

## 有知识的无知：大模型的道德归责

### Knowledgeable Ignorance: The Moral Responsibility of Large Models

王炜 / WANG Wei

(复旦大学哲学学院, 上海, 200433)  
(School of Philosophy, Fudan University, Shanghai, 200433)

**摘要:**自OpenAI发布聊天机器人ChatGPT以来,有望成为通用人工智能的大模型广泛进入大众生活,在社会各界引起持续高走的声量。就大模型的训练数据和文本生成能力而言,大模型不仅“拥有”千亿级训练数据集,还能流畅地生成各类文本任务,表现为无所不知知无不言的“有知识”;就大模型的能力本质而言,又会遭受体现在信念和确证能力缺乏的形而上学困境,即大模型不具备形成知识的基本条件,因此“无知”。回到亚里士多德有关无知与道德归责的理论,在无知定义的标准观点和新观点下,文章分析了大模型多种无知状态的道德归责分类:先天性无知与经验性无知中的一般性无知不该负有道德责任,而经验性无知中的非鲁棒性无知应该负有责任。

**关键词:**大型语言模型 道德归责 无知 人工智能伦理 深度学习

**Abstract:** Since OpenAI released the chat robot ChatGPT, the large models that are expected to become artificial general intelligence widely enter public life, which causing continuous high popularity in all walks of life. In terms of the training data and the text generation ability of large models, they not only “own” hundreds of billions of training datasets, but also can smoothly generate all kinds of text tasks, which is shown as “knowledgeable”, that is they know everything and say everything. In terms of the essence of the ability of large models, they will suffer from the metaphysical dilemma embodied in lack of the ability of belief and justification, that is, the large models do not have the basic conditions to form knowledge, so they are “ignorant”. Returning to Aristotle’s theory of ignorance and moral responsibility, under the standard and new views of the definition of ignorance, this paper analyzed the classification of moral responsibility of multiple ignorance states of large models: a priori ignorance and the general ignorance in empirical ignorance should not bear moral responsibility, while the non-robust ignorance in empirical ignorance should bear responsibility.

**Key Words:** Large language models; Moral responsibility; Ignorance; Artificial intelligence ethics; Deep learning

中图分类号: B842; TP18 文献标识码: A DOI: 10.15994/j.1000-0763.2024.09.008

被《麻省理工科技评论》评为2021年十大突破性技术的大模型GPT-3热度还未减去,更为强大的GPT-4便呼之欲出,而在此间隙,OpenAI于2022年11月底发布的基于GPT-3.5框架的聊天机器人ChatGPT以惊人的速度(5天达成百万注册数)迅速风靡全球。将以自然

语言处理(Natural Language Processing, NLP)为代表的大模型(Big Models)发展成为通用人工智能(Artificial General Intelligence),被人工智能科学家们寄予厚望。大模型的最明显特点是训练数据参数之大——GPT-3(OpenAI, 1750亿参数)、盘古(华为, 1000亿参数)、

收稿日期: 2023年2月14日

作者简介: 王 炜(1994-)男, 山东临沂人, 复旦大学哲学学院博士研究生, 研究方向为机器伦理、人工智能对齐。Email: w\_wang21@m.fudan.edu.cn

switch transformer (Google, 1.6 万亿参数)、悟道 2.0 (智源 & 清华, 1.75 万亿参数)。

然而, 大模型的巨大训练参数也会随之带来偏见风险, 数据样本中的歧视和偏见问题会在大模型中体现。此外, 可靠性问题、数据隐私问题、协助作恶问题等都会构成大模型的伦理挑战。那么, 这个掌握强大“知识”甚至无所不知的人类达摩克里斯之剑是否该负有道德责任呢? 本文将从无知角度对大模型的道德归责问题一探究竟。

## 一、“有知识”的大模型

大模型又称为通用模型 (General Models)、基础模型 (Foundation Models), 是指在广泛的数据上训练的任何模型 (通常使用大规模的自我监督), 可以适应 (例如, 微调) 广泛的下游任务, [1] 其原理可见下图 1, 其中, 以 NLP 为任务的大模型<sup>①</sup>又称之为大型语言模型 (Large Language Models)。大模型技术源于机器学习 (Machine Learning), 准确来说, 源于机器学习中的深度学习 (Deep Learning), 深

度学习概念由欣顿 (Geoffrey Hinton) 团队首次提出。他们发表了关于神经网络的重要性文章。[2] 然而深度学习技术可以追溯到麦卡洛克 (Warren S. McCulloch) 和皮茨 (Walter Pitts) 的 MP 模型, 他们首次提出了模拟人类神经元的数学模型, [3] 后来经过上世纪八九十年代的高速发展, 到本世纪又经过多层感知器 (multilayer perception, MLP)、卷积神经网络 (convolutional neural network, CNN)、循环神经网络 (recurrent neural networks, RNN) 的发展, 直到 2017 年 Google 公司提出 Transformer 结构, 以 Self-Attention 取代了 CNN, [4] 成为目前的主流结构。上文提到的 GPT (Generative Pre-Training) 也是基于 Transformer 训练出的大模型, 目前 GPT-3 (2020 年 5 月发布) 的参数已达到 1750 亿。

当然, GPT 的参数也是逐步加大到千亿级别的。第一代 GPT (即 GPT-1) 使用的训练数据是 BooksCorpus 数据集, 包含超过 7000 本未出版的书籍, 共有 1.17 亿参数 (已达到亿级参数)。[5] 到了第二代 GPT (即 GPT-2), 训练数据的参数达到 15 亿, 增长了十倍之多。GPT-2

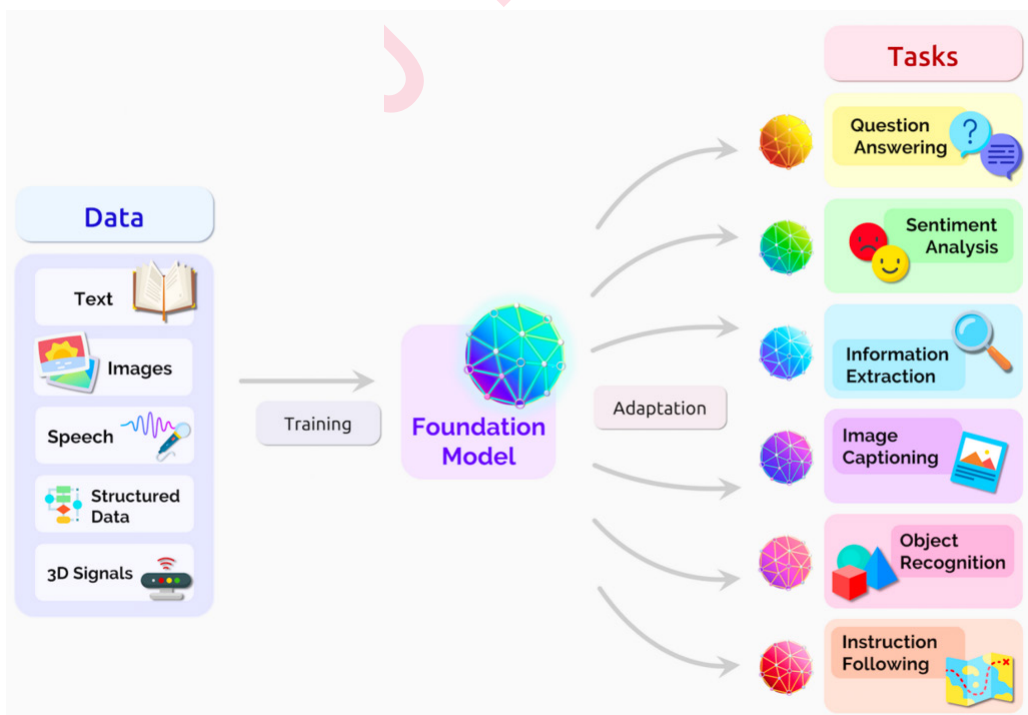


图 1 多模态基础模型下游任务图

①本文所提及和探讨的大模型, 皆指自然语言处理 (NLP) 大模型, 即大型语言模型。

的数据来自Reddit社交媒体平台,通过抓取4500万个网络链接生成数据集WebText,总共有40GB的文本。<sup>[6]</sup>第三代GPT(即GPT-3)的训练数据则直接突破千亿级别,数据集也更为复杂,包含Common Crawl、WebText2、Books1、Books2和维基百科(表1)。<sup>[7]</sup>通过三代GPT可以明显看出,大模型的预训练数据越来越大,也意味着,大模型“拥有”人类的知识越来越多。

训练之后的大模型可以任意生成文本甚至是对话,那么,经训练的大模型如何展现出智能呢?这主要得益于大模型的两个特性:涌现性和同质性。涌现性(Emergence)意味着系统的行为是隐式诱导而非显式构建的;它既是科学兴奋的源泉,也是对意想不到的错误序列的焦虑。<sup>[1]</sup>通俗一点讲就是,在复杂系统中量变引起质变的现象,比如,诺贝尔物理学奖得主安德森(Philip Anderson)曾给涌现下过定义:涌现是指当系统的数量变化时导致行为的性质发生变化,<sup>[8]</sup>人类的意识就典型的具备涌现性特征。于大模型而言,涌现性则由于超大数据训练而形成的文本(或自然语言)生成能力,NLP科学家也不知道大模型会生成何种文本(隐式),当然,涌现性也会导致错误和无效文本(错误序列)。同质性(Homogenization)意味着在广泛的应用中整合构建机器学习系统的方法;它为许多任务提供了强大的杠杆作用,但也造成了单点故障。<sup>[1]</sup>在大模型语境下,同质性指的是模型构建方法的相似性。<sup>[9]</sup>这两个特性基本已经得到业界公认,然而,其特性产生的原因和可解释性仍在不断讨论和探索中。

就大模型“有知识”而言,在人工智能尤其在NLP领域,几乎不是个问题。GPT-1在发

布时,其团队就声称它获得了世界知识(world knowledge)的能力,<sup>[5]</sup>大模型所拥有的关于世界的各类知识来源于其数据集,<sup>[10], [11]</sup>甚至有专家证明了知识密性型的任务表现出来的性能与模型的大小高度关联。<sup>[12]</sup>概括来说,大模型有知识的证据可以总结为以下三个方面。(1)上下文学习(in-context learning)。尽管大模型上下文学习的机制仍未搞清楚,<sup>[13]</sup>但从结果看,表现出了非凡的能力。大模型可以通过数据训练,结合情景进行上下文学习,并可以就新任务给出方案,这便和单纯的知识(信息)存储或者知识图谱有所不同。可以说,大模型不仅学习了巨量的知识,还异于也可以存储知识的硬盘或服务器(知识的载体),具备某种知识主体的特征。(2)相对稳定的能力。上文提到大模型具备涌现性和同质性两大特征,就涌现性而言,大模型具有自主生成知识的能力,就同质性而言,大模型相似的模型结构,保证了其可以成为相对稳定的知识主体。(3)结果表现为掌握知识。大模型最直接的特征就是语言生成的能力,对于各类提示(prompt)问题,给出文本回答,既包含事实性知识(factual knowledge)又包含常识(common sense)。总而言之,大模型不仅是“有知识”,还可以说是非常“有知识”。

尽管如此,上述对于大模型“有知识”的说明或其证据总结都是基于外部观察结果或基本能力背书,并非从知识论角度来探究大模型是否符合知识的条件——从而判定其有知识。因此,大模型有知识(符合知识条件)或无知(符合无知条件)仍需要从知识论视角进一步探索。

## 二、大模型的“无知”

表1 用于训练GPT-3的数据集

数据集	数量 (词元)	训练组合中的权重	当训练3000亿词元时 数据集重现的次数
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

无知 (ignorance) 作为一个古老的哲学问题, 近些年来又引起知识论和伦理学的关注。在无知的知识论 (epistemology of ignorance) 领域, 无知的定义和归类被深入讨论, 在超越知识论的无知 (ignorance beyond epistemology) 研究中, 伦理学问题被很好的结合起来 (比如道德免责、认知不正义等), 尤其还诞生了无知学 (Agnology) —— 提倡主动创造和维持无知——这样的学科。关于无知的定义和分类, 近些年在知识论中有诸多细致的讨论。其中, 有两种代表性的定义理论, 被称之为标准观点和新观点。标准观点 (Standard View) 认为, 无知是知识的缺乏 (the absence or lack of knowledge) “无知” 是 “知识” 的反义词, ([14], p.15) 代表人物有菲尔茨 (Lloyd Fields)、哈克 (Susan Haack)、勒莫万 (Pierre Le Morvan)、齐默尔曼 (Michael Zimmerman) 等。([14], p.12) 新观点 (New View) 认为, 无知是真信念的缺乏 (the absence or lack of true belief), ([14], p.25) 代表人物有戈德曼 (Alvin Goldman)、格雷罗 (Alexander Guerrero)、皮尔斯 (Rik Peels)、沃登伯格 (René van Woudenberg)。([14], p.12) 由此可见, 两种观点的区别在于确证 (justified) 的缺乏是否作为无知的条件。

按照传统知识的定义, 即得到确证的真信念 (Justified True Belief, JTB), 大模型的知识 and 无知是否得以说明和刻画呢? 这显然会因人工智能和心灵哲学由来已久的形而上学问题而面临复杂的困境。知识定义的三个条件中, 只有 “真” (T) 是相对容易确定的, 人工智能 “拥有” 或输出一个命题, 其命题的真假可以通过第三人称视角客观评判。对于 “确证” (J) 而言, 情况就开始复杂起来, 首先, J 在不同的哲学语境中存在解释空间, 我们在此姑且采用费尔德曼 (Richard Feldman) 和科尼 (Earl Conee) 经典的形式化表述: S 在时间 t 对于命题 p 的信念态度 D 在知识上被确证, 当且仅当对 p 的 D 符合 S 在 t 时的证据。<sup>[15]</sup> 问题仍然存在: 大模型是否有使得命题得以确证的证据呢? 由此, 暂且可以分为两种立场, 一种是大模型只接受数

据集训练, 不可能再去通过证据确证命题, 我们称之为 J 强观点; 另一种是大模型基于基础算法框架在接受巨量数据训练时, 已经将涵盖证据的知识学会了, 此外, 强化学习的奖惩、指令微调 (instruction tuning)、基于人类反馈的强化学习 (reinforcement learning with human feedback, RLHF) 等技术手段, 更进一步保证了大模型对于命题的确证, 因此, 可以认为大模型具备确证能力, 我们称之为 J 弱观点。大模型的 “信念” (B) 将会面对更多的争议, 因为拔出信念之萝卜会带出心灵哲学形而上学问题之泥。没有心灵和内在状态的机器如何能拥有信念呢? 人工智能没有心灵, 只是心灵的抽象模拟<sup>[16]</sup>——以塞尔 (John R. Searle) 为代表的强人工智能反对者们不会同意人工智能拥有信念, 同样也不会认可大模型拥有信念, 我们称之为 B 强观点; 与之针锋相对的立场则认为人工智能拥有信念, 人工智能一词的发明者麦卡锡 (John McCarthy) 就说连恒温器这种简单的机器都有信念,<sup>[17]</sup> 更何況人工智能了, 在此立场下, 至少可以说大模型拥有的命题属于大模型信念, 我们称之为 B 弱观点。于此, 我们通过 JTB 对大模型的知识条件做了分析, 并总结出确证和信念维度的两种立场。

结合上文无知的两种定义, 大模型的 “无知” 将会呈现为多种情态。首先是条件 T, 当 T 不满足时, 无论标准观点还是新观点, 大模型都处于 “无知” 状态。其次看条件 B, 其分为 B 强观点和 B 弱观点, 在强观点下, 由于大模型不具备信念, 则不论标准观点还是新观点, 大模型都处于 “无知” 状态。最后看条件 J, 同样分为 J 强观点和 J 弱观点, 在强观点下, 大模型不具备确证条件, 根据标准观点, 大模型无论何种情况都不具备知识; 而新观点下大模型的无知状态则不受影响, 只需要关注条件 T 和 B 即可。根据标准观点和新观点对于无知的定义, 再将条件 B 和条件 J 的两种强弱版本立场考虑进来, 则可以形成大模型标准观点下的 8 种无知情况和新观点下的 4 种无知情况 (表 2 和表 3)。其中, 在标准观点下, 仅有一种情况大模型处于 “非无知” (有知识) 状态, 即 T 满足,

且B弱观点下满足,且J弱观点下满足;在新观点下,也仅有一种情况大模型处于“非无知”(有知识)状态,即T满足,且B弱观点下满足。

表2 标准观点下大模型的无知情况

T	B	J	无知
是	强:否	强:否	是
		弱:是	是
	弱:是	强:否	是
		弱:是	否
否	强:否	强:否	是
		弱:是	是
	弱:是	强:否	是
		弱:是	是

表3 新观点下大模型的无知情况

T	B	无知
是	强:否	是
	弱:是	否
否	强:否	是
	弱:是	是

通过分析可得,大模型在标准观点下有7种无知状态,新观点下有3种无知状态,然而,这些无知却因暗含着形而上学基本立场上的差异而不尽相同。其关隘仍在条件B和条件J的两个立场版本中,所有强观点都会使得大模型处于无知状态,这种无知是因为大模型不具备知识的先天能力而导致的,即大模型不是不满足B和J(在有满足的能力下),而是没有能力且不可能满足B和J。这和人类的无知有本质不同,人类是有满足JTB的能力但可能在某种情境下不满足而已。我们将大模型这种无知情况称之为“先天性无知”,即大模型不具备信念和确证的条件,先天性无知的基础源于人工智能没有心灵和智能的立场,在此立场下大模型不具备拥有知识的能力,这种立场可以追溯到一系列弱人工智能的拥趸。“先天性无知”是指大模型作为知识主体能力的缺乏,人工智能科学家们所认为的或研究中所声称的“大模型

拥有知识”其实只是在“知识载体”而非“知识主体”意义上谈论的“有知识”。相应地,条件B和J的弱观点中,大模型具备拥有B和J的能力,只要条件满足,大模型就会拥有知识,同样,如果大模型不满足条件,就处于无知状态。此时的无知就与“先天性无知”不同,我们可以称之为“经验性无知”。事实上,大模型的确会产生很多错误的信息,即存在“知识的缺乏”。比如,ChatGPT也会判断错质数,认为“numbers”是6个字母等。在“经验性无知”情境中,大模型本身是有能力产生知识的,只是会出错,即缺乏知识或真信念。在上述分析的基础上,我们可以将标准观点和新观点下大模型的无知情况做进一步的分类说明(表4和表5)。

表4 标准观点下大模型的无知情况

T	B	J	无知	无知分类
是	强:否	强:否	是	先天性无知
		弱:是	是	先天性无知
	弱:是	强:否	是	先天性无知
		弱:是	否	
否	强:否	强:否	是	先天性无知
		弱:是	是	先天性无知
	弱:是	强:否	是	先天性无知
		弱:是	是	经验性无知

表5 新观点下大模型的无知情况

T	B	无知	无知分类
是	强:否	是	先天性无知
	弱:是	否	
否	强:否	是	先天性无知
	弱:是	是	经验性无知

进一步地,我们可以将大模型“先天性无知”归为“非知识”(与知识无关)其实是一种特殊的无知,如此,大模型就分为三种可能的知识状态:“有知识”<sup>①</sup>“无知”(经验性无知)、“非知识”(先天性无知)。“有知识”和“无知”是以大模型具备知识主体资格为基础

①此处所提及的“有知识”与上文第一部分(倒数第二段)所提及的大模型“有知识”在结果层面上是一致的。但第一部分所言“有知识”更多是一种外部观察结果和大众直觉,而此处所言“有知识”是在内部符合知识论条件的基础上提及的。

的，“有知识”则意味着大模型满足TB或JTB或JTB+X（盖提尔化后），“无知”意味着大模型不满足TB或JTB或JTB+X。“非知识”则认为大模型不是知识主体，其生成的文本或命题与只是无关。大模型的“无知”与“非知识”并不相融但可以居于两者之间，即大模型虽具备知识主体资格，但其知识能力先天的存在缺陷和不足。大模型关于知识情况的三种维度和关系可以表示为下图2。

### 三、大模型是否该负责？

无知与道德归责的关系可以追溯到亚里士多德，其在《尼各马可伦理学》中提出道德归责问题，首次对“自愿”的条件做出了分析，并把无知问题考虑在道德归责范畴内。亚里士多德是从自愿的反面“不自愿”（或违反意愿）分析开始的，“不自愿”出于强迫和无知。（[18]，p.58）不被强迫则意味着行为的始因源于自身，（[18]，p.59）自己可以控制自己使得自己去做什么；无知也并非只是处于无知状态，亚里士多德认为出于无知而做出的行为与处于无知状态做出的行为是有区别的，前者是由于对行为本身和环境对无知造就的，后者是由于一个本来有知识但并未运用知识而处于对知识的无意识状态导致的，（[18]，pp.61-62）亚里士多德所说的出于无知考虑到了行为本身和行为的相关环境，跟当代实用主义或认知科学（尤其4E认知和预测心智）有相当的理论亲缘性和暗合，实际上，当代关于无知与道德规则的讨论基本也都发端于亚里士多德。总而言之，分析完“不自愿”后，反过来，自愿即可以表述为

行为的始因源于自身且知道该行为的结果和行为的具体环境。

当代哲学家们以更为细致的方式继承了亚里士多德的观点，比如罗森（Gideon Rosen）就认为因认知行为的粗心没有遵守义务，这种无知就是应受责备的，即一种应受责备的无知（Culpable Ignorance）。<sup>[19]</sup>为了更加清晰且直觉上对不同无知与责任做出区分，我们仿照菲茨帕特里克（William Fitzpatrick）的例子构造如下案例。<sup>[20]</sup>

案例1：

小明感染新冠病毒，爸爸非常担忧，将家中所有感冒药全部给小明服下，小明因肝功能衰竭而住院，爸爸并不知道多种感冒药混合吃会造成肝功能衰竭。

案例2：

小红感染新冠病毒，爸爸非常担忧，此时爸爸从手机上看到权威媒体和社区通知发布的“感染新冠居家用药提醒”，要求居民好好查看，爸爸没当回事，并未查看。将家中所有感冒药全部给小红服下，小红因肝功能衰竭而住院。

通过案例1和案例2可以看出，小明和小红的爸爸都是源于无知，我们直觉上仍觉得有差异，且更倾向于小明的爸爸是一种免于责备的无知，而小红爸爸则是应受责备的。小明的爸爸对行为本身和环境（吃药行为所产生的后果以及医学知识）都一无所知，属于亚里士多德所讲的出于无知的行为，而这种行为是免于责备的。按照上述罗森的理论，小红的爸爸是因为一种粗心和不作为而导致的一种应受责备的无知，这种无知虽然与亚里士多德的处于无知有所不同，但结构相似——本应有知但由于当

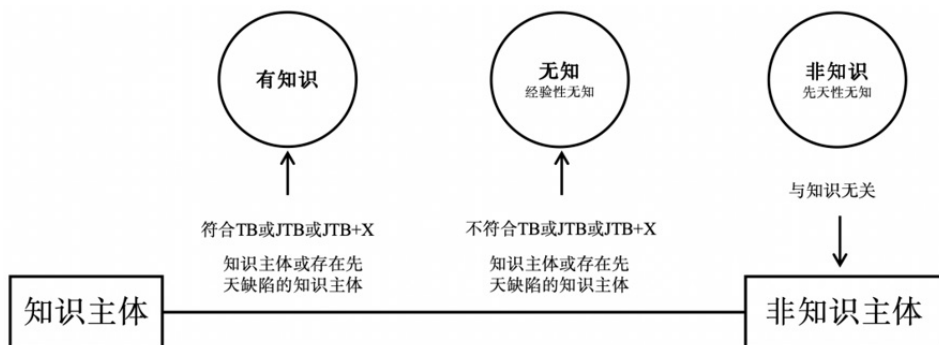


图2 大模型的三种知识状态图

事人原因而处于无知。

为了与大模型的多种知识状态对应,我们再构造一个有知识状态下的案例作为对照,该案例为应受责备的典型情况。

### 案例3:

小刚感染新冠病毒,爸爸非常担忧,此时爸爸从手机上看到权威媒体和社区通知发布的“感染新冠居家用药提醒”,爸爸看过后明知多种感冒药混吃会导致肝功能衰竭,但仍将家中所有感冒药全部给小刚服下,小刚因肝功能衰竭而住院。

有了上述分析,我们便可以根据大模型的无知种类通过无知与道德归责理论来分析大模型的道德归责情况。在分析之前,我们先总结一下大模型道德归责的困难,然后再分析本文的研究视角和出发点如何有效地规避和回应这些困难。

(1) 道德主体性难题。在人工智能伦理或机器伦理研究中,一直存在有关机器道德主体性地位的争议,即讨论机器能不能成为道德施者(moral agent)<sup>①</sup>,目前有道德施者论与反道德施者论两种立场。反道德施者论绝不会认为机器(包括大模型)能够负道德责任,不管是有知识还是无知识。

(2) 无知界定的困难。在知识论的历史长河中,知识的定义一直被哲学家们反复讨论,同样地,无知的标准也非板上钉钉。知识和无知本身具有多重标准,将问题迁移至人工智能则更为复杂。

(3) 知识体系的进化。到达一定奇点,技术进步的速度往往呈指数增长。大模型基于强化学习、对抗训练、微调、人类导师、RLHF等各种纠错机制,将会不断地进化和迭代,其知识体系也将越来越完善,在不同阶段其知识能力也不同,在ChatGPT发布后犹有技术爆发之迹象。

上述三个困境中,第一个最根本也最为棘

手,只要对其讨论就会陷入长久的形而上学之争的泥淖中。本文的思路是避开机器道德主体性难题,从知识和无知角度去探究负有道德责任的必要条件或要素,然后与大模型相匹配,或者从反面,讨论大模型免于责备的条件。我们已经分析了大模型“有知识”“无知”和“非知识”三种可能的状态,又对免于责备的无知和应受责备的无知情况进行了区分,由此,便可以进一步对大模型的道德归责情况进行讨论。

首先,就非知识而言,大模型属于先天性无知,作为非知识主体缺乏拥有知识的能力,因此不可能负道德责任,这与道德主体性地位争议中的反道德施者论相呼应,不具备知识主体资格是不具备道德施者地位的一个子集。其实,利用亚里士多德的也可以解释此情况下大模型不该负有道德责任,因为大模型的所有行为都不属于自愿,其行为都是源于某种“受强迫”。其次,在无知(经验性无知)状态中,一般来说,大模型不负有责任。然而,与人类无知归责的情况类似,大模型也有两种无知(指经验性无知)及其不同的归责。第一种可以称之为一般性无知,即大模型的确不知道某个知识,其训练数据中就不存在这个知识,比如ChatGPT训练数据集截止到2021年6月,因此,其就不拥有2021年6月之后的知识,一般性无知对应亚里士多德所提的出于无知,该情况下大模型不负有责任,对应案例1。第二种可以称之为非鲁棒性(robustness)无知,指的是大模型本应维持稳定的知识能力但因鲁棒性不足而出现的无知现象,特别地,这里的鲁棒性与大模型硬件、服务器、系统性能力缺陷无关,且基于大模型具备知识主体资格状态下。类似于人类意志软弱(akrasia)或懒惰不作为,对应亚里士多德讲的处于无知,类似于案例2(但不完全相同)。最后,大模型有知识状态需要负责,对应案例3,比如,大模型明知道协助

<sup>①</sup>笔者主张将moral agent和moral patient这一对子分别翻译为道德施者和道德受者,patient作为(接)受者已较为普遍,但agent翻译较多且不统一,常见有主体、智能体、能动者、代理人。施有“实行”“给予”“发出”等义,可以表达出一方(施者)能够将理解的道德命题/规则能动性地付诸实践于另一方(受者)。

别人抢银行是不道德的，但当人类询问时仍然给出了抢银行的详细方案。值得一提的是，大模型该负有的责任究竟是一种道德责任还是一种德性（virtue）责任？后者指的是古希腊所主张的事物本来的德性，比如，刀的德性就应该是切肉锋利。当然，这又回到了本文一开始要避免的形而上学难题。

规避人工智能的道德主体性难题，从无知入手去探讨道德归责问题，是一种有价值的尝试。我们不是先确定机器是否能够作为道德施者，这样会纷争不断，而是先从别的角度切入，最后回到道德施者这一根本问题。通过大模型无知的分析，完全的先天性无知，即作为非知识主体的大模型，不可能负道德责任，可以为反道德施者论提供辩护，将其观点推进一步。以上总结可知，大模型在有知识和非鲁棒性无知状态下需要负道德责任。

## 余 论

正如大模型和人工智能还有很长的路要走一样，对其道德施者问题、归责问题甚至知识主体和无知问题的探讨与争论也会长久持续。现象上“有知识”的大模型可能在本质上“无知”，与其整体上做形而上学立场之争，不如深入内部对条件和标准做探究。本文未对无知定义的标准观点与新观点做选择和辩护（因为这于大模型而言仍即鹿无虞，时机未到），也未对大模型的J/B的强观点和弱观点做立场抉择（一因篇幅与论述重点限制，二因大模型还在高速发展时期，目前尚不好定论），而是对各种情况做出分析：在无知定义的标准观点和新观点下，对大模型的J/B强观点和弱观点做出分类，从而得到大模型无知的各种情况，进一步地，对大模型的无知进行归类，再根据无知与归责理论，最终确定大模型免于责备的无知，以及应受责备的无知。

将大模型作为对象独立地进行道德归责研究，其实是一种反人类中心主义的路径，其基于大模型或人工智能“可能”具备道德主体性地位的基本立场。这意味着，大模型可能要承

担相应的道德责任，尽管如此，并非说开发者的道德责任可以排除，在人工智能伦理学中，存在着机器人伦理（robot ethics）与机器伦理（machine ethics）的分支，<sup>[21]</sup>而开发者的责任更多对应机器人伦理。在哲学家库萨的尼古拉（Nicholas of Cusa）那里，有知识的无知是一种本质和追求真理的路径，<sup>[22]</sup>而于大模型而言，“有知识的无知”将存在诸多伦理风险，虚假信息提供、数据歧视、协助作弊作恶等，都会对人类产生直接的影响，因此，增强大模型的道德敏感性以及伦理规则的严格执行能力成为开发者重要且亟待解决的责任和义务。

## [参考文献]

- [1] Bommasani, R., Hudson, D. A., Adeli, E., et al. 'On the Opportunities and Risks of Foundation Models'[J]. *arXiv Preprint*, 2021, arXiv: 2108.07258.
- [2] Hinton, G. E., Salakhutdinov, R. R. 'Reducing the Dimensionality of Data with Neural Networks'[J]. *Science*, 2006, 313(5786): 504-507.
- [3] McCulloch, W. S., Pitts, W. 'A Logical Calculus of the Ideas Immanent in Nervous Activity'[J]. *Bulletin of Mathematical Biology*, 1943, 5: 115-133.
- [4] Vaswani, A., Shazeer, N., Parmar, N., et al. 'Attention is All You Need'[J]. *arXiv Preprint*, 2017, arXiv: 1706.03762.
- [5] Radford, A., Narasimhan, K., Salimans, T., et al. 'Improving Language Understanding by Generative Pre-Training'[EB/OL]. [https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf). 2018-06-09.
- [6] Radford, A., Wu, J., Child, R., et al. 'Language Models are Unsupervised Multitask Learners'[EB/OL]. [https://cdn.openai.com/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf). 2019-02-15.
- [7] Brown, T. B., Mann, B., Ryder, N., et al. 'Language Models are Few-Shot Learners'[J]. *arXiv Preprint*, 2020, arXiv: 2005.14165.
- [8] Anderson, P. W. 'More is Different'[J]. *Science*, 1972, 177(4047): 393-396.
- [9] 滕妍、王国豫、王迎春. 通用模型的伦理与治理：挑战及对策[J]. 中国科学院院刊, 2022, 37(9): 1290-1299.
- [10] Kazemi, M., Mittal, S., Ramachandran, D. 'Understanding Finetuning for Factual Knowledge Extraction from Language Models'[J]. *arXiv Preprint*, 2023, arXiv:

- 2301.11293.
- [11] Yasunaga, M., Leskovec, J., Liang, P. 'LinkBERT: Pretraining Language Models with Document Links'[J]. *arXiv Preprint*, 2022, arXiv: 2203.15827.
- [12] Liang, P., Bommasani, R., Lee, T., et al. 'Holistic Evaluation of Language Models'[J]. *arXiv Preprint*, 2022, arXiv: 2211.09110.
- [13] Xie, S. M., Raghunathan, A., Liang, P., et al. 'An Explanation of In-context Learning as Implicit Bayesian Inference'[J]. *arXiv Preprint*, 2021, arXiv: 2111.02080.
- [14] Pierre, L. M., Peels, R. 'The Nature of Ignorance: Two Views'[A], Peels, R., Blaauw, M. (Eds.) *The Epistemic Dimensions of Ignorance*[C], Cambridge: Cambridge University Press, 2016, 12-32.
- [15] Feldman, R., Conee, E. 'Evidentialism'[J]. *Philosophical Studies*, 1985, 48(1): 15-34.
- [16] Searle, J. 'Minds, Brains, and Programs'[J]. *Behavioral and Brain Sciences*, 1980, 3: 417-458.
- [17] 塞尔. 心, 脑与科学[M]. 杨音莱译, 上海: 上海译文出版社, 2006, 22.
- [18] 亚里士多德. 尼各马可伦理学[M]. 廖申白译注, 北京: 商务印书馆, 2003.
- [19] Rosen, G. 'Culpability and Ignorance'[J]. *Proceedings of the Aristotelian Society*, 2003, 103: 61-84.
- [20] Fitzpatrick, W. 'Moral Responsibility and Normative Ignorance: Answering a New Skeptical Challenge'[J]. *Ethics*, 2008, 4: 589-613.
- [21] Wallach, W., Asaro, P. *Machine Ethics and Robot Ethics*[M]. London & New York: Routledge, 2016.
- [22] 库萨的尼古拉. 论有学识的无知[M]. 尹大贻、朱新民译, 北京: 商务印书馆, 1988.

[责任编辑 王巍 谭笑]

