# 人工智能伦理引论

杜严勇 著

Artificial Intelligence Ethics

人工智能伦理引论，
**使人工智能与人类和谐相处**

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

献给我的父母
杜永茂和李桂珍

# 序：重要的关注

人工智能，现在已经是一个非常热门的概念了。从各种媒体的报道来看，无论是国家政策、科学研究，还是产业开发，都紧密地围绕着这个虽然早已有之但却在近年来越发成为热点的话题，颇有不扯上人工智能就会落后，就会被时代抛弃的感觉。甚至连通俗的网络文学中，都有许多人工智能的主题。连带地，也出现了许多与严格或不严格意义上的人工智能研发应用有关的争议，例如关于人脸识别与监控的广泛应用引发的争议。

其实，关于研究开发人工智能的争议早已存在。在最根本的层面上，这实际上是一个涉及发展观的问题，即人们在强调发展时，是为了什么而发展以及怎样才算是理想的发展的选择的问题。同时，这也是一个科学观的问题，即人们发展科学技术并将之付诸应用在根本上是为了什么，以及对科学技术的不确定性和可能带来的风险的认识的问题。正像本书中作者提到的，2014 年底，英国广播公司报道，著名理论物理学家霍金（Stephen Hawking）曾表示：“人工智能的全面发展可能导致人类的灭绝。”就算没有这样极端，人们至少也无法否认像人工智能的应用会带来的失业等几乎已经初见端倪的后果。

就作序者本人的立场来看，应该说是比较极端地反对像现在这样一窝蜂地极力研发人工智能的。我想，持有类似观点的人也应该还有一些，更有许多人或许是由于对可能带来的问题未加深入思考，或持一种强科学主义的信念而拥护人工智能。正像本书作者所指出的，在我国学者的讨

论中，从事人工智能研究的学者几乎普遍对人工智能的将来持乐观态度，认为目前的人工智能技术不足以威胁人类。而江晓原教授则对之反驳说："我们都知道'养虎遗患'的成语，如果那些养老虎的人跟我们说，老虎还很小，你先让我们养着再说，我们能同意这样的论证吗？"

但无论有什么样的争议，我们毕竟生活在现实的世界中。在这个现实的世界上，存在着国家利益、经济利益的竞争，存在着不可阻挡的要发展更新、更高、更强科技的在观念上和利益上的追求，就像过去许多哪怕人们已经明确认识到本是威胁人类的科学技术应用，如核武器和生化武器之类的发明，也仍然不可阻挡地被发明出来付诸应用。人工智能显然也是一样，在现实中，其发展似乎也无可阻挡。但也正因为如此，面对现实和未来的人工智能发展，对其存在的伦理研究，就有着更加迫切的需要！

就像曾一度在学界流传的一则笑话一样，说不同的院校因不同的传统，对想象中可能要发布的一个研究开发制造椅子的课题持有不同的态度，有的不管三七二十一直接就把椅子造了出来，有的关心这样课题能带来多少篇 SCI 论文的发展，有的转去回顾经典作家关于造椅子的相关论述，更有的要论证：为什么要造椅子？必要性在哪里？本书的研究虽然不是对发展人工智能的必要性的研究，但至少是面对人工智能的发展可能给人类社会带来的最根本的伦理变化的重要研究。

不过，虽然迫切地需要这样的研究，但研究起来又是非常困难的。伦理的研究在整个人文领域的研究中本来就是最为困难的，甚至关于人类的伦理问题，至今也仍然存在着众多的不同立场、观点和理论，更不用说对于人工智能这样新的对象的伦理研究了。这里面既涉及人文，也涉及科学和技术，需要研究者有更广阔的视野和理论准备。在本书中，作者收集、汇总、分析了国内外众多研究者的相关工作（仅此便已经是非常有价值的

研究了），讨论了军用机器人、情侣机器人和助老机器人等几个当下被突出关注的案例，尤其是讨论了其中涉及的传统的与新出现的伦理问题，并在更一般性伦理学背景下，针对人工智能会涉及的伦理问题进行了讨论，其中像对作者比较倾向的自反性伦理治理观点的讨论，对友好人工智能概念的问题的讨论，都是很有特色的。而且，作者的一些结论性的观点，也颇有深意，如："人工智能安全问题从源头上看是由人工智能技术造成的，这应该使科学家认识到，科学技术研究并不是无禁区的，我们需要理性地发展人工智能。""科学家应该意识到，科学与伦理有着不同的选择与评价标准，在进行机器人设计时，应该将科学与伦理结合起来，而不只是在科学的范围内从事研究。"

当然，作者的一些观点本人也并不是完全都能同意，此书也还存在着一些不足（如在内容和结构上并未明确区分人工智能和机器人的不同，尽管这两者间存在着密切的关联），但这些并不影响此研究的重要意义。毕竟这还只是国内学者对此重要领域和问题进行的开创性的系统研究，而对于人工智能的伦理研究将需要有更多、更深入、更细化的后续工作。而书中提出和讨论的问题，对于后续的研究者显然具有启发性的意义，仅此，应该说，就已经是非常有价值的著作了。

是为序。

刘　兵

2020 年 1 月 26 日于北京清华园荷清苑

# 目　录

# 导　论

## 一、时代背景与研究意义

目前，学术界关于人工智能与机器人的研究如日中天，兴盛异常。各地纷纷成立关于人工智能的研究机构，许多企业竞相加大对人工智能的投入力度，各国政府相继发布相关战略规划，唯恐在这场科技竞争中处于下风。同时，随着计算机、人工智能与机器人学等科学技术的快速发展，机器人与其他各种人工智能产品越来越多地进入公众的视野当中。机器人不仅在工业、农业、军事、医疗等领域得到广泛应用，而且在家庭服务、社会娱乐等领域也得到了人们越来越多的关注。

从世界的角度看，美国、日本、欧盟等竞相加大对机器人产业的投资力度，强调多方合作，共同推进机器人技术与产业快速发展。比如，美国专门制订了国家机器人发展计划（National Robotics Initiative），目的就是为了促进美国机器人研究与应用，国家科学基金会、国家航空航天局、国家卫生研究院以及农业部等联邦政府部门共同资助机器人发展计划。[①] 2004 年，欧洲机器人研究网络（European Robotics Research Network）出台了欧洲机器人研究路线图，描述了欧洲机器人技术发展的重点领域，并强调要在机器人技术的市场竞争中取得领导地位。[②] 众所周知，日本、韩国也大力发展机器人技术与产业，日本还经常被誉为是“机器人大国”。

我国政府同样高度重视机器人技术的研发。2012 年 4 月，科技部专门制订了《智能制造科技发展“十二五”专项规划》和《服务机器人科技发展“十二五”专项规划》，提出“十二五”期间将重点培育发展工业和服

---

① “National Robotics Initiative,” http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503641&org=CISE.

② “Euron Research Roadmap,” http://www.cas.kth.se/euron/euron-deliverables/ka1-3-Roadmap.pdf.

务机器人新兴产业。2014 年 6 月 9 日，习近平总书记在两院院士大会上的讲话中指出，“机器人革命”有望成为“第三次工业革命”的一个切入点和重要增长点，将影响全球制造业格局，而且我国将成为全球最大的机器人市场。2015 年 11 月，世界机器人大会在北京召开，习近平总书记致信祝贺，李克强总理做出批示，充分显示了国家领导人对机器人技术与产业的高度重视。

我国政府在 2017 年 7 月发布的《新一代人工智能发展规划》中提出，到 2020 年，我国人工智能总体技术和应用与世界先进水平同步，人工智能核心产业规模超过 1 500 亿元，带动相关产业规模超过 1 万亿元；到 2030 年，人工智能理论、技术与应用总体达到世界先进水平，人工智能核心产业规模超过 1 万亿元，带动相关产业规模超过 10 万亿元。①

在世界各国普遍重视机器人产业的历史背景下，考虑到机器人技术可能产生的深远社会影响，将 21 世纪称为“机器人（人工智能）世纪”可能并非夸大其辞。但是，在机器人越来越聪明、使用越来越广泛的时代背景下，人们难免会对机器人可能导致的负面效应感到忧虑。诺贝尔和平奖获得者、核物理学家罗特布拉特（Joseph Rotblat）指出，会思考的计算机、机器人拥有人工智能，使它们可以复制自身，而不加控制的自我复制是这些新技术具有的危险之一。②

与机器人技术一样，飞速发展的现代科学技术在带给我们种种益处与便利的同时，也引发了形形色色的伦理问题。于是，生命伦理、网络伦理、核伦理、信息伦理等种种科技伦理如雨后春笋般涌现。相对而言，机

① 《新一代人工智能发展规划》，北京：人民出版社，2017，第 8—10 页。

② Veruggio Gianmarco, “The Birth of Roboethics,” http://www.researchgate.net/publication/228623299_The_birth_of_roboethics.

器人伦理出现较晚。2004 年 1 月，第一届机器人伦理学国际研讨会在意大利圣雷莫召开，正式提出了“机器人伦理学”（roboethics）这个术语。机器人伦理学研究涉及许多领域，包括机器人学、计算机科学、人工智能、哲学、伦理学、神学、生物学、生理学、认知科学、神经学、法学、社会学、心理学以及工业设计等。2005 年，“欧洲机器人研究网络”专门资助研究人员进行机器人伦理学研究，希望能够为机器人伦理研究设计出路线图。[①] 此后，机器人伦理研究很快得到越来越多的学者的关注。

虽然有学者认为当前的人工智能热不可能持续很长时间，不过人们普遍认为，人工智能将会对人类社会产生深远影响。那么，人工智能究竟应该向何处去？在人类社会深度科技化的历史背景中，想要阻止人工智能的快速发展几乎是不可能的，更为现实的做法是为人工智能的发展进行伦理规制，由此彰显了机器人与人工智能伦理研究的重要性。

我们可以从学术价值与实践意义两个角度来看人工智能伦理研究的重要意义。从学术价值的角度看，人工智能伦理开拓了科技伦理研究的新领域，并且对我们加深与变革——包括人、生命、智能与机器等基本哲学概念——的理解与认识具有重要意义。从实践意义的角度看，对于社会大众来说，人工智能伦理研究有助于建构起人类与智能产品互动的道德观念与道德规范，实现人与智能产品的和谐相处。对人工智能技术的研发人员来说，人工智能伦理设计原则与方法的研究成果，可以为其提供技术研发的伦理依据与理论支撑；而且，人工智能伦理研究也有助于加强科技人员的道德责任感，培养科技人员的道德想象力与实践能力。对科技管理人员来

① Veruggio Gianmarco and Operto Fiorella, “Roboethics: a Bottom-up Interdisciplinary Discourse in the Field of Applied Ethics in Robotics,” *International Review of Information Ethics*, 2006, Vol.6, No.12, pp.2−8.

说，人工智能伦理研究成果可以为相关产业政策的制定提供理论参考。

## 二、研究现状

首先对“机器人”“人工智能”这两个概念做出简单的界定。一般认为，“机器人”（robot）这个词最初出现在捷克作家恰佩克（Karel Capek）于1921年写的剧本《罗素姆的全能机械人》中，源于捷克（或斯洛伐克）单词“robota”，其意义为“努力工作”或“奴役”。现在机器人一般指在很大程度上拥有人类特征的机械。现在广泛使用的关于机器人的术语有安卓机器人、拟人化机器人、自动机器人、半机械人、人类辅助设备、人形机器人、仿人机器人，等等。[①] 本书中采取一种比较宽泛的机器人界定，主要指拥有一定智能，并且拥有人类（或动物）外观甚至外在表情的机器。

与机器人类似，人们对人工智能的定义也变化不定，没有形成共识。李开复、王咏刚认为，人们对人工智能有五种定义，定义一：人工智能就是让人觉得不可思议的计算机程序；定义二：人工智能就是与人类思考方式相似的计算机程序；定义三：人工智能就是与人类行为相似的计算机程序；定义四：人工智能就是会学习的计算机程序；定义五：人工智能就是根据对环境的感知，做出合理的行动，并获得最大收益的计算机程序。[②] 博登（Margaret Boden）认为，人工智能就是让计算机完成人类心智（mind）能做的各种事情。[③] 通常还有强、弱人工智能以及超级人工智能的区分。弱人工智能一般指可以模拟或实现人类智能部分功能的人工智能，

① 巴-科恩、汉森：《机器人革命》，潘俊译，北京：机械工业出版社，2015，第3—8页。
② 李开复、王咏刚：《人工智能》，北京：文化发展出版社，2017，第1—37页。
③ 博登：《人工智能的本质与未来》，孙诗惠译，北京：中国人民大学出版社，2017，第3页。

强人工智能则指可以实现人类智能所具有的大多数甚至全部功能的人工智能，超级智能则是可以完全超越人类智能的人工智能。本书中提及的人工智能概念一般来说包含了强、弱人工智能两个层面，一般不涉及超级人工智能方面。

一般认为，机器人是人工智能研究的一个分支学科，许多讨论人工智能的著作与论文中都有涉及机器人的内容。这两个概念当然是有区别的，但从科技伦理研究的角度看，实际上机器人伦理与人工智能伦理并无实质性区别，基本的观点、理论与方法几乎是通用的。所以，本书并不严格区别这两个概念，为行文或讨论内容的方便，有时单独使用，有时并列使用。

关于机器人与人工智能伦理的思考很早就开始了，但真正引起学术界和公众的广泛关注是最近十余年的事。2005 年，欧洲机器人研究网络资助成立了机器人伦理学研究室（Euron Roboethics Atelier），该研究室于 2006 年 7 月出台了第一个机器人伦理路线图。路线图对机器人研发中涉及的伦理问题进行了系统的评估，以促进跨学科的深入研究。不过，这种路线图不是学术研究，其目的不是为科技研究提供伦理指南，不是简单罗列问题与答案，也不是原则性的宣言，而是根据机器人领域发展现状及其可能的发展情况，为机器人的设计者、制造商以及使用者提供某些伦理分析与建议。① 受此激励，日本、韩国等国家也纷纷着手制定关于机器人设计、使用的法律法规与指导路线。

目前，在西方学者出版的关于机器人与人工智能伦理研究的著作中，比较有代表性的主要有以下几部：美国耶鲁大学伦理学家瓦拉赫（Wendell Wallach）等的《道德机器》从理论上讨论了具有伦理判断能力的机器人

① Veruggio Gianmarco, “The EURON Roboethics Roadmap,” http://www.roboethics.org/atelier2006/docs/ROBOETHICS%20ROADMAP%20Rel2.1.1.pdf.

的必要性、可能性以及实现方法；[①] 美国佐治亚理工学院的机器人专家阿金（Ronald Arkin）的《控制自主机器人的致命行为》考察了军用机器人引发的伦理问题及可能的解决途径。[②] 美国哈特福特大学计算机专家安德森（Michael Anderson）等人主编的论文集《机器伦理》收集了 31 篇论文，从机器伦理的性质、重要性、问题域、实现路径以及前景等五个方面进行了论述；[③] 美国加州州立理工大学的林（Patrick Lin）等人主编的论文集《机器人伦理》收集了 22 篇论文，讨论了机器人的社会影响及其引发的部分伦理问题。[④] 另外，许多英文期刊如《IEEE 智能系统》（*IEEE Intelligent Systems*）、《信息技术与伦理》（*Ethics and Information Technology*）、《人工智能与社会》（*AI & Society*）等也不定期地刊载有关的论文，或者出版专刊。比如，英文期刊《信息技术与伦理》在 2010 年、2012 年、2013 年、2016 年、2017 年和 2018 年开辟专刊讨论机器人与人工智能伦理问题，西方学术界对相关研究之关注可见一斑。

我国学者主要从 2013 年开始发表机器人与人工智能伦理的研究成果。中国知网上能够找到的较早的学位论文是武汉理工大学李俊平的硕士论文《人工智能技术的伦理问题及其对策研究》[⑤]，较早的博士论文则是南开大学王东浩的《机器人伦理问题研究》[⑥]，不过以硕士论文为主，博士论文比较少见。较早发表的机器人伦理研究论文的是王绍源、赵君的《“物

---

① Wallach Wendell and Allen Colin, *Moral Machines: Teaching Robots Right from Wrong* (Oxford: Oxford University Press, 2009). 中译本见瓦拉赫、艾伦：《道德机器：如何让机器人明辨是非》，王小红主译，北京：北京大学出版社，2017。

② Arkin Ronald, *Governing Lethal Behavior in Autonomous Robots* (Boca Raton: CRC Press, 2009).

③ Anderson Michael and Anderson Susan edited. *Machine Ethics* (Cambridge: Cambridge University Press, 2011).

④ Lin Patrick, et al edited, *Robot Ethics* (Cambridge: The MIT Press, 2012).

⑤ 李俊平：《人工智能技术的伦理问题及其对策研究》，武汉理工大学，2013。

⑥ 王东浩：《机器人伦理问题研究》，南开大学，2014 年。