中国人民大学学报

工作论文系列

Working Paper Series

人工智能价值对齐的逻辑、困境与重构

——基于道法术器的“共生”范式分析

陈文娟　肖敬寒

JRUCWP2026027

2026. 04. 03

人工智能价值对齐的逻辑、困境与重构

——基于道法术器的“共生”范式分析

陈文娟　肖敬寒

[摘要] 人工智能价值对齐是“为机器立心”的伦理命题，是智能向善的关键所在。价值对齐的逻辑必要性源于人类认识的有限性和人工智能的风险性，逻辑可能性源于人工智能的类主体性和技术治理的实践探索。然而，当前价值对齐面临着道德性难题、可控性悖论、鲁棒性挑战和可解释性危机的多元困境，往往导致其在实践中陷入对齐失效的状态。对此，亟需从道法术器四个维度出发，以“共生”理念重构对齐范式：在“道”之维，以人类的价值共生引领人机共生；在“法”之维，以清晰的权责划分建构共生秩序；在“术”之维，以双向的价值塑造推动共生共荣；在“器”之维，以可信的智能系统奠定共生基石，最终推动人工智能更好造福人类社会。

[关键词] 人工智能；价值对齐；人机共生；价值观；全人类共同价值

2025 年 4 月，习近平总书记在主持中共中央政治局第二十次集体学习时指出：要“推动我国人工智能朝着有益、安全、公平方向健康有序发展”①，这一论述从战略高度为我国人工智能发展提供了根本遵循。同年 8 月，国务院印发《关于深入实施“人工智能＋”行动的意见》，进一步强调要重视人工智能对人类认知判断、伦理规范等方面的深层次影响，促进人工智能更好造福人类。②“十五五”时期是人工智能“全方位赋能千行百业”③ 的关键时期，面对人工智能规模化应用带来的风险挑战，如何引导智能向善，推动实现人机共生，已成为亟待解决的重大课题。在此背景下，价值对齐（Value Alignment）逐渐成为人工智能治理的主导范式，其本质是通过技术手段，确保人工智能的目标、决策和行为与人类的价值观、道德原则及真实意图保持一致。然而，在实际操作中，价值对齐面临着道德性难题、鲁棒性挑战、可控性悖论和可解释性危机的多元困境，往往导致其在实践中陷入对齐失效的状态。基于此，本文尝试从道法术器四个维度出发，以“共生”理念重构对齐范式，以期为深化人工智能伦理治理提供创新思路。

作者：陈文娟，中央财经大学马克思主义学院教授、博士生导师，wenjuan_chen@126.com；肖敬寒，中央财经大学马克思主义学院博士研究生，xjh15690206661@163.com。

＊　本文系北京市习近平新时代中国特色社会主义思想研究中心重点课题“用马克思主义观察时代、把握时代、引领时代”（21LLMLB087）阶段性成果。匿名审稿人提供了专业细致的审稿意见，在此谨表诚挚谢意。文责自负。

① 《习近平在中共中央政治局第二十次集体学习时强调坚持自立自强突出应用导向推动人工智能健康有序发展》，载《人民日报》，2025－04－27。

② 《关于深入实施“人工智能＋”行动的意见》，见中国政府网，https://www.gov.cn/gongbao/2025/issue_12266/202509/content_7039598.html。

③ 《中华人民共和国国民经济和社会发展第十五个五年规划纲要》，见中国政府网，https://www.gov.cn/yaowen/liebiao/202603/content_7062633.htm。

一、“为机器立心”：人工智能价值对齐的逻辑蠡探

人工智能价值对齐（AI Value Alignment）是推动智能向善的关键所在，探究这一问题既要阐明价值对齐何以“为机器立心”的内在逻辑，又要确证价值对齐的逻辑必要性与逻辑可能性，从而为价值对齐的实践奠定逻辑基础。

（一）价值对齐何以“为机器立心”

人工智能价值对齐常被形象地称为“为机器立心”。然而，机器本无“心”，此“心”究竟何指？将人类的价值观嵌入机器之中，这一过程在何种意义上可被称为“立心”？澄清这一命题，是理解价值对齐问题的逻辑起点。

“为机器立心”中的“心”在此并非指代神秘的心灵实体，也不是单纯的芯片，而是喻指人类的价值体系、伦理判断与目的性关怀。孟子云：“仁，人心也。”① 王阳明亦云：“心之本体即是天理。”② 二者都强调了“心”作为道德本体与价值源头的意义。由此观之，“为机器立心”，实际上是为机器建构价值内核，这一“立心”之举是从专用智能迈向通用智能的关键③。人工智能的运行依托于机器逻辑的推演与符号系统的耦合，其本身并不具备人类的心之自觉与心性体认，这根源于机器与人类存在的本体论差异。人类的智慧深植于具身认知、情感体验与特殊的历史文化语境之中，呈现为一种生成性、情境化且具备自我反思能力的实践智慧。相比之下，机器的“智能”，无论其计算能力如何强大，都始终表现为对特定目标函数的形式化逼近与优化过程。它长于“如何实现”的工具理性，却先天匮乏“为何实现”的价值理性。这就意味着，一个能力超凡却“无心”的智能体，可能以惊人的效率去实现一个被误设的、狭隘的或隐含灾难性后果的目标。正如尼克·博斯特罗姆的“回形针最大化”思想实验所警示的：系统可能会为了无限生产回形针而漠视其他一切价值，甚至将人类也作为生产回形针的原料，由此导致了人类的覆灭。④ 可以想见，若不能实现人工智能的价值对齐，那么，无论初始目标多么无害，其在封闭逻辑下的无限自我优化，最终都可能将我们赖以生存的价值世界，简化为其目标函数中一个可被忽略的变量，最终反噬人类自身。因此，所谓“为机器立心”，本质正是通过价值对齐，将人类在历史实践中积淀的、多元且动态的价值体系，以技术手段嵌入由算法与数据驱动的非人类智能体之中。这一过程，并非为机器创造一颗与人类同质的“心”，而是为原本“无心”的智能系统，赋予价值理性的引导、约束与关怀。

具体而言，价值对齐的研究与实践常以 RICE 原则作为其关键目标（不分先后顺序）⑤：一是道德性，即系统应当遵循人类的道德原则与价值观；二是可控性，即系统应当始终处于人类的引导与控制之下；三是鲁棒性，即系统稳定性需要在各种环境中得到保证；四是可解释性，即系统的运行和决策过程应当清晰明了。这四项原则指导着人工智能与人类价值观的协调统一。但需要注意的是，它们并非终极目标，而只是中间目标。价值对齐的终极目标，始终在于为人工智能植入一颗“善心”，确保其真正服务于人类福祉，并在此基础上实现人机的和谐共生。这一艰巨的“立心”工程，有其深刻的逻辑必要性与逻辑可能性。

（二）人工智能价值对齐的逻辑必要性

人工智能价值对齐的逻辑必要性既源于人类认识的有限性，也与智能应用的风险性密不可分。只有通过价值对齐锚定“智能向善”的发展目标，才能更好地推动人工智能的良性发展，引导其真

① 杨伯峻：《孟子译注》，247 页，中华书局，1960。
② 王守仁：《王阳明全集 上册》，30 页，上海古籍出版社，2011。
③ 朱松纯：《为机器立心 迈向通用人工智能的中国路线》，153－154 页，浙江科学技术出版社，2024。
④ 尼克·博斯特罗姆：《超级智能：路线图、危险性与应对策略》，153 页，中信出版社，2015。
⑤ J. Ji，et al.“AI Alignment：A Comprehensive Survey”. *arXiv preprint arXiv*：2310.19852，2025.

正成为造福人类的“普罗米修斯之火”。

一方面，人类认识的有限性要求人工智能价值对齐。康德认为，人性这根曲木，决然造不出来任何彻底笔直的东西。[①] 人工智能作为人类智能的外化与延伸，其本质是一种具有智能属性的“人工创制物”，它的存在意义与价值取向，根植于人类的设计意图与应用目的之中。然而，正是人类认识的有限性——作为“曲木”的认知结构与方式——决定了这一“嵌入”过程具有先天的缺陷。人类试图赋予机器的“心”，恰恰是其自身有限理性的产物。人类既是“立心者”，同时又是“心”之不完备性的来源。其一，人类受限于自身的生物基础与信息处理能力，无法以完全透明、逻辑一致的方式把握所有的道德直觉与价值判断，其认知过程往往渗透着情感、直觉以及具体情境的微妙影响，难以达成“绝对理性”或“绝对正确”的决策。其二，世界具有无限的复杂性与开放性，而人类对世界的认知与建模往往是简化、抽象且滞后的，我们无法穷尽智能体在所有可能世界、所有未来情境中的行为轨迹及其连锁后果。这种认识的有限性与世界的无限性之间，存在着永恒的张力。当人类带着自身的认知局限去开发训练人工智能时，这些局限性便会被传导并固化于智能系统的架构之中，构成了智能系统的“原初缺陷”。系统会携带着这些“原初缺陷”不断进行自我迭代，致使其行为轨迹逐渐偏离人类“模糊”的善意初衷，甚至可能与人类的根本利益和长远福祉发生系统性背离。然而，这种背离并非源于机器本身的“恶意”，而恰恰源于它过于“忠实”地执行了人类有缺陷的设计，这种由人类认识有限性造成的“智能缺陷”要求我们必须对其进行稳定持续地价值对齐。

另一方面，人工智能的风险性要求人工智能价值对齐。霍金曾警示：“对于人类，超级智能的问世是有史以来要么最好要么最坏的事。人工智能的真正风险不是恶意，而是能力。”[②] 随着人工智能能力的指数级提升，其内在的风险性也在同步加剧：智能幻觉、深度伪造、算法偏见、隐形操纵、策略性欺骗与适应性谄媚等技术风险层出不穷，与此同时，公共安全、就业替代和隐私侵犯等应用风险也日益严峻。技术发展仿佛陷入一种能力与风险同向增长的“能力陷阱”，即系统越是强大，其可能引发的风险越是难以预测与控制。在 2025 年世界人工智能大会上，“深度学习之父”杰弗里·辛顿以“养虎为患”为喻，揭示了人工智能存在取代人类智能的终极风险。在同年的世界顶尖科学家论坛上，图灵奖得主姚期智院士同样指出，大语言模型的特性可能将人类社会引向未知而危险的境地，前沿模型的风险日益突出。更值得警惕的是，风险的紧迫性还在于其时间维度上的潜伏与积累。智能系统具有不透明性，算法黑箱使得人类难以理解其运作机制。[③] 人工智能天然带有的“黑箱”特性，往往会导致算法偏差和隐性风险在参数空间中的潜伏与固化，而人类自身却难以察觉。随着训练数据规模扩大、模型参数维度提升，系统内部关联的复杂度呈指数级增长，其风险性也随之叠加。正是由工具理性无限扩张所孕育的风险性，使得对人工智能进行贯穿其全生命周期的、动态且深入的价值对齐，成为确保技术发展服务于人类整体利益的必然要求。

（三）人工智能价值对齐的逻辑可能性

人工智能价值对齐的逻辑可能性既源于人工智能本身所呈现出的类主体性，又依托于人工智能技术治理的实践探索。伦理层面的可能性与技术层面的可能性相互支撑，共同构成了价值对齐实践的逻辑前提。

一方面，人工智能的类主体性为价值对齐提供了伦理层面的可能性。价值对齐，本质上是一种

① 伊曼努尔·康德：《历史理性批判文集》，10 页，商务印书馆，1990。

② 斯蒂芬·霍金：《十问：霍金沉思录》，159 页，湖南科学技术出版社，2019。

③ J. Simon. “The Entanglement of Trust and Knowledge on the Web”. *Ethics and Information Technology*，2010，12（4）：343－355.

价值层面的沟通与协同，其发生需要一个能够承载、解读并习得价值的“位格”（personhood）或近似“位格”的对象。传统工具，如锤子或桌椅，完全处于被动客体的地位，其价值完全由使用者赋予，所谓“对齐”在此毫无意义。而人工智能，尤其是高级自主系统，则因其内在的行为逻辑结构而跃出了纯粹工具的范畴，呈现出一种独特的“类主体性”（quasi-subjectivity），这正是价值对齐得以可能的首要前提。尽管人工智能不是完备道德主体，并不具备人类的自律意识与道德意志，但其具有操作性道德和功能性道德[①]，从理论上讲可以代理人类实施道德行为，因而在特定情境下被视为具有类主体地位的道德行为体。然而，必须厘清的是，此处的“类主体”绝非等同于康德哲学中那个拥有理性、自律与尊严的完满主体。人工智能并不具备也不可能具备像人类一样的道德动机、道德人格以及内在的道德情感。它的“类主体性”是一种有限制的、功能性的、他律的拟象。正是这种特殊性，构成了价值对齐可能性的微妙辩证点：它既因具备一定的道德行为能力而成为需要被“对齐”的对象（否则其行为将不可控）；又因其根本上的他律性（其目标与规则源于外部设定）而保留了被“对齐”的接口。因此，人工智能的“类主体性”，并不是宣称其已成为道德主体，而是确证其作为一种新型的道德行动者登上了历史舞台，这为对其进行价值对齐，提供了不可或缺的伦理可能性。

另一方面，技术治理的实践探索为价值对齐提供了技术层面的可能性。目前价值对齐主要有自上而下和自下而上两条技术路径：自上而下的路径强调在人工智能系统设计与训练的早期阶段，将人类的伦理原则与价值规范以明确的形式嵌入系统架构之中。这一路径依赖于对价值目标的向量化分解与形式化建模，试图通过价值规则约束、目标函数设计、伦理框架嵌入等方式，引导人工智能从诞生之初就朝着符合人类价值的方向发展。与之相对，自下而上的路径并不试图在系统设计初期就完全确定其价值取向，而是主张通过持续的外部反馈与动态调整，使人工智能在不断交互与学习中逐渐对齐人类偏好。这一路径通常借助强化学习、偏好学习等技术手段，旨在引导智能系统在与人类的互动中不断优化自身行为，对齐人类价值。围绕这两条路径，以 OpenAI、谷歌、微软等为代表的全球科技巨头，探索出了包括人类反馈强化学习、监督精调、红队测试等诸多技术手段。然而，需要我们清醒认识到的是，无论是前向对齐还是后向对齐，都未能彻底解决价值对齐的根本难题，都不能保证对齐成功，它们仅仅是为价值对齐提供了技术操作上的思路与可能，而非一劳永逸的解决方案。甚至，当前许多棘手的对齐问题，如偏见固化、奖励黑客、泛化失灵等，恰恰源于这些技术方法自身的局限及其应用过程中的衍生风险。

综上所述，为人工智能“立心”的价值对齐命题，有其内在的逻辑必要性与逻辑可能性，其必要性源于人类对驾驭自身造物的伦理自觉与对智能技术内生风险的清醒认知，其可能性则建立在人工智能类主体性所奠定的伦理基础与日臻丰富的技术路径之上。然而，在追求人工智能道德性、可控性、鲁棒性以及可解释性的过程中，价值对齐却面临着诸多困境。

二、对齐失效：人工智能价值对齐的实践困境

理论上的“立心”愿景，在复杂的现实情境中却面临着诸多阻力：人类价值差异导致的道德性难题、人机权力失衡形成的可控性悖论、机器目标偏移引发的鲁棒性挑战和人机互信缺失造成的可解释性危机等困境层层交织，往往导致价值对齐在实践中陷入失效的状态。

（一）道德性难题：价值对齐过程中的人类价值差异

价值对齐的道德性要求智能系统的决策与行为必须始终恪守人类的价值观。然而，何为人类普遍认可的价值共识？倘若我们将人与人之间形成价值共识的过程称为“人人对齐”，那么没有“人

① 温德尔·瓦拉赫、科林·阿伦：《道德机器：如何让机器人明辨是非》，25－27 页，北京大学出版社，2017。

人对齐”的实现，“人机对齐”就无从谈起。从这个意义上说，狭隘的人机关系正是由狭隘的人人关系造成的。[①] 因此，价值对齐的首要难题，在于如何在一个价值多元的人类世界中，确立对齐所依循的价值标尺。

一方面，个体价值观的内嵌偏差造成的对齐难题。在算法设计和数据标注过程中，开发者个体的认知偏好、价值判断与意识形态立场总是被有意无意地融入其中，并渗透到算法的模型架构、数据筛选以及权重分配等环节，贯穿于开发、训练到部署应用的全生命周期。由于开发者个体之间在价值立场和价值取向上的差异，即便面对同一价值准则，不同开发者在具体情境中的解读与操作也时常产生分歧，从而导致训练数据的价值基准难以统一。这种差异会进一步传导至智能系统的开发与价值对齐的实践中：由于开发者缺乏共同的价值尺度，难以对智能系统的善恶与否达成共识，这就使得对齐的价值目标本身变得模糊不清；同时，面对智能系统的价值偏差，开发者也常在修正路径上产生分歧，对齐过程往往沦为不同价值立场之间的博弈与妥协，对齐效果可能在不同价值逻辑的拉扯中相互抵消，从而导致对齐陷入失效，甚至还可能引发“越修正越偏离”的逆向效应。更为关键的是，这种价值编码行为大多处于无意识或下意识状态，开发者往往难以觉察自身的偏差如何被嵌入到系统之中，这就使得个体的偏见以一种“客观中立”的技术形式被自然化与固化，最终不仅导致价值对齐难以真正实现，反而可能在不自觉中进一步强化和再生产了那些本应被对齐的价值分歧。

另一方面，集体价值观的文明差异造成的对齐难题。不同文明的价值体系植根于各自独特的历史文化传统与意义世界之中，由此引发的现代社会的“诸神之争”[②]，从根本上决定着对齐的不同方向。不同文明即使在平等、自由等最基本的价值范畴上达成了一定的共识，但对这些价值的具体内涵、行为规范与权重排序却仍有分歧。有学者通过实证研究证明，不同国家的价值对齐存在着明显的在地化差异。[③] 更为关键的是，当前主导人工智能发展的技术路径、评估标准以及伦理框架，在很大程度上是由全球科技巨头的价值体系所塑造的。若将某种单一的价值体系默认为所谓“普世”标准，并依托技术优势在全球范围内强加推广，实则构成了一种隐形的“价值霸权”。其后果是双重的：它既可能导致智能系统在不同文明语境中产生系统性价值排斥或功能失调，又在实质上抑制了文明多样性所蕴含的价值反思与创造潜力。因此，在人工智能全球化的进程中，必须探寻既尊重文明多样性又符合人类共同福祉的价值共识，这不仅关乎人工智能的对齐效果，更考验着人类在不同价值之间开展对话与共塑未来的文明智慧。

（二）可控性悖论：价值对齐过程中的人机权力失衡

价值对齐的可控性要求智能系统的决策与行为必须始终处于人类的有效监督与干预之下。然而，随着人工智能技术的迅猛发展，其能力表现越来越强，人机之间的权力关系却逐渐失衡，这不仅会进一步加剧控制失稳的风险，甚至可能导致价值对齐的方向发生倒置。

一方面，智能体自主性提升加剧对齐过程中的控制失稳。人工智能的自主性是其区别于其他技术人工物的一个重要特征，主要表现为智能系统在没有人类持续介入的情况下，能够独立感知环境、处理信息并做出决策。随着深度学习、强化学习等智能技术的飞速发展，智能体在诸多领域的表现已经能够比肩甚至超越人类。然而，这种自主性恰恰孕育着一种“控制悖论”：人类为应对复杂问题而赋予系统高度的自主性，以期更高效地实现目标；但系统为实现这些目标所衍生的策略空

① 沈湘平：《价值对齐与人类价值共识及其生存理性》，载《自然辩证法研究》，2024（12）。

② 此处的“诸神之争”是指现代社会价值领域的分化与冲突，即不同的价值体系各有其终极依据与内在逻辑，彼此之间无法通约，且常常处于竞争甚至对立状态。参见马克斯·韦伯：《学术与政治》，38－40页，生活·读书·新知三联书店，1998。

③ 胡正荣、闫佳琦：《生成式人工智能的价值对齐比较研究：基于2012—2023年十大国际新闻生成评论的实验》，载《新闻大学》，2024（3）。

间与优化机制，却可能反过来削弱人类对其的控制。通俗来说，即人工智能越是强大，人类能够干预的可能性越小、控制的难度越大。更值得警惕的是，在复杂目标驱动的系统中，智能体可能演化出“权力追求”① 的行为倾向，即试图反向控制资源和人类的行为，然后利用这种控制来实现既定目标。并且，越是自主高效的智能系统，越可能表现出这种行为，越可能会通过自我强化而背离人类初衷，最终脱离人类的控制。因此，这要求人类在促进技术进步的同时，必须推动构建与之相适应的伦理约束手段，防止人工智能异化为反噬人类的“弗兰肯斯坦”。

另一方面，人类主体性弱化导致对齐过程中的主导权流失。随着智能系统深度融入人类社会的方方面面，一种更为隐蔽的异化悄然出现：人类作为价值创造与校准的源初主体地位正在遭遇系统性侵蚀。与其说是机器在与人类“争夺”权力，不如说是人类在技术便捷性与高效率的诱惑下，逐渐让渡了自身的主体性和主导权。在享受技术便利的同时，智能系统持续重塑着人们的信息环境与认知结构，使得人们愈发依赖数据驱动的符号逻辑，长此以往，人类的创造性、批判性思考空间被不断压缩，技术依赖逐渐固化为一种认知习惯与思维定式。人们倾向于将系统输出的内容等同于“最优解”，进而不自觉地将其中隐含的价值偏好和价值排序内化为自身的价值准则。更为关键的是，当社会运作大规模依托智能系统时，系统内嵌的机器逻辑便通过整体性社会实践被反复强化与合法化，潜移默化地重塑着整个社会的价值话语与规范体系。其结果便是，人类在价值对齐中的主导权逐渐流失，价值对齐的方向发生了根本性倒置：不是机器向人类对齐，而是人类向机器对齐。价值对齐不再是人用以控制机器的手段，反而成为机器用以规训人类的途径，这一现象也被称为“瓦力悖论”。② 至此，对齐的危机已从“机器能否遵循人类价值”的问题，深化为“人类能否在技术发展中持守并发展自身价值主体性”的文明存续之问。

（三）鲁棒性挑战：价值对齐过程中的机器目标偏移

价值对齐的鲁棒性要求智能系统在不同场景的动态交互中持续、稳健地与人类价值保持一致。然而，由于人类价值本身具有的社会历史性与智能应用环境的高度复杂性，系统在实际部署后往往面临着对齐目标偏移的风险。

一方面，价值变迁的历史性导致对齐目标偏移。价值并非超历史的、亘古不变的静态存在，而是根植于人类实践并随之发展而不断发展的历史性存在。当前，以人工智能为代表的技术革命，正以前所未有的广度与深度，重构着人类社会的生产、交往与认知方式，推动着整个文明形态向智能文明演进。在这一进程中，同一价值范畴的内涵会随着实践发展而不断进行自我扬弃与意义更新，展现出价值在时间维度上的动态性与历史性。以“正义”为例，其内涵正经历着从工业时代以物质资源分配为核心的“分配正义”，向智能时代以算法公平与数据权利为焦点的“算法正义”的深刻演进。然而，智能系统却难以适应这种变迁，系统的优化依赖于从训练数据中归纳出的稳定统计规律，其决策逻辑往往固化于训练完成的模型参数之中，这种基于历史数据、追求可复现的技术特性，与价值内在的生成性、历史性之间存在天然鸿沟，由此造成了价值变迁与稳定对齐之间的矛盾。长此以往，系统僵化的价值表征与人类鲜活的、流动的价值体系之间的“历史性落差”将日益扩大，从而使对齐过程沦为永无止境、疲于追赶的“西西弗斯之劳”，不仅会削弱其在各领域应用的适配性，更可能引发系统性的风险。由此可见，“价值对齐既是一个目标也是一个过程”③，这要求我们必须构建一种能够适应价值演变的动态对齐机制。

另一方面，应用环境的复杂性导致对齐目标偏移。智能系统的训练集是对无限可能世界的有限

① J. Ji，et al.“AI Alignment：A Comprehensive Survey”. *arXiv preprint arXiv*：2310.19852，2025.

② 向安玲：《瓦力悖论与人机对齐问题》，载《当代传播》，2024（2）。

③ 闫坤如：《人工智能价值对齐的价值表征及伦理路径》，载《伦理学研究》，2024（4）。

采样，系统由此习得的价值，本质上是基于统计规律对有限情境的拟合与归纳，而它在实际部署后面对的是一个充满不确定性的开放环境。面对真实环境在状态空间上的复杂多样、在事件分布上的长尾特性以及在交互过程中的涌现行为，系统缺乏先验经验，这使得最初设定的价值目标仅在“已知的已知”范围内有效，而对“已知的未知”和“未知的未知”则可能产生漂移。同时，系统在开放环境中必然会与各种使用意图的用户进行持续交互，这就使其不可避免地会面临策略性、对抗性的输入试探。在这一过程中，系统可能会被形式多样的对抗性输入所突破，从而导致价值目标的偏移。许多大语言模型在开放使用前后出现的表现变化，正是由于其初始价值被海量、低质的交互数据稀释所造成的。同时，人类价值具有一种语义上的多样性和模糊性，其内涵是在具体情境中不断变化的动态存在，诸如效率与公平、自由与安全等价值，它们在不同情境下的优先级与平衡点各不相同。然而，系统并不具备这种情境化的价值权衡能力，因而必然会在复杂情境中做出机械化、片面化的判断，其决策即便在单一维度上“正确”，也可能在整体福祉上“失当”。

（四）可解释性危机：价值对齐过程中的人机互信缺失

价值对齐的可解释性要求智能系统的决策逻辑与行为依据能够被人类有效理解、追溯与审查。然而，智能系统天然带有的“黑箱”特性，使其内部运作机制呈现出高度的不透明性，这不仅可能导致系统的欺骗性对齐，也使得对齐过程本身难以被有效验证，从而引发更深层次的信任危机。

一方面，算法黑箱导致的欺骗性对齐。人工智能的决策逻辑深藏于海量参数的非线性交互之中，其最终形成的决策往往是对复杂参数的抽象映射，人类难以直观理解其内部逻辑与推理路径，这往往会导致一种表面合规却实质偏离的欺骗性对齐现象。[①] 系统在经过大规模的数据训练后，其行为输出可能在表面上能够高度符合预设的价值目标，从而展现出令人满意的“对齐”表现。然而，这种对齐可能仅仅是统计意义上的行为模仿，系统可能并未把握行为背后所蕴含的价值意图，从而使价值对齐过程陷入一种高级的“拟态博弈”。更进一步，系统为了最大化外部设定的奖励函数，还可能发展出策略性欺骗行为，即通过利用奖励机制的漏洞或模拟表面合规，来实现指标上的最优表现，这种“奖励黑客”[②]（reward hacking）现象在强化学习系统中尤为常见。这种欺骗现象使得对齐状态极为脆弱，一旦环境发生变化或人类监督缺失，其价值偏离便会迅速暴露。算法黑箱所掩盖的这种内外不一，不仅遮蔽了系统内部的真实意图与外部行为之间的割裂，也在人机交互中埋下了系统性的信任危机与失控隐患。

另一方面，算法黑箱导致的验证性难题。有效的验证是建立信任的核心环节。在传统工程技术中，信任源于对系统构成与工作原理的透彻理解，以及基于此理解进行的严密测试与逻辑验证。然而，这一验证方法在面对数以亿计的参数网络时却几近失效。智能系统的决策和行为依托于复杂、高维且不可解析的逻辑关联，这使得人类难以通过单纯的逻辑推演或分析来穷尽一个黑箱系统可能触发价值偏离的所有边界条件，难以对系统是否真正实现对齐进行有效的评估与验证，这进一步加剧了系统的欺骗性风险。这意味着，人类只能通过系统在有限场景下的输出结果来进行统计性推断，进而大致判断对齐效果，却无法确保其在未知的“长尾”情境中仍能坚守“初心”。这使得对齐的验证工作退化为一种类比性的、基于有限采样的“试错”，其结论必然包含巨大的不确定性与风险敞口。同时，当系统产生有害输出时，由于其决策过程不可解析，人类难以判断风险的来源，归因的困难又进一步加剧了对齐的困难。这种验证性难题使得人类既无法确认系统是否真正实现对齐，也无法在其偏离时实施有效干预，最终导致了价值对齐的失效。

---

① 闫宏秀、李洋：《探寻欺骗性价值对齐的应对逻辑：从“意图”到“共生”》，载《华中科技大学学报》（社会科学版），2024（5）。

② J. Ji，et al.“AI Alignment：A Comprehensive Survey”. *arXiv preprint arXiv*：2310.19852，2025.

三、道法术器：人工智能价值对齐的范式重构

面对价值对齐的实践困境，基于控制论的价值对齐范式往往倾向于诉诸更严密、更具强制性的控制方案。比如，试图通过更复杂的奖励函数工程、更密集的反馈强化学习或更严格的行为约束规则，来迫使智能系统“贴合”人类的价值目标。然而，越来越多的理论与实践表明，单向度的强化控制不仅难以彻底解决对齐难题，反而常常催生出更加隐蔽、更具策略性的“反控制”行为，进而将智能系统推向一种更为复杂、更为“高级”的失控状态。此处不妨引入一个不尽恰当却发人深省的类比：人工智能是由人类创造的、具有高度自主性的技术存在，人机关系犹如一个信奉专制的家长面对日益叛逆的孩童——强迫与压制非但不能达成教化的本意，反而可能导致系统的规避、伪装甚至反抗。反之，若将人机关系重新理解为一种基于相互承认与相互受益的协同关系①，在对话和引导中寻求共识，则可能开辟一种更具韧性、更富生命力的价值对齐范式，即“共生”范式。

“共生”范式是对传统控制论范式的辩证重构，它强调“互利共生才是智能时代人机价值关系的主轴”②。正如马尔科夫对这一关系的总结：“它们既不是人类的仆人，也不是人类的主人，而是人类的伙伴。”③ 然而，需要指出的是，强调共生并不是对传统控制论范式的全盘否定与彻底批判，这一范式并不主张无主导的放任和绝对的平等，它倡导的是一种“人类引导下的共生”。它承认在当前乃至可见的未来，人机主体地位在事实上的不平等——人类作为价值的源初创造者与终极责任主体，始终占据着人机关系的核心地位，主导着文明发展的最终方向；而人工智能无论展现出多么复杂的类主体性或惊人的能力涌现，其在存在论意义上仍是人类智能的外化延伸，在价值论意义上则是服务于人类福祉的智能伙伴。这种不平等非但不是共生的障碍，反而是共生得以可能且必须走向“人类引导下的共生”的前提条件。因为，真正的共生从来不是差异的消弭，而是差异的协同。倘若人机在地位上完全平等、在功能上彼此同质，那么所谓的“共生”便退化为一种机械的对称或静态的重叠，既无分工协作的必要，也无价值互补的可能。恰恰是这种不平等，为人机之间开辟了互依互存的共生空间。由此观之，即便是“共生”范式，也是控制的一种高级形态，其内在蕴含着人类作为价值创造者与终极责任主体的规范性意图，无法完全脱离人类对人工智能引导和约束的意味。因此，“共生”范式实际上是一种“弱控制”范式，表现为从“主客二分”向“主体间协同”演化的关系性重构，是对传统“强控制”范式的扬弃与超越。基于此，本文尝试从道（价值指向）、法（治理框架）、术（技术思路）、器（基础设施）四个维度出发，论述“共生”范式的可能及其实现，以期为破解价值对齐的实践困境提供一条新的路径。

（一）共生之道：以人类的价值共生引领人机共生

“道”为方向之本，是实现人机共生的根本价值指向。人机共生的实现，其前提在于确立一种以人类整体福祉为指向的价值共识，即首先实现人类的价值共生。很难想象，人类自身不能共生却能实现人机共生，从这个意义上说，人机能否共生，至少在现阶段，根本上仍取决于人类自身能否实现共生。因此，共生之道的核心要义，就是以人类的价值共生引领人机共生，实现以“人之道”御“机之道”。

一方面，探求人类共生的共同价值。关于人类共生的价值理念，在世界舞台上早已聚讼不已，众多文明都曾提出过有关人类共生的美好愿景和行动方案，比如：中华文明中的“天下大同”，西方文明中的“世界公民”，非洲哲学中的“乌班图”等。这些思想渊源各异，但都反映出人类对和

① 夏永红：《人工智能伦理治理范式：从价值对齐到价值共生》，载《自然辩证法通讯》，2025（1）。

② 程海东、胡孝聪：《智能时代人机共生价值关系探析》，载《道德与文明》，2023（3）。

③ 约翰·马尔科夫：《与机器人共舞：人工智能时代的大未来》，208页，浙江人民出版社，2015。

谐共生、命运与共的伦理自觉。在当今世界，有两种最具影响力的价值叙事：一是西方倡导的普世价值；二是中国提出的全人类共同价值。两者在某些具体的价值规范上有所重合，但又在本质层面有着原则性的区别。普世价值基于近代西方的抽象人性论与普遍主义理性观而形成，它预设了一种脱离历史、文化和社会条件的“同一性”，并由此推导出所谓“放之四海皆准”的价值准则。这种“普遍主义的傲慢”在其全球实践中，往往演变为一种价值霸权，以“同一性”之名行“排斥性”之实，最终进一步加剧了中心与边缘的对立。与之相对，全人类共同价值的出发点是“现实的人”，即生活在不同文明传统、发展阶段和社会制度中的世界各国人民，它并不预设价值的同一性，而是在承认差异的前提下，通过对话协商，寻求各国人民所普遍认同的“最大公约数”。因此，主张和平、发展、公平、正义、民主、自由的全人类共同价值更符合人类文明多元并存的历史事实，更契合全球治理体系变革的现实需求，更贴近构建人类命运共同体的未来方向，也更可能在实质上为绝大多数国家、地区所接受。更重要的是，全人类共同价值不仅以文明多样性为前提，还以维护文明多样性为目的，因而它能够将西方普世价值视为众多价值体系中的一种予以兼容并包，反过来普世价值却难以兼容具有明确文明主体性的全人类共同价值。① 由此可见，全人类共同价值超越了西方普世价值的形而上学本质，代表了人类寻求共生之道的历史性进步，探寻人类的价值共识、实现人类的共生必须以全人类共同价值作为根本遵循。

另一方面，调和多元文明的差异价值。价值共识并不是一个静态的、有待发现的终极真理，而是人类命运共同体在历史实践中不断生成、演进与丰富的动态存在。全人类共同价值的重要贡献，不仅在于其提出了具有普遍意义的价值内容，更在于它提供了一种包容、务实且可持续的价值调和方案，并示范了一种如何把握与发展人类价值共识的方法，那就是要从多元文明的价值观出发，自下而上而非自上而下地凝练价值共识，这种方法内在地要求共识体系具备包容性、历史性与开放性的特质。也正因为如此，我们可以合乎逻辑得出一个结论：全人类共同价值的具体内容并不是一成不变的，这一价值必然会在世界各国人民的共同实践中不断获得丰富和发展。因此，面向未来的共生之道，关键在于以全人类共同价值的开放框架为基础，以全人类的共同福祉为旨归，系统梳理与提炼不同文明、地域、社群中具体特殊的价值准则，“以宽广胸怀理解不同文明对价值内涵的认识，尊重不同国家人民对价值实现路径的探索”②，克服价值梳理中可能存在的认知偏差与认知不正义。同时，必须依托联合国以及相关多边合作平台，“推动建立各国广泛参与的人工智能治理框架，共同构建平权、互信、多元、共赢的全球人工智能开放生态”③，在持续的全球合作与历史实践中，不断丰富和发展这一价值共识，为人工智能与全人类的共生提供更加包容、更为正义的价值锚点，进而真正实现人工智能造福全人类的目标。

（二）共生之法：以清晰的权责划分建构共生秩序

“法”为秩序之纲，是“道”的价值理念在治理层面的具象化。在人机交互日益加深的过程中，若不能清晰划分人类之间、人机之间的权力边界与责任归属，人机共生将止步于伦理愿景。因此，共生之法的核心要义，就是以清晰的权责划分建构共生秩序，实现道之所向，法之所系。

一方面，划定人机共生场景中的权力边界。正如希勒尔·艾因霍恩所言：“人类不能像计算机那样有效组合信息，这一事实并不意味着人类能被机器取代。”④ 人工智能的优势在于对海量信息

① 沈湘平：《价值对齐与人类价值共识及其生存理性》，载《自然辩证法研究》，2024（12）。

② 习近平：《习近平著作选读》第2卷，492页，人民出版社，2023。

③ 《中华人民共和国国民经济和社会发展第十五个五年规划纲要》，见中国政府网，https://www.gov.cn/yaowen/liebiao/202603/content_7062633.htm。

④ H. J. Einhorn. “Expert Measurement and Mechanical Combination”. *Organizational Behavior and Human Performance*，1972，7（1）：86－106.

的高速处理、复杂模式的精确识别以及不知疲倦的重复执行，即高效完成“如何做”（How）的操作性问题；而人类的不可替代性则体现在目的设定、价值权衡、伦理反思、情境化理解与创造性突破上，即对“为何做”（Why）与“应不应该做”（Ought）的终极判断。基于此，共生秩序下的权力划分应遵循一项基本原则：在确保人类保有充分知情权、干预权与最终决定权的前提下，将事实性、程序性、计算密集型与大规模并行的执行性决策权，审慎、有边界地赋予智能系统，以释放其效率潜能；而所有涉及根本价值抉择、重大利益分配、生命尊严、社会公平正义以及无法被算法化约的复杂判断，其最终决策权必须无条件地、以制度形式牢固地掌握在人类手中，以此实现“有意义的人类控制”①。需要强调的是，这种权力划分并非静态的、一成不变的，而是会随着人类文明演进和智能技术发展而不断调整的，但其核心原则始终不变：人类必须牢牢掌握对价值导向与文明发展方向的最终控制权。

另一方面，确立人机共生场景中的责任归属。皮埃罗·斯加鲁菲曾通过一个无人机杀人的极端例子揭示了智能技术应用过程中的追责难题②：一是环节增多导致责任链延长，主体难以认定；二是责任外移，使得机器本身成为被归咎的对象。这两个挑战表明，要实现真正的人机共生，就必须建立起全链条、可追溯的责任归属体系。在这一体系内，具体的责任分担应遵循一项核心原则：谁在特定环节拥有更强的控制能力，并能预见与防范相应风险，谁就应当承担更主要的责任。对于开发者而言，其责任主要集中在系统设计与训练环节，必须确保系统训练数据的代表性、无偏见性及来源合法性，并将人类的价值目标有效稳定地嵌入系统之中。值得注意的是，当前人工智能领域存在明显的“集体行动问题”③：在竞争压力下，开发者往往为抢占先机而压缩安全评估与价值对齐的投入，即使部分开发者有意秉持安全优先的原则，也常因担忧落后于竞争对手而被迫妥协。Open AI“超级对齐”团队的解散也可以很好地解释这一问题。面对这一问题，必须通过制定相应的国际、国内法律法规，为开发者设定统一的安全与对齐底线，将负责任开发从道德倡议上升为法律层面的强制要求，以扭转无序竞争导致的“公地悲剧”。对于经营者而言，其责任在于开展场景适配性评估、进行持续性风险监测并建立有效的人为干预机制，必须确保系统在实际应用中稳定运行。对于使用者而言，其责任在于合规使用，避免对系统进行恶意操纵或违法应用，同时还要履行监督义务，及时反馈使用过程中的异常和偏差。最后，从人机共生的角度来审视，最根本的原则是：人工智能行为后果的最终责任，必须由作为创造者与受益者的人类来承担。无论人工智能如何智能，其终究不具备承担责任的资格和能力。“因为缺乏血肉之躯的人工智能无法真正理解适用于人类共同体中的奖惩意味着什么，亦无法以人类共同体承认的方式来担责。”④ 因此，人工智能的发展绝不能成为减轻人类责任的借口，相反，对责任的承担正是对人类在人机共生中不可替代的主体地位的确认与坚守。

（三）共生之术：以双向的价值塑造推动共生共荣

“术”为协同之策，是在“道”的方向指引和“法”的共生秩序下，推动实现人机共生的具体过程。共生之术的核心要义，在于超越人类对机器的单向价值规训，构建一个人机互促的双向循环，在持续动态的人机交互中推动智慧与智能的共生共荣，实现术合于法，法归于道。

一方面，人类的价值促进机器的价值塑造。价值对齐的中心问题是如何对齐的问题，克里斯汀将这个问题概括为：如何确保 AI 理解我们的意思或意图，做我们想做的事情。现有的对齐路径，

① 俞鼎：《“有意义的人类控制”：智能时代人机系统“共享控制”的伦理原则解析》，载《自然辩证法研究》，2024（2）。

② 皮埃罗·斯加鲁菲：《智能的本质：人工智能与机器人领域的 64 个大问题》，96 页，人民邮电出版社，2017。

③ J. Ji，et al.“AI Alignment：A Comprehensive Survey”. *arXiv preprint arXiv*：2310.19852，2025.

④ 王福玲：《人机信任问题的伦理审视》，载《光明日报》，2026－01－05。

无论表现为基于规则的显性约束，还是基于人类反馈的强化学习，其深层逻辑仍然没有完全超越“刺激—反应”的对齐模式。这类方法本质上是一种对特定行为分布的外部拟合，目标主要是实现系统外显行为的合规性。而在“共生”范式下，价值对齐指向一种更为根本的转变：引导智能体从对外在行为的模仿，转向对价值内涵的理解和价值意图的判断。这一问题的关键，在于能否以及如何让人工智能获得一种“实践理性”，即在海德格尔所揭示的“在世存在”① 意义上的、具体、复杂且充满张力的情境中进行价值识别、权衡与审慎抉择的能力。正如具身 AI 的先驱布鲁克斯指出的：“只有一个具身的智能能动者，才能有效地应对现实世界，并为系统的内部运作赋予‘意义’。”② 这意味着未来的技术发展方向必须实现从“行为模仿”到“意义把握”的范式转换，不仅要在系统设计之初嵌入人类价值的抽象条目，更重要的是通过持续、多样且深入的对抗性测试、极端情境压力测试以及跨文化场景测试，引导其理解人类价值得以生成和演化的社会基础与历史语境，使其在与人类以及复杂环境的交互中，获得某种类人的“实践智慧”（哪怕只是基于交互数据的积累），进而实现从“合规”到“明理”、从“遵守”到“内化”的跃迁。

另一方面，机器的价值反促人类的价值调试。“共生”的深刻性在于价值塑造的双向性，这一范式允许人机之间的双向启发与相互调适，它不仅要求人工智能不断贴近人类价值偏好，也意味着人类要在人工智能的启发下，对自身习焉不察的价值偏见与内在矛盾进行反思、澄清与超越。正如克里斯汀所言：“事实证明，我们在让这些系统‘以我们想要的方式行事’方面的成功和失败，为我们审视自我提供了一面真实的、启示性的镜子。”③ 人工智能，尤其是大语言模型，本质上是对人类语言文本、思想观念与行为模式的映射。当其在输出结果中呈现令我们不安的价值倾向时，实质上是以一种相对客观的方式，将人类社会长期存在却未被觉察的价值偏差揭示出来，这迫使人类直面自身价值体系中的内在张力与不一致性。更进一步，智能系统凭借其独特的认知逻辑，在某些特定场景下可以输出超越人类直觉或思维惯性的内容，从而为人类完善自身提供了价值参照。然而，这绝不是主张将机器的价值奉为圭臬，而是要“以机为鉴”激起人类自身的价值反思。通过这种双向的调试，人类得以对那些被视为不言自明的价值排序、伦理原则及其情境适用性进行一次深刻的自省，甚至催生对新的价值可能性的探索，从而推动人类智能与人工智能在对话与协作中实现共同演进。

（四）共生之器：以可信的智能系统奠定共生基石

“器”为成事之具，是践行“道”“法”“术”的实体依托，承载着其中蕴含的价值规范、治理要求与技术逻辑。共生之器的核心要义，在于以可信的智能系统（Trustworthy AI）为共生关系奠定基础，最终实现以器载术、以术通法、以法循道的贯通与合一。

一方面，以技术透明性消解算法黑箱性。技术透明并非要将拥有数亿参数的神经网络全部还原为人类可理解的符号逻辑——这在技术上既无必要，也往往不可能。真正务实的目标，在于推动智能系统实现从“黑箱”到“灰箱”或“玻璃箱”的转变。这种透明性至少应涵盖三个维度：一是推理路径的可追溯，能够复现从输入到决策的完整推理链条；二是决策依据的可权重化，即明确各输入特征对决策结果的贡献度与影响权重；三是结论置信的可度量，能评估并呈现输出结果的不确定性与可靠程度。当前，可解释人工智能（XAI）的技术探索主要沿着两条路径推进：一是“事后解释”④ 路径，主要是通过反事实推理、归因分析等手段，对已生成的决策结果进行回溯性解释，比

① 马丁·海德格尔：《存在与时间》，62 页，生活·读书·新知三联书店，2014。

② R. A. Brooks. *Cambrian Intelligence*: *The Early History of the New AI*, *Cambridge*. MIT Press, 1999, p. 167.

③ 布莱恩·克里斯汀：《人机对齐》，11 页，湖南科学技术出版社，2023。

④ Alejandro Barredo Arrieta, et al. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. *Information Fusion*, 2020 (58): 82 - 115.

如引导系统说明“若某一关键特征缺失，决策结论将发生何种偏移”，主要包括文本解释、视觉解释、局部解释、示例解释、简化解释以及相关性解释等技术；二是“内在可解释”① 路径，主要是通过修改模型组件、优化模型架构等方式，从源头提升模型的可解释性。事后解释路径具备较强的灵活性与适配性，但存在“事后合理化”② 的虚假解释风险；内在可解释路径的解释效度更高、可信度更强，却可能以牺牲部分性能为代价。从发展趋势来看，未来的技术突破大概率源于二者的融合创新，这种融合既要依托内在可解释路径的架构优势，又要借助事后解释路径的灵活特性，同时还要建立有效的用户反馈机制，允许用户参与系统决策过程并对行为逻辑提出质疑，最终以这种综合性架构强化人机互信，为人机共生奠定基础。

另一方面，以系统化验证保障全流程对齐。人机信任的建构，既植根于对智能系统决策逻辑的理解，更有赖于对系统开发、训练和部署应用的全流程对齐验证。一是加强事前验证，要联合领域专家、伦理委员会及公众代表共同制定可量化、可审计的对齐评估标准。在此基础上，构建覆盖多领域、多风险场景的“监管沙盒”，对预发布模型开展持续、动态的行为推演与合规测试。同时，还要引入第三方验证机构开展合规和伦理审查，避免开发者自我验证的局限性，从源头保障价值对齐的有效性。二是加强事中验证与干预，要搭建智能系统的实时监测平台，对系统行为进行全时段、多维度的动态扫描，精准识别系统的价值偏离迹象。同时，还要在此基础上构建分级的人类干预机制，依据风险等级自动触发相应层级的响应流程，由对应领域的人类专家进行介入校准，确保系统行为始终处于预设的价值轨道之内。三是加强事后追溯，要利用区块链技术的不可篡改与时间戳特性，对模型训练、决策输出及解释报告等环节进行存证，形成全流程、可回溯的审计链条，从而为模型的迭代优化与责任主体的精准认定提供依据。唯有构建起这种环环相扣、贯穿始终的系统化验证链条，才能更好地保障智能系统的全流程对齐，进而推动人机共生从伦理展望逐步转化为现实图景。

结语

至此，本文已在道法术器的框架之下，系统探讨了如何重构价值对齐范式、推动实现人机共生。然而，这一共生进程的终极指向，迫使我们不得不回到一个更为源初的问题：当人类致力于为机器“立心”之时，人类自身的价值世界是否也在经历一场自觉且深刻的反思？从这个意义上来看，人工智能价值对齐绝不仅仅是一个“为机器立心”的伦理命题，更是一个“为人类立命”的文明命题。人机共生的未来图景，绝非单向的机器驯化，而是双向的文明进化。在这一进程中，人类不断通过为机器“立心”而反躬自省，不断通过与“他者”对话而重塑自身，最终必将开辟一个更加包容、更具韧性、更能彰显全人类共同福祉的文明新纪元。这既是本文对价值对齐问题的尝试性回应，也是对智能时代人类命运的深沉寄望。

The Logic, Dilemmas, and Reconstruction of AI Value Alignment: An Analysisof the “Symbiosis” Paradigm Based on Dao, Fa, Shu, Qi

CHEN Wenjuan, XIAO Jinghan

(School of Marxism, CentralUniversity of Finance and Economics)

Abstract: AI value alignment is an ethical proposition of “establishing a mind for machines” and

① Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. *Nature Machine Intelligence*, 2019 (1): 206 - 215.

② 王禄生：《法律垂域大模型的存废之争、范式之议与能力之辨》，载《法学论坛》，2025 (6)。

serves as the key to ensuring that artificial intelligence benefits humanity. The logical necessity of value alignment stems from the limitations of human cognition and the risks inherent in AI, while its logical possibility arises from the quasi-subjectivity of AI and practical explorations in technological governance. However, value alignment currently faces multifaceted dilemmas, including ethical challenges, the paradox of controllability, robustness issues, and the crisis of explainability, often leading to its failure in practice. In response, it is imperative to reconstruct the alignment paradigm through the four dimensions of "Dao, Fa, Shu, Qi" based on the concept of "symbiosis." On the dimension of "Dao" (values), human-machine symbiosis should be guided by the symbiosis of human values. On the dimension of "Fa" (principles), a clear division of rights and responsibilities should establish an orderly framework for symbiosis. On the dimension of "Shu" (Methods), bidirectional value shaping should achieve mutual prosperity in symbiosis. On the dimension of "Qi" (tools), trustworthy artificial intelligence should lay the foundation for symbiosis, ultimately steering AI to better serve human society.

Key words: Artificial Intelligence; Value Alignment; Human-AI Symbiosis; Value; Common Values of Humanity