家政策打开了一扇机会之窗，因为这两个事件减轻了公众参与的阻力，特别是政策制定者和有关当局在这方面的阻力。我们公开呼吁进行自发召集的圆桌讨论会，任何地方的人都可以参加并做出贡献，唯一的要求是以拟定的（1）有利因素，（2）开发与采用，和（3）伦理、监管方面和社会经济影响为轴作为指导。这一流程既是共同制定政策草案初稿的一次实践，又是一个学术界、产业界、政府和社会彼此互动、自专业知识中吸取经验，并相互学习的学习空间。超过7000人参加了相关讨论和网络研讨会，并得到了在最近刚刚闭幕的公众咨询会中将讨论内容与草案进行对比的机会。

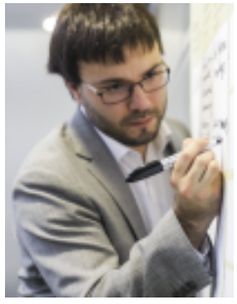
思考未来几十年的AI治理时，应分析与智利的经历类似的经验并加以改进，以设计出承认技术的社会技术性质的制度。各个国家，尤其是新兴国家，应根据自身优势和不足，制定开放的流程，建立自己的发展模式，而不是一味复制国际上的经验。

参考文献

[1] Jasanoff, S.（2005年），《饱受批评的判断：专家合法性的三体问题》，出自《专业知识的民主化（第209-224页），施普林格，多列特克。

[2] Pfotenhauer、S. M., Juhl, J.和Aarden, E.（2019年），《挑战创新的“赤字模型”：在创新势在必行的前提下处理政策问题》，《研究政策》第48(4)期，第895-904页。

作者简介 何塞 · 古里迪 · 比斯托（José Guridi Bustos）



何塞 · 古里迪 · 比斯托是智利经济、发展和旅游部经济与中小企业部的副部长顾问。在此之前，他曾任科学、技术、知识和创新部未来团队的顾问，带头制定了智利的人工智能国家政策。同时，何塞还是智利天主教大学工程学院的一名客座教授。

何塞是经合组织人工智能专家网络、国际投资发展银行fAIr LAC倡议以及未来协会分支机构的成员。他拥有智利天主教大学的工业和系统工程硕士学位（工程学位）。

第六部分：
国家和地区政策进展（中国）

人工智能和国际安全：挑战与治理

傅莹

近年来，人工智能（AI）技术的快速发展带来了巨大的机遇。由于技术变革往往伴随着无法预料的安全挑战，我们必须特别关注AI武器化可能带来的道德和技术危害。如今，人类亟需思考如何有效平衡AI技术所带来的效益与其武器化的安全风险，以及如何探寻适当的AI治理方法。

就目前来看，虽然基于AI的武器系统具有强大的军事效用，但却并不完全可靠，其应用存在大量的潜在挑战。AI固有的技术缺陷可能使攻击者难以控制攻击范围，从而给被攻击者带来过多的附带伤害，并造成平民意外伤亡。同时，基于大数据的算法和训练数据集会不可避免地将偏见带入AI系统的实际应用中，而训练数据集也可能被他国篡改，可能导致AI系统向决策者提出错误的建议，并误导军事指挥官进行错误的部署。此外，AI在可解释性、学习能力和常识认知方面的缺陷将加大人机合作期间发生战场冲突的风险，甚至致使国际危机螺旋式升级。

人类若想应对AI治理的全球共同挑战，就必须进行国际合作，因为任何一个国家都无法独立解决这一问题。各国需要在军事领域保持克制，并携手建立相应的国际治理机制。应禁止使用尚未明晰责任或风险的辅助决策系统。使用基于AI的武器时必须限制其攻击范围，以免造成附带伤害和冲突升级。开发和使用基于AI的武器时必须遵守现行的国际法律法规。应高度重视AI数据安全，对从数据挖掘和收集、数据标注和分类，到数据使用和监控的整个流程进行监管和限制，以防形成错误的模型，导致决策者在此基础上做出错误的判断。

现阶段，我们拥有建立AI安全国际规范的关键机会。中国和美国应在这方面展开对话与合作，因为双方或许有能力为AI治理全球合作献计献策。中国和美国应展开官方讨论，研究如何建立国际规范和制度、在各自的利益和关注点的基础上探索可以合作的领域、交换并翻译相关文件，以及进行政策对话和学术交流。这些措施同时也有助于降低双边关系与全球安全方面的潜在风险。

作者简介 傅莹

清华大学战略与安全研究中心主任、第十三届全国人大外事委员会副主任委员。

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年任驻印度尼西亚使馆公使衔参赞，1998年任驻菲律宾大使，2000年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，傅莹当选为第十二届全国人大代表；2018年当选为第十三届全国人大代表。2013年至2018年任第十二届全国人大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。2018年任第十三届全国人大外事委员会副主任委员。

2018年兼任清华大学战略与安全研究中心主任、清华大学兼职教授。2020年受聘为清华大学人工智能国际治理研究院名誉院长。

中国持续推动人工智能治理全球合作

赵志耘

2020年，突如其来的新冠肺炎疫情在给世界人民健康带来巨大影响的同时，也给全球治理带来了巨大的挑战。基于人工智能技术的各种解决方案，不仅在病毒溯源、疫情防控、疫苗研发等方面发挥了积极的作用，也在加速推动网上购物、在线教育、远程医疗等“非接触经济”全面提速。

与抗击新冠疫情需要全球协作一样，伴随着人工智能技术在全球的加速应用，人工智能的治理也更加需要全球的共同努力。2020年以来，中国围绕落实《新一代人工智能发展规划》和《新一代人工智能治理原则——发展负责任的人工智能》，在强化人工智能治理问题研究、推动人工智能治理实践的同时，也在积极倡导人工智能的治理领域的国际合作。习近平主席在今年二十国集团领导人峰会上特别提出，适时召开人工智能专题会议，推动落实二十国集团人工智能原则。针对数据这一人工智能治理的重点领域，中国政府还发起了《全球数据安全倡议》，明确提出共商、共建、共享来解决全球数字治理问题。

这一系列举措和倡议，充分体现了中国在推动人工智能治理全球合作方面的持续努力。未来，中国将围绕人工智能治理问题，继续搭建平台、拓展渠道，开展更宽领域的国际合作与交流，凝聚更广泛的共识，与全球各界共同推动人工智能的健康发展。

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。

稳步起飞：中国人工智能社会实验全面展开

苏竣

当前，人工智能、大数据、物联网等新技术新应用新业态方兴未艾，特别是在当下全人类共同携手抗击新冠疫情、助力经济高水平高质量复苏的关键时期，人工智能显示出促进经济发展、推动社会重构的巨大威力，成为我们走向未来的最强大的动力之一。但同时，从“技术—社会”学理论的角度来看，从公共政策研究的角度来看，以及从社会公众的感受的角度来看，人工智能在赋能发展的同时也带来了诸多的挑战，引起了法律、隐私、伦理、道德、安全等诸多方面的治理问题。这些问题的规避与化解都需要基于循证方法的科学研究，探索和研判人工智能对人类社会产生的综合社会影响。

自2019年中国的专家学者发出倡议以来，各方积极响应，采用社会实验的方法研究人工智能的综合影响，开始了一场基于循证研究的大规模调查。2020年，在中国克服新冠肺炎疫情影响，通过统筹规划、顶层设计、重点发展、示范推动，人工智能社会实验的学术研究、组织建设、队伍建设、人才培养、基地建设等已经在全国有序展开、稳步起飞，实现了公共政策研究“理论研究—政策建议—政治决策—行政执行—组织实施”的全循环。

从“发出倡议”到“稳步起飞”，人工智能社会实验的开局起步离不开“建设具有人文温度的智能社会”的广泛共识。有人文温度的智能社会，是一种科学技术高度发达、智能技术广泛应用、技术理性和价值理性综合平衡、人—环境—技术和谐共生、社会开放包容、人文精神张扬的以人为本的社会。这种共识植根于中国发展新兴技术的人本宗旨，发端于人工智能社会实验的生动实践，面向人类未来共同的智能社会图景。

在这种社会共识的推动下，中央部委统筹布局，成立人工智能社会实验专家组，编制人工智能社会实验工作规划。清华大学、浙江大学、北京大学、北京师范大学、中国人民大学等高校积极行动，制定人工智能社会实验的方法模式和操作流程，确立人工智能社会实验的伦理规范，并在城市治理、农村电子商务等领域开展先导性实验，综合运用自然实验、准实验、问卷调查等方法，研究人工智能应用过程中对个体和社会组织带来的影响。迄今为止，在中国学术界、产业界与政府部门的密切配合下，在科学抽样和伦理审查的前提下，围绕医疗、教育、养老、环境保护、城市治理、农业农村等领域，在北京、上海、浙江、广东、湖北等十余个省市，建设了上百个人工智能社会实验场景，形成了一批创新性强、亮点突出的案例，为开展长周期、宽领域、多学科人工智能社会实验打下了坚实的基础。

智能社会是人类从未触及的蓝海，又是我们的未来。现在，人工智能社会实验已经平稳起飞，相关工作已经有序开展。在以人为本的价值引领下，基于科学循证的规范研究，人工智能社会实验将为人类社会从工业时代向智能时代的成功转型提供客观准确的事实依据和“有人文温度”的解决方案。

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。

人工智能治理需要“技术创新+制度创新”

李修全

人工智能的加速发展和应用正在对伦理、安全、隐私、公平等各方面治理提出迫切需求，需要新的法律、制度设计和治理规则引导和规范技术发展。人工智能治理问题既是制度设计问题也是技术研发问题，应对人工智能治理对于技术创新的需求日益迫切。

每一类治理问题都蕴含大量的技术研发需求。当前以深度学习为代表的大数据驱动的端到端的机器学习方法，采用黑箱模式，模型内部结构和机理对用户不透明，缺乏对越来越庞大和复杂的机器智能进行理解，对“智能”行为出现错误时就无法做出溯源和解释，给制度设计者进行法律、规范设计带来非常大的困难。研发更具透明性的模型算法，开发可解释、可理解、可预测的机理型智能技术，将为解决责任认定困境、实现伦理约束等治理问题提供技术途径和实现可能。

同样，样本攻击、传感器干扰、深度学习框架漏洞，都可能给智能系统安全带来冲击和挑战。安全可靠的人工智能系统应具有强健的安全性能，能有效应对各类蓄意攻击，避免因异常操作和恶意攻击导致安全事故；隐私侵犯和歧视或偏见等治理，也需要从数据采集、数据保存、知识处理等各环节的技术本源入手破解，用户隐私数据脱敏技术、联邦学习和小数据增量学习正在为应对这些风险挑战提供有益探索，将通过技术创新增强社会公众对技术的信任。

为应对人工智能治理挑战，技术创新与制度创新的融合在2020年正在走向深化，人工智能技术研发者开始将越来越多的精力投入到研发应对安全、隐私、公平问题的技术解决方案。未来，发展负责任的人工智能技术应该成为科研界下一步理论创新和技术研发者关注的重要方向，通过技术途径破解治理难题，更需与法律、行政各方强化责任共担，携手应对，共同保障人工智能产业健康发展。

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10余项研究课题。

发展负责任的人工智能：从原则到实践

王国豫

2020年是全球疫情下人工智能继续飞速发展的一年，也是人工智能伦理原则不断丰富的一年。据非营利组织Algorithm Watch所制定的全球人工智能伦理指南清单（AI Ethics Guidelines Global Inventory），截至2020年4月，已收录伦理准则超过160部，发布主体涉及政府、学界、企业等各类机构和组织，其中许多伦理原则都直接以负责任的人工智能（Responsible AI）为标题。仅由此看来，发展“负责任的人工智能”已在一定程度上成为国际共识。

然而，发展“负责任的人工智能”不仅需要原则和呼吁，更需要行动和实践。为此，首先需要明确责任的主体和客体以及问责的主管（Instance），解决谁负责任（who），对什么事情负责任(what)和向谁负责(whom)的问题。比如，对于诸如人工智能应用中带来的隐私泄露、歧视和自动驾驶中的责任问题，谁是责任主体，是设计师、工程师还是企业的决策者或监管者？是个体还是机构或者国家？当一个行为涉及不同机构的多主体时，责任应该如何分配？当企业利益和社会伦理发生冲突时，如何进行优先排序？尤其是当一方面人工智能的发展涉及国家核心竞争力，另一方面与伦理原则发生冲突时，要不要发展（比如人工智能在军事上的应用）、怎么发展（人脸识别技术）？谁有权力问责、如何问责？如果要使人工智能伦理原则走向伦理实践，就不能回避这些问题。

要回答这些问题，还需要我们加深对这些问题的理论探讨。但是另一方面，更重要的是在发展AI的实践中厘清责任问题，推进负责任的机制建设和文化建设。为此，中国计算机学会于2020年成立了跨学科的“职业伦理与学术道德委员会”，由计算机专家和哲学家共同担任主席和委员。在2020年中国计算机学会的年会CNCC上，委员会举办了“为世界更和谐——信息技术中的职业伦理”专题论坛，一方面从理论层面对信息技术时代计算机和人工智能职业伦理的历史与维度、发展负责任的人工智能的理论路径等进行了分析，同时，也对在高校开展人工智能伦理教育的实践以及人工智能敏捷治理的框架等问题进行了深入讨论和交流。中国计算机学会青年计算机科学家与工程师论坛（CCF YOCSEF）也探讨了如何将公平、透明、可接受性与可持续发展等理念纳入第三代人工智能设计中去的问题。中国计算机学会正在尝试从伦理、机制、教育和技术四个层面推动人工智能的负责任原则转化为工程师的负责任实践。

要将发展“负责任的人工智能”落实到行动，还需要考虑的是，不能仅仅将人工智能看成是一个孤立的、封闭的算法和技术系统。人工智能的伦理问题是人与AI共生、技术与社会交互塑型的社会－技术系统中的伦理问题。发展“负责任的人工智能”需要面对的不仅是AI的技术伦理和工程师的职业伦理问题，而且还要面对社会的经济、政治和文化的多维度语境（场景context）下的不确定性和社会伦理挑战，特别是要考虑在全球伦理的框架下让人工智能造福社会。因此，不仅有必要强化人工智能从业人员的道德敏感性，加强人工智能在应用场景中的责任和规范研究，还必须继续推动人工智能伦理的国际对话和跨国问责机制的构建。

作者简介 王国豫

复旦大学哲学学院教授，复旦大学应用伦理学中心、复旦大学生命医学伦理研究中心主任，博士生导师。教育部长江学者特聘教授。兼任中国自然辩证法研究会科学技术与工程伦理专业委员会副主任，中国计算机学会职业伦理与学术道德委员会共同主席，国际人类表型组计划伦理委员会主任。王国豫长期从事高科技伦理与治理研究，是国家社会科学基金重大项目“高科技伦理问题研究”首席专家，国家重点研发计划重点专项“精准医疗的伦理政策法规框架研究”首席科学家。

推动形成“技术+规则”的治理综合解决方案

王迎春

近年来，在全球范围内，相关组织和机构发布了大量人工智能治理原则和倡议，推动人工智能“以人为本”和“负责任”发展已成为国际共识，如何将这些共识性原则转化为行动方案已是大势所趋。

2020年7月10日，世界人工智能大会治理论坛在上海举办，本届治理论坛以“发展负责任的人工智能”为主题，探讨“技术+规则”双轮驱动的治理综合解决方案，推动人工智能治理原则落地落实。论坛上，发布了“协同落实人工智能治理原则的行动建议”（上海AI治理协同行动9条），提出了“一平台、四工作、四体系”的落实人工智能治理原则的系统化的行动框架建议。

“一平台”即搭建全球合作网络和交流平台，形成全球人工智能治理研究和协作共同体，汲取多元文化智慧，构建多边、多学科、多主体参与的协商机制，探索可通约、非排他的安全保障和利益分享方案，推动形成共商共建共享的全球人工智能治理体系。

“四工作”包括标准规范、行业自律、最佳实践和可信技术。各利益相关方协同研究制定技术标准和应用规范，明确行业准入门槛，保证相关方的合法权益；建立科学研究和企业产品的伦理承诺和审查制度，形成行业内部的合规自律流程和操作指南；不断总结优质实践案例和相关经验，从实践中提炼规范，通过最佳实践引导各方行为；发展算法透明、保护隐私的可信技术，研发合伦理评估等监管技术，通过公共部门和私人部门的合作促进可信方案的推广。

“四体系”包括评估体系、监管体系、人才体系和保障体系。构建合规的指标体系、检测方法和评测平台，开展治理的绩效评估与分级认证；完善公开透明的多部门协同的监管体系，实现对相关产品和应用的全流程监管；推动交叉学科研究，开设治理必修课程，培养复合型人才；建设应对人工智能冲击就业结构的社会保障体系，开发相关保险产品，提升公众的数字化素养。

实现人工智能“以人为本”的美好愿景，需要技术与规则相融合的治理综合解决方案，我们建议全球同行一起携手，围绕典型场景开发与社会价值更加适宜的产品和方案，在积极稳妥的实践中不断达成共识并解决具体问题。

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。