人工智能与智慧

人工智能遇见中国哲学家

宋冰 X 编著

中国传统**儒释道**
如何看待
现代**人工智能**的未来

赵汀阳、张祥龙、何怀宏、
刘晓力、干春松、陈小平等17位学者
×
聚焦中国和全球语境下的
科技哲思

中信出版集团

人心与智能

智慧

宋冰 X 编著

人工智能
遇见中国哲学家

中信出版集团 | 北京

图书在版编目（CIP）数据

智能与智慧：人工智能遇见中国哲学家 / 宋冰编著
. -- 北京：中信出版社，2020.2
ISBN 978-7-5217-1317-6

Ⅰ. ①智… Ⅱ. ①宋… Ⅲ. ①人工智能—关系—哲学
—研究—中国 Ⅳ. ① TP18 ② B2

中国版本图书馆 CIP 数据核字（2019）第 275731 号

智能与智慧——人工智能遇见中国哲学家

编　　著：宋　冰
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲 4 号富盛大厦 2 座　邮编　100029）
承 印 者：北京通州皇家印刷厂

开　　本：880mm × 1230mm　1/32　　印　　张：13.5　　字　　数：303 千字
版　　次：2020 年 2 月第 1 版　　印　　次：2020 年 2 月第 1 次印刷
广告经营许可证：京朝工商广字第 8087 号
书　　号：ISBN 978-7-5217-1317-6
定　　价：69.00 元

序　言

智能与智慧：人工智能遇见中国哲学家

宋　冰

前沿科技时代基础价值观的错位或缺失

以人工智能、基因编辑、大数据与量子计算为核心的前沿科技时代呼啸而至。人工智能和机器人应用的广泛场景已经开始深刻影响人们的生活方式、理念，甚至人之所以为人的思考。全球呼吁建立人工智能与机器人伦理规范的声音此起彼伏，从2015年至今，全球共有超过50份与人工智能和机器人相关的伦理原则与价值声明。其中有如公正、尊严、自由和人性的抽象原则，也有在个人权利基础上的保护隐私、反偏见等原则，以及诸如可解释性、安全性与鲁棒性等以技术性考量为重的原则。由欧盟人工智能高级别专家组起草的《可信赖的人工智能伦理指导》指出，指导这些原则的基础价值观是“尊重

人权、民主与法治”，且推进个人基本权利的方式方法。[①] 中国、美国和新加坡的人工智能发展战略与规划纲要等各类官方文件所触及的基础价值观包括以人为中心、打造国际竞争优势，保持在研发和规则制定上的战略优势以及保障国家安全等理念。

前沿科技之所以具有颠覆性，在于它技术迭代快、发展进路不确定且呈非线性轨迹，同时各项技术的结合往往产生叠加效应，常常会出现意料之外的结果。通过广泛的应用场景，人工智能与机器人等前沿科技全面渗透我们的个人、社会、经济、政治生活，对我们的生活方式和理念有着令我们始料未及的深刻影响。有的人甚至认为这些科技的发展会威胁人类的存续。然而，目前社会上提出的价值观和伦理原则多半是工业革命时代的主流价值观和伦理原则，这些足以应对如此具有颠覆性、迭代迅速、发展前途扑朔迷离的前沿科技吗？此外，近年来提出的伦理与治理原则大都源于西方文明的价值观，是从人类中心主义出发，围绕个人内在价值、个人的能动性和主体性等哲学假定而提出的。虽然在全球化时代，这些价值观或者表征类似价值认同的符号在许多非西方国家（包括中国）已被广泛采用，但这些原则和理念真的可以成为建构人类和其他存在的共同前途的基础价值观吗？

《人类简史》作者、历史学家尤瓦尔·赫拉利曾指出，我们不仅仅在经历技术上的危机，我们也在经历哲学的危机。现代世界是建立

① 参见 High-level expert group on artificial intelligence, Ethics Guidelines for Trustworthy AI, April 2019。

在 17—18 世纪的关于人类能动性和个人自由意志等理念上的，但这些概念正在面临前所未有的挑战。[①] 闻此评述，笔者心有戚戚焉，认为在讨论符合人工智能和机器人研发的伦理原则之时，应该反思我们习以为常或奉为圭臬的哲学理念和框架。如果用“道”与“术”的框架来看的话，当下关于人工智能伦理原则的讨论更多关注于“术”的层面。本书希望在“道”的层面上引发哲学家和科学家的思考。

人工智能与中西哲学

在西方学术界，人工智能与哲学的结缘始于图灵于 1950 年在哲学杂志《心智》上发表的《计算机器与智能》一文。该文可以被看作人工智能的宣言书，提出了后来众所周知的图灵测试。这个测试的方法和结论引起了西方学术界至今尚未有定论的“何为智能”的讨论。如果说人类对机器能力和智能的分析与预测是人工智能哲学思考的发端，我们甚至可以把这种拷问追溯到 19 世纪中叶洛芙莱斯伯爵夫人对智能机器的畅想。她预测机器可以创造有复杂性的乐章，可以表征大自然，并开启科学史的光荣时代。[②] 此后，有关人工智能哲学的讨论主要围绕核心概念即何为智能而展开。人有可能发展出机器智能吗？没有意向性的能力是智能吗？

如果人工智能要像人一样有思想感情、本能反应、灵感和创造

① 参见 https: //36kr.com/p/5202408。

② Margaret A. Boden, *AI: Its Nature and Future*, Oxford University Press, 2016.

力，那么发展机器意识就成为人工智能科学家努力的方向。科学家们对人的意识的产生及其与身体反应和行动间的关系知之甚少，对如何发展机器智能的认识只是在十分初级的阶段。但是不得不说，近几年对人工智能的热议以及脑科学与认知科学的新发现也引发了人们对哲学领域古老的“心身”问题、何为意识、何为自由意志、意识与神经反应活动的关系的讨论。其他相关问题随之而来。“心”实际上就是一种计算系统吗？电脑可以帮助我们认识人的大脑吗？计算的方法可以用到思维、感情、情绪和动机上去吗？人的常识和日常生活可以还原成计算模型吗？有可能发展机器意识吗？它和人的意识一样吗？人的意识的本质是什么？机器，即使是智能机器可以被认为是生命吗？生命的定义是什么？硅基体没有新陈代谢，是否就永远不可能成为生命？没有生命是否就无法发展意识？

人工智能与人工智能哲学作为学术门类一开始就是跨领域、跨学科的。计算机工程、计算机软件的科学家、数学家，认知科学家、心理学家、心理分析师、语言学家和哲学家等相互启发，在对智能的理解、对人脑规律的揭示、建模等问题上相互推进认识。比如说神经病理学家兼精神分析学家沃伦·麦卡洛和数学家沃尔特·皮茨的文章《神经活动中固有思维的逻辑微运算》就启发了联结主义的人工智能的发展。冯·诺依曼于是开始对思想建模。[①] 英国科学家戴维·马尔从脑科学得到启发，将脑科学的发现与分析用于他的视觉

① Margaret A. Boden, *The Philosophy of Artifical Intelligence*, Oxford University Press, 1990.

和运动控制的建模中。脑科学对人工智能研究启发的最新例子是人工智能企业 DeepMind 受到脑科学在动物记忆重现上的研究成果启发而研发了深度 Q 网络。该企业呼吁并推动了脑科学与人工智能研究团队的融合。刘晓力就人工智能研究领域在人类意识与机器意识的演化、研究进路做了细致而深入的研究。

总之，人工智能科学、工程学、计算心理学、认知科学、脑科学、心灵哲学等都在一个互相推动和融合的学术生态圈中。因此，人工智能哲学被认为是关于智能的科学探讨，是一门与心灵哲学、认知科学、认知哲学、语言学、心理学和认识论等紧密联系的学科。①

本书收集的文章大多不属于以上西方学界人工智能哲学领域的学术探讨，而多是中国哲学家对人工智能与机器人等前沿科技对人类及人类社会的影响的思考，以及中国哲学的观念与思维方式如何应对当下科技提出的挑战、如何就修正和应对当下价值观与伦理讨论有所贡献。但也有两篇文章的作者对智能和意识提出了不同的假设与诠释：张祥龙从时间化能力，即广义的意识能力，来分析深度学习方法是否为初步意识的体现；蔡恒进则提出了触觉大脑假说来解释人的意识与智能的进化，从而对人工智能科学家的研究提供灵感和启发。作为人工智能科学家的陈小平，为应对人工智能需克服不确定性环境的挑战，另辟蹊径，与基于精确性的人工智能经典思维不同，提出了“融差性”思维，以寻求人工智能研究的下一个突

① Margaret A. Boden, *The Philosophy of Artifical Intelligence*, Oxford University Press, 1990.

破口。从事类脑人工智能研究的科学家曾毅则认为，下一步人工智能研究应该重点突破利他行为的计算机制，实现利他行为的智能计算模型，从而在设计建模之初就确保未来的超级人工智能不会对人类的存在造成威胁。

如果要找到西方哲学和中国哲学在人工智能和机器人领域广泛的对话基础，我们也许应转向前沿科技时代人类及后人类如何定义善生活、如何实现善生活的伦理基础等问题。在此，西方的德性伦理学和中国哲学遥相呼应，“惺惺相惜”了。香农·沃伦指出：“德性伦理尤为适合当下所谓的‘技术道德世界’，因为它可以帮助人类避免本质主义的弊病、过度泛化和抽象化的倾向。”“德性伦理适用于正在塑造人类条件的无常多变以及开放式的科技发展。”此外，沃伦认为，德性伦理不把人看成原子个体，而是看成在社会语境下承担具体角色、处于某种关系之中以及对他人和环境承担责任的存在体。[①] 如此的思想基础何其类似于儒释道对人、人伦和社会的论述。本书收集的文章更多的是哲学家们就人工智能和其他前沿科技的发展对人之所以为人的理解、对人机关系以及人机和谐共生的伦理框架的思考。

问题的提出与中国哲学家的初步思考

如前所述，本书希望在“道”的层面引发哲学家和科学家的思考。既然前沿科技有如此颠覆性和深度融蚀的力量，我们是否应

① Shannon Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*, Oxford University Press, 2016.

该在提出具体伦理原则和制定伦理框架之前思考一些更根本性的问题？比如，人与机器的本质性区别到底是什么？人性的核心是什么？机器的本质又是什么？将来的生命形态会如何？人机关系的基本框架和原则是什么？机器在智能方面全面超越人类后，人类会对它们失去控制吗？人机的融合是人的异化还是进化？以人为本的各种哲学、伦理和文化框架对转型后的机器与人类还适用吗？人类社会应该坚持的最核心价值观是什么？如何给人工智能和智能机器人植入伦理观念或给予他们伦理教化？哪些核心观念或价值观应被导入人工智能或使其学习？中国哲学应该怎样发展以适合新时代的挑战？中国哲学如何适应转型后的人类和智能机器（类人）？

自 2017 年年底以来，我们就以上问题在北京大学组织了一系列哲学家和人工智能科学家的对谈和工作坊，在此基础上，我们编撰了本书。每一位哲学家和科学家从他们各自的学科背景回应了上述问题。在此，笔者对他们的初步反思做如下简单梳理。

第一组问题为：人的本质是什么？机器的本质是什么？人工智能和智能机器人的发展如何冲击或补充我们对人和机器本质的理解？

要分析超级人工智能会如何影响人之所以为人的思考，我们首先要了解中国哲学传统对“人”的理解。干春松认为，“儒家从人的社会性来理解人，也就是说，每个人自出生的那天开始，就具备其社会身份”。依据同样的思路，赵汀阳认为，人的关系先于人的本

质。“一个人无法根据自然性把自己定义为人，而必须在与他人的关系中才能被定义。”儒家伦理的基础是血缘关系，政治治理遵循家国同构的理念，于是家庭的角色被扩充到社会与政治身份。

儒家对存在于家庭和社会关系中的人的理解使得儒家格外强调诸如仁、义、礼、智、信之类的道德准则，并认为这些才是人之所以为人的“本质”体现。虽然孟子、荀子分别执有人性本善、人性本恶的观点，但他们在人的教化、强调“学以成人”的理念上是相通的。总之，儒家从人的社会关系和道德性来定义人的本质。因此，干春松认为，儒家的这种以血缘为基础的伦理架构使不少儒家学者排斥任何会威胁家庭和人伦的科技发展，任何挑战血脉相承的发展趋势都会被认为危及儒家的伦理基础。

道家认为宇宙间一切事物（包括人）的本质是气。王蓉蓉认为，“‘气’往往被认为是动态、无所不在、穿透一切、改变一切的力量，能给宇宙中的一切带来活力”。气是生命的基础结构和力量……道是通过气表现出来的。生命就是气的流动。道家对人的理解建立在形、气、神三个元素的协同统一之上。“神”是统摄人体和生命的心理和精神上的力量。人是精气神的统一体，也是性和命的统一体。从这点来看，人工智能或智能机器也许有道的流动性，因为它可以在一个反馈循环中自我调适，但是人工智能缺乏“神”，而“神”才是人不同于人工智能的根本所在。换个角度看，人的智能和人工智能的根本区别或许在于人有所谓的“阴阳智能”。阴阳智能的最大特色就是高度的适应性、开放性和包容性，在此间，事物没有绝对的对立性、分立性和确定性，事物不断融合和流转。人工智能

现阶段的发展离阴阳智能的时代还很遥远。

虽说道家不认为人工智能可以有“神”，从而全面超过人类，盖菲则从道教人士的视角出发，对人工智能发展的前景充满遐想，反诘人工智能与人平起平坐或能力超过人类有何不可，甚至人工智能从某种意义上来看，似乎可以与道教的“仙”比拟。人工智能通过数据或许已达到了长生不老的境界，即人在道教中苦苦追求的境界。若超人工智能出现，也许道教的神仙谱系也会迭代，开拓出数据神仙一派。智能数据也给追求长生不老的人类指明了一个下一步进化的方向。

李四龙从佛教的观点出发，认为生命是轮回的，而人是因缘和合的产物，是六道轮回之中的一种生命形态。因果关系也不是线性的直接联系，因缘条件错综复杂，必然性与偶然性混杂。因此他认为生命没有固定不变的理性。如果人工智能的特性是精确性，人工智能对人类并不可怕。但是他同时也指出，佛教不欢迎有情感的智能机器人。

刘丰河从解决人类生死烦恼的根本智慧出发，对人性与人的本质予以剖析。他指出人们大多根据自己的觉知、思维和行为来判断人是什么，我们周遭的世界又是什么。然而，人们所觉知的人的生老病死和世界万物的变化又源于什么呢？刘丰河认为，对一切问题的认识应从存在（或者说本源、自性）出发，万事万物（包括人）都是基于存在。存在才是显现万事万物的本体。因为存在，人才有了觉知、思维和行为的作用。而存在，过去无始，未来无终，无生无灭，于不变中又能起种种变化的作用。人是存在作用的结果，具有存在性，这就是人的本质。

当代中国哲学家多半善于从东西方哲学比较的视角来审视当下的问题与挑战。在思考前沿科技对人类的影响时亦复如是。李晨阳梳理了东西方思想家几千年来对人性的思考，对强人工智能的出现和其对人类中心论的挑战抱着一种开放的态度。他认为，我们人类一直都在寻找和强调我们作为人类的特殊性，一直在寻找人何以成为万物之尺度、万物之灵。我们居住在宇宙中心，是上帝按照他的形象创造的，我们是理性和有智能的。但所有这些论断不断地被科学、心理学的发展挑战。东方思想家所强调的——如孟子的人性的四端论和荀子的人能“群”的特质，在预期中的强人工智能中也未必不能出现。前沿科技的发展是否终于给人的“自知”，尤其是人类独特性的追求画上了句号？我们或许仅仅是各种存在中的一种，是时候思考如何更好地与其他存在共生共荣了。

赵汀阳在对人的概念的理解上，除了从人际关系来定义人，他还认为，人应该具有理性反思的自我意识，即能够反思自己的行为、价值观和思想的合理性。但是理性人的求知欲和对人的主体性的追求让人类深陷于自我毁灭的悖论。“只要具备技术条件，基因编辑和人工智能的出现都是必然的。正如宗教的知识追求培养了宗教的掘墓人，现代的主体性逻辑也同样培养了主体性的掘墓人。”“只要坚持主体性的概念。那么，基因科学创造的超人或者人工智能创造的超级智能就都是合乎逻辑的结果，而这种结果却很可能是对人类主体性的彻底否定。”

张祥龙认为西方的现象学和古代东方哲学都倾向于将生命内涵的时间性看作是人性的源头。他认为人性的鲜明特点就在深长时间

意识出现的同时展现出来了。相比动物，人具有更长远、更生动的想象力、记忆力和筹划能力。因此，“人类才会生出和珍视那些需要长期时间构造力才能感受到的道德智慧，比如孝悌、忠信、仁爱、慈善、公正与合作”。从现象学、哲学、人类学和东方心学的角度，张祥龙认为，人工智能的新进展即深度学习方法的出现显示出这种智能开始有了初步的时间化的能力，也就是自主学习或逐渐改进自身的能力。他认为，“我们或许应该承认人工智能的智能开始有了意识的萌芽，虽然这意识还十分浅薄。但这个突破是真实的，深度学习的‘深度’（多层裂隙中的交织参数联系）构成了我们以前不敢想象机器数字化能够达到的智能程度，即某种程度上的自主学习能力”。在这个意义上，智能与意识这两条看似不同且隔离的人工智能研究探索的进路其实有深层的内在关联性。

总之，各位学者在人工智能是否会挑战人性的问题上观点不一，道家似乎在这个问题上比较坦然，儒家若从血缘伦理的角度看，对人工智能和人机融合发展的趋势十分抗拒，但从源于天人合一、天地人“三才”思路导出的非人类中心主义的角度看，强于人类的存在或物种的出现也未必不可以接受。从究竟的大智慧来看，刘丰河则认为，只要觉悟了，就会发现存在是无生无灭的。人类应该合理利用人工智能，即便人工智能被用于毁灭人类，觉悟者也无所畏惧。从现象学和东方心学的角度看，人工智能已经发展出意识的萌芽，在此阶段，人类和机器尚有往良善互动的方向发展的可能性。若从西方哲学的主体性的角度看，学者多对强人工智能的发展对人类的冲击忧心忡忡、惶恐不安。

第二组问题为：人机关系的前瞻是什么？人工智能和智能机器对人类社会的影响是什么？人类如何应对？

在人机关系以及人工智能和智能机器给人类社会带来的冲击方面，大多中国哲学家对短期的冲击不是特别担心。这种冲击虽然会是巨大的，但其性质和以往的技术革命带给人类社会的问题类似。赵汀阳认为，“根据以往的经验，技术革命总是导致生产性的工作减少，但同时出现了更多服务性的工作以及知识生产的工作。在人工智能时代也很可能将会出现更多服务性的工作和知识生产的工作来解决失业问题”。对某些儒家学者如安靖如来说，在当下和短期内，现有的人工智能应用甚至可以帮助儒家实现“学以成人”的人文理想，这些应用或许可以帮助人们更好、更快地进行品德修持，更好地实现仁、义、礼、智、信，即儒家五大道德追求。

就长期的时间尺度来看，各位哲学家的想法开始分化。赵汀阳认为人类将失去存在的意义。“技术进步似乎并未给人类获得自由创造机会，反而导致了人的异化。”人们在物质充裕而又无所事事的“幸福”生活中忧郁地退化。而且智能高度发达的时代也更可能使社会极端分化，人们可能生活在一个技术寡头的垄断社会中，“结果可能出现一个高科技的新奴隶制”。在人机关系和将来的发展趋向上，赵汀阳呼吁，“任何智能的危险性都不在其能力，而在于自我意识”。于是科学家应该停止为人工智能植入反思功能和自主创造能力。干春松从儒家的角度也对具有自我意识的人工智能和机器人提出警戒，“无法预测的是，这些已经具有自我意识的机器人可能会‘自行定

义’他们自己的意义与生命目标，从而以他们在生理、脑力上的优势确立其支配地位”。他进一步解释，从儒家的角度来看，基因编辑和人工智能都能带来根本性的观念冲突。首先儒家特别强调血缘，其伦理体系的基础是夫妇、父子，然后是家国天下，如果家庭不是建立在夫妇血缘的基础上，那么儒家伦理体系就会整体崩溃，连续数千年的价值观也就成了无本之木。那么，儒家的秩序理论需要做什么样的调整？还能发展出儒家的人工智能版本吗？儒家的危机感深切。

乐观的学者如贝淡宁则认为，从马克思主义、共产主义高级阶段的角度出发，从长远来看，人工智能似乎可以帮助实现共产主义高级阶段，人类可以从劳作中解脱，从而过上诗意而有创造力的生活。从道教的观点看，人工智能是人类了解世界、解析世界产生的技术之一。对强人工智能乃至超人工智能的出现，人类应该欢欣鼓舞，或许人可以更快地找到超越它们的途径。或许人工智能可以反哺人类，人类即可以通过这种独特的方术，达到如盖菲论述的“与道合一”的境界。

同样比较乐观的李晨阳认为，既然人工智能有智能，智能的存在让我们对它们更容易产生亲近感，那么我们至少可以把它们看作是有一定道德能力的存在，至少是道德施受者。但是人工智能不是生命体，也没有痛苦的体验，所以人工智能还不是道德行为者。在他文章的最后，李晨阳更是对把人工智能纳入儒家的差序格局表示了极大的信心。在“亲亲”“仁民”“爱物”的差序格局中，不同发展阶段的人工智能可以适用不同的角色。因其不同层次的道德成就，

我们人类对其承担的义务也不同。

姚中秋从中西主流思想对强人工智能截然不同的反应入手，分析了中西不同信仰理念造就了中西对人工智能风险的理解之别。一神人格教的基本教义是，神造万物与人，神是全知、全能、全善的，人对神只能是绝对服从。由此，神与人之间是主奴关系，神是主人，人是奴隶。西方人以此框架思考人与人工智能的关系，自然将自己放到了造物主的位子上，而担心人工智能强大后会反抗人类。恐惧由此而来。这种二元视域和零和格局思想充斥于西方关于人工智能风险的讨论。姚中秋认为，中国首先没有存在于物之外、人之外的人格神。自尧舜时代开始崇拜的“天”不是人格化绝对的存在，无其独立之体，更没有绝对意志。万物各自生生于天之中，其中也包括人工智能。当然，自然之物有别于人工物品。我们在开发、利用人工智能时，应秉持“正德、利用、厚生、惟和”(《尚书》) 的原则。

说起人与人工智能的关系，姚中秋从中国哲学的角度，认为人与物同出乾坤父母，故人对物不是绝对支配者，而是“相与者”：彼此虽然不同但出自同源，故可以共存而相互往来而有休戚之情。直白地说，“相与者”即朋友。根据中国人的观念，人应当乐见人工智能之超越性发展，因为人工智能可以凭借其卓越的能力代为承担人对人、人对世界的部分职责。人工智能能力越强，越有利于人类，因为两者共在天之中为“相与者”。

对于人工智能发展自我意识，姚中秋也持乐观态度。一方面他认为自主意识绝不只是理智计算，还有情感。“无情”的人工智能是不可能产生自我意识的。另一方面，如果强人工智能有自主意识，