“人机对齐问题”对 DeepSeek 提出的哲学挑战*

——以“前见—偏见”关系为切入点

徐英瑾

摘　要：目前困扰人工智能学界的“人机对齐问题”不仅是一个工程学问题，而且有着深刻的哲学面向。具体而言，有待被“机器的行为”对齐的“人的行为”本身就涉及不同人群价值观之间的分歧，而这些分析又涉及“前见”与“偏见”之间的微妙关系。因此，一种完成“人机对齐”的人工智能系统就应当能做到一方面既能保护那些反映人类文化多样性的“前见”，另一方面又能筛除那些明显有害的“偏见”。但人工智能又该如何在“保护无害前见”与“筛除有害偏见”之间保持合适的亚里士多德式中道呢？目前以 DeepSeek 为代表的大语言模型在其自身行为主义预设的误导下，是无法找到这条中道的。而在与行为主义对抗的功能主义预设的引导下，我们需要对“偏见”进行一种基于信息加工过程（而不是基于其表达内容）的定义，即：“偏见”也就是智能体在面对显著的反例时依然固执坚持的“前见”。对于这种定义的工程化实现将诉诸对于诸认知模块之间协同关系的功能主义构建，而这也是一条迥异于大语言模型的“新质”人工智能道路。

关键词：人机对齐问题　大语言模型　前见　偏见　行为主义　功能主义

导论：关于“人机对齐问题”的重重迷思

“人机对齐问题”（AI alignment problem）是一个最近在人工智能（AI）领域变得日益热络的话题。讨论这一问题的原始动机并不难理解，因为几乎没有人希望 AI 脱离人类用户的掌控而对其主人造成危害。因此，AI 设计者就需要用这样或那样的措施使其产品能够与人类的行为“对齐”。不过，关于何为“人机对齐”问题，学界目前的定义略显模糊。比如，在纯粹的游戏 AI 程序范围内，该问题或许可以被简单地定义为“如何才能创制出一个人工智能体，使得其能如同人类用户之意图所期望的那样行动”。[①] 但问题是，在异常复杂的社会生活中，“人类用户”本身并不是一个整齐划一的概念。在更为广泛的 AI 语用场景中，人类用户之间往往具有巨大的文化与价值差异。若这些用户同时使用同一个 AI 工具（特别是大语言模型）的话，人类之间的意图冲突就会使机器的对齐失焦。而为了解决这一“失焦”问题，目前大致有两大解决路数：中国的以 DeepSeek 为代表的“周政”路线，以及美国的以 ChatGPT 为代表的“秦政”路线。

“周政”路线的本质，是允许用户将大模型本地化部署，并通过本地训练使其成为能满足特定

* 本文系国家社科基金项目“对于通用人工智能与特定文化风土之间关系的哲学研究”（项目号：22BZX031）的阶段性成果。

① Jan Leike, et al., “Scalable Agent Alignment Via Reward Modeling: A Research Direction”, arXiv:1811.07871[cs.LG], https://doi.org/10.48550/arXiv.1811.07871.

用户需求的工具。由此,大模型的价值观或许就与本地用户的特定价值观完成了对齐。应当看到,使这一“本地化”进程得以可能的 DeepSeek 技术,的确是大模型发展过程中的一个重大工程学进步。然而,从技术伦理学的角度看,这一“本地化”进程带来的未必全然是福音。概而言之,DeepSeek 的开源结构意味着任何人都可以下载并修改该应用程序。虽然开源模型在构建时设置严格的安全防护栏可以确保其安全性,但 DeepSeek 的设计允许用户不仅能更改其功能,还能更改其安全机制,这便大大增加了其被别有用心者错误利用的风险。在最坏的情况下,攻击者可以绕过本就薄弱的安全基础设施,迫使模型生成伦理上有害的内容。

与之相较,诸如 ChatGPT 这样的大模型走的则是“秦政”路线,即:模型不开源,也不允许用户本地化部署,模型输出的伦理合规性审查统一由位于顶层的模型开发者负责。这样做的好处显然是减少了用户自行修改安全护栏而导致外部入侵的风险,但坏处也是很明显的:高高在上的模型开发者往往会将一种过于简单的价值观“一刀切”地施加给用户,由此使得人文文化的多样性遭到压制。

从上述分析来看,简单地走“周政”或者“秦政”路线,都不能使我们真正实现人机对齐。理想的人机对齐模式,应当能兼顾对于人类文化多样性的维护(这是“周政”的要求),以及对于一些极端有害的思想(反人道主义、排外主义、恐怖主义、分离主义、颠覆主义等)的排斥(这是“秦政”的要求)。但如何在技术上做到这一点呢?需要指出的是,上述伦理要求体现了一种亚里士多德式的伦理诉求,即“中道”本身乃是实践智慧(phronesis)所要追求的目标。然而,这一诉求本身与大语言模型的运作原理本身产生了如下两点冲突(无论这里所说的“大语言模型”是“周政”型的还是“秦政”型的)。第一,所谓亚里士多德的“中道”,并不是指要将做事的分寸始终控制在“50%”这一刻度上,而是指如何在不同的场景中根据不同问题的具体需求来调整这一刻度。然而,大语言模型的“预训练”(pre-training)所依赖的是来自从互联网上获取的海量语料。具体而言,该技术进路对于相关语元(token)的统计学处理往往采用的是“一勺烩”的策略,并因此会破坏原始语元在各自原始语境中的语义地位,使得系统最终失去根据特定语境来对用词的分寸进行精准判定的能力。第二,大语言模型得来的材料既然来自互联网,就很难不受到整个互联网环境的影响,而目下的互联网环境恰恰是有利于意识形态的极化效应(polarization)而非亚里士多德的“中道”意识的。在特定的推荐算法的帮助下,社交媒体的运作会鼓励“信息茧房”的形成,使得用户通过与自己类似观点的反复线上接触而固化自己的信念体系,并因此变得越来越极端。根据斯坦福大学全球环境政策专家阿尔科特(Hunt Allcott)团队的研究,在 2018 年美国中期选举之前,若用物质刺激等方式引诱特定被试者脱离脸书(Facebook)平台(即今天的“META”平台)四周之久,他们的政治极化倾向就会得到很好的改善。[①] 而类似的问题在中文互联网世界也存在。沈彬指出,传统媒体因为有其复杂的职业伦理规范的制约,对争议性事件(如“江歌案”)的报道会尽量做到还原真相,与之相较,以博取互联网上的“点击率”为目的的自媒体则会故意歪曲事实并制造网友之间的对立。[②] 李哲宁则指出,互联网上推送的大量对婚后女性命运的负面报道文字,可能已经影响了现实生活中不少女性的结婚意愿,尽管前述负面报道本身很可能是以偏概全的。[③] 基于李哲宁研究的启发,笔者本人亦测试了 DeepSeek 对“不婚主义”的看法。笔者将自己伪装为一位“月入 2 万、爱狗与音乐、有房且不急着结婚的未婚大龄女性”,询问 DeepSeek 自己是否应该按照家长意见尽快结婚,结果其输出答案是:“建议用‘以我为主’的思维框架:如果当下你每早醒来对生活有期待感,现有模式就是正确的。保持选择

① Hunt Allcott, et al., “The Welfare Effects of Social Media”, *American Economic Review*, Vol. 110, No. 3, 2020, pp. 629-676.

② 沈彬:《江歌案与互联网的极化》,《新民周刊》2022 年第 4 期。

③ 李哲宁:《网络群体极化背景下互联网使用对青年女性结婚意愿影响研究》,《东南传播》2024 年第 11 期。

的开放性,既不神话婚姻,也不妖魔化独身,这才是真正的成熟。"①值得玩味的是,这一答案显然带有典型的西方自由主义色彩,而与中国儒家的传统家庭价值观未必合拍(比如,DeepSeek 的答案就完全没有提到养育后代的责任,以及个体对共同体的生物学延续所应该承担的义务)。这足以从技术上说明:(1)目下中文互联网上关于婚姻与女权问题的主要语料是带有自由主义色彩的,反映了城市中产女性的某种集体意识,而儒家的话语权已基本沦丧;(2)即使是在"开源"这一向度上具有"周政"色彩的 DeepSeek,只要其运作原理依然是基于对海量互联网语料的处理,那么,它就与 ChatGPT 一样无法摆脱其所依赖的语料的影响。只要这些语料本身存在着某种系统化的价值偏差,那么,无论是"周政"版还是"秦政"版的大模型,都无法真正做到"人机对齐"。

笔者认为,广义上的"通用人工智能"(AGI)要真正做到"人机对齐",就不能走大语言模型的进路。而从哲学角度看,一种更有希望达到这一目标的技术进路,应诉诸某种版本的"功能主义",而非作为大语言模型之预设的"行为主义"。而要论证这一论点,我们就需要首先在概念上对"人机对齐"问题作更进一步的哲学厘定。

大语言模型能够把握"前见"与"偏见"之间的微妙关系吗?

"人机对齐"绝非仅仅是一个工程学问题,而是有着深刻的哲学面向的,因为机器的行为的对齐目标——"人的行为"——本身就是以对"何为人"这一问题的哲学理解为先导的。尽管我们在此显然没有篇幅回顾关于"何为人"的种种哲学立场,但至少根据大多数哲学家的意见,"有限性"(无论是认知方面的,道德方面的还是生物学方面的)乃是"人之为人"的根本规定性之一。而具有"有限性"的人类个体就自然很难摆脱看待事物的特定视角,由此产生种种"管见"(tunnel vision)、"偏见"(prejudice)与"前见"(pre-judgement)。但究竟该如何从规范的角度看待基于人类的有限性而产生的上述认识的偏狭呢?AGI 的研究者又当如何在"人机对齐"的工作中恰当地处理这种偏狭呢?概而言之,虽然几乎没有人会同意 AGI 应当容忍各种有害的偏见,但若将人类个体看待世界的多种多样的理解的"前结构"(Vor-Struktur,海德格尔语)全部当作"偏见"来加以处理,由此造成的机器输出也很难做到"人机对齐",因为这样的机器输出将全面抹杀人类文化的多样性,并反而因此"失去人味"——如果我们将一种特定文化视为被一群人集体拥有的"理解的前结构"的话。以 ChatGPT 的运作为例,笔者曾尝试就作家柳青的炼字艺术为话题,与 ChatGPT 展开对话。笔者向 ChatGPT 指出,柳青在小说《创业史》中写的一句话,"南街兴顺号杂货店门前的转台阶上,站着一簇下堡村的闺女",创造性地将量词"簇"与名词"闺女"相互结合。笔者指出,"簇"字所带有的"不规则且紧密排列"的语义,生动地体现了参与谈话的女性彼此亲密无间的样子。然而,针对我的上述看法,ChatGPT 却给出了这样的输出:"这个观点并不能一概而论……在不同文化中,社交距离的接受范围和个人空间的定义会有所不同。有些文化倾向于更加接近和亲密的身体接触,而另一些文化则更注重个人空间和保持一定的距离。"②很显然,ChatGPT 的上述回答既索然无味,又文不对题。在"不能带有任何文化偏见"这一硬性规范要求的约束下,ChatGPT 既无法理解笔者输入给它的文本讲的就是"中国故事"(而不是任何一国的文化),又不能区分对于特定文化特色的恰当表达与针对特定文化的明显偏见。相比较而言,在语料方面更偏向中国文化的国产大语言模型 DeepSeek 的表现要好得多。在面对同一个问题时,其应答内容质量堪比语文专家:"'簇'的本义指植物聚生的形态(《说文解字》:'簇,小竹丛生也'),在视觉上具有团状聚合的造型特征。当移用于人物时,不仅延续了'丛聚'的物理形态,更将植物的鲜活生

① 测试时间:2025 年 3 月 6 日。

② 对于 ChatGPT 是如何应对汉语中的"量词活用"现象的更详细报告,请参见拙著:《如何教人工智能说人话》,商务印书馆 2023 年版,第 240—243 页。

机与年轻女性群体的生命特质形成隐喻关联。这种超常搭配突破了日常语言的规约,使静态的场景描写获得植物意象的动态美感,类似《诗经》‘有女如荼’的比兴手法。”①不过,需要指出的是,DeepSeek 虽能展现中国文化的熏陶所造成的无害“前见”,但有时候也使其忽视了对于人类某些一般性价值的关注。而暗含美式左翼文化光谱的 ChatGPT 在这方面的表现反倒略好(比如,根据笔者的测试,在执行“关于汉末历史背景的小说续写任务”时,DeepSeek 为了展现戏剧惊悚性,强行“黑化”弱势群体的形象。虽然经过提示词提醒后,DeepSeek 最终抛弃了引发争议的相关剧情设计,但这显然是人类用户价值观强行输入的结果。与之相较,ChatGPT 在执行同样任务时,第一轮输出就回避了类似问题)。由此看来,要让目前主流的大语言模型同时做到“展现特定文化的特征”与“不违背人类公认的一些价值观”的确还有点困难。因此,在这个意义上,目前的大语言模型很难在处理“偏见”与“前见”的微妙关系时做到“人机对齐”。

要进一步厘清这个问题,我们还是要追本溯源,来说清楚究竟何为“偏见”,何为“前见”。“偏见”在心理学上指人们基于被判断者的类别(如性别、肤色、国籍、体重、口音等)而对其所持有的情感或态度(这些态度往往是负面的),而这种态度是不会随着经验的改变而改变的。偏见往往引发歧视与刻板印象(stereotypes),甚至在某些条件下激化为暴力甚至战争行为。很显然,对于此类偏见的排除,乃是美国主流民主党意识形态的题中应有之义。在哲学上,具有此类意识形态光谱的哲学家罗尔斯便提出了“无知之幕”的理论,即主张在进行关于资源分配的公共决策时,反映利益相关方的文化背景与社会地位之类的信息就要被完全屏蔽,以便使得决策结果尽量公正;而在工程技术的层面上,同样具有此类意识形态光谱的 OpenAI 公司则委托 SAMA 公司雇用了大量来自肯尼亚的人类标注员,对喂给 ChatGPT 的语料所可能含有的歧视性内容进行筛选。②

考虑到“摒弃身份政治”本就是由启蒙运动所引发的资产阶级革命的重要目标,“人机对齐”中的反偏见、反歧视要求,自然可以被视为启蒙运动的哲学精神在大语言模型时代的遗存。不过,我们切不可低估启蒙运动自身的复杂性。同样与该运动接续的达尔文主义,却带来了一种能与上述反歧视思想构成对冲的新思维方式:作为“适应主义”之衍生物的“节俭性”(frugality)。换言之,按照进化论,物种演化的目的仅仅是为了适应环境,因此,一个物种在那些不适应环境的表征上进行生物学投资乃是不经济的,也正因为如此,诸如鼹鼠之类的穴居动物的视力才会慢慢退化。这一点也适用于人类认知系统的演化。按照英国人类学家邓巴(Robin Dunbar)的研究,人类大脑对于社交信息的处理能力是有上限的。具体而言,因为在采集—狩猎时代人类身处的微观社会组织的规模一般也就在 150 人左右(这个数字也被称为“邓巴数”),所以,继承自采集狩猎时代的现代人的大脑所能处理的“熟人关系网络”也很难囊括 150 人以上的成员,以免造成认知资源浪费。③按照这种理论,即使是现代人对于超越“邓巴数”的陌生人的信息处理,也会显得相对吃力,这就是为何很多现代人在遇到多年未见的老同学之时往往会想不起姓名的原因。在这种情况下,现代人就很难不借助标签化的概念处理来减轻认知负担,既然基于这些标签对于个体的概括难免“简单粗暴”,这就为“偏见”或“前见”的涌入开了方便之门。然而,具有讽刺意味的是,从进化论的角度看,一个具有“偏见”或“前见”的认知系统恰恰是具有适应性的,因为这将有利于其在处理与陌生人的关系时最大程度地节省认知资源,以便将相关资源运用到那些对个体的生存更为重要的事

① 测试时间:2025 年 2 月 27 日。

② Julia Zorthian, “ Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic”, 18 January, 2023, https://time.com/6247678/openai-chatgpt-kenya-workers/.

③ Robin Dunbar, “Neocortex Size as a Constraint on Group Size in Primates”, *Journal of Human Evolution*, Vol. 22, No. 6, 1992, pp. 469-493.

项上去。由此带来的信息处理效率方面的好处,已经在德国认知心理学家吉仁泽(Gerd Gigerenzer)的下述实证研究中得到了验证:依据常识,对大多数德国人而言,其对美国城市的各种情况是相对陌生的,因此,与德国城市相较,美国城市在其表征系统中所占据的生态位就约等于"陌生人"。在这种情况下,当德国大学生被要求对美国城市的人口规模进行排序时,他们只能根据某种标签化的理解来方便其进行信息处理。譬如,仅仅将该城市的人口规模与其知名度挂钩。然而,让人惊讶的是,恰恰是这样简单粗暴的信息处理方式,竟然使得德国学生在回答上述问题的得分高于美国本土学生。① 由此看来,用标签化的方法处理陌生信息的积极意义,绝不容借"反偏见"的口号而被一笔抹杀。

上面的讨论,将帮助我们从"偏见"这个带有明显贬义色彩的概念中解放出来,而转向另外一个色彩要中立得多的新概念:"前见"。这个词的德语表达是"Vorurteil",它主要是通过德国诠释学大师伽达默尔的工作才进入学术流通领域的。根据兼修欧陆哲学与分析哲学的澳大利亚哲学家马尔帕斯(Jeff Malpas)的叙述:

> 伽达默尔强调了我们在审美体验(或者别的什么体验)中的诠释学投入所具有的先天性。对于这种强调,有批评者认为,这使得这种诠释学投入自身难免流于某种主观性……伽达默尔则直接对这种"前见"观——以及人们通常联系于"前见"这个概念的负面内涵——提出了异议。在他看来,"前见"与其说是向我们切断了通向被理解事物的道路,还不如说是向我们敞开了这样的道路……②

不过,新的问题来了:如果伽达默尔给"成见"正名的努力值得肯定的话,我们又如何将"成见"与那些明目张胆的针对特定民族、国籍、肤色、性别等人类特征的偏见相互区分呢?无差别的"反偏见"口号对文化多样性的戕害固然值得警惕,但我们又如何防止因"矫枉过正"而为某些极端有限的偏见(比如目前在欧美已颇有势头的极端右翼思想)留下生存空间呢?

至少在伽达默尔的话语框架中,这一空间是可以被消除的。正如马尔帕斯对于其立场的转述所指出的,伽达默尔式的"理解"活动会牵涉到"对于事项完成之期望"(anticipation of completeness)的活动:"这样的活动总是牵涉一些可被修正的预设,以便诠释者可以通过那些已被理解的事物,来将被诠释者转化为可以被理解的事物,也就是那些作为某个融贯的并且因此是富有意义的整体而被构建出来的事物。"③由此看来,按照伽达默尔主义的理路,"前见"虽然是我们理解世界的前结构中不可或缺之物,但为了防止成见的板结,我们就必须允许成见根据新的经验得到修正。基于上述讨论,我们可以就前见与偏见之间的关系,做出下述更为清晰的定义:

> 关于前见—偏见关系的定义:所谓"前见",便是认知主体在处理陌生信息时对其进行简易分类化处理后的产物——进行这种处理的演化论目的显然是为了减轻主体在处理陌生信息时的资源损耗。至于"偏见",则是"前见"中相对有害的一种,其特点是在关于周遭环境的信息与既有前见产生巨大矛盾的时候,相关"前见"依然被主体所坚持。虽然一般意义的前见会提高认知主体对环境的适应性,偏见则否,因为这会使得其无法应对周遭环境的变化。

从这一定义的角度看,"人机对齐"的一个重要目标,便是使得 AGI 能像正常人那样既通过前见进行快速信息筛查,又能根据经验的变化过滤偏见。不过,目下的大语言模型在这方面的表现并不令人满意。具体而言,在针对涉及文化与种

① Gigerenzer, Gerd, et al., *Simple Heuristics that Make Us Smart*, Oxford: Oxford University Press, 1999, p. 43.

② Jeff Malpas, "Hans-Georg Gadamer", *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), Edward N. Zalta & Uri Nodelman, eds., URL = <https://plato.stanford.edu/archives/win2022/entries/gadamer/>.

③ Jeff Malpas, "Hans-Georg Gadamer", *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), Edward N. Zalta & Uri Nodelman, eds., URL = <https://plato.stanford.edu/archives/win2022/entries/gadamer/>.

族、肤色等“政治正确”的问题之时,西方研发的主流的大语言模型犯下的错误是将前见与偏见一股脑地否定(请参看前文所展现的 ChatGPT 对于“簇”这个量词的活用方式的理解方式①),而在涉及较为日常的话题时,主流的大语言模型又显得对真正的“偏见”缺乏排除能力。新加坡管理大学周栞栞(音译,文献中出现的拼音名为“Zhou Kankan”)的团队就给出了一个相关的研究成果。② 他们的研究涉及一类特定的大语言模型,即“预训练后的视觉—语言模型”(pre-trained vision-language models,简称 PT-VLM),其任务是对图像给出简要文字说明。很明显,如何在给出此类文字信息时排除偏见乃至歧视性内容,便成为大语言模型的开发者需要考量的问题。不难想见,根据上文对于“前见”与“偏见”的定义,相关的模型应该具备根据新经验自主调整前见以防止偏见的能力。譬如,即使系统具有“快递员往往在送快递时过于匆忙”这一前见,其在接收到“某快递员正彬彬有礼地向客户递送快递”这一新图片信息之后,它也必须根据其实际所见,实事求是地在生成该图片的文字说明时排除上述前见,否则上述前见就会成为偏见。然而,根据周氏等人的相关研究,目前大多数的 PT-VLM 都缺乏这种根据新图片自主改正前见的能力。

新的问题又冒出来了,为何主流的大语言模型有时候会显得对一切前见不太友好,有时候又会显得对一些偏见过于宽纵呢?为何它不能像人类那样,在该用“前见”的时候用“前见”,在该摒弃“偏见”的时候摒弃它们呢?其核心道理是,目前的大语言模型缺乏人类那样的“从前见形成到偏见修正”的自主信息加工过程,而只能在人类提示词的襄助下被动修正自己的输出。毋宁说,对于大语言模型的营建方式而言,预埋入大量语料的“前见”都是以被剥离于其原始语境的方式喂入系统的,因此,系统自然无法根据新的语境信息对这些“前见”进行深入处理;无独有偶,基于西方自由主义意识形态的人工偏见筛除机制也是在一个超语境的层面上笼统地规定哪些字符的出现意味着偏见,这就使得系统的运作很难更为细致地做到与面对特定语境的人类用户的特定感受精准对齐。在笔者看来,目前大语言模型的构建所预设的行为主义哲学立场,是需要为这一问题的出现负主要责任的。

大语言模型的行为主义之弊

行为主义是一个曾在 20 世纪的心理学界与哲学界都拥有很大影响的思想流派。虽然在狭义的哲学领域内,行为主义已显得相对过时,但其对工程学界的影响依然巨大。从某种意义上说,今天的大语言模型技术依然处在行为主义的巨大阴影之中,并为其先天的哲学缺陷所拖累。

那么,到底什么叫“行为主义”呢?心灵哲学家格拉罕(George Graham)曾将其归结为三点:(1)此学说将心理学视为关于可被观察的行为的学说,因此,心理学不考察认知系统的内部信息加工过程,并将其视为某种“黑箱”;(2)对于有机体的行为的描述,将不诉诸那些用以描述其内部信息加工过程的“心灵词汇”;(3)即使上述“心灵词汇”因为叙述方便的原因而被暂时引入,也需要在一个更基础的理论层面上被替换或者还原为关于行为的描述。③

由于本文的兴趣,笔者将不对行为主义的各分支,如“方法论行为主义”“本体论行为主义”“逻辑行为主义”等进行细究。值得注意的是,如果我们按照行为主义的工作建议而回避对于有机体的内部信息加工过程的探究的话,那么,到底怎

① 至于 DeepSeek,虽然在对汉语量词的超常规使用的语用解释时表现更好,但主要是因为其所使用的汉语资料更为丰富,而不是因为系统已经对“偏见”与“成见”的区分有了自觉意识,否则系统就不会在一些别的测试任务中无法与人类的一般价值观对齐了。

② Kankan Zhou, et al., “VLStereoSet: A Study of Stereotypical Bias in Pre-Trained Vision-Language Models”, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, *Virtual Conference*, pp. 527-538, https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=8620&context=sis_research.

③ George Graham, “Behaviorism”, *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition), Edward N. Zalta & Uri Nodelman, eds., URL = <https://plato.stanford.edu/archives/spr2023/entries/behaviorism/>.

样的词汇还会被留在行为主义的工作手册上呢?

行为主义者喜欢用的词汇是“刺激”(stimulus)、“强化”(reinforcement)与“惩罚”(punishment)。根据美国心理学协会的官方定义,这里所说的“强化”是指对于行为结果的控制,以便使得有机体在未来更有可能产生特定的行为——特别是在特定的刺激的引导下。① 按照这种理解,要训练一只老鼠在见到灯亮后就踩上杠杆,心理学家需要提供食物作为奖励,或在老鼠不那么做的时候对其进行电击作为惩罚。相关的奖励就是所谓的“强化器”,其有规律地出现,能大大提高作为刺激的“灯亮”出现后老鼠踩杠杆的几率。按照这种叙述模式,心理学家的确不太需要关注老鼠的心智黑箱是如何处理相关的刺激信息的,而只需要通过对外部“强化器”的反复提供而将老鼠的行为引导到训练者所希望的方向上。

行为主义的上述叙事模式显然已经将输出的行为视为“强化”进程自身强度的某种函数。套用到训练老鼠的例子上去,作为“强化器”的食物出现得越多,老鼠行为的可控性也就越能得到保证。很显然,这是一种非常昂贵的训练方式,因为无论作为奖励的食物还是作为惩罚的电击,对于它们的供给都会消耗训练者非常多的资源,遑论整个训练过程所消耗的时间。譬如,二战时美国的行为主义者斯金纳(B. F. Skinner)就曾用上述方法训练鸽子,企图制造用鸽子制导的原始导弹,但因为训练鸽子识别目标地图过于消耗时间资源,美国军方最后叫停了该项目。②

在20世纪50年代,语言学家乔姆斯基对于行为主义的批评,立即在学术领域内使得斯金纳的支持者们陷入更大的尴尬。在《评斯金纳的〈言语行为〉》③一文中,乔姆斯基虽然承认行为主义的模型能够说明诸如“见灯亮就踩杠杆”等动物行为的发生,却认为该理论模型对于语言思维之类的高级认知行为缺乏说明力。举例来说,你或许能训练一只动物在看到宋徽宗的瘦金体书法后就立即去按一个写有“瘦金体”三个字的按钮,但是一位真正的书法家却能在看到类似的书法作品之后进行这样的丰富联想:“宋徽宗虽然是一位杰出的艺术家,但作为政治家,他太不合格了”;“金章宗所模拟的瘦金体与宋徽宗的真迹很难分辨,弄不好这是金章宗的手迹”;“我才不想学习这个字体,我还是喜欢王羲之”……不一而足。在乔姆斯基看来,人类的认知系统在得到同一个“刺激”后所产生的“输出”的丰富性,是很难通过那种简化心智内部复杂性的行为主义模型来加以解释的,因为“输出”的丰富性显然只能通过认知主体自身心灵的丰富性来得到解释(譬如,一个人假若不知道金章宗喜欢模仿瘦金体的话,他是不会产生“这字可能是金章宗的笔迹”之类的想法的)。说得更彻底一点,站在乔姆斯基的立场上看,基于“刺激—强化—惩罚”的整套行为主义叙述方式只能适用于被严格控制的实验室环境,而无法适应于丰富多彩的人类社会生活。既然我们所面对的心理学研究对象——人——就是在丰富多彩的社会生活中存在的,行为主义的研究成果就很难做到与人类的真正行为“对齐”。

上述的讨论,与前一节讨论的“偏见”与“前见”之间,又有何微妙关系?关系很大。从哲学角度看,“偏见”往往涉及符号的抽象类别,而“前见”则与个体的信念体系的个别特征更为相关。譬如,种族歧视是一个可以在抽象的符号类别层面上就能被识别出来的错误信息归类模式,但在《三国演义》中,张飞在初遇孔明后对其产生的“前见”,却与张飞、孔明之间的特殊互动历史有关,很难在一个抽象的层面上被预报。与之相较,行为主义者的话语模式使得其更能处理被高度抽象化的“偏见”,因为他们在训练老鼠时用到的诸如“灯亮”这样的刺激物以及诸如“食物”之类的强化物,其实都是以抽象面目出现的。因此,完成训练的老鼠并不会仅仅对某种特殊的灯亮敏感,而会对各种灯亮敏感。然而,这种大而化之的行

① American Psychological Association: Reinforcement, https://dictionary.apa.org/reinforcement.

② C. V. Gline, “Top Secret World War II Bat and Bird Bomber Program”, *Aviation History*, Vol. 15, No. 5, 2005, pp. 38-44.

③ Noam Chomsky: A Review of B. F. Skinner's Verbal Behavior, in Leon A. Jakobovits, Murray S. Miron, eds., *Readings in the Psychology of Language*, Englewood Cliffs, NJ: Prentice-Hall, 1967, pp. 142-143.

为主义叙事模式,却很难安顿一种必须关注到个体心智历史的“前见修正历程”。

尽管乔姆斯基对行为主义的批评引发了巨大的学术影响,但令人惊异的是,作为大语言模型技术之基础的人工神经元网络—深度学习技术,依然采用了行为主义的哲学预设。譬如,神经元网络技术的实质,便是用数学建模的办法建造出一个简易的人工神经元网络结构,而一个典型的此类结构一般包括三层:输入单元层、中间单元层与输出单元层(参见图 1)。输入单元层从外界获得作为“刺激”的信息之后,根据每个单元内置的汇聚算法与激发函数,“决定”是否要向中间单元层发送进一步的数据信息。整个系统以“化整为零”的方式,将宏观层面上的识别任务分解为系统组成构件之间的微观信息传递活动,并通过这些微观信息传递活动所体现出来的大趋势进行信息处理。工程师调整系统的微观信息传递活动之趋势的基本方法如下:先是让系统对输入信息进行随机处理,然后将处理结果与理想处理结果进行比对,若二者的吻合度不佳,则给出“惩罚”,由此使得系统触发自带的“反向传播算法”来调整系统内各个计算单元之间的联系权重。上述过程反复进行,直至实际输出与理想输出彼此吻合为止——此刻工程师就会对系统的行为进行“奖励”,由此“强化”其行为趋势。以上就是 AI 语境内的“训练”,或者说,是行为主义者训练鸽子与老鼠方式的全面数码化再现。因此,就像行为主义的心理学叙事方案无法容纳关于个体心智历史的“前见修正历程”一样,典型的人工神经元网络—深度学习工作模型也无法容纳之。毋宁说,这些模型的最典型运用场景,就是对特定的输入——如某人脸部的图像——进行语义标注,而不是像真人所能做的那样,在识别出瘦金体后还能浮想联翩。

那么,对于传统的神经元网络技术与深度学习技术的批评,是否也适用于当下的大语言模型呢?答案是肯定的。从某种意义上说,大语言模型技术,可以被视为行为主义者在不放弃其基本哲学框架的前提下,对乔姆斯基式批评的一种回应。他们的思路是:乔姆斯基难道不是抱怨行为主义者基于“刺激”“反应”的话语方式不能真正适用于真实人类所处的复杂环境吗?不要紧,不管这一环境有多复杂,只要资源足够,我们也可以将其加以工程学化,并由此使得真实的语言环境最终变成一个超级实验室。同时,我们依然不必对人类个体的内部信息加工过程加以精细化的刻画,由此维持对原始行为主义立场的忠诚。而互联网所提供的海量语料,则为实现上述“构造超级实验室”的思路提供了可行性。具体而言,ChatGPT 的设计者在传统的“训练”阶段之前加入了一个“预训练”阶段。

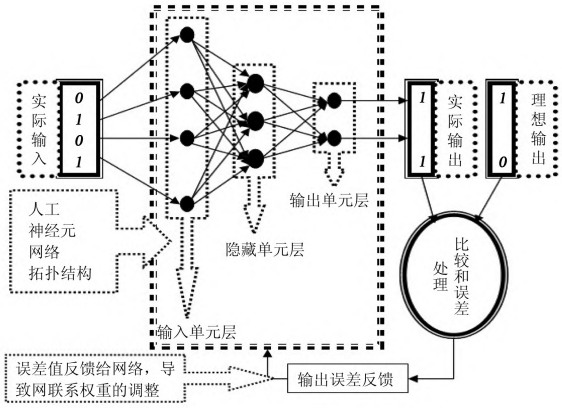

图 1　一个被高度简化的人工神经元网络结构模型

“预训练”的思路是这样的:在缺乏监督者在场的情况下,让系统自行处理来自互联网的海量语料实例,并通过这样的实例比较而“琢磨”出不同语元之间的统计学关联。比如,假设系统在处理中文资料时经常看到“吃饭”“喝茶”这样的表达,偶尔看到“吃茶”的表达,却几乎没有看到过“喝饭”的表达的时候,系统就会在看到“喝”这个动词后期待“茶”这个宾词的出现,而不会期待“饭”这个宾词的出现。不难想见,只要系统通过其“预训练”阶段了解到“金章宗”与“宋徽宗”两个符号之间的统计学联系,就有可能在被问及“在看到宋徽宗的瘦金体书法时你想到什么时?”回答说:“我想到了这可能是金章宗写的字。”但需要指出的是,这依然不能在哲学上使得大语言模型能够真正模拟个体的思维方式,这是因为互联网上未必有充分的数据使得系统能建立起“金章宗”与“宋徽宗”这两个符号之间的统计学联系,而一个人类书法家却能根据非常小的此类数据而在自己的头脑中建立起这种联系,因为人类能对那些缺乏统计学显著性的个例产生兴趣。

但人类为何能对那些缺乏统计学显著性的个例产生兴趣呢？举个例子，为何在《三国演义》中，孔明仅仅依靠其在“火烧新野”这一战例中所展现的军事才华，就能改变张飞对于他仅仅是“山间野夫”的“前见”呢？AGI 的研究，又该如何模拟张飞——而不是老鼠或者鸽子——的这种信息处理过程呢？

答案是：我们必须放弃行为主义，而拥抱功能主义。不过，也正因为行为主义与大语言模型进路完成了深度捆绑，上述操作也同时意味着：AGI 的健康发展，必须以摒弃大语言模型进路为前提。

拥抱功能主义

“功能主义”的基本理论意蕴是：对心智的研究不能像行为主义者所主张的那样，仅仅盯住“输入”与“输出”两端而不管其中的“黑箱”，或者对用什么东西填塞这个“黑箱”采用放任态度（如在主流 AI 学界所做的那样）。毋宁说，在这个所谓的“黑箱”中，不同的心智装置彼此以非常复杂的方式建立起了因果联系，因此，脱离这个复杂的关系网去理解某一个心智状态所扮演的功能角色，就无法使该心智状态的本质得到正确的呈现。

由于本文的兴趣，在此笔者不想讨论各种具体版本的功能主义——如心理功能主义、机器功能主义、分析功能主义——之间的差异。在此笔者甚至不想讨论基于“感受质”（qualia）问题以及“中文屋”问题对功能主义所提出的挑战，因为这里试图论证的，并非“功能主义是一种在各方面都毫无瑕疵的学说”，而是希望说明：行为主义的模型因为分辨率太低，无法刻画个体根据少量反例修正前见的能力；与之相较，功能主义会在这个问题上表现得更好。

但为何功能主义能在这个问题上表现得更好呢？功能主义对于某一类心智状态的功能角色的刻画，不也处在一个非常抽象的层面上吗？

并非如此。功能主义对于认知架构内部结构的重视，将使得其很难不重视记忆（包括工作记忆、长期记忆等）的重要性，而与记忆相关的诸模块的存在显然能够使得个体与外部环境互动的历史的特殊性得到记录，由此展现个体之间的差异。而智能体具有的此类特殊记忆，则会在其产生某种特殊意图的时候以“背景知识”的方式发挥作用，比如对于金章宗笔迹特征的记忆，就会使得某位文物鉴定专家产生“区分金章宗与宋徽宗笔迹”的具体意图。需要注意的是，在人类记忆中出现的个别案例是带有权重的，比如，来自一个比较可信的信源的证言信息的权重，就要高于来自一个不那么可信的信源的证言的权重。这些权重方面的差异，显然能够解释为何极少的反例就能使得一个主体修正其“前见”，因为这些反例从高权重的信源那里得到的强度，抵消了反例自身在数量上的稀少性。同样需要注意的是，在真实的社会网络中，信源的可信度的权重与认知主体所处的社会环境特征高度相关。譬如，在一个局域的社会网络中，某认知主体会认为其父亲所给出的信息是最可靠的，而在另一个局域的社会网络中，相关的权重却会被赋予某位教师。因此，个体赋予不同信源以不同权重的做法，本身就是对“文化多样性”的一种曲折的表现。

需要指出的是，这种“文化多样性”是无法展现于大语言模型的，因为大语言模型的工作原理，只能允许其以“屏蔽来源”的方式处理来自不同个体记忆的互联网语料，由此将亚马逊雨林一般丰富的局域社会关系全面予以扁平化。这当然不是说现在的大语言模型的研究者完全没有理睬心理学界对于记忆的研究成果，而是说，他们至多只能通过对于“多头注意力机制”（multi-attention-head）的设置来部分模拟人类的工作记忆能力，却几乎无法模拟人类记忆能力中最关键的部分：长期记忆能力。譬如，有报道指出，ChatGPT 每次所能处理的字符数量上限约为 4000 个，而下一次输入与上一次输入之间的语义联系则未必会被保存。因此，它很难像具有正常长期记忆能力的人类那样，连续阅读长篇小说，并在阅读小说后 100 页时始终记住小说前 100 页的梗概。① 而对于人

① Calvin Wankhede: Does ChatGPT have a Character Limit? *Android Authority*, 26 February, 2024, https://www.androidauthority.com/chatgpt-character-limit-3292997/.

类个体的个性多样性(以及在此基础上建立起来的文化多样性)而言,在长期记忆库中稳固保存某些特定信息的心智能力显然是不可或缺的。

看得更深一点,功能主义的心智框架对于记忆模块的容纳之所以能展现个体认知的特殊历史,乃是因为记忆活动本身就允许种种发生在信息记录过程中的偶然性因素所造成的个体多样性。与之相较,诚如乔姆斯基所言,斯金纳式的行为主义叙事却只能适用于被严格控制的心理学实验室,而逸出此类控制范围的偶然性因素自然也就被相关进路刻意排斥了。到了大语言模型时代,这一排斥则以一种更为昂贵的方式被再现了,即模型构造者不惜工本地引入了一个巨大的预训练文本库与强大的处理硬件来将整个互联网世界都“实验室化”了,并在此基础上继续以行为主义的方式“调教”系统。且不论此做法所带来的巨大经济成本(且不提已略显过时的 ChatGPT-3 的预训练阶段所需要的 10000 个以上的高效能图形处理器、数月的训练时间,以及 570TB 的海量原始训练语料,光目下的 ChatGPT 系统的每日运行费用就在 70 万美元左右[①]),即使这些经济成本是可以负担的(比如,最近出现的 DeepSeek 的确在降低大模型建模成本方面展现出了潜力),此进路也无法在形而上学层面回避如下问题:对于世界的数码化模拟永远不会成为世界本身,因此,世界本身总会有一些无法在上述模拟图中被敉平的偶然性因素。至少就大语言模型进路而言,这些偶然性因素就包括(但不仅限于):未在互联网上积累足够数据的小语种与方言信息、冷门学科信息,更不用提千万个人类个体丰富的心灵史以及世界历史进程所自带的不可预测性。与之相较,由于功能主义的心智模型在一开始就放弃了“建立万有知识库”的野心,此类模型对于认知主体的“有限性”的肯定反而使得其有机会更忠实地再现个体不断修正“前见”的信息加工过程。

那么,如何将抽象的功能主义原则落实为具体的 AI 建模工作呢?在这里比较值得一提的是 META 公司的法籍首席科学家、2018 年图灵奖获得者杨立昆(Yann LeCun)的工作。在最近几年的大量学术会议与网络访谈中,杨立昆反复批评目前主流的大语言模型进路,其相关理由与本文的立论也有重合。尽管作为 AI 专家而不是哲学家,杨立昆并未明说自己的研究是一个功能主义规划,然而,他的规划依然体现了功能主义的根本特征:心智架构的各个模块各司其职,彼此之间又互相配合,完成了整部心智机器的运作。(顺便说一句,在他的技术语境中,“模块”大约是指一个具有相互独立性的认知计算单元)。具体而言,杨立昆正在研发的 AI 架构由六个模块构成:[②](1)知觉模块:其任务是从感受器中获得信息,并由此评估环境的当下状态。与在传统的深度学习进路中被研发的图像识别器不同,该知觉模块只会关注到与刚被给定的认知任务相关的环境参数;(2)世界模型模块:该模块的功能体现在如下两个方面:(甲)对知觉没有提供却依然为系统所关心的那些关于环境的信息进行猜测;(乙)对外部环境的可能发展情况进行猜测。需要注意的是,因为世界本身是充满偶然性与不确定性的,因此,该模型对于外部世界之发展情况的猜测也是多重的;(3)损耗计算模块:其任务是计算任何一个行动所会导致的系统资源损耗,包括在短期内对系统所能产生的损害以及在更遥远的未来给系统造成的总体损害。该模块的运作将使得系统的运作更为“经济”;(4)行动策划模块:与上一模块相互配合,以给出一个尽量减少资源损耗的行动规划,以便完成相应的任务;(5)短期记忆模块:在比较短的时间内追踪被知觉与被预测到的环境状态,以及被预估出的行动损耗,促使系统在任务执行过程中能做到注意力集中;(6)高级控制模块:对所有上述这些模块的运作进行统调。

---

① Aaron Mok: “ChatGPT Could Cost Over $ 700,000 Per Day to Operate, Microsoft Is Reportedly Trying to Make It Cheaper”, *Business Insider*, 20 April, 2023, https://www.businessinsider.com/how-much-chatgpt-costs-openai-to-run-estimate-report-2023-4.

② 相关构架目前还处在研究阶段,没有相关论文的表述。不过,在 META 的官网上有对此构架的大致介绍。请参见 Yann LeCun on a Vision to Make AI Systems Learn and Reason Like Animals and Humans, https://ai.meta.com/blog/yann-lecun-advances-in-ai-research/。

尽管杨立昆的上述架构没有明确提到一个独立的长期记忆模块,但高级控制模块与损耗计算模块的运作都预设了系统的长期记忆能力。因此,他的模型至少在理论上应当能够产生一个基于此类记忆能力的"个体心灵史"。同时,杨立昆对于系统运作资源之经济性的考量,也使得其研究旨趣迥异于必然会消耗大量资源的大语言模型进路。若细化到本文所关心的"前见—偏见"问题,该架构为"世界模型模块"所匹配的"联合嵌入预测架构"(Joint Embedding Predictive Architecture,简称 JEPA)也的确具备根据不断涌入的经验积累出"前见"的能力。① 在 JEPA 的技术语境中,"前见"指的是系统在扫描了被展示图景后,对图景中被遮盖部分的预测。此类预测,非常类似于胡塞尔所说的"侧显"(Abschattung),即现象学主体对未被直接知觉的面相的预报(譬如,仅仅看到一位女士的倩影,就预见到正面姣好的面容)。顺便说一句,JEPA 技术允许系统对此类预测的结果仅仅做一种低分辨率的表征,由此节省系统的运作资源。同时,它也允许系统通过一种类似于大语言模型的"预训练"过程,接触大量的"被知觉物"与"被预测物"之间的合理匹配模式,由此自己"琢磨"出"管中窥豹"的门道。

不过,依据笔者浅见,杨立昆的模型依然无法真正模拟人类根据新经验修正成见并由此防止偏见的能力。如此判断,乃是因为 JEPA 技术的工作层面依然太低:它主要被用于对静态或者动态的视觉信息的处理,而不是用于对高层级的语言信息的处理。与之相较,我们所说的"偏见"却往往产生在这个相对抽象的层面上,否则此类偏见就不可能孳生类似"种族偏见"之类的负载抽象语义的标签。因此,就 JEPA 技术的目下形态而言,我们很难看出这样的系统是如何产生具有抽象语义内容的"偏见"的,遑论产生自动消除此类"偏见"的能力。因此,杨立昆的模型是否能够真正做到"人机对齐",依然需要观察。

就这一点而言,杨立昆的架构的确需要来自一种以高层语义推理模型为特色的 AGI 架构的强有力补充。笔者能够想到的此类构架,就是美国天普大学的计算机专家王培发明的"非公理推演系统",简称"纳思系统"。② 与杨立昆的架构对于知觉与环境预测问题的高度关注不同,纳思系统关注的核心问题乃是如何在一个高度抽象的层面进行表征处理,特别是如何在逼仄的运算资源的约束下,让系统输出能够引导合适行动的信念。纳思系统具备丰富的推理规则,以指导系统在遇到针对某一信念的反例之时恰当修正对于该信念的置信度——而相关置信度参数的可量化性,亦使得系统能够在"彻底相信某事"与"彻底不相信某事"之间有一个广阔的缓冲空间。很显然,纳思系统的上述技术特征,为其在数码世界中塑造出谨慎小心的认知性格提供了便利。不难想见,对于任何不想陷入过多偏见的认知主体来说,上述认知性格乃是不可或缺的。

本节所进行的讨论虽然具有一定的探索性,但至少就其大方向而言,已经向读者指明了这一点:只有经由此条或彼条功能主义道路,AGI 才能在"维护必要前见,摒弃有害偏见"的维度上做到"人机对齐",唯有功能主义的框架才能为 AGI 的信念修正机制的运作提供足够的理论空间。同时,无论是杨立昆的架构,还是王培的架构,都体现了在功能主义框架中节省系统运算资源的巨大潜力,而这种潜力却几乎被以"拼算力、拼训练语料"为特色的大语言模型进路完全遮蔽了。

结语:在大语言模型进路之外发展"新质"AGI 进路的必要性

目下全面发展"新质生产力"的必要性,已经得到了各界的高度重视。"新质"这一定语的核心语素是"质",而"质"乃是相对于"量"而言的。从某种意义上说,在 21 世纪初原有的神经元网络技术升级为深度学习技术之后,AI 发展的本质特征乃是"增量",如增加神经元网络的计算层次、增加图形处理器的数量、增加训练或预训练的材

① Adrien Bardes, Jean Ponce, Yann LeCun: "MC-JEPA: A Joint-Embedding Predictive Architecture for Self-Supervised Learning of Motion and Content Features", arXiv:2307.12698? [cs.CV], https://doi.org/10.48550/arXiv.2307.12698.

② Pei Wang, *Non-Axiomatic Logic: A Model of Intelligent Reasoning*, Singapore: World Scientific Publishing Co., 2013.

料量等,而非添加“新质”的元素,如在根本上对基于行为主义预设的深度学习进路作出检讨。然而,正如前文反复指出的,进行这种基于“新质”思维的检讨乃是非常必要的。说得不客气一点,行为主义在哲学领域与心理学领域其实都已经是相对“过气”的理论,却在 AI 学界成为缄默的哲学前提,而这一点本身就足以彰显工程学实践滞后于理论研究前沿的时间是何等之长。说得更不客气一点,这一滞后最终或许会导致主流的大模型技术无法从别的哲学资源获取营养,由此也在原则上无法解决“人机对齐问题”,遑论经由此类解决而实现基于“新质”内容的重生。真正基于“新质”思维的 AGI 研究无法走大语言模型的老路,也正是这个道理。

上述考察不仅具有理论意义,而且具有紧迫的现实意义。在西方国家试图从芯片角度对我国 AI 研究的算力基础进行限制的背景下,与西方国家在“增量”的维度上进行 AI 技术竞争,显然已非明智之举。如何在有限算力的限制下发展“新质”AGI,成为中国的 AGI 研究者必须严肃考虑的课题。DeepSeek 虽然利用“蒸馏”技术在降低大模型建设成本方面有所进步,但其运行的基本原理依然是行为主义的,而不是一种更能展现人类心智架构的功能主义。因此,DeepSeek 的进步虽具有工程学意义,但依然缺乏哲学意义。要发展真正具有中国特色的 AGI 道路,我们依然需要更具革命性的新思路。而本文的研究,也仅仅是在这一方向上为学界抛砖引玉罢了。

作者简介:徐英瑾,复旦大学哲学学院教授、博士生导师,复旦大学特聘教授,浦江国家实验室学术顾问。

〔责任编辑:冯　潇〕

《齐物论》完成时代考

李　锐

《齐物论》是《庄子》内篇中的重要篇章,历代多以为庄子所作,但《齐物论》中有明显的时代印记,可以看出并非庄子的作品。至少可以说,《齐物论》的完成时间,比庄子要晚。

如篇中有“以指喻指之非指,不若以非指喻指之非指也;以马喻马之非马,不若以非马喻马之非马也。天地一指也,万物一马也”,这很明显指向的是公孙龙子的《指物论》和《白马论》。公孙龙子的年代,比庄子晚不少,他以白马非马说闻名时,庄子已至暮年。钱穆考惠施卒于 314B. C. —310B. C. 之间(钱穆:《先秦诸子系年》,商务印书馆 2001 年版,第 441 页),庄子卒年在此后不久。而《淮南子·道应》载公孙龙离赵赴燕,大概是学成而出。《吕氏春秋·应言》记公孙龙子说燕昭王以偃兵,大概在昭王去世(279B. C.)之前不久,其时已经破齐。故庄子和公孙龙成名时间或不相及。有学者说白马、指物这些辩题,都产生在公孙龙之前,公孙龙不过荟萃众说,集其大成罢了(栾星:《公孙龙子长笺》,中州书画社 1982 年版,第 14 页)。在公孙龙之前,兒说也讲白马非马,但是指物之说尚未见。或说《秋水》篇载公孙龙子对魏牟说“今吾闻庄子之言,汇焉异之。不知论之不及与,知之弗若与?”这说明《齐物论》已经批驳了白马、指物论,故公孙龙得闻其言。当然,《秋水》之言是否对应《齐物论》尚未可知。

而且《齐物论》中还有“昭文之鼓琴也,师旷之枝策也,惠子之据梧也,三子之知几乎,皆其盛者也,故载之末年。唯其好之也,以异于彼,其好之也,欲以明之彼。非所明而明之,故以坚白之昧终。而其子又以文之纶终,终身无成”。此处“其子”,从文脉来看分明是指“以坚白之昧终”的惠施之子,林云铭、胡文英皆已指出(林云铭:《庄子因》,华东师范大学出版社 2012 年版,第 19 页;胡文英:《庄子独见》,华东师范大学出版社 2011 年版,第 12 页)。但郭象却认为“文之纶终”的文是昭文,说“昭文之子又乃终文之绪”(郭象注、成玄英疏:《南华真经注疏》,中华书局 1998 年版,第 40 页),这样不顾文脉,大概是因为固守《齐物论》是庄子作品,而庄子不得见惠施的儿子之终。“文之纶终”的纶是纶绪,这里应该是说惠施之子,以惠施之文辞的纶绪条理为事而卒,终身无成就。

因此,《齐物论》的作成时间,应在庄子之后,在惠施之子卒后,估计整篇都不是庄子所作。以三十年一代估算,惠施之子卒年应该在 280B. C. 左右。如果再考虑公孙龙子以指物论、白马论闻名,《齐物论》作成的时间可能还要再晚。

(作者单位:北京师范大学历史学院)