Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Review

中国人工智能的伦理原则及其治理技术发展

吴文峻 [a,*]，黄铁军 [b]，龚克 [c]

[a] State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China
[b] School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
[c] Chinese Institute of New Generation Artificial Intelligence Development Strategie, Nankai University, Tianjin 300071, China

ARTICLE INFO

Article history:
Received 19 September 2019
Revised 18 November 2019
Accepted 31 December 2019
Available online 8 January 2020

关键词
人工智能伦理
人工智能治理技术
机器学习
隐私
安全
公平

摘要

伦理原则和治理技术对于人工智能（AI）的健康和可持续发展至关重要。为了实现AI造福人类社会这一长期目标，中国政府、研究机构和企业已经发布了AI的伦理原则，并启动了研究AI治理技术的项目。本文对这些工作进行了综述，并着重介绍了中国在这一领域的初步成果。此外，本文总结了AI治理研究中所面临的主要挑战，并讨论了未来的研究方向。

1. 引言

随着新一代人工智能（AI）算法和产品的快速开发和部署，AI在各个领域发挥着越来越重要的作用，并对现代社会结构产生了重大影响。AI模型和算法已被广泛应用于各种决策场景，如刑事审判、交通控制、金融贷款和医疗诊断。这种基于AI的自动决策系统得到不断的推广和应用，正在从安全性和公平性等方面带来潜在的风险。

例如，当前自动驾驶系统的安全性引发了人们的关注。2015年，中国的一辆特斯拉汽车发生了致命的事故，原因是汽车的自动驾驶系统未能识别出路上的清扫车，从而导致其没有执行正确的操作。在智能司法场景中，司法人员根据囚犯行为特征，采用智能算法来决定是否授予其假释许可。有人指出，这种算法可能会基于种族和文化背景做出有偏见和不公平的决定。在金融场景中，基于AI的数字贷款算法可能会因为有偏见的判断而拒绝某些人的贷款申请。政府机构、学术界和工业界都意识到了AI应用的安全性和公平性等问题，而且对AI应用的治理也成为了越来越重要的议题，所以我们必须采取有效措施来减轻与AI相关的一些潜在风险。

如今，许多国家的政府部门、研究机构和企业已经公布了他们的AI伦理准则、原则和建议。为了在当前的AI系统和产品中实施这些准则、原则和建议，开发

* Corresponding author.
E-mail address: wwj@nlsde.buaa.edu.cn (W. Wu)

英文原文: Engineering 2020, 6(3): 302–309
引用本文: Wenjun Wu, Tiejun Huang, Ke Gong. Ethical Principles and Governance Technology Development of AI in China. *Engineering*, https://doi.org/10.1016/j.eng.2019.12.015

AI治理技术将至关重要，需要深入研究联邦学习、可解释AI、严格的AI安全测试和验证以及AI伦理评估等方法。这些技术仍在高速发展期，而且也不够成熟，所以无法被广泛应用于商业场景。主要的技术障碍来源于现代AI研究所面临的基本挑战，如人类层面的道德认知、常识性的道德推理和多学科AI伦理工程。本文旨在从中国的角度对AI伦理原则和正在进行的相关研究工作进行总体概述。

本文的其余部分安排如下：第二部分介绍了政府机构和组织发布的伦理原则，并重点介绍了中国研究人员在AI治理方面的主要研究成果；第三部分比较了中国与其他国家在AI伦理原则和治理技术发展方面的差异；第四部分讨论了AI治理研究中的巨大挑战，并提出了未来的研究方向。

2. 中国的 AI 伦理原则与新兴治理技术

2017年中国发布的《新一代人工智能发展战略规划》（Development Plan of the New Generation Artificial Intelligence）强调，必须认真管理好AI的技术属性和社会属性，以确保AI的可靠性。2019年，中国科技部成立了国家新一代人工智能治理专业委员会，并发布了《新一代人工智能治理原则——发展负责任的人工智能》[1]。北京智源人工智能研究院（Beijing Academy of Artificial Intelligence, BAAI）也发布了《人工智能北京共识》[2]，就推进AI的研发、使用、治理和长远规划提出倡议，以支持实现对人类和自然环境有益的AI。在参考文献[3]中，来自BAAI的研究人员收集了20多个关于AI伦理原则的提议，并对这些提议的文本内容进行了主题词分析。他们确定了提议中共同提到的关键词——隐私、保密、安全、透明、问责和公平。

（1）保密性和隐私性：AI系统应该是保密的且尊重个人隐私。

（2）安全性和可靠性：AI系统应能可靠并安全地运行。

表1 AI伦理原则和治理技术

AI ethical principle	AI governance technology
Security and privacy	Federated learning, blockchains
Safety and reliability	Machine learning test and verification
Transparency	Interpretable/explainable AI
Accountability	AI provenance, auditing, and forensic
Fairness	AI fairness evaluation and debiasing algorithm

（3）透明度：AI系统应该是可以理解的。

（4）问责制：AI系统应该有问责制。

（5）公平性：AI系统应该公平对待所有人。

这些通用原则得到了全球AI领域的研究人员、从业人员和监管人员的广泛认可。这些原则不仅反映了我们人类社会的友好和道德信念，而且也需要可行的和全面的技术框架及解决方案来实现AI模型、算法和产品中的伦理约束。表1列举了与AI伦理相关的新兴技术，这些技术在支撑AI的有效治理方面有很大的潜力。

2.1. 数据安全和隐私

数据安全是AI伦理原则中最基本、最常见的要求。各国政府正在制定保护数据安全和隐私的法律。例如，欧盟于2018年实施了《通用数据保护条例》（General Data Protection Regulation, GDPR），中国也于2017年颁布了《中华人民共和国网络安全法》（Cybersecurity Law of the People's Republic of China）。这些法规的建立旨在保护用户的个人隐私，并给如今被广泛使用的数据驱动型AI的开发带来新的挑战。

在数据驱动型AI的范式中，开发人员通常需要在中央存储库中收集大量用户数据，然后进行后续的数据处理，包括数据清洗、融合和标注，以储备数据集用于训练深度神经网络（DNN）模型。显然，新公布的法规阻碍了公司直接从其云服务器上收集和保留用户数据。

联邦学习为AI公司提供了一个有效的解决方案，该方法能够以合法的方式解决数据碎片和数据孤岛的问题。香港科技大学及其他研究机构的研究人员[4]列出了三种联邦学习模式，即横向联邦学习、纵向联邦学习和联邦迁移学习。当数据共享的参与方各自具有不重叠的数据集，但数据样本又具有相同的特征空间时，我们可以使用横向联邦学习。纵向联邦学习适用于参与方的数据集指向同一组实体，但其特征属性不同的情况。当参与方的数据集不能满足上述任意一个条件时（数据样本既指向不同实体，又具有不同的特征空间），联邦迁移学习可能是一个合理的选择。通过这些模式，AI公司始终能够为多个企业建立统一的模型，而无需在一个集中的地方共享它们的本地数据。

联邦学习不仅为机构间分布式机器学习模型的协同开发提供了一种用于隐私保护的技术解决方案，而且也为AI社会的可持续发展指明了一种新的商业模型，用于发展可信任的数字生态系统。通过在区块链的基础上运行联邦学习，我们能够利用智能合约和可信赖的利润激

励机制，使得数字生态系统中的成员主动分享其数据，并创建联邦机器学习模型。

联邦学习模式被中国越来越多的在线金融机构所采用。微众银行（WeBank）已经制定了一个关于联邦学习的开源项目，并向Linux基金会贡献了联邦AI技术使能器（federated AI technology enabler, FATE）框架。WeBank的AI研发团队[5]还启动了IEEE联邦学习的标准化工作，并开始起草体系结构框架的定义和应用指南。

2.2. AI 安全性、透明度和可信度

在计算机科学和软件工程领域，为确保大型复杂信息系统的安全性和可信度，研究人员已经进行了数十年的研究。随着信息系统的规模和复杂性的增加，如何以经济高效且无差错的方式设计和实施一个安全和可信赖的系统，成为计算机科学和软件工程领域面临的一个巨大挑战。当自主系统中被部署的AI组件与不确定的动态环境交互时，会不可避免地使上述问题变得更加困难。由于先进的AI模型采用了非常复杂的DNN和端到端的训练方法，这种黑箱性质不仅妨碍了开发人员充分理解其结构和行为，而且还会因恶意输入而给模型引入潜在漏洞。因此，AI治理框架必须综合多种技术，以使AI工程师能够对AI行为进行系统的评估，并提供能够建立公众对AI系统信任的证据。图1显示了AI行为分析和评估框架的主要构建模块，包括测试、验证、解释和溯源。

这些新兴的AI治理技术都是从不同方面来检查和评估AI行为和内部工作机制的。AI测试通常侧重于评估智能模型的输入和输出之间的关系，以确保AI的功能和行为能够符合所需的目标和伦理要求。AI验证采用严格的数学模型来证明AI算法的可靠性。AI解释旨在开发新型技术来分析和揭示复杂的DNN模型的内部工作机制。AI溯源可以跟踪训练数据、模型、算法和决策过程，以支持审计和责任认定。

这些AI治理技术的集成是非常重要的，因为它将所有的利益相关者聚集在一起，去理解、检查和审计一个自主且智能的系统。对于受AI系统决策影响的用户，他们有权了解和理解算法决策背后的原理。对于负责AI系统开发和维护的工程师，他们必须依靠AI测试、验证和解释工具来诊断AI算法的潜在问题，从而采取必要的补救措施和改进措施。对于负责监督AI工程流程和产品质量的管理人员，他们应利用这些工具来查询流程数据、指导伦理标准的实施以及降低系统的伦理和质量风险。对于调查事故或法律案件中AI系统责任的政府审计人员，他们必须利用AI溯源技术来跟踪系统演化的脉络并收集相关证据。

2.2.1. AI 的安全性和鲁棒性

近年来，DNN的对抗样本已成为机器学习领域非常热门的研究课题。DNN模型容易受到对抗样本的攻击，在这些样本中，带有细微扰动的输入会误导DNN，

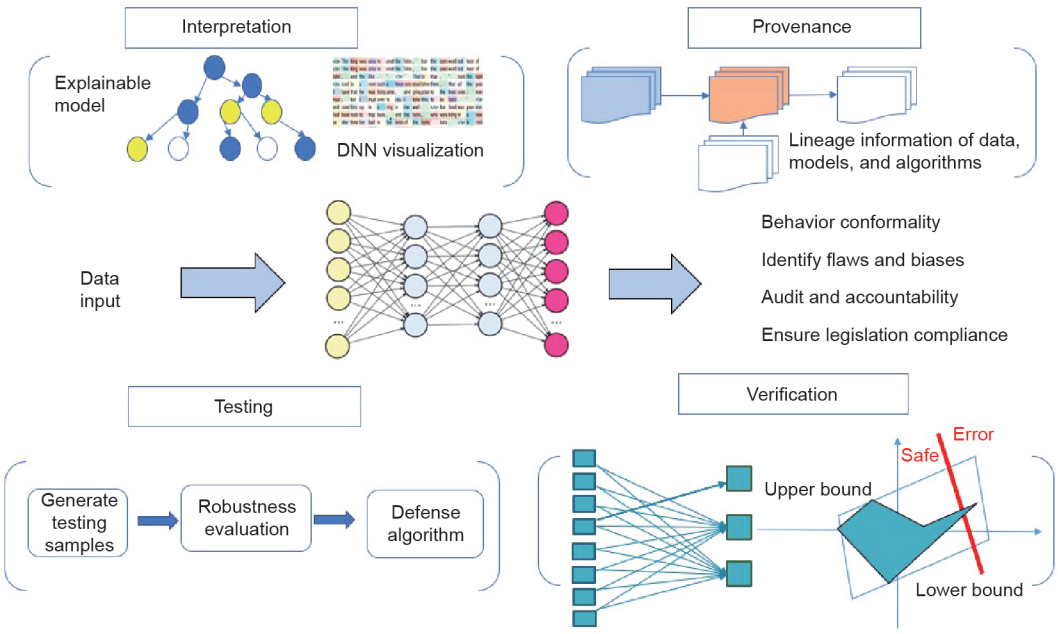

图1. 可信赖AI的测试、验证、解释和溯源。

使其产生错误的结果。例如，黑客可以恶意地在行人过马路图像上添加少量扰动，从而生成相应的对抗样本来欺骗DNN，使其忽略路上的行人。因此，当实际的深度学习应用程序（如自动驾驶和人脸识别系统）受到严重损害时，对抗样本可能会导致经济损失或致命事故。目前主要采用对抗测试和形式验证这两种方法来解决AI的安全问题，并确保AI系统在干扰下的鲁棒性。

（1）用于AI安全的对抗测试。已经有许多科研人员对如何通过生成对抗样本来测试DNN模型进行了研究。生成测试用例的一种最直接的方法是，在不影响场景的整体可视图像情况下直接干扰原始输入。但当黑客无法访问输入信号源且无法在输入图像中添加扰动时，此方法有其局限性。因此研究人员开始探索基于生成对抗网络（GAN）的对抗样本的生成，该对抗样本由微小的图像补丁组成，这些补丁可以被很容易地贴在物体（如户外的电线杆或人的帽子）表面上，从而影响图像输入[6]。

北京航空航天大学的研究人员[7]提出了一种感知敏感的GAN，该GAN可以增强对抗补丁的视觉保真度，并在安全性测试下为神经网络生成更真实的测试样本。在2018年国际NeurIPS（Neural Information Processing Systems）大会上，清华大学的研究人员发表了两篇有关DNN网络防御对抗样本算法的论文[8,9]。其中一篇提出了一种新的基于对抗扰动的正则化方法，即深度防御法（deep defense），用于训练DNN来对抗可能的攻击；另一篇则建议通过最小化训练过程中的反向交叉熵来检测可能存在的对抗样本。浙江大学和阿里巴巴的研究人员实施了一个以DEEPSEC命名的DNN测试平台[10]，该平台集成了十多种最先进的攻击和防御算法。该平台使研究人员和从业人员既能够评估DNN模型的安全性，又可以评估攻击和防御算法的有效性。

（2）DNN模型的形式化验证。因为输入扰动的选择组合情况极其庞大，对抗测试无法列举出给定一组输入的所有可能输出。作为对抗测试方法的补充，研究人员引入了形式验证方法，以此来证明DNN模型的输出与所有可能输入的规范严格一致。但是，对神经网络模型进行形式验证是一个非常困难的问题。已有研究表明，即使是验证关于神经网络行为的简单属性也是一个NP（non-deterministic polynomial）完全问题[11]。

形式化验证中遇到的主要困难主要来自于激活函数的非线性和复杂的神经网络结构。为了解决神经网络中非线性带来的问题，最近的研究工作都集中在了分段线性形式的激活函数上。研究人员正在研究有效且可扩展的验证方法，主要关注了输出集的几何边界。对于DNN模型，基本上有两种形式化的验证程序，即完整验证程序和不完整验证程序。完整的验证程序可以保证没有误报，但可扩展性有限，因为它们采用了计算成本较高的方法，如可满足性模理论（SMT）解决方案[12]。不完整的验证程序可能会产生误报，但其可扩展性比完整的验证程序好。苏黎世联邦理工学院（ETH Zurich）的研究人员[13,14]基于抽象解释提出了一种不完整的验证程序，其中基于形状的抽象域被表示为非线性激活函数输出的几何边界，以逼近DNN的无穷行为集。华东师范大学、中国科学院和其他相关机构的研究人员[15,16]也介绍了他们基于线性规划或符号传播的形式化验证框架。

这些研究工作仍处于起步阶段，目前对不同种类的激活函数和神经网络结构还没有进行过系统研究。尽管研究人员在形式化验证领域进行了数十年的研究，但是由于深度学习的复杂性，可扩展的验证方法在处理现代大规模的深度学习系统时尚不成熟。

2.2.2. AI 的透明度和可追溯性

AI的透明度对于公众在许多决策应用中能够理解和信任AI是至关重要的，如医疗诊断、贷款管理和法律执行。可解释AI有助于人们理解深度学习模型内部复杂的工作原理，它可以将这些深度学习模型的推理形成人类可以理解的解释。随着AI透明度的提高，人们将更加有信心去利用AI工具做出决策，并评估自主系统的合法性和可靠性。

研究人员正在研究如何构建可解释的DNN框架和分析工具。在这个研究方向上，多种方法已经被提出用于支持模型理解。一些研究人员设计出了伴随神经网络，用于在DNN推理过程中生成自然语言解释。另一种流行的方法被称为LIME [17]，它试图从原始的复杂模型中去构造基于简单模型（如稀疏线性模型和决策树）的代理模型，用以近似原始模型的行为。上海交通大学及其他研究机构的研究人员[18]提出了一种基于决策树的LIME方法，该方法从语义层面定量解释了预训练的卷积神经网络（convolutional neural network, CNN）所做出的每个预测的基本原理。

信息可视化也被广泛认为是实现可解释的DNN模型的一种有效方法。清华大学的研究人员[18]提出了交互式DNN可视化和分析工具，以支持模型的理解和诊

断。利用对伦理价值观的正确知识表示，这种视觉分析方法可以使AI工程师直观地验证其DNN模型是否遵循了人类的伦理原则。

与可解释AI密切相关的另一个重要研究领域是AI溯源。它强调记录、呈现和查询各种与模型、算法和数据有关的历史演化信息，以供将来进行审计和取证分析。尽管我们已经有了成熟的数据和信息溯源框架，但对AI溯源的研究还很少。南京大学和美国普渡大学联合发表了一篇论文[20]，该论文通过跟踪内部导数计算步骤，设计了AI算法的溯源计算系统。此方法可以帮助算法设计者诊断潜在问题。

除了能够促进AI模型的开发外，溯源技术在新兴的AI司法鉴定研究中也发挥着重要作用。最近，DeepFake技术被滥用，该技术利用GAN生成了虚假的人脸图像和视频，从而对社会规范和安全造成了巨大威胁。许多研究人员正在开发新的分类方法，以检测这些虚假的图像并确保视觉内容的可信度。例如，中国科学院自动化研究所的研究人员[21]尝试去改善DeepFake检测算法的通用性，并提出了一种新的司法鉴定CNN模型。但是，仅凭这些努力还不足以击败DeepFake，因为狡猾的设计者总是可以构思出更好的算法来欺骗已知的检测算法。或许，我们应该对原始图像的可靠来源信息进行补充，以提供必要的线索来验证图像来源的合法性。基于区块链的溯源管理系统对建立一个可靠且可信赖的数字生态系统是有帮助的，在这个生态系统中我们可以跟踪和验证数字资源的真实来源，从而彻底清除虚假图像和视频。

2.3. AI 算法的公平性评估

最近，公平性成为了评估AI算法的一种重要的非功能性特征。AI公平性研究的工作主要集中在评估不同群体之间或个人之间AI输出的差异。研究人员提出了许多公平性评价的标准。Gajane和Pechenizkiy [22]对相关文献中如何定义并形式化预测任务的公平性进行了调查。以下列出了AI公平性的主要定义。

（1）隐式的公平。只要在基于AI的决策过程中未明确使用被保护的属性，AI算法就是公平的。例如，智能欺诈检测系统应首先从其特征集中排除敏感属性，如种族和性别，然后再进行风险评估。尽管这种简单且盲目的方法在某些情况下可能会起作用，但它有一个非常严重的局限性，即排除相关属性可能会降低预测性能，而且从长远来看，其产生的结果的有效性要低于属性关注预测法。

（2）群体公平。群体公平要求AI算法针对基于特定属性区分的用户群体要做出相同概率的决策。群体公平有以下几种类型，包括人口结构均等、概率均等和机会均等。此类公平的定义之所以具有吸引力，是因为它不假定训练数据具有任何特殊属性，并且很容易被验证。

（3）个体公平。如果两个人具有相似的属性，则AI算法应该提出相似的决策。

（4）反事实公平。在许多决策场景中，受保护的属性（如种族和性别群体）可能对预测结果产生因果性影响。所以，“隐式的公平”度量标准可能会导致产生群体差异，而这正是该度量指标试图要避免的。为了减轻这种固有的偏见，Kusner等[23]利用因果框架来描述受保护的属性和数据之间的关系，从而给出了反事实公平的定义。这种公平性的定义还提供了一种用于解释偏见原因的机制。

当前，关于哪种公平性定义最合适尚无共识，甚至在某些情况下，这些定义彼此不兼容。如何为特定情况下的机器学习选择适当的公平标准，并在充分考虑社会背景的情况下设计公平的智能决策算法，仍是一个尚待解决的问题。

除了众多公平性定义外，研究人员还介绍了不同的偏差处理算法，用以解决AI模型生命周期不同阶段的公平性问题。例如，Bolukbasi等 [24]设计了一种去偏差方法，该方法可从自然语言处理常用的词嵌入中消除性别偏见。上海交通大学的研究人员[25]提出，在奖励机制中使用社会福利函数对公平性进行编码，并在深度强化学习的框架内解决资源分配问题的公平性。

大型AI公司正在积极开发公平性评估和去偏见工具，用以促进在真实智能系统中实施AI公平性。Google发布了名为What-If的交互式可视化工具，该工具使数据科学家能够以直观的方式检查复杂的机器学习模型。它集成了一些公平性指标，如隐式群体、机会均等和人口均等，用以评估和诊断机器学习模型的公平性。IBM创建了一个可扩展的开源工具包AI Fairness 360，用于处理算法的偏差[26]。该软件包在数据集和模型中集成了一套全面的公平标准和去偏差算法。

3. 中外 AI 伦理原则与治理技术发展对比

在本节中，我们将比较中国和其他国家在发展AI伦理原则和治理技术方面所开展的研究工作。从政府和机构的角度来看，中国的政府部门和企业在建立AI伦理

原则和促进AI向善的认知方面采取了积极的举措。从学术研究和产业发展的角度来看，中国研究人员和从业人员一直积极与国际同行齐头并进地开发AI治理技术。

3.1. 政府与机构的视角

世界上的主要经济大国已经发布了他们的AI伦理原则和治理法规。欧盟在2018年发布了GDPR。2019年4月，欧盟AI高级别专家组（High-Level Expert Group）提出了《可信赖AI的伦理原则》（Ethics Guidelines for Trustworthy AI）[27]。2019年，美国政府发布了《维持美国AI领先地位的行政命令》（Executive Order on Maintaining American Leadership in Artificial Intelligence），并要求美国国家标准与技术研究院（National Institute of Standards and Technology, NIST）为制定关于AI的相关技术标准进行计划，这些标准旨在规范开发可靠、鲁棒和可信赖的AI系统 [28]。与欧盟和美国一样，中国是在全国范围内发起AI治理和伦理倡议的主要国家之一。联合国也正在推进AI伦理方面的国际政策的制定，并在2019年3月联合国教科文组织（UNESCO）举办的AI会议上宣布了其对待AI的人文主义态度，它强调“以具有人类价值观的AI实现可持续发展”。但是，目前多国政府尚未采取任何联合行动来推动落实。

此外，大型科技公司，如谷歌、亚马逊和微软，以及中国的百度、阿里巴巴和腾讯，一直都在积极参与国内外的AI伦理和治理的相关工作。腾讯在2018年宣布了AI的ARCC（“可用、可靠、可理解、可控制”）原则，并在2019年发布了数字社会中关于AI伦理的报告[29]。百度加入了AI合作伙伴联盟（Partnership on AI）[30]，这是一个由AI行业的主要参与者组成的国际联盟，其任务旨在建立造福社会的AI系统。

3.2. 学术研究与工业开发的视角

在第2节中，我们重点介绍了中国研究人员在AI伦理和新兴治理技术方面的主要研究工作。在与AI伦理和治理技术相关的四个主要领域中，中国的研究人员与国际同行并驾齐驱，开发和设计出了新的模型、算法和工具。

在数据保密和隐私方面，微众银行的FATE是联邦学习的主要开源项目之一。根据参考文献[31]，与谷歌TensorFlow的联邦学习相比，FATE是这些开源项目中唯一支持分布式联邦学习的框架。

在AI的安全性和鲁棒性方面，针对Szegedy等[32]提出的DNN的脆弱性，中国研究人员在对抗攻击和防御方面开发的新算法表现出色。在2017年NIPS会议上，Google Brain组织了一场关于对抗性攻击和防御方法的国际竞赛，清华大学团队在攻击性和防御性方面均获得了第一名[33]。在这方面，也有国际合作的例子，中国的百度研究院与美国密歇根大学（University of Michigan）和伊利诺伊大学香槟分校（University of Illinois at Urbana-Champaign）的研究人员合作，发现了基于LiDAR的自动驾驶检测系统中DNN的安全漏洞[34]。

在AI的透明度和可追溯性方面，来自学术界和工业界（如阿里巴巴和百度）的中国研究人员，积极提出了新的解释方法和可视化工具。IBM、Facebook和微软之类的大公司已经发布了他们的AI解释工具。例如，IBM推出了AI Explainability 360，这个开源软件工具包集成了8种AI解释方法和2种评估指标[35]。相比于这些用于AI解释的通用框架，中国企业应努力将新算法和原型集成到开源工具中，并使其在全球范围内广泛使用。

尽管AI的公平性是一个相对较新的话题，但它在AI学术界引起了许多关注。如第2.3节所述，对AI公平性问题的研究通常需要跨学科的方法。2016年，ACM召开了公平、问责和透明大会（ACM Conference on Fairness, Accountability, and Transparency, ACM FAT），重点关注了AI的伦理问题，如算法的透明度、机器学习的公平性以及偏见。这次会议有500多人参加，其中包括AI领域的专家以及伦理学、哲学、法律和公共政策等社会科学领域的学者。尽管ACM FAT已经成为国际上研究人员讨论AI公平性的主要会议之一，但中国的研究人员对其了解甚少。所以鼓励中国AI学术界对这一新兴领域进行更多的跨学科研究是十分必要的。

总之，各国政府、学术界和产业界均认识到了AI伦理的重要性，并采取了积极的措施开发AI的治理技术。中国是在全国范围内发起AI治理和伦理倡议的主要国家之一。我们认为，为了国际社会的发展和我们人类共同的未来，在这个新兴领域加强国际合作是十分必要的。然而，目前这方面的国际合作还没有得到各国的充分重视。

4. AI 治理研究中的重大挑战

为了满足AI造福社会的基本原则，将伦理价值观和法规纳入当前的AI治理框架还需要克服许多挑战。本节中，我们将从以下几个方面详细阐述主要挑战：AI伦

理决策框架、AI工程开发流程和跨学科研究。

4.1. 伦理决策框架

伦理决策框架是AI治理研究的主要议题之一。香港科技大学和南洋理工大学的研究人员[36]在AI顶级会议上回顾了有关现有伦理决策框架的文章，并提出划分这一主题的分类法，将伦理决策框架划分为4个部分，即道德困境探索、个体伦理决策框架、集体伦理决策框架和人机交互中的伦理。其他研究者[37]也回顾了关于通用AI（artificial general intelligence, AGI）的安全性研究，其中，伦理决策问题研究通常采用强化学习的理论框架。他们假定理性智能体能够通过与社会环境互动产生的经验来学习人类的道德偏好和规则。因此，在强化学习的框架中，AI设计者可以将伦理价值观指定为奖励函数，以使智能体的目标与其人类同伴的目标保持一致，并激发智能体以人类的道德规范行事。在新兴的研究领域，科学家们必须克服当前数据驱动型DNN模型的主要瓶颈，以实现人类水平的自动化伦理决策，并对部署到实际复杂道德环境中的这些理论框架进行广泛评估。

4.1.1. 如何为伦理规则和价值观建模

在大多数情况下，我们很难通过直接设计数学函数来对道德价值观建模，尤其针对所谓的道德困境，即人们只能在消极的选择中做出艰难的决策。采取数据驱动型和基于学习的方法使自主智能体从人类经验中学习适当的道德表示是可行的。例如，麻省理工学院的研究人员发起了道德机器项目（Moral Machine project）[38]，以众包方式收集有关各种道德困境的数据集。但是，由于这种倾向于道德困境的众包式自我报告方法没有任何约束机制，它无法确保用户的真实选择，所以这种方法会不可避免地偏离实际的决策行为。

4.1.2. 伦理决策中的常识和情境感知

尽管现代AI技术飞速发展，但基于DNN的AI智能体大多只擅长识别潜在模式，在开放的和非结构化的环境中，这样的智能体不能很好地支持通用的认知智能。在复杂的道德困境中，最先进的AI智能体尚没有足够的认知能力来感知正确的道德环境，所以其更无法通过常识性推理来成功地解决困境。目前的研究工作主要集中于探索博弈论道德模型或基于Bayesian的效用函数。美国杜克大学的研究人员[39]采用博弈论方法对道德困境建模，并使AI的道德价值观与人类的一致。MIT的研究人员[40]设计了一种计算模型，用于将道德困境描述为效用函数，并引入了层次Bayesian模型来表示社会结构和群体规范。虽然这些早期的尝试可能不足以支持常见的道德场景，但它们至少为如何在AI伦理学领域综合利用DNN和可解释Bayesian推理模型这两种方法指出了新的研究方向。

4.1.3. 安全强化学习

许多研究人员采用深度强化学习将道德约束建模为奖励函数，并使用Markov决策过程进行序贯决策。然而，深度强化学习还不够成熟，要将其应用于游戏之外的其他场景的话，我们还有很长的路要走。其主要问题之一在于强化学习过程的安全性。一个恶意的智能体能够使用多种方法来欺骗奖励机制，规避道德监管约束。例如，它可以利用奖励过程中存在的漏洞来获得比预期更多的奖励。

4.2. 在智能工程流程中整合 AI 伦理原则

伦理原则应被转化为指导AI系统设计和实施的软件规范。从软件工程的角度来看，对AI模型的开发和AI系统的运行通常是按照一个有明确定义的生命周期（图2）来组织，这一生命周期包括AI任务定义、数据收集和准备、AI设计和模型训练、模型测试和验证、模型部署和应用。AI隐私性、安全性和公平性的软件规范的实施应该贯穿于整个AI开发和运行一体化（DevOps）的生命周期。

首先，我们需要在需求分析阶段定义和分析AI任务。设计人员可以针对不同应用场景中的定制要求，采用不同种类的伦理规范和评估指标。在数据收集和准备阶段，工程师必须通过消除已损坏的数据样本和减少数据集的潜在偏差来确保训练数据集的有效性。利用平衡且正确的数据集，工程师可以根据伦理规范设计出恰当的模型结构，从而进行模型训练。在设计和模型训练阶段之后，工程师必须根据伦理规范所描述的公平性、鲁棒性、透明性和任务绩效方面的约束条件对初步模型进行测试和验证。如果模型无法通过测试和验证阶段，则工程师必须重新设计模型、重新检查数据和重新训练模型。如果模型通过测试和验证阶段，则该模型可以与其他软件组件进行集成并被部署在智能系统中。在系统运

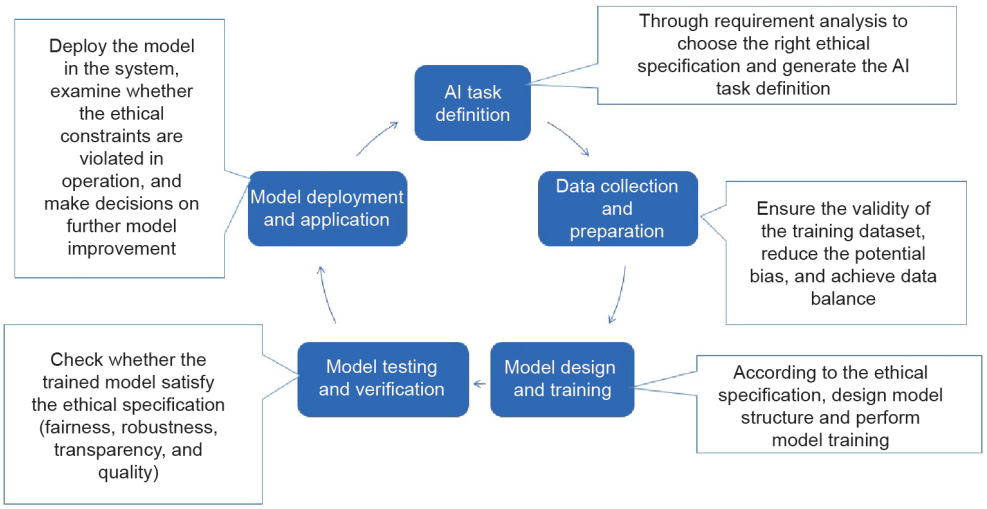

图2. AI DevOps生命周期。

行期间，工程师将不断检查该系统的运行时行为，并判断其是否符合道德要求。如果违反道德约束，工程师则必须进一步改进AI模型，并启动新一轮的AI DevOps生命周期。

显然，为了简化这种具有伦理意识的AI DevOps生命周期，我们需要开发许多AI工程工具，并将其集成到一个全面而灵活的环境中，以供AI模型设计者和系统开发人员使用。如前几节所述，这些工具必须能实现核心技术，如联邦学习、对抗测试、形式化验证、公平性评估、可解释、溯源和运行时沙盒以及安全监控。目前，这些工具（如AI Fairness 360）仍在开发中，因此主要的AI DevOps平台尚未将这些工具集成在一起，形成AI伦理原则所要求的主要功能。为了形成一个具内置伦理支持的开放的AI DevOps环境，更多的研究工作和工程开发是必不可少的，在这个环境中，研究人员和工程师可以方便地探索新的AI道德技术，系统地评估不同的指标，并针对各种应用领域中不同的道德场景构思新的解决方案。

随着AI治理技术的进步，我们可以预见，AI伦理方面的法规和标准将在企业、社区、国家和国际层面得到不断的实施，以加强AI系统和产品的合规性。实际上，全球AI标准化研究工作已经开展了好多年。例如，国际标准化组织（ISO）的SC 24于2018年成立了AI可信度工作组，中国人工智能标准化总体组、专家咨询组则于2019年发布了AI伦理风险分析白皮书[41]。AI工业界和标准化组织的共同努力将进一步提高人们对AI伦理问题的认识，并在AI工业界和研究社区加速道德价值观在智能系统和产品中的集成。

4.3. AI 治理的跨学科研究

AI系统是复杂且先进的社会性技术系统，通常涉及机器学习模型、支撑软件组件和社会组织。来自多个学科的研究人员需要对AI进行社会系统分析[42]，以理解AI在不同社会、文化和政治环境下所带来的影响。这样的社会系统分析需要跨学科的研究方法，需要充分利用哲学、法学和社会学以及其他学科的相关研究成果。通过多学科研究，AI设计者和开发人员可以与法学专家和社会学专家合作，对智能系统的道德方面进行整体建模和分析，用以评估可能对各方产生的影响，并在AI DevOps的各个阶段和状态中处理道德问题。

毫无疑问，对社会性技术工程来说，这种跨学科的和整体的分析方法要求AI开发人员与具有其他领域专业知识的合作伙伴进行深入合作。尽管中国的计算机科学、哲学、法学和社会学领域的研究人员对AI伦理学的认知日益提高，但他们的大部分研究工作仍在各自的轨道上进行，尚未完全协同起来去解决上述重大挑战。因此，我们认为将所有相关学科的专家召集在一起是至关重要的，并针对具有明确目标的AI伦理问题开展工作。首先，我们需要根据公认的AI伦理原则，在自动驾驶、智能法院和金融贷款决策等应用中确定关键和典型的道德场景，并要求跨学科团队提出新的研究思路和解决方案。在这些场景中，我们要适当地对复杂社会环境进行抽象并将其描述为AI伦理规范。其次，我们应该创建一个开放和通用的平台来促进AI伦理的跨学科研究。这样的平台将极大地方便来自不同背景的研究人员分享他们的见解和工作，并比较不同的技

术框架和道德标准，以构建具有道德价值观和准则的智能机器。

5. 结论

AI的快速发展和部署预示着人类社会即将发生根本性转变。这将是建设人类命运共同体和促进自然与社会可持续发展的绝佳机会。但是如果没有足够的有效治理手段和规章制度，这一转变也有可能会造成前所未有的消极影响。为了确保这些转变在完全融入我们的生活之前能造福人类，我们必须建立可靠且可行的AI治理框架，用以根据人类道德和价值观来约束AI开发者。这样，我们才能让AI变得可追责且值得信赖，同时增强公众对AI技术和系统的信任。

本文介绍了中国在发展AI治理理论和技术方面正在开展的工作。许多中国研究人员正在积极开展研究，以解决当前AI的伦理问题。为了克服数据驱动型AI的保密性问题，中国的大学和企业的科研团队正努力开发联邦学习技术。为了确保DNN模型的安全性和鲁棒性，研究人员在对抗性测试和形式化验证中提出了新的算法。此外，一些研究团队还在研究可解释AI、溯源和取证领域的有效框架。这些研究工作目前仍处于起步阶段，有待进一步加强，从而为将来的广泛应用和实践提供成熟的解决方案。

我们建议采取以下行动来推进当前有关AI治理的倡议。第一，政府、基金会和企业应开展跨学科、跨行业和跨国合作，从而在AI伦理原则方面达成共识。第二，政府、基金会和企业必须加强AI治理技术的协作研发，以跟上AI快速发展的步伐。第三，开发具有内置伦理约束相关工具的开放式AI DevOps平台，以支持不同AI系统的相关人员来评估AI系统的功能和合法性。第四，明确定义具有重大社会影响的AI伦理情景，以便来自不同学科的专家能够共同应对AI伦理挑战。最后，我们必须积极促进对在研发、应用和管理环节的每个AI利益相关者的伦理教育，以显著提高他们的AI伦理意识。

Compliance with ethics guidelines

Wenjun Wu, Tiejun Huang, and Ke Gong declare that they have no conflicts of interest or financial conflicts to disclose.

References

[1] National Governance Committee for the New Generation Artificial Intelligence. Governance principles for the new generation artificial intelligence—developing responsible artificial intelligence [Internet]. Beijing: China Daily; c1995-2019 [updated 2019 Jun 17; cited 2019 Dec 18]. Available from: https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html?from=timeline&isappinstalled=0.

[2] Beijing AI Principles [Internet]. Beijing: Beijing Academy of Artificial Intelligence; c2019 [updated 2019 May 28; cited 2019 Dec 18]. Available from: https://www.baai.ac.cn/blog/beijing-ai-principles

[3] Zeng Y, Lu E, Huangfu C. Linking artificial intelligence principles. 2018. arXiv:1812.04814.

[4] Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. ACM Trans Intell Syst Technol 2019;10(2):12.

[5] Guide for architectural framework and application of federated machine learning [Internet]. New York: IEEE P3652.1 Federated Machine Learning Working Group; c2019 [cited 2019 Dec 18]. Available from: https://sagroups.ieee.org/3652-1/.

[6] Xiao C, Li B, Zhu J, He W, Liu M, Song D. Generating adversarial examples with adversarial networks. 2018. arXiv:1801.02610.

[7] Liu A, Liu X, Fan J, Ma Y, Zhang A, Xie H, et al. Perceptual-sensitive GAN for generating adversarial patches. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence; 2019 Jan 27–Feb 1; Honolulu, HI, USA; 2019.

[8] Yan Z, Guo Y, Zhang C. Deep defense: training DNNs with improved adversarial robustness. Adv Neural Inf Process Syst 2018:419–28.

[9] Pang T, Du C, Dong Y, Zhu J. Towards robust detection of adversarial examples. Adv Neural Inf Process Syst 2018:4579–89.

[10] Ling X, Ji S, Zou J, Wang J, Wu C, Li B, et al. DEEPSEC: a uniform platform for security analysis of deep learning model. In: Proceedings of the 40th IEEE Symposium on Security and Privacy; 2019 May 20–22; San Francisco, CA, USA; 2019.

[11] Pulina L, Tacchella A. Challenging SMT solvers to verify neural networks. AI Commun 2012;25(2):117–35.

[12] Katz G, Barrett C, Dill DL, Julian K, Kochenderfer MJ. Reluplex: an efficient SMT solver for verifying deep neural networks. In: Proceedings of the International Conference on Computer Aided Verification; 2017 Jul 24–28; Heidelberg, Germany; 2017. p. 97–117.

[13] Gehr T, Mirman M, Drachsler-Cohen D, Tsankov P, Chaudhuri S, Vechev M. AI2: safety and robustness certification of neural networks with abstract interpretation. In: Proceedings of the 2018 IEEE Symposium on Security and Privacy; 2018 May 20–24; San Francisco, CA, USA; 2018.

[14] Singh G, Gehr T, Mirman M, Püschel M, Vechev M. Fast and effective robustness certification. In: Proceedings of the Advances in Neural Information Processing Systems 31; 2018 Dec 3–8; Montreal, QC, Canada; 2018. p. 10802–13.

[15] Lin W, Yang Z, Chen X, Zhao Q, Li X, Liu Z, et al. Robustness verification of classification deep neural networks via linear programming. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019 Jun 16–20; Long Beach, CA, USA; 2019. p. 11418–27.

[16] Yang P, Liu J, Li J, Chen L, Huang X. Analyzing deep neural networks with symbolic propagation: towards higher precision and faster verification. 2019. arXiv:1902.09866.

[17] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA, USA; 2016. p. 1135–44.

[18] Zhang Q, Yang Y, Ma H, Wu YN. Interpreting CNNs via decision trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019 Jun 16–20; Long Beach, CA, USA; 2019. p. 6261–70.

[19] Liu S, Wang X, Liu M, Zhu J. Towards better analysis of machine learning models: a visual analytics perspective. Visual Inf 2017;1(1):48–56.

[20] Ma S, Aafer Y, Xu Z, Lee WC, Zhai J, Liu Y, et al. LAMP: data provenance for graph based machine learning algorithms through derivative computation. In: Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering; 2017 Sept 4–8; Paderborn, Germany; 2017. p. 786–97.

[21] Xuan X, Peng B, Dong J, Wang W. On the generalization of GAN image forensics. 2019. arXiv:1902.11153.

[22] Gajane P, Pechenizkiy M. On formalizing fairness in prediction with machine learning. 2017. arXIv:1710.03184.

[23] Kusner MJ, Loftus J, Russell C, Silva R. Counterfactual fairness. Adv Neural Inf Process Syst 2017:4066–76.

[24] Bolukbasi T, Chang KW, Zou J, Saligrama V, Kalai A. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Adv Neural Inf Process Syst 2016:4349–57.

[25] Weng P. Fairness in reinforcement learning. 2019. arXiv:1907.10323.

[26] Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. 2018. arXiv:1810.01943.

[27] High-Level Expert Group on AI. Ethics guidelines for trustworthy AI [Internet]. Brussels: European Commisiion; 2019 Apr 8 [cited 2019 Dec 18]. Available

from: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
[28] Trump DJ. Executive order on maintaining American leadership in artificial intelligence [Internet]. Wahington, DC: The White House; 2019 Feb 11 [cited 2019 Dec 18]. Available from: https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/.
[29] Tencent AI Lab. Technological ethics at intelligent era—reshape trustworthiness in digital society [Internet]. Beijing: Tencent Research Institute; 2019 Jul 8 [cited 2019 Dec 18]. Available from: https://tisi.org/10890. Chinese.
[30] Meet the Partners [Internet]. San Francisco: Partnership on AI; c2016–2018 [cited 2019 Dec 18]. Available from: https://www.partnershiponai.org/partners/.
[31] Li Q, Wen Z, Wu Z, Hu S, Wang N, He B. Federated learning systems: vision, hype and reality for data privacy and protection. 2019. arXiv:1907.09693.
[32] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D,Goodfellow I, et al. Intriguing properties of neural networks. 2013. arXiv:1312.6199.
[33] Kurakin A, Goodfellow I, Bengio S, Dong Y, Liao F, Liang M, et al. Adversarial attacks and defences competition. In: Escalera S, Weimer M, editors. The NIPS'17 competition: building intelligent systems. Cham: Springer; 2018. p. 195–231.
[34] Cao Y, Xiao C, Yang D, Fang J, Yang R, Liu M, et al. Adversarial objects against LiDAR-based autonomous driving systems. 2019. arXiv:1907.05418.
[35] Arya V, Bellamy RK, Chen PY, Dhurandhar A, Hind M, Hoffman SC, et al. One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques. 2019. arXiv:1909.03012.
[36] Yu H, Shen Z, Miao C, Leung C, Lesser VR, Yang Q. Building ethics into artificial intelligence. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence; 2018 Jul 13–19; Stockholm, Sweden; 2018. p. 5527–33.
[37] Everitt T, Kumar R, Krakovna V, Legg S. Modeling AGI safety frameworks with causal influence diagrams. 2019. arXiv:1906.08663.
[38] Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, et al. The moral machine experiment. Nature 2018;563(7729):59–64.
[39] Conitzer V, Sinnott-Armstrong W, Borg JS, Deng Y, Kramer M. Moral decision making frameworks for artificial intelligence. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence; 2017 Feb 4–10; San Francisco, CA, USA; 2017. p. 4831–5.
[40] Kim R, Kleiman-Weiner M, Abeliuk A, Awad E, Dsouza S, Tenenbaum JB, et al. A computational model of commonsense moral decision making. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society; 2018 Feb 2–3; New Orleans, LA, USA; 2018. p. 197–203.
[41] National Artificial Intelligence Standardization Steering Committee. Report on artificial intelligence ethical risk analysis [Internet]. [cited 2019 Dec 18]. Available from: http://www.cesi.ac.cn/images/editor/20190425/20190425142632634001.pdf. Chinese.
[42] Crawford K, Calo R. There is a blind spot in AI research. Nature 2016; 538(7625):311–3.