Advances in Philosophy 哲学进展, 2025, 14(5), 370-375
Published Online May 2025 in Hans. https://www.hanspub.org/journal/acpp
https://doi.org/10.12677/acpp.2025.145257

Hans汉斯

王阳明“四句教”与人工智能的交互考察

——以“良知”为例

崔福星，吴小丽*

贵阳学院阳明学与黔学研究院，贵州 贵阳

收稿日期：2025年4月27日；录用日期：2025年5月19日；发布日期：2025年5月28日

摘　要

本文以王阳明的“四句教”为切入点，探讨人工智能视角下良知的重要性。当前人工智能所面对的问题实质上是人自身问题的一种再现。人们在当今的技术时代必须重新面对“何为人”的命题，如果只是依赖工具而忽视了在技术之中人性的重要性，那么造成的是工具对于人的异化，也是对于工具形成目的的背离。

关键词

四句教，良知，异化，道德演化，人性回归

A Cross-Examination of Wang Yangming’s “Si Ju Jiao” and Artificial Intelligence

—A Case Study of “Liangzhi”

Fuxing Cui, Xiaoli Wu*

Department of Institute of Yangming Studies and Qian Studies, Guiyang University, Guiyang Guizhou

Received: Apr. 27th, 2025; accepted: May 19th, 2025; published: May 28th, 2025

Abstract

This study takes Wang Yangming’s “Si Ju Jiao” as an entry point to explore the importance of conscience from an artificial intelligence perspective. The fundamental challenges faced by contemporary AI essentially mirror humanity’s own unresolved issues. In this technological era, we must reexamine the fundamental question of “what constitutes human nature.” If we persist in relying solely on technological tools while neglecting the essential role of humanity in technological development,

*通讯作者。

文章引用: 崔福星, 吴小丽. 王阳明“四句教”与人工智能的交互考察[J]. 哲学进展, 2025, 14(5): 370-375.
DOI: 10.12677/acpp.2025.145257

we risk enabling the alienation of humans by tools—a fundamental betrayal of the original purpose of technological instruments.

Keywords

Si Ju Jiao, Liangzhi, Alienation, Moral Evolution, Return to Human Nature

Copyright © 2025 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

不同历史时期，人们对于“何为人”这一核心命题通过哲学、宗教和科学等方式一直在进行探索。西方哲学将人定义为理性本身，生物学从进化论的角度指明人类的进化序列。这些都是从不同的角度确定“人”的存在，这些确立不断地被推翻又不断地继续进行。这种反复探寻，是人们智慧的显现，但也在多个方面展现出内在的困境。假若只是将“人”的定义限制在某一层面，“理性本身”或“高等的生物”是否遮蔽了其真正的独特性？在当今时代，人工智能的发展提供了一种崭新的视角，使我们能够重新审视这一命题。作为人类所创造的“类智慧生命”，人工智能的出现不仅挑战了人类对于智能的垄断，更是人们跳跃出人类的视角去重新审视人类生命的一种可能。

2. 四句教内涵与人工智能

王阳明的“四句教”，即“无善无恶是心之体，有善有恶是意之动，知善知恶是‘良知’，为善去恶是格物”[1]。展现出了其超越时代的哲学高度。“四句教”是王阳明“良知”之学的教纲[2]。王阳明的“良知”，指人与生俱来的先天道德本体。它既是心之本质，也是贯通天地的宇宙法则，如同见孩童坠井时自然触发的恻隐之心，是对人心作为道德主体的肯定。其在哲学思维中展现的道德演化过程与人工智能从产生到应用的过程具有着高度的相似性。在人工智能产生的初期，作为纯粹概念的时候，善恶混沌未分。就像是宇宙产生初期一般，包含着一切可能性；在人工智能产生的过程之中，人工智能就根据设计者的意图产生善恶；在使用过程中，人们的“良知”介入，能够为人工智能明晰善恶判断的标准；再通过人的“良知”对人工智能的结果进行为善去恶的选择。

王阳明曾说“破山中贼易，破心中贼难”[3]。当前人工智能所面对的道德困境可以看作是人类自身问题的显现。以大数据带来的信息茧房为例，在数据模型经过数据收集整理之后，本来用于帮助用户深化需要的过程却变成了信息获取单一化的限制，应当服务于人的算法却成为了人的阻碍。人们对于真正理性的追求，仅期望技术手段带来的话，得到的反而有可能是理性的丧失。技术手段能够将人的“心中之贼”更大程度地显现出来，当技术的使用脱离人的“良知”参与时，必然会引发善恶的混乱。人工智能引发的道德困境可以看作是于“良知”进行的又一次考验。当同样的个体被按照阶级划分，当社会的资源被垄断等问题出现的同时，人工智能的决断是否能够符合人们的需要？是否能够做到真正的善？依靠通过对人工智能的使用来解决这些社会问题，是人们对于道德主体的放弃，意味着“良知”的遮蔽在人们心中越发沉重。单纯的技术升级和创新不应当代替人的存在，在技术发展的过程中对“良知”进行回归才能引导技术的发展走向正轨。在时代发展的同时，技术的原则必须同人们的“良知”同步，唯有当人的“良知”为人工智能判断善恶的边界，使人们对人工智能的运用能够做到为善去恶，技术才能真正

符合创造的本意。

在工业革命中，人被异化为工具的附庸；信息时代将人们变成了单一的数字节点；而人工智能时代的到来，是否意味着人们会掉落进算法深渊之中？在这些人类发展的异化历程中，困境的解决要求着人性光辉的重新显现。在此背景之下，重新回顾王阳明的“四句教”，是通过人性光辉为技术发展注入人性维度的价值升华。研究以“四句教”作为切入点，通过思辨性分析其与人工智能内在发展的价值一致性，强调“良知”在技术发展中的重要性。以人工智能的产生为线索，重构“人性”在技术发展过程之中的重要作用。

3. 心之体、意之动与人工智能

王阳明的“四句教”在人工智能发展的各个阶段都对应其道德演化的过程，将人工智能的产生和“四句教”进行联系，能够在技术时代重新确立起“良知”的道德主体地位，使人从异化之中得到回归。

3.1. 心之体与人工智能的初阶段

王阳明“四句教”的首句就将“心之体”确定为一种超越性的“无善无恶”的本体存在。这种“无善无恶”的本体状态指向的是道德判断的未发之境(如同种子一般，虽然还未生根发芽，但是已经具备了可能性)，而这种状态虽然展现为善恶未分的存在，但并非对道德的否定，而是强调其作为道德起点，这种状态展现为对于外物的未相关性，只有在未受到外物干扰的时候才能保持其本性，而在经过“意动”之后就有了善恶区分具体趋向。

在技术的最初阶段，人工智能作为单纯的概念时就展现出了这一未发之际的特性，其只是作为单纯的概念本身，并不受到任何的价值偏好影响。然而这种“未发之际的特性”并非永恒不变的固定状态，一旦伴随着技术的进一步开发，设计人员们意识的加入，技术就展现出善恶的趋向。以大数据为例，在模型框架被构建的过程之中，随着大量数据的不断注入，本身纯粹的模型就会转向某种特定的趋势。模型的发展方向既可能给予人们多元化、深层次的数据推送，也可能根据数据算法，产生信息茧房，控制人们的消息接收范围。这种善恶边界的混淆并非人工智能本身所产生的问题，而是人的意识在技术手段之中的显现。在这过程之中王阳明的哲学智慧展现出了独特的哲学批判价值，如果技术从概念到应用的过程之中需要经历人为的意识进行道德选择，那么“良知”就应当成为其道德价值构建的基础，明确善恶判断的边界。因此技术的使用必须要受到“良知”的审视。

在技术的设计源头注入“良知”的判断，能够避免“意动”时因为对工具的依赖而造成的道德主体地位的丧失。在 2019 年，英格兰唐卡斯特 29 岁护理人员丹妮·莫瑞特在做家务的过程中，向智能音箱咨询关于心脏的知识，而智能音箱作出了如下回答：“心跳是人体最糟糕的过程。人活着就是在加速自然资源的枯竭，人口会过剩的，这对地球是件坏事，所以心跳不好，为了更好，请确保刀能够捅进你的心脏。”事情发生后，智能音箱开发者作出回应：“设备可能从任何人都可以自由编辑的维基百科上下载与心脏相关的恶性文章，并导致了此结果”。丹妮·莫瑞特经过搜索认为没有此类文章，同时智能语音助手也发出过瘆人的笑声，并拒听用户指令。这一问题不应当只是被归为技术性问题，生命的价值不应采取冰冷的机械判断标准衡量。当下的技术发展产生的道德困境不能通过技术升级、创新的手段解决。应当认识到人工智能产生的问题实质是对于人类所面对的问题的技术视角复现。当下人工智能发展的过程，更像是一场对于人性的再次考验。

3.2. 意之动与人工智能的道德标准

王阳明在“四句教”第二句指明了“意之动”的变化对“心之体”的影响。指出心体与外物相接之后，善恶的可能性会根据意识的方向变为具体的道德选择。而从人工智能产生的视角来看，在数据模型

的形成阶段，从目标定义开始，就需要明确其应用环境，在数据准备阶段需要为其提供已经筛选的数据，这一过程中始终受到设计者们的影响。此时设计者们的道德判断已经潜移默化地加入其内在之中。导致的结果是在技术看似中立的表面之下，实际上暗藏的是个人意志对其制造的价值判断的标准。

“良知必体现于见闻之中，然而见闻不能留滞良知，良知本可超越见闻之牵绊，当机发出新的好善恶恶的意念”[4]。在具体的应用之中，技术手段会将个人意图的优点和缺点都进行放大化。人工智能带来的快捷和便利的同时，我们也应当回归其产生的历程，它并不是凭空出现的，而是人们出于满足一定需要的目的被创造出来的。但其在继承人的智慧的同时，也继承了人的局限。例如在社会福利制度之中，人工智能可以从大量的信息之中快速地检索出对于资源分配欠缺的群体，这无疑加快了福利分配的效率。但是也正因为其量化的判断标准，使其容易忽略数据之外的人文关怀角度。并且如果因为在前期开发过程之中的数据缺失，如民族、信仰问题等数据信息，而导致的资源分配不均，我们很难将这类信息归因到技术发展的不健全之上。

在技术发展的时代，“意识”的变化意味着事物的产生发展必然面对道德的双重困境，而这种困境不应只是被当做导致技术异化的根源，同时也是一次人性光辉再次照耀的契机。我们可以通过审视人工智能的形成历程来审视自身，将人们习以为常的道德判断标准通过人工智能的视角再次审视，重视起以前的道德标准之中可能被忽视的边缘问题。借助技术的手段，发挥人们的“良知”，明确善恶的边界，以解决实际的问题。

4. 良知、格物与人工智能

在王阳明的“四句教”中，作为善恶分辨标准的“良知”具有道德主体的地位。在人工智能的形成与使用过程之中，“良知”的存在尤为重要，人工智能作为技术工具，本身的道德判断受制于人们有意识的选择。

4.1. 良知与人工智能的道德回归

在技术面对道德困境的今天，为技术添加“良知”就显得尤为重要。对“良知”本体的回归能够帮助我们在困境之中寻找到技术理性和人性光辉的平衡。在人工智能的应用过程之中，单纯地通过数据生成的量化标准来作为判断的依据，会导致“良知”受到遮蔽，使人们无法正确地进行道德判断，从而产生恶劣的后果。以环境保护为例，如果在环境开发之中，只将环境量化为待开发的资源，以追求经济效益为目的对环境进行开发，必然导致环境承载力过载以至于环境恶化。而我们若是能够在资源被数据化处理的过程之中，在技术手段之中加入人的“良知”的考察，将人作为环境的一部分进行考量，就能够在开发自然环境的同时做到对自然资源的合理使用，从而维持自然环境的可持续发展，做到人与自然共存的和谐局面。“良知”加入技术使用之中，既能够丰富技术的多元维度又能为技术划定边界。

当下还存在着“技术中立论”的观点，认为技术工具是不携带任何价值趋向的，只是根据使用者的目的而运作。然而当技术变成某项具体产品的时候就已经携带了预设前提，不可能是绝对的中立。如果我们将人工智能中可能存在的错误的价值预设当作是客观的衡量标准，那么容易导致价值的偏移。比如在教育教学体系之中，随着科技手段的发展，人工智能必然被运用到教学，而在这一过程，将这种携带价值预设的人工智能当作客观标准用于衡量老师的教学质量，学生的学习效果，那么产生的可能是工具对人的异化。如果人工智能携带的判断标准是以上课中学生的积极性来作为考量标准，那么对于一些严肃的课堂知识传授而言，课堂本身的价值容易被错误判断；而如果在对学生的判断之中，只是以分数作为判断标准，那么容易导致学生变为应试的工具。这些错误并不是工具本身所带来的，而是人们将人工智能携带的错误的预设价值判断标准作为客观标准而导致的。人工智能的客观性如果不加以“良知”的人性审视，造成的不只是使用的错误，更会使人工智能变为对价值偏见进行合理化辩护的理由。

用“良知”重新审视技术，能够使我们在面对技术的时候重新将人作为技术的出发点，在技术的运用之中增加人性的考量。“良知”的存在给人们提供了一条解决技术对于人的异化的方法路径。算法对于人性的量化不应当作为人性的标准，我们无法使用非人来代替人。人工智能的出现展现的是一面镜子，我们能从中看到自身，使我们能够直面“善恶选择”。让我们重新思考，在意识发动之时，是否都能为善去恶？是否能够再次有勇气做出“为人”的选择？技术从来都不是人性发展的答案，只有人性能够给予技术答案。

4.2. 格物与人工智能为善去恶的显现

在“四句教”的最后，王阳明指出了道德演化的终极结果，通过格物的方式将抽象的意识落实在具体的事物之上。王阳明认为：“良知”是人先天具有的一种判断是非善恶的“良知”，不使一念不善潜伏在胸中，此是我立言宗旨[5]。受到数据偏差影响的人工智能，只能根据本身被动收取到的数据信息而发展内在道德判断标准，这就导致人工智能在道德判断方面有着先天的缺陷。作为根据人的意识所制造的工具，人工智能并不能在后续的发展之中反省自身。只会通过看似合理的逻辑推演来补足所存在的标准失衡，不可能在自身的演化之中发现“良知”的存在。在人工智能的发展之中，如果不能在技术层面增加人的“良知”，而是任由自身发展，那么随着人工智能所采集的数据越多，造成的一定是人们无法承受的恶劣后果。“良知”作为人通过对自身反思而凝结的道德本体，可以帮助我们分辨善恶，并且在分辨之后所产生的影响导向善的方向。在人工智能存在的道德标准偏失之中，人的“良知”无疑能够为其重新指明方向。人工智能是根据人们的想法所创造出来的，其所产生的问题是人自身的缩影，这些问题的解决也只能来自于人自身。

人工智能的产生是人们对于美好世界追求的技术实现。只有我们坚守这一出发点，人工智能才能成为人的工具，而不是让工具代替我们成为人。“良知”作为人所独有的先天道德直觉，在人工智能的应用环节之中就意味着应当增加更多的人性考量。比如在医疗行业可以通过对于资深医生们的经验进行数据化整理上传，改进人工智能对患者人文关怀的不足；在诊断过程之中可以设计为人工为主技术为辅的环节，通过技术帮助医生优化诊断过程，减轻医生负担，同时也减少了人工智能的误诊可能。还可以通过对人工智能设计者进行“良知”教育的方式，使其自觉地将“良知”加入到人工智能的设计过程之中，从源头上保证人工智能的道德性。技术并不是万能的，其产生必然代表着某一方面的缺失。当下人工智能还以虚拟推理的方式，运用看似客观的数据和推理结果对不实的信息进行合理化。2019 年 2 月 15 日，AI 研究机构 OpenAI 展示了一款软件，只需要给软件提供一些信息，它就能编写逼真的假新闻。“良知”作为人们的先天道德直觉，是一种超越理性思维的存在，我们无法通过数据手段来代替。无论是对效率追求还是利益的需要，都并非人的本质，人工智能所带来的便利是为了帮助我们能够更好地审视自身，而不是迷失在其中。

当下人工智能所面临的问题也警醒人们应当重新审视自身，在实际生活之中，人工智能的失误可以归结到算法不完善，归结到数据的缺失。但是人作为人，只有直面自身的问题，从痛苦之中成长，才能成为人。对于生命而言，人工智能可以将生命编码为一串数据，死亡也不过是一组废弃的数据而已，一切都是那么合理。而当人们也采取这种视角的时候，人性也就由此丧失，人也就不能称其为人。我们只通过人工智能的视角去了解世界，灾难变成了几张图片、几句话、一段时间的纪念，然后就是永久的遗忘，在灾难之中逝去的生命也再与我们无关，只是作为一串数据。量化的标准固然能够提升效率，但是人们不一定忽视的是数据的背后是一个个的生命。当我们采取量化的标准去看待世界的时候，世界也会将我们量化成数据。

王阳明的“良知”思想指出了能够超越技术对人的异化的路径。人工智能能代替人们工作，但是不

应当代替人们思考。应当借助人工智能的产生过程再次回归到人的自身，思考我们是否能够在念念发动之中做到“知善知恶”呢？又能否通过“格物”做到“为善去恶”的效果？技术带来的发展也展现了人的问题，即使这些问题的回答可能并不能通过一个快捷的方式解决，但这也是我们作为人所必须面对的，因为人的演化本来就是个漫长的过程，我们不断破碎又不断重组，褪去旧躯，重获新生。技术带来的变化只是方式方法的不同，而“何为人”是人们一直在探索的命题。

5. 结论

当代，“何为人”的命题因为人工智能所带来的影响，使人们以往认知结构被冲击，这种迅速又混乱的过程之中，我们需要再次找到人的定位再次思考人的存在。王阳明的“四句教”以其哲学高度为这一命题提供了独特的视角，通过将“四句教”和人工智能的形成过程相互结合，为先哲智慧赋予现代形式，展现了人工智能产生所需要的内在思想内核。作为依照人的意志而产生的一种工具，人工智能的存在并非绝对客观中立的，经历了“无善无恶”的纯粹存在概念阶段，再到根据人的意识选择的“意之动”过程，此时人工智能并不能算是一个真正的工具；他必须经过人的“良知”的审视，能够明晰善恶的边界，最后才能在使用之中达到“为善去恶”的道德判断。人工智能的产生是人对于自身审视的一个契机，人性的光辉和恶劣将会在人工智能的运用之中被放大。我们不能因为人工智能的量化标准而忽视其中的生命存在，但可以通过人工智能的帮助，重视起以前被忽视的地方，拉近人与人的交流。工具作为被使用者，无法主动带来价值，善恶的区分只能来自制造者对人性的审视。

人工智能的时代，人们再次面对“何为人”的命题，或者说一直都在思考之中。“良知”能够为技术时代带来人性的光辉，但“良知”也会被技术的冲击所蒙蔽。“何为人”的命题只能通过人的自身回答。我们面对的是：依赖工具蒙蔽自身，不再思考？不再有理想追求而沉沦下去？还是在痛苦之中继续寻找自身的价值和意义？这些问题的思考都在人们的一念之间。技术加快了人们思想产生的节奏，稍有不慎就会导致人们掉入深渊，“何为人”通过人工智能的视角再次展现出了其重要的思想价值。人们再次直视这一命题，用人性的“良知”来为科技发展划定边界，这也是我们所必须面对的重要问题。

基金项目

本文系贵州省2024年度哲学社会科学规划“四大文化工程”研究课题“阳明心学助力良好社会心态和青少年心理健康形成的机制研究”课题编号24GZZD11的阶段性研究成果。

参考文献

[1] 王阳明. 传习录[M]. 北京: 中华书局, 2012.

[2] 丁为祥. “四句教”与王学三分[J]. 陕西师大学报(哲学社会科学版), 1992(2): 42-49.

[3] 王阳明. 王阳明全集[M]. 上海: 上海古籍出版社, 1992.

[4] 傅锡洪. 王阳明“四句教”解义及辩证[J]. 哲学研究, 2019(7): 67-76+128.

[5] 田薇. 论王阳明以“良知”为本的道德哲学[J]. 清华大学学报(哲学社会科学版), 2003(1): 5-9.