中国人民大学学报

工作论文系列

Working Paper Series

人工智能道德主体地位的伦理探析

王福玲

JRUCWP2026004

2026. 01. 20

人工智能道德主体地位的伦理探析

王福玲

［**摘要**］在人工智能伦理研究中，道德地位问题是一个根本性的哲学议题。随着人工智能系统广泛介入人类生活实践并展现出某种形式的自主性，传统伦理学中以“道德行动者”和“道德承受者”为核心范畴来界定道德地位的二元分析框架，逐渐显示出内在局限性，难以充分回应人工智能带来的新型伦理挑战。因此，拓展和重构道德地位理论已是当务之急。一种新的道德地位模型主张区分道德行动者和道德主体的概念，将道德主体概念嵌入传统二元结构中，建构一个集“道德行动者-道德主体-道德承受者”的三元道德地位层级理论，并赋予人工智能道德主体的地位。该进路能够更准确地反映人工智能技术在当前及未来人类生活中扮演的角色，回应技术发展的现实需求，有利于构建和谐的人机合作关系。

［**关键词**］人工智能　道德主体 道德行动者 道德承受者

人工智能技术以迅猛的速度发展并应用于社会生活的各个领域，由此也带来诸多伦理困惑。2010 年 Paro 机器人在日本完成户籍登记。2017 年机器人索菲亚被赋予沙特阿拉伯公民的身份。2024 年美国 14 岁少年与聊天机器人 Character AI 进行长期对话互动后自杀身亡。这些事件引发学界和大众的广泛关注和深入讨论。我们应该如何看待人工智能？问题的实质可以归结为人工智能的道德地位问题，相关研究成果可谓汗牛充栋。学者们主要是在道德行动者（moral agent）和道德承受者（moral patient）[①] 这个传统二元框架下进行探讨。有的学者为人工智能作为道德行动者的身份进行辩护，有的学者聚焦人工智能作为道德承受者的身份及其规范性意蕴。有的学者在二者之间徘徊，提出准道德行动者（或准道德主体）的身份界定。这些研究深入而广泛，尽管未能形成共识，却也引发了人们关于人工智能道德地位问题持续而深刻的思考。笔者尝试扩展传统应用伦理学中讨论道德地位问题的范畴，提出一种关于道德地位的三元层级理论，以期更准确地分析和回应人工智能道德地位问题。

作者：王福玲，中国人民大学伦理学与道德建设研究中心暨哲学院副教授，wfl@ruc.edu.cn。

＊　本文受到北京市社会科学基金重点项目“生命伦理学中的脆弱性问题研究”（23ZXA002）、中国人民大学 2025 年度中央高校建设世界一流大学（学科）和特色发展专项引导资金资助。

① 在笔者看来，将 moral agent 和 moral patient 分别翻译为道德主体和道德客体不太妥当。agent 强调意向性、自主性、责任能力。在目前人工智能的相关研究中，越来越多的学者主张人工智能具有主体性，但是无论学者们在主体性问题上能够达成多大程度共识，较少有学者主张人工智能具有责任能力，可以且应当成为责任承担的主体。鉴于此，我们应该将 moral agent 与汉语中的“道德主体”区别对待。另外，笔者也认为 moral agent 翻译为道德行动者比道德行为者更合适。因为，在哲学、行为科学和社会科学中，“行动”（action）一词有其特定的意义，与行为（behavior）有区别。行为是自动的反射性活动，而行动则是有意向、有目的、有意识和对行动主体有意义的活动。

一、问题的实质及其重要性

道德地位是指一个实体根据其内在特质或在伦理关系中扮演的角色，而被赋予某种道德考量的资格。在传统应用伦理学中，关于道德地位的讨论一直是动物伦理学家们争论的焦点，他们创造性地提出一对范畴——道德行动者和道德承受者，这对范畴为我们理解道德地位问题提供了一个很好的分析框架。

道德行动者是指“所有具有以下这些能力的存在：他能够按照道德或非道德的标准去行事；能够具有责任和义务；能够对自己的行为负责。在这些能力中，最主要的能力是能够判断是非；能够进行道德思考，即：能够对可供选择的各种行为路径的正误进行道德上的思考和权衡并给出理由；能够依据上述理由作出决定；有能力作出必要的解决并有意志力来实施那些决定；而且能对自己未能实施决定向他人作出回答。”① 这些能力中最核心的是道德责任能力。能够为自己的行为承担相应的道德责任就意味着该主体具有自主性以及前述判断思考和行动的能力等等。在此，我们看到，道德行动者的阈值是比较明确的，也是比较高的，其标准是“承担道德责任的能力”。

道德承受者是指获得道德关怀的实体。哪些实体能够被视为道德承受者呢？纵观人类社会的发展历程，我们会发现，道德承受者的范围是在不断扩展的。在早期人类社会中，女人、儿童、外邦人都是被排除在道德承受者之外的。换言之，女人、儿童、外邦人的地位与物等同，可以被随意对待。随着人类文明的进步，人人平等观念深入人心，这些人逐渐进入道德关怀的范围内，享有道德承受者的身份。传统关于道德地位问题讨论的一个重要贡献就是将道德关怀的范围不断拓展，从人类拓展到非人动物，乃至生态环境。由此可见，道德行动者属于道德承受者，既具有对其他道德承受者的义务，又具有被道德对待的资格。道德承受者的范围从道德行动者扩展到人类共同体中的所有成员，进一步扩展到非人动物乃至生态环境等，这一进程恰恰与人类文明的历程相吻合。②

传统应用伦理学对道德地位的讨论聚焦道德承受者。玛丽·安妮·沃伦（Mary Anne Warren）说：“拥有道德地位就是在道德上应给予关怀，或拥有道德资格。拥有道德地位的实体就是这样的实体，即道德行动者对它负有、或能够负有道德义务。如果一个实体拥有道德地位，我们就不能为所欲为地对待它；在进行慎思的时候，我们有道德义务赋予它们的需要、利益或福祉一定的分量。”③ 在此，沃伦所说的道德地位主要是作为道德承受者的身份。概言之，道德地位首先是指一个实体获得道德对待（关怀、帮助等）的资格，某实体拥有道德地位意味着道德行动者在行动时有必要将该实体的需要、利益或福祉考虑在内。④

数智时代，人工智能技术迅猛发展并被广泛应用于生活的诸领域中。医疗诊断、司法辅助、情感陪伴等，人工智能正以前所未有的广度与深度介入人类生活实践。关于人工智能道德地位的讨论也成为伦理学研究的一个热点话题。我们该如何对待这样一类与人类生活紧密相关的智能机器人呢？我们可以无缘无故地踢一只机器狗吗？我们可以对家里的照护机器人进行拳打脚踢，以此来发泄自己的情绪吗？可以虐待性爱机器人吗？换言之，人类有道德地对待人工智能⑤的义务吗？人工智能可以作为道德承受者吗？与此同时，人工智能不仅影响我们的决策方式、工作模式与交往形态，更在潜移默化中参与意义建构、价值生成乃至主体性塑造的过程。它仿佛在极力褪去自身仅仅被视为“物”的身份，这一现象逼迫人类不得不开始进一步审视，除了作为道德承受者是否可能的

① 保罗·沃伦·泰勒：《尊重自然：一种环境伦理学理论》，8页，首都师范大学出版社，2010。

② 杨通进：《道德关怀范围的持续扩展》，载《道德与文明》，2020（1）。

③ Mary Anne Warren. *Moral Status*：*Obligations to Persons and Other Living Things*. Clarendon Press，1997，p. 3.

④ 汤姆·比彻姆、詹姆士·邱卓思：《生命医学伦理原则》，71页，科学出版社，2022。

⑤ 这里讨论的人工智能主要是指像照护机器人、性爱机器人这样一些与人类具有较强互动性的人形智能机器人。

问题外，人工智能是否可以同时作为道德行动者呢？在人工智能伦理中，道德地位的界定是我们讨论其他诸如人工智能的可说明性问题、人工智能价值对齐、人工智能安全性等问题时最终都会诉诸的原点。人们对该概念的理解是否能够达成共识，以及能够在多大程度上形成共识都将直接影响上述问题的解决方案。它为我们讨论责任归属问题以及当下对高新科技进行伦理治理的方案提供了重要的理论依据。

道德地位问题不仅是人工智能伦理中的一个基本问题，同时也是应用伦理学中的一个元伦理问题。有学者质疑道德地位这个概念的有用性和必要性，在他们看来，道德理论应该直接指导人们如何对待个人或他物。但现有的道德地位理论依旧过于简单和抽象，不能具体的、确切地指导在不同情境中应该如何抉择。① 但诚如汤姆·比彻姆（Tom L. Beauchamp）和詹姆士·邱卓思（James F. Childress）所坚持的“这些观点恰当地警示了我们道德地位理论的局限性。尽管如此，道德地位仍然至关重要，它应该被仔细地分析，而不是被忽视或被轻视。”② 道德地位问题的复杂性恰恰彰显了现实世界的丰富性和多样性，目前道德地位理论的局限性不能成为我们放弃该概念的理由。雪莱·卡根（Shelly Kagan）在探讨动物道德地位问题时也强调了道德地位问题的重要性，他说：“在没有充分考虑（道德——笔者注）地位的重要性之前，我们对伦理学的理解——不仅是动物伦理学，而是整个伦理学——都将会是混乱而残缺的。”③ 回顾历史上奴隶制度下奴隶所遭受的非人待遇，以及生物医学研究历史上那些受试者所遭受的虐待，我们就会看到道德地位标准的缺陷以及人们对这一问题的忽视将带来灾难性的后果。随着人工智能等新型多元主体的呈现和介入，人类的生活实践将变得更加复杂，这也要求我们对道德地位问题给予持续的关注和深入的探讨。

二、传统道德地位范畴的解释困境

道德行动者和道德承受者这对经典的二元范畴是否足以解释当下人机深度互动给人类生活世界带来的结构性变迁？面对人工智能技术发展带来的诸多伦理挑战，我们是否能够在这一传统框架下给出恰当的、具有前瞻性的伦理回应？笔者将从这对范畴出发，考察人工智能作为道德承受者的可能性，辨析其作为道德行动者的局限性，在此基础上，揭示传统二元范畴在分析人工智能道德地位问题时面临解释力不足、无法恰当回应人机交互中涌现出的新型伦理现象等问题。

（一）人工智能作为道德承受者的可能性

与传统应用伦理学中讨论的对象，如动物、植物等自然生物不同，人工智能是人类发挥自身主动性创造出来的“产品”。因此，就其属性来说，它是人工“物”，类似于古代社会中人们发明的斧子、锄头等工具。在西方主客二分的二元论传统中，作为“物”的工具与人类主体具有本质差异。根据康德的道德哲学，作为物的存在可以仅仅被视为工具，作为人格的存在则应该被视为存在自身就具有价值的东西，是目的自身，而不仅仅是工具。④ 因此，我们对“物”没有直接的道德义务。尽管如此，这并不意味着我们可以任意践踏这些仅仅作为“物”的存在，相反，康德强调人类对它们负有间接义务。例如，一匹为主人效劳多年的马应该得到主人的善待，这并不是因为它是一匹

① 萨克斯（Sachs B.）认为，关于道德地位的主张是没有必要且令人困惑的，因为这场争论实际上是关于何种属性能够为特定的权利提供辩护的问题。西尔弗斯（Silvers A.）提出了更加尖锐的批评，在她看来，那些以生物属性和心理属性为标准的道德地位理论忽视了主体在拥有这些特征上的差异性，同时也忽略了主体发展这些特征的潜力，由此不可避免地将一些个体错误地排除出去。从关系视角建构的道德地位理论尽管具有一定的现实指导意义，但由于“关系”概念的模糊性，相关理论缺乏一个令人信服的理论基础。Sachs B. “The Status of Moral Status”. *Pacific Philosophical Quarterly*, 2011, 92; Silvers, A. “Moral Status: What a Bad Idea!”. *Journal of Intellectual Disability Research*, 2012. 56 (11).

② 汤姆·比彻姆、詹姆士·邱卓思：《生命医学伦理原则》（第 8 版），93 页，科学出版社，2022。

③ Shelly Kagan. *How to Count Animals*. Oxford University Press, 2019, p. 303.

④ 康德：《道德形而上学奠基》，62－63 页，人民出版社，2013。

马，而是因为这样做有利于呵护主人的道德情感，进而有利于德性的养成。[①] 遵循该逻辑，我们同样也可以推论，在未来人机共存的社会，人类应该善待机器人，而不能虐待它们，或者说人类对待它们的行为方式应该体现出人类的道德。因此，诸如踢打机器狗、对照护机器人拳打脚踢、虐待性爱机器人的行为应该受到道德谴责。其原因不是因为它们会感到疼痛，而是因为这样做不利于人类道德的养成。在关于人工智能道德地位的讨论文献中，一些学者遵循这一进路主张人工智能应该被纳入道德关怀范围内。

马克·考科尔伯格（Mark Coeckelbergh）基于更广泛的社会背景拓宽了这一论证路径，进而强化了人工智能作为道德承受者的理由。他从美德伦理出发，强调德性和恶习往往表现为一种习性。我们不是通过个体一次性的行为来判断的，而是根据他长期反复的行为方式进行评价的。同时，个体德性和恶习的养成不仅仅关涉个体的内在意向，更是社会结构和观念在个体身上的内化，进而通过个体行为方式表现出来。这个视角对于理解人工智能道德地位来说非常重要。虐待性爱机器人错在哪里？不是因为对机器人造成了伤害，也不仅仅是因为虐待行为反映出该个体的恶习，而是因为该行为映射出更广泛的滋生虐待的社会环境。这个环境容忍或纵容对女性的性虐待，仿佛她们是机器一般。当一个人虐待性爱机器人时，我们需要反思，他“虐待”的行为是怎么形成并内化为习惯的。考科尔伯格强调美德的社会性维度，他指出德性、恶习不仅仅是个体意向的事情，更是社会观念、机制逐渐内化为个体习性的过程。这一方面揭示出个体德性养成的困难，另一方面也意味着个体德性的养成需要得到社会环境的支持。一个良好的社会环境有助于个体德性的养成。因此，在人工智能道德地位问题上，我们应该以什么方式对待机器人，所关涉的是我们希望如何对待他人，以及如何被他人对待的问题。我们希望个体拥有什么样的德性，就应该在培养个体内在德性意向的同时，也注重营造一个有德性的社会环境。在人工智能已经深度嵌入人类社会生活的当下，将人工智能纳入道德关怀的范围，有助于营造和谐的社会环境，进而塑造个体德性。正如玛蒂娜·罗斯布拉特（Martine Rothblatt）在《虚拟人》中所倡导的，“当我们做到像尊重自己一样尊重他人（即虚拟人，笔者注），并将这一美德普及至世间各处时，我们就为明日世界的到来做了最好的准备。”[②]

考科尔伯格对于诉诸直接义务[③]论证道德承受者地位的进路提出质疑，在他看来，“关于道德地位的一种哲学讨论不应该仅仅聚焦关于道德地位的直接论证，还应该思考道德地位何以成为一个问题，道德承受者是如何被建构的。”[④] 笔者赞同考科尔伯格的观点，关于道德地位的讨论是在人类语言内部展开的，因此，关于某实体是否拥有某种道德地位的讨论离不开人类的主观判断。就此而言，关于道德地位的讨论中，真正有意义的问题并非该实体在本体论层面是什么，而是它与人类处于何种关系架构中，人类应该以何种态度看待该实体，应该以何种方式对待该实体。

（二）人工智能作为道德行动者的局限性

人工智能可以作为道德行动者吗？如前所述，道德行动者的核心要素包括：自主性和道德责任能力。拥有自主性的存在应该为自己的行为承担道德责任。目前，关于人工智能的讨论中，许多文

---

① 《康德著作全集》第6卷，454页，中国人民大学出版社，2007。

② 玛蒂娜·罗斯布拉特：《虚拟人》，317页，浙江人民出版社，2021。

③ 汤姆·雷根（Tom Regan）在探讨动物权利问题时，将过去的理论探索大致区分成间接义务论和直接义务论两个进路。在间接义务论看来，我们对动物的道德关怀并非源于动物本身，不是因为动物本身具有何种与道德相关的属性，而是因为动物与道德行动者以不同方式绑定在一起，我们对动物的道德关怀实质上可以归结为对道德行动者的义务。在直接义务论看来，动物本身具有某种与道德相关的属性，他们的需求和利益值得我们以道德的方式对待。参见T·雷根：《关于动物权利的激进的平等主义观点》，载《哲学译丛》，1999（4）。

④ Mark Coeckelbergh.“The Moral Standing of Machines：Towards a Relational and Non-Cartesian Moral Hermeneutics”. *Philos. Technol*，2014，27：66.

献谈及人工智能的自主性问题。在笔者看来，这里的自主性概念是极其有限的。准确地说，只是一种类比意义上的自主性，同时，人工智能也不具备承担道德责任的能力。就此而言，人工智能不可能成为道德行者①

欧盟《人工智能法案》序言第12条指出，“不同程度的自主性”是指人工智能系统被设计为在行动上在一定程度上独立于人类参与，并且具备在无人类干预的情况下运行的能力。② 这里的自主性体现在能够独立处理信息、做出决策。例如，人工智能系统可以通过分析大量数据生成内容、规划任务，整个过程无需人类直接干预。同时，人工智能还可以通过深度学习、自我迭代等方式，根据环境变化，自主调整策略，展现出更高程度的自主性。然而，诚如邓晓芒在分析人工智能的本质时指出：“在人工智能中，思维规律被外化为机器式的数码传动装置，它的‘自动性’是假象，其实是早就设定好了的，哪怕设定时不一定预计到它的后果，甚至结果还会出乎设计者的预料，但其中的逻辑关系（包括模态逻辑或概率关系）是能够算出来的。”③ 就目前人类对人工智能发展的有限认识来看，学界普遍认为人工智能本质上是对人类智能的模仿和学习。人工智能的自主性是有限的，因为它还受到设计者设计的目标和规则的约束。自主性是哲学、伦理学中的一个核心概念，是道德责任可归咎性的重要依据。

与自主性紧密相关的是道德责任概念。人工智能是否具有承担道德责任的能力？这是判定人工智能是否能够成为道德行动者的关键因素。对此，笔者将从以下两个视角展开讨论。第一，我们是否应该将道德责任追溯并赋予人工智能？第二，人工智能是否有能力承担相应的道德责任。

首先，在严格意义上，道德责任、义务这些范畴都只是针对人类这样有意志且不纯粹的个体而言的。作为有意识的自主性存在，人可以对自己的行为和能力有清醒的认知，并对自己的行为进行反思和评价，这是承担道德责任的前提。当人类共同体中的成员，例如婴幼儿、精神障碍患者等不具备或丧失这些能力时，人们不会在法律和道德层面追究这类个体的责任。他们不是“回应性态度”的适当对象。就目前人工智能的发展水平而言，人工智能尚不能形成这种理性认知。因此，在人机合作的情境中，只能将道德责任归咎于人。概言之，自主性是道德责任可归咎性的理论前提，类比意义上的自主性不能成为归咎道德责任的依据。

其次，退一步说，纵然我们可以在类比的意义上将责任归咎于人工智能，但问题的关键是，人工智能能够承担道德责任吗？法比奥·托隆（Fabio Tollon）主张我们可以对人工智能进行道德评价，并将相应责任归属于人工智能。他说：“进步主义解释则采纳了这样的观点：无论行为是否‘可惩罚’，实体都可能对行为承担道德责任。正如我所论证的，我们完全可以在不附加‘必须接受惩罚’（如精神病患者案例）这一条件的情况下，认定主体具有道德责任。”④ 在笔者看来，这种阐释看似对责任的界定和归属问题给出了一种解决方案，但它不具有实践意义，因为人工智能不具备承担道德责任的能力。在人类历史上，人类共同体中的责任个体是通过接受不同形式的惩罚来承担相应责任的，比如，谴责、施加痛苦、流放、罚款、没收财产、剥夺自由乃至生命等手段。这些惩罚手段直接作用于责任主体所关切的利益。正因此，奖惩才有意义和功效。从根本上来说，这种承

① 需要注意的是，人工智能不能作为moral agent这一结论，是笔者运用应用伦理学中moral agent和moral patient这对范畴分析人工智能道德地位问题时得出的结论，也就是说，笔者这里所言moral agent与人工智能领域中moral agent内涵不同。如前文所述，在笔者讨论的语境中，moral agent的核心要素是自主性和责任能力，就此而言，人工智能成为moral agent面临很大的挑战。关于二者的详细区别，笔者将另辟文进行讨论。

② 参见 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689

③ 邓晓芒：《人工智能的本质》，载《山东社会科学》，2022（12）。

④ Fabio Tollon. “Do Others Mind? Moral Agents without Mental States”. *South African Journal of Philosophy*, 2021, 40 (2): 192.

担责任的能力源自人性，即人是具有肉身的，因而是脆弱的理性存在者。人类是拥有血肉之躯的理性存在者，因此人类才有被保护、被尊重、被满足的“需求”，满足或剥夺与这些需求紧密相关的东西才构成了对人类行为的奖惩。那么，适用于人类社会规范的这种责任、惩罚机制是否适用于人工智能呢？根据目前人类对人工智能的认知，我们无法证明人工智能具有上述关切，也无法判断适用于人类行为规范的道德和法律对于人工智能本身来说是否有意义。因此，我们不可能在严格意义上对人工智能进行奖惩。尽管在人工智能领域的研究中，设计者们在设计程序、训练数据时会运用到奖惩机制，让人工智能能够得出更精准的数据和结论。但这里的奖惩也并非上述承担道德责任意义上的奖惩。鉴于此，笔者主张人工智能不能成为道德行动者，只有人类才有资格作为责任，尤其是道德责任的承担者。

（三）面临的解释困境

如前所述，沿用传统的二元范畴来审视人工智能的道德地位，可得出如下启示：在日益深化的人机共存社会中，人类在使用人工智能产品，尤其是具备高度交互能力的智能机器人时，应该以合乎伦理道德的方式对待之。然而，在此传统分析框架下，人工智能被排除在道德行动者的范畴之外，仅被视为道德承受者，亦即伦理语境中的被动接受者。尽管这一界定重申了人类作为道德主体的核心地位及其不可推卸的伦理责任，却未能充分捕捉当代人工智能对人类生活方式、社会结构乃至价值体系所具有的建构性影响，亦缺乏对这一新兴现象的理论解释力。

人工智能产品越来越像人，它可以不断模拟人类的认知、行为，乃至情感表达，且越发逼真。未来社会，人工智能将成为人类生产生活不可或缺的伙伴，它可以协助人类处理大量事务，大大推动社会的发展。与此同时，它不仅影响人类的生存和生活方式，甚至在无形中塑造人类的思维方式，影响人类建构意义的过程。人类也将在与人工智能更加紧密的互动关系中重新定义人类自身。这一进程速度飞快，影响巨大，不断逼迫人类思考：人工智能是什么？确切的说，人工智能是什么本身已经不再重要，重要的是我们应该将人工智能视为什么？我们应该以什么样的态度对待人工智能？无论它是什么亦或不是什么，它都已经展现出对人类深刻的影响力，这是至关重要、且不容小觑的。鉴于这一现象，有学者提出“准道德行动者”“不完全的道德行动者”等术语尝试给人工智能更精确的定位。[①] 在笔者看来，这一现象表明，传统讨论道德地位的二元范畴已经不足以涵盖像人工智能这样的实体。而“准”“不完全”这些表述反而令人与人工智能的界限变得模糊了。

人工智能对人类产生的影响已经远远超过传统社会中我们对于外在物的理解。在这种时代背景下，仅仅将人工智能视为一般“物”或道德承受者的认知容易催生道德盲点，忽视一些重要的伦理问题，导致人类对人工智能技术可能潜藏的风险反应迟缓，也不利于建立健康持续的人机合作环境。有研究揭示，“人们在收集与选择数据时，各种偏见往往会随着人类自身的介入而渗透在数据中，数据带有原始性偏见；用具有偏见性的数据再去训练算法，又会产生更深的算法偏见与算法歧视等一系列问题。”[②] 在医疗诊断、招聘决策等领域，人工智能系统在应用过程中，会不可避免地自动学习并延续人类既有的偏见，甚至存在进一步放大这些偏见的风险。与人类相比，人工智能系统对数据中微小偏差的敏感度更高，它能够精准捕捉并利用这些细微的偏差信息，构建出更“精准”的预测模型。然而，这种基于偏差数据的“精准”预测实际上却进一步加剧了偏见的传播与固化。具体而言，人类在将带有偏见的数据输入人工智能系统后，在后续与人工智能系统的交互合作过程中，这些原本隐藏于数据背后的偏见，会逐渐被客观化、数据化。经过人工智能系统的处理与反馈，这些偏见以一种看似科学、客观的形式呈现出来，变得更加难以甄别与纠正。由此可见，人

① 代欣玲等：《生成进路下人工智能的道德主体地位》，载《自然辩证法研究》，2022（8）.

② 董春雨：《从机器认识的不透明性看人工智能的本质及其限度》，载《中国社会科学》，2023（5）。

类与人工智能系统之间的互动，并非简单的信息传递与决策辅助过程，而是一个复杂的偏见交互与放大过程。仅仅承认人工智能作为道德承受者的身份容易让人们忽视人工智能与人互动过程中隐藏的潜在风险。

人工智能技术迅猛发展并广泛应用到人类生活的每一个角落。护理机器人、伴侣机器人等逐渐走进千家万户，悄然改变着人类的生活图景。在这一过程中，人与人之间的黏性逐渐被稀释，原本深厚的情感纽带逐渐松弛。人们之间的情感依赖开始从人与人之间，转向一部分人对机器人的单向情感依恋。人工智能在满足人类多样化需求的同时，已然成为人类情感寄托的对象、日常交往的伙伴，甚至是不可或缺的工作助理、生活中的亲密搭档。与此同时，人们的生育观念、婚姻观念、养老观念乃至对死亡的认知，都在人工智能的影响下悄然发生变化。可以说，人工智能不仅重塑了人类的生存方式与生活模式，更深入到人类建构意义的核心领域，对人类的精神世界产生深远影响。鉴于人工智能对人类社会产生如此巨大的影响，将其仅仅视为“工具”、“对象”或“客体”的态度并不是适应未来人机合作模式的最佳方案。相反，这种认知和态度不仅折射出人类在科技发展过程中的傲慢心态，更可能在无形中放大人类的狂妄自大。

三、一种替代方案：道德地位三元层级理论

当人工智能以迅雷不及掩耳之势侵入人类生活时，我们越是需要冷静地判断，“它”是什么，“我们”是什么，清晰的身份界定有利于进一步探讨责任的归属，更能够为人工智能的发展保驾护航。鉴于此，笔者主张，我们需要拓展传统道德地位的二元范畴，区分道德行动者和道德主体（moral subject），建构一个集道德行动者、道德主体和道德承受者为一体的三元层级理论，在此基础上赋予人工智能道德主体地位。这一立场有利于在未来人机共存的社会中营造健康有序的人机共存共赢的合作氛围。

（一）重释“道德主体”：超越主客二分的伦理视角

在汉语的习惯性表达中，人们倾向于用道德主体和道德客体来表达传统道德行动者和道德承受者这对范畴的内涵。事实上，在更广泛的哲学领域，道德主体和道德客体这对范畴的使用范围更广，它可以根据不同语境分别运用在认识论和价值论层面。在讨论人工智能道德地位问题时，大多数学者会在相同意义上使用道德主体和道德行动者，有时甚至直接用道德主体来翻译英文文献中的moral agent。[①] 然而，笔者主张，在讨论人工智能问题时，我们最好区分汉语常用的“道德主体”和英文中 moral agent 这两个概念。在严格限制人工智能成为道德行动者的同时，将道德主体赋予人工智能。与道德行动者相比，道德主体这一概念不必承载责任能力这一要素，尤其是道德责任。它可以指那些具有一定自主性，但不能为自己的行动承担责任的实体。当代动物伦理学家马克·罗兰（Mark Rowlands）也尝试通过区分道德行动者、道德主体和道德承受者这三个概念来讨论动物的道德地位问题。在他看来，动物可以作为道德主体，因为动物可以基于道德理由行动。[②] 尽管罗兰试图通过证明动物也具有道德能力，可以基于道德理由去行动的观点受到争议，但他通过分析归责的条件这一视角区分道德行动者和道德主体的思路却对我们分析人工智能的道德地位问题具有启发意义。

词源上，subject 是对希腊语 hypokeimenon 的翻译。在亚里士多德哲学中，subject 是指存在

① 例如贾向桐等学者在“论当代智能体人工道德主体辩护的逻辑与超越”一文中运用的是“道德主体”这一术语，对应的英文是 moral agent。代欣玲等学者也将 moral agent 翻译为道德主体。参见贾向桐、冯枫添：《论当代智能体人工道德主体辩护的逻辑与超越》，载《学习与探索》，2025（3）。代欣玲、彭小兵、程鹏：《生成进路下人工智能的道德主体地位》，载《自然辩证法研究》，2022（8）。

② Mark Rowlands. “Moral Agent，Patients，and Subjects”. In *Can Animals Be Moral*? Oxford University Press，2012. pp. 71－98.

论上的主体或实体，在逻辑上则指谓述判断中的主词。在近代哲学中，subject 被赋予了主动性的内涵，在行动意义上指行动的发出者。据此，我们可以将 moral subject 理解为道德行动的主动发出者，该主体不必然是人。在哲学中，agent 主要流行于英语语境中，不仅指行动的发出者，更强调行动的主导者。① 概言之，moral agent 是道德行动的主导者，moral subject 是道德行动的发出者，moral agent 可以是 moral subject，但 moral subject 并不必然是 moral agent。与行动者（agent）相比，主体概念可以指代的对象范围更广。纵观主体概念的历史演变也可以看到，主体直接定位在人身上是很晚时候才发生的。“从主体概念进入哲学之时算起，在近两千年的时间内与‘人’并没有什么根本上的关联。”② 道德行动者这个概念能够清晰地界定那些具有自主性和责任能力的主体。对于那些具有一定的自主性，却无法对自己的行为承担责任的主体，他们的道德地位则是相对模糊的。他们不能作为道德行动者存在，而仅仅作为道德承受者的身份却无法恰当反映出这类主体的独特性及其在交互性的生活实践中对人类生活的重要影响。鉴于此，我们不妨将道德行动者和道德主体区分开来，并将道德主体赋予这类具有一定自主性，但无法对自己行为承担道德责任的主体，这就包括婴幼儿、人工智能等具有一定自主性的实体。我们可以说，所有人都是道德主体，在这个意义上，人人享有平等的道德地位。但是，只有人类中那些能够承担责任的主体才是道德行动者。概言之，笔者尝试在传统道德行动者和道德承受者这对范畴中嵌入一个介于二者之间的道德主体概念来分析人工智能的道德地位问题。

在该理论中，道德行动者、道德主体和道德承受者是基本范畴。人类共同体中那些能够承担道德责任的人是道德行动者。人类共同体中无法承担道德责任的人，如儿童等群体以及具有一定自主性的人工智能是道德主体。那些通过间接义务和直接义务路径能够证明应该被道德对待的实体是道德承受者。三者的位置是一个金字塔结构，位于顶端的是道德行动者，中间是道德主体，底端是道德承受者。道德行动者是道德主体，也是道德承受者。据此，人工智能不仅是道德承受者，同时还可以作为道德主体。理智健全的成年人类当然是道德承受者，同时也是道德主体，最重要的是，就目前人类的生活实践而言，有且仅有理智健全的成年人类是道德行动者。

随着社会的发展，哪些属性或何种关系将成为决定道德地位的关键属性或重要关系，这将会是一个动态的变化过程。与此同时，随着个体自身的发展，当那些潜在的关键属性逐渐变成现实属性时，他们的道德地位也可能会随之改变，随着人类生活方式和需求的变化，一些新型的重要关系也将浮出水面，这些因素都将影响相关实体的道德地位。

（二）确立人工智能道德主体地位的优势

在科技浪潮奔涌向前的当下，赋予人工智能道德主体的身份，既是可行的，也是必要的。如同在我们的法律体系和伦理规范中所规定的，未成年人、精神病患者不需要为他们的行为负责，但这并不影响他们作为道德主体的身份。同样，人工智能虽然不具有独自承担道德责任的能力，但并不影响赋予其道德主体的身份。赋予人工智能道德主体地位更有利于推动人工智能的健康发展，在未来人机合作中建立和谐的人机共生关系，有利于人类采取恰当的伦理姿态对待人工智能。

首先，人工智能道德主体的身份可以时刻警醒人类在享受人工智能带来的便利与进步的同时，深刻认识到人工智能的独特性与局限性，避免过度神化或妖魔化人工智能。目前，人工智能已经或即将拥有类似人类的自主行动与判断能力。我们不能再以高傲的姿态俯瞰人工智能，尽管它终究不是人。但事实上，它已经在诸多领域开始扮演类似人类的道德主体角色，给人类生活带来深远的影响。就此而言，赋予人工智能道德主体地位，犹如为人类敲响一记振聋发聩的警钟，能够极大地提

① 感谢聂敏里、李科政两位老师在该词源解释上为笔者提供的建议。

② 张志伟：《主体概念的历史演变》，载《教学与研究》，1996（5）。

高人类的风险意识和责任意识。这一举措是基于对未来人机关系深刻洞察的必然选择。通过赋予人工智能道德主体地位，人类能够更加审慎地思考人机关系的边界与规范，从而构建一个更加和谐、可持续的人机共生社会。有人或许担忧，提升人工智能的道德地位会威胁到人类自身的地位。在笔者看来，这一担忧是多余的。确立人工智能的道德主体地位要求人类重新审视自身，这是人类再次理解自身本质的一个契机，它要求人类在更高程度上发挥人作为道德行动者的主体性。

其次，从关系的视角出发，赋予人工智能道德主体地位为理解人工智能在道德层面的角色提供了独特且富有洞察力的视角。根据戴维·贡克尔（David J. Gunkel），判定某一实体是否具备道德地位的核心要素，并非该实体自身所固有的属性，而是该实体所嵌入的关系网络。[①] 脱离具体关系来孤立地确定道德地位，这种做法本质上是对道德的背离。道德并非抽象、孤立的存在，而是深深植根于实体间的交互关系之中。因此，当我们在探讨人工智能的道德地位问题时，首要的任务便是深入考察人工智能与人的关系。我们需要细致剖析人工智能是如何嵌入人类生活，影响人类行为方式并塑造人类思维模式和价值实践的。当下社会，人工智能已经深度参与到人类社会的运转之中，它已经不仅仅是执行任务的机器，而且与人类形成了错综复杂的关系网络，影响人类的决策，更为关键的是它参与到人类建构意义的过程。在这一过程中，人工智能展现出一种交互主体性。[②] 这种主体性不是基于人工智能的内在意识，而是通过与人类的互动而呈现出来的行动能力。关系视角为我们探讨人工智能的道德地位问题提供了别具一格的探究模式。通过关系进路，我们能够更加全面、深入地理解人工智能与人类之间的互动关系，洞察人工智能对人类社会产生的深远影响。

综上所述，赋予人工智能道德主体地位，是应对人工智能时代挑战，规范人类行为，促进人机和谐共生的必然选择。这一主张是基于一种更深层次的人文关怀和伦理责任感，彰显了人类在科技发展进程中的伦理自觉与责任担当。人工智能道德主体的身份可以时刻警醒人类在对待高度智能化的技术时，不能仅仅将其视为被动的对象加以操控。相反，我们应当以一种更加审慎、尊重的态度研发、设计并使用人工智能产品。这种态度最终指向的并不是人工智能本身的利益，而是人类社会整体的道德进步。我们对待技术的态度，在很大程度上反映了我们如何理解人性、价值与责任。换言之，人工智能道德主体地位的确立体现了人类的一种伦理姿态，旨在通过这种身份的界定来提升人类在技术发展中的道德自觉，而非实质性地赋予人工智能权利与义务。总之，作为一种反思性的伦理策略，赋予人工智能道德主体地位，不是为了模糊人与机器之间的界限，而是为了在技术飞速发展的时代背景下，唤醒人类对自身行为的伦理反思，从而实现更加负责任、可持续的人工智能发展路径。

Ethical Explorationon Artificial Intelligence as Moral Subject

WANG Fuling

（1. Center for Ethical Studies，Renmin University of China；
2. School of Philosophy，Renmin University of China）

Abstract：In the field of AI ethics，the question of moral status constitutes a fundamental philosophical issue. As AI systems increasingly intervene in human life and exhibit forms of autonomy，the traditional ethical framework that defines moral status primarily through the binary categories of moral agent and moral patient has gradually revealed its internal limitations. This framework

① 戴维·贡克尔：《关系转向：21世纪及之后的技术-伦理》，载《杭州师范大学学报》，2024（1）。

② 殷杰：《生成式人工智能的主体性问题》，载《中国社会科学》，2024（8）。

proves insufficient in addressing the novel ethical challenges posed by the development and deployment of AI technologies. Therefore, it is imperative to expand and reconstruct existing theories of moral status. A new model proposes a conceptual distinction between moral agent and moral subject, introducing the notion of moral subject as an intermediary category within the traditional binary structure. This leads to the construction of a tripartite hierarchical theory of moral status comprising moral agent, moral subject, and moral patient. By positioning artificial intelligence within the category of moral subject, this theoretical approach better reflects the evolving role of AI in both current and future human life. It responds more effectively to the practical demands of technological development and lays the groundwork for fostering a harmonious cooperative relationship between humans and intelligent systems.

Key words: Artificial intelligence; Moral subject; Moral agent; Moral patient