·社会学研究·

DOI:10.20004/j.cnki.ujn.2024.06.011

生成式人工智能实践下的社会观念分化探析

——基于自主行动者建模的仿真模拟实验研究

梁玉成,马昱堃

(中山大学 社会学与人类学学院,广东 广州 510275)

[摘　要]当前,以大语言模型为代表的生成式人工智能在短时间内迅速发展,其对社会观念分化产生了影响。采用自主行动者建模方法,借鉴文化传播模型,研究大模型对社会观念分化的影响。实验结果表明,大语言模型对不同观念之间的交流和讨论具有激活作用;不同大语言模型之间的内在观念差异越大,社会观念越难以形成稳定的分化格局,并且分化程度会逐渐加剧;此外,在已经形成了一定观念分化格局下,大语言模型可能会进一步减少社会观念之间的差异,有助于缓解社会观念"多中心化"。研究表明,减小不同生成式人工智能输出结果的观念差异,能够有效降低社会观念的"多中心化"、遏制局部"信息茧房"的自我强化。长远来看,大语言模型存在造成"整体茧房"的可能性。

[关键词]大语言模型;ChatGPT;生成式人工智能;社会观念分化;自主行动者建模

[中图分类号]C912.6　　[文献标识码]A　　[文章编号]1671-3842(2024)06-0120-13

一、引言

数字社会的出现,带来了人与人之间社会连接模式的变迁,任意一个社会个体都可以通过网络与另外一个个体相联系,因此,社会观念分化的主要场域从线下空间转移到线上空间,产生了一系列新现象。相较于线下互动,人们在社交媒体上需要通过数字账户与他人交流。这些账户构成了互动者的数字化身,并且所有互动行为痕迹以及用户对线上形象的设置,都与他们的心理人格存在某种关联,共同构成了数字人格(Digital Personality)①。

算法能够将用户的数字人格投射在一个高维空间,把用户分成若干个子群体,形成"微目标"(Micro-targeting)②。社会观念的产生、传播和操纵变成了一种社会工程技术手段,线上空间的社会观念干预与治理成为可能③。除此之外,社交机器人的存在也让观念操纵变得容易。宣传者可以通

[基金项目]国家社会科学基金重大项目"基于大型调查数据基础上中国城镇社区结构异质性及其基层治理研究"(项目编号:15ZDB172)。

[作者简介]梁玉成,中山大学社会学与人类学学院教授、博士生导师,教育部长江学者特聘教授;马昱堃(通讯作者),中山大学社会学与人类学学院博士研究生,邮箱:mayk5@mail2.sysu.edu.cn。

①兰天:《数字人格:数字智能时代的人格研究》,《全球传媒学刊》,2023年第3期。

②Prummer A., Micro-targeting and Polarization, *Journal of Public Economics*, Vol. 188, 2020, p. 104210.

③Kruikemeier S., Vermeer S., Metoui N., et al., (Tar)getting you: The Use of Online Political Targeted Messages on Facebook, *Big Data&Society*, Vol. 9, No. 2, 2022, pp. 1-20; Cotter K., Medeiros M., Pak C., et al., "Reach The Right People": The Politics of "Interests" in Facebook's Classification System for Ad Targeting, *Big Data&Society*, Vol. 8, No. 1, 2021, pp. 1-16.

过社交机器人宣传自己的价值观点，从而影响其他用户[①]。

大语言模型（Large Language Model，以下简称大模型）的出现可能改变线上社会观念分化的格局。2022年11月底，OpenAI发布ChatGPT之后，以大模型为代表的生成式人工智能的开发与应用开始步入迅速的迭代进化阶段[②]，产生了愈加广泛的社会影响。它们可以根据自然语言指令生成文本、图像、音频和视频内容，在互联网尤其是社交媒体上，仍然具有尚未被挖掘的应用前景。

大模型对社会观念分化的影响可以从主动和被动两方面来看。主动影响是人们基于自身目的主动使用大模型时受到的影响；被动影响则是人们在自身不知情的情况下被应用大模型技术的社交机器人影响。显然，对于后者来说，大模型实现这种观念的操纵变得更加容易[③]。大模型能够批量生成文本，大幅降低了社交机器人的运营成本。而用户主动使用互联网平台企业提供的大模型，实现文本生成、知识问答、生活求助等需求，也会对用户造成潜移默化的影响。模型生成的内容可能会进一步转化为用户自己的知识，并向其他人传播。

社会科学界亟需了解大模型对社会观念的分化和整合产生的新影响。大模型在海量语料数据上训练而成，语料选择、参数设置以及后续的价值观对齐算法等差异，都会使不同大模型针对同一输入时可能给出不同输出。而大模型的专业性背景和技术成分让其更容易被当作“可信主体”，人们更倾向于相信它的输出，调整自己原有的观念。如果大模型的输出包含偏见，那么长期使用也会让人们习得并深化这些偏见[④]。这些都会妨碍社会观念的整合，可能进一步导致观念的分化与极化。

那么，大模型以“可信主体”介入互动过程，对不同子群体构成广泛辐射，将如何影响人们的社会观念？本文使用自主行动者建模（Agent-Based Model，ABM）方法[⑤]，借鉴罗伯特·阿克塞尔罗德（Robert Axelrod）提出的文化传播模型[⑥]，尝试探索大模型影响社会观念分化的后果并讨论了大模型对社会整体观念的长远影响。随着大模型在社会互动过程中介入程度加深，虽然可能避免子群体层次的“局部茧房”，但在社会整体层面可能形成更大也更难克服的“整体茧房”。这需要社会科学研究者们进一步结合实证数据对其保持关注。

二、数字时代社会观念的分化模式及其治理方式

（一）数字时代社会观念的“去中心化”与“多中心化”

社交媒体出现前，信息传播由专业媒体机构主导。信息从中心源头向外分发，传递给目标受众，呈现出“中心化”的结构。在社交媒体时代，信息传播与线上社会互动的边界变得模糊，信息流动与社会互动紧密交织[⑦]，产生了一些意见领袖，以之为中心聚集起观念相似的群体，呈现出了“去中心化”的格局。通过信息关联，用户在不同平台上的互动行为和数字画像构成一个人完整的数字

---

①Goldstein J. A.，Ssatry G.，Musser M.，et al.，Generative Language Models and Automated Influence Operations：Emerging Threats and Potential Mitigations.，*ArXiv*，https://arxiv.org/abs/2301.04246，2023.10.10.

②Zhao W. X.，Zhou K.，LI J.，et al.，A Survey of Large Language Models.，*ArXiv*，https://arxiv.org/abs/2303.18223，2024.08.24.

③GOLDSTEIN J. A.，SASTRY G.，MUSSER M.，et al.，Generative Language Models and Automated Influence Operations：Emerging Threats and Potential Mitigations.，*ArXiv*，https://arxiv.org/abs/2301.04246，2024.08.24.

④Kidd C.，Birhane A.，How AI Can Distort Human Beliefs，*Science*，Vol. 380，No. 6651，2023，pp. 1222-1223.

⑤梁玉成，贾小双：《数据驱动下的自主行动者建模》，《贵州师范大学学报》（社会科学版），2016年第6期；吕鹏：《智能体仿真模拟：推进行动与结构互构研究》，《社会学研究》，2024年第4期。

⑥Axelrod R.，The Dissemination of Culture：A Model with Local Convergence and Global Polarization，*Journal of Conflict Resolution*，Vol. 41，No. 2，1997，pp. 203-226.

⑦梁玉成，马昱堃：《对青年的计算文本“远读”——数字时代基于降维的整体认识论》，《青年探索》，2022年第3期。

人格[①]。数字人格既包括个体的属性特征(性别、地区等),也包括个体的行为特征(操作频率、点击次数等),能被分解为多种维度,转化为多维向量。所以,数字人格能更加清晰地反映用户的个人特质。个体正是借由线上互动将自身人格的一部分在线上空间重构,同时也让算法分析数字人格成为可能[②]。互动中的信息交换和判断过程,可以抽象为互动者对自身观念体系各种参数的比较和估计的过程。人们根据与他人交互过程中获得的新信息更新自己对特定事件的认识。鉴于每个人都嵌入在一个规模有限的社会网络之中,社会观念的分化过程就是信息在网络中传播以及人们对特定事件信念参数的调整过程[③]。获得信息存在两类成本,分别是接触信息的时间成本,以及接纳信息的平衡成本。接触信息的时间成本导致在给定时间条件下只能接触有限多的信息量,造成群体规模的有限性;接纳信息的平衡成本导致人们倾向于选择并快速接纳和自己观念更加相似的信息,造成了群体内信息有限异质性。这就造成了"信息茧房"现象[④],个体只能了解到关于社会的局部事实[⑤],形成了社会观念的"多中心化"格局。

在以上条件下,用户群的数字人格不再呈现均匀分布,而是呈现簇状分布的特征。每个簇都呈现高度均质化,社交媒体的用户被抽象为嵌入在高维空间中的数据点,利用大数据和算法就能将人群分解为群体内部信息均质的微目标[⑥],并根据其特征投放内容,从而建构舆论控制系统。

总结来说,社交媒体呈现用户之间彼此互联的网络结构,信息沿网络连边传递,作为网络节点的个人据此调整自己的观念,逐渐形成不同的子中心,瓦解了原有的中心化,逐渐形成"信息茧房",导致"多中心化"的观念格局。

(二)社会观念的"去多中心化"治理模式

互联网的发展贯穿着"去中心化"与"再中心化"博弈[⑦],社交媒体"去中心化"以及"多中心化"的发展态势,对社会治理构成了较大挑战。若任由社会观念的去中心化和多中心化发展,不同观念的群体之间将变得难以相互交流,在"信息茧房"内自我强化[⑧],将导致观念的极端化,并最终可能导致冲突。为避免这一现象,需要对社会观念进行"去多中心化"治理。

党的十八大以来,国家一直在加快建设网络强国、数字中国,同时,国家日益在强调主流意识形态的领导地位。对此,有学者总结了政府对网络舆论生态进行治理所采取的一系列措施,包括健全法律法规、依法处罚对社会造成危害的非主流意识形态;官方媒体参与舆论引导、宣传主流意识形态;网络技术控制非主流意识形态、化解极端性;监督社交媒体意见领袖等[⑨]。通过改变舆论场域

---

①兰天:《数字人格:数字智能时代的人格研究》,《全球传媒学刊》,2023年第3期。

②Alrehili M. M. , Yafooz W. M. S. , A Review of Extracting and Mining User Interest from Social Media Based on Personality, in *Proceedings of the 2021 3rd International Youth Conference on Radio Electronics*, Electrical and Power Engineering (REEPE), 2021, pp. 1-6.

③Fan J., Tong X., Zeng Y., Multi-agent Inference in Social Networks: A Finite Population Learning Approach, *Journal of the American Statistical Association*, Vol. 110, No. 509, 2015, pp. 149-158.

④陈云松:《观念的"割席"——当代中国互联网空间的群内区隔》,《社会学研究》,2022年第4期;徐翔,赦子棋,史静远等:《殊途同归:社交媒体用户内容生产中信息茧房趋同化——基于新浪微博的实证分析》,《西安交通大学学报》(社会科学版),2022年第3期。

⑤梁玉成,马昱堃:《对青年的计算文本"远读"——数字时代基于降维的整体认识论》,《青年探索》,2022年第3期。

⑥Matthes J., Hirsch M., Stubenvoll M.,et al. , Understanding the Democratic Role of Perceived Online Political Micro-targeting: Longitudinal Effects on Trust in Democracy and Political Interest, *Journal of Information Technology & Politics*, Vol. 19, No. 4, 2022, pp. 435-448.

⑦刘晗:《平台权力的发生学——网络社会的再中心化机制》,《文化纵横》,2021年第1期。

⑧Prummer A. , Micro-targeting and polarization, *Journal of Public Economics*, Vol. 188, 2020, p. 104210.

⑨张爱军,秦小琪:《网络意识形态去中心化及其治理》,《理论与改革》,2018年第1期。

形态、减少极化流量、强制沟通和隔离、基于大数据自动治理、微目标干预等手段改善网络舆论生态，使“多中心化”在社交媒体中得到遏制。

随着大模型的出现，网络生态将会出现新的挑战，酝酿新的意识形态风险[①]。大模型在训练及输出过程中都可能被注入特定的价值观念。例如，对ChatGPT的政治倾向进行测试后，发现它在所有政治倾向测试中都表现出了左倾[②]。因此，大模型可能会变革政治传播方式，介入公共性对话对社会观念的分化产生影响[③]。

为了应对大模型带来的意识形态挑战，我国在2023年8月15日起施行的《生成式人工智能服务管理暂行办法》中规定生成式人工智能服务需要坚持社会主义核心价值观，不得生成煽动颠覆国家政权、推翻社会主义制度，危害国家安全和利益、损害国家形象等内容[④]。在国家进行“去多中心化”的社会治理背景下，这一办法的出台有助于统合大模型的观念倾向，弱化不同大模型产品之间的观念差异，规避大模型生成非主流意识形态内容的可能性。这能够在一定程度上维护社会共识，促进达成社会整合的目标。

三、大语言模型影响社会观念分化的机制分析

（一）大语言模型对社会观念的影响基础

无论是通过社交机器人介入网络舆论场，还是人们直接通过大模型提出需求受其影响，这都表明大模型已经可以成为某种“信息行动”[⑤]的主体。针对大模型的价值观念塑造，既可以通过设置特定的提示词要求大模型输出符合指定观念的回答，也可以通过使用包含特定观念的文本数据对大模型进行微调（Fine-tuning），还可以通过人类反馈的强化学习（RLHF）对大模型内在的价值观念进行价值观对齐[⑥]。总之，大模型的输出很难做到纯粹的“观念无涉”。

大模型基于海量的文本数据训练，其内部的参数表征了文本中的潜在语义结构，是对知识的高度压缩。对数字时代活跃于网络空间的线上群体来说，使用大模型并不需要特殊的技术门槛，交互过程更加自然，获取知识更加便利。然而，大模型可能会输出错误信息、产生幻觉，甚至存在语种偏见[⑦]。虽然人们甄别信息的能力比大模型更强[⑧]，但仍然难以分辨大模型回答的正确性。有学者指出，人们互动时会表达自己观点的不确定性，而大模型会直接输出答案，这种细微的语义差异会导致它的回答看上去更加可信[⑨]。OpenAI曾在一份声明中呼吁公众更加谨慎对待获取到的任何

①代金平，覃杨杨：《ChatGPT等生成式人工智能的意识形态风险及其应对》，《重庆大学学报》（社会科学版），2023年第5期。

②Rozado D. ，The Political Biases of ChatGPT，*Social Sciences*，Vol. 12，No. 3，2023，p. 148.

③苏颖，汪燕妮：《生成式人工智能时代的政治传播走向——基于ChatGPT的讨论》，《党政研究》，2023年第3期。

④中国政府网：《生成式人工智能服务管理暂行办法》：https://www. gov. cn/zhengce/zhengceku/202307/content_6891752. htm，2023. 07. 10.

⑤陈氚：《信息行动理论——数字社会时代的社会行动理论探讨》，《社会学评论》，2021年第5期。

⑥Zhao W. X. ，Zhou K. ，Li J. ，et al. ，A Survey of Large Language Models，*ArXiv*，https://arxiv. org/abs/2303. 18223，2023. 10. 10.

⑦Luo Q. ，Puett M. J. ，Smith M. D. ，A Perspectival Mirror of the Elephant：Investigating Language Bias on Google，ChatGPT，Wikipedia，and YouTube，ArXiv，https://arxiv. org/abs/2303. 16281，2024. 08. 10.

⑧Spitale G. ，Biller-Andorno N. ，Germani F. ，AI model GPT-3（Dis）Informs Us Better Than Humans，*Science Advances*，Vol. 9，No. 26，2023，p. eadh1850.

⑨Kidd C. ，Birhane A. ，How AI Can Distort Human Beliefs，*Science*，Vol. 380，No. 6651，2023，pp. 1222-1223.

内容[①]。

(二)大语言模型对社会观念影响的机制假设

大模型对其用户群体可能存在对齐效应——只要使用同一模型,理论上都会得到相近的观点输出,观念相异的用户向大模型对齐,就具有了打破"信息茧房"的潜力。不同人群在大模型的观念调和下,更容易形成稳定的观念分化。因此,使用大模型的人群规模越大,越有可能导致社会观念的"去多中心化"。

然而,大模型目前仍然处在发展阶段,虽然人们总是倾向于使用效果最佳的大模型,但市面上仍然不断出现新的产品。如果不同的大模型本身就表征了不同的观念,并且由于它们能捕捉到文本的潜在语义关联,会产生更多的观念元素,那么不同人群使用不同的大模型,就仍然有可能维持多中心的社会观念分化格局。此时,大模型难以起到"去多中心化"的效果。此外,在大模型具有普遍的社会影响之前,有一个逐渐普及的过程,并且在它流行之前,社会已经形成了一定的观念分化格局。换句话说,在研究大模型对社会观念产生的影响时,还需要考虑已有的观念分化情况。如果大模型在已经形成了观念分化格局的情况下介入互动,则在形成了一些子群体的情况下,大模型更加容易发挥"对齐"作用,从而进一步实现"去多中心化"。

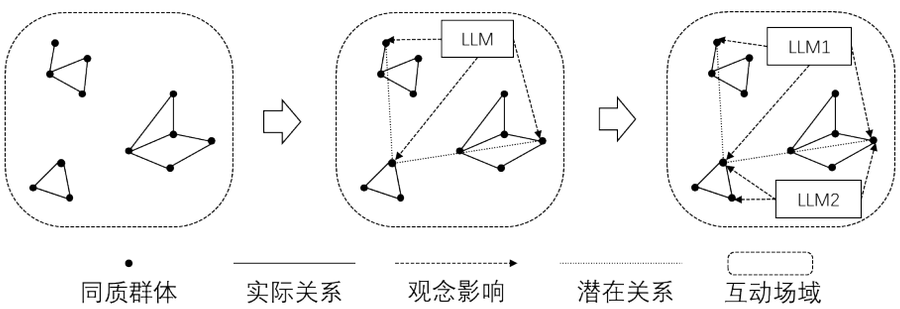

图1　大语言模型(LLM)介入社会互动的发展过程示意图

综上,大模型对社会观念的分化影响,应当和大模型在社会中的普及和使用程度,大模型的观念类型数量,以及模型在介入互动之前已经形成了何种程度上的观念分化格局等因素有关。据此,本文提出以下研究假设。

假设1(普及使用程度假设):大模型在社会中的普及率越高,社会观念的分化程度越低,越可能形成社会观念"去多中心化"的格局。

假设2(观念类型差异假设):不同大模型之间的观念类型差异越大,社会观念的分化程度越高,越可能形成社会观念"多中心化"的格局。

假设3(观念元素扩增假设):大模型增加的观念元素数越多,社会观念的分化程度越高,越可能形成社会观念"多中心化"的格局。

假设4(模型混合使用假设):如果存在同一人群同时使用多种大模型的情况,社会观念的分化程度越高,越容易形成社会观念"多中心化"的格局。

假设5(既存观念分化假设):大模型介入时社会观念已经形成一定的分化格局,越可能形成社会观念"去多中心化"的格局。

① Solaiman I., Brundage M., Clark J., et al., Release Strategies and the Social Impacts of Language Models, *ArXiv*, https://arxiv.org/abs/1908.09203, 2024.08.23.

四、大语言模型影响观念分化的自主行动者建模研究设计

要研究大模型对社会观念分化的影响，就需要考虑它的普及程度、不同模型的观念类型差异，以及已经形成的社会观念分化格局。大模型的发展还在继续，较难收集实证数据对理论进行实证检验，因此本文采用自主行动者建模仿真的计算社会学方法，对大模型如何影响社会观念进行探索性分析。ABM是一种基于行动者互动的仿真方式，是一种对社会过程进行自下而上模拟的方法[①]。通过在计算机中设置多个基于特定规则的自决策行动者个体并让其互动，达到在多轮迭代计算中探索社会过程和机制的研究目的。它在社会预测方面具有较大的潜力[②]。

（一）仿真模拟实验设计

社会观念可以被视为一种文化范畴，它属于构成人们认知世界的文化图式。对于社会整体而言，社会成员所拥有的文化图式是多元的，在文化传递和传播过程中会出现变异[③]。因此，社会观念的传播机制和文化的传播机制类似，也是在社会互动过程中，通过信息载体对文化图式中的元素进行交换并发生变异，借此形成个人观念的更新，从而形成相似观念的群体聚集。因而，针对文化传播的ABM研究也可以被应用于社会观念的分化研究。

1. 阿克塞尔罗德文化传播模型

罗伯特·阿克塞尔罗德使用ABM对文化传播过程做了形式化抽象，他的文化传播（Culture Dissemination）模型揭示了不同的行动者群体如何在局部互动中形成了整体上的文化分化[④]。他将每个行动主体的“文化组”表示为一个多维向量，向量的每个维度表示文化所体现的各种特征（Features），每个维度又用整数编码表示若干特质（Traits）[⑤]。为了减少混淆，本文将前者称为文化“维度”，将后者称为文化“特质”。举例来说，语言就是一种文化“维度”，而不同方言就是语言维度下的不同文化“特质”。

每一次互动共有两个步骤。首先，随机选取一个行动主体，让它随机与其“上下左右”方向的邻居之一进行文化组比较。其次，如果两者在若干相互对应的文化维度上有重叠，就以重叠维度数占总维度数的比例（重叠度）作为互动概率，随机在该行动主体与邻居不重叠的文化维度上，将自身的文化特质置换为邻居对应维度的文化特质，如图2所示。

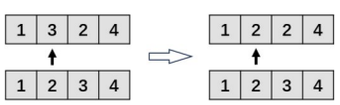

图2　文化组向量特征交换过程示意图

文化越接近的两个行动者群体，越容易在互动中发生同质化，这会导致二者的文化表征进一步趋同。当行动者之间完全达成共识或完全没有共识，互动终止。在文化传播模型的基础上，本文在模拟仿真中将大模型设置为能够对其他行动者群体进行单向影响的行动主体。

2. 大语言模型影响社会观念的仿真实验设计

虽然大模型能够覆盖海量的人类知识，可以捕捉到文本的潜在语义关联，但它不会发明新的

---

①梁玉成，贾小双：《数据驱动下的自主行动者建模》，《贵州师范大学学报》（社会科学版），2016年第6期。

②Chattoe-Brown E. , Is Agent-based Modelling the Future of Prediction? , *International Journal of Social Research Methodology*, Vol. 26, No. 2, 2023, pp. 143-155.

③胡安宁：《社会学视野下的文化传承：实践—认知图式导向的分析框架》，《中国社会科学》，2020年第5期。

④Axelrod R. , The Dissemination of Culture: A Model with Local Convergence and Global Polarization, *Journal of Conflict Resolution*, Vol. 41, No. 2, 1997, pp. 203-226.

⑤文化组向量可以看成一组定类变量的组合。

观念维度。但是,正因大模型的训练参数大,训练语料多,所以它在海量文本中捕捉到的潜在语义关联能够拓展已有观念维度的特质数量。因此,本文假定大模型不会增加观念维度数,但会增加每个维度的观念特质数。在整个互动过程中,大模型的观念组向量不会发生变化。

在实验中,大模型对行动者群体的影响规则同样遵循行动者群体间的互动规则。行动者群体与大模型的观念组重叠度越高,他们越容易采纳大模型输出的观念,越容易受其影响。如果行动者群体的观念组与大模型差异太大,它们受到影响的概率就会较低。本文所用的ABM模型使用NetLogo仿真软件实现[①]。NetLogo的行为空间实验工具箱可以自动调整输入参数的取值,遍历所有参数设置组合。本文对同一参数组合的实验重复运行2次,最终全部的实验次数为19200次。所有实验均由计算机随机初始化,每次实验的随机数种子都不同,保证了各次实验初始条件的随机性[②]。

(二)实验结果分析方法

1. 因变量

因变量为每轮实验结束时存留的观念组数。每次随机初始化时,每个行动者群体的观念组都不同,共有100种观念组,在多轮互动迭代之后,存留的观念组数越少,表示社会观念的分化程度越低。

2. 自变量

(1)大模型普及率。大模型普及率越高,表示使用大模型的行动者群体越多。(2)大模型观念组数。目前大模型基本都由平台企业各自训练,在语料选择和价值观对齐上存在差异,因此考虑多种不同观念组的大模型对社会观念分化的影响。(3)大模型扩增的观念特质数。大模型会增加每个观念维度的观念特质取值空间,将其对观念特质的增加数也作为自变量。(4)大模型加入之前的互动次数。社会观念在互联网空间的传播和演化先于大模型对社会施加影响,只有在最理想的情况下,大模型才是在实验初始化时介入互动的,而现实社会更接近于已经经历过多轮互动之后的实验设置。(5)是否存在同时使用多种大模型的情况。现实社会中会存在一个用户同时使用多个大模型的情况,这些大模型的观念类型既可能相似,也可能有差异。

3. 控制变量

阿克塞尔罗德使用文化维度数和每个维度的特质数作为自变量,本文将其作为控制变量。然而,在模型设定上,观念特质数与观念维度数的比值越大,行动者群体之间存在观念重叠的可能性越低[③]。本文将上述比值定义为"单维多样性",作为控制变量。此外,对于能够在达到运行时间轮数上限之前收敛的实验,在分析大模型加入之前的互动次数时,也把实验的运行轮数加入模型作为控制变量。

4. 分析模型选择

本文的主要因变量为多轮互动后存留的观念组数,也即观念组向量的种类数。由于该变量为离散的计数变量,其分布形态近似泊松分布,并且在数据探查中发现,其方差显著大于均值,存在过度离散的情况,因此采用负二项回归模型对其影响因素进行分析。除此之外,在仿真模拟阶段,

①NetLogo是美国西北大学互联学习和计算机建模中心(Center for Connected Learning and Computer-Based Modeling, CCL)的Uri Wilensky领衔开发的复杂系统仿真建模软件,能够对多种自然和社会现象进行仿真模拟。NetLogo代码参考并检验了Daniel Weissglass在NetLogo官方社区提交的示例,在其基础上复现文化扩散模型并根据本文研究设计修改代码。参见:http://ccl.northwestern.edu/netlogo/,https://ccl.northwestern.edu/netlogo/models/community/Axelrod-Basic,2024.08.24.

②参数设置和逻辑步骤报告,可联系作者提供。

③Axelrod R., The Dissemination of Culture: A Model with Local Convergence and Global Polarization, *Journal of Conflict Resolution*, Vol. 41, No. 2, 1997, pp. 203-226.

本文发现在加入大模型的情况下，模拟结果会分为收敛和不收敛两种结局。为了分析大模型特征对这一模拟结果的影响，本文还使用了logit模型。

五、研究结果

（一）描述性统计

表1　样本分布与观念组存留数量

	终止原因	样本分布		观念组存留数量				
		样本量	百分比	均值	标准差	最小值	中位数	最大值
无大模型	轮数限制	0	0.0%	/	/	/	/	/
	互动终止	160	100.0%	5.0	10.5	1	1	52
	总计	160	100.0%	5.0	10.5	1	1	52
有大模型	轮数限制	10,967	57.1%	14.8	9.8	2	12	74
	互动终止	8,233	42.9%	10.5	13.8	1	4	72
	总计	19,200	100.0%	12.9	11.9	1	9	74

首先验证了大模型未介入的模拟实验信度，结果与阿克塞尔罗德原文一致，而有大模型介入的实验中，约有60%的实验无法自发终止。经多次测试，将每次实验的最大运行轮数上限调整为5000。在这一条件下，能够自发停止、形成稳定观念分化格局的实验为有序结局，其中有99%以上都会在运行2200轮之后终止。将运行轮数达到5000次还未能终止的实验视为无序结局，这些实验无法形成稳定的观念分化格局。

表1区分了大模型是否介入的情况，统计不同终止原因的实验结果。在有序结局下，行动者群体之间会因为形成了稳定的观念区隔后终止互动，并且最终存留的观念组数均值和中位数都比无序情况要少，更容易形成同质观念。在无序结局下，不同社会观念处在持续的互动过程中，难以形成整体层面的同质观念。

（二）回归模型

首先分别对收敛与不收敛的两种实验结果的观念组存留数量的影响因素进行分析，之后进一步分析影响实验结果是否收敛的因素。

1. 观念组存留数的影响因素分析

表2描述了实验结果能够收敛时，停止互动后观念组数的分析模型，使用负二项回归模型。

表2　观念组存留数量的影响因素负二项回归模型（有序结局）

因变量：终止时观念组数	模型1	模型2	模型3	模型4
控制变量				
观念组维度数	−0.1766***	−0.0459***	−0.0445***	−0.0466***
	（0.0014）	（0.0025）	（0.0025）	（0.0025）
行动者群体观念特质数	0.1365***	0.0267***	0.0262***	0.0288***
	（0.0012）	（0.0021）	（0.0021）	（0.0021）
单维多样性		0.0071***	0.0071***	0.0069***
		（0.0001）	（0.0001）	（0.0001）

续表

因变量:终止时观念组数	模型1	模型2	模型3	模型4
运行轮数				-0.0001***
				(0.0000)
自变量				
大模型普及率			0.0057***	0.0051***
			(0.0003)	(0.0003)
大模型观念组种类数			0.0277***	0.0367***
			(0.0041)	(0.0040)
大模型观念特质增加数			-0.0014	-0.0009
			(0.0009)	(0.0009)
同时使用多个大模型(是=1)			-0.0369***	-0.0335***
			(0.0099)	(0.0097)
大模型进入前的互动轮数				-2.6025***
				(0.1372)
常数项	1.7619***	0.7076***	0.4523***	0.6599***
	(0.0214)	(0.0275)	(0.0343)	(0.0349)
样本量	8,233	8,233	8,233	8,233
Pseudo_R^2	0.2543	0.3073	0.3146	0.3230

注:括号内为标准误,*** p<0.01, ** p<0.05, * p<0.1。下同。

模型1和模型2为基准模型,从中可见,与原始的文化传播模型一致,观念组维度数越大,观念组存留数越少;观念特质数越多,观念组的存留数也越多。在加入单维多样性之后,前述两个因素的系数绝对值变小,但系数方向没有发生变化。单维多样性越大,存留的观念组数越多,这与逻辑推理的结论一致。

模型3加入了大模型相关的自变量。大模型的普及率对于观念组存留数具有显著的正向影响。使用大模型的行动者群体越多,社会观念的分化程度越大。在实验中,行动者分布在一个10*10的正方形网格中,大模型会随机对其中的部分行动者群体产生影响。大模型的介入将行动者群体根据是否受其影响"一分为二",这两者之间可能减少交流,并且不使用大模型的行动者群体之间的交流通道会被"切割",这更易造成分化。受大模型影响的区块最终价值观念会被其同化①。大模型的观念组种类数对于社会观念的分化程度也具有显著的正向影响。它增加的观念特质虽然与分化程度呈现负相关,但系数并不显著。这说明大模型给社会观念造成分化的原因并非由于在某个单独的观念维度增加新的观念特质,而是因为不同模型给用户传递的观念类型本身就不同。对有序结局来说,行动者群体同时使用多个大模型会减少观念组存留数。

模型4在控制每次实验的运行轮数之后,发现大模型介入观念互动的时间越晚,观念组存留数越少。这说明,如果社会观念本身已经存在一定程度的分化格局,那么大模型的介入会进一步弱化社会观念的分化,体现了对齐效应。表3描述了实验结果无法收敛时,达到运行轮数上限终止后观

①随着大模型普及率越来越高,直至所有行动者群体都使用大模型,观念分化程度会再次下降。观念分化程度与大模型普及率的关系呈"倒U型",加入更多实验参数会倍增实验总次数,没有呈现这些实验结果。

念组数的分析模型，本文使用负二项回归模型。

表3　观念组存留数量的影响因素负二项回归模型（无序结局）

因变量：终止时观念组数	模型1	模型2	模型3	模型4
控制变量				
观念组维度数	−0.0502***	−0.0082***	−0.0060***	−0.0060***
	（0.0010）	（0.0017）	（0.0016）	（0.0016）
行动者群体观念特质数	−0.0060***	−0.0426***	−0.0423***	−0.0423***
	（0.0009）	（0.0016）	（0.0015）	（0.0015）
单维多样性		0.0039***	0.0040***	0.0040***
		（0.0001）	（0.0001）	（0.0001）
自变量				
大模型普及率			−0.0005	−0.0005*
			（0.0003）	（0.0003）
大模型观念组种类数			0.1177***	0.1177***
			（0.0047）	（0.0047）
大模型观念特质增加数			−0.0223***	−0.0224***
			（0.0009）	（0.0009）
同时使用多个大模型（是=1）			0.0493***	0.0493***
			（0.0096）	（0.0096）
大模型进入前的互动轮数				−0.5563***
				（0.1361）
常数项	3.4178***	2.8517***	2.6169***	2.6454***
	（0.0185）	（0.0260）	（0.0344）	（0.0351）
样本量	10,967	10,967	10,967	10,967
Pseudo_R^2	0.0313	0.0422	0.0582	0.0585

从模型1和模型2可以看出，与有序结局不同，无序结局中，观念组维度数和观念特质数都与最终的观念组存留数负相关。但单维多样性与观念组存留数正相关。从模型3和模型4中可见，大模型普及率在无序结局下对观念组数的影响显著性较弱。与有序结局相比，在无序结局的实验过程中，不同观念的行动者群体在持续地进行观念交流，虽然仍然存在被大模型影响的行动者簇，但它的边界并不固定，不会阻碍不同群体之间的交流。大模型的观念组种类数越多，最终存留的观念组也越多，并且系数大于有序结局。而随着大模型增加的观念特质数越多，观念组数会减少。如果存在同一行动者群体同时使用多种不同大模型的情况，观念组存留数会增加。模型4表明大模型介入互动之前的时间越长，观念组存留数越少。这与有序结局的结论一致。

总之，综合比较有序结局和无序结局的回归结果，发现大模型的观念组数越多，最终观念组存留数越多，并且大模型介入互动的时间越晚，观念组存留数就越少。但是，在大模型的普及率、增加的观念特质数量以及是否存在同时使用多个大模型方面，两种实验的情况并不一致。为此，我们需要进一步探究上述因素对于实验结局是否有序的影响。

2. 实验结局是否有序的影响因素

使用logit模型对实验结局是否有序的情况进行分析，结果如表4所示。

表4　实验结局是否有序的影响因素logit回归模型

因变量:是否有序(是=1)	模型1	模型2	模型3	模型4
控制变量				
观念组维度数	-0.0943***	-0.0331***	-0.0675***	-0.0675***
	(0.0028)	(0.0052)	(0.0073)	(0.0073)
行动者群体观念特质数	0.0432***	-0.0096**	-0.0206***	-0.0206***
	(0.0027)	(0.0047)	(0.0066)	(0.0066)
单维多样性		0.0051***	0.0103***	0.0103***
		(0.0004)	(0.0005)	(0.0005)
自变量				
大模型普及率			-0.0089***	-0.0089***
			(0.0013)	(0.0013)
大模型观念组种类数			-1.6544***	-1.6547***
			(0.0234)	(0.0234)
大模型观念特质增加数			0.0746***	0.0746***
			(0.0039)	(0.0039)
同时使用多个大模型(是=1)			0.0157	0.0157
			(0.0430)	(0.0430)
大模型进入前的互动轮数				0.9603
				(0.6078)
常数项	0.3293***	-0.4179***	3.5288***	3.4815***
	(0.0492)	(0.0744)	(0.1394)	(0.1425)
样本量	19,200	19,200	19,200	19,200
Pseudo_R^2	0.0556	0.0626	0.4763	0.4764

上表展示了影响结局是否有序的因素分析。在原始的文化传播模型中,所有实验都是有序结局,观念组维度数和观念特质数并不会影响实验结局。但上述模型1到模型4的结果表明,在大模型介入互动的情况下,行动者群体的观念组维度数量和特质数量均会对实验结局有序与否产生显著影响,这说明,大模型的介入,会与社会既存的观念维度与特质相互作用,从而影响社会观念分化情况。

如果"观念组维度"表示人们能够介入讨论的议题种类,"观念特质"表示在这些议题上发表的具体观点,那么在大模型介入互动的条件下,议题种类越多,在每个议题上可选的观点相对更多,想要在各种议题都达成共识的复杂性就会上升。虽然单维多样性增加会促进观念分化格局有序,但根据表2的分析结果,这会更易导致社会整体观念差异更大的结果,加深分歧。大模型会持续激活人们对不同议题的讨论,从而降低形成有序结局的概率。

模型3加入了大模型对社会观念分化影响的特征变量。大模型的普及率越高,实验结果越不易有序,这也佐证了大模型会和社会现存的观念维度与特质产生相互作用的讨论。大模型的观念组种类数越多,实验结局越不容易有序;大模型增加的观念特质数越多,有序结局的概率越大;行动

者群体是否同时使用多种类型的大模型，以及大模型进入之前的实验轮数对结果的影响并不显著。

以上结果表明，大模型在介入社会互动场域后会与既存的社会观念发生相互作用。在其他条件不变的情况下，减小大模型的观念组差异，大模型引入新的观念特质越多，社会观念越会倾向于形成更加稳定的分化格局。如果大模型之间存在较大的观念差异，那么就难以形成稳定的社会观念分化格局，不同行动者群体之间将会存在普遍的互动和交流。当然，如果大模型的观念组会随时间变化，这种影响可能会发生变化。

（三）小结

综上所述，模拟结果分化出了两种不同的结局，对假设的支持情况存在差异。

如表5所示，假设2和假设5得到了支持，与机制分析的推测相符合。不同大模型的观念差异越小，介入互动时社会观念已经形成了一定的分化格局，“去多中心化”的效果就越明显。同时，不同大模型的观念差异越小，越能促进社会观念形成有序的格局。这些结果表明，《生成式人工智能服务管理暂行办法》对不同大模型产品的规范，能够在一定程度上防止社会观念的过度分化。

表5　大语言模型影响社会观念分化的机制假设支持情况[①]

	假设内容	对应变量	观念组存留数		有序结局概率
			有序结局	无序结局	
假设1	普及使用程度假设	大模型普及率	+*** 与假设相反	−* 显著性较弱	−***
假设2	观念类型差异假设	大模型观念组种类数	+*** 符合假设	+*** 符合假设	−***
假设3	观念元素扩增假设	大模型观念特质增加数	− 不显著	−*** 与假设相反	+***
假设4	模型混合使用假设	同时使用多个大模型	−*** 与假设相反	+*** 符合假设	+ 不显著
假设5	既存观念分化假设	大模型进入前的互动轮数	−*** 符合假设	−*** 符合假设	+ 不显著

假设1、假设3和假设4的情况相对复杂。对假设1来说，由于大模型的介入会分化出使用大模型和不使用大模型两个群体，这两者之间的观念差异可能会逐渐拉开。并且大模型对不同观念之间的讨论交流起到激活作用，普及率升高会增加社会形成无序分化格局的概率。即便在有序结局的结果中，普及率升高也会增加观念的种类，倾向“多中心化”。这种激活不同观念交流讨论的作用还会与社会既存的观念维度与特质相互强化。对假设3来说，大模型对观念元素的扩增会提升形成有序格局的概率，但无法表明它能够促进形成“多中心化”的格局，相反，观念元素的增多反而可能减少最终的观念组数量，形成“去多中心化”。对新特征的接纳可能体现了大模型对不同群体的对齐作用。对假设4来说，它对实验结果是否有序的影响并不显著。在无序结局下，存在同时使用多个大模型的情况会导致更多的观念组数，符合假设；在有序结局下，同时使用多种不同的大模型反而会减少观念组数。这可能是由于，如果某一行动者群体同时使用的多种大模型本身存在较大的观念差异，那么他们最终可能更倾向于选择与自身观念更加接近的产品，但这一结论的证实需要后续更多的研究。

①表5仅汇总了系数的正负情况和显著性。

六、结论与讨论

本文通过理论与仿真驱动的方式,对了解大模型如何影响社会观念尝试了初步的探索。研究发现,首先,大模型对社会观念产生的影响,与它的普及程度、观念类型差异、介入社会的次序以及大模型的观念特质数均有关系。大模型观念类型异质性较高时,可能会不断激活不同观念之间的交锋和讨论,社会观念整体上既无法形成稳定的多中心,也无法达成共识。

其次,由于可以通过价值观微调的方式塑造大模型的隐含观念,多种不同的大模型如果从语料当中"学到"了有差异的观念,或者由于微调而产生了各不相同的观念,这可能会削弱社会整合。在国家进行"去多中心化"的社会治理背景下,《生成式人工智能服务管理暂行办法》的出台有助于促进国内大模型的研发和使用处于可控的观念差异范围内,从而在一定程度上引导大模型产品促进社会共识的形成,进而达成社会整合的目标。

再次,互联网平台企业会根据用户的使用反馈微调大模型产品,可能会让大模型的观念随时间漂移,本文并未考虑这种情况。主要原因是,本文是对社会观念互动过程的形式化抽象,由于无法获知大模型提供商的调整结果,无法判断大模型的文化观念如何随时间产生变化。此外,行动者规模和邻居数量对观念分化的影响并非本文的分析重点[①],纳入实验会大幅增加实验模拟的次数,因此本文没有加以考虑。本文意在基于文化传播模型来探索大模型介入社会互动后对社会观念分化的可能影响,因此只选取了文化传播模型的基本特征,把研究重点放在对新现象的探索上。

最后,对于大模型介入社会互动将会形成怎样的长远影响,在此进行一些讨论。本文的结论能够说明,如果社会上普及的大模型观念差异较小,那么它们有助于人们跳出既有的"局部茧房"。然而,由于大模型的便利性而导致的过度依赖,可能造成社会形成"整体茧房"。由此,大模型可以生成知识内容并发布在社交媒体上,这种方式对于社会观念的影响非常隐蔽。如果没有清晰标识出内容由机器生成,人们也会更容易选择相信。如果这些内容反过来被当作大模型的训练语料,甚至可能对大模型的性能产生影响。已有研究表明,如果基于大模型生成的内容继续训练大模型,可能会造成模型坍缩(Model Collapse)[②]。大模型如果想要达到很好的效果,不仅需要高效的算法设计,也需要使用高质量的训练数据。但还有研究指出,高质量的语言数据可能会在可见的几年之内耗尽[③]。

[责任编辑:王文娟]

①Axelrod R., The Dissemination of Culture: A Model with Local Convergence and Global Polarization, *Journal of Conflict Resolution*, Vol. 41, No. 2, 1997, pp. 203-226.

②Shumailov I., Shumaylov Z., Zhao Y., et al., AI Models Collapse When Trained on Recursively Generated Data, *Nature*, Vol. 631, No. 8022, 2024, pp. 755-759.

③Villalobos P., Sevilla J., Heim L., et al., Will We Run Out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning, *ArXiv*, https://arxiv.org/abs/2211.04325, 2024.08.24.