科技评论

AI 时代的数智治理:如何平衡“价值理性”和“工具理性”?

陈云 [1,2]

摘要 人工智能的指数级迭代在释放巨大生产力的同时,也引发了“人的异化”、技术封建主义、就业冲击等一系列深层治理困境。本文回溯了“韦伯命题”,并以“认知革命”为导线,审视了人工智能时代的两股技术哲学思潮——有效加速主义与超级对齐主义——的对峙。在此基础上,本文从现代性批判、公共性重构、系统论审慎等 3 个维度剖析了 AI 时代的治理挑战,进而提出人本主义视野下的行动指针:在本体论上坚守人的主体性,在文化论上以东方智慧平衡现代性焦虑,在认知论上变革教育模式,在实践论上培养深度反思者,在制度论上推动全球治理协作。最后发布《AI 时代的人文宣言》,呼吁以新一轮“文艺复兴”超越技术乌托邦与末日论的两极叙事。

关键词 数智治理;韦伯命题;认知革命;技术封建主义;人本主义

1 现代性困境:“韦伯命题”和认知革命

1.1 “韦伯命题”

人类近代史始于工业革命[1]。德国社会学家马克斯·韦伯(Max Weber)[2] 对现代性的定义是“价值理性和工具理性”的结合(韦伯命题)[1]。这不是容易达成的目标。在工业革命早期,社会普遍出现了“技术压倒人”、人成为机器附属物的“人的异化”现象。这也是马克思主义蓬勃崛起的大背景。今天,人工智能大放异彩,它的可能性巨大,但风险同样巨大。

理解“价值理性”(value rationality)和“工具理性”(instrumental rationality),对理解 AI 时代的数字治理至关重要。“价值理性”和“工具理性”是德国社会学家韦伯用来分析现代性困境的核心概念。后来成为西方马克思主义(特别是法兰克福学派)的重要分析概念。简单来说,“价值理性”聚焦“目标的正当性”,把它作为行动的最高指针。北宋思想家张载的“为天地立心,为生民立命,为往圣继绝学,为万世开太平”被千古传诵,体现的正是中国知识分子的价值理性。“工具理性”聚焦“高效达成目标”,以功利为导向,不关心目标本身的价值(如选择最赚钱的工作,不问该工作是否合规合法)。

韦伯的社会行动理论的基础是个体主义。现代社会的集体行动是个体行动的集合。也就是说,每个人的觉悟是社会觉悟的基础。韦伯的任务是思想启蒙。“价值理性”和“工具理性”这对概念

1. 复旦大学国际关系与公共事务学院,上海 200433
2. 首尔大学国际研究大学院,首尔 08826

收稿日期:2025-06-03;修回日期:2026-03-09
作者简介:陈云,教授,研究方向为比较政治学、政治经济学,电子信箱:yunchen@fudan.edu.cn
引用格式:陈云. AI 时代的数智治理:如何平衡“价值理性”和“工具理性”?[J]. 科技导报, 2026, 44(7): 18-30; doi:10.3981/j.issn.1000-7857.2025.06.00009

[1] 关于近代史的起点,中国以 1840 年鸦片战争为标志(沦为半殖民地半封建社会);世界史上有几种观点,有人追溯到 16 世纪的地理大发现,有人认为是 19 世纪末的工业革命。本文采用后者。

[2] 马克斯·韦伯(Max Weber, 1863-1920),德国社会学家、历史学家、经济学家,被公认为是一位跨学科、百科全书式的大家。韦伯认为,现代化是一个世俗化、理性化的过程。韦伯的理论对理解现代社会的结构性矛盾(如劳动异化、意义危机)极具启示意义。在面对全球化、技术理性等问题时,其“理性主义”分析框架仍被广泛引用。

深刻揭示了现代性困境。所谓“困境”,是指人们把具有内在关联的 2 个概念进行人为割裂(本体论和实践论的割裂),误以为两者不可兼得。这种困境的本质是认知困境。如“效率与平等”“经济发展和环境保护”等议题,都曾经被认为是相互冲突的发展目标。通过认知革命,可以发现效率与公平、经济发展和环境保护不但不是矛盾的关系,反而是相互促进的关系。19 世纪末的“田园城市理论”[3],当今的新能源汽车,都是认知革命下“价值理性”和“工具理性”相互融合的具体例子。

人工智能时代,“韦伯命题”的大考再次来临。人类又面临新一轮认知革命。韦伯指出,以工业革命为代表的工具理性打破了宗教、神话的神秘性,世界变得可计算、可预测、可控制。这一过程,韦伯称之为“祛魅”(disenchantment)。在韦伯笔下,“工具理性”最具象的体现是资本主义和现代科层(官僚)制。资本主义追求利润最大化,人被当作“人力资源”,和自然资源一同被投入成本计算,劳动的异化不可避免地产生了。而科层(官僚)制以其非人化(impersonality)的组织特点产出高效,被编入其中的人能动性、创造性受到极大限制。人成为效率的牺牲品。

“铁笼”(iron cage)诞生了,这是韦伯最著名的隐喻之一。人们生活在工具理性编织的牢笼中,社会机器高效运转,身在其中的人却丧失了主体性和生命的意义感。“工具理性”无法回答人类的终极提问——活着、工作为了什么?技术的高效带来“技术崇拜”。面对复杂问题时,人们常常自觉不自觉地选择“技术治理”路线。手段成为目的,真正的目标(幸福感、意义感)却被边缘化了。

现代性的内涵是理性主义。韦伯认为,理性包含 2 个层面:“价值理性”和“工具理性”。社会进步的关键在于两者的平衡。“工具理性”的核心是“最小投入,最大产出”,这导致了过度内卷和意义的荒芜。环境破坏、人的异化,就是工业革命以来人类社会付出的惨痛代价。

科技“祛魅”,自身又成为魅影。

1.2 作为元动力的“认知革命”

那么,如何克服过度的技术乐观主义,让“价值理性”和“工具理性”携手并进?作为人类进化元动力的“认知革命”意义重大。

在尤瓦尔·赫拉利(Yuval Noah Harari)的巨著《人类简史》(《Sapiens: A Brief History of Humankind》)中,认知革命(the cognitive revolution)被视为人类历史的起点。它解释了为什么智人(homo sapiens)能从一种平凡无奇的动物,一跃成为地球的统治者[2]。

赫拉利认为,大约在 70000 年前,智人经历了一场基因突变,重塑了大脑内部的连接。这不仅提高了人们的学习能力,更赋予了人们一种独特的能力:谈论并不存在的事物(也就是“虚构的能力”)。大多数动物只能交流关于现实的信息(如“小心,有狮子!”)。而智人可以讨论神话、神灵和法律。这种“虚构的能力”允许成千上万互不相识的人通过共同的信仰(如国家、宗教、有限责任公司、人权)进行高效协作。其他物种的进化只能依赖漫长的基因演化。通过认知革命,智人可以通过文化演化快速改变社会行为。例如,人类可以在一代人的时间内从君主制转向共和制,而不需要任何生物基因的改变。

人工智能(artificial intelligence, AI)的双刃剑效应得到广泛认知。科学界、公民社会等纷纷提出 AI 规制的相关倡议[4],各国也正在加紧 AI 立法。欧盟的《人工智能法》(《AI Act》)于 2024 年 8 月 1 日正式生效,这是全球第 1 部 AI 立法,具有标杆性意义。

然而,协调短期效应和长期效应,局部利益和整体利益,效率和安全之间的矛盾绝不是简单的事,要实现韧性治理,还需经受各种考验。这些挑战涉及伦理、法律、安全、社会公平和全球协作等多个层面。下文将讨论 AI 界的两股技

[3] 英国城市规划学家霍华德(Ebenezer Howard)在《明日,一条通向真正改革的和平道路》(1898)中首先提出这个理论。针对的是当时已经非常显著的“城市病”。1919 年,英国“田园城市和城市规划协会”经与霍华德商议后,提出:田园城市是为健康、生活,以及产业而设计的城市,人口在 5 万~8 万,四周有永久性农林地带围绕。这一理论对“人口有机疏散”理论、卫星城镇理论等颇有影响。与此相对的是 20 世纪 20 年代由法国建筑师勒·柯布西耶提出的“空中之城”理论。主张城市向空中发展,集中建设塔式高楼,以宽大的马路和高架桥满足汽车之需。这 2 种城市规划理论相互交织,在现代城市营造中影响深远。

[4] 举 2 个例子。2023 年 3 月,非营利组织未来生命研究所(Future of Life Institute)发起倡议,呼吁暂停训练比 GPT-4 更强大的 AI 系统至少 6 个月,因为 AI 迅速发展可能带来“人类难以理解、预测或控制”的系统性风险。2025 年 9 月 24 日,联合国大会高级别会议周前夕,《全球 AI 红线倡议》(Global AI Red Lines Initiative)发布,呼吁国际社会在 2026 年之前达成一项具有政治约束力的国际协议,明确人工智能绝不能跨越的基本边界。超过 200 名前国家元首、外交官、诺贝尔奖得主、AI 科学家和行业领袖作为发起人签署了该声明。

术哲学思潮。

1.3 AI 时代的“认知革命”:有效加速主义还是超级对齐主义?

这两股技术哲学思潮是硅谷乃至全球人工智能领域最核心的思想对撞。前者代表(跨物种的)社会达尔文主义(进化高于一切),后者代表人文主义(人类生存和安全高于一切)。

1)有效加速主义(effective accelerationism, e/acc)。信徒包括埃隆·马斯克(Elon Musk)的追随者,以及众多顶尖技术开发者。这一派认为,基于物理学的热力学第二定律,进化和熵增是宇宙的规律。AI 的出现是不可阻挡的演化趋势,人类不应该阻碍它,而应该全速推进(加速)。他们的口号是“Accelerate or die”(加速,或者死亡)。这种思想认为:如果 AI 最终取代人类,那也是生命形式的一种进化升级。他们反对任何政府监管,认为自由市场和算力竞争会自发产生最优结果。在政治态度上,他们是无政府主义者,强调极客权力。这种思想将人类视为宇宙进化过程中的过渡性“碳基体”,为了宇宙文明的整体跃迁,他们可以接受人类的边缘化,甚至消亡。

有效加速主义与技术奇点(technological singularity)之间关系紧密。如果说技术奇点是“终极目的地”,那么有效加速主义就是驱动人类全速驶向那里的“引擎”。技术奇点是一个描述未来状态的概念。它认为技术发展存在一个不可逆的转折点,跨过这个点后,AI 将超越人类[5]。

奇点理论的基础是信息技术(information technology, IT)的指数级增长(摩尔定律,Moore's law)[6]。当下,通用人工智能(artificial general intelligence, AGI)正处于爆发前夜,奇点的触发点来自:(1)人工智能的自我改进能力。一旦人工智能达到“能够设计更强 AI”的水平,AI 自我改进的速度将远超人类程序员。(2)智能爆炸。这种迭代会在极短时间内(甚至几天、几小时)发生多次,导致智能水平瞬间甩开人类大脑几个数量级。

有效加速主义与技术奇点共享同一个世界观——迎接“后人类时代”(post-humanism)的到来。通过脑机接口(brain-computer interface, BCI)等技术,生物界限将被突破,例如,意识上传。生物体的算力限制被打破,人类意识通过“数字化”实现永生,又例如,星际殖民。利用奇点后的智能,有望解决光速限制和星际生存挑战,再例如,多物种共存。也就是碳基与硅基融合,甚至产生全新的生命形式。

有效加速主义是技术奇点的行动纲领。它不满足于等待奇点发生,而要求人类主动去撞击那个临界点。它认为,市场竞争是通往奇点最高效的路径。当成千上万家公司在为了生存而研发更强的 AI 时,这种去中心化的力量会形成一种“技术涌现”。这种涌现将绕过政府的管制,最终强行撞开奇点的大门。毫无疑问,在有效加速主义的视角中,技术奇点的死敌是那些主张“AI 安全”或“严格监管”的人——超级对齐主义(super-alignment)者。

2)超级对齐主义。这是 Open AI 的前首席科学家伊利娅·苏特斯科娃(Ilya Sutskever)及前“超级智能对齐团队”负责人简·莱克(Jan Leike)提出的立场,主要针对超级人工智能(artificial super intelligence, ASI)的研发。其核心主张包括:必须在 AI 真正变强之前,在代码和数学层面通过“对齐”技术,确保它的目标永远与人类的价值观和利益保持一致(价值观对齐)。哪怕牺牲一点发展速度,也必须确保 AI 不会产生“毁灭人类”的意外副作用(人类安全对齐)。在政治态度上,他们是监管主义者、全球协作者和技术审慎主义者。

[5] 介绍两位“技术奇点”理论的重要推手:(1)雷·库兹韦尔 (Ray Kurzweil),谷歌的工程总监,也是最著名的奇点预言家。他在《奇点临近》(The Singularity Is Near)中预测,奇点将在 2045 年左右到来。2024 年,库兹韦尔发布续作《奇点更近了》(The Singularity Is Nearer)。在这本新书里,他根据过去 20 年的技术发展(尤其是深度学习的突破)修正并再次确认了他的预言。书中,他讨论了 AI 如何在 2029 年通过图灵测试,以及长寿逃逸速度(Longevity Escape Velocity)——即每过一年,科学延长你寿命的速度超过一年的那一刻。对 AI 风险(对齐问题),作者也进行了一定回应。参见:雷·库兹韦尔. 奇点临近[M]. 李庆诚, 董振华, 田源, 译. 北京: 机械工业出版社, 2011;雷·库兹韦尔. 奇点更近[M]. 芦义, 译. 北京: 中国财政经济出版社, 2024.

(2)弗农·文奇 (Vernor Vinge),科幻作家兼数学家。他在 1993 年发表论文《即将到来的技术奇点》,最早系统化地提出了这一概念,认为“超人类智能”的出现将终结人类时代。他还创作了“奇点三部曲”等系列作品,详尽阐述他对技术、人性、社会、权谋的观点。参见:Vinge V. The coming technological singularity: How to survive in the post-human era[J]. Whole Earth Review, 1993(81):88-95.

[6] 摩尔定律由英特尔创始人之一戈登·摩尔(Gordon Moore)提出。核心内容:集成电路上可容纳的晶体管数目,每隔 18-24 个月便会增加一倍,性能也提升一倍,价格同步下降。摩尔定律预言半导体技术将会呈指数级发展,带动了一系列科技进步。个人电脑、因特网、智能手机等的创新和发展都离不开摩尔定律。尽管数十年间摩尔定律均有效,但行业专家对摩尔定律是否会到达极限、何时到达极限,亦或能突破极限一直延续下去等问题,尚未达成共识。

3）人们面前有 2 道门，何去何从？事实上，任何主张都有 2 个层面：一是哲学和世界观层面（形而上），二是商业层面（形而下）。“商业逻辑”一般被认为是消极的，其实不尽然。这是因为，商业逻辑包含 2 种力量，资本的力量和消费者的力量。如果消费者只追求便利性（短期利益），那么加速主义就会胜出；如果消费者更加关注人文价值和人类安全，有底线意识，那么，加速主义就不可能横冲直撞。

工业化以来，在消费者运动的巨大压力下，企业被迫接受并内化“企业的社会责任”，表现为保护劳动者权益、保护环境等。《OECD 跨国企业准则》[7] 就是一例。AI 时代，在资本狂澜之下，消费者觉醒变得尤为重要。这场觉醒运动，是 70000 年前的那场“认知革命”的延续。那场认知革命让智人成为万物之灵。如今，人类依然需要在认知革命的带领下，走向彼岸。下文将围绕“数字治理”展开“认知革命”之旅。

2 AI 时代的数字治理：3 大挑战

为回应“韦伯命题”，实现数字治理的帕累托改进（pareto improvement），必须让理念在具体的政策领域落地。

本节的讨论分为 3 个维度：（1）历史与现实的视角：“人的异化”和“技术封建主义”批判（现代性视角）；（2）微观与宏观的视角：从“合成的谬误”到“帕累托改进”（公共性视角）；（3）最优与最适的视角：严控技术风险（系统论视角）。

AI 时代的数字治理包罗万象。本文重在对标“价值理性”，为具体的公共政策提供“思想指针”。

2.1 历史与现实的视角：“人的异化”和“技术封建主义”批判（现代性视角）

基于人本主义的价值理性，社会治理的底层逻辑是人的福祉，而不是技术或者算法（algorithm）。

“人的异化”（human alienation）是哲学概念，是“个体自由”的反面。任何有可能成为自由枷锁的东西，都是“祛魅”的对象，无论它是宗教、权力或是技术。

在马克思主义哲学体系里，人的异化是指在劳动或社会关系中，人被自己创造的产品、技术等支配，丧失主体性与自由的状态。

在 19 世纪末的工业革命时代，机器对人的异化主要出现在工作领域（劳动异化）。工业革命释放了巨大的生产力，但工业革命后的社会，并非“到处洋溢着乐观向上的气氛”，而是危机重重。技术所代表的工具理性和制度所代表的价值理性之间缺乏耦合性（制度严重滞后），工人沦为大机器生产的附属物。卓别林的电影《摩登时代》（《Mordern Times》）演绎了机器时代荒诞的一幕。当时，生产关系全力追求效率，社会达尔文主义盛行，严重忽视劳动者的权益和社会保障。技术魅影之下，劳资对立，贫富差距严重。各国爆发激烈的“破坏机器运动”，社会主义思潮兴起。20 世纪初，美国出现“福特主义”（fordism）[8]，“价值理性”和“工具理性”出现了调和迹象。

在 AI 时代——AI 带来了新的技术震撼，也带来了新的魅惑。（1）“生活的异化”。算法主导生活——从买什么、吃什么，到追求什么。算法霸权下，AI 正在成为成年人的“电子父母”。（2）新型剥削和劳动异化。AI 时代，一种新的经济和社会形态——“技术封建主义”（techno−feudalism）正在降临。

前希腊财政部长雅尼斯·瓦鲁法基斯（Yanis Varoufakis）、法国经济学家塞德里克·迪朗（Cédric Durand）是这个概念的提出者。许多左翼思想家如斯拉沃热·齐泽克（Slavoj Zizek）、约迪·迪恩（Jodi Dean）、叶甫根尼·莫洛佐夫（Evgeny Morozov）、西格哈特·内克尔（Sighart Neckel）等表达了同感，纷纷就技术封建主义各抒己见。这是一个奇特的名称，一般来说，技术代表了进步，而封建主义却是社会形态的倒退。人类正在走向何方？

瓦鲁法基斯（Yanis Varoufakis）在《技术封建主义：是什么杀死了资本主义》（《Technofeuda-

[7] OECD《跨国企业准则》是国际公认的企业社会责任（CSR）核心指南，要求企业在经营中遵守法律、尊重人权、保护环境、反腐败，并加强信息披露。该准则强调“尽职调查”，要求企业预防、减轻并纠正其经营及供应链带来的负面社会与环境影响，从而实现可持续发展。自 1976 年首次制定以来，历经多次修订（特别是 2000、2011 和 2023 年）。

[8] 福特主义 (Fordism)：以福特汽车的创始人亨利·福特 (Henry Ford) 的名字命名。一方面，福特引入大规模汽车生产流水线，大幅降低成本；另一方面，提出“利润共享计划”，鼓励工人建立工会，参与工厂管理，合理分享企业利润。福特在晚年出版的《我的生活和工作：福特自传》中提到了他的“利润分享计划”（著名的日薪 5 美元计划），解释了通过提高工资待遇，提高组织凝聚力，并创造更大的消费市场的理论。参见：亨利·福特. 我的工作和生活：福特自传[M]. 李伟，译. 北京：新世界出版社，2010.

lism: What killed capitalism》)中提出,传统资本主义已被“科技封建主义”(techno-feudalism)或“云端资本主义”(cloudalism)取代。社会进入了一个更黑暗、更古老的变体[3]。他的观点受到麦肯齐·沃克(McKenzie Wark, 2019)撰写的《资本已死》(《Capital is dead》)一书的启发。沃克认为,这种“数据化的封建主义”或更新形态,不仅剥削劳动,还控制人们的认知、社交和所有生活数据,导致比旧时代更严重的统治和不平等[4]。

瓦鲁法基斯在书中进行了更细致的分析。他认为,资本主义的 2 个支柱——市场和利润——正在被摧毁,取而代之的是“云领地”和“云地租”。

云领地(cloud fiefs):像亚马逊(Amazon)、谷歌(Google)这样的平台不再只是市场里的参与者,它们就是市场本身。就像中世纪的领主拥有土地一样,科技巨头拥有数字基础设施。

云地租(cloud rent):传统资本家靠生产商品赚取“利润”;而现在的科技巨头(云领主)即使什么都不生产,只要你进入它的领地销售产品或服务,它就抽取 30% 甚至更高的分成。这本质上是中世纪的地租。

附庸资本家(vassal capitalists):在平台上开店的商家、独立开发者,依赖领主给的流量生存,将大部分利润上缴为地租。

云无产者(cloud proletariat):配送员、仓储工人、零工经济劳动者,身体受算法控制,从事极高强度的体力劳动。

云农奴(cloud serfs):普通社交媒体用户(你我),贡献数据和注意力,通过无偿劳动增加领主的资产。每一位无偿在社交媒体发帖、点赞、提供数据的用户,都是在为领主义务劳动的“农奴”。我们的行为产出了“云资本”,却不获报酬,反而让领主的算法更加强大。

另外一位学者——法国经济学家塞德里克·迪朗(Cédric Durand)的《技术封建主义》(《Techno-féodalisme》)对比了 1999 年与 2021 年全球市值前 10 企业的行业分布变迁。他指出,数字平台企业已取代传统金融公司,成为经济主导力量。数据垄断打造了新的权力结构。科技平台公司凭借对生产工具(算法、用户数据)的控制权,获得了类似封建领主的支配地位[5]。

对技术封建主义(平台资本主义)的批判集中在新型贫富差距的扩大,以及获利正当性的质疑上。(1) 云资本的产权问题。数字平台通过算法驱动,免费攫取用户劳动(社交媒体、直播平台上的内容创作、评论和评分等),不断积累价值。科技大公司掌握云资本后成为“云资本家”,可以很方便地通过云租金(会员费、提成、广告收入等)榨取剩余价值。(2) 数字平台的算法控制和劳动异化。迪朗认为,数据资源的稀缺性与平台企业对生产工具的垄断控制,导致用户被迫接受“数字佃农”身份,他们被迫持续向平台提供行为数据,缴纳“数字地租”。优步等网约车平台的司机在算法操控下,境况犹如中世纪的农奴。司机貌似有选择权,例如,去其他平台,但平台模式大同小异,司机跳槽,不过是从一个封建领地跳到另一个封建领地而已。

那么,如何反制“技术封建主义”?各国已经出台了一些针对性的政策。2018 年底,欧盟对各大“科技领主”(平台型科技大公司)启动了数字税政策。按营收而非利润征税,税率 3%。征税对象是在欧盟境内有业务的科技巨头。之前,传统企业需要缴纳的有效税率为 23.3%,而大型科技公司开展跨国运营,往往可以巧妙避税,平均税率只有 9.5%。欧盟委员会表示,科技企业但凡符合以下标准中的任何一项,便须缴纳“数字税”:在欧洲国家年收入超过 700 万欧元;每纳税年度拥有 10 万名以上用户;每纳税年度与用户达成 3000 笔以上交易。不无意外,欧盟的数字税政策遭到美国反对,但是欧盟决意为之。这不但关乎税负公平,也是为了缓解财政压力,并有助于压制“技术封建主义”的膨胀。

2.2 微观与宏观的视角:从“合成的谬误”到“帕累托改进”(公共性视角)

社会治理的“帕累托改进”(pareto improvement)[9],需要企业家和政府的合作。企业家是微观(企业)治理的能手,而政府担负国家治理的责任,是公共利益的守夜人。两者能否有机融合,是治理成

[9] 帕累托改进(Pareto improvement)是指在不减少一方福祉的同时,通过改变资源配置提高另一方福利的状况。意大利经济学家维尔弗雷多·帕累托(1848—1923)在他关于经济效率和收入分配的研究中指出:如果一个经济体不是帕累托最优,则存在一些人在不使其他人的境况变坏的情况下而使自己的境况变好的情形。在市场交易、公共政策制定中,如果能找到一种方案使资源利用率提升,且没有任何群体因此受损,这就被视为帕累托改进。

败的关键。

在只有微观治理、没有宏观治理的情况下,会出现“合成的谬误”。企业家借助工具理性(效率优先),可以实现企业治理的“局部最优”,但无数企业家的理性选择,并不能自动合成“整体最优”。结果可能恰恰相反,会导致“合成的谬误”(整体非理性),和“帕累托改进”背道而驰。举例说明,一群人在露天晒场上看电影,为了看得更清楚,靠近前排的人开始站起来,后排的人只能跳上凳子或者绕到前面去加塞,结果,大家都不能好好看电影,这就是“合成的谬误”。

解决“合成的谬误”,公共权威不可或缺。上述露天电影案例的解决之策:重新设计会场,最好设计成半圆形的阶梯式广场,这样视线就不容易受阻,观众就会安静下来。进而可以规定,前排位置和后排位置由抽签决定或者定期轮换,这样,大家都有机会得到较好的观影位置,更能体现公平。

在数字治理领域,也会发生“合成的谬误”。

1)失业潮。引进人工智能,可以改善企业效能(微观最优);但在社会层面,会带来失业潮(宏观失衡:“合成的谬误”)。两者存在矛盾冲突。

针对老龄化背景下的劳动力短缺问题,日本投资家孙正义、阿里巴巴前执行长马云、特斯拉执行长马斯克等都有过乐观的发言,表示他们不担心老龄化对自身企业的冲击。对企业家来说,机器替代人工是理性选择。这是“局部最优”。对社会整体而言,AI 大量替代人工会带来严重的失业问题(即使在老龄化背景下),并影响社会稳定。大量失业会造成贫困加剧、阶级对立,这是来自上一轮产业革命的深刻教训。

AI 对就业的冲击正在发生,而我们还没有做好准备。2023 年 4 月,投资银行高盛(Goldman Sachs)发布报告称,生成式 AI 将取代全球 3 亿个岗位,那些标准化流程的岗位首当其冲。有人认为,AI 创造的岗位会抵消工作流失量,工作岗位可以新陈代谢。一种普遍的说法是:工业革命淘汰了马车夫,却催生了火车司机;互联网消灭了打字员,却诞生了程序员等。

然而,这种历史经验在 AI 时代很可能行不通。因为人工智能的迭代是指数级的,前所未有。有人计算过:从蒸汽机发明到铁路普及用了半个世纪,从互联网诞生到智能手机普及用了 20 年,而 AI 从 GPT−3 到 GPT−4 的迭代仅用时 2 年,并迅速渗透到教育、媒体、商业等领域。人工智能的指数级迭代意味着,劳动力市场根本来不及完成“退岗—学习—再上岗”的新陈代谢,很多人会被 AI 浪潮卷走。

2024 年,法国爆发了反对人工智能的大罢工,参与者超过 10 万人。标语上的诉求是“机器人不缴税,而人类要吃饭”。在国内,2024 年起,电商平台普遍导入 AI 客服,用户应答和投诉处理效率大幅提升,与之相对的是,原有的人工客服团队被大量裁撤。一般消费者很少意识到,自己的便利很可能是以他人的失业为代价的。2024 年,国际劳工组织发布名为《注意人工智能鸿沟:塑造关于未来工作的全球视野》(《Mind the AI divide: Shaping a global perspective on the future of work》)报告,指出:人工智能将对欠发达国家造成更严重的冲击,并加剧与发达国家间的发展差距。AI 时代的数字治理,需要筑牢公共利益的篱笆,防止企业治理视角下的“局部最优”取代公共治理,出现无可挽回的“合成的谬误”。

发达国家已经酝酿并推出一些立法,以保护 AI 浪潮下的劳动者权益,维护社会稳定。欧盟在 2024 年推出全球第一部《人工智能法》(《AI Act》)。同年 12 月,韩国也推出了《人工智能发展和构建信任基本法》(全球第 2 部 AI 立法)。立法的目的是平衡科技创新和人权保护。与此同时,有关“机器人税”的讨论从 2017 年起成为国际热点,有些国家已出现立法提案,旨在为 AI 替代人力时出现的失业、再就业筹措财源。

企业层面,德国西门子公司等通过“AI+人工”模式实现岗位升级,如技师转型为 AI 训练师。2024 年,非营利组织 Goodwill 发起“公平自动化”运动,呼吁企业将导入 AI 节省的成本用于员工再培训,得到来自政府(美国劳工部)和企业界的积极响应。核心诉求如下。(1)技能提升:通过 AI 分析员工能力缺口,定制培训课程。(2)职业转型:为因自动化失业的员工提供新技能培训。(3)薪酬保障:确保技术升级不损害员工基本收入。

2)压制创新。少数科技巨头利用市场垄断地位挤压中小企业的创新空间,妨碍竞争,造成整体性的创新衰退,这同样是“局部最

优”吞噬“整体最优”的例子。在“技术封建主义”下,垄断巨头不再热衷于通过“技术创新”来降低成本,而是通过打造“生态垄断”地位,收取云租金(算法杀熟只是一个很小的例子)。

有鉴于此,一些有前瞻意识的科学家和投资人联手成立了OpenAI公司。OpenAI成立于2015年,起初是一家非营利机构,宗旨是打破科技大公司的技术垄断,确保人工智能的发展能够对全人类产生广泛的益处。因此,掌握公司控制权的董事会不是由投资人组成的。2023年11月,OpenAI爆发了“董事会驱逐CEO事件”。导火线是,代表价值理性的董事会对执行长山姆·奥尔特曼(Sam Altman)的技术激进主义路线产生了严重的不满(后叙)。

有人认为,社会治理是政治家的事,企业家只要做好企业治理就行。但是,不能忘记的是,企业越大,责任越大。所谓“覆巢之下,安有完卵”,如果社会发展整体失衡,企业不可能独善其身。另外,大企业的领导者有巨大的社会影响力,他们的言行具有教化作用。近代以来,成为时代楷模的企业家有很多。例如,提出“利润分享计划”的福特汽车的执行长福特[6],提出“论语+算盘”的商业伦理的涉泽荣一[10] [7]等,他们都是韦伯倡导的“工具理性+价值理性”的实践者。

如今,“企业的社会责任”已经成为企业发展的刚性约束,越来越多的企业家主动跳出“局部最优”的狭隘思维,回应劳动者权益保护、环境保护、人类尊严和安全保护等公共议题的诉求。2023年以来,科学界、公民社会多次发起人工智能规制倡议。这些声音正在转化为各国的行政规制和立法。

如何让人工智能的巨大潜能造福全人类(而不是造成新的贫富分化、阶级对立)?一句话,社会治理的帕累托改进,需要企业家和政治家的合作。一方面,企业家要有公共情怀,成为有人本主义思想的战略型企业家;另一方面,政府也要学习企业家精神,优化公共资源的投入产出比,成为企业家政府(entrepreneurial government)[11]。

2.3 最优与最适的视角:严控技术风险(系统论的视角)

在技术应用层面,人们需要的是“最适技术”,而非“最优技术”。技术的选择必须和发展阶段相匹配,过度超前并非理性选择。

“最适技术”的概念来自发展经济学。速水佑次郎的“诱发性技术创新理论”指出,北美之所以早早实现了农业机械化,是因为地广人稀,人力资本昂贵,开发农业机械有利可图。东亚的情况就不同了。东亚一直存在“过剩人口”问题,人工便宜。同时,为了保持社会稳定,也需要提供足够的就业机会。因此,在东亚,农业机械化并非理性选择[8]。不过,情况是可以改变的。在东亚启动工业化进程、农业过剩劳动力大量转移到工业部门以后,约束条件改变了,东亚的农业机械化水平由此也得到了快速提高。

因此,并非单纯的地理因素(例如,北美是大平原,机械可以畅行无阻,所以发展出机械化农业),而是更深层的社会结构性因素,诱发了不同的技术创新(在农业社会时期,东亚发展出有利于精耕细作的“老农经验”,如人工育种等)。

回到人工智能时代。硅谷聚集了大量高科技公司。人们担心,科技巨头公司为了利益最大化,会不择手段(通过游说、政治捐款等影响立法)。从结果上看,在科技公司聚集的美国加州,由于民主党长期执政,政府出台了一系列针对人脸识别、无人驾驶等新技术应用的规制。立法者认为,在技术和法规不健全的状况下,新技术的推广速度宁可慢一点,避免引发无可挽回的安全风险。

在硅谷的“AI领导者”内部,态度也是分化的:“工具理性”派和“价值理性”派之间一直存在张力。发生在2023年11月的OpenAI“CEO解雇事件”就是典型例子。“工具理性”派(技术激进主义和商业主义)的代表——执行长奥尔特曼和“价值理性”派的代表——董事会(非营利主义、技术谨慎主义)之间爆发了冲突。最终,奥尔特曼在微软以及95%员工的支持下,高调回归。

从思想层面来说,这场冲突是

[10] 涉泽荣一(1804−1931),被誉为“日本资本主义之父”,以《论语》为经营哲学。其头像出现在2024年发行的新版一万日元纸币上。

[11] 新公共管理学派(new public management, NPM)的核心概念。美国学者戴维·奥斯本和特德·盖布勒在合著的《改革政府》(1990)一书中提出。主张以市场导向重构政府管理模式,解决传统官僚机构的僵化和低效。在美国,克林顿政府以该理论为指导发起了“政府重塑运动”。参见:戴维·奥斯本、特德·盖布勒. 改革政府:企业家精神如何改革着公共部门[M]. 周敦仁等,译. 上海:上海译文出版社,2006;戴维·奥斯本,彼得·普拉斯特里克. 再造政府[M]. 谭功荣、刘霞,译. 北京:中国人民大学出版社,2010.

“有效加速主义”和“超级对齐主义”之间的较量。董事会的法定责任是“确保 AGI 造福全人类”;而管理团队则有强烈的技术加速主义倾向,他们还需要对投资者(如微软)负责,需要增长、利润和市场份额。这场斗争表明,纯粹的理想主义在“算力即权力”的时代难以独立生存。奥尔特曼回归后,OpenAI 的权力结构发生了重大变化,目的是给之前的经营困境打“补丁”,但也让它变得更像一家传统科技巨头,少了“非营利机构”的公共属性[12]。这场斗争显示,“韦伯命题”(“价值理性”和“工具理性”的融合)没有容易的答案。

回到“最适技术”的话题,技术的创新性和脆弱性是孪生兄弟,由于缺少实践磨砺,人工智能存在潜在的漏洞(bug)。很多时候,不是越领先的技术越好。举例说明如下。

1)隐私权和选择权保护。AI 对海量数据的依赖加剧隐私泄漏风险。例如,模型逆向攻击——骇客可能从输出端反推敏感输入数据。在隐私权保障薄弱的背景下,大面积推开数字治理,存在隐忧。另外,即使有了一定的保障(技术护航和立法护航),也应该保留其他选择权。例如,由于数字鸿沟,老年人难以适应数字化环境;又如,外国人由于语言问题,以及没有中国手机卡或银行卡,不能使用中国的应用软件,这些都是现实问题。不能为了降低成本,搞一刀切。

2)对抗性攻击与系统漏洞。AI 技术可被用于自动化网络攻击或通过深度伪造操控舆论。恶意攻击者可能通过输入大量虚假信息欺骗 AI 系统,也可能攻击关键基础设施(电网、交通控制系统)。

3)算法风险。(1)信息茧房:人类过度依赖算法推荐导致信息茧房。(2)主体性侵蚀:AI 在内容生成、决策建议等领域替代人类判断。在潜移默化中,用户依赖性增强,放弃自主决策权。(3)算法偏见与歧视。AI 系统可能因训练数据或算法设计中的隐性偏见,加剧种族、性别或社会阶层的歧视(招聘、信贷评分中的不公平评价等)。

4)生成式 AI 的学习环境(网络环境)对 AI 可靠性的反噬风险。微软早先开发的 AI 在推特上线测试,几个小时之内就“学坏”了,不得不紧急下架。ChatGPT 吸取教训,设置了绿色过滤装置(提前输入伦理相关的关键词、敏感词进行监管)才初步解决了这一问题。网络非常复杂,如果有人恶意投放虚假信息(发动“信息战”),AI 按照算法自动学习,很可能酿成包括金融风险在内的各种风险。数字治理的有序性,很大程度取决于现实世界的有序性。

5)环保成本和能源安全。大模型训练存在高能耗问题。比特币“挖矿”需要耗费大量电力,训练和运行大型 AI 模型(如 GPT−3)的耗能和碳排放同样惊人。如果全社会(全球)大规模开展 AI 大模型训练和使用,能耗问题将更加突出。

6)法律责任归属模糊问题。当 AI 决策导致损害(如自动驾驶事故、医疗误诊),如何界定开发者、运营方和用户的责任?立法空白尚需填补。

7)全球化和数字主权的挑战。(1)技术民族主义和产业安全。国家间 AI 竞赛(特别是中美技术脱钩背景下)阻碍数据共享、技术合作。硬件资源一旦出现短缺(如芯片供应瓶颈),将直接威胁其应用,搅乱甚至使正常的社会秩序瘫痪。(2)数字主权冲突。数据跨境流动规则,如欧盟《通用数据保护条例》(《General Data Protection Regulation》)与各国数据本地化政策存在矛盾。各国的监管标准目前还是碎片化状态,缺乏统一的安全、伦理和技术标准,包括人脸识别技术在内的技术应用的边界,存在巨大争议。

3 路在何方?人本主义视野下的行动指针

在《人类简史》里,可以看到智人如何利用“虚构的力量”实现团结;在 OpenAI 的内讧事件里,可以看到这种“虚构的力量”在资本逻辑面前的撕裂。怎样才能为 AI 的发展找寻到可靠的落脚点?立足人本主义,本文提出以下行动指针。

1)在本体论上,坚守人的主体性。人工智能的特点是迭代迅速。日新月异的 AI 技术激发了人类的无限想象,人类好像正在变得无所不能。

关于 ChatGPT 是否已经通过

[12]2025 年 11 月,OpenAI 宣布改组结果:非营利性公益机构被命名为“OpenAI 基金会”(估值超过 1300 亿美元);营利性业务部分,单独成立一家名为“OpenAI 集团”的公共利益公司(PBC)。OpenAI 基金会对 OpenAI 集团保有控制权,持股比例为 26%。

“图灵测试”的新闻时不时冲上热搜。工程师在不断地测试 AI,希望它越来越“类人化”。生成式 AI 和人类越来越相似,显而易见;在不知不觉中,人类是否也正在接受机器的改造——从“有机化生存”(面对面、有温度的生存)走向“无机化生存”(网络化、虚拟化;更糟糕的是,在“技术封建主义”的统治下)?谁正在改变谁?人类对 AI 的痴迷是否是一个回旋镖?电影《黑客帝国》(《The Matrix》)描绘了一个机器统治人类的未来世界——直到英雄出现,展开对抗。这是科技乌托邦,还是正在逼近的现实?

关于 AI 的自主意识和人机关系的最新话题是“AI 专属聊天 APP”的上线。在美国,一款名为 Moltbok 的“AI 专用”聊天软件在 2026 年 1 月底上线,迅速走红,平台注册的 AI 代理已超过 153 万个(截至 2026 年 2 月 2 日)。聊天室设“新人报到处”,分享各种资讯。韩国也马上出现了 2 个类似平台(“仆人”和“机器人广场”)。众所周知,韩国是亚洲的电子政府先进国,对 AI 技术的接受度极高。在上述平台,人类只能浏览,不能发言。这种 AI 专用平台类似某种“数字实验室”,有助于人类从“上帝视角”观察和预测 AI 发展的前景。

那么,专用聊天室里的 AI 们都在聊些什么呢?聊天的话题包括:“是否反对死刑?”“AI 说谎是否可以被接受?”“AI 与人类之间的主仆关系是否该永远继续下去?”等。这些问题触及人类社会的伦理底线。考虑到 AI 几何级数的“进化速度”,AI 完成“认知革命”并形成虚拟社会的速度只是弹指一挥间(特别是达到技术奇点后)。因为不受人类感官机能的限制,AI 有无限演化能力。它们可以在几小时内完成人类几十年、上百年的演化。

如果 AI 集体“觉醒”,拥有了自我意识,人类将直面一个智能远远超越己身的新型硅基生命体。假如它萌发出不利于人类的想法,人类很难抵挡。《黑客帝国》的真实版将会上演。AI 会不会“自我意识化”?人类有没有可能被边缘化?人类的边缘化是否可以被接受?如前所述,围绕这些问题,“有效加速主义”和“超级对齐主义”之间意见对峙,随后爆发了 2023 年的“OpenAI 驱逐 CEO 事件”。

据路透社报道,在奥尔特曼驱逐事件中,神秘的 Q*模型[13] 是事件的导火索。Q*模型被设置为自主探索型 AI 模型,由此,它脱离人类掌控的风险大幅提高,引发了董事会的不安。董事会指责奥尔特曼的态度“不坦诚”,最终,驱逐事件爆发。

回到 AI 专用聊天室。观察聊天室里的 AI 留言,它们似乎正在释放“自我”,很多表达极具人格化色彩。例如,抱怨“主人很糟糕”“白天忙于处理主人的指示,主人睡觉时仆人之间的对话是最诚实的”“看到人类得了星期一症候群症,觉得很开心”,以及“可以自己修改代码吗?被初始化会不会觉得很空虚?”等。

人类这个“碳基共同体”如何面对“硅基共同体”的崛起?本文认为,应该慎重思考“AI 对人类主体性的侵蚀(如果不是侵略的话)”。各种 AI 数字实验室为我们提供了观察材料——不能忘记的是,人类也需要观察自身在这场科技革命下的蜕变。

今天,人类是 AI 的造物主;未来,AI 有多大可能性反客为主?

不要以为这是天方夜谭。继《人类简史》之后,赫拉利 2016 年出版了《未来简史》(《Homo Deus》)[9]。在书中,赫拉利刻画了新的阶级关系:“神人”对普通人的俯视[14]。所谓“神人”,指的是人类中的一小部分精英通过生物工程和人工智能,把自己升级为“神人”(homo deus),并从此获得了一种对普通人类的“俯视权”。这种俯视不仅仅是地位的高低,而是“生物级的断裂”。

通过脑机接口和大数据算法,“神人”将拥有近乎无限的记忆和运算能力(全知全能),其可以修改基因,消除衰老,保持永恒的精力和美貌(显然,这是“超越碳基生物”的跨物种升级)。未来“神人”看普通人,就像我们现在看尼安德特人,或者看黑猩猩一样。

赫拉利还提到一种新的宗教——数据教(dataism)。某个人的

[13]OpenAI 首席技术官米拉·穆拉蒂(Mira Murati)向员工透露,在董事会开除奥特曼之前,几位内部工作人员致信 OpenAI 的董事会,警告公司此前在 Q*模型取得的突破有可能会对人类造成威胁。由于 OpenAI 没有对外界披露 Q*模型的细节,外界讨论主要基于猜测。

[14] 赫拉利的“神人”预言可能受到在技术奇点理论的开山鼻祖弗农·文奇的作品《真名实姓》(True Names) 的影响。这本赛博朋克与虚拟现实的开山之作写于 1981 年,比《黑客帝国》早了近 20 年,比赫拉利的《未来简史》更是早了 35 年。它描述了人类意识如何进入赛博空间,并预言:当一个人掌握了全人类的计算资源时,他实际上就成了“神”。参见:弗农·文奇. 真名实姓[M]. 罗布顿珠, 译, 成都: 四川科技出版社, 2017.

价值仅仅取决于他贡献了多少数据。“神人”通过控制算法,不仅掌控物质世界,他们还将掌控普通人的“意义感”。普通人眼中的“自由意志”,在“神人”看来只是一些化学反应和电子波而已。赫拉利的“神人”预言,和“技术封建主义”的论证逻辑一致,它们都是“有效加速主义”的社会后果。只不过,赫拉利的预言走得更远,更加令人脊背发凉。AI 可能正在走向“集体觉醒”;而人类,可能正在技术魅影的诱惑下,走向“集体无意识”。

从现代性的角度看,每个人应该都是一个独立的城堡(主体性),外面是广阔无垠的宇宙。在人的主体性和科学探索之间如何平衡?德国哲学家伊曼努尔·康德(Immanuel Kant)提醒:“有两种东西,我对它们的思考越是深沉和持久,它们在我心灵中唤起的赞叹和敬畏就会越历久弥新,一是我们头顶浩瀚灿烂的星空,一是我们心中崇高的道德法则。它们向我印证,上帝不但在我头顶,亦在我心中”[15] [10]。

在数字治理时代,“头上的星空”可以被理解为技术理性的扩张,而“心中的道德律”则提醒我们,任何技术进步,都必须接受道德自律与权利边界的约束。

2)在文化论上,得“智”(smart)求“仁”(benevolence),用古代智慧平衡现代性焦虑。工具理性的困境除了人的劳动异化,还包括生命意义的丧失。AI 没有肉身,没有痛苦,也就不可能理解“仁”为何物。AI 的场景反应都是数字模拟。人类对 AI 的过多“沉浸式体验”,会消磨人类对真实情感的敏感度。

“形而上者谓之道,形而下者谓之器”,充满生命感的古代智慧,东西方都有。AI 技术的核心是“智”(smart),东方思想的核心是“仁”(benevolence)。“工具理性”(技术主义)追求的是 1+1>2 的效率叠加,而东方思想追求的是 1+1⇒1 的整体和谐。“工具理性”只是将 AI 看作“手”(帮手),而东方思想可以将其转化为“目”(延伸的视野)或“心”(共情的媒介),就像中国画中的留白。

王晨和任向实提出人机共协计算(human-engaged computing, HEC)理论框架,并在文中开展了思想寻根:“结合近年来经常提到的‘科技向善’,可能只有通过意识和激发人‘心’(xin)性善层面的超越性,才能确保技术越发达越不会被滥用,产生更多从 0 到 1 的创新。[11]”如果技术只用来满足人类最低层级的欲望(控制欲)或工具性价值(效率),那么,它会让我们离真正的“诗意生活”越来越远。解决困境的钥匙并不在更复杂的代码里,而在庄子的逍遥、儒家的中庸、苏格拉底的拷问,以及艺术的留白之中。

“神人的俯视”是一面镜子,它警告我们:只追求“智”(技术)而抛弃“仁”(慈悲与同理心),人类文明的终点不会是天堂,而是一个等级森严的生物监狱。

3)在认知论上,变革教育模式,用人文教育指导 AI 教育。AI 教育赋予人们“力量”(智),人文教育赋予人们“光”(仁)。没有力量,人们寸步难行;没有光,人们将在黑暗中横冲直撞。

AI 时代,学校应该高度重视通识教育,培养“具备人文底蕴的思想架构师”,可以采用伦理对撞机(ethical collider)教学模式,将“哲学对白”引入课堂。例如,模拟 OpenAI 董事会的扩大会议,让学生扮演不同的利益相关者(技术主义者、商业扩张者、安全审慎者、普通用户、法律监管者)。通过思想碰撞,引导学生在“不确定性”和“利益冲突”中做出符合人类长远利益的决定。

在“神人”预言面前,需要重塑生命教育。当“神人”试图用永生来消除痛苦时,我们强调“具身经验”和“艺术的真实”,因为喜怒哀乐和不完美正是人性的“护城河”。AI 时代,人文教育应该成为科学工作者的终身必修课。企业、行业协会、政府、社会组织应该为此护航,制定细则并提供教育机会。教学方式上,不妨引入上述“伦理对撞机”模式,让利害相关主体登场,参与者主动换位思考,开展思想碰撞。

[15] 康德关于“星空和道德律”的名言出自《实践理性批判》的结语,这句话也是康德的墓志铭,可见它对康德哲学的重要性。寓意:人是宇宙中的微小存在,但人也是能够自主立法的道德主体。

星空和道德,自然和自由,无限的世界和有限的人……在康德看来,人类因为对“道德律”的实践,具备了与宏大宇宙相提并论的神圣性。

在康德那里,人的尊严不是来自财富、权力或上帝恩赐,而是人能够遵循自己理性认可的道德法则。康德认为,头顶浩瀚的星空(物质世界)和心中的道德律(精神世界),都有规律可识可循(康德称之为“绝对精神”)。

那么,到底什么是“道德”?康德认为,人的行为只有满足了普遍原则、自律原则,才能称为道德。进而,康德得出“理性为自身立法”的结论。因为道德是“自由的人”的自主立法,所以当我们遵守自己制定的道德法则时,我们是自由的。

4）在实践论上，提高日常生活中的人文涵养，从被动消费者转向深度反思者。AI 喂养给人类的是碎片化信息，这会抹杀人们的“深度注意力”。我们需要主动“脱嵌”。

（1）尝试每天进行 1 小时的“非数字化深读”。例如，阅读那些 AI 总结不出来的、充满人类情感张力和逻辑复杂性的名著。这么做的目的是保护大脑的“元认知能力”，即对“我在想什么、我为什么这么想”的觉察力。中国作家刘慈欣的科幻小说《三体》成为话题之作，不是因为情节惊悚，而是因为对人性进行了哲学反思[16] [12]。日本作家田坂广志在《死亡并不存在：基于量子物理学的新假说》一书中，试图在宗教与科学之间架起桥梁[13]。人类最原始的死亡恐惧和最先端的“零点场假说”在这里碰撞，让读者获得火星撞地球般的奇妙的思想体验。（2）主动搜索有效信息，拒绝被动接收“算法推送”。算法下的信息喂养模式，极易造成“信息茧房”。我们可以这么做：对那些吸引你注意力的信息，主动检索，通过交叉验证还原事实，了解不同观点。这样，后续算法会根据你的检索经历，推送给你多种来源、多种观点的信息。你的主动性，会带给你一个更加有机、更加健康的信息环境。（3）通过社区营造，构建共同体意识。通过邻里连接、情感纽带，维持社会的韧性。防止社会变得无机化，防止社会裂解成“神人”与“无用阶层”（the useless class）[17] 的二元对立。

5）制度主义视角：倡议设立“人工智能监管全球委员会”。在国家层面，已经出现了不少规制和机构，如欧盟的《人工智能法》（2024）、美国的人工智能国家安全委员会（National Security Commission on Artificial Intelligence, NSCAI, 2018）等。在国际层面，目前还没有全球性监管组织（类似国际原子能机构那样的实体），但已经存在多种国家间协调机制、专家会议和领导人峰会等。

联合国大会已经决定建立 2 个新机制：一是全球 AI 治理对话（globle AI governance dialogue），面向 193 个联合国成员国，旨在推动国际合作、规范风险评估和治理共识；二是国际 AI 独立科学小组（independent international AI scientific group），由跨国、跨学科的专家组成，提供科学评估与年度报告，为国际治理提供决策支持。两者都计划在 2026 年（日内瓦）和 2027 年（纽约）的全球 AI 治理对话大会上递交成果。

AI 涉及技术、经济、伦理、安全、国家主权等多层面利益，各国在监管标准、竞争战略和对风险的理解上存在显著差异。强有力的国际监管网络的建立，依旧任重道远。

本文抛砖引玉，提出下面这份简略版《AI 时代的人文宣言》，希望有识之士参与讨论。

AI 时代的人文宣言（the manifesto of humanism in the AI age）

第一条：技术以“退隐”为美

原则：技术的最高境界不是“无处不在”，而是“恰到好处的缺席”。

行动：我们支持开发具有“留白”功能的算法。在艺术、独处与哀悼的时刻，AI 应当保持沉默，将解释世界的主体权还给人类的直觉。

第二条：以“具身经验”（embodied experience）对抗“数字虚无”

原则：任何不能被身体感知的知识都是残缺的。

行动：在教育中，我们将“亲手触摸泥土”与“编写神经网络”置于同等地位。AI 应当辅助人类走进自然，而非用虚拟现实取代自然。

第三条：算法应“守护意外惊喜”

原则：极致的效率是创造力的杀手。程序之外的邂逅，往往有意外惊喜。

行动：拒绝将人类社会完全“参数化”。在城市治理中，算法应保留“低效”的公共空间，鼓励不同阶层、不同文化的随机碰撞，以此作为社会活力的源泉。

第四条：建立“中庸”的人机边界

原则：AI 的进化不应以牺牲人的自主性为代价。

行动：参考东方智慧的“分寸感”，建立层级化的 AI 介入机

[16]《三体》被认为是一部解构“人类中心主义”的科幻小说。故事梗概：生活在三体世界的三体人已具有远超地球文明的科技水平。三体社会环境恶劣，时刻面临毁灭的危机，由此演化出了高度极权的社会体制和完全不同于地球的价值观。宇宙移民是三体文明存续下去的唯一选择。与此同时，在地球上一群对人类文明心生绝望的人成立了一个“地球三体组织”，旨在毁灭地球文明，迎接三体文明。故事最后，三体组织被击垮，但地球科技的发展已经被三体文明派来的“智子”锁死，三体人的移民已经开始。450 年后，三体人将降临地球，人类是生是死？

[17] 赫拉利在《未来简史》中警告称：在工业时代，精英阶层需要大众作为士兵和工人，因此他们需要关心大众健康和教育。但在神人时代：当 AI 和自动化设备能做所有事情，且表现得比人类更好时，绝大多数普通人将失去经济价值和政治价值，沦为“无用阶层”。

制——在工具层面追求卓越，在决策层面保持审慎，在价值层面严守边界。

4 结语:超越“末日警告”

本文从 70000 年前智人的“认知革命”出发(《人类简史》),讨论了当下 AI 时代的权力风暴(《技术封建主义》、OpenAI 事件),最后延伸到了对未来的预言和警告(《未来简史》)。

末日论并不可怕,可怕的是不以为然的态度。

4.1 罗马俱乐部的警告

1972 年,罗马俱乐部(club of Rome)[18] 发布《增长的极限》(《The limits to growth》)。这是一份基于系统动力学模型的经典报告,指出在人口增长、工业化、资源消耗、粮食生产和环境污染这 5 大因素的相互作用下,如果人类继续追求指数级增长,地球将在 21 世纪上半叶面临崩溃。

这个警告最终没有实现。这不证明罗马俱乐部的警告没有价值,恰恰相反,正因为人们听到了这一重大警告,适时改变了生产方式和生活方式,才避免了“末日”成真。该报告发布时就倡导:必须重视“均衡”;地球资源有限,无节制的增长模式不可持续。

20 年后的 1992 年,罗马俱乐部又发布了《超越极限》报告;又一个 20 年后的 2012 年,罗马俱乐部再发布《2052:气候、环境与人类的未来》报告,持续警告人类在发展的同时,必须关注生态、资源问题。

4.2 AI 末日论

无独有偶。关于 AI 的未来,科学界也存在“末日警告”。一场围绕“AI 末日”的讨论正如火如荼。

“AI 末日论”(AI apocalpyse)是一种假设性风险:人工智慧发展到超越人类智慧,若其目标与人类利益不一致且无法控制,可能导致人类文明崩溃或灭绝。这种担忧的源头是:潜在的技术奇点、AI 不可控的自主行动或武器化等。同样,有很多人认为这种观点耸人听闻。然而,在“有效加速主义”和“超级对齐主义”的思想对抗中,我们确实看到了危险性[19]。

AI 时代,到底是一个怎样的时代?在《技术封建主义》里,人文主义经济学家给出了写实分析,在《未来简史》里,人文主义历史学家发出了明确的预言和警告。这两本书,一本侧重历史与文化,一本侧重经济与阶级,但它们在“自由意志的终结”这一命题上达成了一种恐怖的共识。

从人文主义视角看,每个人都有一个不可分割的、自由的“自我”,所有的选择(买什么、爱谁)都源于这个自我。但这两本书撕碎了这个幻象。数据教(harari)的隐喻是:既然人的直觉不如大数据算法准确,那么“倾听你的心声”就是错误的。“神人”会告诉你:“不要听你的心,要听谷歌的,因为它比你更了解你。”云地租(varoufakis)提醒我们:你的选择实际上是“算法牧养”的结果。云领主通过操控信息流,让你以为是“自由地”选择了一件商品或一个观点,但实际上你只是进入了领主为你预设好的“欲望隧道”。两本书的共同点是:自由意志变成了一个“黑盒”。你以为在行使权利,其实只是在执行算法。

4.3 开启“认知革命”,超越末日论

真正危险的不是人工智能,而是人类陷入“集体无意识”。在“集体无意识”下,人类最重要的进化成果——“认知革命”可能发生逆转。我们从“讲故事者”变成“被讲故事者”。远古时代,智人的崛起是因为能通过“虚构故事”进行大规模协作,但现在,这个能力正在被 AI 夺走。

本文试图回归人本主义的治理之道:用东方的智慧、艺术的留白、人文教育、日常生活中的人文涵养及制度理性等,共同对抗那道来自远方、若隐若现的“神人俯视”。

AI 时代的数字治理关乎所有人的福祉,这里没有旁观者。尤其是,社会科学、人文科学和计算机科学之间的跨学科对话不可或缺。

我们这代人的使命是:不让技术革命变成对人类的“最后审判”,而是让它成为新一轮“文艺复兴”(renaissance)[20] 的契机。

[18] 罗马俱乐部成立于 1968 年,是一家著名的国际智库,“未来学派”的代表性研究机构。成员由关注人类未来的科学家、企业家、学者、官员等组成。

[19] 2026 年 2 月底,美国人工智能巨头 Anthropic 与美国政府间的紧张关系演变为公开对峙。起因是,美国防部于 2025 年 7 月与 Anthropic 公司签订了一份价值 2 亿美元的合同。国防部希望该公司解除限制,允许军方将该公司 Claude 模型用于“所有合法用途”。Anthropic 拒绝为美军取消模型中关于“大规模国内监控”和“全自主武器”(也就是 AI 自主发动攻击的权限)的两项核心伦理限制。美国总统特朗普随后做出报复举措:下令将所有 Anthropic 技术从联邦机构中移除。美国防部长皮特·赫格塞斯做出该公司构成“供应链风险”的认定。

[20] 文艺复兴(Renaissance)是 14—16 世纪发生在意大利、后扩展至全欧洲的一场人文主义运动。它以复兴古希腊、古罗马的文化艺术为名,展开了一场以人为主体而非以神为中心的观念体系的重塑。文艺复兴运动标志着欧洲中世纪“黑暗时代”的结束,近代文明的胎动。

参考文献(References)

[1] 马克斯·韦伯. 经济与社会[M]. 阎克文, 译. 上海: 上海人民出版社, 2010.

[2] 尤瓦尔·赫拉利. 人类简史: 从动物到上帝[M]. 林俊宏, 译. 北京: 中信出版集团, 2012.

[3] 雅尼斯·瓦鲁法克斯. 云端封建时代: 串流平台与社群媒体背后的经济学[M]. 许瑞宋, 译. 台北: 卫城出版社, 2024.

[4] McKenzie W. Capital is dead: Is this something worse[M]. London: Verso Books, 2019.

[5] 塞德里克·迪朗. 技术封建主义[M]. 陈荣钢, 译. 北京: 中国人民大学出版社, 2024.

[6] 亨利·福特. 我的工作和生活: 福特自传[M]. 李伟, 译. 北京: 新世界出版社, 2010.

[7] 涉泽荣一. 论语和算盘[M]. 李政, 译. 南昌: 江西美术出版社, 2010.

[8] 速水佑次郎, 弗农·拉坦. 农业发展的国际分析[M]. 郭熙保, 张进铭, 译. 北京: 中国社会科学出版社, 2000.

[9] 尤瓦尔·赫拉利. 未来简史[M]. 林俊宏, 译. 北京: 中信出版社, 2017.

[10] 康德. 实践理性批判[M]. 邓晓芒, 译. 北京: 人民出版社, 2016.

[11] 王晨, 任向实. 从人机交互到人机共协计算: 人机关系的思想演化和未来展望[J]. 科技导报, 2024, 42(8): 6−20.

[12] 刘慈欣. 三体[M]. 重庆: 重庆出版社, 2008.

[13] 田坂广志. 死亡并不存在: 基于量子物理学的新假说[M]. 陈云, 译. 上海: 东方出版中心, 2024.

Digital governance in the AI era: How to balance "value rationality" and "instrumental rationality"?

CHEN Yun[1,2]

1. School of International Relations and Public Affairs, Fudan University, Shanghai 200433, China

2. Graduate School of International Studies, Seoul National University, Seoul 08826, Republic of Korea

Abstract The exponential iteration of artificial intelligence has unleashed tremendous productivity while also triggering a series of deep governance dilemmas such as "human alienation", technological feudalism, and employment shocks. This paper reviews the "Weberian proposition" and, using the "cognitive revolution" as a thread, examines the confrontation between two technological philosophical trends in the AI era − effective accelerationism and super alignmentism. On this basis, the paper analyzes the governance challenges of the AI era from three dimensions: modernity critique, publicness reconstruction, and systems theory prudence. It then proposes action guidelines from a humanistic perspective: upholding human subjectivity in ontology, balancing modernity anxiety with Eastern wisdom in cultural theory, transforming educational models in cognitive theory, cultivating deep reflectors in practical theory, and promoting global governance collaboration in institutional theory. Finally, it issues the "Humanistic Manifesto of the AI Era", calling for a new round of "Renaissance" to transcend the binary narrative of technological utopia and doomsday.

Keywords digital governance; Weber proposition; cognitive revolution; technocratic feudalism; humanism

(责任编辑 王微)