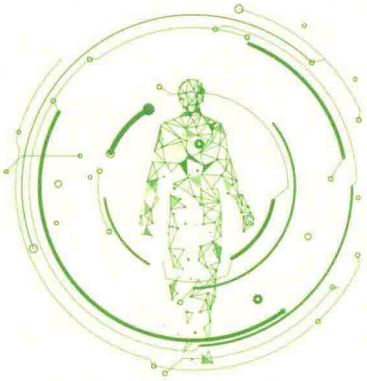
互联网、大数据与人工智能伦理丛书

人工智能与大数据伦理

李　伦／主编



ETHICS OF ARTIFICIAL INTELLIGENCE AND BIG DATA

科学出版社

互联网、大数据与人工智能伦理丛书

人工智能与大数据伦理

李 伦/主编

ETHICS OF ARTIFICIAL INTELLIGENCE AND BIG DATA

科学出版社

北京

图书在版编目（CIP）数据

人工智能与大数据伦理 / 李伦主编. —北京：科学出版社，2018.12
（互联网、大数据与人工智能伦理丛书）
ISBN 978-7-03-059872-1

Ⅰ. ①人… Ⅱ. ①李… Ⅲ. ①人工智能-技术伦理学-研究 ②数据处理-技术伦理学-研究 Ⅳ. ①B82-057

中国版本图书馆 CIP 数据核字（2018）第 271598 号

丛书策划：侯俊琳 邹 聪
责任编辑：邹 聪 张 楠 / 责任校对：韩 杨
责任印制：张欣秀 / 封面设计：有道文化
编辑部电话：010-64035853
E-mail：houjunlin@mail.sciencep.com

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com
北京建宏印刷有限公司印刷
科学出版社发行 各地新华书店经销
*
2018 年 12 月第 一 版 开本：720×1000 B5
2018 年 12 月第一次印刷 印张：26
字数：350 000
定价：128.00 元
（如有印装质量问题，我社负责调换）

丛书出版获以下项目支持

国家社会科学基金重大项目“大数据环境下信息价值开发的伦理约束机制研究”（17ZDA023）
国家社会科学基金一般项目“开源运动的开放共享伦理研究”（17BZX022）
湖南省高等学校“双一流”学科建设项目湖南师范大学哲学学科
大连理工大学学科建设项目“人工智能伦理问题研究”

互联网、大数据与人工智能伦理丛书
编委会

丛书序

互联网、大数据和人工智能是当代及未来发展的驱动力。互联网拓展了人类的生存空间，大数据是21世纪的“新石油”，人工智能成了社会发展的引擎。世界各国纷纷将互联网、大数据和人工智能的发展上升至国家发展战略层面。同时，互联网、大数据和人工智能的发展面临诸多现实的伦理和法律问题，如网络安全、个人隐私、数据权益和公平公正等。关于这些问题的伦理学研究常常是制定相关法律法规和政策的前置议程。和发达国家一样，我国在制定互联网、大数据和人工智能发展战略时，也极其重视伦理和法律问题的研判与应对。

2015年7月，国务院发布《国务院关于积极推进“互联网+”行动的指导意见》，要求加快“互联网+”相关立法工作，落实加强网络信息保护和信息公开的有关规定，加快推动制定网络安全、个人信息保护、互联网信息服务管理等法律法规，逐步完善相关标准规范、信用体系和法律法规。

2015年8月，国务院发布《促进大数据发展行动纲要》，高度重视数据共享、数据安全和隐私保护。《促进大数据发展行动纲要》要求明确数据共享的范围边界和使用方式，厘清数据共享的义务和权利，加强安全保障和隐私保护，界定个人信息采集应用的范围和方式，明确相关主体的权利、责任和义务，加强对国家利益、

公共安全、商业秘密、个人隐私、军工科研生产等信息的保护，加强对数据滥用、侵犯个人隐私等行为的管理和惩戒，推动数据资源权益相关立法工作。

2017 年 7 月，国务院发布《新一代人工智能发展规划》，对人工智能伦理问题研究提出了明确要求，将人工智能伦理法律研究列为重点任务，要求开展跨学科探索性研究，推动人工智能法律伦理的基础理论问题研究。《新一代人工智能发展规划》指出人工智能可能冲击法律与社会伦理、侵犯个人隐私、挑战国际关系准则，要求加强前瞻预防与约束引导，最大限度地降低风险，确保人工智能安全、可靠、可控发展。最引人注目的是，《新一代人工智能发展规划》关于人工智能伦理和法律制定了三步走的战略目标：到 2020 年，部分领域的人工智能伦理规范和政策法规初步建立；到 2025 年，初步建立人工智能法律法规、伦理规范和政策体系；到 2030 年，建成更加完善的人工智能法律法规、伦理规范和政策体系。

技术发展与社会环境戚戚相关。在大力发展高新技术的同时，必须高度重视可能的社会风险和伦理挑战，必须加强技术伦理学研究。技术伦理学是规范性的，也是建设性的。技术伦理学研究旨在揭示技术发展面临的伦理难题，为技术发展清理路障，同时为技术发展提供价值指引，确保技术在造福人类的轨道上发展。

随着互联网、大数据与人工智能对人类社会影响的普遍化，其伦理问题不再只是寓于哲学伦理学圈内的议题，已成为政界、业界、学界和公众高度关注的公共话题。对这些问题的研究也不再局限于哲学伦理学方法，搭建多学科交叉研究和交流的平台势在必行。

李伦教授主编的这套丛书是搭建这种平台的一种尝试。这套丛书将运用伦理学、法学、社会学和管理学等的理论与方法，关切人类未来，聚焦互联网、大数据和人工智能面临的现实问题，如网络内容治理问题、网络空间数字化生存问题、数据权和数据主权问题、隐私权和自主权问题、数据共享和数据滥用问题、网络安全和信息安全问

题、网络知识产权问题、大数据价值开发的伦理规范问题，以及人工智能的道德哲学、道德算法、设计伦理和社会伦理等问题，并为治理互联网、大数据和人工智能的伦理问题提供对策建议。

郭东明

中国工程院院士，大连理工大学校长

代 序

给人工智能一颗良芯

——人工智能伦理研究的四个维度*

李 伦 孙保学

随着大数据技术的发展、算法的进步和机器算力的大幅提升，人工智能在众多领域不断攻城略池，赶超人类。同时，围绕人工智能产生的伦理问题也越来越突出，成为全社会关注的焦点。近年来，国际学术界对人工智能道德算法、价值定位、技术性失业和致命性自主武器等问题展开了广泛的讨论，一系列富有卓见的研究成果相继问世。然而，学界关于人工智能伦理研究的问题框架尚未达成共识，其正处于探索研究范式的阶段，人工智能伦理的教育与教学究竟该从哪些方面展开仍然缺少框架性的指导。其实，设计和研发安全可靠的人工智能不是单纯的技术层面的要求，也需要道德层面的引导和规制。人工智能的发展应增进人类福祉，与人类的道德价值相符合。要做到这一点，必须给人工智能一颗良芯（良心）。这意味着以打造良芯（良心）为中心任务的人工智能伦理研究应从

* 原载《教学与研究》2018 年第 8 期。

“机芯”（机器之芯）和“人心”（人类之心）两个维度来展开。“机芯”研究主要是指人工智能道德算法研究，旨在使人工智能拥有“良芯”，使之成为道德的人工智能或道德的机器（moral machine）。“人心”是指人工智能研发者和应用者应当具有“良心”，使人工智能的设计合乎道德，避免恶意设计，并确保人工智能的善用，使之造福人类社会。也就是说，“人心”研究主要涵盖人工智能设计伦理和社会伦理等方面的研究。“机芯”和“人心”研究存在诸多基础性问题，需要我们在道德哲学层面做出回应或拓展。因此，人工智能伦理研究包括人工智能道德哲学、人工智能道德算法、人工智能设计伦理和人工智能社会伦理四个维度。

一、人工智能道德哲学

随着人工智能和自主系统的决策能力的不断提升，它们可能对传统的道德哲学构成挑战。传统的伦理观念和道德理论的适用性问题将变得越来越突出，其中一些伦理原则或道德规范有可能会失效或部分失效，甚至在某些领域引起道德哲学的真空。不过，这反过来可能会倒逼伦理学家进行更加深刻的思考，修正已有的道德概念和道德理论，提出适应智能时代的全新的道德哲学体系，从而促进道德哲学的整体发展。从这种意义上讲，人工智能道德哲学并非替代传统道德哲学，而是对传统道德哲学的延伸和扩张。

那么，我们究竟该从何种意义上谈论人工智能的道德哲学呢？按照尼克·波斯特洛姆（Nick Bostrom）和埃利泽·尤德考斯基（Eliezer Yudkowsky）的观点，“创造思维机器的可能性提出了一系列的伦理问题，这些问题既与确保这种机器不伤害人类和其他道德上关联的存在者有关，也与机器自身的道德地位有关”[1] 316。智能机器不伤害人类和其他道德上关联的存在者是设计伦理的要求，而机器自身的道德地位问题则是人工智能道德哲学的范畴，相关讨论可以从人工智能的主体地位、道德责任和人机关系等方面展开。

第一，构建人工道德主体（artificial moral agents，AMAs）是否是

可能的？谈论人工智能伦理是否应该把 AMAs 地位的确立作为前提？不过，按照计算机伦理学的开创者詹姆斯·摩尔（James Moor）的观点，我们可以根据智能机器的自主性程度划分出四种人工道德主体：有道德影响的主体、隐式的道德主体、显式的道德主体和完全的道德主体。[2] 因此，即使人工智能系统并没有被内置道德规范，只要其决策算法能够产生伦理影响，即可将其视为具备道德推理能力，可以成为道德受体或道德关怀对象。此外，讨论的更为前沿的话题是：拥有自由意志或自主能力是否可以作为人工智能和自主系统成为道德主体的前提？如果人工智能有资格作为道德主体，那么人性中的情感和意识等因素是否是人工智能具备道德的必要条件？的确，拥有自我意识并且具备人类情感的人工智能会更像人，但将情感和自由意志赋予人工智能有可能会打开“潘多拉的魔盒”，给人类带来灾难性的后果。在处理人类错误方面，人类社会积累了丰富的经验，但在如何应对人工智能机器的错误方面，我们的社会可能尚未做好充分准备。如果不良情绪被植入智能系统，它们可能不是给人们带来便利，而是给人们制造更多的麻烦。关于类似问题的探讨对传统的道德哲学提出了挑战，我们已有的关于“自由意志”“自主”“理性”“道德主体”“道德受体”“道德地位”等概念可能需要被重新定义。这可能导致新的道德哲学的产生，或导致传统的道德哲学发生转向，如人工物伦理学的转向。“与许多人的直觉想法相反，技术不再是让我们的生存变得便利的简单的中立工具。在这些技术完成它们功能的同时，已经产生了更多的效应：它们框定着我们该做什么以及我们如何体验世界，并且以此方式，它们积极参与到我们的生活中。”[3] 1

第二，人工智能的道德权利和责任分配问题。在直觉上，多数人可能会认为 AMAs 并不应享有与人类相同的权利。在他们看来，AMAs 是服务于人的福祉的，而不应该被设定为追求自身的福祉。人工道德主体的界定困难导致相应的责任、义务和权利的认定变得复杂。如果人工道德主体以人类伙伴的身份服务于社会，那么它们的权利和义务边界如何划定，需要人类慎重对待以平衡人类的同理心。如果人工智能具有道德决策能力，同时获得了道德主体的地位，那么它

的设计者、制造者和使用者是否应当为它们的过失负担相应的责任呢？如果它们能够承担一定的责任，那么人们是否会设法将一些道德责任推卸给人工智能来承担呢？面对这种情形，AMAs 可能沦为替人类行为免责的一种工具，它们的权利保障可能更是无从谈起。如果 AMAs 能够为其不当行为承担完全责任，那么责任的边界和限度的划定问题将变得越来越突出。毕竟，道德谴责和否定评价如何对 AMAs 产生实际的道德约束在技术上仍然是个难题。如果道德约束对于 AMAs 是不充分的，那么它们承担相应的法律责任是否是可行的？如果 AMAs 需要承担一定的刑事责任，那么究竟何种惩罚对它们具有威慑力（例如，清除记忆、机身销毁等是否等同于机器的死亡，限制机器人人身自由是否有效）？同时，人类共同体在情感上能否接受这种形式的惩罚？如果人工智能没有基本的财产权，那么它们以何种形式承担民事责任？是社会整体承担还是多主体分担？这些都是值得深入探讨的问题。

第三，人类道德与机器道德的关系问题。通常，人们基于后果来判定某人行为是否道德，但深入思考人们会认识到，主体的心灵状态（意图、动机或目的等）在评价或判定某人行为是否道德时发挥着重要作用。目前，主流的道德理论基本上都是以人类为中心提出的。随着人机互动时代的到来，这些道德理论可能需要卢西亚诺·弗洛里迪（Luciano Floridi）提出的所谓“分布式道德”来补充。根据这种观点，局部层面的智能体相互作用，可以导致系统整体或宏观层面的道德行为和集体责任感的增加。[4] 另外，虽然 AMAs 必须以人类价值为目标被建构起来，但面对多样化的文化价值系统、多元化的宗教传统和无数的伦理理论，我们究竟该选择哪种作为底层的设计框架呢？按照瓦拉赫和艾伦的观点，弱人工智能的道德实际上是一种“操作性道德”，由设计者和使用者赋予其意义；随着复杂程度的增加，人工智能将会表现出所谓的“功能性道德”，它将使其自身具备道德决断能力并能够影响道德挑战。[5] 6 这种划分有助于我们明晰所谈论的到底是指何种道德。不过，如果纯粹从实用主义视角出发，那么人工智能是否有资格成为道德主体可能并不是最紧迫的问题，只要它们能够像道德主体一般行动，尽可能地减少道德摩擦，而不惹是生非。从现实层面来

讲，那些在道德上具有卓越表现的智能机器会有更高的公众认可度，能够更加顺利地融入人类社会。这种可接受性的动力学来源可能来自以下两个方面：一方面，机器道德的研究要求人们深刻理解人类道德的发生、起源和机制，在这种意义上建构人工智能道德，可能要求我们更多地思考如何从学习、进化和发展的角度去考察培养人工智能的道德感；另一方面，机器道德的研究对传统的道德哲学发挥“反哺”功能，它能够激发传统的伦理学研究做出革新，甚至催生全新的伦理学体系。

二、人工智能道德算法

我们已经生活在算法时代。几十年前，人们对于“算法”的认识主要局限于数学和计算机领域。“当今，文明社会的每个角落都存在算法，日常生活的每分每秒都和算法有关。算法不仅存在于你的手机或笔记本电脑，还存在于你的汽车、房子、家电以及玩具当中。”[6]3 事实上，算法已经成为当今信息社会的一种基础设施，它们能够引导甚至支配人们的思维与行动。因此，算法越来越多地参与到人类的道德生活中将是一种必然趋势。当然，有人会主张道德是不可计算的，反对用算法刻画道德，甚至认为道德领域是技术扩张的禁忌之地。不过，人工智能对社会结构的渗透作用势不可挡，算法的影响早已深入文明社会的每个角落。

人工智能相关技术的核心主要体现在决策算法方面。人工智能道德算法的研究主要是指那些在道德上可接受的算法或合乎伦理的算法，它们使自主系统的决策具有极高的可靠性和安全性。从这种意义上讲，道德算法是实现人工智能功能安全的一项基本原则和技术底线。那么，人工智能如何才能成为一个安全可靠的道德推理者，能够像人类一样甚至比人类更加理性地做道德决策呢？在此，我们首先需要明确人工智能是如何进行道德决策的，也就是人工智能的道德推理机制问题。目前，相关研究主要沿着以下三个方向展开。

一是理论/规则驱动进路。这也被称为自上而下的进路，它是将特定群体认可的价值观和道德标准程序化为道德代码，嵌入智能系统，

内置道德决策场景的指导性抉择标准。在这方面，人们最容易想到的是20世纪40年代艾萨克·阿西莫夫（Isaac Asimov）提出的“机器人三定律”：①机器人不可以伤害人，或者通过不作为，让任何人受到伤害；②机器人必须遵从人类的指令，除非那个指令与第一定律相冲突；③机器人必须保护自己的生存，条件是那样做与第一、第二定律没有冲突。[5] 1 从理论上讲，伊曼纽尔·康德（Immanuel Kant）的道义论、约翰·密尔（John Mill）的功利主义和约翰·罗尔斯（John Rawls）的正义论等都可以成为理论驱动进路的理论备选项。早期机器伦理尤其是符号主义的支持者对理论/规则驱动模式情有独钟，但技术瓶颈和可操性难题使得这种研究进路日渐式微。例如，有限的道德准则如何适应无穷变化的场景在技术上始终是个难题；如何调和不同的价值共同体对于不同的道德机器的需求存在种种现实困境；如何将内置某种道德理论偏向的人工智能产品让消费者接受也是一个社会难题。[7] 近年来，新兴起的一种综合运用贝叶斯推理技术和概率生成模型的研究方法为自上而下式的研究进路带来了曙光。[8]

二是数据驱动进路。这种自下而上的进路要求对智能系统进行一定的道德训练，使其具备类人的道德推理能力，并利用学习算法将使用者的道德偏好投射到智能机器上。从根本上说，这是一种基于进化逻辑的机器技能习得模式。这种研究进路的支持者普遍持有一种道德发生学视角，主张道德能力是在一般性智能的基础上演化而来的，或者道德能力被视为智能的子类而不是高于普通智能的高阶能力。因此，他们是运用与人类的道德演化相似的进路展开研究的。如今，以深度学习（deep learning）为代表的数据驱动进路获得了更多的拥趸。在人工智能道德算法的研究中，道德学习将成为人工智能获得道德判断能力的关键。有学者提出，人工智能可通过阅读和理解故事来“学会”故事所要传达的道德决策模式或价值观，以此来应对各种复杂的道德场景。[9] 105 不过，以深度学习为核心架构的决策算法是基于相关性的概率推理，不是基于因果性的推理，这使得道德推理似乎呈现出一种全新样态。实际上，道德决策算法并不是独立运作的，需要和其他算法系统联合决策。当然，单单依靠学习算法是不够的，甚至是有

严重缺陷的。机器的道德学习严重依赖于训练数据样本和特征值的输入，不稳定的对抗样本重复出现可能导致智能机器被误导，基于虚假的相关性做出道德判断，从而做出错误的道德决策。而且，算法的黑箱特征使得道德算法的决策逻辑缺乏透明性和可解释性。另外，道德学习还会受到使用者价值偏好和道德场景的影响，做出与使用者理性状态下相反的道德判断。例如，人们不自觉的习惯动作往往在理智上是要克服的，而机器学习很难对此做出甄别。从这个意义上说，人工智能既有可能“学好”，也有可能“学坏”。

三是混合式决策进路。自上而下和自下而上的二元划分过于简略，难以应对复杂性带来的种种挑战。混合式的决策模式试图综合两种研究进路，寻找一种更有前景的人工智能道德推理模式。按照当前学界的普遍观点，混合式决策进路是人工智能道德推理的必然趋势，但问题是二者究竟以何种方式结合。我们知道，人类的道德推理能力是先天禀赋和后天学习的共同结果，道德决策是道德推理能力在具体场景中稳定或非稳定的展现。与人类的道德推理不同，人工智能的道德推理能力在很大程度上是人类预制的，但这种能力并不能保证人工智能能够做出合理的道德决策，因为道德决策与具体场景密切相关，而场景又极其复杂和多变。因此，当智能机器遭遇道德规范普适性难题时，到底该如何解决？这不仅是技术专家所面临的难题，更是对共同体价值如何获得一致性的考验。面对具体场景，不同的道德规范可能发生冲突，也可能产生各种道德困境，究竟该以哪种标准优化道德算法，价值参量排序的优先性问题将会变得越来越突出。按照瓦拉赫和艾伦的分析，在众多的候选理论中，德性理论将有可能成为一种最有前景的开发人工智能道德决策能力的模型。[5] 102-109 德性伦理将人们对后果与责任的关注转向对品质和良习的培养，因为后者是好行为的保证，而这种道德良知的获得被认为恰恰需要混合式决策进路来完成。

三、人工智能设计伦理

人工智能设计伦理需要从两个维度展开：一方面，设计和研发某

种人工智能产品及服务之前，设计者或制造商要有明确的价值定位，能够对其所产生的伦理和社会影响有总体预判；另一方面，人工智能在为消费者提供服务的过程中如果出现价值偏差，系统内置的纠偏机制能够有效地防控危害继续发生，防止危险的发生。例如，数据挖掘技术可能将隐藏在数据中的偏见或歧视揭示出来并运用于行动决策中，但机器自身很难像人类一样自觉地抵制一些个人偏见或歧视，这要求通过技术手段和社会手段联合消除这种偏见或歧视。

那么，如何才能设计出符合人类道德规范的人工智能呢？概言之，面对在智力上日趋接近并超越人类的人工智能，设计者要设法赋予其对人类友善的动机，使其具备特定的道德品质，做出合乎道德的行为。人工智能的设计应使它们能够充分发挥特定的功能，同时又遵从人类道德主体的道德规范和价值体系，不逾越法律和道德的底线。但是，人工智能不可能自我演化出道德感，其工具性特征使得人们对它们的利用可能出现偏差。因此，在设计人工智能产品和服务时，尤其应当努力规避潜在地被误用或滥用的可能性。如果某些个人或公司出于私利而设计或研发违背人性之善的自主系统，那么公共政策的制定机构有必要提前对这些误用和滥用行为采取法律与伦理规制。如果具有高度自主性的系统缺乏伦理约束机制和价值一致性，那么其在尚不成熟阶段被开放使用，后果令人担忧。因此，为人工智能系统内置良善的价值标准和控制机制是必要手段，这是保证智能系统获得良知和做出良行的关键。

人工智能应该更好地服务于人类，而不是使人类受制于它，这是人工智能设计的总体价值定位。2016 年，国际电气和电子工程师协会（Institute of Electrical and Electronics Engineers，IEEE）发布《以伦理为基准的设计：在人工智能及自主系统中将人类福祉摆在优先地位的愿景》（第一版），呼吁科研人员在进行人工智能研究时优先考虑伦理问题，技术人员要获得相应的工程设计伦理培训。IEEE 要求优先将增进人类福祉作为算法时代进步的指标。人工智能设计伦理是解决安全问题的必要措施，旨在保证优先发展造福人类的人工智能，避免设计和制造不符合人类价值和利益的人工智能产品及服务。2017 年 12 月

12 日，IEEE《人工智能设计的伦理准则》（第二版）在全球同时发布，进一步完善了对设计者、制造商、使用者和监管者等不同的利益相关方在人工智能的伦理设计方面的总体要求和努力方向。

人工智能的产品制造者和服务提供商在设计与研发人工智能系统时，必须使它们与社会的核心价值体系保持一致。在这方面，IEEE《人工智能设计的伦理准则》（第二版）阐述的人工智能设计“基本原则”为我们提供了很好的启示：第一，人权原则。算法设置应当遵循基本的伦理原则，尊重和保护人权是第一位的，尤其是生命安全权和隐私权等。第二，福祉原则。设计和使用人工智能技术应当优先考虑是否有助于增进人类福祉，是否避免了算法歧视和算法偏见等现象的发生，维护社会公正和良序发展。第三，问责原则。对于设计者和使用者要明确相应的权责分配追责机制，避免相关人员借用技术推卸责任。第四，透明原则。人工智能系统的运转尤其是算法部分要以透明性和可解释性作为基本要求。第五，慎用原则。要将人工智能技术被滥用的风险降到最低。尤其是在人工智能技术被全面推向市场的初期，风险防控机制的设置必须到位，以赢得公众的信任。[10]

人工智能的设计目标是增进人类福祉，使尽可能多的人从中受益，避免造成“数字鸿沟”。人工智能技术的革命性进展可能会改变当前的一些制度设计，但这种改变不能偏离以人为本的发展结构，应当坚守人道主义的发展底线。技术设计不能逾越个人对自身合法权益的控制权，制度和政策的设计应当维护个人数据安全及其在信息社会中的身份特质。如果人工智能系统造成社会危害，那么相应的问责机制应当被有效地调用。实际上，这是要求相关管理部门能够提前制定规则和追责标准，使人工智能系统的决策控制权最终由人类社会共同体掌握。按照阿米塔伊·埃齐奥尼（Amitai Etzioni）和奥伦·埃齐奥尼（Oren Etzioni）的建议，从技术自身的监管角度考虑，应当引入二阶的监督系统作为人工智能系统的监护人程序，以此来避免人工智能技术的潜在风险。[11] 对于社会监管而言，人工智能系统的政府监管机构和相关行业的伦理审查委员会对人工智能系统的使用数据及信息具有调取权，并且能够对系统的安全性和可靠性进行风险评估、测试、审