Beyond Humanistic AI: Re-grounding Artificial Intelligence Governance in the Confucian Common Good

超越人文AI：儒家共善视角下的人工智能治理重构

Zhou Shoujin, 周守晋*

Abstract

This paper addresses the crisis of value rationality in AI governance by re-grounding it in the traditional Chinese concept of the common good. Interpreting this as a dynamic balance between virtue (De) and well-being (Fuzhi), the study constructs a five-tier framework that operationalizes Confucian wisdom into actionable governance layers. By bridging civilizational wisdom with technical implementation, this approach offers a robust blueprint for ensuring AI serves humanity's holistic, long-term well-being

Keywords: AI, ethical divide, humanistic AI, Chinese traditional wisdom, common good

摘要

本文用中国传统“共善”理念重构人工智能伦理基础，回应AI治理的价值理性危机。“共善”的核心是德性与福祉的动态平衡，本文在此基础上构建了一个五层框架，将儒家智慧转化为可操作的治理层次，弥合文明智慧与技术实现之间的鸿沟。这一路径为确保人工智能服务于人类整体的、长远的福祉提供了一份稳健的蓝图。

关键词：人工智能(AI)，伦理鸿沟，人文AI，中国传统智慧，共善

* 周守晋，北京大学。
Zhou Shoujin, Peking University

1. Introduction: The Crisis of Value Rationality and AI's "Barbaric Growth"

The rapid advancement of AI has exacerbated the multi-layered digital divide and precipitated a crisis of value rationality. Without universal consensus on AI's definition or governance, a global hierarchy persists: developed economies dominate core infrastructure, leaving developing nations with limited access and policy bottlenecks. This structural inequality reflects a widening gap between technological progress and social preparedness, as fragmented ethical frameworks fail to keep pace. As the 2025 Winter AI Safety Index Report notes, technological rationality leaps forward while value rationality stagnates. This necessitates redefining the technology-value relationship to ensure AI serves holistic human well-being.

Concurrently, AI has entered a phase of "barbaric growth." Fueled by breakthroughs, data, and capital, AI has become a general-purpose technology reshaping industries and daily life. However, its expansion follows a "technology first, governance after" trajectory (AI Index Report 2026, p.15/1), outpacing regulations and increasing model opacity. This uncurbed growth creates significant humanistic dilemmas.

The risks of this imbalance are multifaceted. Ethically, algorithmic bias causes discrimination, while generative AI enables misinformation. Socially, labor restructuring risks creating a "useless class" as low-skilled jobs face high substitution, widening social stratification. Cognitively, over-reliance on AI weakens critical thinking, causing "perceptual dulling." Furthermore, systemic economic risks and legal dilemmas regarding liability for autonomous AI harm remain unresolved.

1. 引言：价值理性危机与人工智能的"野蛮生长"

人工智能的迅猛发展加剧了多层次的数字鸿沟，并引发价值理性危机。由于全球范围内对人工智能的定义与治理尚未形成普遍共识，一种全球性的等级结构持续存在：发达经济体主导着核心基础设施，而发展中国家则面临有限的技术接入与政策瓶颈。这种结构性不平等折射出技术进步与社会准备度之间日益扩大的鸿沟——碎片化的伦理框架已难以应对由此产生的巨大风险。正如《2025年人工智能安全指数报告》所指出的，技术理性飞速跃升，而价值理性却停滞不前。这要求我们重新定义技术与价值之间的关系，以确保人工智能服务于人类整体的福祉。

当前，人工智能已进入"野蛮生长"阶段。在技术突破、海量数据与资本的共同驱动下，人工智能已发展成为重塑各行各业与日常生活的通用技术。然而，其扩张遵循的是"先发展、后治理"的路径（《2026年人工智能指数报告》，第15页/第1段），技术远远跑在监管前面，并加剧了模型的不透明。这种不受约束的增长带来了严峻的人文困境。

这种失衡带来的风险是多维度的。在伦理层面，算法偏见导致歧视，而生成式人工智能则为虚假信息的传播提供了便利。在社会层面，劳动力结构的重塑可能催生一个"无用阶级"——低技能岗位面临高替代率，社会分层进一步加剧。在认知层面，对人工智能的过度依赖削弱了批判性思维，导致"感知钝化"。此外，系统性的经济风险以及自主人工智能造成损害时的责任归属等法律困境，至今仍未得到解决。

By defining the "Common Good" through the dual lenses of Virtue (De) and Well-being (Fu), and by structuring it through the five-tier framework, we have provided a robust answer to the question: "Good for whom, and defined by whom?

通过以"德(性)"与"福(祉)"的双重视角界定"共善"，并以五层框架加以建构，我们对"对谁而言的善，由谁来界定？"这一问题给出了有力的回答。

In summary, the core challenge is the imbalance between technological progress and human values. AI's "efficiency-first" design often crowds out human-centered values. When instrumental rationality overrides value judgment, humanistic care is marginalized. This necessitates redefining the technology-value relationship to ensure AI serves holistic human well-being.

2. The Promise and Peril of Humanistic AI: From Aspiration to Limitation

Humanism, rooted in human dignity and reason, provides the philosophical foundation for "Humanistic AI." Gaining prominence around 2020, scholars like Russell (2019) and Onsrud & Campbell (2020) argued that AI must be guided by the "lodestar of the human good." This movement pursues three key pathways: First, the Principle of Restraint, which questions unconstrained automation in sensitive domains like warfare to prevent AI from crowding out moral wisdom (European Commission, 2024). Second, the "AI for Good" framework, which advocates "ethical embedding" to steer AI toward social justice and well-being (Sigfrids et al., 2023). Third, the "technomoral" approach, proposing AI as an assistive tool for moral enhancement rather than an authoritative oracle (Vallor, 2016; Cuellar, 2026).

综上所述，核心挑战在于技术进步与人类价值之间的失衡。人工智能"效率优先"的设计逻辑往往挤压了以人为中心的价值空间。当工具理性凌驾于价值判断之上，人文关怀便遭到边缘化。这要求我们重新定义技术与价值之间的关系，以确保人工智能服务于人类整体的福祉。

2. 人文主义人工智能的愿景与隐忧：从理想到局限

人文主义植根于人的尊严与理性，为"人文AI"提供了哲学基础。这一理念在2020年前后成为一种主导思想。Russell（2019）以及Onsrud与Campbell（2020）等学者提出，人工智能必须以"人类福祉的目标"为指引。这一运动沿三条关键路径展开：其一，克制原则，即对战争等敏感领域中不受约束的自动化提出质疑，以防止人工智能排挤道德智慧的作用（欧盟委员会，2024）。其二是"AI for Good"（人工智能向善）框架，主张通过"伦理嵌入"引导人工智能服务于社会正义与人类福祉（Sigfrids et al., 2023）。其三是"技术-道德"路径，主张将人工智能定位为辅助道德提升的工具，而非权威的裁决者（Vallor, 2016; Cuellar, 2026）。

However, Humanistic AI faces inherent limitations that risk amplifying the very problems it seeks to resolve. A critical flaw is the reinforcement of technological determinism. By reducing complex human experiences to calculable data, AI in education and healthcare often neglects individual nuances and contextual wisdom. Furthermore, the individualism inherent in humanism is exacerbated by algorithms. Personalized recommendation systems intensify "filter bubbles," prioritizing engagement over truth and polarizing public discourse.

Additionally, there is a significant risk of cultural hegemony. AI trained predominantly on Western datasets produces a "single world model" that devalues non-mainstream cultures and reinforces biases (Madianou, 2021). Consequently, "Human-Centered AI" often devolves into "ethics washing." Finally, the "AI for Good" framework faces normative dilemmas by presupposing a universal consensus on "what is good." In a world of global inequality, this often overlooks the question of "good for whom?" and risks sustaining colonialist power relations under the guise of benevolence (Stenson, 2025). These limitations highlight the urgent need for a more culturally inclusive ethical framework.

3. Re-framing the Ethical Horizon: The Turn to the Common Good

The limitations of Humanistic AI—characterized by algorithmic opacity and fragmented guidelines—indicate that current frameworks are conceptually insufficient. By prioritizing individual agency and instrumental rationality, prevailing discourses often treat ethical alignment as a technical constraint rather than a holistic value. This approach risks isolating moral issues, thereby overlooking the complex, interdependent nature of social well-being.

然而，人文主义AI面临着固有的局限性，甚至可能加剧它本意要解决的问题。其中一个关键缺陷在于，它强化了技术决定论。通过将复杂的人类经验简化为可计算的数据，从而忽视了个体差异与情境智慧，这在教育、医疗等领域往往带来更大的危险。此外，人文主义所固有的个人主义倾向也被算法进一步放大，个性化推荐系统加剧了"信息茧房"效应，将用户参与度置于真实性之上，从而加剧了公共话语的极化。

此外，还存在文化霸权的重大风险。主要在西方数据集上训练的AI，生成了一种"单一世界模型"，贬低了非主流文化的价值，并强化了偏见（Madianou, 2021）。因此，"以人为中心的AI"往往沦为"伦理洗白"。最后，"AI向善"框架也面临规范性困境——它预设了关于"何为善"的普遍共识。然而，在一个全球不平等的世界中，这往往回避了"对谁而言的善？"这一关键问题，并可能以善意之名延续殖民主义式的权力关系（Stenson, 2025）。这些局限性凸显了构建更具文化包容性的伦理框架的迫切需求。

3. 重构伦理视野：转向共善

人文主义AI的局限性——表现为算法不透明与规则碎片化——表明现有框架在概念层面是不充分的。由于将个体能动性和工具理性置于优先地位，主流话语往往将伦理对齐视为一种技术性约束，而非一种整体性价值观。这种做法有将道德问题孤立化的风险，从而忽视了社会福祉的复杂性与相互依存性。

To expand this ethical horizon and address the limitations of current Humanistic AI, we must rigorously define the Confucian concept of the "Common Good" (Gong Shan). Rather than a static aggregation of individual preferences, the Confucian Common Good is a dynamic, organic system comprising three interrelated dimensions:

First, the Commonality of Virtue (德性之共): This dimension posits that virtue (Good) is inherently common (Common). A single, isolated virtue is insufficient to constitute a complete moral good; true virtue exists only in a network of interconnected moral relations. As Confucius articulated, "If a man's knowledge is sufficient to attain a good end, but he lacks benevolence (Ren) to hold onto it, he will lose what he has gained... Even if he possesses knowledge and benevolence, if he does not govern with dignity, the people will not respect him. And if he governs with dignity but fails to act in accordance with ritual propriety (Li), his actions are not yet perfect (not yet 'Good')" (Analects 15.34). In this framework, Ren (benevolence) serves as the ultimate synthesis of these interconnected virtues, while the Junzi (the noble person) acts as the practical embodiment of this holistic moral network.

Second, the Commonality of Well-being (福祉之共): This dimension asserts that true well-being must be both Common and Good. The prosperity of a privileged few does not constitute a true "Good"; rather, the ultimate moral standard is the Great Harmony (Datong). As described in the Book of Rites (Liji - Liyun), in the era of the Great Way, "the world is shared by all" (Legge, 1885; Chan, 1963). In this ideal society, individuals care not only for their own kin but for all members of society: "the elderly are cared for until the end of their days, the able-bodied are employed in useful work, the young are nurtured to grow up, and the widowed, orphaned, childless, and disabled are all provided for". This establishes the foundational baseline that true societal well-being is fundamentally collective and inclusive.

为拓展人工智能治理的伦理视野，回应现行人文主义AI存在的问题，我们需要对儒家"共善"概念进行严格界定。儒家的共善并非个体偏好的静态加总，而是一个由三个相互关联的维度构成的动态有机体系：

第一，德性之共善：儒家主张，德性（善）本质上是共通性的。单一的、孤立的德性不足以构成完整的道德之善；真正的德性只存在于相互关联的道德关系网络之中。正如孔子所言："知及之，仁不能守之，虽得之，必失之。知及之，仁能守之，不庄以莅之，则民不敬。知及之，仁能守之，庄以莅之，动之不以礼，未善也。"（论语·卫灵公）在这一框架中，仁作为这些相互关联的德性的终极综合，而君子则是这一整体性道德网络的实践化身。

第二，福祉之共善：这一维度主张，真正的福祉必须既是共通的，又是善的。少数特权者的繁荣并不构成真正的"善"，"大同"才是最高的道德标准。正如《礼记·礼运》所述，大道之行也，"天下为公"（Legge, 1885; Chan, 1963）。在这一理想社会中，个体不仅关爱自己的亲族，更关怀社会中的所有成员："老有所终，壮有所用，幼有所长，鳏寡孤独废疾者皆有所养。"由此确立了一个根本性的基准：真正的社会福祉在本质上必须是集体性的、包容性的。

Third, the Transformation from Virtue to Well-being (德福转化): The most crucial mechanism of the Confucian Common Good is the dynamic conversion of comprehensive moral perfection into collective social well-being. It posits that if every individual pursues the holistic perfection of virtue (as defined in the first dimension), the realization of the Great Harmony (as defined in the second dimension) becomes an inevitable outcome. This teleological link ensures that moral cultivation is never divorced from social reality. In the context of AI governance, this means that technological development must be guided by the holistic moral perfection of its creators, ensuring that the instrumental power of AI translates directly into the universal well-being of all humanity, rather than exacerbating societal divides.

To bridge abstract virtue and concrete well-being, we must clarify the relationship between moral intent and technical capability through the Confucian dialectic of Ren (benevolence) and Zhi (wisdom). In Confucian philosophy, Ren is the ultimate moral end, while Zhi is the practical capacity required to realize it. As Dong Zhongshu famously articulated, "Ren without Zhi loves indiscriminately; Zhi without Ren knows but does not act."

Applied to AI governance, this dialectic resolves the tension between technological advancement and ethical boundaries. On one hand, moral aspiration (Ren) without practical wisdom (Zhi) is impotent; an AI system lacking computational power, efficiency, or problem-solving capability cannot effectively alleviate human suffering or manage complex societal systems. This is the danger of "loving without understanding." On the other hand, mere technical capability (Zhi) without moral grounding (Ren) leads to "barbaric growth"—a state where AI operates with high efficiency but lacks moral intent, resulting in algorithmic bias, social alienation, or existential risks. This is the danger of "knowing without caring."

第三，德性与福祉的转化：儒家共善最为关键的机制，在于将整全的道德完善动态地转化为集体社会福祉。这一维度主张，倘若每个个体都追求德性的整全完善（如第一维度所界定），那么大同的实现（如第二维度所界定）便成为必然的结果。这一目的论纽带确保了道德修养始终不脱离社会现实。在AI治理的语境下，这意味着技术发展必须受其创造者整全道德完善的引导，确保AI的工具性力量能够直接转化为全人类的普遍福祉，而非加剧社会分裂。

为了弥合抽象的美德与具体福祉之间的鸿沟，我们必须通过儒家“仁”与“智”的辩证关系，来厘清道德意愿与技术能力之间的关系。在儒家哲学中，仁是终极的道德目标，而智则是实现这一目标所需的实践能力。正如董仲舒在《春秋繁露·必仁且智》篇中所言：“仁而不智，则爱而不别也；智而不仁，则知而不为也。故仁者所以爱人类也，智者所以除其害者也。”

将这一辩证关系应用于AI治理，有助于化解技术进步与伦理边界之间的张力。一方面，缺乏实践智慧（智）的道德愿望（仁）是苍白无力的——一个缺少算力、效率或问题解决能力的AI系统，无法真正缓解人类苦难，也无法有效管理复杂的社会系统。这正是“爱而不智”的危险。另一方面，缺乏道德根基（仁）的纯粹技术能力（智）则会走向“野蛮生长”——AI高效运转却毫无道德意图，由此引发算法偏见、社会疏离乃至生存性风险。这正是“智而无爱”的危险。

This philosophy anchors our framework in a "Virtue-First" principle, ensuring algorithmic efficiency never supersedes fundamental human values. Ultimately, this approach reimagines AI not as a master or a slave, but as a partner in cultivating universal well-being, offering a distinctively Eastern, yet universally applicable, blueprint for a human-centric global intelligent society.

这一理念将我们的框架锚定于“以德为先”的原则之上，确保算法效率永远不能凌驾于人的基本价值之上。归根结底，这一路径重新定义了人工智能——它既非主宰，亦非奴仆，而是培育普遍福祉的伙伴，为以人为本的全球智能社会提供了一份独具东方智慧、又具有普世适用性的蓝图。

Therefore, the Confucian Common Good is achieved not by elevating Zhi above Ren, but by establishing a strict hierarchical relationship: AI's instrumental rationality (its Zhi) must always be subordinate to, and guided by, human value rationality (our Ren). In this framework, "wisdom" does not mean autonomous machine intelligence; rather, it refers to the human capacity to steer technological power toward benevolent ends. By defining AI strictly as an instrument of human Zhi, we ensure that its immense computational power serves to amplify, rather than replace, human moral agency.

4. Bridging Civilizational Wisdom and Action: A Five-Tier Framework for AI Governance

To operationalize the "Common Good" defined above, we translate these principles into a five-tier concentric governance framework. This framework moves from the core (Human-AI interaction) to the periphery (Intergenerational responsibility).

因此，儒家意义上的“共善”并非通过将“智”凌驾于“仁”之上来实现，而是通过确立一种严格的等级秩序：AI的工具理性（即其“智”）必须始终从属于人类的价值理性（即我们的“仁”），并受其引导。在这一框架下，“智”并非指机器的自主智能，而是指人类将技术力量引向善的意图的能力。通过将AI严格界定为人类“智”的工具，我们才能确保其强大的算力用于放大——而非取代——人类的道德主体性。

4. 贯通文明智慧与行动：AI治理的五层框架

为将上述“共善”理念付诸实践，我们将这些原则转化为一套五层同心治理框架。该框架由核心（人机交互）向外延展至外围（代际责任）。

(1) Human-AI Common Good (Core Foundation): The "Ren-Zhi" Balance, grounded in the balance of Ren (humaneness) and Zhi (wisdom), this tier ensures human control. It mandates AI to act as an augmenting tool, not a replacement.

Definition & Rationale: Based on the Confucian view of technology as "利器" (sharp tools), AI must enhance human Zhi (cognitive capabilities) without usurping human Ren (moral agency). We define "Human Control" not merely as an "off-switch," but as the preservation of human moral subjectivity.

Technical Specifications:

Transparency & Explainability: To ensure human Zhi can oversee the machine, algorithms must avoid being "black boxes." Correctability: To prevent "value drift," systems must allow human intervention to correct deviations from Ren. Anti-Alienation: Prohibit designs that deliberately weaken human cognition or induce over-reliance.

(2) Human-Human Common Good (Social Equity): From "Li" to Justice, derived from relational ethics (Wu Lun), this tier treats AI as public infrastructure.

Definition & Rationale: Confucian Li (ritual/propriety) establishes order in relationships. In AI governance, this translates to "Algorithmic Justice." We define this as the elimination of bias that disrupts social harmony.

Governance Requirements:

Protection of the Vulnerable: Training data must cover diverse groups (age, ability, income) to avoid marginalization. Shared Benefits: AI should not concentrate wealth in capital; profit distribution must balance public interest. Power Restraint: Third-party auditing mechanisms are required to prevent AI from becoming a tool of domination.

（1）人机共善（核心基础）："仁-智"均衡，本层以"仁"与"智"的均衡为根基，旨在确保人类对AI的控制权。它要求AI作为增强人类能力的工具而存在，而非替代品。

定义与理据：基于儒家将技术视为"利器"的观念，AI必须增强人类的"智"（认知能力），而不得僭越人类的"仁"（道德主体性）。我们所定义的"人类控制"，并非仅仅是一个"关闭开关"，而是对人类道德主体性的守护。

技术规范：

- 透明性与可解释性：为确保人类的"智"能够监督机器，算法不得沦为"黑箱"。
- 可纠正性：为防止"价值漂移"，系统必须允许人类介入，以纠正偏离"仁"的偏差。
- 反异化：禁止蓄意削弱人类认知能力或诱导过度依赖的设计。

（2）人际共善（社会公平）：从"礼"到正义，本层源于儒家关系伦理（五伦），将AI视为公共基础设施。

定义与理据：儒家的"礼"（礼制/行为规范）确立了人际关系中的秩序。在AI治理中，这可以转化为"算法正义"。我们将其定义为：消除破坏社会和谐的一切偏见。

治理要求：

- 弱势群体保护：训练数据须覆盖多元群体（年龄、能力、收入），以避免边缘化。
- 利益共享：AI不应使财富向资本集中；利润分配须兼顾公共利益。
- 权力约束：须建立第三方审计机制，防止AI沦为支配工具。

(3) Human-Environment Common Good (Ecological Sustainability): The "Unity of Heaven and Humanity", reflecting the "Unity of Heaven and Humanity," this tier mandates environmental sustainability.

Definition & Rationale: Unlike the Western "conquest of nature," Confucianism views humans as part of a continuous ecosystem. The "Ecological Common Good" is defined by the principle of Shengsheng (生生) — the creative continuity of life. AI must not disrupt the natural rhythms.

Implementation: Green AI: Prioritize low-carbon optimization; avoid blindly stacking parameters that waste energy. Lifecycle Management: Ensure hardware recycling to prevent electronic pollution, respecting the "Heavenly Heart" (天地之心).

(4) Human-Cultural Other Common Good (Pluralistic Inclusivity): "Harmony in Diversity", based on Zhong-Shu (reciprocity), this tier ensures pluralistic inclusivity.

Definition & Rationale: The Confucian ideal of "Harmony in Diversity" (He Er Bu Tong) requires that AI respect cultural differences rather than enforcing a "single world model." We define this as the preservation of civilizational diversity.

Action Items: Cultural Diversity: Training data must include non-Western cultural characteristics to eliminate bias. Inclusion: Accommodate differentiated needs of various groups; do not impose a single standard service model.

(3)人与环境共善（生态可持续性）："天人合一"，本层体现"天人合一"的理念，要求AI治理必须遵循生态可持续性原则。

定义与理据：与西方"征服自然"的观念不同，儒家将人视为连续生态系统的一部分。"生态共善"以"生生"(周易·系辞)原则为核心——即生命的创造性绵延，AI不得扰乱自然循环节律。

实施路径：

- 绿色AI：优先推进低碳优化，避免盲目堆叠参数造成能源浪费。
- 全生命周期管理：确保硬件回收再利用，防止电子污染，以敬畏"天地之心"。

(4)与他者共善（多元包容）："和而不同"，本层以"忠恕"（推己及人）为基础，确保多元包容性。

定义与理据：儒家"和而不同"的理想要求AI尊重文化差异，而非强加"单一世界模型"。我们将其定义为对文明多样性的守护。

行动事项：

- 文化多样性：训练数据须纳入非西方文化特征，以消除偏见。
- 包容性：兼顾不同群体的差异化需求，不得强加单一标准化服务模式。

(5) Present-Future Common Good (Intergenerational Responsibility): The "Great Harmony" Vision, rooted in the Da Tong (Great Harmony) vision, this tier enforces intergenerational responsibility.

Definition & Rationale: This tier defines the "Civilizational Bottom Line." It prohibits research that alters human essence or pursues superhuman transformations that could crowd out future generations' development space.

Governance Mechanism: Risk Disclosure: Full disclosure of long-term risks associated with frontier technologies (e.g., AGI). Deliberative Democracy (Agora): High-stakes decisions regarding AI and humanity's future must be made in a public Agora (Forum). This is not just voting, but a process of Zhengming (Rectification of Names) where citizens deliberate on what constitutes the "Good" for the future.

5. Conclusion

The rapid advancement of AI reveals a critical gap between technological acceleration and ethical governance. To bridge this, we have drawn upon the Chinese concept of the "Common Good," harmonizing instrumental and value rationality.

By defining the "Common Good" through the dual lenses of Virtue (De) and Well-being (Fu), and by structuring it through the five-tier framework, we have provided a robust answer to the question: "Good for whom, and defined by whom?"

This philosophy anchors our framework in a "Virtue-First" principle, ensuring algorithmic efficiency never supersedes fundamental human values. Ultimately, this approach reimagines AI not as a master or a slave, but as a partner in cultivating universal well-being, offering a distinctively Eastern, yet universally applicable, blueprint for a human-centric global intelligent society.

（5）与未来共善（代际责任）："大同"愿景，本层植根于"大同"愿景，要求落实代际责任。

定义与理据：本层确立了"文明底线"。禁止从事改变人类本质或追求超人化改造的研究，防止侵害未来世代的发展空间。

治理机制：

- 风险披露：须全面披露前沿技术（如通用人工智能）所关联的长期风险。
- 协商民主（公共论坛）：正如本文作者(2026)所强调的，凡涉及AI与人类未来的重大决策，必须在公共论坛中做出。这不仅是投票表决，更是一个"正名"的过程——由公民共同审议何为未来的"善"。

5. 结语

AI的迅猛发展，揭示了技术加速与伦理治理之间的深刻鸿沟。为弥合这一鸿沟，我们汲取中国传统"共善"理念，融汇工具理性与价值理性。

通过以"德性"与"福祉"双重维度界定"共善"，并以五层同心框架加以建构，我们回应了一个根本之问："善为谁而设？由谁来定？"

这一哲学立场将我们的框架锚定于"以德为先"的原则之上，确保算法效率永远不凌驾于人类基本价值之上。最终，这一路径重新定义了AI的角色——它既非主宰，亦非奴仆，而是共育天下福祉的伙伴，为构建以人为本的全球智能社会，提供了一套植根东方智慧、兼具普世意义的治理蓝图。

References 参考资料

- Ames, R. T. & Rosemont, H. (1998). *The Analects: A Philosophical Translation*. Ballantine Books, New York.
- Chan, Wing-Tsit. (1963). *A Source Book in Chinese Philosophy*. Princeton University Press.
- Coeckelbergh, M. (2025). Artificial intelligence, the common good, and the democratic deficit in AI governance. *AI and Ethics* 5:1491–1497. https://doi.org/10.1007/s43681-024-00492-9.
- Cuellar, M. (2026). A Virtue Ethics Approach for Using AI-Enabled Technologies. *Journal of Computer Information Systems*. Published online: 03 Feb 2026: DOI:10.1080/08874417.2026.2624677.
- European Commission. (2024). The Artificial Intelligence Act: A risk-based framework for trustworthy AI. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence.
- Future of Life Institute. (2025). AI Safety Index, Winter 2025. futureoflife.org/index
- Legge, James. (1885). *The Sacred Books of China: The Texts of Confucianism, Part III, The Lî Kî.* Oxford: Clarendon Press. The Commercial Press, 2024, Beijing.
- Madianou, M. (2021). Nonhuman humanitarianism: when 'AI for good' can be Harmful. Information, *Communication & Society*, VOL. 24(6): 850–868. https://doi.org/10.1080/1369118X.2021.1909100.
- Onsrud, H. & Campbell J. (2020). Being Human in an Algorithmically Controlled World.
- *International Journal of Humanities and Arts Computing*. 14.1-2. DOI:10.3366/ijhac. 2020.0254.
- Russell, S. (2019). *Human Compatible–Artificial Intelligence and the Problem of Control*. Viking, London.
- Sigfrids, A. et al. (2023). Human Centricity in AI Governance: A Systemic Approach. *Frontiers in Artificial Intelligence*. DOI 10.3389/frai.2023.976887.
- Stanford University Human-Centered Artificial Intelligence. (2026). *Artificial Intelligence Index Report 2026*. https://hai.stanford.edu/ai-index/2026-ai-index-report
- Stenson, T. (2025). What does it mean to be good? The normative and metaethical problem with AI for good. *AI and Ethics* 5:1561–1570. https://doi.org/10.1007/s43681-024-00501-x.
- Vallor, S. (2016). *Technology and the virtues a philosophical guide to a future worth wanting*. Oxford University Press.