计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210600034

人工智能治理理论及系统的现状与趋势

朝乐门　尹显龙
数据工程与知识工程教育部重点实验室(中国人民大学)　北京 100872
中国人民大学信息资源管理学院　北京 100872

摘　要　人工智能(Artificial Intelligence,AI)治理是解决 AI 挑战的主要手段。AI 治理的主要目的是充分发挥人工智能带来的优势和有效降低人工智能导致的风险,并通过整合技术、法律、政策、标准、伦理、道德、安全、经济、社会等多个方面的影响因素,最终建设负责任的人工智能(Responsible Artificial Intelligence,RAI)。AI 治理可以从智能个体治理、智能群体治理以及人机合作与共生系统的治理等 3 个方面,分技术层、伦理层、社会及法律层等 3 个层面进行。AI 治理的主要关键技术有 4 种:可理解性人工智能、防御对抗性攻击技术、建模及仿真技术和实时审计技术。从谷歌、IBM 和微软等公司的 AI 治理实践来看,产业界主要关注的是 RAI 研发,在 AI 系统的可解释性、隐私保护和公平性检查等方面已出现一些专用组件工具。目前,AI 治理需要研究的科学问题有:软件定义的 AI 治理、AI 治理关键技术、大规模机器学习中的 AI 治理评价、基于联邦学习的 AI 治理、AI 治理的标准制定、增强人工智能与人在回路型 AI 训练等。

关键词:人工智能;可理解性人工智能;负责任人工智能;人工智能治理
中图法分类号　TP391

AI Governance and System:Current Situation and Trend

CHAO Le-men and YIN Xian-long
Key Laboratory of Data Engineering and Knowledge Engineering (Renmin University of China),Beijing 100872,China
School of Information Resource Management,Renmin University of China,Beijing 100872,China

Abstract　The main purpose of AI governance is to take advantage of AI and reduce the risk. AI governance also aims to build a responsible AI via embracing the influencing factors such as technology,law,policy,standard,ethics,morality,safety,economy,as well as society. AI governance has three aspects:individual intelligent governance,group intelligent governance,human-computer cooperation and symbiotic system governance,which can be divided into three levels:technical level,ethical level,social and legal level. There are four key technologies for AI governance,which are intelligible AI,defense against adversarial attacks,modeling and simulation,and real-time audit. The industry is mostly concerned about developing a responsible AI in that by studying the actual practice of AI governance from leading companies like Google,IBM and Microsoft. Furthermore,tools like interpretability,privacy protection and fairness check for AI systems are already in use. At present,the main research topics on AI governance includes software-defined AI governance,key technologies of AI governance,AI governance evaluation in large-scale machine learning,AI governance based on federated learning,standardization of AI governance,enhancement on artificial intelligence and human-in-the-loop AI training.

Keywords　Artificial intelligence,Explainable artificial intelligence,Responsible artificial intelligence,AI governance

1　引言

随着人工智能在公共医疗、信息与通信、环境保护、交通、立法与政策制定、经济与社会、教育、政府等多个领域的广泛应用[1],人们开始关注其带来的挑战、风险及负面影响,进而人工智能治理成为了当今社会的新关注点。目前,AI 带来的主要挑战表现在技术实现、法律规范、社会以及伦理等 4 个方面[2],如图 1 所示。其中,技术实现方面的主要挑战包括 AI 安全、系统及数据的质量与集成、性价比和专业化程度及技能;法律及规范方面的挑战主要表现在自动智能系统的治理、负责任和问责能力以及隐私与安全;社会方面的挑战主要表现在劳动力的取代及变革、AI 的社会接受度与信任,以及人机交互的转型;伦理领域的挑战主要体现在 AI 对人类行为提出新规则、AI 与人在价值判断上的兼容性、AI 带来的道德困境以及 AI 歧视。

解决上述 4 种挑战的主要手段是将治理理论及实践引入

到稿日期:2021-05-03　返修日期:2021-06-28
通信作者:朝乐门(chaolemen@ruc.edu.cn)

人工智能的研发和应用。从治理对象看，人工智能的治理分为两大类：1)人机合作场景下的人工智能的治理，其难点在于权衡人与计算机的分工协作；2)自治智能群体场景下的人工智能治理，其难点在于控制机器与机器之间的协同计算。

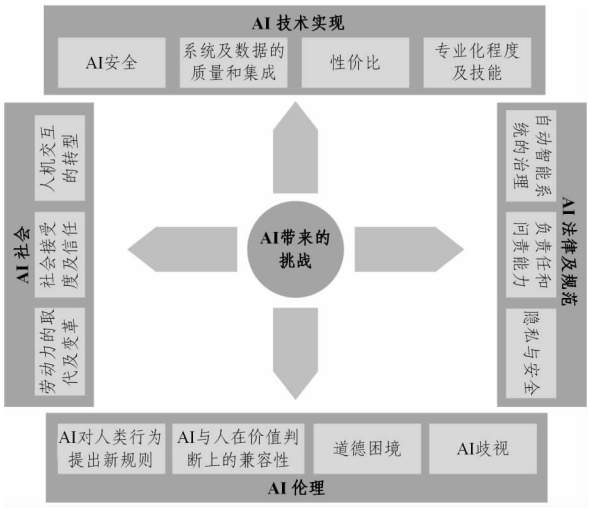

图 1 AI 的四大挑战模型[2]

Fig. 1 Four main challenges model for AI[2]

本文第 2 节主要讨论人工智能治理(AI Governance，简称 AI 治理)的内涵，包含其定义及治理内容；第 3 节采用分层思想分析 AI 治理框架，为 AI 治理的理论研究与实践应用提供方法论指导；第 4 节探讨 AI 治理的关键技术，提出 AI 治理所需的关键技术；第 5 节调查谷歌、IBM 和微软的 AI 治理实践，分析 AI 治理系统的研发与应用现状；最后总结全文，提出 AI 治理的未来研究应优先关注的科学问题。

2 人工智能治理的内涵

2.1 人工智能治理的定义

目前，人工智能治理的定义方法有多个，尚未达成共识，但可以分为三大类。

(1)能力说：认为 AI 治理是一种能力，比较典型的是 IBM 的定义方法——AI 治理是指导、管理和监视组织机构的 AI 活动能力，主要包括跟踪、记录和审计等活动的过程[3]。IBM 对 AI 治理的定义强调的是一种以模型跟踪和记录为基础，以实现人工智能的透明、可信和合规为目标的治理能力。

(2)过程说：认为 AI 治理是一种过程，比较有代表性的是 BasisAI 的定义方法——AI 治理是以可解释、透明和遵从伦理的方式，指导 AI 的设计、发展和部署的一种框架和过程，主要包括行动指南(原则)和系统的过程[4]。BasisAI 的 AI 治理目的是加快 AI 的采纳和确保 AI 的使用能够负责任。与 IBM 的定义方法不同的是，BasisAI 的定义强调 AI 治理的可解释性、透明性和遵从伦理规范。

(3)方式/方法说：认为 AI 治理是一种方式或方法。人工智能的政策类文件(如 BIC，APPGAI 2017a)通常主张 AI 治理是促进 AI 的优势及减轻 AI 风险的一种方法[5]。方式/方法说与能力说和过程说的不同点在于，其强调的是 AI 治理的手段和工具。

但是，无论能力说、过程说还是方式/方法说，目前人们对 AI 治理的责任、目标和内涵的认识是相对一致的——AI 治理的责任为充分发挥人工智能带来的优势和有效降低人工智能导致的风险，通过整合技术、法律、政策、标准、道德、伦理、安全、经济、环境、教育和社会等各方面的影响因素，最终建设成人类可以信任的 AI——负责任的人工智能，如图 2 所示。其中，计算机科学是 AI 治理的主要手段和工具，主要贡献表现在两个方面：1)确保人工智能及其算法的可解释性，即可理解性人工智能的实现；2)确保人工智能应用的安全与可靠，即 RAI 的实现。

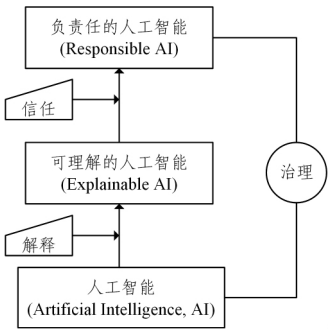

图 2 人工智能、可理解性人工智能与负责任的人工智能的关系

Fig. 2 Relationship among Artificial Intelligence, Explainable AI and Responsible AI

目前，很多组织机构相继提出了自己的 RAI 的原则或思想，比较有代表性的是波士顿咨询公司(Boston Consulting Group，BCG)提出的 RAI 的“6+1”原则(见图 3)[6]。其中，“6”代表的是问责、公平与平等、数据及隐私治理、社会与环境影响、安全与鲁棒性、透明及可理解性；“1”代表的是将人(Human)放在整个框架的中心位置，进而打通上述 6 个原则。再如，H20[7]认为负责任的 AI 思想至少包括可理解性 AI、可解释性机器学习、AI 伦理、AI 安全、以人为中心的 AI 以及与相关法律要求的合规性等内容。

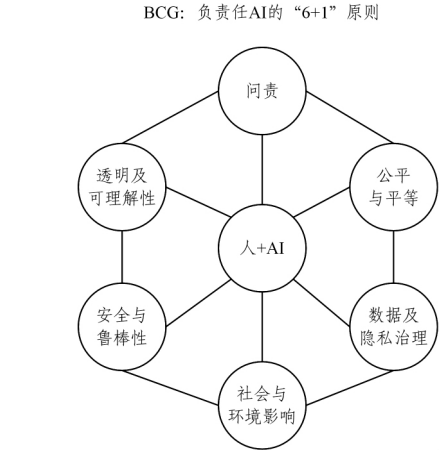

图 3 BCG 负责任 AI 的“6+1”原则[6]

Fig. 3 BCG “6+1” principles of responsible AI[6]

AI 治理可分为通用 AI 治理和面向领域的 AI 治理两种。其中，面向领域的 AI 治理的含义与所依赖的领域有着密切联系。例如，文献[8]提出人工智能在自然保护领域的错误应用将导致不良影响，并建议通过改进算法度量和伦理监管降

低其负面影响，进而建立面向自然保护的负责任的人工智能。

2.2 要素

目前，学术界对 AI 治理研究内容的描述差异较大。例如，Dafoe(2018)认为 AI 治理至少涉及 3 类要素[9]：1)技术类，主要关注如何理解人工智能的技术输入、机会和限制，为政策类和愿景类治理提供基础；2)政策类，主要关注企业、政府、事业、研究者和其他干系人之间的竞争与合作关系，进而满足他们各自的诉求；3)愿景类，主要关注的是 AI 治理的理想状态、结构及布局，通过基础设施、法律和标准的建设，优化人工智能的治理。Kuziemski 等认为，AI 治理的主要内容有 3 个[10]：1)激励以合规为中心的人工智能创新；2)AI 对社会赋能；3)增强 AI 治理结构的互操作性。

人们对 AI 治理研究的内容存在不同认识的主要原因在于其讨论的层次和视角不同。为了更好地揭示 AI 治理的本质内容，我们可以将 AI 治理分为 3 个不同的层次进行讨论，如图 4 所示。

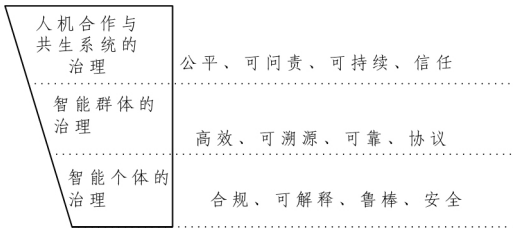

图 4 数据治理的内容

Fig. 4 Content of AI Governance

(1)智能个体的治理。其主要关注的是如何通过 AI 治理，确保智能个体自主决策活动或行为的合规、可解释、鲁棒和安全性。

(2)智能群体的治理。智能群体的治理是 AI 治理的一个重要内容，通常建立在智能个体的治理之上，并为人机合作与共生系统的治理提供支持。智能群体的治理重点在于保障其高效、可溯源、可靠和自治协议。智能群体治理是需要以联邦学习(Federated Learning)[11]为代表的多边缘智能体能够在无须全局共享数据的前提下，可共同训练新算法的技术。

(3)人机合作与共生系统的治理。其主要治理对象为根据特定应用场景，明确人与机器之间的主导地位，尤其是 AI 在工作流程中能够取代人的程度[12]。增强人工智能(Augmented AI)是目前人机合作与共生系统治理的一个重要关注点。人机合作与共生系统的治理重点在于确保其公平、可问责、可持续和可信任。

3 人工智能治理的框架

目前，AI 治理研究普遍采用分层分析方法，不同观点间的主要区别在于所分出的层次数量及命名方法不同。例如，Wirtz 等认为监管理论(regulation theory)是 AI 治理框架的基础，并采用分层设计方法提出了 AI 集成治理框架[13]；Gasser 等提出了 AI 治理的层次模型，该模型将 AI 治理分为 3 个主要层次，即技术层、伦理层以及社会及法律层[14]，如图 5 所示。

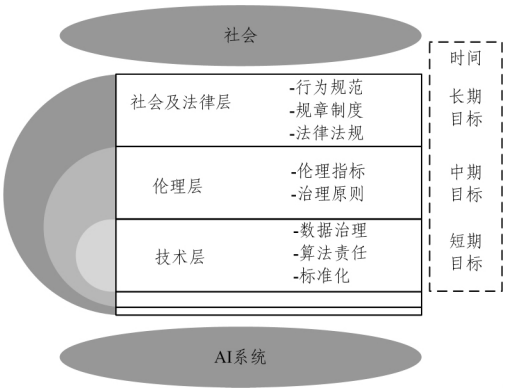

图 5 AI 治理的分层模型[14]

Fig. 5 Layered model for AI Governance[14]

3.1 技术层

技术层是 AI 治理的基础，也是 AI 治理的短期目标所关注的层次，主要内容包括：数据治理、算法责任和相关标准制定。数据和算法是现阶段人工智能的主要技术手段，因此 AI 治理应重视数据治理和算法责任。同时，标准化是实现数据治理和算法责任的重要手段。

数据治理和 AI 治理虽然是两个不同的概念，但在具体应用中二者是密不可分的，需要统筹和集成。传统的“数据治理”关注数据安全性、主数据管理、数据质量、数据体系结构和元数据管理等活动，然而“数据＋AI 治理”需要在“数据治理”的基础上至少新增两个功能模块：机器学习模型管理和负责任 AI 治理[15]。

3.2 伦理层

伦理层位于技术层和社会及法律层之间，是 AI 治理的中期目标所关注的层次，涉及 AI 治理的伦理指标和原则。基于知情与同意的隐私保护技术是伦理层上 AI 治理的重要策略。相比技术层和社会及法律层，伦理层的 AI 治理需求更为复杂，不确定性较高，通常无法直接使用精确的计算机协议或社会法律法规，需要结合不同目标群体的需求进行动态调整。

Lei 等提出了衔接 AI 治理原则以及技术的隐私保护框架，该框架利用监督机制、匿名机制、分区机制为用户在虚拟社区中产生的隐私资源提供隐私保护，并且明确了在隐私资源流通的不同环节(感知、存储、传输和处理)中各参与者(隐私内容产生者、虚拟社区、访问者)能够获取的隐私权[16]。

Schiff 等在分析超过 80 篇来自政府和非政府组织制定的 AI 文件的基础上，提出 AI 伦理、政策和治理相关文件的成功与否主要取决于 5 个因素[17]：与特定法律和政策的结合度、内容描述的具体化程度、对外共享及曝光度、执行过程的强制性程度，以及有效监督和快速迭代及持续更新。目前，伦理层的 AI 治理相关指南(或规定)中存在的主要问题在于其术语内涵模糊、描述过于抽象以及可操作性差，需要采用制定可操作性强的标准和法律规范、引入伦理专家委员会和跨学科领域的培训等手段来弥补上述局限[18]。

3.3 社会及法律层

社会及法律层建立在伦理层之上，主要关注的是 AI 治理的长远目标，具体涉及 AI 治理的行为规范、规章制度和法

律法规。社会及法律层的 AI 治理难点在于平衡 AI 与人之间的主从、分工与合作，以及权衡 AI 干系人之间的价值观与利益冲突。

相关政策的制定和实施是在社会及法律层面开展 AI 治理的重要手段。Theodorou 等认为，目前已经产生了许多关于 AI 的高级伦理准则，需要结合现有伦理、法律和文化价值观制定具体政策[18]。Writz 等在提出的 AI 治理框架中强调了公共政策层是重要的组成部分[13]。目前，社会及法律层中存在的一种曲解是将算法、AI 和机器人技术的法律法规制定工作相互分离[19]，因此未来亟待对其进行集成与融合。

图 2 给出了通用的 AI 治理框架，可为面向特定领域 AI 治理提供基础。近年来，面向特定领域的 AI 治理也受到了一定程度的重视。例如，文献[20]提出了一种面向健康护理领域的 AI 治理框架——GMAIH(Governance Model for AI in Healthcare)，主要强调 AI 治理中的公平性、透明度、可信度和问责能力。

此外，AI 治理可分为组织机构、国家、国际和全球 4 个不同层次，每个层次都有自己的特殊性。例如，文献[21]探讨了 G20 国家的 AI 治理现状；文献[22]对比分析了国际 AI 治理的两种架构模式——集中治理和分散治理，分析了其优缺点，并提出分散治理将会是国际 AI 治理的主流趋势；文献[23]研究了全球 AI 治理问题，并建议将国际标准作为全球 AI 治理的工具。

4 人工智能治理的技术

AI 治理的相关技术较多，但就现阶段的 AI 治理而言，具有重要意义的关键技术有以下 4 种。

4.1 可理解性人工智能

与其他技术系统(如互联网等)不同的是，AI 治理必须以可解释性为前提。普华永道(PwC)的一项全球 CEO 调查(Global CEO Survey)显示，84%的 CEO 认为基于 AI 的决策需要解释才能被信任[24]。因此，可解释性或可理解性人工智能是 AI 治理的关键技术之一。

Gunning 等认为可理解性人工智能(Explainable Artificial Intelligence，XAI)是一种能够向人类用户解释自己的逻辑依据，能够描述自身优势和劣势以及能够传达自己未来如何行动的系统[25]，其主要强调了对 3 个基本问题的解释能力：逻辑依据、优缺点和未来行为。XAI 系统应遵循 3 个基本原则：1)具备能够自我解释的能力和可理解性；2)能够解释已经、正在以及未来所做的事；3)能够透露正在运行的工作中相关的重要信息[26]。与传统的人工智能相比，可理解性人工智能具有面向特定的受众提供必要的细节信息和上述 XAI 应遵循的 3 个基本原则等特征。

AI 治理涉及可理解性人工智能的 4 个要素[27]：能够了解如何得出特定答案、能够为模型所提供的答案提供正当理由、能够为人的决策活动提供新信息以及能够量化预测结果的可靠程度。

目前，可理解性人工智能的研究主要包括：解释方法、解释用户界面、解释评价和解释心理机理等科学问题。以 DARPA 的 XAI 项目[24]为例，其研究分为 3 个主要领域：1)可解释性模型的学习方法；2)更有效的解释界面的设计；3)有效解释的心理需求。

4.2 防御对抗性攻击技术

AI 为传统信息安全技术提出了新挑战，如何应对 AI 的对抗性攻击(adversarial attack)是 AI 治理需要解决的关键性安全技术。因此，AI 治理需要引入新的防护、检测和响应机制及技术，进而确保 AI 模型和决策的安全。

针对 AI 的对抗性攻击主要是利用 AI 技术及系统的鲁棒性差的弱点，分析 AI 模型及其实现中的漏洞，进行违背于 AI 设计者初衷的行为与活动。目前，防御 AI 对抗性攻击的主要技术有对抗性训练(adversarial training)、梯度隐藏(gradient hiding)、防御性蒸馏(defensive distillation)、特征压缩(feature squeezing)、可迁移性的封锁、Defense-GAN、MagNet、高层特征降噪器(High-level Representation Guided Denoiser，HGD)和基函数变换(Basis Function Transformations)等[28]。

4.3 建模及仿真技术

人工智能的 FAT(Fairness，Accountability and Transparency)，即如何保障人工智能应用的公平、可问责和透明是人工智能治理的一个难点[29]。AI 治理应引入建模及仿真(Modeling and Simulation，M&S)[30]技术，以模型为仿真的基础，生成和优化用于智能体决策的数据集与算法。

值得一提的是，面向 AI 治理的建模及仿真活动需要将人放入目标系统，并重视模拟和仿真过程中人的交互作用。因此，人在回路型仿真(Human-In-The-Loop simulation，HITL 仿真)的技术[31]将在 AI 治理中得到广泛应用。

4.4 实时审计技术

AI 治理的被审计对象为 AI 应用中的公平性、是否存在偏见以及责任主体。虽然目前对 AI 审计的研究不多，但是也有一些新的积极探索。以开源工具包 Aequitas 为例[32]，该工具包提出了一种机器学习算法中的偏见与不公平性的度量方法，并提供 Python 接口。此外，DARPA 正在从用户满意度、心理模型、任务性能和适当信任 4 个维度研究 XAI 的理解效果评价问题[25]，并提出了评价框架(evaluation framework)。该评价框架主要涉及用户满意度、对原始模型的忠实度、解释信度、效度及效率和解释的可扩展性。

实时的数据溯源和证据保留是 AI 治理中实现可问责性的重要前提。AI 治理中的自动审计依赖于其数据溯源以及证据保留的能力。

5 人工智能治理的系统

目前，AI 治理相关的应用越来越多，比较典型的是加拿大银行 AI 治理[33]、芬兰财政部 AuroraAI[34]、UCARE.AI[35]、智能迪拜[36]、TAIGER AI 治理[37]、西班牙电信的 AI 原则[38]等。从系统及解决方案的提供商来看，比较典型的有谷歌公司(Google Inc.，简称谷歌)、IBM 公司(International Business Machines Corporation，简称 IBM)和微软公司(Microsoft，简称微软)。AI 治理典型案例的分析如表 1 所列。

表1 AI治理典型案例的分析

Table 1 Typical case analysis for AI Governance

	谷歌	IBM	微软
AI治理目标	实现Responsible AI	实现Trusted AI	实现Responsible AI
AI(治理)原则	1)对社会有利 2)避免产生或加大不公偏见 3)支持安全 4)向人类负责 5)纳入隐私设计 6)支持科学的高标准 7)限制有害或滥用	1)具有可解释性 2)对决策公平 3)应用无害 4)透明化设计 5)隐私安全	1)具有公平性 2)安全和可靠 3)安全且尊重隐私 4)具有包容性 5)系统透明化 6)对AI系统负责
AI治理关键技术	1)大规模机器学习中的公平性保障 2)融入伦理原则的机器学习模型 3)可解释的机器学习系统的设计原则	1)公平性和可解释性AI技术 2)AI模型的标准化文档	1)Fairlearn和AI公平性检查表 2)InterpretML和数据集的数据表 3)Microsoft SEAL
AI治理工具与系统	1)公平指数(Fairness Indicators) 2)Explainable AI 3)TensorFlow Privacy 4)Model Cards工具集	1)IBM Cloud Pak ® for Data 2)AI Fairness 360 3)AI Explainability 360 4)AI Adversarial Robustness 360 5)AI FactSheets 360	1)Azure ML 2)Fairlearn 3)InterpretML 4)错误分析工具(Error Analysis) 5)SmartNoise 6)Microsoft SEAL 7)Presidio
AI治理在企业内外的应用	COVID-19、名人识别API、谷歌翻译	KMPG毕马威、Regions Bank	对话机器人

5.1 谷歌AI治理及系统

谷歌AI治理的目标是实现负责任的AI。谷歌曾提出[39],在伦理和法律监督的前提下,AI能够使社会经济、决策支持更加公平和安全,通过AI治理可以实现能够充分发挥潜能的负责任AI。

谷歌提出了AI治理的7项原则[40]:1)AI的社会向善,并且其正面影响应该远大于可能出现的负面影响;2)避免肤色、性别、政治立场以及伦理信仰等方面产生偏见;3)部署前进行充分测试,部署后持续监控,保证AI系统的安全性;4)AI系统应该始终在人类的控制范围内运行;5)保护用户的隐私数据;6)支持AI跨领域合作,并帮助相应领域突破技术瓶颈;7)限制AI可能带来的危害并避免其滥用。

谷歌提供的AI治理解决方案中涉及3个关键技术,即大规模机器学习中的公平性保障、融入伦理原则的机器学习模型和可解释性机器学习系统的设计原则。谷歌研发团队于2019年提出的一项用于减少偏见的正则化技术——MinDiff框架[41],支持大规模机器学习中的公平性保障。考虑到机器学习系统易违反伦理原则,谷歌主张伦理原则可以通过约束模型融入到机器学习模型中,其优势在于约束模型只对相关的输入做出反应[42]。哈佛和谷歌研究团队通过分析决策集的不同复杂性对模型可解释性的影响,提出可用于指导可解释机器学习系统开发的通用原则[43]。

谷歌提供了可全面部署和训练模型的系统并融入了AI治理原则的端到端开源机器学习平台——TensorFlow。此外,他们还提供了一些支持人工智能治理的部分功能的工具集,如用于公平性原则的公平指数(fairness indicators)、用于可理解性人工智能的Explainable AI、用于保护隐私数据的TensorFlow Privacy,以及用于报告机器学习模型的来源、用途和伦理评估的Models Cards等工具系统。

目前,谷歌的AI治理解决方案在企业内外都有初步应用。在企业内,通过公平性原则减少谷歌翻译的性别偏见[44];在企业外,联合哈佛全球健康研究所改善新冠疫情的公共预测能力,并为美国新冠疫情的预测公平性做出诊断分析[45]。

5.2 IBM AI治理及系统

IBM的AI治理目标在于构建一种对所有人有益的、透明和可信的AI解决方案,使用户可以充分了解、信任和合理使用AI技术,进而实现可信任的AI(Trusted AI)。

IBM给出了AI治理的5项原则[46]:1)对AI产生的决策结果应该能够解释其原因,并且解释能力需达到能够被非专业用户理解的程度;2)通过应对偏见和增加包容性,帮助用户得出更加公平的决策;3)AI系统需要保证其输出结果的安全性;4)确保AI模型和服务生成的整个流程足够透明;5)对AI系统用户的隐私和数据权益提供保护,并向用户声明所使用的保护方式。

目前,IBM的AI治理技术主要集中在AI的公平性和可解释性以及AI模型的标准化文档[47],并提供了开源工具包AI Fairness 360,AI Explainability 360以及AI Factsheet 360。AI Fairness 360于2018年9月发布,作为一个开源的工具箱,不仅集成了多个降低偏见的算法,还提供了用于测试偏见的数据集以及公平性指标[48];AI Explainability 360于2019年8月发布,该工具包主要用于支持机器学习模型的可解释性并集成了可用于解释机器学习模型的算法,其中包括了由IBM自主研发的两种算法,即通过列生成的布尔分类规则和对比性解释法[49];AI Factsheets 360是由IBM于2020年9月发布的构建AI标准化文档的网站,该网站提供了模板文档以及构建文档所需的主要步骤及活动[50]。

IBM还提供了其他工具包和平台,如针对AI的对抗性攻击,IBM提供了专用工具AI Adversarial Robustness 360以及一种基于云的AI开发平台——IBM Cloud Pak © for Data,其中内置了支持AI构建、管理和监控的组件,如Watson Knowledge Catalog,Watson OpenScale和Watson Studio等。

在企业外,IBM为毕马威(KMPG)提供IBM Cloud Pak © for Data和Watson OpenScale服务,用于解决AI面向用户的透明计算问题[51]。

5.3 微软 AI 治理及系统

微软的 AI 治理目标是实现以人为本，能够造福所有人且负责任的 AI(Responsible AI)。微软设立了负责任 AI 办公室(Office of Responsible AI, ORA)、以太委员会(Aether Committee)和面向工程的负责任 AI 策略(Responsible AI Strategy in Engineering, RAISE)等 3 个团队，共同推进其 AI 治理目标的实现。

微软提出了 AI 治理的 6 项原则[52]：1)应该具有公平性，平等对待所有人；2)应该保证 AI 的可靠性以及其产生的结果为无害；3)尊重用户的隐私以及保证 AI 系统的安全性；4)应该赋予每个人参与权；5)将系统透明化，使非专业人士也能够理解 AI；6)使 AI 系统受控于人的监管下，使人能够对系统负责。

微软的 AI 治理关键技术主要涉及 AI 系统的公平性、可解释性以及隐私安全。微软提供 Fairlearn 开源包，支持减少因模型导致的性别、肤色、年龄等特征的偏见及其负面影响；与 48 位从业者联合设计出一个 AI 公平性检查表(AI Fairness Checklist)，帮助 AI 研发团队在开发周期的每个阶段进行检查，并在 AI 系统部署之前预测公平性问题[53]；提供 InterpretML 开源包，用于理解模型的行为以及对预测结果进行解释；研发数据集数据表(DataSheet for DataSet)，帮助用户在模型的训练前检查数据集是否符合需求[54]；提出 Microsoft SEAL 开源动态加密技术，防止用户数据被非法获取[52]。

此外，微软研发的 Azure ML 支持构建和部署负责任的机器学习(Responsible Machine Learning)服务，同时也提供了 Fairlearn, InterpretML, Error Analysis, SmartNoise, Microsoft Seal 和 Presidio 等工具箱和开源包。目前，微软在企业内将其 AI 治理解决方案应用于对话机器人的开发，帮助用户利用 AI 治理的原则开发一个负责任的对话机器人[55]。

结束语 AI 治理研究将会超出人工智能技术本身的学科范畴，需要来自计算机科学、数据科学、法学、伦理学及具体应用学科等多个领域专家的共同参与。其中，计算机科学为 AI 治理提供了技术和工具手段，尤其是对 AI 治理中的可解释性、透明、安全性和可靠性等核心问题的解决方案提供了主要理论依据。从计算机科学的角度看，除了 AI 技术本身的研究外，AI 治理的未来研究应优先关注的科学问题有：

(1)软件定义与面向特定应用场景的 AI 治理解决方案；

(2)AI 治理关键技术的突破，包括可解释性人工智能、防御对抗性攻击技术、建模及防战技术和实时审计技术；

(3)大规模机器学习中的隐私保护、偏见检测与歧视预测、治理效果评估等专用工具及其一体化平台的研发；

(4)基于联邦学习的 AI 治理的安全保障和可信计算技术；

(5)面向 AI 治理的技术与数据标准以及互操作协议的制定；

(6)增强人工智能以及人在回路型(Human-In-The-Loop)AI 训练方法。

AI 治理和基于 AI 的治理是两个不同的科学问题，前者的治理对象是 AI 智能体，而后者的治理对象不仅限于 AI 本身，还涉及经济、社会、政治、文化、环境、科技和教育等多个领域。本文的讨论边界是 AI 治理，并未讨论基于 AI 的治理。相比 AI 治理的研究，基于 AI 的治理的研究需要讨论的问题更为广泛。例如，微软的 Microsoft AI for Earth 主要针对的是基于 AI 的环境治理。因此，我们可以进一步关注基于 AI 的治理，将 AI 治理放入基于 AI 治理的框架内进行深入研究。

参 考 文 献

[1] SHARMA G D, YADAV A, CHOPRA R. Artificial intelligence and effective governance: A review, critique and research agenda [J]. Sustainable Futures, 2020, 2: 100004.

[2] WIRTZ B W, WEYERER J C, GEYER C. Artificial Intelligence and the Public Sector—Applications and Challenges[J]. International Journal of Public Administration, 2019, 42(7): 596-615.

[3] AI governance: Ensuring your AI is transparent, compliant, and trustworthy [EB/OL]. [2021-05-15]. https://www.ibm.com/analytics/common/smartpapers/ai-governance-smartpaper/#ai-governance-delivers.

[4] AI Governance: The path to responsibleadoption of artificial intelligence [R/OL]. [2021-05-15]. https://www.asianscientist.com/wp-content/uploads/2020/07/AI-Governance-Whitepaper-Basis-AI.pdf.

[5] ULNICANE I, KNIGHT W, LEACH T, et al. Framing Governance for a Contested Emerging Technology: Insights from AI Policy[J]. Policy and Society, 2020, 40(2): 1-20.

[6] DURANTON S, MILLS S. Responsible AI: Leading by Example [EB/OL]. (2021-02-03) [2021-05-15]. https://medium.com/bcggamma/responsible-ai-leading-by-example-c25a8a0a98ea.

[7] Responsible Machine Learning [EB/OL]. [2021-05-15]. https://www.h2o.ai/responsible-ai/.

[8] WEARN O R, FREEMAN R, JACOBY D M. Responsible AI for conservation[J]. Nature Machine Intelligence, 2019, 1(2): 72-73.

[9] DAFOE A. AI governance: a research agenda[EB/OL]. (2018-08-27)[2021-05-15]. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf.

[10] KUZIEMSKI M, PALKA P. AI governance post-GDPR: lessons learned and the road ahead[J/OL]. 2019. http://diana-n.iue.it:8080/handle/1814/64146.

[11] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: Challenges, methods, and future directions[J]. IEEE Signal Processing Magazine, 2020, 37(3): 50-60.

[12] GHALLAB M. Responsible AI: requirements and challenges [J]. AI Perspectives, 2019, 1(1): 1-7.

[13] WIRTZ B W, WEYERER J C, STURM B J. The dark sides of artificial intelligence: An integrated AI governance framework for public administration[J]. International Journal of Public Administration, 2020, 43(9): 818-829.

[14] GASSER U, ALMEIDA V A. A layered model for AI governance[J]. IEEE Internet Computing, 2017, 21(6): 58-62.

[15] ADLER S. From Data Governance to AI Governance: How to successfully make the shift? [EB/OL]. [2021-05-15]. https://

www. aidataanalytics. network/data-science-ai/whitepapers/fromdata--governance-to-ai-governance-how-to-successfully-make-the-shift.

[16] LEI Y，DUAN Y，SONG M. Technical Implementation Framework of AI Governance Policies for Cross-Modal Privacy Protection[C]// International Conference on Collaborative Computing：Networking，Applications and Worksharing. Springer，Cham，2020：431-443.

[17] SCHIFF D，BIDDLE J，BORENSTEIN J，et al. What's Next for AI Ethics，Policy，and Governance? A Global Overview[C]// Proceedings of the AAAI/ACM Conference on AI，Ethics，and Society. 2020.

[18] THEODOROU A，DIGNUM V. Towards ethical and socio-legal governance in AI[J]. Nature Machine Intelligence，2020，2(1)：10-12.

[19] WACHTER S，MITTELSTADT B，FLORIDI L. Transparent，explainable，and accountable AI for robotics[J]. Science (Robotics)，2017，2(6)：1-5.

[20] REDDY S，ALLAN S，COGHLAN S，et al. A governance model for the application of AI in health care[J]. Journal of the American Medical Informatics Association，2020，27(3)：491-497.

[21] POMARES J，ABDALA M B. The future of AI governance [J/OL]. [2021-05-15]. https://www. global-solutions-initiative. org/wp-content/uploads/2020/04/GSJ5_Pomares_Abdala. pdf.

[22] CIHON P，MAAS M M，KEMP L. Fragmentation and the Future：Investigating Architectures for International AI Governance[J]. Global Policy，2020，11(5)：545-556.

[23] CIHON P. Standards for AI governance：international standards to enable global coordination in AI research & development [R]. Future of Humanity Institute University of Oxford. 2019：1-41.

[24] A practical guide to Responsible Artificial Intelligence (AI) [R/OL]. [2021-05-15]. https://www. pwc. com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai/responsible-ai-practical-guide. pdf.

[25] GUNNING D，AHA D. DARPA's explainable artificial intelligence (XAI) program[J]. AI Magazine，2019，40(2)：44-58.

[26] GUNNING D，STEFIK M，CHOI J，et al. XAI—Explainable artificial intelligence[J]. Science Robotics，2019，4(37)：1-2.

[27] JIMÉNEZ-LUNA J，GRISONI F，SCHNEIDER G. Drug discovery with explainable artificial intelligence[J]. Nature Machine Intelligence，2020，2(10)：573-584.

[28] CHAKRABORTY A，ALAM M，DEY V，et al. Adversarial attacks anddefences：A survey[J]. arXiv：181000069，2018.

[29] YEUNG K，HOWES A，POGREBNA G. AI governance by human rights-centred design，deliberation and oversight：An end to ethics washing [M]. The Oxford Handbook of AI Ethics，Oxford University Press，2019：1-27.

[30] ZEIGLER B P，MUZY A，KOFMAN E. Theory of modeling and simulation：discrete event & iterative system computational foundations [M]. Academic Press，2018.

[31] ROTHROCK L，NARAYANAN S. Human-in-the-loop simulations [M]. Springer，2011.

[32] SALEIRO P，KUESTER B，HINKSON L，et al. Aequitas：A bias and fairness audit toolkit[J]. arXiv：181105577，2018.

[33] TORRIE V. AI Governance in Canadian Banking：Fairness，Credit Models，and Equality Rights [J]. Credit Models，and Equality Rights，2020，36(1)：5-38.

[34] Implementation of the nationalAuroraAI programme [EB/OL]. [2021-05-15]. https：//vm. fi/en/auroraai-en.

[35] REMOLINA N，SEAH J. How to Address the AI Governance Discussion? What Can We LearnFrom Singapore's AI Strategy? [J]. SMU Centre for AI & Data Governance Research Paper，2019(8)：1-18.

[36] SALEM F. A Smart City for public value：Digital transformation through agile governance-the case of 'Smart Dubai'[J]. World Government Summit Publications，Forthcoming，2020(5)：1-70.

[37] LEE D. TAIGER featured as an exemplary model for AI Ethics and Governance practices by IMDA and PDPC [EB/OL] (2020-10-19) [2021-05-15]. https：//taiger. com/articles/taiger-featured-as-an-exemplary-model-for-ai-ethics-and-governance-practices-by-imda-and-pdpc/.

[38] AI PRINCIPLES OF TELEFÓNICA [EB/OL]. [2021-05-15]. https：//www. telefonica. com/en/web/responsible -business/our-commitments/ai-principles.

[39] Perspectives on Issues in AI Governance [R/OL]. [2021-05-15]. https：//ai. google/static/documents/perspectives -on-issues-in-ai-governance. pdf.

[40] AI Principles 2020 Progress update [R/OL]. [2021-05-15]. https：//ai. google/static/documents/ai-principles-2020-progress-update. pdf.

[41] PROST F，QIAN H，CHEN Q，et al. Toward a better trade-off between performance and fairness with kernel-based distribution matching[J]. arXiv：191011779，2019.

[42] WANG S，GUPTA M. Deontological ethics by monotonicity shape constraints[C]// International Conference on Artificial Intelligence and Statistics. PMLR，2020：2043-2054.

[43] LAGE I，CHEN E，HE J，et al. Human evaluation of models built for interpretability[C]// Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. 2019.

[44] KUCZMARSKI J. Reducing gender bias in Google Translate [EB/OL]. (2018-12-6) [2021-05-15]. https：//blog. google/products/translate/reducing-gender-bias-google-translate/.

[45] PFISFER T. Google Cloud，Harvard Global Health Institute release improved COVID-19 Public Forecasts，share lessons learned [EB/OL]. (November 17，2020) [2021-05-15]. https：//cloud. google. com/blog/products/ai-machine-learning/google-and-harvard-improve-covid-19-forecasts.

[46] IBM，Artificial Intelligence [EB/OL]. [2021-05-15]. https：//www. ibm. com/artificial-intelligence/ai-ethics-focus-areas.

[47] TUCKER E，VAIDYANATHAN R. AI Governance：Drive compliance，efficiency and outcomes from your AI lifecycle [EB/OL]. (2020-05-26) [2021-05-15]. https：//www. ibm. com/blogs/journey-to-ai/2020/05/ai-governance-drive-compliance-efficiency-and-outcomes-from-your-ai-lifecycle/? mhsrc = ibmsearch_a&-mhq=AI-Governance.

[48] VARSHNEY K R. Introducing AI Fairness 360 [EB/OL]. (2018-09-19) [2021-05-15]. https：//www. ibm. com/blogs/re-

search/2018/09/ai-fairness-360/.

[49] MOJSILOVIC A. Introucing AI Explainability 360 [EB/OL]. (2019-08-08) [2021-05-15]. https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/?mhsrc=ibmsearch_a&mhq=IBM%20Explainability%20360.

[50] HIND M. IBMFactSheets Further Advances Trust in AI[OL]. (2020-07-09) [2021-05-15]. https://www.ibm.com/blogs/research/2020/07/aifactsheets/?mhsrc=ibmsearch_a&mhq=AI%20Factsheet%20360.

[51] SOKALSKI M. Artificial Intelligence in Control with WatsonOpenScale [EB/OL]. [2021-05-15]. https://www.kpmg.us/alliances/kpmg-ibm/ai-in-control-watson-openscale.html.

[52] Research Collection: Research Supporting Responsible AI [EB/OL]. (2020-04-13) [2021-05-15]. https://www.microsoft.com/en-us/research/blog/research-collection-research-supporting-responsible-ai/.

[53] MADAIO M A, STARK L, WORTMAN V J, et al. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai [C]// Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020.

[54] GEBRU T, MORGENSTERN J, VECCHIONE B, et al. Datasheets for datasets[J]. arXiv: 180309010, 2018.

[55] Responsible bots: 10 guidelines for developers of conversational AI [OL]. (2018-09) [2021-05-15]. https://www.microsoft.com/en-us/research/publication/responsible-bots/.

CHAO Le-men, born in 1979, Ph.D, associate professor, Ph.D supervisor. His main research interests include data science and big data analysis.

专栏特邀编审

邢春晓 博士，研究员，博士生导师，清华大学信息技术研究院、清华大学互联网产业研究院和清华大学新型城镇化研究院副院长。主要研究领域为数据库和数据仓库、大数据和知识工程、人工智能、软件工程、区块链技术、智慧城市、智慧医疗、数字图书馆和电子政务关键技术等。发表学术论文350余篇，其中SCI 40余篇、EI 200余篇，软件著作权23项，获得发明专利40项，教育部科技成果1项。作为主要负责人承担了国家973项目、国家自然基金重点项目、国家863重点项目和目标导向项目、国家高科技产业化CNGI项目、国家科技支撑计划项目等。

朝乐门 副教授，博士生导师。中国人民大学数据工程与知识工程教育部重点实验室研究员、校友工作办公室副处长（副主任）、数据科学50人、国家级一流本科课程《数据科学导论》负责人、中国计算机学会信息系统专业委员会委员、全国高校人工智能与大数据创新联盟专家委员会副主任、国际期刊《Data Science and Informetrics》副主编。主持完成国家自然科学基金、国家社会科学基金等重要科学研究项目10余项；参与完成核高基、973、863、国家自然科学基金重点项目等10余项。

张桂刚 中国科学院自动化研究所副研究员，硕士生导师。中国计算机学会信息系统专业委员会委员、副秘书长，中国自动化学会会员，中国人工智能学会会员。主要研究方向为人工智能、航空发动机/飞机智能健康管理。出版专著5本，授权发明专利20余项，在SCI/EI期刊上发表论文80余篇。主持国家重点研发计划课题、国家自然基金面上项目、北京自然基金项目、航空科学基金项目等20余项。

黄梦醒 博士，教授，博士生导师，海南大学信息学院院长，海南省杰出人才，海南省“515人才工程”第一层次人选，海南省南海名家，海南省“信息感知融合与智慧服务”人才团队带头人，国家重点研发计划项目首席科学家，中国人工智能学会智能空天系统专委会常委，中国计算机学会高级会员，中国计算机学会信息系统专委会委员。主要研究领域为大数据与智能信息处理、多源信息感知与融合、人工智能与智慧服务等。以第一作者和通信作者发表学术论文150余篇，其中在SCI、EI期刊上发表110余篇；获得国家授权发明专利16项；获得软件著作权92项；出版专著4部，译著2部；获得海南省科技进步奖3项。主持和承担国家重点研发计划项目、国家科技支撑计划及国家自然科学基金项目等国家及省部级项目20余项。