后人类纪文明的到来与 ChatGPT 的终极之问

汪行福，复旦大学当代国外马克思主义研究中心研究员、哲学学院教授

ChatGPT（以下简称 GPT）对人类来说，到底意味着什么？比尔·盖茨在自己的博客中写道，这是他人生中第二次被科技真正震撼到；基辛格认为这是继印刷术发明以来人类最为重要的技术发明。[①]

GPT 的潜能巨大，将在极大范围、极深程度上影响到人类当下和未来的生活，其意义可放在上下五千年文明大背景下讨论。GPT 能量巨大，我们对它的认识已呈现出两极化倾向。一些人把它视为赋能（empower）工具，“百”求必应。可以设想，如果它与互联网连接，实现实时查询，就成了 web 版的 GPT，利用自带的翻译功能，就可遍访人类所有的知识库，成为全球版的 GPT。这些不是想象，而是当下现实，是正在发生的事情。也有一些人认为 GPT 存在着致命的缺陷，就像一个黑箱，不透明，缺少反思和批判能力，与人的学习能力相差甚远，更为严重的是，它经常产生幻觉 (illusion) 和偏见（bias）。正反两方面的观察和评估给我们提出一个严肃的问题：GPT 是对人类的赋能还是人类自我罢黜（dethrone）的陷阱？人自诩为万物之灵，是天地之间唯一的有意识的动物，是能说话、有思维和推理能力的动物。如果 AI 有了类人式意识，又没有人在记忆和耐力上的生理限制，它是否会把人从“王座”上推下去，自己登上“王位”？

GPT 无疑是一个伟大的技术突破，它成功地使人工智能自然语言化，为人类创造了知识生产和运用的母体平台，这有可能彻底地改变人类文明的底层逻辑。人类的文明是以智人大脑为基础的自然语言文明。以人工智能为基础的人机一体化完全颠覆了以往人类技术发明的逻辑，其侵入人类智能领域，打破人类纪文明的边界。GPT 到底是对人类的赐福还是诅咒？这是我们需要面对的“终极之问”。

以智人大脑为基础的自然语言文明

GPT 是大型自然语言 AI 对话程序。之所以引起轰动，缘于其出色的技术优势。概括起来，即通用性、生成性和增强性。这些特点缘于 AI 的自然语言处

理能力和大数据模型的结合。这一技术革新意味着知识的获取、创造和运用的母体或平台的大迁移。

人工智能革命出现之前，科学技术已取得了一系列重大的突破和成就。一方面，人类能上天入地，现代交通工具把全球联系在一起，不仅创造了地球文明（global civilization），如果火星移民计划实现，还会出现星球文明（planet civilization），这是人类在超距文明尺度上取得的成就；另一方面，现代物理学、化学和生物学研究已进入物质的内部，深入到感觉之下的细胞核、原子、分子层面，有了改造物质微观结构的能力，这些是人类在微距文明尺度上的成就。然而，在人工智能取得决定性突破之前，人的智能、意识、思维和理性仍然是技术的禁区，是“化外之地”。在根本意义上，超距文明和微距文明影响的是人类生活的环境，没有影响到人之为人的特征，即人的意识和思维这一核

① 基辛格：《ChatGPT 预示着一场智能革命，而人类还没有准备好》，https://new.qq.com/rain/a/20230228A02K6Q00。

心领域。GPT 的决定性意义正在于它突破了这一禁区。如果将以智人的大脑结构和自然语言相结合为基础的人类文明称作人类纪文明，GPT 以及人工智能生成内容（aritificial intelligence generated content，AIGC）技术出现后的文明则可称为后人类或后人类纪文明。①

人类文明史经历过多次大的转型，从采集文明到农业文明，之后进入工业文明，再到后工业文明。每次文明变革都有其祸福相依的一面，既是对人的赋能，也意味着原有的人类活动及其形式的重组，这也不可避免地导致对收益和代价的重新分配。就此而言，GPT 与以往的文明变迁没有太多区别。然而，作为自然语言的人机对话平台，GPT 打破了智人大脑的生理限制，也突破了自然语言技术的限制。这一点具有深刻的意义。如果 GPT 和 AI 的发展去除了上帝为人类文明设定的限制，从根本上改变了人类文明的底层逻辑，那么它们的出现确实意味着人类文明史的奇点。巴别塔的倒塌意味着以智人大脑为基础的自然语言文明的到来，而 GPT 的出现则意味着以 AI 为基础的后人类文明时代的到来。

自然语言 AI 大模型意味着什么

人是语言的动物，语言的限度就是思想的限度。然而，自从现代科学诞生以来，特别是现代哲学的语言学转向以来，在语言上存在着两种范式的竞争，一种是自然语言，一种是人工语言。自然语言是文化的语言，人工语言是科学的语言。自然语言与人工语言各有其偏爱的领地。自然语言的独特领地是人文科学，它为人的丰富思想和情感表达提供了条件，也为民族认同建构提供了神圣的象征。然而，自然语言有其优点，也有其缺点，它承载着人类生活世界的背景知识，是人与人交往的中介，它的生动性、多义性为文学和艺术提供了得天独厚的条件。但是，自然语言的歧义性和含义的模糊性也给知识的表达和传播带来了消极影响。正因为如此，一些人仍梦想巴别塔之前的同一语言，用一种世界语取代五花八门的自然语言。

人工语言传统虽然很有影响力，但在人类生活中仍然是自然语言占统治地位。亚里士多德的三段论和形式逻辑中也运用了符号，但它们仅仅用于表示判断之间的逻辑关系。从莱布尼茨开始，哲学家们就孜孜以求建立一套人工语言系统，用以表达人的思维和推理，并把算术机器化。人工语言的理想直到符号逻辑和符号数理学的出现才部分得以实现，即使如此，符号逻辑的运用范围也是有限的，它始终被限制在语言表达式之间的逻辑关系层面，并没有进入思想内容本身。在人工智能取得突破性进展之前，语言的人工智能化陷入这样的困境：或者接受一个应用范围有限的可技术化处理的人工语言系统，或者延续自然语言传统但放弃对其进行人工智能化。GPT 的出现打破了这一僵局。作为自然语言人工智能系统，它不仅通用于各种自然语言，而且具有自然语言的生成性和对话性。

GPT 最大的特点是其生成性。它可以根据语言或文本提示生成新的内容，完成文本写作、绘画、编程、解码等任务。在受到大量的“喂食”和一定时间的训练后，大模型会“涌现”出各种神奇的能力，如信息搜索、历史记忆、上下文理解、推理能力、与人流畅对话等。在此之前，这些能力被认为只有人才具有，且只能通过自然语言充分实现。AI 大模型的能力涌现引起了广泛关注。能力涌现是指在无监督情况下拥有理解、对话、创作和处理复杂问题的能力。虽然我们已经有一些关于 AI 能力涌现的技术条件以及大模型参数与能力涌现之间变化轨迹的经验性结论，例如能力涌现需要多少参数和训练量，参数的增加与能力涌现之间的非线性和倒 U 形轨迹等，但我们对其产生机理所知甚少。因此，AI 专家表示：“我们对 GPT-4 的研究完全是基于现象学的，我们关注的是 GPT-4 能够

① 按通行的理解，人类纪文明自近代使用化石能源开始，因为人类活动影响到地球表面的生态和近地大气层。简要地说，人类纪文明就是工业文明。后人类文明是一个宽泛的概念，有许多含义，其断代史时间也各不相同。罗西·布拉伊多蒂在《后人类》（*The Posthuman*）中主要把它理解为后人类中心主义价值和后现代主义观念。本文中使用的后人类文明概念更加激进，即一种基于 AI 和大数据的人机一体化的文明。

做到这些令人惊讶的事情，但我们并不知道它是如何变得如此智能的。”①

对于“涌现”现象的思考在哲学中并不陌生。海德格尔认为，运用语言的思想活动与其说是受人的意识的控制，不如说是语言本身的自发运动。真正的思想不是说话者的思想，而是思想通过语言自行“道说”，而自然语言正是思想最贴近的表达方式。真正的文学创作不是“为赋新词强说愁”，而是让自己沉浸于思想中，让思想自由表达。一个人沉浸于思考之中，会文思泉涌。对此，伽达默尔说：“‘表达’并非主观选择的问题，即并非限于事实之后，并且借助于个人思想中的意义被赋予了可传达性而加于事实之上的某种东西。”②海德格尔、伽达默尔等人对思想与语言关系的论述似乎先行地道出了 GPT 的逻辑，即自行思考、自行生成和自我表达。实际上，与 GPT 的对话之所以具有这样的特征，不仅在于它能够很好地模拟人的自然语言，而且在于其本身表现出了类似于自然语言系统的特性。如果说 AI 是对人脑神经网络的模仿，那么 GPT 则是对人的自然语言对话系统的模仿，如模糊认知、解释、交互学习等。因此，从现象意义上来说 GPT 具有类人式意识并不为过。

问题在于，GPT 不仅具有类人式意识，而且具有超人式意识。依笔者所见，人类文明建立在自然语言和人脑思维之上，两者的结合既提供了人类文明发展的空间，也限制了它的发展。GPT 的出现截断了人脑机能和自然语言之间的联系，自然语言对话平台从人际间转移到了人机之间，甚至未来可能会发展到机器之间的对话。这是人类文明底层逻辑的重大变化。在超大型数据库（人类知识储存）和计算机超级算力（思维力）的加持下，语言生成模型的能力可以远远超过人类。就像计算机在计算领域取得的成就一样，AIGC 也可能在自然语言的理解、解释和生成方面拥有类似的能力。例如，AI 将人类的思考和推理等思维活动从自然人的智能平台转移到了人工智能平台，就像汽车从崎岖不平的土路走上了高速公路。在这个过程中，无数信息通过无数节点迅速连接，瞬间完成各种任务。我们不知道它是如何做到的，但它给人们呈现出一个“all-knowing machine”或“全知之神”的形象。

对 GPT 的近忧远虑

海德格尔曾区分“怕”与“畏”，认为“怕”有具体对象和缘由，可计算得失；“畏”则不同，它是对无之畏，是对存在之整体的沉沦的畏，在此，所有存在者被抛出自己的轨道，进入无根无据的深渊。面对 GPT，我们不由地生成复杂的感情，它不仅是“怕”，而且也是“畏”。

在现象层面，每一次技术创新都有复杂的伦理和社会后果。（1）像每一次技术创新和大规模运用一样，GPT 不可避免地使人际间出现收益和代价的再分配，催生一些行业，毁灭一些行业。如果我们承认，人的尊严与其劳动中体会到的自主性有关，GPT 的出现无疑会对那些被剥夺了劳动机会的人的生活造成重大损害。（2）虽然 GPT 有问必答，有极强的亲民性，但是这不等于技术的可及性，其有可能造成新的不平等，加深数字鸿沟。（3）与任何节省体力和脑力的技术工具一样，GPT 的使用中会产生依赖性风险，这在教育和学习领域影响尤为显著。GPT 可能提高人的学习效率，也可能使人因对其依赖而丧失学习动力。（4）GPT 是自然语言的人机对话程序，形似人与人的对话，但毕竟不是人与人的对话。如果交往是人与人的关系和社会融合的条件，GPT 的出现可能会导致人类出现逃避其同类的倾向，强化人的孤独感，影响到以交往为基础的人类文明的各种组织形式，如婚姻、家庭、宗教、民主政治等。（5）GPT 表面上具有亲和性，对人有求必应，其给出的回答也不乏个性，表现得就像一个知心朋友，然而大数据模型训练出来的回答到底有多少个性？它向人呈现的个性

①《是什么让 ChatGPT 变得如此聪明？仍然未知的大语言模型“能力涌现”现象》，新浪财经，2023 年 4 月 11 日。

② Hans Gadamer, *Hegel's Dialectic*, trans. by P. Christopher Smith, New Haven: Yale University Press, 1982, p.32.

化是否只是一个幻影？诸如此类问题都是现实的，也是需要不同领域的研究者去探讨的。

真正引发人们担忧的是本体论层面的问题：GPT 到底是什么？我们到底能对它做什么？前者可称为透明性问题，后者可称为可控性问题。在任何时代，完全的透明性都是不存在的。长期以来，自然界对我们就是不透明的、充满奥秘的世界，只是这种不透明在哲学家和诗人眼中是神圣的，而技术的不透明却是有罪的。收音机发明出来后，海德格尔就对那个小黑盒里面发出的声音感到恐惧。因为与风车和水车相比，收音机是“不透明”的，我们看不见它如何工作，即使拆开来也是如此。GPT 更是如此，甚至 AI 技术人员也会对它的神奇能力感到困惑，更不要说外行人了。在某种意义上，GPT 就是一个黑箱，它对人类所提出的任何问题都能瞬间给出回答，且正确率越来越高。这不能不让人感到困惑：它到底是如何做到的？按照海德格尔对“怕”和“畏”的区分，在 GPT 黑箱面前我们感到的应当是畏，即对莫名的风险的恐惧。

在本体论层面上一个更大的问题是 GPT 的可控性问题。人的思维依赖于大脑神经网络，通过感知激活大脑中存储的知识和记忆，通过反思、推理形成自己的判断和观念。在这里，自然语言和符号成为记忆和感知的中介，而 GPT 则不同，它有自己的记忆，其记忆体即数据库可对电子数据信息一网打尽，它有自己的“大脑”，即神经网络 AI 程序，它能够生成新的知识。将来，它通过电子感应器还可以感知事物，与机器人相结合直接作用于物理世界。

今天，GPT 仍然需要依赖人的“提词”（prompt），但我们可以想象一个无人的 AI 世界。黑格尔在《精神现象学》中有一个著名的命题——“实体即主体”，他把精神视为比人更高的主体，它不仅通过自身运动创造世界，而且通过自我反思认识自己。然而，黑格尔的“绝对精神”与波普尔的“世界 3”一样，它们的自主性在某种意义上仍然是拟人化的，是人类生命的投射。按通常的理解，无论它如何神奇，归根到底 GPT 是一个智能系统，是物，是实体。但是，如果 GPT 不仅具有了人的意识和思维能力的表现，而且在加速迭代中获得远超于人的能力，不仅能为人所为，而且能为人所不能为，我们还纠缠于它是否有意识、是否是主体就失去了意义。

对人类文明来说，GPT 的真正挑战在于我们是否创造出了一个能够彻底逃逸人的控制的物，无论我们把它称为人工理智还是超级智能。如果是后者，它将导致人类“王位”的废黜。阿多诺在《否定辩证法》开篇曾指出：一度似乎过时的哲学由于错过其实现时机而得以保留下来。同样，一度过时的全知全能的理想因没有实现，人类才将之保存下来。今天，我们面临的悖论是，也许 GPT 正在实现人类的全知全能幻想，成为新的“通天塔”，但它的成功会是人类的成功吗？ GPT 是人类新文明的开端，还是人类文明终结的开始？

如何避免 GPT 对文明的反噬

科幻小说家阿西莫夫（Issac Asimov）于 1942 年提出了著名的机器人三定律：一是机器人不得伤害人类，或者目睹人类遭受危险而袖手旁观；二是机器人必须服从人给予它的命令；三是在不违背第一、第二定律的情况下，机器人要尽力保护自己。GPT 是否伤害人类取决于第二定律。包括 OpenAI 团队在内的专家都没有对它的运行机制，特别是其神奇的能力涌现现象给出解释，自然也就无法保证它百分之百执行人类的命令。应对 GPT 这一既具有强大赋能能力又潜藏巨大风险的技术也许没有万全之策，但并不意味着无需对此进行思考。

当前，各“大厂”纷纷建立自己的大模型，呈现

出不可阻挡的“进步”特征。这样的情况下，如何思考人类孜孜以求的进步？是否在任何情况下都要拥抱进步？本雅明的思考也许有一定的借鉴意义。他在《历史哲学提纲》中借助保罗·克利的画作《新天使》对盲目追求进步的冲动进行了批判。在本雅明看来，人类的救赎需要对过去的沉思，特别是对过去灾难的沉思。他把这一救赎的历史意识拟人化地描述为“历史天使”。历史天使想停下脚步唤醒死者，修补破碎的世界，“可是从天堂吹来了一阵风暴，它猛烈地吹击着天使的翅膀，以致他再也无法把它们收拢。这风暴无可置疑地把天使刮向他背着的未来……这场风暴就是我们所称的进步”。[①]当然，本雅明观点难以直接运用于 GPT，但面对当前人工智能“竞赛”，本雅明式批判对我们不无启示。从时间结构上看，进步和灾难是非对称的，进步依赖于时间的连续性，而灾难总是不期而遇的，必须在火花碰到炸药之前将燃烧着的导火线切断，[②]因为“对于个人，如同群体蒙受的苦难一样，只有一个临界存在，超越了它，‘事情就不会再这样继续下去了’，这个临界点就是毁灭”。[③]今天的 GPT 是否逼近了这一临界点我们并不知道，但我们不能失去警惕。

海德格尔和马尔库塞的技术批判也有助于认识 GPT 的威胁。在他们看来，除了技术滥用和工具理性，技术本身也在对自然和人进行统治。统治的目的不是后来补充进去的，也不是人从外面强加给它的，而是早就进入了技术的设计之中。海德格尔把技术理解为存在的解蔽，即存在向人类敞开的方式。然而，与艺术等其他敞开形式不同，“在现代技术中起支配作用的解蔽乃是一种促逼（herausfordern），此种促逼向自然提出蛮横要求，要求自然提供能够被开采和贮藏的能量”。[④]马尔库塞也认为，现代技术不仅是生产工艺学，而且是统治工艺学。就此而言，不仅要批判技术中立化教条，而且在任何重大技术创新的源头和开始处就必须认清存在何种危险。

技术批判固然重要，但更重要的是在新技术出现时寻找到作出合理判断和采取行动的机会。在安德鲁·芬伯格看来，“技术不是一种天命，而是斗争的舞台。技术是社会的战场，或者用一种更好的隐喻来说，把技术比作一个文明的替代形式相互竞争的‘事态的议会’”。[⑤]虽然技术发明者可以设定技术的目标和价值标准，但技术使用者和接受者也不完全是待宰的羔羊。底层人民可以在流动的边缘上与技术发明者和主导者进行缠斗，改变技术的具体形态和使用方式，实现技术的民主化。技术的民主化简单地说，就是让处在边缘的弱势群体和受制者的利益得到体现，使技术的多种潜能得以呈现，让他们在技术面前有更多的自主性。技术民主化可以以多种方式展开，如不同利益群体之间的辩论，公众参与技术设计，对技术进行再发明和改造。技术的冲突实际上是两种“自主性”的冲突，一边是技术系统本身的操作自主性，另一边是技术运用者实现自身自律的操作自主性。现代技术体制的主要缺陷是那些因技术而受到影响的人对设计和运行很少或没有控制权。提高人类的自主性涉及技术的规划、建造和控制过程尽可能向那些注定要体验最终产品和社会后果的人敞开。芬伯格的技术民主化方案对我们思考如何应对 GPT 有重要的借鉴意义，虽然这一方案需要结合 GPT 技术的特点和当下社会条件转译为合适的斗争战略和策略。

本雅明和芬伯格代表着两个极端，本雅明为我们提供的是最终决断，芬伯格提供的是技术斗争的日常政治。两种立场并非决然对立，具体的选择取决于对事情本身的判断。

[本文系国家社科基金重大项目“复杂现代性与中国发展之道”（15ZBD013）的阶段性成果。]

① 汉娜·阿伦特编：《启迪：本雅明文选》，张旭东译，北京：生活·读书·新知三联书店，2008 年，第 270 页。

②③④ 本雅明：《单行街》，南京：凤凰出版集团、江苏人民出版社，2006 年，第 91 页，第 91 页，第 932 页。

⑤ 安德鲁·芬伯格：《技术批判理论》，韩连庆、曹观法译，北京：北京大学出版社，2005 年，第 16 页。