可否认，机器人伦理学反思会用到哲学、经济学、社会学、管理学、统计学、心理学、文化研究等诸多学科的理论、方法和案例。换言之，机器人伦理学并不属于伦理学的分支，而是属于整个哲学的应用分支。从某种意义上说，机器人拟人论对机器人研究、设计、制造和应用的支配，催生了最主要的机器人伦理学问题，机器人伦理学研究应想方设法解决这些问题，实现“机器人伦理软着陆”，使机器人产业能够健康稳健地发展。

第四，机器人伦理学应具备前瞻性、情境性、动态性。机器人伦理学通常落后于机器人技术的发展。① 面对机器人拟人化引发的不确定性与极大社会影响，伦理先行尤为关键。在机器人研发阶段，应始终贯穿伦理考量，进行伦理审查与法规制定，主动预测潜在风险，兼顾短期与长期伦理问题。此外，机器人相关伦理问题具有较强的情境依赖性。例如，医疗机器人更多涉及隐私、决策等伦理问题，而工业机器人主要与失业、安全等伦理问题相关。机器人在不同应用场景涉及的伦理规范、价值观和社会影响可能大相径庭，因此机器人伦理学需要针对具体情境进行深入探讨。机器人伦理学的框架需要根据最新的技术成果和社会需求做出相应的调整，积极回应技术创新带来的伦理挑战，建立实时反馈机制，确保灵活性与适应性。

如何将机器人安放于人类社会之中？*

——一种哲学进路的思考

刘　鹏

机器人小朵是儿童科幻小说《卧底机器人》中的主要角色，她经历了一场时间更久、难度更高的“图灵测试”，需要在人类的学校中“生活”一整个学年

① 参见［美］帕特里克·林、凯斯·阿布尼、乔治·A. 贝基主编：《机器人伦理学》，薛少华、仵婷译，人民邮电出版社 2021 年版，第 13 页。

* 本文系中国科学院学部科技伦理研究项目“数字技术的伦理研究”（XBKJLL2023001）、江苏省社会科学基金重大项目“新一代人工智能重大哲学与逻辑问题研究”（24ZD006）的阶段性成果。

而不被发现真实身份，并最终取得了成功。小说用几句话就能解决可以困扰人类数十年、数百年的难题——小朵下载一个“爱情补丁”，就可以感受到痛苦、嫉妒等情感，而且这种情感是内生性的，是基于算法产生的非算法反应。正是凭借这种“主动性”，小朵通过与人类的互动，逐渐找到其在寄居家庭、学校和社会中的位置。于是，人与机器人各得其所，一种新的社会秩序得以确立。①

正如小朵的故事展示的，当任何一种新技术出现时，人类都必须为之在社会中找到一个合适的位置。这一寻找的过程不仅意味着给予技术一处可以安身的物理空间，而且意味着在技术与人类社会协同演进的过程中实现对技术和人类社会的持续性调整甚至重构。然而，因为目前的技术包括机器人技术都缺少科幻小说赋予小朵的这种主动性，所以无法指望技术本身能够完成这类重构性的工作，担负责任的只能是人类自身。

一、虚无与方向：新技术给人类带来了什么？

科学和技术在创造新世界的同时，也在破除着旧世界，而且这种破除伴随着旧秩序的逐渐解体，正是在此意义上，比利时哲学家基尔伯特·奥托瓦（Gilbert Hottois）创造了“技性科学”（technoscience）一词。在奥托瓦看来，技性科学终结一切：它并不是在“沉思”外部世界，而只是关注世界“在技术宇宙中的功能性和可操控性”，打破边界和重构边界成为其核心任务，这就带来了传统本体论的终结；技性科学的创造性打开了各种可能性，继之而来的则是一种极度开放而又极度模糊的未来，前者来自其操作性蕴含的力量，后者则代表着意义的丧失，于是，力量的彰显和意义的退场成为同一过程的两个侧面；技性科学打破的不仅是物的边界，也是人的边界，它不断改变着对人类生死、情感、精神等层面的界定，甚至也在改变着对人本身的界定（如“赛博格”），于是，传统意义上的人也被终结了；既然物与物、人与物的边界被破除，可能性取代了确定性，开放性代替了目的性，那么，传统的伦理学由于“太过人性”，无法应对这种可能性和开放性，故而也就不再适用了。②

① 参见［英］戴维·埃德蒙兹、休·弗雷泽：《卧底机器人》，李玮译，中信出版社2021年版。

② 参见 Gilbert Hottois, Technoscience: Nihilistic Power versus a New Ethical Consciousness, in Paul T. Durbin, ed., *Technology and Responsibility*, Dordrecht: D. Reidel Publishing Company, 1987, pp.72−84; Gilbert Hottois, Definir la Bioethique: Retour aux Sources, *Revista Colombiana de Bioética,* Vol.6(2), 2011, p.103。

机器人技术是技性科学的代表，如今已被广泛应用于人类的生产和生活领域，进入家庭也不再是幻想。当然，在机器人技术能否像电力技术一样给人类社会带来一场彻底的变革这一问题上，仿照布鲁诺·拉图尔（Bruno Latour）《法国的巴斯德化》一书的标题，我们可以说世界的电力化，但对世界的机器人化（并不是说人类被机器人取代，而是说机器人彻底重构整个人类社会），现在尚言之过早。不过，机器人技术在与人工智能紧密结合后带来的冲击是显而易见的，这是因为与传统技术相比，智能机器人的学习能力、自主能力、适应能力更强，不论对管理者还是使用者来说都具有更大的不确定性。在此意义上，与一般的技性科学相比，人工智能与机器人技术带来的终结更加彻底。① 特别是当机器人进入日常生活后，人们首先会对之产生好奇，其后则可能产生恐慌。

面对担忧与恐慌，以本质主义的人性观回避机器人技术的发展不过是一种鸵鸟策略。从人类科技史和产业史的发展历程看，人类对技术的忧虑一直存在，就如同面对火车的出现，有人会说“这对上帝和我都是一件同样不愉快的事情”。② 面对今天的人工智能和机器人，忧虑之声更甚。这类恐慌尽管表现各异、原因不同，但究其本质，可以分为以下几个方面：对人类生存权的恐慌，如对机器人的使用可能带来的失业问题以及从长远来看的物种替代的忧虑；对人类价值独特性的恐慌，如人形机器人引发的“恐惑谷”效应（尽管人形机器人与机器狗在本质上并无多大差别，但它们给人类带来的情感体验和心理感受大相径庭）以及机器人对原初独属于人类的某些角色的替代导致的价值丧失感；对技术失控的恐慌，如机器人技术本身的不确定性尤其是机器人自主决策能力的提升可能带来的安全恐慌；对社会失序的恐慌，例如人们可能会忧虑资本对技术的滥用，而这种忧虑又在自媒体的语境中被不断放大，此外，数据权利、隐私权等方面的问题也可能引发人们的担忧。当上述恐慌与传统社会结构相耦合，又会产生关于公正公平、群体歧视、公共安全等方面的忧虑。

① 尽管“长期来看，人类与机器、线上与线下、虚拟与真实等方面的区别，都会逐渐消失”之类的观点似乎有些骇人听闻，但至少边界的模糊已是不争的事实。参见［英］杰米·萨斯坎德：《算法的力量：人类如何共同生存？》，李大白译，北京日报出版社 2022 年版，“导论”，第 ii 页。

② ［英］斯诺：《两种文化》，陈克艰、秦小虎译，上海科学技术出版社 2003 年版，第 21 页。

要有效应对这些恐慌，就必须认识到，机器人进入人类社会，并不仅仅是为人类社会增添了一种新的技术物，而且是增添了一种新的社会角色。这种角色的融入需要社会进行全方位的动态重构，这种重构的最终结果应该是一个容纳了机器人、内部呈现有机结构状态的新社会。要推进这种重构，就必须认清机器人技术可能给人类带来的一系列伦理风险及其根源，而后再采取行动。

二、愿景与悖论：机器人伦理风险的发生机制

技术应服务于人，这是人们的共识。对机器人技术来说同样如此。然而，初衷和目标并不等同于结果，更不能消除在目标达成过程中可能经历的各种波折。要减少乃至规避机器人伦理风险的发生，就必须深刻认识这些伦理风险的发生机制。总括而言，此类风险的发生机制可归结为如下四对矛盾。

第一，认识论困境，即技术的完备性与技术落地应用的急迫性之间的矛盾。英国技术哲学家大卫·科林格里奇（David Collingridge）指出了技术应用的一个困境：技术的社会后果在技术发展的早期阶段难以预料，而当某些负面的后果出现时，技术已经成为经济和社会结构的一部分，对技术的控制将变得异常困难。这进而就产生了一个难题："当改变轻而易举时，改变的必要性却不甚明了；当改变的必要性显而易见时，改变却又成了一项耗钱、耗力、耗时的事情。"① 这一困境的根源就在于技术的完备性与技术应用的急迫性之间的矛盾。对机器人来说同样如此，公众对获得机器人服务的渴望、企业对机器人产业化的需求以及国家间机器人产业的竞争，使应用机器人的愿望日益迫切；而机器人装配的智能系统可能会出现幻觉、虚构等问题，其做出的决策建议可能会面临情境适配等难题，特别地，其存在安全漏洞进而遭受黑客入侵可能会带来严重后果。这就导致了技术的完备性与应用急迫性之间的矛盾。

第二，规范落地困境，即对道德机器的美好愿望与道德算法的能力限度之间的矛盾。在医疗辅助、工业制造、家庭陪护等方面，机器人已经开始发挥作用。然而，机器人的行为一旦涉及道德判断，它该如何做出道德选择？事实上，在真正的机器人被应用于工业生产之前，阿西莫夫就在其小说中提出"机器人三定律"，并将之作为规范机器人行为的根本道德准则，此后作为

① 参见 David Collingridge, *The Social Control of Technology*, London: Frances Pinter, 1980, p.19。

补充又提出了“第零定律”，这些定律在小说中当然发挥了非常好的道德约束作用。然而，小说终归是小说，小朵与人类尽管存在身体构成上的差异，但最终实现了“硅基道德”与“碳基道德”的共存，而这一过程中被快速跳过的部分恰恰是现实中的最大难题。该如何为机器人赋予道德呢？只要我们判断机器人“是否拥有智能的唯一方法是通过其输出或行为进行判断”，① 亦即只要机器人不存在内在的道德自生机制，那么，我们就只能将人类的道德规范转变为机器人的算法，并试图以此塑造一种“道德机器”，② 进而实现“价值对齐”。这一方案看起来非常吸引人，但也会面临向谁对齐、如何对齐的难题，前者在同一价值共同体内会面临价值整体性和代表性的困难，在不同共同体间则会面临价值差异性的难题，后者则需应对规则的算法有限解与道德情境的无限多样性之间的矛盾。

第三，权责分配困境，即机器人的行动能力与权利主体及责任主体难以匹配的矛盾。与传统技术相比，机器人具有自主或半自主工作的能力，这使行动与权利的关系变得更加复杂。例如，机器人在数据收集与处理之后才能做出决策，从而对环境做出反应，这些数据并不全然是以规则形式呈现的指令，也可能是通过“感知”系统获得的，这就涉及数据权利和数据隐私的问题，甚至在很多时候会使人们陷入“不让渡则不能使用”的困境。同样，机器人的应用也会面临责任分配的难题。在电影《2001：太空漫游》中，机器人哈尔 9000 最后杀死了人类宇航员，哈尔该对此负责吗？在机器人已经成为现实的今天，尽管人们通常将自由意志等作为责任主体的基本特征，但只要机器人还感受不到“痛苦”，惩罚也就毫无意义。由此，机器人的权与责及机器人的设计者、管理者和使用者之间的权责关系，同样挑战了传统伦理关系的责任分配模式。

第四，角色实现困境，即机器人的服务角色与人类的主体角色之间的矛盾。技术在本质上通过对人的部分能力的替代或放大，从而实现服务于人的目的。然而，当这种替代或放大发展至一定程度时，可能会导致对人本身的某些角色的替代。此类故事在历史上不断上演，尽管人们每次在面临技术革新时都会喜忧参半，但机器人引发的替代恐慌是史无前例的。从微观层面看，

① ［美］杰夫·霍金斯、桑德拉·布莱克斯利：《新机器智能》，廖璐、陆玉晨译，浙江教育出版社 2022 年版，第 14 页。

② 参见［美］温德尔·瓦拉赫、科林·艾伦：《道德机器：如何让机器人明辨是非》，王小红主译，北京大学出版社 2017 年版。

机器人的算法程序在某种程度上可能会蕴含“脚本”特征。[①] 这里的“脚本”并非技术术语，而是一个政治术语，它能够以“规约”的方式预设人类的行动方式，由此人们可能会被算法引导乃至“操控”，甚至可以说，“代码就是力量”。[②] 在此意义上，服务与被服务的关系发生了颠倒，服务者与被服务者的身份也就发生了对调，于是，古老的政治关系在新技术条件下呈现了新的权力结构。从更长远的尺度看，人们的替代性担忧可能会加深至物种替代的程度。尽管学者们对此看法不一，但许多人认为物种替代不过是个时间问题，[③] 就如同《罗素姆万能机器人》的最终结局一样。

三、分立与杂合：人类社会该如何安放机器人？

机器的引入虽引发了卢德运动的反击，但机器最终得以安放于人类社会之中；面对机器人的出现，无视与反对同样不可取。人与技术并非本性对立，亦非天生一致，人与技术和谐关系的达成需要人类付出努力。同样地，机器人在如小朵一样获得自主意识之前也不能完成“认知自己”的哲学任务，因此，认识机器人这一任务的承担者只能是人类，只有认识到这一点，才可能构建出一种和谐的人机关系。要完成这种构建，须注意如下四个方面。

第一，认识论上的谦逊性。从历史来看，科学与技术的确定性只能是一种理想，不确定性才是现实。这就要求我们保持对科学的谦逊态度。牛顿在对科学研究方法的讨论中，明确意识到了归纳问题的存在，[④] 但科学和哲学的一个差别就是，哲学总想在找到一个确定、可靠的起点后才开始行动，科学则在行动中不断寻找新的起点。在此意义上，牛顿只是要求我们保持一种

① 参见 Madeleine Akrich and Bruno Latour, A Summary of a Convenient Vocabulary for the Semiotics of Human and Nonhuman Assemblies, in Wiebe E. Bijker and John Law, eds., *Shaping Technology / Building Society: Studies in Sociotechnical Change*, Cambridge, Massachusetts: MIT Press, 1992, pp.259−264。

② ［英］杰米·萨斯坎德：《算法的力量：人类如何共同生存？》，李大白译，北京日报出版社 2022 年版，第 117 页。

③ 即便承认按照当前的技术进路完全的物种替代是不可能的，对替代的恐慌也是真实的。

④ 如牛顿所言，“在实验哲学中，由现象通过归纳推得的命题，在其他现象使这些命题更为精确或者出现例外之前……应被认为是完全真实的，或者是非常接近于真实的”。参见［英］牛顿：《自然哲学的数学原理》，赵振江译，商务印书馆 2011 年版，第 478 页。

认识论上的谦逊，如此才能获得一种更加开放的态度，最终才可能达至一种更优的科学。对待机器人和人工智能同样应该如此。美国通用电气公司通过设想一场有关机器人的噩梦，提出了“谦逊 AI”（humble AI）的呼吁：“这对你而言简直就是一场噩梦：被赋予了人工智能的各种机器，变得比其创造者还要聪明，它们接管人类并试图将人类从自身中拯救出来。呜呼哀哉！直至最后，它们才发现聪明并非智慧。面对苏格拉底、亚里士多德以及阿尔伯特·爱因斯坦说过的那句话‘知之愈多，愈知己之无知’，机器嗤之以鼻。机器一意孤行，最终使文明在大家头顶轰然倒塌。”① 机器人的设计者要保持谦逊，以一种开放的态度制造出更加完美的产品；使用者要以一种反思性的态度对待机器人，对机器人的能力边界始终保持认识论上的敏感。特别需要注意的是，机器人的决策是依据一定的数据做出的，但决策过程可能存在不可解释性和不透明性，这就需要对具体决策的情境适用性做出判断。

第二，价值观上的包容性。机器人有价值观吗？至少目前尚无人类意义上的内在价值观。从技术层面来说，机器人的价值观不过是各种价值规范的数字形态，或者说，它实际上是人类设定的一个规则集合。就此而言，伦理意义上的善恶问题便被转换为对错问题。这进而会衍生三个子问题。一是规则设计问题，机器人遵循的价值规则被转换为设计者的规则，在此意义上，设计者需要秉承一种开放包容的价值立场，考虑到不同文化、不同群体在群体特征和偏好等方面的差异，避免现实社会中的不公正现象延伸到机器人的“意识”中。二是规则遵循问题，规则即便是合理的，也会面临遵循难题。一方面，正如维特根斯坦指出的，即便是单条规则在执行过程中仍然会存在极度开放的认知偏差，这种偏差一旦落实为决策可能会使机器人采取全然不同的行动；另一方面，如果采取增加规则的方式来缩小规则的外延范围，那么就会进一步产生在具体场景中不同规则之间的相容问题，就如同机器人三定律仍然面临执行困境一样。面对后一类问题，设计者需要强化规则的场景敏感性，虽然场景难以穷尽，但就如图灵指出的，人工智能也需要像人一样经历不断学习的过程，尽管就目前的技术进路而言两者的学习机制仍是不同的。三是数据匹配问题，即便规则是合理的，也是可遵循的，仍可能面临数据来源与数据应用的匹配问题。这类问题常常会表现出群体歧视的特征，这种歧

① Judgment Call: Why GE Is Experimenting with “Humble AI”, https://www.ge.com/news/taxonomy/term/8480, 2025 年 3 月 14 日。

视产生于作为数据来源的群体与作为数据应用对象的群体之间的现实差异。例如，医疗机器人的训练数据如果都来自年轻人，那么将这些数据运用于老年人时很可能会发生误诊。由此，事实性的数据由于应用范围超出了其最初的空间边界，故而带来了价值性的后果。要想将价值负面后果转变为价值正面后果，一方面可以通过限定机器人的应用边界，从而减少数据应用扩展的范围，另一方面可以通过增强数据的代表性，从而增强机器人判断与决策的包容性。由此可见，机器人的“心灵”设计是一项高度社会性的工作，在此意义上，“软件工程师将越来越多地成为数字生活世界的社会工程师”。①

第三，治理观上的协同性。一方面，与传统技术一样，机器人的功能实现尽管具有开放性，但其可能范围与设计阶段密不可分；另一方面，传统技术的使用往往发生在封闭场景中，换句话说，互动仅仅发生于在场的人与物之间，而机器人则是直接或间接存在于一个更大的技术网络中，这就使人机互动的空间边界发生了扩展，甚至可以潜在地扩展到全球范围。在此意义上，机器人的治理更加需要多元主体的参与，是“健全多方参与、协同共治的科技伦理治理体制机制”② 的重要组成部分。这种协同不仅发生在宏观领域如政策与行业规范的制定中，而且发生在机器人应用的微观场景中。例如，必须充分重视用户的反馈。传统观点往往将普通公众视为认识论上的无知者，进而忽视其知识的认识论地位，实际上，公众是技术使用的直接参与者，他们在使用过程中形成的知识尽管可能被贴上地方性的标签，但仍然具有重要的反馈价值，对改进机器人在具体场景中的应用能力非常重要。再如，特别是在具备互动功能的社交机器人领域，应以外力强化企业的伦理责任，这是因为企业强化伦理责任的内在动力是不足的。显然，高度负责的社交机器人很可能会被用户视为喜欢“说教”而被弃用，于是，道德底线越低的机器人可能越会受到某些用户的欢迎，这就是为什么各种形式的“AI 聊天致死”“AI 怂恿杀人”事件时有发生。因此，必须协调企业在“义”与“利”之间的均衡，进而推动企业合理塑造机器人的“义利观”。

第四，伦理观上的开放性。在传统上，伦理学研究的是人与人之间的关系。然而，在人类的个人生活和社会生活都已经被技术深度渗透的今天，必

① ［英］杰米·萨斯坎德：《算法的力量：人类如何共同生存？》，李大白译，北京日报出版社 2022 年版，第 244 页。

② 《中共中央办公厅 国务院办公厅印发〈关于加强科技伦理治理的意见〉》，https://www.gov.cn/zhengce/202212/content_6688372.htm，2025 年 3 月 14 日。

须将技术物纳入伦理学的视野。学界常用非人类中心主义称谓此种伦理学立场。这里的转变不仅仅涉及认识论和价值观，实际上已经涉及本体论层面。人们对机器人是否可以作为道德主体一直争议不断，但从人机关系的角度看，人机双方一直处于互构状态中。人类既是机器人的设计者，也是机器人现实行动的反馈者，因此人对机器人的影响不言而喻，而机器人同样在重构着人，这种重构不仅发生在行为层面，而且发生在认知层面，甚至会潜移默化地发生在价值层面。可以说，“人类正在适应机器，就像机器正在适应人类一样”。① 一些调查发现，人类开始对机器人甚至机器狗做出情感反应，甚至与类人机器人建立了非常私人的关系。② 在此意义上，人类不再是有着封闭边界的传统主体，机器人也不再是惰性的客体，两者在人机互动中不断重构自身，都是拟主体和拟客体。③ 按照此种理解，伦理学不仅仅是将机器人纳入对人类伦理关系的理解中，也不再是从主客两端出发分析处于两者之间的人机互动，而是应该从本体论的层面认可人与机器人是同一行动过程的两个产物，承认由中间生发出两极，这是因为，正是互动生发了伦理关系，进而不断地再生着人与机器人。世界原本就是一个“杂合”并且不断“再杂合”的世界，分立的主体与客体不过是一个幻象。由此，伦理学的范围大大扩展，理论与现实之间的关系也更加顺畅。

综上所述，当机器人进入寻常百姓之家时，我们要做的并不是简单地为之布置一个插座以备充电、安排一个空间以供摆放，而是应该推动一种新型人机关系的建立。这种关系的建立势必与那些已经深嵌于人类社会的其他技术物一样，牵一物而动全身，需要我们围绕人机关系构造一种新型的伦理、经济、政治等全方位的关系。由此，虽然诸如机器人之类的新技术带来了虚无，但这不过意味着先定方向的丧失，进而意味着人类应承担更重要的不断重塑方向的责任，“未来远不是我们完全无法掌控的、象征性的存在，而是一种可以被设计和建造的事物”。④ 因此，对待机器人，我们必须“超越批

① ［美］奥利·洛贝尔：《平等机器：数字技术创造美好未来》，苏苗罕、王梦菲译，上海人民出版社 2024 年版，第 304~305 页。

② 参见［美］帕特里克·林、凯斯·阿布尼、乔治·A. 贝基主编：《机器人伦理学》，薛少华、仵婷译，人民邮电出版社 2021 年版，第 217~221 页。

③ 参见［法］布鲁诺·拉图尔：《我们从未现代过：对称性人类学论集》，刘鹏、安涅思译，上海文艺出版社 2022 年版，第 106~114 页。

④ ［英］杰米·萨斯坎德：《算法的力量：人类如何共同生存？》，李大白译，北京日报出版社 2022 年版，“导论”，第ⅶ页。

判”，[①] 需要在认清新技术可能带来的各类问题的基础上，开启建设性的工作。如同拉图尔所说的，对人类社会中新技术带来的各种“怪物”和各种新型的“弗兰肯斯坦”，要热爱它们，要“像对待我们的孩子一样悉心照料”它们，不能因为技术会犯错就放弃，也不能溺爱技术而对之疏于管教。[②] 只有这样，才能在潘多拉魔盒的底部找到希望。

克服恐惑谷效应：跨文化视域下的中国人机协存社会*

程　林

人工智能伦理研究者斯文·尼霍姆（Sven Nyholm）曾假设，如果对无具体面容的人形机器人进行改造，赋予其中国人的外貌、名字和行为方式，那么人们对待改造前后的机器人的态度就会非常不同。[③] 不难理解的是，名称、外观、背景故事和设计理念会影响我们对机器人的感知，进而影响人机关系和伦理。与之类似，是以拟人化乃至仿真化为宗旨来设计机器人，还是出于某种异化模拟带来的不安而对拟人化产生抗拒，这既是机器人设计的现实问题，也是人文学界关心的审美、心理与伦理话题。机器人恐惑现象或曰“恐惑谷效应”（uncanny valley effect）的存在，可能会从审美、心理和情感方面导向对类人机器人乃至人机共存观念的排斥。恐惑谷效应亦与社会文化、人机关系和伦理联系紧密，这在既有研究中尚未得到足够重视。本文从机器人

① ［荷］马克·舒伦伯格、里克·彼得斯编：《算法社会：技术、权力和知识》，王延川、栗鹏飞译，商务印书馆 2023 年版，第 13 页。

② 参见 Bruno Latour, Love Your Monsters: Why We Must Care for Our Technologies as We Do Our Children, in Michael Shellenberger and Ted Nordhaus, eds., Love Your Monsters: Postenvironmentalism and the Anthropocene, Oakland: Breakthrough Institute, 2011。

* 本文系广东外语外贸大学外国文学文化研究院 2025 年度招标课题“文明互鉴视域下的东西方 AI/ 机器人文化研究”（25ZBKT03）的阶段性成果。

③ 参见［瑞典］斯文·尼霍姆：《序言·致中国读者》，《人与机器人：伦理、行动与拟人论》，刘铮译，上海交通大学出版社 2024 年版，第 2 页。

·笔谈·

机器人伦理学前沿问题

刘永谋等

【主持人语】经过几十年的发展，机器人技术已达到某种“临界点”，表现为人形机器人（humanoid robot）技术加速发展。工业和信息化部在2023年10月印发的《人形机器人创新发展指导意见》开篇即高屋建瓴地判断，人形机器人“有望成为继计算机、智能手机、新能源汽车后的颠覆性产品”。人形机器人的科技研发和应用同样面临科技风险和科技伦理问题，受到全社会的广泛关注。由此，机器人伦理学成为近年来的热门研究领域，尤其是应用伦理学重要的“理论增长点”。本次笔谈由6篇文章组成，聚焦机器人伦理学发展前沿，抛砖引玉，以期推动该领域研究的进一步发展。刘永谋和白英慧讨论拟人论意识形态在机器人伦理学建构中的基础作用，分析机器人拟人论在该领域流行的原因及启示。刘鹏考察了机器人伦理风险的发生机制及治理原则，基于此指出人类社会应通过与机器人的互动、互构，构建一种新型的人机有机结构。程林在跨文化视域下考察了机器人拟人化及恐惑现象，对中国人机协存观念和机器人设计理念提出了建议。谭笑考察了小数据主义技术路线在隐私保护和权力结构均衡等方面的优势，论证了由于所需知识类型不同，这一路线不太适用于社交机器人领域。孙圣引入拟人化分析，阐述了机器人伦理模型的惩罚缺漏不可避免，以此反驳直接将之作为判决标准的可行性，提出拟人度概念，用以划分人机共存社会发展阶段，并论证了划分何以可能。杨庆峰和朱清君考察了人形机器人导致的社会角色的临时替代和永久替代问题，并探讨了由人形机器人角色替代带来的伦理学挑战。

（刘永谋，中国人民大学哲学院教授、博士生导师）

【关 键 词】机器人　人形机器人　伦理学　机器人伦理学

【作者简介】刘永谋，中国人民大学哲学院教授、博士生导师；白英慧，中国人民大学哲学院博士研究生。刘鹏，南京大学哲学学院教授、博士生导师。程林，广东外语外贸大学外国文学文化研究院教授、阐释学研究院兼职研究员。谭笑，首都师范大学政法学院教授。孙圣，西北师范大学哲学与社会学院副教授、硕士生导师。杨庆峰，复旦大学科技伦理与人类未来研究院教授；朱清君，复旦大学社会发展与公共政策学院博士研究生。

【中图分类号】B829　【文献标识码】A

【文章编号】2097−1125（2025）06−0005−55

机器人伦理学的拟人论基础*

刘永谋　白英慧

近年来，人工智能、传感器、机器人控制与动力学、云计算与物联网、人工肌肉与柔性材料等技术的不断突破，使机器人更加灵活与自主，并使其应用场景愈加多元化、精细化。同时，机器人的研究、设计、制造和使用面临人类失业、隐私泄露、情感欺骗、责任分配等十分棘手的伦理问题，这些问题对机器人技术的发展及人类社会生活产生了不可忽视的重大影响，必须结合具体情境加以认真研究。在此背景下，机器人伦理学（roboethics）兴起并持续火热。顾名思义，机器人伦理学是研究有关机器人的伦理问题的学问。凯斯·阿布尼（Keith Abney）总结了机器人伦理学研究对象的三层含义：第一，机器人技术专家的职业道德；第二，为自动化机器人编写的道德规范的代码程序，即机器人自己的而非人类的准则；第三，机器人在具备进行伦理推理的自我意识能力时自行选择的伦理准则。① 换言之，机器人伦理学研究

* 本文系国家社会科学基金重大项目“现代技术治理理论问题研究”（21&ZD064）的阶段性成果。

① 参见［美］帕特里克·林、凯斯·阿布尼、乔治·A. 贝基主编：《机器人伦理学》，薛少华、仵婷译，人民邮电出版社 2021 年版，第 35 页。

Abstracts

Frontier Issues in Roboethics
Liu Yongmou et al.

【Abstract】 After decades of development, robotic technology has reached a certain "critical point", manifested in the accelerated development of humanoid robot technology. The Ministry of Industry and Information Technology issued the *Guidelines for the Innovation and Development of Humanoid Robots* in October 2023, which made a high-level judgment at the outset that humanoid robots "are expected to become a disruptive product after computers, smartphones and new energy vehicles". However, the research, development, and application of humanoid robots also face scientific and technological risks as well as scientific and technological ethics issues, drawing widespread attention from the whole society. As a result, robot ethics has emerged as a hot research field in recent years, particularly as a significant "theoretical growth point" within applied ethics. This special issue features six invited articles that focus on the cutting-edge developments in roboethics, aiming to spark further discussion and advance research in this field. Liu Yongmou and Bai Yinghui discuss the fundamental role of anthropomorphic ideology in the construction of roboethics, and analyze the reasons for the prevalence of anthropomorphism in this field and its implications. Liu Peng examines the mechanisms of the occurrence of robot ethical risks and the principles of governance. Based on this, he points out that human society should build a new type of human-machine organic structure through interaction and mutual construction with robots. Cheng Lin explores the anthropomorphism and uncanny valley phenomenon of robots from a cross-cultural perspective, offering suggestions for the Chinese-style human-machine co-existence and robot design. Tan Xiao examines the advantages of the small data technology roadmap in terms of privacy protection and power structure balance, and demonstrates that this roadmap is not very suitable for the field of social robots due to the different types of knowledge required. Sun Sheng introduces anthropomorphism analysis to demonstrate the inevitability of the retribution gaps of ethical models of robots, thereby refuting

the feasibility of using them directly as a criterion for judgement. He proposes the concept of degree of anthropomorphism as the demarcation of development stages of human-machine coexisting societies and justifies its plausibility. Yang Qingfeng and Zhu Qingjun investigate the temporary and permanent substitution of social roles caused by humanoid robots, exploring the ethical challenges posed by the role substitution of humanoid robots.

(Liu Yongmou, Professor and PhD Superviser, School of Philosophy, Renmin University of China)

【Keywords】robot; humanoid robot; ethics; roboethics

Promoting the Common Values of Mankind: China's Contribution to the Advancement of International Rule of Law

Li Lin

【Abstract】The proposal of new ideas regarding the common values of humanity has not only provided a new value foundation for people of all countries to join hands in building a community with a shared future for mankind, but also offered strong value guidance for China's participation in promoting theoretical, institutional and practical innovations in the international rule of law. Since the founding of the People's Republic of China, particularly since the 18th National Congress of the Communist Party of China, China has upheld the banner of human values and civilizations and made significant contributions to safeguarding world peace and promoting the development of the international rule of law. Hence, as a responsible major power emerging on the global stage and engaging in international affairs, China must adeptly employ the rule of law. In order to employ the rule-of-law thinking and rule-of-law based approach to boost the building of a community with a shared future for mankind, greater emphasis should be placed on the coordinated advancement of domestic and foreign-related rule of law. China is also required to promote the common values of humanity, advance foreign-related rule of law initiatives, and actively engage in the development of the international rule of law, thereby contributing more Chinese wisdom and strength to the progress of international rule of law.

【Keywords】common values of humanity; domestic rule of law; foreign-related rule of law; international rule of law; rule of law civilization