基于 AI 基础设施与 MLSys 视角的马克思恩格斯科技观当代价值审视

李宇哲 SA25011049

摘要—作为一名深耕于人工智能基础设施构建与机器学习系统（MLSys）领域的研究生，我的科研工作不仅聚焦于算法效能的提升与工程实现的优化，更深刻地嵌入在算力资源的调度、数据要素的治理以及 AI 平台生态的协作网络之中。当前，大模型技术的爆发式增长使得“技术进步”具象化为生产力的质的飞跃，但与此同时，资源的高度垄断、劳动模式的重构以及随之而来的治理难题也日益凸显。若将上述现象置于马克思恩格斯科学技术观的理论框架下进行考察，将有助于我们跳出“技术决定论”或“技术中性论”的窠臼，从唯物辩证法的高度深刻洞察技术演进、资本运作与社会结构之间的动态张力，从而更好地引导技术服务于人的价值实现与社会的良性发展。

I. 马克思恩格斯科技观的核心内涵及其方法论启示

首先，马克思与恩格斯深刻揭示了科学技术与社会生产力之间的辩证互动。机器体系并非孤立于社会之外的机械装置，而是社会化大生产的有机组成部分。一方面，科学技术正加速转化为“直接生产力”，深度渗透至生产的各个环节；另一方面，生产关系的变革与分工的细化又反向驱动了技术创新的需求。置于当代语境，这意味着技术迭代并非单一学科的线性延展，而是由科研体制、产业集群、基础设施建设及市场逻辑共同编织的系统性工程。

其次，技术从来不是抽象且中立的器具，而是深深嵌入特定生产关系之中的社会存在。马克思在剖析机器大工业时敏锐地指出，机器在提升劳动生产率的同时，若受资本逻辑主导，极易转化为强化劳动控制、压缩必要劳动时间并攫取剩余价值的工具。这一观点为我们提供了重要的方法论：在评估任何一项新技术时，不能仅局限于性能参数的考量，更需剖析其部署背后的制度框架、控制权归属以及利益分配机制。

再次，生产力与生产关系的矛盾运动是推动社会演进的根本动力。技术突破往往会冲破旧有生产方式的藩篱，催生新的组织形态与利益格局；反之，滞后的制度安排亦会阻碍技术潜能的充分释放。辩证法既承认技术进步的客观驱动力，也强调其社会后果取决于矛盾双方的博弈与调整。

最后，人的自由全面发展是评价技术进步的终极尺度。恩格斯在探讨人与自然关系时凸显了人的主体能动性；马克思则在异化劳动理论中警示：当劳动过程及其产品被外在力量所主宰，技术进步可能异化为统治人的“物的力量”。因此，科学技术观不仅关注“技术可行性”，更需追问“技术为谁服务”以及“技术如何促进人的解放”。

II. AI Infra / MLSys：当代“机器体系”的重构与生产力质变

从 AI 基础设施与 MLSys 的专业视角审视，大模型时代的革命性突破不仅在于算法层面或者系统层面的创新，更在于一种新型“机器体系”的建立：算力硬件（GPU/专用加速器 [2]、网络互联、存储系统）、数据要素（采集、清洗、标注、合规审查）、工程架构（分布式训练与推理框架、编译级优化、容错与扩缩容机制）以及协作生态（跨组织研发、开源社区）共同筑就了现代智能生产的物理底座。

1）算力与系统工程：固定资本的当代形态演进在经典工业时代，资本凝结为厂房与机器；而在 AI 产业中，数据中心、异构计算集群、训练 [5] 以及推理侧框架 [4] 则构成了高度资本密集的固定资本形态。MLSys 研究中对吞吐量、延迟、能效比及通信开销的极致追求，本质上旨在提升这一庞大固定资本的周转效率与利用率。诸如张量并行 [5]、流水线并行 [1]、参数切分及混合精度训练等技术路径，实则是在硬件物理极限下对“生产流程”的精细化重组，这既是工程学的胜利，也是生产力形态的迭代。

2）“科学转化为直接生产力”的具象化随着模型规模与工程复杂度的耦合，科研成果向生产系统的转化周期被极度压缩：新型优化器、并行策略、RAG（检索增强生成）、蒸馏及量化技术，能够迅速映射为成本曲线

的优化。这生动印证了马克思的论断：知识不再仅仅是解释世界的理论，而是通过代码、算子库、自动化流水线等形式，直接蜕变为决定生产效能的核心要素。

3）社会化协作与“隐形劳动”的价值大模型的诞生并非个别天才的独舞，而是高度社会化的集体劳动成果：涵盖了算法研究、系统工程、产品设计、数据处理及安全运维等多个维度。特别是数据标注、RLHF（人类反馈强化学习）中的反馈提供等环节，虽常处于产业链边缘，却是保障模型智能与安全的基础。这使得“协作劳动”与“价值分配”间的张力愈发显著：模型产权与收益的归属，与承担重复性劳动及潜在风险的主体之间，构成了当代技术体系中不容忽视的社会问题。而价值工程一概念，也成为清华大学的章明星教授的工作 Mooncake[3] 的独特视角。将大模型推理中的关键要素 KVCache，转变为价值工程中的“价值”，并将其作为系统优化的主要对象，也大力推进力大模型推理系统的发展，并助理 Kimi 产品的落地。

III. 资本逻辑下的技术异化：效率崇拜与结构性困境

在现行生产关系下，技术进步往往服务于效率最大化、市场扩张及风险控制。AI Infra/MLSys 的诸多技术目标（如降本增效、自动化运维）虽与此逻辑契合，但也引发了辩证的负面效应。

1）资源集聚与技术垄断趋势大模型训练对算力、数据及高端人才的极高门槛，导致行业呈现明显的寡头化特征：少数巨头垄断了底层集群与分发渠道，并通过生态壁垒形成技术锁定。从辩证法视角看，这既是生产力发展带来的规模经济效应，也可能固化不平等的生产关系，使得中小创新者因缺乏基础设施支持而难以将技术构想社会化。

2）劳动的替代、重构与潜在异化 AI 系统在替代重复性脑力劳动的同时，也催生了提示词工程、数据治理等新岗位。然而，在实际执行中，劳动过程往往被进一步量化与流程化，即便是研发人员也可能被裹挟进“以算力消耗与迭代速度为核心”的绩效体系，面临新的异化风险：人被迫去适配机器的频率与平台的规则，而非技术服务于人的创造性发展。这与马克思关于机器体系可能加剧分工固化的分析具有高度的现实相关性。

3）技术外部性：能耗挑战与治理成本 MLSys 的优化常聚焦于推理成本的降低与吞吐量的提升，但若缺乏宏观层面的制约，可能陷入“杰文斯悖论”：单位计算成本的下降反而诱发更庞大的使用需求，导致总能耗激增及内容生产过剩。此外，数据版权争议、隐私泄露风险及算法偏见等问题，进一步揭示了技术系统并非价值中立的计算黑盒，而是深度交织着社会权力结构的“制度性技术”。

IV. 解放的维度：在矛盾运动中探寻技术演进路径

马克思恩格斯的科技观并非简单的“反技术主义”，而是主张在剖析技术与生产关系矛盾的同时，寻找通向新社会组织形式的物质基础。结合 AI Infra/MLSys 领域，我们可以从以下维度探索技术的“解放向度”。

1）公共性与普惠性：打破技术壁垒的工程实践开源框架、开放权重模型及标准化 API 的普及，在一定程度上打破了知识垄断，促进了技术的民主化传播；而公共算力平台与科研云的建设，则有助于缓解算力集中对底层创新的压制。对于 MLSys 研究者而言，提升系统效率不应仅服务于资本增值，同样的优化技术亦可用于降低中小实体的训练门槛，推动技术红利的普惠化。

2）以人为本的评价体系：超越单一的“性能崇拜”传统的基准测试（Benchmark）往往过度追求精度与速度，但辩证法提示我们：指标的设定本身即蕴含价值导向。将隐私保护、可解释性、系统安全性、绿色低碳及公平性纳入系统设计的核心指标，实质上是在技术层面回应生产关系的矛盾，将“技术可行性”升华为“社会可接受性”。这不仅是工程伦理的要求，更是技术走向成熟的必由之路。

3）重塑劳动协作关系：正视“隐形劳动”要消解技术体系中的异化现象，必须赋予数据处理、内容审核及安全治理等环节以合理的劳动尊严与权益保障。模型智能的涌现并非源自算法的神秘力量，而是社会化协作劳动的结晶。将这些隐形劳动显性化、制度化，是确保技术发展回归“人的全面发展”这一初心的关键步骤。

V. 结语

立足于 AI 基础设施与 MLSys 的研究前沿，重温马克思恩格斯的科学技术观，结论清晰而深刻：科学技术确为解放生产力的伟力，但其社会效应并非由技术本身自动生成，而是取决于其所嵌入的生产关系、组织架构及分配制度。大模型时代的“机器体系”在构筑新质生产力的同时，也伴生了资源集中、劳动异化及治理外部性等深层矛盾。唯物辩证法的当代价值在于，它将这些矛盾视为可被分析、介入乃至改造的对象：通过推动开源生态与公共基础设施建设、确立以人为本的系统评

价指标以及完善协作劳动的保障机制，我们有望引导技术进步走向人的解放而非新的奴役。对我个人而言，这一理论视角不仅是课程学习的收获，更将成为未来从事 AI 系统研究时不可或缺的价值罗盘与方法论指引。

致谢

诚挚感谢老师本学期的悉心指导。在修读本课程之前，我曾对技术时代研习此类辩证法与政治理论课程的必要性心存疑虑。然而，经过系统的学习与反思，我愈发深刻地意识到：在一个 AI 算力高度受控于资本逻辑的时代，个人的技术贡献究竟是通往自我价值实现与社会财富创造的桥梁，还是异化为资本机器中新型“白领”的生产工具？这门课程为我解答这一困惑提供了关键的理论钥匙，并将持续启发我在未来的科学探索与工程实践中保持清醒的批判意识与人文关怀。

参考文献

[1] Deepak Narayanan, Aaron Harlap, Amar Phanishayee, Vivek Seshadri, Nikhil R Devanur, Gregory R Ganger, Phillip B Gibbons, and Matei Zaharia. 3d parallelism: Combining data, model and pipeline parallelism for deep learning. *arXiv preprint arXiv:2105.14500*, 2021.

[2] NVIDIA Corporation. CUDA Downloads. https://developer.nvidia.com/cuda-downloads, 2025. Accessed: 2025-12-29.

[3] Ruoyu Qin, Zheming Li, Weiran He, Jialei Cui, Feng Ren, Mingxing Zhang, Yongwei Wu, Weimin Zheng, and Xinran Xu. Mooncake: Trading more storage for less computation — a KVCache-centric architecture for serving LLM chatbot. In *23rd USENIX Conference on File and Storage Technologies (FAST 25)*, pages 155–170, Santa Clara, CA, February 2025. USENIX Association.

[4] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020.

[5] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.