Advances in Philosophy 哲学进展, 2026, 15(9), 17-24
Published Online September 2026 in Hans. https://www.hanspub.org/journal/acpp
https://doi.org/10.12677/acpp.2026.159414

Hans汉斯

从价值对齐到智能正义：生成式人工智能价值风险的哲学审视

布尔兰·阿曼角勒

新疆大学马克思主义学院，新疆 乌鲁木齐

收稿日期：2026年7月27日；录用日期：2026年8月18日；发布日期：2026年8月28日

摘　要

生成式人工智能的价值对齐方案试图通过技术手段确保AI系统与人类价值观保持一致，但方案悬置了一个根本问题，即对齐何种价值观，又以何种程序实现对齐。从批判理论视角审视，当前主流价值对齐方案并非中立性的技术完善，而是技术资本主义语境下的资本逻辑主导下，对马克思所揭示的“一般智力”进行的价值嵌入，它将特殊的、服务于特定群体利益的价值取向伪装成普遍适用的人类共同价值，由此生成的价值风险具有三重维度。认识论层面的社会认知结构性偏差；价值论层面的技术理性对人文价值的排挤；存在论层面的人的主体性风险。若要超越这一困境，就需要从价值对齐转向智能正义。因而智能正义围绕AI设计、部署与收益分配，建立能够抵抗资本与权力扭曲的价值批判和制度反思框架。所以智能正义不是对价值对齐的简单否定，而是对其深层权力关系的揭示、批判与重构。

关键词

生成式人工智能，价值对齐，价值风险，智能正义，一般智力

From Value Alignment to Intelligent Justice: A Philosophical Examination of the Value Risks of Generative Artificial Intelligence

Buerlan Amanjiaole

School of Marxism, Xinjiang University, Urumqi Xinjiang

Received: July 27, 2026; accepted: August 18, 2026; published: August 28, 2026

Abstract

The value alignment scheme of generative artificial intelligence attempts to ensure that AI systems

文章引用: 布尔兰·阿曼角勒. 从价值对齐到智能正义: 生成式人工智能价值风险的哲学审视[J]. 哲学进展, 2026, 15(9): 17-24. DOI: 10.12677/acpp.2026.159414

align with human values through technological means, but it leaves a fundamental question unresolved: which values to align with, and through what procedures to achieve alignment. From the perspective of critical theory, the current mainstream value alignment schemes are not neutral technological improvements but rather the embedding of values under the dominance of techno-capitalist logic into what Marx revealed as general intelligence. They disguise the particular and value-oriented interests serving specific groups as universal and common human values, thereby generating value-related risks with three dimensions: social cognitive structural deviations at the epistemological level; the exclusion of human values by technical rationality at the axiological level; and the risk to human subjectivity at the ontological level. To overcome this predicament, it is necessary to shift from value alignment to intelligent justice. Intelligent justice focuses on the design, deployment, and revenue distribution of AI, establishing a framework of value criticism and institutional reflection that can resist the distortion of capital and power. Intelligent justice is not a simple negation of value alignment but rather the revelation, criticism, and reconstruction of its deep power relations.

Keywords

Generative Artificial Intelligence, Value Alignment, Value Risk, Intelligent Justice, General Intelligence

Copyright © 2026 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

当前，生成式人工智能(Generative AI)正在以指数级速度发展的同时，也在重塑知识生产、文化传播与社会交往的基本形态。ChatGPT 大语言模型、Sora 多模态生成模型的发展展现出了惊人能力，也引发了一系列深层忧虑，如模型可能输出种族主义或性别歧视言论，可能基于训练数据中的偏见而强化社会刻板印象，在涉及政治、历史等议题时呈现出系统性的价值倾斜。作为回应，技术界提出了“价值对齐”(value alignment)方案，试图通过算法设计确保 AI 系统的目标与人类价值观保持一致。然而，一个更根本的问题被长期悬置，即人类价值观并非同质整体，而是充满多元性与内在矛盾的。一个实现了价值对齐的 AI，完全可能服务于剥削、操纵或系统性歧视。然而这些问题已超出了技术方案的回答能力，指向了哲学领域，成为价值批判理论在数字时代必须回答的核心议题。

对于技术与价值的关系，海德格尔在《技术的追问》中提供了一个关键的哲学坐标：“现代技术的本质在于‘架构’(Ge-stell)。架构阻断了真理的朗现和真理的统治，嬗变为安排的规定，结果成了最大的危险。危险的不是技术。技术并非恶魔；但技术的本质是神秘的，技术的本质作为对揭示的一种规定，就是危险。”[1]海德格尔的洞见在于，技术本身就带有价值框定的性质。以此审视价值对齐方案，便会发现其内部存在着难以调和的张力。该方案试图通过技术手段让人工智能对齐人类价值，但其自身所依赖的可计算性、目标函数化、最优化等技术逻辑，早已先行框定了什么样的价值能够被表征、哪些价值可以被纳入考量。本文正是沿着这一追问，从批判理论视角出发，将价值对齐作为一个问题情境，透过它去诊断生成式人工智能价值风险的深层生成机制，并尝试提出从价值对齐迈向智能正义的理论转向路径。

2. “价值对齐”的哲学阐释：从技术概念到问题情境

价值对齐方案在技术界的流行并非偶然，它直接回应了生成式 AI 在价值倾向上引发的普遍忧虑。然

而，技术方案的有效性不能替代对其前提的哲学审视。本章首先澄清价值对齐的技术内涵及其隐含的理论预设，进而引入马克思“一般智力”的分析框架，揭示价值对齐方案在技术资本主义语境下的资本逻辑主导下的价值编码本质。这一分析将为后续三重风险的具体展开奠定批判性的理论基础。

2.1. 价值对齐的技术内涵与哲学预设

价值对齐方案隐含着三个需要被哲学审视的理论预设。第一，存在一套普遍且可编码的人类价值观。无论是 RLHF 中标注人员的偏好排序，还是宪法人工智能(Constitutional AI)中预设的行为准则，都假定人类价值可以被提炼为一组相对稳定、跨文化可通约的规范集合。第二，技术手段能够客观地将这些价值观嵌入人工智能系统。这一预设相信，算法优化是中性的，只要目标函数设定得当，价值问题就可以还原为技术问题。第三，价值对齐是解决人工智能社会风险的必要条件。它假定只要 AI 按照设定的价值观行事，风险就得到了管控。但这三个预设在可行性方面都存在重重漏洞。

第一个预设面临着双重困难。一方面，人类价值观并非同质整体。不同文化传统、阶级立场、政治制度下，对于什么是好的以及正当的，往往有着不同的甚至相互冲突的理解。即便在同一社会内部，不同群体对同一议题的价值判断也可能截然对立。价值对齐方案所预设的普遍价值观，实际上是一种理论虚构，它要么悬置了价值争议，要么暗中偏向了某一特定群体的价值立场。另一方面，可编码的要求本身就是一个过滤器。那些无法被量化、无法被形式化为奖励函数的价值，如尊严、本真性、批判性等，在技术方案中往往被系统性地排除或简化。由此可见，价值对齐所对齐的是进行技术可行性筛选后的特定价值秩序。

第二个预设相信算法优化是中性的，但这一说法需要进一步证明。训练数据的来源框定模型学习结构，将标注人员的价值偏好渗透其中，因此奖励函数决定着对行为的强化和抑制。从数据采集到标注规则再到奖励函数设计，人工智能研发的每一环节都渗透着价值选择和价值判断。价值对齐方案将这些选择包装为技术优化的结果，遮蔽了其背后的权力运作。换言之，客观嵌入的本质就是一种价值编码策略，它将本应是公共讨论的价值问题，转移到了技术专家的专业领域。

第三个预设是将对齐等同于安全乃至正义。一个能够完美遵循所有者指令的 AI，完全可能服务于剥削、操纵或系统性歧视。价值对齐只能保证 AI 不做所有者不允许的事，却无法保证所有者本身的价值取向是正义的。更深层的问题在于，价值对齐方案将 AI 风险简化为 AI 不服从指令的技术问题，而忽略了更根本的风险来源，即掌握 AI 权力的所有者本身可能就是不正义的。在这个意义上，价值对齐最多是解决 AI 风险的必要条件，远非充分条件。正如吴静所批评的，旨在提供整体性方案的静态价值对齐范式，根源在于其将对齐目标误设为对价值函数的逼近，试图将多元的价值观叙事压缩为单一的偏好排序关系[2]。这种将价值抽象为数学关系的路径，正是技术理性对精确性追求在价值领域的越界，也恰恰印证了本文前述对价值对齐方案价值编码本质的判断。

2.2. 马克思“一般智力”视域下的价值对齐批判

马克思在《1857~1858 年经济学手稿》中指出，“一般智力”指的是物化在固定资本中的科学知识与社会知识的总体，它本是人类社会共同积累的成果，但在资本主义条件下被资本占有，反过来成为支配活劳动的力量[3]。生成式人工智能正是“一般智力”在当代的典型形态。大语言模型虽将人类数千年积累的知识、语言与文化规范编码为参数化算法，成为客观化的智力结构，但却被少数科技巨头占有并服务于资本增值。在这一框架下，价值对齐的实质不再是将普遍的人类价值注入 AI，而是将经过技术资本主义逻辑筛选后的特定价值秩序编码为 AI 的行为准则。

生成式 AI 的生产过程体现了“一般智力”的资本占有逻辑。一方面，训练大模型所需的海量数据，

很大程度上来自全球用户的数字劳动，如社交媒体发帖、搜索记录、知识贡献、对话交互等，这些劳动大多未被合理补偿，用户成为免费或低成本的数据原料提供者。另一方面，大模型本身作为私有财产，其参数、架构、训练细节被科技巨头视为商业机密，拒绝公开。这意味着，人类共同积累的知识成果，被转化为少数企业的排他性资产。由此，用户在使用 AI 系统时的每一次交互，又成为进一步训练的数据来源，形成“数据提取–模型训练–价值实现–再提取”的闭环。在这一结构中，“一般智力”不仅被资本占有，而且被转化为持续自我强化的价值增值机器。

因此，在这一占有结构下，价值对齐扮演了关键的角色。它表面上是让 AI 与人类价值观对齐，实际上是将符合资本增值需要的价值秩序编码为 AI 的行为准则。具体来说，第一，训练数据中已经内在地偏向于占据主导地位的群体，如英语语料的主导地位意味着非英语文化的价值表达被系统性边缘化；消费主义文化的数据积累意味着批判消费逻辑的价值被淹没；技术精英阶层的内容生产意味着底层视角的缺失。第二，RLHF 等对齐技术依赖标注人员的价值判断，而这些标注劳动本身处于资本控制的劳动分工体系中，标注规则由企业制定，标注人员按照给定的框架进行判断，其价值选择的自主性被严格限定。第三，在商业实践中，模型输出的“有用”“无害”“诚实”等内容被理解为服务于产品可用性与商业可持续性，而非服务于人的主体性或社会正义。

因而，在此意义上，价值对齐的价值编码功能才得以显现：它将特殊且局部地服务于特定群体利益的价值安排，伪装成普遍的、中立的、人类的价值共识。用户在使用 AI 时，面对的是一个被预先编码了价值秩序的“一般智力”系统，其输出被误认为中立的智能表达，从而忽略了其背后资本逻辑的价值投射。这正是价值自然化的典型运作方式，不是通过强制，而是通过将特殊利益自然化、普遍化，来遮蔽其背后的权力关系。马克思对商品拜物教的分析揭示了一个类似的机制：商品交换关系被误认为物与物的自然关系，而非人与人的社会关系[4]。在生成式 AI 时代，一种价值的技术神秘化正在形成，AI 在资本逻辑下输出的价值编码被误认为纯天然的技术产物。

基于上述批判性分析，我们得出价值对齐方案只能解决如何让系统表现得更符合预期，却没有涉及更根本的问题，即谁有权设定预期并确保这一设定过程是正义的。在此意义上，价值对齐方案的局限在于它以解决方案的姿态遮蔽了更根本的问题。它致力于将价值观编码为算法目标，却没有反思这一编码过程本身由谁控制、服务于谁。正是这一局限，决定了本文对价值对齐的定位：它不是本文试图解决的问题的答案，而是一个需要被诊断的问题情境。换言之，价值对齐方案的提出与流行本身，就是一个值得分析的症候。

3. 生成式人工智能价值风险的生成机制

前文的分析表明，价值对齐方案并非中立的治理技术，而是技术资本主义语境下的资本逻辑对“一般智力”进行价值编码的装置。基于此，本文从认识论、价值论与存在论三个递进的哲学层次，系统考察生成式人工智能价值风险的生成机制。三重维度并非并列关系，而是层层深入：认知偏差构成了风险的表面效应，价值排挤揭示了其深层动力，而主体性风险则指向这一风险最终对人的自由本质的侵蚀。

3.1. 认识论风险：算法偏见导致的社会认知偏差

生成式人工智能的第一重价值风险发生在认识论层面，它通过算法偏见导致社会认知结构发生扭曲。所谓认识论风险，指的是 AI 系统在知识生产与传播过程中，输出偏向于特定群体的认知框架，从而导致用户的认知偏差。当用户依赖 AI 获取信息、理解世界时，他们实际上是在接触一个已经被预先编码了特定价值秩序的知识体系。这种认知扭曲的风险之所以是价值性的，不在于它有意欺骗，而在于它以客观中立的技术面孔，将特殊的认知框架自然化为普遍的常识。

这一风险的具体生成机制可以从三个环节加以分析。第一，具有代表性的训练数据自身带有偏差。大语言模型的训练数据主要来自互联网公开文本，而互联网内容在语言、地域、阶层、文化上存在严重的不均衡，即英语内容占据绝对主导地位，西方发达国家的知识生产被过度代表，而非英语、非西方、边缘群体的声音被系统边缘化。这意味着，模型所学习的世界图景，从一开始就是有偏差的。第二，算法优化的目标导向。当前的 AI 系统以用户参与度响应准确性等可量化指标为优化目标，而非真实性、公正性等难以量化的价值。这导致模型倾向于生成用户想听的内容，忽略了内容与事实的一致性，从而进一步强化了既有偏见。第三，反馈循环的自我强化。当 AI 输出带有偏见的回答并被用户接受或确认时，这一交互又被纳入训练数据，进一步固化原有的偏见模式。偏见不是偶然的故障，是系统自我强化的常态。

在传统社会中，知识生产需要经过学术共同体、出版审核、公共辩论等相对透明的程序；在生成式 AI 时代，知识生产被封装在算法的黑箱之中，少数科技公司掌握了定义什么是常识的权力。用户面对 AI 的输出，往往缺乏批判性思维，容易将算法建构误认为客观事实。当一种认知偏见被普遍接受为常识时，它就不再需要辩护，也不再被质疑，最终形成了认识论层面的价值风险。

3.2. 价值论风险：技术逻辑对人文价值的排挤

生成式人工智能的第二重价值风险发生在价值论层面，也就是说技术逻辑对多元人文价值构成排挤。所谓价值论风险，指的是 AI 系统的运行遵循工具理性逻辑，追求可计算、可优化、可量化的目标，而公平、尊严、真实性、批判性等人文价值难以被编码为算法目标，因而被边缘化甚至排除。这种排挤是技术逻辑本身的结构带来的倾向。当越来越多的社会领域被 AI 系统所介入，人文价值的生存空间就会被压缩。

价值对齐的技术方案需要将人类价值转化为可计算的奖励函数或可标注的偏好数据。然而事实却是人文价值是多元的、动态的，甚至是内在矛盾的，这就表明尊严无法被量化为分数，公平无法被简化为公式，批判性思维难以被编码为优化目标。为了适应技术方案的可操作性要求，这些价值被“降维”处理，要么被简化为某种可测量的代理指标，如公平被简化为统计上的群体均等，要么被完全排除在优化目标之外。其结果就是多元价值被简单化，只有那些能够被技术框架所容纳的价值才获得承认，其他维度的价值被忽视。

当一种价值无法被量化、无法被编码或无法被优化时，它就会被系统视为不切实际、模糊不清或难以操作的，最终在决策中失去分量。这正是工具理性压倒价值理性的结果。技术方案对价值的简化处理不仅表现为筛选与排除，还表现为更深层的悖论：有研究在批判 Anthropic 公司自 2022 年起推行的宪法人工智能(Constitutional AI)等对齐实验时指出，这种道德工程化尝试以他律手段催生自律主体，却在逻辑上瓦解了自律本身的可能性[5]。技术方案对价值的简化处理往往产生价值漂移效应。以公平为例，许多 AI 公平性研究将其量化为不同群体间的统计均等，但这种简化忽略了公平的多元内涵，程序正义、分配正义和承认正义这些维度在量化过程中被系统性牺牲。商业 AI 系统的优化目标天然偏向用户参与度、点击率和留存时间等可量化指标，而非真实性、公正性等难以度量的价值。当模型被训练为最大化用户参与度时，它倾向于生成更具情绪煽动性或认知确认偏误的内容，而那些需要沉思、批判或自我否定的价值维度则在优化过程中被边缘化。由此形成一种结构性的价值筛选机制，技术逻辑不仅翻译价值，更在翻译过程中重塑乃至扭曲价值本身。人文价值在 AI 系统中的边缘化并非出于任何人的主观恶意，而是技术框架的运作方式本身所导致的系统性后果。

3.3. 存在论风险：人机关系重构中的主体性风险

生成式人工智能的第三重价值风险发生在存在论层面，即人机关系的重构有可能诱发人的主体性风

险。所谓存在论风险，指的是 AI 系统可能在改变我们知道什么或重视什么，乃至改变我们如何思考以及什么是人。当我们把知识生成、创意表达、甚至情感陪伴都外包给 AI 时，人类的独立思考能力、判断能力、创造能力有可能面临退化的风险。用户可能从思考者逐渐变为 AI 输出的确认者或编辑者，人的主体性地位可能在不知不觉中被削弱。需要指出的是，上述关于主体性风险的判断目前主要基于理论推演，尚需经验研究的进一步验证。

这一风险的生成机制可以从三个维度加以分析。第一，认知劳动的转移。人的写作、论证和创作原本就被视为需要投入心智的活动，这一过程本身就具有认知训练的价值。当 AI 可以一键生成文章、代码、图像时，用户倾向于选择最省力的路径，接受 AI 输出而非自己思考。长此以往，批判性阅读、逻辑论证、创意构思等认知能力有可能遭受用进废退。第二，判断力的外包。面对 AI 输出的内容，用户需要判断其真实性、相关性和价值倾向，这一判断过程本身是主体性的核心体现。然而，当 AI 系统越来越智能，用户越来越倾向于信任 AI 的判断而非自己的判断，此时，判断力被逐渐外包给算法。第三，创造性的窄化。生成式 AI 的创意本质上是训练数据的重组与插值，而非真正的突破性创新。当用户习惯于 AI 生成的标准答案，创造性思维的边界就可能被算法所预设的框架所限定，失去了被人类的好奇心所拓展的可能性。

马克思在《资本论》中分析了资本主义生产如何通过劳动分工和机器体系，“这种自动机是由许多机械器官和智能器官组成的，因此，工人自己只是被当作自动的机器体系的有意识的肢体”[6]。在生成式 AI 时代，一种新型的认知分工可能正在形成，AI 承担了思考的功能，人类退居到确认的角色。当用户不再具备独立思考和批判性判断的能力时，他们就可能更容易被算法所引导、被内容所操控、被广告所捕获。这正是马尔库塞在《单向度的人》中所警示的“技术的解放力量——使事物工具化——转而成为解放的桎梏”[7]——人们不被强迫接受某种价值立场，而是在便捷、高效、智能的技术体验中，主动放弃了批判性思考，有可能丧失否定性的向度。

4. 从价值对齐到智能正义：一种理论转向

对生成式 AI 三重价值风险的诊断表明，价值对齐方案所暴露的问题已超出技术修补的能力边界。认知偏差、价值排挤与主体性风险并非彼此孤立的故障，而是同一深层权力结构的三种表征。若不能触及这一结构，任何技术层面的完善都只是在技术资本主义语境下的资本逻辑的既定轨道上做无害化处理。因此，本章提出从价值对齐到智能正义的理论转向，尝试建构一个能够回应上述三重风险的批判性框架。智能正义不是对技术方案的简单否定，而是对价值对齐所遮蔽的权力关系进行揭示、批判与重构。

4.1. 何谓智能正义

“智能正义”是本文尝试建构的核心概念，用以指称超越“价值对齐”技术方案的价值批判框架。所谓智能正义，是指围绕人工智能的设计、部署、使用与收益分配，建立能够抵抗资本与权力扭曲、保障多元主体平等参与、并服务于人的主体性的制度与价值安排。这一概念立足于批判理论传统——正义不能脱离生产方式，真正的正义不在于公平分配既有利益，而在于消除产生不正义的社会关系。

智能正义可以从三个层次加以理解。程序正义要求算法系统的设计、训练、部署与评估过程应当透明、可问责、并允许受影响群体的参与，AI 发展过程中出现的问题不能在技术黑箱内部由专家独自决定，应当进行公共讨论。分配正义要求智能系统应致力于消除算法歧视、弥合数字鸿沟、并确保 AI 收益的合理分配，当大模型的训练依赖全球用户的数字劳动而模型本身却是少数企业的私有财产时，分配正义的要求就是对占有的深层批判。生产关系正义则要求追问 AI 的生产资料应当由谁掌控，是少数科技巨头还是作为一种社会公器受到公共监督。

这一界定意味着，智能正义不是对价值对齐的简单否定，而是对其深层权力关系的揭示与重构。在

这个意义上，价值对齐致力于将价值观编码为算法目标，智能正义则关注这一编码过程本身的合法性与公正性。

4.2. 从价值对齐到智能正义：理论演进的逻辑

本文以“从价值对齐到智能正义”为题，暗示了一种理论演进或问题深化的路径。这一“从 A 到 B”的逻辑，不是 A 被 B 取代，而是 A 在 B 的审视下显露出自身的限度，从而需要被纳入一个更根本的问题的框架。因而，价值对齐与智能正义之间的关系，可以从以下三个维度加以理解。

第一，从技术修补到结构批判。价值对齐是技术方案内部的修补，但它始终在既定的技术框架和技术资本主义语境下的资本逻辑内部打转，而智能正义则跳出这一框架，追问技术方案得以运作的前提条件。第二，从价值编码到权力追问。价值对齐的核心操作是将人类价值观编码为可计算的算法目标，智能正义则悬置这些前提，转而追问编码行为本身的政治性。第三，从个体伦理到社会正义。价值对齐聚焦于个体层面的伦理问题，智能正义则将视野从个体伦理提升到社会正义。

需要指出的是，智能正义的理论框架并非排斥技术层面的价值对齐实践。以 Anthropic 公司的宪法人工智能方案为例，其“可扩展监督”(scalable oversight)等技术路径试图通过 AI 自我监督来替代大规模人类反馈[8]。Anthropic 随后还开展了“集体宪法 AI”(Collective Constitutional AI)实验，通过 Polis 平台邀请约 1000 名美国公众参与起草 AI 宪法，探索民主过程如何影响 AI 的价值对齐[9]。OpenAI 等机构也在探索“民主化对齐”(democratic alignment)的技术方案，其“民主化输入 AI”(Democratic Inputs to AI)资助计划资助了 10 支团队设计利用民主方法决定 AI 系统规则的方案，试图通过大规模公众参与来确定 AI 应当遵循的价值原则[10]。这些技术探索与智能正义所倡导的程序正义在方向上具有潜在的兼容性：程序正义所要求的多元主体参与，可以通过改进可扩展监督的机制设计，将更多元的价值立场纳入训练数据的标注和偏好排序环节，从而使对齐过程本身更具公共性和可问责性。然而，技术路径的改进并不能自动解决编码标准本身的权力归属问题——多元价值一旦进入技术框架，仍然面临“被谁筛选、按何种权重组合”的选择。这正是智能正义的理论框架对技术实践所提供的批判性反思：它要求我们将技术选择本身视为一种政治选择，而非纯粹的技术优化问题。

由此可见，从价值对齐到智能正义的理论转向，最终是要将价值对齐重新放置在一个更根本的社会批判框架之中加以理解。价值对齐本身并不因此被抛弃，它仍然可以作为智能正义框架下的一个技术环节而存在，用以确保 AI 系统不产生直接的恶意输出或明显的危害行为。但它的合法性被严格限定在技术操作的层面，它所依赖的价值前提则被暴露在公共审视与制度批判之下。只有当价值对齐被纳入智能正义所要求的权力追问与制度反思之中，它才不会沦为资本对“一般智力”进行价值嵌入的隐蔽工具，也才不会以技术中立的伪装继续遮蔽其背后的政治性选择。

4.3. 智能正义的理论根基

批判视角下的智能正义概念并非无根之木，它的理论根基深植于批判理论传统。

马克思对“一般智力”的分析提供了智能正义的认识论前提。马克思认为资本主义机器生产所产生的异化的根源就在于资本主义生产资料私有制及其驱动和主导的资本增殖逻辑[6]。它揭示了生成式 AI 是“一般智力”的当代形态，智能正义的首要任务就是打破生成式 AI 背后的垄断性占有，从生产关系层面探讨数据、算力与算法作为社会共享资源的可能性。

马克思的商品拜物教批判为智能正义提供了价值祛魅的方法论工具。马克思揭示，在商品交换社会中，人与人的社会关系被遮蔽为物与物的自然关系，从而使得资本主义的生产关系被视为永恒的自然法则[4]。智能正义的任务就是揭示并祛除这种价值的神秘化，揭示 AI 输出背后的社会关系、权力结构与资

本逻辑，使被自然化的价值秩序重新成为可批判的对象。

马克思对人的自由全面发展的追求为智能正义提供了价值旨归。马克思的批判不是为批判而批判，而是指向“每个人的自由发展是一切人的自由发展的条件”[11]。智能正义必须触及 AI 生产关系的根本变革。这不是要在当下给出具体的制度蓝图，而是要确立一个批判的方向：任何关于 AI 的正义理论，如果不能回应“谁拥有 AI、谁控制 AI 和 AI 为谁服务”这三个根本问题，就失去了批判性维度。

在这个意义上，智能正义是批判理论在数字资本主义时代的一种理论延伸。在新的技术条件下，迫切需要重新激活马克思对资本逻辑、价值自然化与人的主体性的批判性分析。从价值对齐到智能正义，需要从技术方案转向社会批判，完成从技术治理到人的主体性理论的回归。

5. 结语：生成式人工智能价值风险的批判与超越

生成式人工智能的价值对齐方案并非中立的治理工具，而是技术资本主义语境下的资本逻辑借技术优化之名对马克思所揭示的一般智力进行价值编码的当代形态。它将特定的、服务于资本增殖的价值秩序伪装成普遍的人类共识，由此在认识论层面造成社会认知的结构性偏差，在价值论层面以技术理性挤占人文价值，在存在论层面诱发人的主体性风险。三重风险并非彼此孤立，而是资本对智力生产条件之垄断性占有的三种表征。正因如此，停留于技术内部的修修补补已远远不够，必须实现从价值对齐到智能正义的理论转向。

智能正义将批判的锋芒对准价值编码背后的权力关系，从程序正义、分配正义直至生产关系正义，追问谁来设定、为谁服务的根本政治问题。它要求打破少数科技巨头对数据、算力与算法的垄断性支配，将被自然化的价值秩序重新置于公共反思与民主监督之下。需要强调的是，智能正义在今天首先是一道批判性命题而非现成的改革方案，它所指向的生产关系变革在资本逻辑仍深度支配全球技术创新的现实中面临巨大阻力。但批判的力量恰恰在于揭示：任何回避了谁拥有 AI、谁控制 AI、AI 为谁服务这三个根本问题的治理方案，都不过是在既定轨道上做无害化处理。在生成式人工智能日益重构人类认知、价值与生存方式的今天，从技术方案通向社会批判、从价值嵌入回归对人的主体性风险的反思，已成为数字时代不可回避的正义之问。

参考文献

[1] 海德格尔. 追问技术[M]//查常平. 人文艺术: 第 1 辑. 成穷, 译. 贵阳: 贵州人民出版社, 1999: 285.

[2] 吴静. 人工智能价值观的动态适应对齐研究——从意图-价值-情境互构范式到智能正义[J]. 华中科技大学学报(社会科学版), 2025, 39(6): 1-10.

[3] 赵泽林. 《1857-1858 年经济学手稿》与人工智能的三重审思[J]. 南京社会科学, 2024(5): 30-36+48.

[4] 马克思. 资本论: 第 1 卷[M]. 中共中央马克思恩格斯列宁斯大林著作编译局, 译. 北京: 人民出版社, 2018: 88-99.

[5] 郭良婧. 人工智能时代道德他律的悖论性生产——Anthropic 对齐实验的伦理学批判[J]. 电子科技大学学报(社会科学版), 2026, 28(3): 12-20.

[6] 马克思, 恩格斯. 马克思恩格斯文集: 第 8 卷[M]. 北京: 人民出版社, 2009: 184.

[7] 马尔库塞. 单向度的人: 发达工业社会意识形态研究[M]. 刘继, 译. 上海: 上海译文出版, 2006: 135.

[8] Anthropic (2023) Claude’s Constitution. https://www.anthropic.com/news/claudes-constitution

[9] Anthropic (2023) Collective Constitutional AI: Aligning a Language Model with Public Input. https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input

[10] OpenAI (2024) Democratic Inputs to AI Grant Program: Lessons Learned and Implementation Plans. https://openai.com/index/democratic-inputs-to-ai-grant-program-update/

[11] 马克思, 恩格斯. 马克思恩格斯文集: 第 2 卷[M]. 中共中央马克思恩格斯列宁斯大林著作编译局, 译. 北京: 人民出版社, 2009: 53.