大语言模型的信任建构

胡晓萌　　陈力源　　刘正源

摘　要：以 ChatGPT 为代表的 AI 大语言模型技术快速兴起，在颠覆现在内容生产方式和智能技术范式的同时，也由于幻觉、虚假内容等问题带来了信任危机。该技术甚至因为信任危机问题遭到抵制和封杀。尽管业界已在可信 AI 方面积极开展了大量的技术实践，但公众对 AI 的信任度仍未显著提升。因此，要解决信任问题，不仅需要厘清信任与可信任的关系，还需要从大语言模型的技术本质出发进行探究。对大语言模型技术的信任应是认知信任，认知信任不仅包含技术信任与人际信任的动态交互，而且是建立在有效监督基础上具有合理性的信任。大语言模型信任的建构路线主要包括以可解释性为核心的信任要素体系，以政府主导的 AI 治理体系为基础、多元主体协同的信任主体和信任环境，以及培养人们正确信任观的信任认知三个模块。

关键词：人工智能；大语言模型；信任；可信任

中图分类号：TP18　　**文献标识码：**A　　**文章编号：**1003-0751(2024)05-0171-06

以 ChatGPT 为代表的 AI 大语言模型(LLMs)是一项具有革命性的技术。它不仅可以像之前的人工智能一样进行分类、预测，还可以通过自然语言与人类对话，生成文本、图像、视频、可执行代码等各种形式的内容，这将对人们的生产生活和社会发展产生深远影响。但人工智能开发和应用阶段如果出现错误，可能会带来灾难。现在大语言模型面临诸多信任挑战，如人们越来越无法分辨出 ChatGPT 生成的内容与人类生成的内容；大语言模型存在幻觉问题，会生成错误、具有诱导性的内容，人们该如何分辨并信任大语言模型生成的内容；大语言模型还存在偏见、歧视、隐私侵犯、有害言论等多方面伦理风险，继而带来一系列信任危机，甚至遭到业界的抵制和封杀。信任是人工智能发展的一个核心问题，人与技术之间信任关系更是技术发展趋势与人类未来的一个核心问题[1]。AI 大模型是一种变革性技术，但它只有在得到信任的情况下才能充分发挥潜力。过去人工智能科学家和科技企业提出可信人工智能的技术框架并开展了大量研发实践，然而，即使科技企业认为他们的人工智能系统是值得信赖的，个人或团体是否愿意冒险并给予它们信任仍然悬而未决[2]。可见，从可信到信任，仍然存在较大鸿沟。AI 大语言模型正在引领通用人工智能甚至超人工智能的到来，但是信任问题已经成为大语言模型技术创新与应用的阻碍。IBM 调研全球近 1000 名企业高管，研究表明阻碍其企业部署应用 AI 大模型的主要因素是信任问题[3]。可见，建立对大语言模型的信任成为技术创新发展与应用的必要前提。

一、从可信 AI 到信任 AI 需跨越的鸿沟

虽然业内积极开展了可信 AI 的实践探索，但距离达成信任 AI 仍存在距离。因此，需要厘清信任与可信任的关系，并且回到技术本身，从其技术特性和

收稿日期：2024-02-27

基金项目：中国博士后科学基金第 73 批面上资助项目(2023M732381)。

作者简介：胡晓萌，男，清华大学社会科学学院博士后研究人员(北京　100084)。陈力源，男，通讯作者，同济大学上海市人工智能社会治理协同创新中心研究员(上海　200082)。刘正源，男，清华大学社会科学学院助理研究员(北京　100084)。

人机关系来寻找建立信任的入口。

1.可信AI的发展现状及困境

2017年,何积丰院士在全球范围内首次提出“可信人工智能”(Trustworthy AI)概念。可信AI研究主要包含安全性、可解释性、公平性、隐私保护等方面,这些方面也成为评价一个人工智能系统是否值得信赖的技术评估标准[4]。可信人工智能已经成为学术界和产业界的共识[5]。一是可信人工智能领域论文发表量在人工智能论文发表量占比逐步提升,其中,美国、中国、英国是全球可信人工智能论文发表的主要国家,这与国家人工智能科研实力、产业发展和应用水平密切相关。二是可信人工智能的专利申请量在快速增长。在专利技术分布上,隐私保护、安全性方面的专利占比较大,可解释性和公平性的专利占比较小。三是各国政府、国际标准组织和行业组织积极推动建设可信人工智能原则和标准,如欧盟《可信人工智能伦理指南》、美国《促进政府使用可信人工智能》、ISO/IEC《信息技术 人工智能 人工智能可信度概述》、IEEE《可解释性人工智能结构框架指南》和《金融服务可信数据和人工智能系统》以及中国人工智能产业发展联盟的《可信AI操作指引》等。这些原则和标准,都在指引从业者如何研发设计出一个可信的人工智能系统。四是人工智能产业积极落地可信AI,贯穿研发、应用、运营等全部流程。如可信AI在智能医疗设备、智能驾驶、智能机器人等产品设备领域均已落地,建立了安全性、合规性、可解释性等兼具的功能体系。

虽然业界在可信人工智能方面做了大量努力,但人们对人工智能的信任状况仍然不容客观。2022年4月,香港消费者委员会发布的一项研究结果显示,亚洲市场上只有31%的消费者表示他们信任人工智能[6]。2023年,毕马威与昆士兰大学就人们是否信任人工智能联合开展了全球调查研究,选取了17个国家和地区超过17000人的样本数据,结果显示,有50%的人不信任人工智能,其中,人们通常不信任、不接受人工智能在人力资源中的应用[7]。

2.信任与可信任的混淆与关系厘清

跨越可信任到信任的鸿沟,需要厘清信任与可信任之间的关系。可信AI着眼于“是什么让人工智能值得信任”,其理论前提是“我们信任可信任的事物”。然而,信任的运作方式并不如此,这就不免存在混淆信任与可信任的问题[8]。

西方学者认为信任是一种心理状态,包括一方对另一方意图或行为的积极预期以及愿意接受损失的意愿。信任涉及委托人和受托人的关系以及委托人对受托人履行期望的信念,它包含两种要素:信心、承诺。然而,人们可能信任完全不值得信任的事物,也可能不信任完全值得信任的事物,所以信任与可信任是两码事[9]。一是人们信任人工智能可能与可信AI的几大要素毫无关系。如企业在设计AI语言助手时,会从声音、外形等角度考虑,使人们更信任AI。机器人的声音要变得“可爱”,这样才能为人们所接受和信任,而女声比男声有优势,这就是为什么智能助手中女声占绝大多数。又如,近期在一场图灵测试中,麻省理工学院团队在1966年开发的AI聊天系统ELIZA中打败了OpenAI开发的最先进的大语言模型GPT4,获得了更高的通过测试成功率,一些测试者认为ELIZA表现得太差,以至于不可能是当前先进的AI大模型,更可能是一个故意不合作的人类[10]。二是信任和可信度是脱节的。某事可能完全值得信任,但人们仍然不会信任它,因为信任是可以赢得的,但也必须被给予。也就是说,可信AI理论预设的“可信任会自动实现信任”并不一定会实现。开发人员在AI产品的可信方面做出大量努力之后,用户可能仍然不信任它,更不用说使用它,这会让开发人员非常有挫败感。三是可信存在被泛化的问题,与建立信任的初衷不一致。不管是可信AI的技术要求,还是各国出台的可信AI指南、原则的文件中,“可信”都与一系列原则相关联。这里似乎表明了一切被认为“好”的东西都应该促成信任,并且这些“好”的东西被视为AI值得信任的必要条件。这会导致人们期望可信AI能够符合人们认为在人际、社会和政治层面上应该符合的公平、正义、可控等所有原则,从而将可信变成了一个不适用或不可操作的流行词。

综上可见,信任与可信任的关系并不是可信AI理论预设的“可信任的AI会获得人们的信任”,但是这也并不能完全否定可信AI为建立信任的积极作用,应当重新思考人们在何种意义上、何种情况下会信任AI,人与AI的信任关系是什么,除了现有的可信AI,还需要做什么。

3.作为一种认知技术的大语言模型

为什么可信人工智能领域的诸多努力似乎并没有达成人们信任人工智能的目的?可能需要从人工智能这项技术的本身来寻求答案。以自动驾驶汽车为例,人对自动驾驶汽车的信任不同于人对传统汽车的信任。在传统汽车的信任场景中,人们只需要知道汽车是安全可靠的,便可以信任地驾驶这辆汽

车,这种对技术的信任相对简单、易把握;而在自动驾驶汽车的信任场景中,自动驾驶汽车被信任的是执行一项认知任务,如行驶决策等,这里的信任则包含了更多复杂的要素。可见,需要被信任的人工智能不同于以往的技术,它能够执行数据分析、语言理解和生成、知识获取、概率评估、分类、预测、推理和映射等认知任务。人工智能本质上是一种认知技术,是对人的认知能力的延伸和增强。以 ChatGPT 为代表的大语言模型作为一种前沿的认知技术,正在引领一场知识和文化的革命。这些模型通过处理和生成自然语言,不仅重塑了知识的生产、获取和传播方式,而且正在渗透到人类文明的核心,影响着人类的思维、信仰和文化构成。大语言模型在几乎涵盖了互联网和人类所有知识库的大数据基础上进行训练,具备理解和生成自然语言的能力,它不仅能够捕捉自然语言的复杂性和多样性,而且能够在某种程度上理解语义和语境。语言是构成人类认知和文化的基石,也是人们传递信息、信仰和价值观的主要工具。而 ChatGPT 和更先进的 AI 大模型正在以超过人类水平的能力掌握语言,能够理解、生成和处理自然语言就有能力参与到文化创造和传播中。

人类以“知识—信念”为基础建立了一套认知框架,而文化是在这套认知框架基础上建立起来的“认知符号系统”。它不仅是一套传统、习俗和道德观念,还是一种高度复杂的信息编码方式,可以在人类社会传播,并影响个体和群体行为。文化是人们用来解释、理解和互动现实世界的一种“模拟现实”。每一种文化都是对现实的一种解释和评价,它涉及价值观、信仰、人类关系甚至人类与自然、宇宙的关系。而未来所有知识,无论是过去的、现在的还是未来的,都有可能通过单个通用大模型获得。目前,大模型破解了人类语言,成为知识的权威来源,并逐步展示出其作为文化构建者的潜力。不仅未来有可能成为文化的一部分,甚至可能改写人类文化和认知的基础规则。那么,当人们使用大语言模型做某项工作时,按照唐·伊德(Don Ihde)的人与技术关系理论[11],人与大模型可能呈现为三种关系:一是具身关系[(人—大模型)→世界],即大模型作为人的功能性延展,去认知世界并实践;二是诠释关系[人→(大模型—世界)],即人类认知世界,是经过大模型的表征、转换或诠释;三是它异关系[人→大模型—(世界)],侧重技术的自主性,即大模型成为认知的客体时,透过大模型展现的世界就变成了一种技术人工物。作为认知技术,大语言模型重塑了人类的认知方式,促使我们重新思考和定义人机关系及其信任基础。

二、认知信任:一种对大语言模型的特殊信任

以认知信任概念为基础,对大语言模型这样一种认知技术的信任是一种基于人际信任和技术信任的动态交互,是一种基于有效监督的具有合理性的认知信任。

1.认知信任的构成

托斯顿·威尔霍尔特(Torsten Wilholt)在探讨科学作为集体认知事业成功运作所必需的要素时提出了认知信任(Epistemic Trust)概念。他认为,认知信任是一种严格根据接收者作为信息提供者或知识传递者的能力来分配的信任[12]。按照认知信任的定义,拉蒙·阿尔瓦拉多(Ramón Alvarado)认为,既然人工智能作为一种认知技术能够扩展增强人类的认知能力,那么人们信任人工智能时就是在相信它作为信息提供者或认知增强工具的能力,因此人们给予人工智能的信任应是一种认知信任[13]。

认知信任包括四个组成部分:信念、沟通、依赖和信心[14]。其中,信念和沟通具有独特的认知属性,依赖和信心是构成任何类型信任的核心。信念是认知信任的基础,指信任方对被信任方提供的信息或知识的真实性和可靠性持有的认知态度。这种信念基于对被信任者知识、能力或诚信的评估。沟通是建立和维持认知信任的关键机制,涉及信息的传递和接收。良好的沟通能够增强信任,因为它帮助信任方理解被信任方的知识基础、推理过程和结论,如人工智能的透明度和可解释性。依赖是信任过程中的一个核心要素,指信任方对被信任方履行其承诺或提供的信息进行依赖。在认知信任中,这种依赖不仅是行动上的,也是认知上的,用户对自动驾驶汽车驾驶决策的信任体现了对该技术的依赖。信心是信任方对被信任方将如预期行动的确信度,这种确信基于对被信任者过去行为的观察、评价和预期的一致性。一个长期提供准确预测的 AI 天气预报系统,能够使用户对其预报内容持有高度信心。AI 大语言模型信任涉及的相关主体不仅包含用户与 AI,还包括技术研发机构、研发人员、机构管理者等。因此,信念、沟通、依赖和信心等认知信任的四个组成部分也将在各相关主体的关系之间有不同的含义与呈现。

2.技术信任与人际信任动态交互

对大语言模型的认知信任不仅包含了原有的技术信任,还包括了“人—技术—人”关系之中的人际信任。技术信任通常指人们在使用某项技术工具时,不仅要求该技术工具能够满足他的实际需要,而且需要确定技术是安全可靠的,强调个体对技术性能的可靠性、稳定性和预期表现的信任。在这种信任中,技术被视为一个静态对象,其信任度基于技术的历史表现和预期的未来性能。然而,AI大语言模型技术极其复杂且具有黑箱属性,技术专家也无法完全确定其内部的工作原理,更何况普通用户。只有信任才是应对技术不确定和不能控制未来的至关重要的策略[15]。

以可信AI框架为基础,大语言模型具有稳健性,在帮助人们完成任务这一点被认为是可靠的,但此时对大语言模型的技术信任却不是静态的、确定的,抑或是在人们无法充分理解技术的情况下充满了随机性。人—机—环境交互系统的主客体界限常常模糊,具有个别性、人为性、异质性、不确定性、价值与事实的统一性、主客相关性等特点,其中充满了复杂的随机因素的作用,不具备重复性,因此在充满变数的人—机—环境交互系统中,存在的逻辑是与各种可能性保持互动的同步性[16]。比如,人们在与ChatGPT对话时,在不同时间问同样的问题,每次得到的答案都不是完全一样的,这便是由于模型的随机性、训练过程中数据的多样性、输入的变化以及模型运行环境变化等多重因素导致的。因此,用户与大语言模型之间的信任建立存在一个相互调节的过程[17],不仅要求大语言模型自身的可信度,还需要用户给予信任,即用户主观感知与大语言模型客观能力的调节匹配。

在“人—技术—人”关系框架中,用户主观感受、给予信任无法隔绝人际信任产生的影响。人际信任是基于个人之间的关系,涉及对他人善意、正直、可靠性和能力的信任。与技术信任不同,人际信任是一个动态的交互过程,它依赖于持续的社会互动和沟通。人际信任建立在个体间相互了解和信任的基础上,包括了对他人行为的预期和对其意图的理解。例如,一个团队成员可能会基于过去的合作经验和对方展现的诚信来信任其他成员。在该关系框架中,人际关系不仅包含人与人之间的关系,更涉及技术背景下人与人之间的关系。在使用者、研发者、管理者等之间的人际信任中,大语言模型技术本身已经成为一个重要的变量深度介入人际信任之中[18]。在充满随机性的环境中,人与大语言模型之间的信任需要人际信任来支撑,人与人之间的信任也面临被调节甚或被技术导引、规制的情形。

对大语言模型的认知信任是技术信任与人际信任的综合,即人对技术的心理预期与技术自身效能的一种混合交互、相互调节的结果。人在多大程度上信任人工智能,该对人工智能分配多少信任。在这种模式中,人们对AI的信任不仅基于技术本身的性能和可靠性,也涉及对技术背后人类研发团队的信任,包括开发者的意图、责任感和能力。这种信任是动态的,因为它需要考虑到技术的发展、学习能力、在不同情境下的表现以及用户与技术之间的互动和沟通。

3.基于有效监督具有合理性的信任

人与人之间的人际信任可以是无监督的,即无条件信任、盲目信任,但人对大语言模型的认识信任如果是无监督的,可能导致两种不良情形:一是没有可信度的信任,即错误的信任。大语言模型能够帮助人们提升生活工作效率,但是它的智能化、自动化特性可能导致人类过度依赖机器,进而侵蚀人的自主性。典型的便是ChatGPT的幻觉问题。它生成一个看似正确实则错误的答案,让过度信任AI的人产生幻觉,将事实上错误的答案当作是正确的。同时,还存在一种情况,就是开发大语言模型的组织往往标榜其系统是可信任的,而用户由于该组织过去有较高的商誉而缺乏谨慎态度,选择信任这些组织。但可能由于技术的局限性导致这些大语言模型并不值得信任,那么人们就会因为错误信任而引发潜在风险。二是值得信赖但不被信任,即不合理的不信任。如果技术悲观主义者或风险偏好较为保守的人们不信任实际上值得信赖的技术,就会承担错失利用好技术的机会成本。对新兴技术的不信任不仅限制了其在现有领域的应用,还可能抑制未来创新的动力。当开发者和研究人员感觉到公众对他们的工作持怀疑态度时,可能会减少对具有潜在的革命性技术的研发和投入。

要保障建立对人工智能合理的信任,人工智能开发及应用的组织必须克服两个关键问题:一是信息问题,组织需要可靠且一致地评估AI系统是否值得信任,以提供人们是否应该信任它的证据基础;二是沟通问题,组织需要将他们的证据传达给其他用户,并在适当的复杂性水平上转译这些证据,以便他们可以相应地决定是信任或是不信任。另外,还可以建立如AI审计师等第三方保障供应商去促成信

任的建立[19]。

三、大语言模型信任的建构路线

建构大语言模型信任的可行路径主要包括信任要素、信任主体和信任环境、信任认知三个方面。

1.搭建以可解释性为核心的信任构成要素

建立对大语言模型的信任应当包括技术信任的信任要素和人际信任的信任要素。其中,技术信任主要包括功能性、可靠性、安全性、可解释性等要素;人际信任主要包括机构商誉、企业社会责任和形象、公平性、企业 AI 治理体系和制度等。在诸多信任要素中,需要以可解释性作为核心,它不仅是技术本身的可解释性,也是信任的可解释性。可解释性作为价值载体,内含透明性、安全性、可理解性、可责性和公正性,并指向技术的可验证、可审核、可监督、可追溯、可预测与可信赖等多个方面,连结与贯通技术信任、人际信任的要素体系,促进技术信任与人际信任的动态交互。托马拉和霍夫曼也曾提出作为一种“看不见的假设”的道德信任,委托人与受托人的信任基于共同的权利与义务。“看不见的假设”被转化为语言陈述的过程涉及三方面:需要提高信任的可解释性、涉及可理解的信任形式以及期望实现的形式是否可被接受。提高信任的可解释性意味着受托人与委托人的关系高度透明,对其意向、善意等有深入认识[20]。

通过 AI 系统的可解释性,可以增加用户的信任。也就是说,AI 大模型只有有效地解释自己,才能取得用户的信任,从而产生高效的人机协作[21]。根据利益相关者的动机、意图和要求不同,AI 解释内容、方法/模型也应不同,应建立基于相关利益者分类的可解释性框架[22]。这里涉及:由谁解释?一般为大语言模型的技术提供者;向谁解释?向利益相关者包括用户、政府和社会等进行解释;解释什么?主要包括技术原理、功能等专业性解释,价值与影响等社会普遍关注的价值性解释,涉及安全解释、责任解释、公平解释、影响解释等具体方面;如何解释?各种解释需融入 AI 大模型全生命周期。

目前以可解释性为核心要素建立信任的路径已经获得广泛认同。2023 年,谷歌、微软、OpenAI 等公司签署人工智能安全自愿承诺协议,其中,重点就是建立公众对开发该技术的公司的信任以及它们的工作方式和收集的信息是透明的。协议规定了提供有关其产品功能和局限性的明确信息等旨在建立公众信任的一系列义务。

2.以政府主导的 AI 治理体系为基础建构多主体协同的信任环境

任何信任的建立都离不开一个良好的信任环境。信任环境包括一系列的社会、文化、经济和技术因素,这些因素共同作用,形成了个体或集体之间建立信任的基础[23]。建构大语言模型的信任则需要以政府主导的 AI 治理体系为基础,多元主体共同参与,以达到有效监督。一是政府治理和监管对建立大语言模型的信任至关重要。有实证研究显示,人们相信当监管工具到位时,人工智能更值得信任[24]。目前各国和地区加紧出台相关法律法规和治理政策,比如,2024 年 3 月,欧盟出台了全球首部人工智能立法《人工智能法案》,旨在为可信人工智能发展设立独特的法律框架。二是多元主体参与的治理机制和工具将为大语言模型的信任建立提供有效抓手。政府的治理与监管受限于技术和人力投入,因此需要行业、企业以及第三方机构等多元主体参与。2021 年英国政府发布了《建立有效的人工智能认证生态系统的路线图》,计划在未来 5 年培育一个世界领先、市场规模达到数十亿英镑的 AI 认证行业,通过中立第三方的 AI 认证服务(包括审计、影响评估、认证等)来评估 AI 系统的可信合规性。我国在《科技伦理审查办法(试行)》中也明确提出建立人工智能的伦理审查制度,由独立的伦理委员会开展伦理审查,确立人工智能的研发与应用主体责任与问责流程。行业协会和科技企业也积极开展自律举措,践行负责任的人工智能发展理念,努力赢得用户信任。

3.培养人们对 AI 大语言模型的信任认知能够合理分配信任

建立对大语言模型的信任,不仅需要技术研发与应用组织通过提升技术可信度和机构可信度来积极争取,还需要用户基于对技术的认知并给予合理的信任。人们对 AI 系统的认知和理解程度与能够给予信任密切相关[25]。理解和培养人们对 AI 系统的信任是一个重大挑战,因为人们缺乏对技术的理解和认知[26]。

用户作为大语言模型技术的使用者,也是技术风险的直接承担者。教育和提高公众对 AI 技术的理解至关重要。用户需要能够客观地评估技术的优势和潜在风险,避免因盲目信任或过度怀疑而做出不合理的决定。正确的认知可以帮助用户在理解技术的同时,对其潜在的影响和风险有一个合理的评

估。同时,信任是一个认知可能风险,并且愿意接受由此带来损失的过程。建立和维持信任需要一个持续的、动态的过程,旨在平衡信任和对风险的认识。而人们往往是风险厌恶型,需要建立正确的信任观。这就需要社会传播普及关于大语言模型的信任与不信任的问题与案例,消除公众对信任的误解。

参考文献

[1]闫宏秀.用信任解码人工智能伦理[J].人工智能,2019(4):95-101.

[2]BRAUN M, BLEHER H, HUMMEL P. A leap of faith: is there a formula for "Trustworthy" AI? [J]. The Hastings Center Report, 2021(3):17-22.

[3]IBM 商业价值研究院.生成式 AI 市场现状[R/OL].(2023-07-14)[2023-12-15].https://www.ibm.com/downloads/cas/7KX4Q06G.

[4]陶大程.可信 AI 的前世今生[J].智能系统学报,2021(4):601.

[5]中国信息通信研究院.可信人工智能产业生态发展报告 2022[R/OL].(2022-09-06)[2023-12-15].http://221.179.172.81/images/20220906/3741662451673713.pdf.

[6]香港消费者委员会.道德与信心共融促进电子商务人工智能发展[R/OL].(2022-09-02)[2023-12-16].https://www.consumer.org.hk/f/initiative_detail/415531/435406/AI%20Study%20-%20Presentation%20(Chinese).pdf.

[7]KPMG. Trust in Artificial Intelligence: A global study[R/OL].(2023-02-24)[2023-12-16].https://assets.kpmg.com/content/dam/kpmg/au/pdf/2023/trust-in-ai-global-insights-2023.pdf.

[8]REINHARDT K. Trust and trustworthiness in AI ethics[J]. AI and Ethics, 2023(3):735-744.

[9]DAUKAS N. Epistemic trust and social location[J]. Episteme, 2006(1-2):109-124.

[10]JONES C, BERGEN B. Does GPT-4 pass the turing test? [J]. arxiv preprint arxiv: 2310.20216, 2023.

[11]IHDE D. Technology and the lifeworld: from garden to earth[M]. Bloomington and Indianapolis: Indiana University Press,1990:72.

[12]WILHOLT T. Epistemic trust in science[J]. The British Journal for the Philosophy of Science, 2013(2):233-253.

[13]ALVARADO R. What kind of trust does AI deserve, if any? [J]. AI and Ethics, 2023(4):1169-1183.

[14]MCCRAW B W. The nature of epistemic trust[J]. Social Epistemology, 2015(4):413-430.

[15]什托姆普卡.信任:一种社会学理论[M].程胜利,译.北京:中华书局,2005:32.

[16]郭雷.系统科学进展 3[M].北京:科学出版社,2023:123-158.

[17]闫宏秀,宋胜男.智能化背景下的算法信任[J].长沙理工大学学报(社会科学版),2020(6):1-9.

[18]闫宏秀.负责任人工智能的信任模塑:从理念到实践[J].云南社会科学,2023(4):40-49.

[19]GOV.UK. Introduction to AI assurance[EB/OL].(2024-02-12)[2024-02-12].https://www.gov.uk/government/publications/introduction-to-ai-assurance/introduction-to-ai-assurance#ai-assurance-in-context.

[20]TUOMELA M, HOFMANN S. Simulating rational social normative trust, predictive trust, and predictive reliance between agents[J]. Ethics and Information Technology, 2003(3):163-176.

[21]朱松纯."为人文赋理,为机器立心",朱松纯教授在图灵大会阐释"通用智能·人机共生"[EB/OL].(2023-08-11)[2024-01-12].https://mp.weixin.qq.com/s/JYyCI49hDV3KalPButaoRg.

[22]MCEERMID J A, JIA Y, PORTER Z, et al. Artificial intelligence explainability: the technical and ethical dimensions[J]. Philosophical Transactions of the Royal Society A, 2021(2207): 20200363.

[23]FUKUYAMA F. Trust: the social virtues and the creation of prosperity[M]. New York: Simon and Schuster, 1996:90.

[24]王俊美.人工智能获得信任的前提是监管到位[N].中国社会科学报,2023-03-06(A03).

[25]HOFFMAN R R, MUELLER S T, KLEIN G, et al. Metrics for explainable AI: challenges and prospects[J]. Frontiers in Computer Science,2023(5):1-15.

[26]AKULA A R, WANG K, LIU C, et al. CX-ToM: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models[J]. Iscience, 2022(1):1-29.

The Trust Construction of Large Language Models

Hu Xiaomeng Chen Liyuan Liu Zhengyuan

Abstract: The rapid rise of AI large language model (LLM) technologies, represented by ChatGPT, is revolutionizing content production and intelligent technology paradigms. However, issues such as hallucinations and false content have led to a crisis of trust, with the technology even facing resistance and bans. Despite the industry's active technical practices in the field of trustworthy AI, public trust in AI has not significantly improved. Therefore, to address the trust issue, it is not only necessary to clarify the relationship between trust and trustworthiness but also to consider the technological essence of LLMs. As an epistemic technology, trust in LLMs should be epistemic trust, which includes the dynamic interaction of technical trust and interpersonal trust, and is based on reasonable trust established through effective supervision. Accordingly, a trust construction route for LLMs is proposed, which mainly includes a trust element system with being explainable at its core, a trust subject and environment with a government-led AI governance system as the basis and the collaboration of multi-subjects , and the trust cognition that cultivates a correct trust perspective among people.

Key words: artificial intelligence; large language models; trust; trustworthy

责任编辑:白 杨