• 专题：人工智能伦理的适配难题 •

编者按：

随着人工智能体与人类现实相照面，在其智能性为人类带来巨大福祉之时，也引发了其道德性如何实现的全面思考。由于人工智能体不是传统意义上的道德主体，如何与之适配成为了人工智能伦理研究的重点与难点。本专题的三篇文章从不同的角度为人工智能伦理的适配问题作出谋划。李平安一文对弱人工智能下机器伦理之可能性进行了疏正。通过对比西方道德理论和先秦道家经典《老子》一书的观点，作者以“弱主体”甚至“无主体”作为德的效应体，来避开西方道德理论在机器伦理的算法设计与主体问题上的适配难题，这为弱人工智能下解决机器道德理论带来新的视野和可行的思路。寿步的文章聚焦于人工智能伦理主体的“正名”问题。借助“循名责实”的方法，追溯了通常被译为“主体”的agent一词在人工智能领域的使用状况，在行为体与智能行为体的实指不一的情况下，作者提倡以“行为体”作为agent的译名并通过“行为体社会”来捕捉人工智能的意向性特征，从而营建起行为体向智能行为体转化的桥梁，为探索人工智能的伦理主体提供更为精准的定位。汪琛、孙启贵与徐飞的文章全面考察了人工智能伦理研究在医疗领域的发展趋势。采用科学计量学方法对医疗人工智能伦理研究文献的主题分布、演进脉络与知识基础进行了系统的梳理和分析，发现研究范式呈现出从规范性伦理研究转向建构性多学科研究的特征，据此提出下阶段医疗人工智能伦理研究的要点问题，如传统伦理学理论的深度融合、伦理准则的治理转向以及心理学方法的介入等，是研究新技术与传统道德伦理理论相适配、融合的重要参照。

（专题策划：李平安）

弱人工智能机器伦理的“老”路新探

——在多释与简单原则之间

Between the Principles of Multiple Interpretations and Simplicity

A Further Discussion on Constructing the Machine Ethics of Weak-AI from the Viewpoints of *Laozi*

李平安 /LI Pingan

（东南大学人文学院，江苏南京，211189）
(School of Humanities, Southeast University, Nanjing, Jiangsu, 211189)

摘　要：机器伦理之塞有二：一是理论中无主体性的机器缺乏适配的道德基础，二是实践中游移于多

收稿日期：2023年3月3日
作者简介：李平安（1994–）男，山西临汾人，东南大学人文学院博士研究生，研究方向为道家哲学、比较哲学。Email: leeping_an@163.com

释原则与简单原则间的具体算法设计患于得失进退。在诸道德理论中，以“无我”“自然”为要目的《老子》德思想，避开了西方道德理论嵌入机器伦理时的主体性难题，以“失道后仁”定位机器伦理于与问题相照面的伦理拯救，通过“三宝”之“慈”领摄多释原则、之“俭”活化简单原则、之“不敢为天下先”守正进退之矩，对弱人工智能的机器伦理之可能性予以疏正，最切时下弱人工智能阶段机器伦理理论设计之亟需。

关键词：机器伦理　弱人工智能　《老子》　多释原则　简单原则

Abstract: There are two problems facing the machine ethics in the stage of weak AI: one is that the machine without subjectivity lacks a suitable moral foundation in theory, and the other is that the specific algorithm design that shifts between the principle of multiple interpretations and that of simplicity lacks gains and losses in practice. Among the moral theories, the moral thoughts in *Laozi* with “selflessness” and “naturalness” as their main points avoid the problem of subjectivity when western moral theories were embedded into machine ethics, thus positioning machine ethics in the ethical salvation of the problem by “sticking to kindness after losing virtues”, guiding the principle of multiple interpretations with “kindness”, revitalizing the principle of simplicity with “frugality”, and complying with the rules of progress and retreat by “daring not taking the lead in the world”, the latter three of which are the essence of “Three Treasures” in *Laozi*. These thoughts would clarify the possibility of machine ethics and meet the urgent need for a theoretical design of machine ethics in the weak AI stage.

Key Words: Machine ethics; Weak AI; *Laozi*; Principle of multiple interpretation; Principle of simplicity

中图分类号：TP242; B82-057　　**文献标识码：**A　　DOI: 10.15994/j.1000–0763.2023.12.001

一、弱人工智能机器伦理之塞

机器伦理（Machine Ethics）一词源于科幻文学，最初只是科幻作家对强人工智能时代人机关系、规则的设想；受益于相关技术发展，当前弱人工智能愈发展现出步入强人工智能的潜力，“机器伦理”亦随之成为与人类社会相照面的现实问题，其实际含义也发生更新。在技术伦理中，机器伦理意指一个新兴的研究领域：不同于以规范人类对技术人工物的使用或考察技术的社会效应为研究范畴的传统技术伦理，机器伦理试图“为机器增加道德维度”[1]或使机器能够在与人类、其它机器交互时采取合乎人类道德的处置方式，[2]探索机器从人工智能行为体（Artificial Intelligent Agents）向人工道德行为体（Artificial Moral Agents）转变的可能性。

在当前的研究范式中，研究者主要借助人类社会的道德理论尝试兼容、适配机器伦理。在“何为机器伦理”的问题上，机器伦理总是某种人类道德理论的延展，如首次系统性阐释机器伦理的安德森（Michael Anderson）以经典伦理学理论为范本设想了基于边沁功利主义与基于罗素义务论的两种机器伦理模型；[1]在“怎样实现机器伦理”的问题上，机器道德往往只是被嵌入机器的人类道德，如瓦拉赫（Wendell Wallach）以道德敏感性、自主性为变量将道德行为体区别为操作性道德（Operational Morality）、功能性道德（Functional Morality）与完全道德行为体（Full Moral Agency），将机器道德作为机器中一种被嵌入的特殊功能。（[3]，pp.9–10）在受益于人类社会的道德理论之时，机器伦理也面临着与之兼容、适配时的不适：以人类社会的道德理论视之，机器是不具备心灵、意识的技术人工物，没有自由意志、更不能承担责任，缺乏形成道德能动性的基本要素，不是恰当的道德主体。[4]–[6]此外，“人工道德行为体不仅面临道德理论本身的争议，也受到道德理论在实际操纵中的计算限制”，[7]现有的道德理论被认为不能为机器伦理的实现提供行之有效的具体方案。

但是，上述批评并不构成对机器伦理的反驳，仅复述了弱人工智能与强人工智能的差异。

弱人工智能机器不能成为自主的道德者，是因为它尚且不是自主的智能者。但是，人工智能的强、弱之分只是描述了问题，不能提供解决方案。在弱人工智能阶段，机器伦理不是以自足的状态出场的，而是以问题意识的形式因其自足状态的缺场而在场的：新闻工作者安格文（Julia Angwin）的团队发现，大数据与机器学习算法的结合不仅未能助益数据的客观化，反而成为了“永久化歧视的潜在来源”；[8]微软研究院首席研究员克劳福德（Kate Crawford）发现智能算法所提供的风险预测正在加剧“工作、家庭以及司法中的不平等”。[9]算法偏差是否真实、普遍地存在仍需专门研究与确认，但可以肯定的是：弱人工智能的道德表现并不符合人类对它的道德预期，并已经影响了强人工智能的构想路径。波斯特伦（Nick Bostrom）以道德效能为预期指出，基于遗传算法的人工神经网络（Artificial Neural Network）虽然具有更强的适用性，但更符合人类现代社会准则的、具有透明性与可预测性的决策树（Decision Tree）更应成为人工智能的发展方向。[10]

也就是说，对机器道德能力的预期在相当程度上影响了其智能能力的实现方式。在机器伦理中，就先在性而论，以智能能力为特征的人工智能行为体较人工道德行为体所具备的只是时间先在性，而以道德能力为目的的人工道德行为体反而具有逻辑上的先在性。因此，机器伦理不宜作为强人工智能的专属物，将弱人工智能纳入机器伦理的研究范畴不仅是“善法”先行之必需，也决定着未来强人工智能的实现方式。

尽管如此，弱人工智能的确不具备主体要素，对其进行道德思考的可能性依赖于道德视角的转变，要么将主体道德降格为行为体道德，要么采取一种不同于西方传统的道德理论。与此同时，弱人工智能机器伦理的实现方式，应区别于传统技术伦理中以人类为规范对象的实践方法，而以机器算法作为规范对象。在西方道德理论奠基于“主体”与“自由”之前，先秦道家经典《老子》已然描绘出一幅以“无我”与“自然”为特色的德图谱，将“弱主体”甚至“无主体”作为德的效应体，更以“慈”“俭”统摄德目设计中的多释原则、简单原则，不仅为弱人工智能机器伦理的道德理论基础带来新的视野，更为机器算法设计中德目统筹的实现提供了可行的思路。

二、以“辅万物之自然”开启弱人工智能机器伦理的新视野

一般来说，机器伦理被认为是强人工智能的专属物。在西方的道德理论中，康德以自由作为道德法则的存在基础，[11]研究者大多据此以道德意愿（善）与道德能力（理性）作为道德存在的必要条件。在此思维下，机器伦理也相应成为了强人工智能的专属物：以道德意愿而言，弱人工智能不具备意向性[4]故而不具备意愿的能力，（[3]，p.95）也就无法成为自由王国的成员；以道德能力而言，道德主体被认为应具备至少五种理性能力，即“对道德正误的判断能力……对判断理据的权衡能力……对权衡结果的决策能力……对决策的实现能力……对未实现原因的解释能力”，[12]这亦是弱人工智能所不具备的。因此，弱人工智能无法进入机器伦理范畴的推理是简明易见的：

（1）大前提：如果具备机器伦理，那么是强人工智能；

（2）小前提：弱人工智能不是强人工智能；

（3）结论：弱人工智能不具备机器伦理。

但是，如果机器伦理无须以强人工智能为前提，那么原本的大前提应变换为“如果是强人工智能，那么可能具备机器伦理”，原本小前提对大前提后件的否定就变成了对前件的否定，不再具备形式上的有效性。那么，机器伦理是强人工智能的专属物吗？

1.“道废有仁”：定位机器伦理的存在状况

将道德视为理智的专属物，这一传统可追溯至苏格拉底著名的“美德即知识”。色诺芬（Ξενοφών）将苏格拉底的推理记录为：所有生物都以它们所认为的好为目的；如果不知道什么是好的，就会因为错误目的而做坏事；如果知道什么是善的，就会因为正确目的做好事。[13]

在本性向善的预设下，恶行来自于对善恶的无知状态。可以发现，苏格拉底认为道德判断与道德行为间具有一致性。不过，二者间的一致性依赖于向善的预设。康德在《道德形而上学奠基》中以理性的智能能力并不能总是达成其目的为由，推断理性之目的在于向善的意愿而不是智能能力，但此推理依赖于宗教性的目的论证明。无独有偶，叶树勋发现，在前《老子》时期的“早期文献的多数语境中”，“德”往往专用于描述统治者的品行，及至老子才“出于挽救时弊的问题意识”对“德”予以“经世致用”的重诠。[14] 机器伦理的滞塞与此相类，机器之所以被认为不具备道德属性，既有其自身因素，也源于道德被依附于一些特殊的存在或存在者。

道德固然需要理智的介入，但若只以“德”论之，理智的参与便不再是必要条件。在西方伦理思想史中，此类观点可追溯至亚里士多德以“习惯（ethos，Habit）”为“伦理（ēthikē，Ethic）”的词源，提出道德美德（也译为实践美德、伦理德性，Moral Virtue）由风俗、习惯袭承而来，是事物本就具有的功能。（[15]，p.23）苗力田在对亚里士多德作品的译介中补充道：“德性并非灵魂的专利，物体也同样具有自己的德性。”[16] 这就为更为广泛的伦理论域提供了可能。但是，其美德理论是形上思维的典型，而弱人工智能的伦理显现并不是本体功能的外化，而是在生活中与人照面的恼人问题。相较于此，《老子》中不仅有“玄德”“上德”的说法，也存在“孔德”“下德”以及“不德”“无德”（[17]，p.108、215、156）的用例，将“德”的描述对象从君王、圣人拓展至了宇宙万物，显示出了更为多样、多维的“德”之存在状况。

于弱人工智能的机器伦理而言，要获得其理据的完备性还须回到善一致性的预设之处，不过，是完全相反的方向。《老子》十八章言：“大道废，有仁义。”（[17]，p.145）天地不仁、道废有仁是对善一致性预设的彻底否定。作为人类道德德目，仁并不是天地大道演绎的产物，而是对大道废弃的补救。道德判断与道德行为也不再总是相互对应的，《老子》十八章又言：“智慧出，有大伪。”（[17]，p.145）一个具有道德认知能力、道德知识的行为者不一定会作出道德行为，作出的道德行为也不一定出于道德意愿。如果对上述两种道德理论的存在状况加以比较，《老子》的设想显然更切合机器伦理在当下的实际境遇。

首先，在本体论状况上，机器伦理源于与道德预期不符的现实问题、而不是机器实体的某种属性。如果机器伦理源于机器实体的属性，弱人工智能是不具备这种属性、能力的；但即便是强人工智能的机器伦理，也需要预设强人工智能本身含有向善的动机，但这种做法会导致强、弱人工智能的区分并不具备真实的意义、为机器嵌入善法更显画蛇添足，因为这种做法意味着机器的伦理状态已于预设的向善性中、预定的和谐性中实现。

其次，在生成论过程中，机器是人工的造物，其拥有的属性是从无至有被创造出来的，其目的是借助其工具性能帮助人类解决问题。机器的伦理属性也是如此，它是被人类创造、由算法设计的，用以弥合机器道德表现与人类道德预期间的疏离，是面向问题作出的解答方案，亦即“道废有仁”的人类谋划。冈克尔（David J. Gunkel）表达过相似的观点，“机器人或任何实体是否应当拥有道德/社会地位做出决定……不取决于事物是什么，而取决于现实的社会情形中我们怎样与之联系、反馈”，[18] 机器的道德属性来自于人机交互间的关系，而非出于机器某种已有的属性。

不同的是，当冈克尔以一种肯定思维将现存关系作为判断道德地位之依据时，《老子》以“道废”为背景将之视为现存的问题。大道废弃，理想的和谐状态变得混乱，现存的状态并不能提供对事物本真状态的启示，如若盲目因循，极易陷入从问题到问题的往复循环。故此，《老子》十九章有：“绝圣（智）弃智（辩），民利百倍；绝仁（伪）弃义（诈），民复孝慈；绝巧弃利，盗贼无有。此三者以为文不足，故令有所属：见素抱朴，少私寡欲，绝学无忧”。（[17]，p.147）依循现存关系而建立、出现的圣智、仁义、伪诈、巧利不能引导人类社会重

回健康、良性的状态，而应在万物的素朴本性中葆养其本真；同样，在人机现存关系中出现的疏离，即意味着对弱人工智能的道德预期不具备有效规范效能，单纯依循这种道德预期，必然不能实现人机的和谐共存。而应回到机器伦理以机器为轴心的观察视野中，在机器中育成机器的伦理状态。

2.“万物之自然”：演绎弱人工智能机器伦理的新视野

《老子》的道德理论之所以能为弱人工智能的机器伦理提供支持，离不开其以“万物”为域的理论视野。“万物”的胸怀并不为《老子》或先秦道家独有，前有《易》之“大哉乾元，万物资始”，[19]后有阳明先生“大人者，以天地万物为一体者也”，[20]可以说，“万物”视野贯穿于中国传统智慧的始终，是千年来滋养中国传统道德伦理观念的重要思想养分。在《老子》中，“万物”概念更有着别样的意蕴，借助“自然”与“辅”的谐律，从受限于自我的自己而然拓展至了辅助万物自我形成的“辅万物之自然”，（[17]，p.301）从而在两个方面使弱人工智能阶段的机器伦理得以可能：

其一，“万物之自然”。意味着作为“万物”一员的弱人工智能机器具有其自身的运演规律以及与由道而来、以道为归的本真状态。《老子》将“万物”与“道”的动态关系描述为“大道氾兮，其可左右。万物恃之以（而）生而不辞……万物归焉而不为主。”（[17]，p.203）“道”对“万物”公正而无所偏私，万物由其滋养、生生不息、并以之为归宿。需要注意的是，“道”并非预定和谐。在现实的场景中，“道”不仅是隐居其后的，在直观的层面更是已然废弃的，这反而使“万物”并不会被笼罩于某种固定的、肯定的独断论之下，能够以“道法自然”（[17]，p.169）的方式被赋予更具包容性、多样性的自性。在“自然”的演绎下，《老子》之“德”也运生出独具特色的思想意趣，郝然称其“是对天命论彻底的颠覆。”[21]得益于此，弱人工智能机器伦理也不必依赖于善一致性的预设，只须根源于其自身的自然之性。

其二，“辅”的出现。“辅”是辅助者对被辅助者的帮助，承转了弱人工智能机器伦理从主体理论向行为体理论的变更。不同于人类智能，人工智能是不具备意识的存在，因其主体性的缺失而不符合西方道德理论中的道德主体的基本特征。为了使人工智能也能被纳入道德考虑的范畴，阿斯凯尔（Amanda Askell）以被动性的现象意识分析替代主动性的自觉意识分析，提出人工智能可以通过奖励、惩罚的机制实现道德刺激与道德反馈，以道德行为体的样态作出对人类、其它机器的道德关怀。[22]不过，奖惩机制下非此即彼的行为模式并不能适应传统道德困境在人工智能背景下的再次“升级”。以“电车难题”为例，如果轨道拉杆的控制者是人类，理性的道德抉择往往会因人类本能的应激反应被回避，在作出理智判断与相应的行为前，仓促间的人类控制者会因身体本能作出下意识的反应从而不被指责；但如果控制者是人工智能，那么一切选择都已被算法预设，无论怎样选择都意味着，有人在事故发生前就已然在算法中失去了生命权。

《老子》之“辅”是对此现状的改变。一方面，“辅”认可了被辅助者的积极作用，不具备主体性的人工智能不是道德问题的局外人，其在感知、判断、行动、数据互联等方面的卓越性能是守护人类道德实践的重要载体，甚至在一些场景中是不可替代的；另一方面，“辅”需要辅助者的在场，在“道废”的背景下，仅依循任何参与者的现存状况进行处置难以实现对“道”的契合，具备一定道德知识的人类有义务对机器伦理作出预先的设计与不断的完善，辅助弱人工智能机器作出恰当的道德选择、趋近其伦理状态，以育成人工道德行为体的实现。当然，这并非易事。

三、以“慈”“俭”统筹弱人工智能机器伦理的算法设计

在《老子》的设计中，辅助者是与道亲和的“圣人”“善为道者”或“有道者”，尽管他们“微妙玄达（通），深不可识”，（[17]，p.179、301、336、129）但在具体的操作领域，祈待

圣人的出现与帮助是不切实际的。越过认知的过程而意图直达真理是《老子》二十四章所批评的“馀食赘形”，（[17]，p.167）赋予事物以超出其自身的内容只会凭添负担。目前来说，以理想化的道德理论适配问题化的机器伦理是困难的：一方面，功利主义无法协调“无穷无尽的计算与现实道德决策中有限计算资源之间的矛盾”；另一方面，“当多个道德规则产生冲突时”，义务论难以在“考察时代的社会发展、生活习俗、文化传统、意识形态等”因素的情况下“确定道德规则之优先级”。[23]在此情况下，相较于确认机器伦理之具体德目、内容、序列并将其嵌入机器之中，在面向问题的动态过程中析理机器伦理的存在形式更符合当前的认知阶段。即，如果机器伦理存在，其德目、内容是如何、或应当如何存在并相互联系、统一的。

由于机器不具备主体能力，其伦理状态不是其自身伦理要素的简单加和，而存在于设计者、机器、应用情境等多维联系之中，其伦理要素的结合必然需要辅助者的介入。以机器决策的实现方式而言，其实质是通过算法赋予输入值与输出值相关性，因此，其伦理要素的相关性也反映在数据间的相关性上，并由算法所赋予。但目前为止，尚无一种算法可以在所有的应用情境中都表现出较其它算法的明显优势。不同算法原则所依赖的世界观、方法论也存在较大差异。当前的算法设计中，多释原则与简单原则是统筹算法设计的最底层逻辑，设计者在二者间的选择决定了庞杂的算法设计具体采取离散建模、还是连续建模。[24]同样，意欲在算法中实现机器伦理诸要素的统合，也须要设计者在多释原则、简单原则间觅求平衡。

1. 以“慈”处“多”：与算法设计中的多释原则相协调

多释原则源于古希腊哲学家伊壁鸠鲁，认为如果有多种理论可以描述经验结果，那么就应当将其全部保留。[25]在机器学习领域，即采用多种模型处理单一问题，是集成学习算法的基本理据之一。目前，自动驾驶安全系统对“离手检测”功能的保留，在L2及以下自动驾驶级别中完全禁止驾驶员的手脱离方向盘，正是通过人、机的双重决策模型为车辆行驶安全护航。在道德领域中，道德困境往往来自于单一道德理论所易产生悖谬的特殊情境，其解决方案也常常需要多种道德理论的结合。

一般来说，善是道德理论的最高原则，但由于历史、地域等因素的复杂影响，不同人类群体对善的认知是存在差异的。在此情况下，差异性的善只能作为机器伦理算法设计所采用的多种模型之一，而不能作为多种算法分类、赋权、协调的统合原则。面对不同算法生成的数据集，设计者会加入“硬”或“软”的投票策略进行处理，但无论是多数服从少数、或先行对不同数据予以差异性的加权，都难免出现某个德目对其它德目、某种道德理论对其它道德理论的先在性优势，从而不得不选择决策树之“白箱”的算法设计方式。其弊端在于存在被“牺牲”的道德要素以及面临连续性过强、类别过多、差异性过小及预定之外等情况时缺乏及时、有效的处理能力。

因此，多释原则应建立在比善更具包容性、从而更具可行性的统合原则之上。《说文解字》有“善，吉也”，（[26]，p.103）又“慈，爱也”。（[26]，p.423）从字义的分疏可见，“善”偏重于结果而“慈”着笔于发端，在作用的程序环节上，“慈”较“善”更为恰当。《老子》将慈作为“三宝”之首，以“婴儿”“赤子”喻指民众的理想状态；而将生成、辅育的形象称为“母”“玄牝”“雌”；（[17]，p.310、108、274、73、98、108）圣人对百姓、万物的作为被记为“圣人皆孩之”。（[17]，p.253）除此之外，通行本之“民复孝慈”在更早的竹简本中为“民复季子”，（[17]，p.147）文本的溯洄澄明了“慈”在《老子》中不是被辅育者的德目，而从属于辅助者。

不言而喻，在算法设计中，算法从属于智能机器、却是由编程者赋予的，这正是一种《老子》中母子间的关系。父母生育子女是让子女成为它自己，而非以父母为蓝本的复刻，《老子》亦有“功成事遂，百姓皆谓我自然”（[17]，p.141）的说法。与以善为原则的道德理论不同，“慈”并不是将某一价值体系嵌入既有的载体，

而是对不同载体本有的价值观念加以引导、辅育。正因如此，“慈”能够消解单一道德理论的价值束缚，借由辅助者之“德”实现被辅者之“得”。

那么，“慈”应该如何应用呢？机器算法对多道德模型的选择在本质上属于算法的加权问题。《老子》认为“人之道，则不然，损不足以奉有馀”，（[17]，p.336）算法所面对的样本及其特征往往是不均匀的，多数量的样本特征能够得到充分学习，而少数量的样本特征却难以被充分学习，从而造成了算法偏见。[26]对于此类问题，《老子》提出“损有馀而补不足。”（[17]，p.336）实际上，研究者对这种设计理念不会感到陌生：所莱曼尼（Ava Soleimany）的研究团队认为，人工智能中面部识别之所以出现歧视现象，是因为对数量上较少的样本特征而言，其数量不足以支持算法生成分类标准，其分类标准来往往自于人类的既有认知（如肤色、脸型）；不过，其反面也意味着存在尚未被人类充分识别的样本特征，通过将输入数据的映射拓展为向量分布而不只是向量，样本可以由机器学习解码、重构并生成新的样本特征，使其得到同样的加权基础，以此修正造成偏差的样本分类。[27]

2. 以“俭”致“广”：与算法设计中的简单原则相协调

简单原则，又称“奥卡姆剃刀原则”，由圣方济各会修士奥卡姆的威廉（William of Occam）提出，并深刻影响了机器学习领域中的分类学习，认为应采用“与样本数据一致的最简单的假设”。[28]算法的实现能力受到机器实体的性能约束，复杂的道德嵌入无疑会加重运算负担，加剧能耗的同时也会造成机器输出的延迟、滞后。在提高机器实体的物理性能之外，简化算法可以使机器运算更多的数据、处理更复杂的道德情景。

不过，高效、简约虽已是共识，但关于什么是“最简单的”的分歧依然存在。奥卡姆剃刀剥离了理论中冗余的实体概念，也在一定程度上剥离了概念的实体性，[29]黄俊维指出：“科学家在谈论奥卡姆剃刀的时候，却并非针对科学理论背后的本体论预设，而是关注一个理论或假说的复杂程度，以及它们所依赖的前提的数量”。[30]但对于机器伦理而言，不论是本体论式的简单原则、或句法式的简单原则，都难与机器伦理相适配。由于强人工智能尚属构想、而弱人工智能缺乏自足性，在机器伦理的实体性存疑之时，任何对其本体论状况的推测与描述都是与简单原则相悖的理论冗余。在实际使用中，简单原则的使用需要将奥卡姆剃刀原则倒置，将观察、描述的对象从机器伦理的实体性转向对机器伦理的问题意识、需求及预期。从被辅助者视角移向辅助者视角，利用其逆反命题——如增实体、务有必需，将必要性作为衡量“最简单”的判别标准，要求算法设计与其目的严格匹配。在国内的案例中，淘宝网的架构演化常常被视为简单原则的范例，其早期所使用的LAMP架构（Linux+Apache+MySQL+PHP）具有“开源、免费、简单”的特点，在需求多变的情况下可以凭借“工程师学习周期短”“沉没成本低”等优势实现快速的技术方案响应。[31]即便如此，简单原则依然存在含混不清的地方，将对“最简单”的困惑转向了对“必需”的困惑。

在《老子》中亦有“不得已而用之”（[17]，p.195）的主张与简单原则相应和，更在此基础上对何为“必需”进行了求索。《老子》虽将“俭”作为“三宝”之一，但用字仅出现三次。《老子》用以表达俭约之意的还有“啬”字，并将之作为“长久”的根由：“治人事天，莫若啬；夫唯啬，是以早服；早服是谓重积德，重积德则无不克；无不克则莫知其极，莫知其极，可以有国。有国之母，可以长久。”（[17]，p.288）人心有欲，节俭是与本性欲望是相违的，但“啬”不仅有节约、节俭的意思，《说文》有：“啬，爱濇也，从来从亩。来者，亩而藏之”，（[26]，p.171）“濇”即“滞涩”，将其作为“爱”的对象是因为爱惜收获的谷物而贮藏于谷仓之中，故农夫也被称为“啬夫”。（[26]，p.171）可见，节俭不是作减法，而是在“爱”的发动下珍藏已有的收获。

“俭”与“爱”的密切关系也体现在《老

子》“三宝”的次序中，前有爱之“慈”、后有爱之“不敢”，“俭，故能广”即是爱惜自己所有、积少成多之“广”，也承连爱惜万物所有，从个体的“勇”至天下“器长”之“广”。失去此前提，即使拥有卓越的自然语言处理能力的ChatGPT也会蕴含巨大的伦理风险，佩里戈（Billy Perrigo）曝出OpenAI训练ChatGPT时造成了大量数据标注员的心理创伤，技术产品迭代的背后是被忽视的、对庞大劳动力的依赖。[32]人类希望得到机器的“善待”是机器伦理研究最为朴素的起点，[33]不过，免于饥饿即要爱惜食物、珍视食物即要体恤民力，这同样是朴素的道理。“甘其食，美其服，安其居，乐其俗”（[17]，p.345）是《老子》对“三宝”之“俭”的现实刻画。相较于创造新的可欲之物，《老子》意图将欲望的对象返回至已有之物，将无穷尽的贪念返回至对本有之物的爱惜。新出现的“必需”是建立在已有之上的，苛刻于“必需”的计较本源于对已有之物的珍视。故此，简单原则的施用不应是禁欲主义或道德上的原教旨主义，而是爱惜人类已经创造出的美好生活及已付诸实践的道德意愿、道德行为，这是创新、发展所不应舍弃的起点。

四、结语：守正机器伦理的进退之道

长久以来，理智能力被认为是人类在自然界拥有特别地位的原因，在此思维下，机器伦理的实现也常常被寄托于强人工智能的出现、或经由人工智能行为体向人工道德行为体过渡。理智能力、智能能力总是先在于道德能力或伦理状态的。但在《老子》中，“圣人”之所以成为万物的“器长”，并非因其优先的地位，反而是由“不敢为天下先”成就的。万物万形，即使在人类族群中亦会因地理、历史等因素产生出不同的文化形态、伦理道德观念。仅因循一形一理，必使人类族群中的其它文化形态失去本有的价值与话语权，使万物中的其它成员失去本有的颜色与生命力。退于天下之后，以万物为先，才能因循万物的已有之形、本有之理。因此，构建弱人工智能机器伦理，不应仅是将人类某种文化形态下的道德理论嵌入其中，而应以“慈”统合多释原则加以包容、以“俭”领摄简单原则予以爱护，并将其纳入机器算法的设计环节。

《韩非子·解老》有言：“慈母之于弱子也，务致其福，务致其福则事除其祸，事除其祸则思虑熟，思虑熟则得事理……圣人之于万事也，尽如慈母之为弱子虑也”。[34]欲育成其弱人工智能机器伦理，也应如慈母为弱子忧思，围绕现阶段弱人工智能所显露的机器伦理问题意识构建讨论域与实践场，在辅助弱人工智能伦理良性态的过程中促成研究者、设计者、使用者的伦理知识储备、道德意识养成，以此守正机器伦理的进退之道。

［参 考 文 献］

[1] Anderson, M., Anderson, S., Armen, C. 'Towards Machine Ethics: Implementing Two Action-based Ethical Theories'[A], Michael, A., Susan, L. A., Chris, A. (Eds.) *Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics*[C], Menlo Park: The AAAI Press, 2005, 1–7.

[2] Anderson, M., Anderson, S. 'Creating An Ethical Intelligent Agent'[J]. *AI Magazine*, 2007, 28(4): 15–26.

[3] Wallach, W., Allen, C. *Moral Machines: Teaching Robots Right from Wrong*[M]. New York: Oxford University Press, 2009.

[4] Johnson, D. G. 'Computer Systems: Moral Entities But Not Moral Agents'[J]. *Ethics and Information Technology*, 2006, 8: 195–204.

[5] Nath, R., Sahu, V. 'The Problem of Machine Ethics in Artificial Intelligence'[J]. *AI and Society*, 2020, 35(1): 103–171.

[6] Sparrow, R. 'Killer Robots'[J]. *Journal of Applied Philosophy*, 2007, 24(1): 62–77.

[7] Allen, C., Varner, G., Zinser, J. 'Prolegomena to Any Future Artificial Moral Agent'[J]. *Journal of Experimental & Theoretical Artificial Intelligence*, 2000, 12(3): 251–261.

[8] Julia, A., Jeff, L., Surya, M., et al. 'Machine Bias'[EB/OL]. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. 2016–05–23.

[9] Kate, C. 'Artificial Intelligence's White Guy Problem'[EB/OL]. http://nyti.ms/28YaKg7. 2016–06–25.

[10] Nick, B. *Superintelligence: Paths, Dangers, Strategies*[M]. Oxford: Oxford University Press, 2016, 4.

[11] 康德. 康德著作全集（第5卷）[M]. 李秋零 译, 北京: 中国人民大学出版社, 2007, 5.
[12] Taylor, P. *Respect of Nature: A Theory of Environmental Ethics*[M]. Princeton: Princeton University Press, 1986, 14.
[13] Xenophon. *Memorabilia Oeconomicus Symposium Apologia*[M]. Cambridge, Massachusetts: Harvard University Press, 2007, 225.
[14] 叶树勋. 老子对“德”观念的改造与重建[J]. 哲学研究, 2014,（9）: 55–62.
[15] Aristotle. *The Nicomachean Ethics*[M]. Oxford: Oxford University Press, 2009.
[16] 亚里士多德. 亚里士多德选集·伦理学卷[M]. 苗力田 编, 北京: 中国人民大学出版社, 1999, 26.
[17] 陈鼓应.《老子》今注今译[M]. 北京: 商务印书馆, 2007.
[18] Gunkel, D. J. *Robot Rights*[M]. Cambridge, MA: MIT Press, 2018, 170.
[19] 周振甫. 周易译注[M]. 北京: 中华书局, 2013, 3.
[20] 王守仁. 王文成公全书·卷之七[M]. 王晓昕 赵平略 点校, 北京: 中华书局, 2015, 305.
[21] 郝然.《老子》德论[EB/OL], http://www.cssn.cn/zhx/zx_llx/202009/t20200916_5183531.shtml. 2020–09–16.
[22] Askell, A. 'My Mostly Boring Views About AI Consciousness'[EB/OL]. https://askellio.substack.com/p/ai-consciousness. 2022–02–21.
[23] 阮凯. 机器伦理何以可能：现有方案及其改良[J]. 自然辩证法研究, 2018, 34（11）: 53–58.
[24] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016, 17.
[25] Asinis, E. *Epicurus' Scientific Method*[M]. Ithaca, NY: Cornell University Press, 1984, 321.
[26] 许慎. 说文解字（中华书局影印）[M]. 徐铉杨 校, 陈昌治 刻, 北京: 中华书局, 2014.
[27] Soleimany, A., Amini, A. A., Schwarting, W., et al. *Uncovering and Mitigating Algorithmic Bias Through Learned Latent Structure*[M]. Honolulu, HI, USA, 2019
[28] Blumer, A., Ehrenfeucht, A., Haussier, D., et al. 'Occam's Razor'[J]. *Information Processing Letters*, 1987, 24(6): 377–380.
[29] Ockham, W. *Predestination, God's Foreknowledge, and Future Contingents*[M]. Indianapolis: Hackett, 1983, 46.
[30] 黄俊维. 简单性原则的潜在检验辩护[J]. 自然辩证法通讯, 2020, 42（10）: 31–37.
[31] huashiou. 服务端高并发分布式架构演进之路[EB/OL], https://segmentfault.com/a/1790000018626163. 2019–03–13.
[32] Perrigo, B. 'Exclusive: OpenAI Used Kenyan Workers on Less Than $2 per Hour To Make ChatGPT Less Toxic'[EB/OL]. https://time.com/6247678/openai–chatgpt–kenya–workers. 2023–01–20.
[33] Moor, J. H. 'The Nature, Importance and Difficulty of Machine Ethics'[J]. *IEEE Intelligent Systems*, 2006, 21(4): 18–21.
[34] 王先慎. 韩非子集解[M]. 钟哲 点校, 北京: 中华书局, 1998, 151.

[责任编辑 王巍 谭笑]