人工智能“以人为本”伦理准则反思

——一种马克思主义哲学的分析思路

潘恩荣　孙宗岭

摘要：针对人工智能可能产生的负面影响，西方社会提出的人工智能“以人为本”伦理准则事实上已经吸纳历史教训和马克思主义思想，但仍然存在着两个局限——逻辑性悖论和人性漏洞。前者指向人工智能（机器）本身的问题，西方社会力图用伦理准则规范人工智能，但又要求“人工智能不能具备独立的自我意识，且不能具备辨识善恶的伦理能力”。后者指向“人工智能（机器）的资本主义应用”问题，但在西方社会背景中却难以堵住这个漏洞。马克思关于工业革命和机器大工业批判等研究为人工智能伦理准则提供了新的“以人为本”路径。

关键词：人工智能；以人为本；逻辑性悖论；人性漏洞；机器大工业批判

中图分类号：N02　　**文献标识码**：A　　**文章编号**：1000—8691（2022）06—0030—10

一、引　言

18 世纪 60 年代，现代机器首次崛起全面增强了人类的劳动能力，却没有达到“以人为本”的效果，反而导致了机器异化、贫富分化、社会分裂和冲突。21 世纪初，以人工智能为代表的新机器“二次崛起”[①]。作为第四次工业革命的代表性技术，人工智能已经开始对人类生活生产进行广泛赋能，又一次全面增强了人类的劳动能力，“智慧城市”“智慧物流”新业态蓬勃发展。然而，越来越多的事实表明，人工智能应用带来了算法歧视、就业替代和隐私泄露等新老问题，冲击着现有社会秩序。鉴于现代机器与第一次工业革命的历史发展经验，人们不禁发出疑问，以人工智能为代表的新机器和第四次工业革命是否重蹈历史覆辙造成新的社会分裂？还是以其“智能”弥合分裂走向“以人为本”？

针对上述疑问，西方各国政府、企业、科研机构、组织以及联合国纷纷提出人工智能伦理准则，倡导将“以人为本”（Human-centered）作为人工智能发展的宗旨。机器能够增强人类的劳动能力，其根源在于前者在特定方面的能力超过人类的能力。那么，人工智能“以人为本”伦理准则指的是，在人与“超人”的（人工智能）机器之间，凡事应该坚持“以（凡）人为中心”，确保“人”的各项权利和权力不受损害。然而，在“人与机”关系中，“超人”的（人工智能）机器可能“以（凡）人为中心”吗？如果可

基金项目：本文是国家社会科学基金重大项目“当代新兴增强技术前沿的人文主义哲学研究”（项目号：20&ZD045）、国家社会科学基金项目“技术哲学社会人工物研究”（项目号：19BZX027）的阶段性成果。

作者简介：潘恩荣，男，浙江大学马克思主义学院教授，博士生导师，主要从事技术哲学和人工智能伦理研究。
孙宗岭，男，浙江大学马克思主义学院博士研究生，主要从事马克思人机关系思想研究。

① 人工智能机器等数字化技术标志着机器的二次崛起，参见 [美] 埃里克·布莱恩约弗森、安德鲁·麦卡菲：《第二次机器革命》，蒋永军译，北京：中信出版社，2016 年，第 VIII 页。

能，又需要怎样的条件？为了回答上述问题，本文首先梳理人工智能对人类生活和社会秩序的可能影响，接着归纳分析西方社会对人工智能的态度及其相应的局限，最后基于马克思的机器大工业批判探讨克服人工智能负面影响的思路。

二、人工智能可能的影响

机器介入且改变人类生活和社会秩序是现代社会的基本特征之一。自 18 世纪英国第一次工业革命以来，人类历史上已经发生了至少三次明显的机器革命，即机械化革命、电气化革命与信息化革命。由此引发的规模宏大的“机器换人”过程，在西方社会引发了人类生活和社会秩序的颠覆性变迁，经历两百年之后最终形成相对稳定的“人与机”关系。

人工智能是一种新的机器形态，其特殊性在于一种革命性的“两自能力”，即以深度学习技术为核心的人工智能具备革命性的“自主学习能力”及相应的“自我进化”能力。[①] 人工智能凭借其“两自能力”再次冲击传统的“人与机”关系，颠覆性地介入和改变人类生活和社会秩序。与以往的非智能“机器换人”过程相比，人工智能的冲击发生了明显变化，引发了一系列新的影响。概括来说，人工智能可能的影响可归纳为近期、中长期与远期三类。

人工智能的近期影响主要是“失业问题”与“隐私问题”，其原理是人工智能短时间冲击“人与机”关系在社会层面引发的应激反应。“失业问题”是一个历久弥新的老问题，“随着技术的发展，机械化、自动化、人工智能的程度高了，技术肯定会取代一部分人力，在创造一些新的行业和就业岗位的同时，把一些传统的行业和岗位淘汰掉”[②]。麻省理工学院和波士顿大学发布的研究显示：1990—2007 年，每千名美国工人中增加 1 个机器人，全美就业人数比例下降 0.2%。[③] 社会对机器革命引发的创伤已深入社会文化意识，因而，一旦有新技术和新机器出现，社会层面便会出现相关的应激反应。因此，2016 年“阿尔法狗”（Alpha GO）事件之后，人们的第一反应是人工智能可能带来新一轮规模宏大的“机器换人”效应和相应的失业问题。“隐私问题”是一个因人工智能而规模性爆发的老问题。在西方社会，尤其是标榜“自由”的社会氛围下，个人隐私是一种优先级别非常高的社会事务。在互联网社会，虽然有巨量的数据沉淀，但个人隐私信息也淹没在数据汪洋中。虽然偶有出现侵犯隐私问题，只要不是规模性社会事件，隐私问题并未成为令人担忧的社会问题。然而，当人工智能作为世界新一轮科技革命和产业变革通用技术且具备“两自能力”的时候，个人隐私信息在人工智能数据挖掘过程中无所遁形。因此，大规模隐私泄露已经成为事实而不只是一种风险。在国内，隐私泄露引发各种“大数据杀熟”事件，最终导致国内《个人信息保护法》出台明令禁止此类行为；在国外，隐私泄露引发欧盟设立《通用数据保护法》等苛刻的法律以保护个人隐私。

人工智能的中长期影响主要是“歧视问题”，如数据歧视和算法歧视等，以及由此引发的社会分裂和分化问题。一方面，“数据歧视”放大了人类社会文化原有的分裂情绪，可能加剧社会分化。当前的人工智能的核心深度学习技术，可以不通过人类专家的帮助自行在大规模数据学习中建立抽象的特征规则。[④] 然而，数据并不必然具有中立性。人类的社会文化偏见沉淀于大数据之中，人工智能算法不同程度上学习并放大了人类社会文化中的歧视情绪。例如，COMPAS 是一种在美国广泛使用的存在明显偏见的算法，该系统预测的黑人再次犯罪的风险远高于白人。[⑤] 也就是说，大数据算法实现预测的过程，内含人

① 早期关于人工智能能力及其可能影响的研究，参见潘恩荣、阮凡、郭喨：《人工智能“机器换人”问题重构——一种马克思主义哲学的解释与介入路径》，《浙江社会科学》2019 年第 5 期。

② 韩天琪：《技术进步会让我们失业吗？》，《中国科学报》2017 年 11 月 13 日。

③ D. Acemoglu and p.Restrepo,“Robots and Jobs: Evidence from US Labor Markets”, *National Bureau of Economic Research Working Paper Series*, No.23285, 2017.

④ 周志华：《关于深度学习的一点思考》，于剑、封举富、张敏灵等编：《机器学习及其应用 2019》，北京：清华大学出版社，2019 年，第 1—9 页。

⑤ 王燃：《大数据证明的机理及可靠性探究》，《法学家》2022 年第 3 期。

类认知的歧视情绪。另一方面，“算法歧视”是人为将某种情绪嵌入人工智能算法之中，进而导致有偏见的结论，可能激化社会矛盾。如利用人工智能进行分类筛选时，由于设计者对某些人群带有某些负面情绪，可能出现人工智能系统做出不利于后者的预测。

人工智能的远期影响是人类命运问题，如人工智能奇点问题、无用阶级问题和人类种族存续问题。“人工智能奇点问题”主要担忧未来人工智能有可能出现“奇点”，如人工智能诞生自我意识，进而出现不可控的风险。早在 20 世纪 50 年代，波兰数学家斯塔尼斯拉夫 • 乌拉姆在冯 • 诺伊曼的悼文中就担忧过：“技术正以其前所未有的速度增长，……人类社会的发展正逼近某个奇点，一旦超越了这个奇点，我们现在熟知的人类社会将变得无法预测。”[①] 雷 • 库兹韦尔（Ray Kurzweil）将“技术奇点”引入人工智能领域。他指出，随着人类知识与人类技术结合度不断提升，知识共享将可以完全实现，人类与机器将形成深度融合，文明将超越人脑的限制，技术奇点出现。[②]“无用阶级问题”指的是人工智能持续加剧社会分化最终导致绝大部分人成为对社会经济“无用的普通人”，而另外一小部分人成为“超级人类”。[③]“人类种族存续问题”指的是人类可能过于依赖人工智能而引发某种退化，出现“主奴互换”，导致人类被人工智能奴役甚至灭绝。首先，由于越来越依赖便捷的人工智能技术，人类可能失去自身有用的技能和思维能力。发达的互联网使得信息的接受与传达分外简单，在长期的碎片化信息接受习惯下，人们思考能力与对于事物的分辨能力将面临下降风险。数字化和智能化服务给人类在生产生活中提供便捷的同时，使得人们直接参与生产生活的机会变少，思考和劳动能力在不明显或无意识的状态下渐趋弱化。其次，由于人工智能的“两自能力”不断增强，人类自主性和对人工智能的最终掌控权遭到不断侵蚀。人工智能已经在许多方面替代了原本仅属于人类的脑力行为，如许多社交网站或者购物网站通过用户数据来进行计算，智能化选择性地推送产品给用户。这种看似智能便捷的做法其实在某种意义上限制了人类的自主权和决策权，人类不自觉地将思维和选择的能力转移到人工智能身上。长此以往，人的取向或者判断将被人工智能限制在设定好的界限里，产生“信息茧房”效应。同时，高智能的定制化推送和信息获取，会使人们弱化甚至丧失自主思考和决策能力。如果人工智能奇点来临，在西方“自由至上”思想的影响下，人工智能很可能翻身做主人。例如，斯蒂芬 • 威廉 • 霍金认为，人工智能会在将来的某一天赶上甚至超过人类，人工智能发展可能导致人类灭绝。[④] 如果人工智能永远不会产生自我意识，在人类深度依赖人工智能而失去自主思考能力的背景下，整个人类社会也可能因为没有新的突破而陷入退化和消亡过程。

三、西方社会对人工智能的态度及其局限

西方社会对人工智能的态度总体上可概括为“以人为本”，强调人工智能须“以人类为中心”，不能僭越“人”的地位和权利。据词频统计，人工智能伦理准则的关键词包括以下各类：为人服务（For Human）、可持续性（Sustainability）、合作（Collaboration）、共享（Share）、公平（Fairness）、隐私（Privacy）、问责（Accountability）、透明（Transparency）、安保（Security）、安全（Safety）和人工智能长远发展（Long Term AI）。[⑤] 其中，前七类直接与人相关，后四类主要针对人工智能技术，但也间接与人相关。因此，西方社会的人工智能伦理准则总体上将归宿点聚焦于“人”的存在和发展，即确保人工智能安全、可控、可靠地朝有利于人类与社会发展的方向发展。

在人工智能“以人为本”伦理准则中，西方社会对“人”的理解有多重涵义。首先，西方社会对“人”的理解是从“以神为本”走向“以人为本”。前现代时期，西方社会遵循基督教体系“以神为本”的思路，

① 朱彦明：《奇点理论：技术“复魅”世界？——批判地阅读库兹韦尔的〈奇点临近〉》，《科学技术哲学研究》2020 年第 6 期。

② [美] 雷 • 库兹韦尔：《奇点临近》，李庆诚、董振华、田源译，北京：机械工业出版社，2011 年，第 9 页。

③ [以色列] 尤瓦尔 • 赫拉利：《未来简史：从智人到神人》，林俊宏译，北京：中信出版社，2017 年，第 315 页。

④ S. Hawking, *Brief answers to the big questions*, New York: Bantam Books, 2018, p.182.

⑤ 参见“链接人工智能准则平台”，https://www.linking-ai-principles.org/cnkeywords。

神与天使是“超人”，天使和人是神创造的“产品”，是神的奴仆。从“神”的角度看，当天使具备独立意识时，天使会拒绝神的旨意，挑战神的权威。在基督教一神论“以神为本”的设定下，这是神不可接受的结果。近代西方文化代表性诗歌——《失乐园》完整地表达了这样的场景，构建了人“从何而来”的概念。

其次，西方社会人文主义思潮构建了“自然人”概念，带有“人权不可侵犯、至高无上”的神圣属性。人文主义将“自然人”的价值、尊严和生命等“自然属性”充分地拔高和神圣化，替换了中世纪基督教“神”或“神圣人”概念。洛克的自由主义与社会契约思想影响了后来康德的“人为自然立法”、法国的“天赋人权”等思想，最后影响了美国《独立宣言》（1776）和法国《人权宣言》（1789），成为近代西方社会的一个基础性伦理价值。在当前人工智能伦理准则中，“自然人”含义更多地体现为人类的福祉、尊严、自主、自由和公平等关键词。

再次，西方社会科学主义思潮构建了“理性人”概念，带有某种“全能的”（Almighty）的神圣属性。科学主义可追溯到古希腊的“逻格斯”（Logos）思想，启蒙运动后集中体现在以笛卡尔为代表的理性主义思想。“理性人”虽然替代了宗教的“神”或“神圣人”，但也保留了某些“类神”特征，如认为人具有无限理性或完全理性。在 18 世纪第一次工业革命过程中，“理性人”得到了进一步的衍生。在亚当•斯密那里，“理性人”表达为一种“理性经济人”概念，每个人都谋求自身利益最大化且具备无限的理性。在当前人工智能伦理准则中，“理性人”含义更多地体现为人类的有益、隐私、最终掌控权、共享、合作和问责等关键词。

最后，西方社会人工智能伦理准则的“以人为本”思想不仅承袭了“西方中心论”思想和“个人自由主义”思想，还进一步表达了人类与人工智能之间的某种“人类种族中心主义”情绪。经历宗教改革（16 世纪）、资产阶级革命（17 世纪）、自然科学革命（17 世纪）和工业革命（18—19 世纪）之后，西方社会涌现出一种“西方式现代化”的文化信念，即“新教—白人—自由”的意识形态。其中，“新教”部分主要表达了“一神论”与（西方）人作为“神”创造的对象和其地位是“一神之下、万物之上”等思想；“白人”部分主要表达了“种族主义”思想，即白色人种优先于其他有色人种；“自由”部分主要表达了“个人自由主义”和“个人人权至上”等思想。当新一代基于深度学习技术的人工智能兴起之后，上述“现代化”意识形态迅速占领了人工智能伦理准则的主流位置。因此，人工智能“以人为本”伦理准则直白地阐明“人类中心论”思想——人类相对于人工智能的优先性，不仅提倡保障人的权力和权利，还表达了对人工智能可能作为新“种族”的恐惧，以及人工智能必须置于人类掌控之下的思想。

然而，西方社会提倡的人工智能“以人为本”伦理准则存在两方面的局限，可能会导致伦理准则陷入自相矛盾的困境。一个局限是逻辑性悖论，即西方社会力图用伦理准则规范人工智能，但最后不接受具备伦理能力的人工智能。另一个局限是人性漏洞。

（一）逻辑性悖论

首先，西方社会对人工智能的态度深受“自然人”传统及其背后的基督教文化“人神关系”的影响。人工智能首先是人类的“产品”，诚如人类是神的“产品”一样，前者应该完全服从后者并遵循以后者为中心的伦理准则。当“人与机”关系处于“人为主、机为奴”的“主奴关系”时，在人工智能增强人类劳动能力的同时，“以人为本”是可能的，也是可行的。非智能的现代机械机器如同一般的天使，完全听命于控制者。当前弱人工智能如同具有一定自主性却没有独立自我意识的天使，仍然不会主动“反抗”其创造者的权威和最终控制权。

其次，人工智能的能力超越人类能力将引发人类对人工智能的担忧、恐惧和反抗。近代西方文化用“自然人”替代了“神”的位置后，强调人权高于一切，从而难以接受（某些）能力超越人类的人工智能。基于“神—天使（奴仆）”思想，“自然人”认为“人与机”关系是天然的“主奴”关系，即人为“主”、人工智能为“奴”。这是“人与机”对立的“主奴关系”理解模式。如果人工智能的能力超越人，机器便具备了“超

越”或“替代”人的可能性。第一次工业革命的历史表明，机器增强只有部分的“以人为本”功能效果，即其中一部分人得益，而另一部分却未分享到红利，于是后者群起而攻机器。例如，在哈格里夫斯研发珍妮纺纱机期间，邻居担忧失业而强行冲进其住所，捣毁珍妮纺纱机样品和砸碎家具，[①] 后期此类事情演化为“卢德运动”(Luddite Movement)。同理，在人工智能机器崛起的第四次工业革命期间，当人工智能机器出现了一些“超人的智能”，人们也会开始担忧人工智能机器。尤其是 2016 年“阿尔法狗”战胜人类围棋职业高手后，人们猛然发现原来机器智能胜过“人类智慧”，原来处于奴仆地位的人工智能可以在智力上超越原来处于主人地位的人类。社会文化意识中的一些恐惧和担忧情绪涌现，推动了各种人工智能“以人为本”伦理准则爆发式出台。

再次，具有独立自我意识的人工智能逻辑上将反叛人类主人。神与天使的关系以及它们与人类的关系喻示了人与人工智能的关系。既然被创造出来的人类可以拥有独立的自我意识，那么，被人类创造出来的人工智能在逻辑上也可能诞生独立的自我意识。然而，在西方社会意识中，具备独立自我意识的天使最终反叛了神；同理，具备独立的自我意识的人工智能逻辑上也将反叛人类。因此，即便是刚刚出现一点“自主性”的人工智能也引发了西方社会的担忧，因为前者引发了人类不可解释的“认知黑箱”，即当前的人工智能有远离人类控制的倾向，人与人工智能机器的“主奴关系”出现松动。[②] 人工智能“奇点”理论指向的就是极端情况下的“主奴互换”：人工智能机器涌现出独立自主的意识，且机器无论从体力上还是在智力上都是“超人的”。“人们开始忧虑人工智能可能引发人类命运的担忧，动摇人类作为万物之灵的地位”[③]，人工智能时代的“新卢德主义”思想崭露头角。未来强人工智能机器如同具有自我独立意识的高级天使，能够主动“反抗”主人权威。因此，逻辑上，在西方基督教文化和自然人思想背景下，人类最终不会接受强人工智能机器。

最后，人类也不会接受具备伦理能力的人工智能机器。从西方“神—天使（奴仆）—人”之间的关系看，神既不接受具备独立意识的天使，也不接受具备“辨识善恶能力”的人类。在《伊甸园》的叙事中，当人类始祖受诱惑而食用了具有“辨识善恶能力”的智慧树果实之后，神开始担心具备“辨识善恶能力”的人类进一步追求生命树上的永生果实，最后威胁到神的地位，于是把人类从伊甸园中驱逐到地面上。“辨识善恶的能力”本质上是一种伦理能力。既然基督教文化中神不接受具备伦理能力的人类，那么，按照近代以来西方对“自然人”的理解，人类也不接受具备伦理能力的人工智能机器。否则，人工智能就可能进一步“追求生命树上的果实”，要求人工智能（机器人）的生命权，进而要求“人机平权”，甚至可能要求“天赋机权”，最终动摇神圣的“自然人”地位和权力。从这个角度来看，西方人工智能伦理准则把“以人为本”作为最高原则之一，其实质并不是培养具备伦理能力的人工智能机器，而是要求人工智能机器处于人类奴仆的状态。那么，西方社会提出和倡导的人工智能伦理准则将走向自我否定。

（二）人性漏洞

无论人工智能是否具备独立的自我意识，从“理性经济人”角度看，人类可以冒风险接受人工智能机器出现“主奴互换”。用西方的话语来说，这是人类与魔鬼做交易，不管是小魔鬼（弱人工智能）还是大魔鬼（强人工智能），都是极度危险的事情。按照“理性经济人”的思维，只要机器增强比人类更有“能力”获得利润空间，无论是生产、流通还是交换场景中，一定会出现“机器换人”现象。低利润空间中的“机器换人”是合乎理性的经济行为。尽管一部分人未必能享受到人工智能增强带来的福利，但仍有一部分人类精英可以得益于人工智能增强。然而，资本的本性是追逐高利润并且呈现无克制的趋势，“一旦有适当的利润，资本就胆大起来，……有 50% 的利润，它就铤而走险；为了 100% 的利润，它就

① [英] 罗伯特·艾伦：《近代英国工业革命揭秘：放眼全球的深度透视》，毛立坤译，杭州：浙江大学出版社，2012 年，第 293—294 页。

② 潘恩荣、曹先瑞：《面向未来工程教育的人工智能伦理谱系》，《高等工程教育研究》2021 年第 6 期。

③ 潘恩荣、张为志：《无科学，不哲学》，《中国科学报》2018 年 12 月 24 日。

敢践踏一切人间法律；有 300% 的利润，它就敢犯任何罪行，甚至冒绞首的危险。如果动乱和纷争能带来利润，它就会鼓励动乱和纷争”[①]。因此，基于资本逻辑发明和使用人工智能容易走向极端，进而引发一些不可逆的社会事件。一种极端情况是，在高利润空间的诱惑下，“理性经济人”就敢冒“与魔鬼做交易”的风险接受“主奴互换”式的机器换人，不惜让人工智能主宰人类命运。例如，在波音 737MAX 坠机案例中，在高额商业利润的刺激和竞争下，波音设置人工智能自动驾驶系统的“决策权限”高于人类飞机驾驶员，但没有严格地验证自动驾驶系统的可靠性和安全性并执行知情同意伦理原则。当智能机器决策正确时，“理性人”和（神圣的）“自然人”都接受并“享受”这样的状态，前者获得高利润，后者获得高质量服务。但是，当智能机器决策错误导致灾难发生时，（神圣的）“自然人”开始拒绝接受这样的人与智能机器的“主奴互换”状态，表现为西方发达国家同样纷纷禁飞波音 737MAX 机型。事实上，这是人类对人工智能的禁用。另外一种极端情况是，如果运气足够好，“理性经济人”不断地成功实施“机器换人”最后走向人工智能大工业。在弱人工智能条件下，绝大部分人类可能成为对社会发展“无用”[②]的低阶群体，而一小部分人类因为掌控人工智能而成为高阶人群；在强人工智能条件下，人类整体上反而是“无用阶级”。

四、“以人为本”新思路

西方在人工智能伦理准则和立法的先发优势已经开始影响到中国设立人工智能伦理准则和立法的思路。然而，西方社会提出并倡导的人工智能“以人为本”伦理准则存在两个局限。如果完全照搬与沿用西方社会的“以人为本”人工智能伦理准则，中国人工智能伦理治理将囿于逻辑性悖论和人性漏洞带来的困境。

马克思关于第一次工业革命和机器大工业批判的研究为人工智能伦理准则提供了“以人为本”的新线索。《资本论》可以看作“欧洲工业化背景下关于技术进步、技术创新与制度变革之间的相互作用和影响的批判和反思”[③]。当前西方社会人工智能“以人为本”伦理准则的思想根基是“西方式现代化”的文化信念，《资本论》恰是以完成工业革命的英国为例批判和反思“西方式现代化”。那么，马克思的思想有助于应对人工智能“以人为本”伦理准则的两个局限。

西方社会人工智能“以人为本”伦理准则的两个局限之间的关系是“人工智能本身与人工智能资本主义应用”的关系。逻辑性悖论指向人工智能（机器）本身的问题，西方社会力图用伦理准则规范人工智能，但又要求“人工智能不能具备独立的自我意识，且不能具备辨识善恶的伦理能力”。前者指的是“以人为本”的人工智能应当是“弱人工智能”而非“强人工智能”；后者指的是人工智能伦理准则的目标之一就是人工智能能够辨识善恶，即伦理能力是人工智能的固有功能之一。人性漏洞则指向“人工智能（机器）的资本主义应用”问题，也是“人工智能异化”问题，要求“人类社会本身需克制资本逻辑，避免人类使用人工智能的行为走向极端”，但在西方社会背景中却难以实现。

（一）应对西方人工智能伦理准则局面

从马克思的思想角度看，上述关系本质上是“机器与机器的资本主义应用”之间的关系在人工智能维度的再现。那么，基于马克思的思想，尤其是《资本论》及其手稿关于工业革命和机器大工业批判的思想，有三条路径能够应对西方社会人工智能“以人为本”伦理准则的两个局限。

首先，根据历史唯物主义的原理，人工智能“以人为本”伦理准则及其实践的范围只能是“弱人工智能”。历史唯物主义的基本原理之一是社会存在决定社会意识。那么，当前社会中关于人工智能的现实是什么？

① 这段话是马克思引用自托·约·邓宁《工联和罢工》1860 年伦敦版第 35、36 页的内容，参见《马克思恩格斯文集》（第 5 卷），北京：人民出版社，2009 年，第 871 页。

② [以色列] 尤瓦尔·赫拉利：《未来简史：从智人到神人》，第 315 页。

③ 更多参见潘恩荣：《创新驱动发展与资本逻辑》，杭州：浙江大学出版社，2016 年，第 3—6 页。

关于人工智能是否出现自我意识、是否引发奇点效应的问题，人工智能科技专家主流的意见是：在理论上，人工智能涌现自我意识是可能的，但在现实中，“哲人们担心的事情一百年都不会发生”[①]。因此，社会现实是当前及未来相当长的一段时间内，人工智能只是“弱人工智能”而非“强人工智能”。那么，逻辑性悖论前半部分关于“强人工智能”的担忧，是没有社会现实基础的虚假的社会意识，虽然有意义但不是当前人工智能“以人为本”伦理准则应该做的事情。

其次，根据人的自由而全面发展思想，人工智能“以人为本”伦理准则及其实践欢迎具备伦理能力的人工智能。关于人工智能是否具备“辨识善恶”的伦理能力的问题，这是人工智能自身发展、人工智能伦理和治理追求的终极目标之一。但是，回顾工业革命以来西方社会对先进科技的态度，就会发现，掌控机器和先进科技的群体，在资本逻辑的主导下，为了资本增殖和扩大再生产可以肆无忌惮地“机器换人”，罔顾工人身心健康和安全。如果人工智能具备“辨识善恶”的伦理能力，那么，人工智能将会阻碍甚至反抗机器的资本主义应用、资本主义生产方式及其相适应的生产关系和分配关系。马克思强调“人的本质不是单个人所固有的抽象物，在其现实性上，它是一切社会关系的总和”[②]。因此，关于对“人”的理解，要超越西方社会“自然人—理性人”的理解模式，从而超越西方社会对人工智能“以人为本”伦理准则的理解。在马克思看来，近代西方文化神圣的“自然人”的理解模式在方法论上存在一个错误：不是从人的现实的社会实践活动和经济关系出发，而是从人的“自然属性”或所谓的“天赋人性”出发去探究社会道德关系。[③] 同理，从“理性人”角度理解“人”在方法论上也存在一个重要的缺陷：基于“理性经济人”假设，人理解社会关系的方法论最终会走向基于资本逻辑的方法论。因此，马克思对“人”的理解不是从天国降到人间，而是从（生物学）“自然人”向“抽象人”的超越，始于现实的人——“具体人”，经过分裂的社会阶级的人——“社会人”，止于理想社会环境中的人——“抽象人”，形成一种“人”之四象的理解模式。在马克思的理解中，“具体人”并不是近代西方的“自然人”或“理性人”的衍生，而是“现实的人”，即“从事实际活动的人”，或者是资本主义生产方式下的“劳动者”与“资本家”。第一次工业革命期间，由于机器崛起强势介入“人与人”之间的关系，“具体人”进一步演化为“社会人”概念。“劳动者”与“资本家”这两个人类群体演化为对立的两大社会阶级——无产阶级和资产阶级。两个社会阶级相互饱含敌意、价值对立和利益冲突，这是机器增强导致社会分裂的后果。因而，马克思的“社会人”概念也是分裂的。为了消除如此不合理的社会分裂现象，以及再次面向全人类增强“以人为本”，马克思提出了一种“抽象人”概念，强调人的解放和自由而全面发展。这样的“人”的概念不是现实生活中的“具体人”，而是在理想主义社会环境中的自由而全面发展的人。总地来说，近代西方文化对“人”的理解是一种对“人与人”关系的理解，而马克思对“人”的理解模式是一种“人与机”关系。近代西方文化视域的焦点在于“抽象人”，即作为神圣的“自然人”或“理性人”拥有某种神圣属性；马克思视域的焦点在于“具体人”，即在机器大规模社会应用的背景下“现实的人及其境遇”。同理，在当前人工智能大规模社会应用的背景下，“现实的人”不能仅仅发展人与机器的关系以实现资本增殖和扩大再生产，还要发展人与人、人与机、机与机之间的关系实现所有人的全面发展。那么，人工智能“以人为本”伦理准则及其实践就需要具备伦理能力的人工智能，关注“现实的人及其境遇”，既能发挥掌控人工智能的人类群体的个人追求，又有利于公众共享人工智能发展带来的机遇和福利。

最后，根据辩证发展思想，人工智能“以人为本”伦理准则及其实践需要合情合理地使用人性特点，但不能使之成为人性漏洞。马克思已经证明，当资本介入生产过程，基于“理性经济人”思维的“资本逻辑”逐渐掌控全局，“机器异化”或“机器的资本主义应用”使得机器成为一种“生产剩余价值的手段”[④]。由此推理，在“理性经济人”思维背景下，人工智能也将进入“异化”或“机器的资本主义应用”的状态。

① 潘恩荣、曹先瑞：《面向未来工程教育的人工智能伦理谱系》。

② 《马克思恩格斯文集》（第 1 卷），北京：人民出版社，2009 年，第 505 页。

③ 余达淮：《历史唯物主义：马克思主义伦理学的根本方法》，《光明日报》2018 年 10 月 29 日。

④ 《马克思恩格斯全集》（第 23 卷），北京：人民出版社，1972 年，第 408 页。

那么，关于人工智能“以人为本”伦理准则是否可能的问题，主要考察“超人”的人工智能是否处于异化或资本主义应用的状态。

（二）弱人工智能的异化

弱人工智能的异化或资本主义应用与马克思理解的机器异化或机器的资本主义应用是同质的。“异化”（Entfremdung）的原意是“疏远、脱离”，在机器大工业过程中有四种具体表现形式：人与劳动产品相异化、人与劳动相异化、人与人的类本质相异化、人与人相异化。[①] 由于人工智能在机器性能指标上有着明显的提升，人工智能的“异化”程度更深，实践“以人为本”伦理准则的难度提高。但是，只要人工智能尚未诞生独立的自我意识，即出现强人工智能，人工智能的“异化”仍然只是因“疏远而脱离”，而不是因“独立而脱离”人类的最终控制权。

第一，人工智能加深人与自身的劳动产品相异化的程度。传统非智能机器的异化或资本主义应用使得人与自身的劳动产品相异化，工人生产的产品不归工人（机器的使用者）占有，被资本家（机器的所有者）无偿占有。人工智能进一步从体力和智力方面辅助或取代人，同时加剧了这种无偿占有。人使用人工智能生产的产品不归使用者占有，也不归机器本身占有，而归机器的所有者占有。例如，腾讯机器人 Dreamwriter 自动撰写的文章可作为“法人作品”受著作权保护，但著作权归人工智能所有者（法人）占有，而不是使用或设计人工智能的人所有，或者所有人共同所有。[②]

第二，人工智能加深人与劳动相异化的程度。在传统机器异化状态下，劳动成为一种外加在人身上的“体力型”强制手段，人在劳动中不是获得作为“人”的肯定，而是否定。人工智能加速了生产行为，远远超出了“具体人”（机器使用者）的生产速度，甚至超过了“具体人”的理解范围，产生“可解释性”难题。于是，劳动成为一种外加在人身上的“智力型”或“智力型 + 体力型”的强制手段，“具体人”越来越感觉不到劳动属于他自己。人工智能不仅大规模进行物质生产，还开始进行精神生产，如作曲、作画、作诗、作视频等。即便是人工智能在创造性方面没有超过人类艺术家，但是前者“模仿”的水平越来越接近后者。与第一次工业革命机器生产的规模效应一样，人工智能创作也将使物质产品和精神产品越来越价廉物美。那么，对于劳动者而言，他们已经无须、无理由进行生产，包括物质生产与精神生产。

第三，人工智能加剧人与自身的类本质相异化的程度。“有意识的劳动”是人的生命活动的标志之一，也是人区别动物的关键特征之一。但是，当前的人工智能有了一定的“自主决策空间”。在人工智能增强人类劳动能力的情况下，部分劳动成为人们“有意识的劳动”之外在的“黑箱”。人们只看到输入与输出，却不知道“黑箱”中的劳动为何以及如何的过程，引发了人工智能“可解释性”难题。因此，由于人类劳动出现“黑箱化”状态，人与动物的区别、人与机器的区别被模糊化了。

第四，人工智能加剧人与人相异化的程度。当人工智能加剧人与劳动产品、劳动和自身的类本质相异化，按照马克思的逻辑，叠加结果就是加剧“人与人”之间的异化。极端情况下，绝大部分“具体人”事实上成为人工智能生产过程中的“无用者”。在传统私有制社会中，这些无用者或被社会抛弃；在高福利社会或未来机器人社会，这些无用者或被社会“保护”。这种极端情况放大到整个社会层面，就会出现赫拉利口中的“无用的阶级”[③]。

机器异化的原因之一是近代西方的“自由竞争”理念和社会环境。近代西方的“自然人”概念中，“自由”是基本人性之一，因而“自由竞争”也是人性之一。马克思恩格斯的研究表明，“自由竞争”是机器大工业发展的必要条件之一[④]，使得资本主义生产的内在规律强制资本家按资本逻辑行为[⑤]，如以

① 《马克思恩格斯文集》（第 1 卷），第 163 页。

② 广东省深圳市南山区人民法院一审审结原告深圳市腾讯计算机系统有限公司诉被告上海盈某科技有限公司侵害著作权及不正当竞争纠纷一案，参见 https://www.sohu.com/a/379897207_120054912。

③ [以色列] 尤瓦尔 • 赫拉利：《未来简史：从智人到神人》，第 288 页。

④ 《马克思恩格斯文集》（第 1 卷），第 680—681 页。

⑤ 《马克思恩格斯文集》（第 5 卷），第 312 页。

机器生产替代手工生产或者不停地对机器进行更新换代。因此，无论是“自由竞争”还是“资本逻辑”，人工智能“以人为本”伦理准则及其实践需要“辩证”使用、敏捷治理，既要发挥它们的特性扩大再生产、解放和发展生产力，又要严格预防它们失控引发垄断和资本无序扩张等社会问题。

在“辩证”使用人工智能的基础上，马克思“重构个人所有制”的思路有望从根本上解决人工智能的异化。马克思指出，在自由竞争的背景下持续推动科学技术的大规模使用，社会化大生产与私人所有制之间的张力越来越大，以至于最终导致资本主义体系被炸毁。[①] 当前的弱人工智能是第四次工业革命的通用技术，可以对所有的科学技术和产业进行进一步赋能，进而引发马克思所言的“张力越来越大的问题”。关于人工智能伦理及相应准则的热烈讨论本质上也是为了应对“张力”越来越大的一种现象。鉴于以往工业革命过程中机器异化或资本主义应用引发社会不良结果的历史事实，以及马克思提供的关于技术进步与制度变迁的“可解释”理论，西方社会的人工智能“以人为本”伦理准则事实上已经吸纳历史教训和马克思主义思想。例如，前面提到的十一类人工智能伦理关键词中，“为人服务”和“隐私”主要是近代西方“自然人”和“理性人”思想的体现，可持续性、合作、共享、公平、问责都是避免或反对人工智能带来的社会福利被私有制占有的结果。从马克思主义视角看，后面这些人工智能伦理准则都有针对“人性漏洞”的意图，体现了一定的“为人民服务”或“为全人类服务”的主张，而不是单纯的“为少数人服务”。在西方社会背景下，社会化大生产与私人所有制之间的张力可以缓解但无法根除。但是，在中国式现代化的背景下，社会主义现代化有望通过“重构个人所有制”匹配社会大生产，从根本制度上充分调动人性特点但不会成为人性漏洞。

五、小结与展望

西方社会人工智能“以人为本”伦理准则的两个局限，事实上，也是当前人工智能发展的两个障碍。《资本论》关于工业革命和机器大工业批判的思想为破解上述两个局限提供了新的线索（见表 1）。关于“强人工智能”的担忧，根据历史唯物主义的原理，即社会存在决定社会意识，人工智能“以人为本”伦理准则及其实践的前提是“弱人工智能”而非“强人工智能”。关于是否需要具备伦理能力的人工智能的顾虑，根据人的全面发展思想，当前的人工智能发展及人工智能“以人为本”伦理准则都欢迎具备伦理能力的人工智能，以便更好地关注和应对“现实的人及其境遇”。关于“人性漏洞”的问题，这主要是自由竞争和资本逻辑极端化使用可能产生的重大社会问题。在西方社会私有制背景下，完全堵住“人性漏洞”是不可能。基于辩证思维审视“人性漏洞”，治标的方式是合情合理地驱使人性特点发明和使用人工智能，寻找最大社会可接受的人工智能，最大程度实现“以人为本”的人工智能；治本的方式是人类社会仍然需要“重构个人所有制”，从根本上消除人工智能社会大生产与个人私有制之间的张力，彻底实现“以人为本”的人工智能。

表 1　西方人工智能“以人为本”伦理准则的局限与应对线索

西方“以人为本”的局限	马克思的批判对象	对象的描述	“以人为本”新线索
逻辑性悖论	机器问题	强人工智能	坚持历史唯物主义原理，澄清当前社会存在的是弱人工智能而不是强人工智能
		具备伦理能力的人工智能	坚持人的全面发展思想，尤其关注“现实的人及其境遇”
人性漏洞	机器的资本主义应用问题	自由竞争和资本逻辑的极端化使用	坚持辩证思维发明和使用人工智能，既发挥自由竞争和资本逻辑的正面效应，又克制相关的负面效应

如果未来人工智能“进化”成为“强人工智能”，人工智能的自我意识得以诞生，西方社会的人工智能“以人为本”伦理准则将失效。虽然马克思口中的“机器”概念只能覆盖弱人工智能，但是有一些重要论述

① 《马克思恩格斯文集》（第 5 卷），第 874 页。

仍然可以延伸分析强人工智能。

强人工智能是一种“隐喻”，指的是人工智能发展到接近甚至超过人的能力的状态。从近代西方“人”的理解模式看，强人工智能已经成为一种标准的“理性人”。那么，机器异化现象则可能出现“突变”。马克思认为，“人正因为是有意识的存在物，才把自己的生命活动，自己的本质变成仅仅维持自己生存的手段”[①]。以此推论，一旦人工智能成为“有独立自我意识的存在物”，强人工智能如同人一样首先要维持“自己”的生存和发展。此时，人工智能机器的“自主性”和人类社会的“资本逻辑”集成，可能涌现出一种强人工智能和机器人社会的“自我增殖”逻辑。在第一次工业革命期间，基于“理性经济人”思维，资本家成为“资本人格化代表”，奉行“资本逻辑”，推动非智能机器异化或资本主义应用，最终导致社会分裂为资产阶级和无产阶级。那么，未来强人工智能的“自我增殖”逻辑可能催生“机器人”成为新的“资本人格化代表”，导致未来社会全新的机器异化或机器的资本主义应用，造成社会分裂为新的有用阶级和无用阶级。一种社会分裂形态是机器人资本家与人类资本家共存，如英国式现代化是传统贵族与新兴资产阶级共享权力。社会分裂状况类似赫拉利的描述：有用阶级是具有“独立自我意识”的人工智能机器与极少部分人类精英的集合；无用阶级是绝大多数的普通人类，可能还要加上其他无自我意识的机器。另一种社会分裂形态——机器人新贵社会，则是彻底“主奴颠倒”的社会形态。机器人资本家独掌权力，彻底排挤人类资本家直至“破产”而归入无用阶级。社会分类状况类似霍金的描述，有用阶级是“独立自我意识”的强人工智能机器人，无用阶级是人类，也许还包括其他无自我意识的机器人或“破产的机器人资本家”。此时，机器人阶级与人类阶级进入更深刻的种族竞争状态。人工智能机器是否增强 “以人为本”伦理准则超出人类的认知范围。也许那个时候的人类需要通过马克思主义意义上的“革命”来保障“以人为本”，如《黑客帝国》和《终结者》描绘的场景。

Reflections on the Human-centered Ethical Code of Artificial Intelligence: An Analytical Path of Marxist Philosophy

PAN En-rong, SUN Zong-ling

(School of Marxism, Zhejiang University, Hangzhou, 310058)

Abstract: In view of possible negative effects of artificial intelligence (AI), the “human-centered” ethical code of AI proposed by Western society has in fact absorbed historical lessons and Marxist thoughts, but there are still two limitations, i.e. logical paradox and human deficiency. The former points to the problem of AI (machine) itself. Western society has tried to regulate AI with ethical norms, but requires that AI cannot have independent self-consciousness or the ethical ability to distinguish good and evil. The latter points to the problem of “capitalist application of AI”, but is difficult to plug the loophole in the context of Western society. Marx’s researches on the Industrial Revolution and the critique of machine industry provides a new “human-centered” path for the ethical codes of AI.

Keywords: Artificial Intelligence, Human-centered, Logical Paradox, Human Deficiency, Criticism of Machine Industry

[责任编辑：谢雨佟]

① 《马克思恩格斯文集》（第 1 卷），第 162 页。